Black Box Attack on Machine Learning Assisted Wide Area Monitoring and Protection Systems

Milan Biswal, Satyajayant Misra, and Abu S. Tayeen
Computer science, New Mexico state University
Las Cruces, NM, USA 88001
Email: milanb@nmsu.edu, misra@cs.nmsu.edu, tayeen@nmsu.edu

Abstract—**The applications for wide area monitoring, protection, and control systems (WAMPC) at the control center, help with providing resilient, efficient, and secure operation of the transmission system of the smart grid. The increased proliferation of phasor measurement units (PMUs) in this space has inspired many prudent applications to assist in the process of decision making in the control centers. Machine learning (ML) based decision support systems have become viable with the availability of abundant high-resolution wide area operational PMU data. We propose a deep neural network (DNN) based supervisory protection and event diagnosis system and demonstrate that it works with very high degree of confidence. The system introduces a supervisory layer that processes the data streams collected from PMUs and detects disturbances in the power systems that may have gone unnoticed by the local monitoring and protection system. Then, we investigate compromise of the insights of this ML based supervisory control by crafting adversaries that corrupt the PMU data via minimal coordinated manipulation and identification of the spatio-temporal regions in the multi-dimensional PMU data in a way that the DNN classifier makes wrong event predictions.**

Keywords: Black box attack, wide area monitoring systems, adversarial machine learning, PMU data analytics.

I. Introduction

The wide area monitoring, protection, and control (WAMPC) system is a critical monitoring infrastructure for the power transmission system in the smart grid. The secure, reliable, resilient, and economical operation of the bulk power system is primarily dictated by the WAMPC applications [1]. The availability of real time synchrophasor data from a group of phasor measurement units (PMUs) have enabled many critical WAMPC applications, such as stability monitoring, wide area control, and supervisory protection [2]. The abundance of high resolution measurements paves the way for the data-driven intelligent applications. Further, the revolutionary advancements in *machine learning* [3] have shown potential to harness the data, to create highly reliable and efficient systems.

Previous research elaborates the concept of a supervisory protection and event diagnosis framework based on machine learning, for the WAMPC system [4]. The application can be deployed at the control center and identifies dynamic disturbances in the power systems by processing the data from multiple PMUs in a given region. This forms a supervisory layer that automatically detects any mis-operation that goes undetected by the local monitoring and protection infrastructure [5]. Depending on the type of the event and the state of the system, either a manual or automated response is initiated, enhancing the resilience of the grid. One such application utilized a machine learning classifiers and a set of features that were chosen with the knowledge of the domain [4].

Deep neural networks (DNN) have been quite successful in solving complex challenges in a range of domains, including the smart grid [3]. Deep convolutional neural networks (CNN) belonging to the family of DNN, have a distinct advantage of automatic feature extraction and classification for multidimensional data (so far, widely used for images). To the best of our knowledge, in the power system domain, despite this automatic extraction advantage of CNNs and their ready applicability to WAMPC, CNNs have not been studied.

Motivation and Contributions: Motivated by this, we propose a deep CNN classifier to efficiently classify different disturbances in the power systems. This classifier is trained on a set of ground truths (disturbance data with labels). Then, the trained classifier is deployed as a WAMPC application to perform disturbance identification and supervisory protection task. We demonstrate the efficiency of the proposed deep CNN classifier with a simulated dataset of PMU data comprising of different power system disturbances.

However, our major contribution is probing the susceptibility of this deep CNN classifier to adversarial perturbations. We present a procedure for careful crafting of PMU data manipulation attacks that can *mislead* our effective classifier. We discuss this crafting technique to craft adversaries that perform minimal manipulation of data transmitted by a small, select subset of PMUs in a way that the resultant event classification of the classifier becomes incorrect. This type of attack is a *black box attack*, where the adversary does not have any knowledge of the classifier and does not have any access to the training data, but conjectures to identify the important components of the data using simulated data and then proceeds to perturb the real readings to result in the misclassification.

