

Dynamic Detection of False Data Injection Attack in Smart Grid using Deep Learning

Xiangyu Niu

Department of Electrical Engineering
and Computer Science
University of Tennessee, Knoxville
Knoxville, Tennessee 37996
Email: xniu@vols.utk.edu

Jinyuan Sun

Department of Electrical Engineering
and Computer Science
University of Tennessee, Knoxville
Knoxville, Tennessee 37996
Email: jysun@utk.edu

Abstract—Modern advances in sensor, computing, and communication technologies enable various smart grid applications. The heavy dependence on communication technology has highlighted the vulnerability of the electricity grid to false data injection (FDI) attacks that can bypass bad data detection mechanisms. Existing mitigation in the power system either focus on redundant measurements or protect a set of basic measurements. These methods make specific assumptions about FDI attacks, which are often restrictive and inadequate to deal with modern cyber threats. In the proposed approach, a deep learning based framework is used to detect injected data measurement. Our time-series anomaly detector adopts a Convolutional Neural Network (CNN) and a Long Short Term Memory (LSTM) network. To effectively estimate system variables, our approach observes both data measurements and network level features to jointly learn system states. The proposed system is tested on IEEE 39-bus system. Experimental analysis shows that the deep learning algorithm can identify anomalies which cannot be detected by traditional state estimation bad data detection.

I. INTRODUCTION

The future smart grid is designed to operate more reliable, economical and efficient in an environment of increasing power demand. This goal, however, is achieved by incorporating with a tremendous increase of data communications which lead to great opportunities for a various of cyber attacks. Thus, ensuring cyber security of the Smart Grid is a critical priority. Although a large number of countermeasures have been published, such as communication standards (e.g. IEC 61850-90-5 [1]), regulation laws (e.g. Colorado Regulations (CCR) 723-3), cryptographic implementations (e.g. secure channel [7]), and official guidelines (e.g. NISTIR 7628 Guidelines [8]), current smart grid still remains vulnerable to cyber attacks.

To prevent cyber attacks, legacy grid relies on traditional security scheme (e.g., firewall and general intrusion detection system). Intrusion detection systems (IDS) are able to generate alarms for potential intrusions by consistently monitoring network traffic or system logs. Although there are a number of studies on general IDS in network security community, limited effort has been made specifically to smart grid. At the same time, the risk of attacks targeting data availability and integrity in the power networks is indeed real. A notable work is by Liu et al. [16], which proposed a type of attacks, called false data injection (FDI) attacks, against the state

estimation in the power grid. In such attacks, the attackers aim to bypass existing bad data detection system and pose damage on the operation of power system by intentional changing the estimated state of the grid systems. Therefore, there is an urgent need of effective smart grid specific intrusion detection systems.

To address the above issues, two schemes have been widely studied to detect FDI attacks [4], [14]: One way is to strategically protect a number of secure basic measurements. Kim et al. [12] propose a greedy algorithm to select a subset of base measurements and the placement of secure phasor measurement units. Bi et al. [3] characterize the problem into a graphical defending mechanism to select the minimum number of meter measurements which cannot be compromised. The other way of defending FDI attack is to verify each state variables independently. Liu et al. [15] formulate a low rank matrix separation problem to identify attacks and propose two optimization methods to solve the problem. Ashok et al. [2] present an online detection algorithm that utilizes statistical information and predictions of the state variables to detect measurement anomalies.

Recently, machine learning algorithms have been broadly adopted to the smart grid literature for monitoring and preventing cyber attack of power systems. Ozay et al. [17] generate Gaussian distributed attacks and use both supervised and semi-supervised machine learning methods to classify attacks. Similarly, Esmalifalak et al. [5] devise a distributed support vector machines based model for labeled data and a statistical anomaly detector for unsupervised learning cases. He et al. [10] employs Conditional Deep Belief Network (CDBN) to efficiently reveal the high-dimensional temporal behavior features of the unobservable FDI attacks. However, existing works mainly focus on finding bad measurement at certain state, no prior studies have been conducted over the dynamic behavior of FDI attack. Besides, detecting FDI attacks is considered as supervised binary classification problem in [5], [10] which are incapable of detecting dynamically evolving cyber threats and changing system configuration.

Recent breakthrough in GPU computing provide the foundation for neural network to go "deep". In this paper, we develop an anomaly detection framework based on neural

network to enable the construction of a smart grid specific IDS. More specifically, a recurrent neural network with LSTM [11] cell is deployed to capture the dynamic behavior of power system and a convolutional neural network [13] is adopted to balance between two input sources. An attack is alerted when residual between the observed and the estimated measurements is greater than a given threshold.