The rest of the paper is structured as follows. In Section II, we discuss the background on the PMU infrastructure and their vulnerability to cyber attacks. In Section III, the methodology and the proposed architecture of the WAMPC application is presented. An overview of black box attack on machine learning models is presented in Section IV. The experimental setup and data creation are explained in Section V. In Section VI, the performance of our machine learning model is evaluated under adversarial attack. Section VII concludes the paper with

978-1-7281-3103-0/20/$31.00 ©2020 IEEE

a discussion of the scope for future work.

II. Background

The PMUs are primarily placed in the transmission system at the generating stations, major junctions, and at substations. However, they can also be placed in the distribution grid or in any bus of interest. Essentially, a group of PMUs in an area are connected to a regional phasor data concentrator (PDC), and a set of PDCs are connected to a super PDC, to form a hierarchical information-flow network that conveys the complex dynamics of the power systems [6]. The PDC is either housed at the control center or transmits data to the nearest control center, which hosts the WAMPC applications. The measured PMU data and control commands are transmitted over a wide area communication network between the PMUs and their corresponding control centers. The IEEE recommended the *C37.118* standard for PMU data communication, which relies on TCP/IP protocol for flexible and timely data delivery, but the applications still remain susceptible to cyber-attacks [7].

A. PMU data manipulation attacks

The PMU data manipulation attack (PDMA) is a type of cyber attack where an adversary can alter some of the PMU data. The resultant data manipulation may affect the WAMPC applications at the control centers by introducing errors in the control or decision support system [8]. This sophisticated cyber attack utilizes rich domain knowledge to carefully craft organized adversaries that remain blind to data integrity check algorithms and cause damage to the power grid without getting detected. This is the same as the false data injection attack (FDIA), in which an adversary stealthily compromises measurements from electricity grid sensors (e.g. PMUs) in a coordinated fashion, and is capable of evading detection by the power system bad data detection module.

The 2015 Ukraine Blackout is a classic example of an adversarial opening that can lead to FDIA [9]. The adversaries loaded malicious firmware into the communication network field gateway devices. The password-protected access to substation control center and ICS network was likely gained via keystroke loggers. The control center data manipulation is not difficult to achieve after gaining credentialed access. It can be foreseen that the adversaries can locally manipulate device parameters and corrupt measurements by building direct access to field devices, such as PMUs. The man-in-the-middle attacks can intercept the measurements and maliciously manipulate their values by infiltrating into the communication networks [10]. The success of Stuxnet attack rings the alarm that even private wide-area networks or closed networks with strict confidentiality are vulnerable to cyber security threats [11].

Although most gross errors and outliers in the PMU measurements can be detected by bad data detection during state estimation, one can still circumvent it by maliciously selecting a set of measurements with an intent to compromise [12]. Once these PMU measurements are manipulated, the resulting corrupted data may mislead the control centers to take unwanted actions (or not take any actions), resulting in outage events of varying intensities.

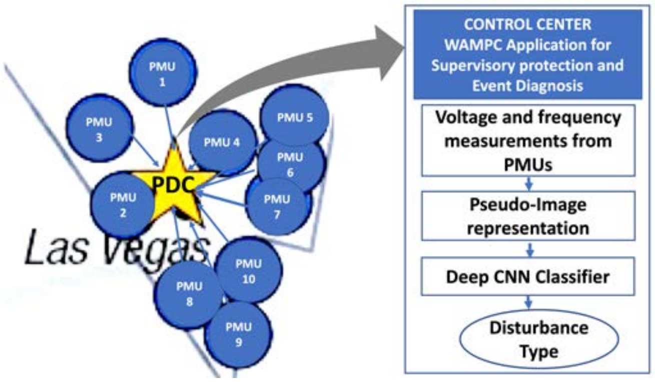

PMU 1
PMU 3
PMU 4
PMU 5
PMU 6
PMU 2
PDC
PMU 7
Las Vegas
PMU 10
PMU 8
PMU 9
CONTROL CENTER WAMPC Application for Supervisory protection and Event Diagnosis
Voltage and frequency measurements from PMUs
Pseudo-Image representation
Deep CNN Classifier
Disturbance Type

Fig. 1: Illustration of the proposed framework for the WAMS application in the WECC system, the actual location of the PMUs in Las Vegas area shown over a cropped portion of the NASPI PMU location map [13].