Moreover, attackers with sophisticated domain knowledge may continually manipulate the power grid state estimation without being detected causing extensive damages. As such, we want to bridge the gap between network anomaly detector and FDI attacks detection mechanism. Unlike other works which separate two detectors, our framework combines both network traffic characteristics and time-series data measurements with help of convolution neural network to equalize between two inputs. With the help of the proposed neural network structure, our anomaly detector demonstrates highly accurate detection performance.

We organize the rest of this section as follows: Section II introduces the background of FDI attack and neural network. Section III presents our combined detection system along with the static and dynamic method to detect FDI attack in Section III-B and Section III-C. Section IV presents the case study on IEEE 10-machine 39-bus power system. Finally, we conclude our work in Section V.

II. BACKGROUND

A. False Data Injection Attack

In a power system, the state is represented by bus voltage magnitudes $V \in \mathcal{R}^n$ and angles $\theta \in ([-\pi, \pi])^n$, where n is the number of buses. Let $z = [z_1, z_2, \dots, z_m]^T \in \mathcal{R}^m$ be the measurement vector, $x = [x_1, x_2, \dots, x_n]^T \in \mathcal{R}^n$ be the state vector, and $e = [e_1, e_2, \dots, e_m]^T \in \mathcal{R}^m$ denote the measurement error vector. We describe the AC measurement model as follows:

$$z = h(x) + e \quad (1)$$

In analyzing the impact of data attack on state estimation, we adopt the DC model obtained by linearizing the AC model where the relationship between these m meter measurements and n state variables can be characterized by an $m \times n$ matrix \mathbf{H} . In general, the matrix \mathbf{H} of a power system is a constant matrix determined by the topology and line impedances of the system.

$$z = \mathbf{H}x + e \quad (2)$$

Typically, a weighted least squares estimation is used to obtain the state estimate as $\hat{x} = \min_x \frac{1}{2}(z - \mathbf{H}x)^T \mathbf{R}^{-1}(z - \mathbf{H}x) = (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1} z$ where \mathbf{R} is the covariance matrix.

Let z_a represent the vector of observed measurements that may contain malicious data. z_a can be represented as $z_a = z + a$ where $a = (a_1, \dots, a_m)^T$ is the malicious data added to the original measurements. Let x_a denote the estimates of x using the malicious measurements z_a . Then x_a can be

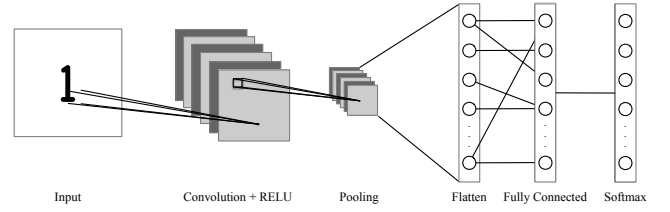


Fig. 1. An example of architecture for image classification with convolutional neural network.

represented as $\hat{x} + c$, where c is a non-zero vector representing the impact on the estimate from the malicious injection and \hat{x} is the estimate using the original measurements. In this paper, for target FDI attackers, we assume the attacker has enough inside information to constructing x_a while random FDI attackers only have partial information.

B. Convolutional Neural Network

Convolutional Neural Networks (CNNs or ConvNets) are a category of Neural Networks that have been successful in processing image and video signal such as identifying objects in real time video and style transfer for images as visualized in Figure 1.

CNNs describe the most classic form of neural network where multiple nodes are arranged in layers such that information only flows from input to output. We use three main types of layers to build a CNN architecture: Convolutional Layer, Pooling Layer, and Fully-Connected Layer.

In this way, CNN transforms the original input layer by layer from the original tensor to the final output which can be class score. In particular, each convolutional layer and fully connected layers perform transformation that is a function of both the parameters (weights and biases) and activations in the input volume. The parameters in the CNN will be trained with gradient descent optimization algorithm to minimize the loss between the outputs that the CNN computes and the labels of training dataset.

C. Recurrent Neural Network

Recurrent neural networks, or RNNs [19], are a family of neural networks for processing sequential data. Contrasting from convolutional network which is specialized for processing high dimensional tensors such as image, a recurrent neural network is a neural network that is specialized for processing time-series values $x^{(1)}, \dots, x^{(t)}$. RNNs are ideal for long sequences without sequence-based specialization.

Recurrent neural networks use the following equation to define the values of hidden units.

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta) \quad (3)$$

where h represent the state and $x^{(t)}$ refers the time-series input at time t .