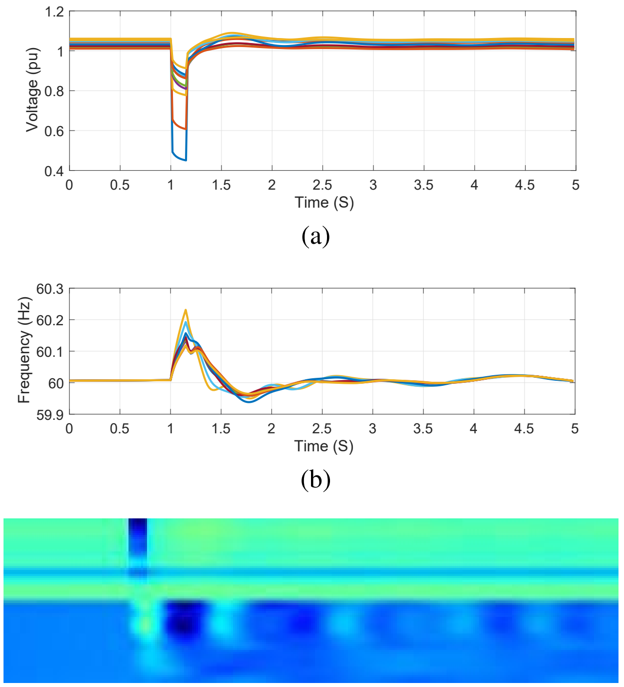
III. Framework and Methods

The framework for the proposed WAMPC application for supervisory protection is shown in Fig. 1. A set of PMUs in a geographical area (around Las Vegas) are connected to a local PDC. The control center has access to the data stored in the PDC and this data is fed as input to the application. The raw PMU data were subjected to pre-processing and bad data removal.

A. Data Preparation

The data streams from multiple PMUs are combined to form a two-dimensional matrix representation. Although the PMUs measure current, voltage, and frequency, the proposed decision support system in the WAMPCS relies on the voltage and frequency waveform to recognize disturbances. Motivated by the significant reliability of deep CNN in classifying images [14], [15], we create pseudo-color images of the two dimensional PMU data matrix. The spatio-temporal PMU data voltage and frequencies were individually normalized and quantized to 256 intensity levels of each red, green, and blue colors. The pseudo-color images for the voltage and frequencies were appended along with the spatial coordinates to create an image that maps the spatio-temporal PMU voltage and frequency data.

The result of the conversion of spatio-temporal PMU data to pseudo-images from raw PMU data through colormap is illustrated in Fig. 2, where the graphs in Figs. 2a and 2b represent the voltage and the frequency data of 10 PMUs proximate to an event sampling at a rate of 60 frames per second, for a duration of 5 seconds, and the third figure (Fig. 2c) represents the corresponding colormap obtained by putting the continuous readings of each PMU as a row of the colormap. The resultant color pseudo-images were saved as JPEG files which were fed as input to the deep CNN classifier.



(a)

(b)

(c)

Fig. 2: The PMUs' (a) voltage and (b) frequency waveforms corresponding to a 3-phase fault. The corresponding colormap image representations of (c) simulated PMU data

B. Deep CNN Model

The DNN is composed of a sequence of layers that are individually parameterized by a set of weights and each layer consists of units called neurons. The DNN for classification of different types of data can be conceptualized as a nonlinear functional mapping that takes an input and returns a class label. The non-linearity in the network is introduced by the activation function associated with each neuron. The CNN is a type of feed forward DNN that exploits the hierarchical pattern in data and decomposes complex patterns into smaller, simpler and sparser patterns. CNNs are particularly suitable for analyzing images, as they automatically extract reduced dimensional features and feed them to the following layers of the network to identify different classes of images. We have adopted an LeNet [16] type CNN architecture with a convolutional layer (with 5×5 kernels and ReLU activation functions), followed by a max-pooling layer, and a two-layer fully connected network with softmax activations, as shown in Fig. 3.