Schuster et al. [20] shows a bi-directional deep neural network. At each time-step t , bi-directional RNN maintains two hidden units, one for the forward propagation and another

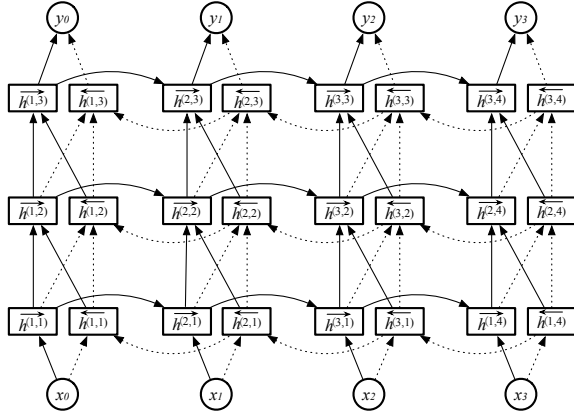


Fig. 2. Computation of a typical 3 layers bidirectional recurrent neural network.

for the backward propagation. The final result, y_t , is generated through combining the score results produced by both hidden units. Figure 2 shows the bi-directional network architecture, and (4) show the formulation of a single bidirectional RNN hidden layer.

$$\vec{h}(t) = f(h^{(t-1)}, x^{(t)}; \vec{\theta}) \quad (4)$$

$$\overleftarrow{h}(t) = f(h^{(t+1)}, x^{(t)}; \overleftarrow{\theta}) \quad (5)$$

1) *Long Short-Time Memory*: Currently, the most commonly implemented RNNs fall into the class of long short-time memory (LSTM) neural networks [11]. Different from vanilla RNN with single gate, LSTM exhibits notable performance gain for preserving long time dependencies while also keeping short time memories. Each LSTM cell involve three gates to which are input gate i , output gate o , the forget gate f . The information flow of LSTM cell is as follows:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (6)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (7)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (8)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \quad (9)$$

$$h_t = o_t \circ \sigma_h(c_t) \quad (10)$$

where $\sigma_g(\cdot)$ and $\sigma_c(\cdot)$ represent the sigmoid and tangent function respectively, and \circ denotes the element-wise product. Here, c and h stand for the cell state vector and hidden unit vector.

III. DETECTING DATA LEVEL FALSE DATA INJECTION ATTACK

Various research on static FDI attack detection method has been published. A common assumption is a threat model where the attackers have knowledge of the power system topology; however, can only inject a limited number of bad data points which is shown in Figure. 3a. In this threat model, FDI attack can be mitigated if a proportion of the comprised substation is below a certain threshold. Moreover,



(a) Limited power FDI attacks

(b) Unlimited power FDI attacks

Fig. 3. Different scenario for dynamic FDI attacks.

data measurements are often redundant for estimating the actual state. This threat model is widely adopted in existing works. Nonetheless, we stress this threat mode by: 1) removing the limitation of the number of measurement data that are corrupted; and 2) assuming the attackers have basic understanding of the aforementioned static detection mechanism in (11).

Figure 3b shows the dynamic FDI attack that is focused in this work. The attack starts at $t = 3$ and the measurements of both bus 2 and 3 have been compromised. Static method will fail in this scenario, for the reason that two thirds of the measurements have been modified from $t = 3$ to $t = 6$. A sophisticate attacker can deliberately generate a false event based on a real event and inject it to the power grid. As the result, it is unlikely to detect this attack only based on static method which can cost catastrophe results if control center makes false actions.

A. The Combined Attack Detection Method

In this section, we provide an overview of our proposed system for detecting FDI attacks in Figure 4. Our proposed detection mechanism mainly consists of a static detector and a deep learning based detection scheme. The static detector can be an State Estimator (SE) or any aforementioned FDI attack detector [2], [3], [5], [10], [15], [17], [12] which is built independently beyond our dynamic detector. As mentioned in the previous section, the dynamic detector takes two input sources. While the data level features are explicit, the network packages are captured by tcpdump and each network packet includes header and data payload, with unique features which are defined in NSL-KDD dataset [22]. The NSL-KDD dataset has 41 features which are categorized into three types of features: basic, content based and traffic-based features. It should also be mentioned that some features are generated based on a fixed window (default is 2 second) which will remain consistent within the window.

Our dynamic detector is employed to recognize the high-level time-series features of the FDI attacks. To achieve this goal, our time-series method consists of two essential mechanisms: offline training and online detection. The offline training is trained based on historical measurement and can be potentially facilitated by outsourcing to public machine learning cloud services. Unlike other methods which are designed under the assumption that the physical status of the power system does not change overtimes, our system will collect real-time measurement data to support offline training and the prediction model will update after retrain is completed.

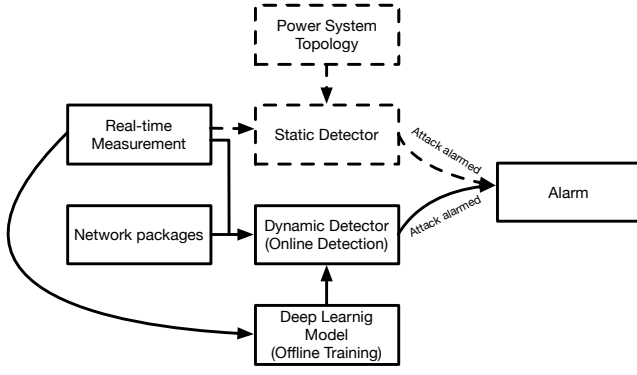


Fig. 4. The overview of our proposed deep learning based FDI attacks detection system.

B. State Estimator Method

In bad data injection attack, one method is to use Chi-squares test. Once bad data are detected, they need to be eliminated or corrected, in order to obtain the correct states. There are two important hypothesis that are the largest normalized residual (LNR) and the $J(\hat{x})$ performance index:

$$J(\hat{x}) < \tau \quad (11)$$

where $J(\hat{x})$ follows a chi-square distribution and τ is a preset threshold. The threshold can be obtained from the χ^2 distribution. If $J(\hat{x}) > \tau$, bad data will be suspected. For DC model, $\text{diag}(\sigma_i^2, 0) = I$, the traditional bad data detection approaches often reduce to l_2 -norm of the measurement residual [16]:

$$\|z - H\hat{x}\|_2 < \tau \quad (12)$$

C. Dynamic Detection Method

In [9], the authors formulate the bus voltage magnitudes, angles and states of measuring devices together as system states in Markov Decision Process (MDPs). In our method, we extend to a recursive model where the decision not only depends previous one state but previous n states where the loss is as follows:

$$\eta = L(\phi(s_t), f(\phi(s_{t-1}, \dots, s_{t-n-1}), \theta), \tau) \quad (13)$$

where ϕ, θ are parameters need to be turned and τ is the threshold that is needed to decide whether the attack has been started.

Figure 5 shows the structure of our stacked dynamic detection model. Specifically, the input of the model are the time-serious power system data, the features will be passed to several LSTM layers to learn high dimensional temporal features. Previous works [6], [10] characterize FDI attacks as a binary classification problem which looks promising in the experimental setting, since the datasets to be tested can be manually tuned for different scenarios. In real world implementations, power system data is highly unbalanced, thus,

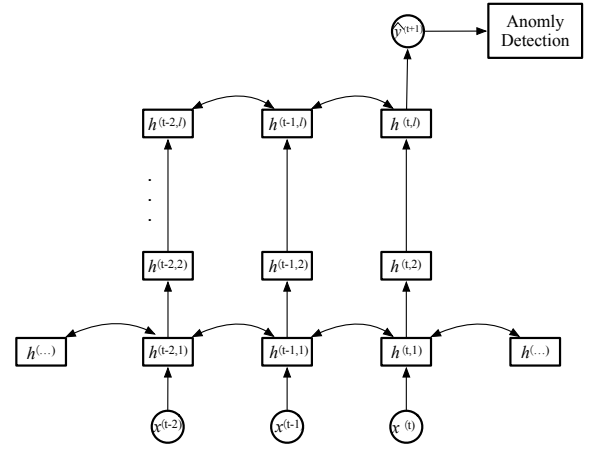


Fig. 5. The time-serious dynamic detection method based on RNN.

binary classification methods will inevitably have low recall even the overall accuracy is high. However, for evaluating IDS, recall is often more important than accuracy since any cyber attack can cost catastrophe results.

In general, our dynamic anomaly detector takes the input of time-series $\dots, x^{(t-1)}, x^{(t)}, \dots$, learn their higher dimensional feature representations, and then use those features to predict the next data point $\hat{x}^{(t)}$. Furthermore, the predicted data point can be used to classify if $x^{(t)}$ is anomalous by checking the similarity between the actual data $x^{(t)}$ and predicted data $\hat{x}^{(t)}$.

Having presented single source FDI attacks detection model, we now introduce a framework that combines FDI attacks detector with network intrusion detection system. This framework is dealing with a case when an IDS that relies on data measurement fails to detect the start of FDI attacks. Accordingly, if the fabricated injection data are derived from a legitimate measurement in our threat model, data level detectors may fail to determine if current network is intruded or not. In this case, to increase the overall performance of time-series anomaly detection model, a combined attack detection method is proposed in this paper.