IV. Adversary Crafting

A. Threat and Attack Model

Cyber-security of a system is based on three major pillars, confidentiality, integrity, and availability (CIA) [17]. We assume that the adversary launches a malicious attack by intelligently modifying the PMU measurements, violating the integrity. Authentication and non-repudiation of information are critical to maintain the integrity in the smart grid [18] and Section II-A briefs the threats that may compromise these aspects.

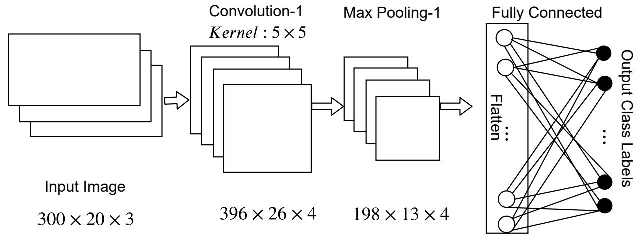


Fig. 3: Architecture of the deep CNN classifier

The attack assumptions made in this paper are based on the susceptibility of PMU measurements to adversarial manipulation. There are four approaches to manipulating PMU measurements: (i) compromising PMUs locally with either physical access or by manipulating the transducer readings; (ii) intercepting and forging data packets during their transit to the control center; and (iii) modifying data at control center database [19]. (iii) modifying data at control center database [19]. Note that if an attacker gains physical access to the PMUs, the sensed data can be manipulated at the output of the voltage and current transducers or by infecting the PMU firmware.

(iv) In a more realistic situation, an attacker may start with an Internet Protocols (IPs) scan to identify all the connected hosts and to find any open or vulnerable ports that are accessible and can help the attacker get access to the PMU [20]. The attacker may also use phishing [21] and password pilfering attacks [22] to gain access to the control center. *Modbus* and distributed networking protocol 3.0 (DNP3) are two communication protocols used for PMU data communication, which are vulnerable to scanning attacks [23]. Modbus transmits data in plaintext, with no encryption, making the data vulnerable to sniffing or tampering by hackers.

B. Adversary crafting

The weights and biases of the CNN are determined during the training phase in a way that minimizes a cost function. A DNN model can be expressed as a multidimensional function: $F : X \mapsto Y$, where X is the input data and Y is an output vector of the class labels. The goal of the adversary is to craft an input $X^* = X + \delta x$ by introducing a perturbation δx, that minimizes the following objective function.

$$\arg\min_{\delta x} \|\delta x\| \; s.t. \; F(X + \delta x) = Y^* \qquad (1)$$

Here, Y^* is the new class label specified by the adversary. The adversaries craft attacks by finding δx that such that the DNN does not predict the actual class of the event (predicts something else, or cannot differentiate it from the norm). The fast-gradient sign method efficiently computes the adversarial perturbations for a given classifier [24]. However, it sometimes lead to sub-optimal solutions and does not provide a close approximation of the optimal perturbation. The signed change in derivative is the foremost method for crafting adversaries, however it does not ensure minimal perturbation of the input. The *DeepFool* algorithm alleviates these limitations and computes optimal perturbations that can alter the class label [25]. Therefore, in this paper, we have used the *DeepFool* to craft adversaries.

Our major objective is to craft adversaries that could mislead the ML classifier to predict incorrect class labels from the data. The adversaries are crafted over the PMU data pseudo-images. The change in a pixel value of the image by the adversary corresponds to the change in the data point recorded by the PMU that maps to that that pixel.

V. Simulation

A. Data creation

In order to carry out a comprehensive evaluation, we created a simulated PMU dataset with sufficient number of disturbance events under various operating conditions. We used the positive sequence load flow (PSLF) dynamic simulation tool [26] from General Electric (GE) for creating simulated PMU events corresponding to different event classes. In order to simulate actual operating conditions of the grid, we used a 2008 Heavy Summer load flow base case for the WECC system [5]. A total of 299 time-synchronized voltage and frequency probes were placed in the WECC model at key buses, which approximately correspond to the actual location of PMUs in the WECC system. The data rate for the PMUs were fixed at 60 frames per second. The experimental setup is explained in a greater details in [5]. For the experiments in this paper, we considered 3 different types of events, fault, loss of generation, and synchronous motor switching off.

Test cases for fault were simulated at the buses with base voltage greater than 120 kV. The fault duration was kept 0.15 s. For simulating the loss in generation the generators above 300 MW were switched off at the trigger point. Synchronous motors were switched off one at a time to create events labeled as synchronous motor-off.

The voltage phasors and frequency readings corresponding to each disturbance event were recorded by all the 299 PMU probes. Not all the PMUs in the network will capture the prominent disturbance signatures. The PMUs closer to the disturbance will capture event signatures, which fade with increasing electrical distance from the source, more faithfully compared to the farther PMUs. Therefore, we considered the 10 PMUs with the most strong signature, that are close to the disturbance. In the real world scenario, each control center will have access to one or more PDC, which aggregates measurements from a group of PMUs in its geographic area.

We consider a disturbance pattern length of 5 s—0.5 s before the trigger and 4.5 s after the trigger. Voltage and frequency data streams from 10 PMUs at a sampling rate of 60 frames per second, were aggregated to form pseudo color images, as mentioned in Sec III-A. Each image has a dimension of $[300 \times 20 \times 3]$ comprising of 300 time points, 10 voltage and 10 frequency measurements, and 3 fundamental color intensities. Each image represents the instance of a disturbance. The data-set consisted of 344 instances of faults, 140 instances of loss of generation, and 21 instances of synchronous motor switching events (a total of 505 images).

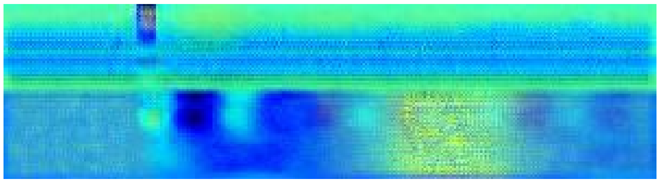


Fig. 4: The color-map pseudo-image representations of the simulated PMU data with adversarial perturbation.

TABLE I: Case I: Classification accuracy for the binary classifier

Class	Accuracy (%)	Accuracy after adversarial attack (%)
Fault	100	89.2
Loss of Generation	100	93
Synchronous motor - OFF	94	82

VI. Results and Interpretation

The data-set comprising of the pseudo-images were fed as input to the CNN classifier. In order to simulate the black box setting, we implemented the CNN classifier in two different environments, *Keras* [27] and *Tensorflow* [28]. We trained the CNN using the Keras back-end, then generates adversarial images and used them to attack a CNN trained on TensorFlow. As the attacker does not have access to the parameters of the *TensorFlow* model, this setup represents a black box attack. We used 10-fold cross validation to assess the performance of the classifier. The data set was divided into 10-equal groups with elements chosen at random. At each fold, 9 groups were combined to form the training set and one group was used for testing. The reported accuracy is an average of the 10-folds.

We conducted two sets of experiments. In the first case, we considered a binary classifier, that is trained on a set of disturbance and steady state data. The goal of the adversary in this case was to create perturbations such that the the disturbance waveform is classified as a normal steady state phenomenon. The classification performance without and with adversarial crafting are summarized in Table I. We note that the results without adversarial crafting illustrates the use of our deep CNN classification framework for classification of disturbances. The high accuracy in non-adversarial scenario demonstrates the efficacy of our framework in classification. It can be observed that the minimal adversarial perturbations can blind the classifier to some of the disturbances.

In the second case, we crafted attacks such that the true class label of a disturbance was mis-classified as another disturbance, e.g., a fault is classified as a generation loss or synchronous motor switching off event. The overall classification accuracy of the CNN classifier was 93.8%, with adversarial manipulations the accuracy dropped to 82%. Although the reduction in accuracy is only 11% that still means that on an average 35 events were misclassified, which in our opinion is 35 too many due to the critical nature of the events (total events studied was 344). The masking or mis-classification of a crucial event like a fault may result in the control actions that are not desired and may lead to operational disruptions and cascaded failures. Fig. 4 shows a pseudo-image corresponding to fault data shown in Fig. 2, with adversarial perturbation, that is predicted as a generation loss by the deep CNN classifier.