Specifically, the schematic structure of the combined framework is given in Figure 6. As seen in the figure, the combined framework is rather straightforward. An alternative method to combine data level information and packet level features is directly concatenate the input vector. However, because the dimension between the data measurements and network packet features differ significantly, direct concatenation may have minimal improvement than aforementioned time-serious methods. Alternatively, each level features are transformed by a convolutional neural network before concatenation as shown in the figure. The purpose of adding additional convolutional neural network is to equalize the dimension between data measurement and packet level features and their respect weights are learned using gradient descent (Adam algorithm is used in our experiments). Inception deep learning architecture [21] is advised when possible.

convolutional neural network and recurrent neural networks. Our model learns normal behavior from normal data and is unrelated to certain attack, and thus can detect unseen attacks. Additionally, our two-level detector is robust using hybrid features and can detect attack when state vector estimator fails. We provided key insights about various factors that impact the performance of the proposed algorithm. We presented a detailed case study of the proposed algorithm on the IEEE 39-bus system.

REFERENCES

- [1] IEC 61850-90-5. Communication networks and systems for power utility automation part 90-5: use of iec 61850 to transmit synchrophasor information according to iec 61850-90-5, 2012.
- [2] Aditya Ashok, Manimaran Govindarasu, and Venkataramana Ajjarpur. Online detection of stealthy false data injection attacks in power system state estimation. *IEEE Transactions on Smart Grid*, 2016.
- [3] Suzhi Bi and Ying Jun Zhang. Graphical methods for defense against false-data injection attacks on power system state estimation. *IEEE Transactions on Smart Grid*, 5(3):1216–1227, 2014.
- [4] Rakesh B Bobba, Katherine M Rogers, Qiyan Wang, Himanshu Khurana, Klara Nahrstedt, and Thomas J Overbye. Detecting false data injection attacks on dc state estimation. In *Preprints of the First Workshop on Secure Control Systems, CPSWEEK*, volume 2010, 2010.
- [5] Mohammad Esmalifalak, Lanchao Liu, Nam Nguyen, Rong Zheng, and Zhu Han. Detecting stealthy false data injection using machine learning in smart grid. *IEEE Systems Journal*, 2014.
- [6] Cheng Feng, Tingting Li, and Deepthi Chana. Multi-level anomaly detection in industrial control systems via package signatures and lstm networks. In *Dependable Systems and Networks (DSN), 2017 47th Annual IEEE/IFIP International Conference on*, pages 261–272. IEEE, 2017.
- [7] Mostafa M Fouda, Zubair Md Fadlullah, Nei Kato, Rongxing Lu, and Xuemin Sherman Shen. A lightweight message authentication scheme for smart grid communications. *IEEE Transactions on Smart Grid*, 2(4):675–685, 2011.
- [8] NIST Smart Grid. Introduction to nistir 7628 guidelines for smart grid cyber security. *Guideline*, Sep, 2010.
- [9] Yingshuai Hao, Meng Wang, and Joe H Chow. Likelihood analysis of cyber data attacks to power systems with markov decision processes. *IEEE Transactions on Smart Grid*, 2016.
- [10] Youbiao He, Gihan J Mendis, and Jin Wei. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. *IEEE Transactions on Smart Grid*, 8(5):2505–2516, 2017.
- [11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [12] Tung T Kim and H Vincent Poor. Strategic protection against data injection attacks on power grids. *IEEE Transactions on Smart Grid*, 2(2):326–333, 2011.
- [13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [14] Gaoqi Liang, Junhua Zhao, Fengji Luo, Steven R Weller, and Zhao Yang Dong. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, 8(4):1630–1638, 2017.
- [15] Lanchao Liu, Mohammad Esmalifalak, Qifeng Ding, Valentine A Eme-sih, and Zhu Han. Detecting false data injection attacks on power grid by sparse optimization. *IEEE Transactions on Smart Grid*, 5(2):612–621, 2014.
- [16] Yao Liu, Peng Ning, and Michael K Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1):13, 2011.
- [17] Mete Ozay, Inaki Esnaola, Fatos Tunay Yarman Vural, Sanjeev R Kulkarni, and H Vincent Poor. Machine learning methods for attack detection in the smart grid. *IEEE Transactions on Neural Networks and Learning Systems*, 27(8):1773–1786, 2016.
- [18] MA Pai, T Athay, R Podmore, and S Virmani. Ieee 39-bus system. 1989.
- [19] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.
- [20] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. *Cvpr*, 2015.
- [22] Mahbod Tavallaei, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. A detailed analysis of the kdd cup 99 data set. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*, pages 1–6. IEEE, 2009.