VII. Conclusion and Future Work

We presented a machine learning assisted WAMPC application for disturbance detection in the transmission grid. The data driven control center application processes PMU measurements in a geographical area and is based on the deep CNN classifier. The vulnerability of the deep CNN classifiers to adversarial manipulations in the PMU data, was shown. We demonstrated a black box attack on the deep CNN classifier, where an adversary modifies the input PMU data selectively. The selective manipulation of the PMU data is dictated by its crafted adversarial pseudo-image representation. The change in pixel values of the image directly correspond to the spatio-temporal modifications of the PMU data. We demonstrated an intelligent adversary crafting technique that can attack sophisticated machine learning based decision support systems in the smart grid without being detected.

The future work will focus on a more complete evaluation with other disturbances in the power system such as load switching and capacitor switching. The other focus will be on techniques to detect the adversarial manipulations and on enhancing robustness of the machine learning models deployed in WAMPC systems.

References

[1] A. Ashok, M. Govindarasu, and J. Wang, "Cyber-Physical Attack-Resilient Wide-Area Monitoring, Protection, and Control for the Power Grid," *Proceedings of the IEEE*, vol. 105, no. 7, pp. 1389–1407, July 2017.

[2] "The Modern Grid Strategy: A Vision for the Smart Grid," june 2009.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[4] M. Biswal, Yifan Hao, P. Chen, S. Brahma, H. Cao, and P. De Leon, "Signal features for classification of power system disturbances using pmu data," in *2016 Power Systems Computation Conference (PSCC)*, June 2016, pp. 1–7.

[5] M. Biswal, S. M. Brahma, and H. Cao, "Supervisory Protection and Automated Event Diagnosis Using PMU Data," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1855–1863, Aug 2016.

[6] "Use of synchrophasor measurements in protective relaying applications," Report of PSRC Working Group, 2012. [Online]. Available: http://www.pes-psrc.org/Reports/UseofSynchrophasorMeasurementsinProtectiveRelayingApplications_final.pdf

[7] "IEEE standard for synchrophasor data transfer for power systems," pp. 1–53, Dec 2011.

[8] J. Wang, D. Shi, Y. Li, J. Chen, H. Ding, and X. Duan, "Distributed framework for detecting pmu data manipulation attacks with deep autoencoders," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 4401–4410, July 2019.

[9] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 ukraine blackout: Implications for false data injection attacks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3317–3318, July 2017.

[10] Y. Yuan, Z. Li, and K. Ren, "Quantitative analysis of load redistribution attacks in power systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1731–1738, Sep. 2012.

[11] D. Kushner, "The real story of stuxnet," *IEEE Spectrum*, vol. 50, no. 3, pp. 48–53, March 2013.

[12] X. Liu, Z. Bao, D. Lu, and Z. Li, "Modeling of local false data injection attacks with reduced network information," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1686–1696, July 2015.

[13] "NASPI PMU map March 2017," 2017. [Online]. Available: https://naspi.org/node/749

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[16] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] V. Y. Pillitteri and T. L. Brewer, "Guidelines for smart grid cybersecurity," NIST Interagency/Internal Report (NISTIR), Tech. Rep. 7628 Rev 1, 2014.

[18] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on cyber security for smart grid communications," *IEEE Communications Surveys Tutorials*, vol. 14, no. 4, pp. 998–1010, Fourth 2012.

[19] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong, "A review of false data injection attacks against modern power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1630–1638, July 2017.

[20] E. D. Knapp and R. Samani, *Applied cyber security and the smart grid: implementing security controls into the modern power infrastructure*. Newnes, 2013.

[21] H. Holm, W. R. Flores, and G. Ericsson, "Cyber security for a smart grid - what about phishing?" in *IEEE PES ISGT Europe 2013*, Oct 2013, pp. 1–5.

[22] A. Stefanov and C. Liu, "Cyber-power system security in a smart grid environment," in *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*, Jan 2012, pp. 1–3.

[23] K. Coffey, R. Smith, L. Maglaras, and H. Janicke, "Vulnerability analysis of network scanning on scada systems," *Security and Communication Networks*, vol. 2018, 2018.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

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

[26] GE, "PSLF." [Online]. Available: https://www.geenergyconsulting.com/practice-area/software-products/pslf

[27] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[28] A. Martín et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/