Learning-Based Defense of False Data Injection Attacks in Power System State Estimation

Arnav Kundu
Electrical Engineering
Texas A&M University
arnav1993k@tamu.edu

Abhijeet Sahu
Electrical Engineering
Texas A&M University
abhijeet_ntpc@tamu.edu

Katherine Davis
Electrical Engineering
Texas A&M University
katedavis@tamu.edu

Erchin Serpedin
Electrical Engineering
Texas A&M University
eserpedin@tamu.edu

***Abstract*—The electric power grid has evolved immensely with time and the modern power grid is dependent on communication networks for efficient transmission and distribution. Since communication networks are vulnerable to various kinds of cyber attacks it is important to detect them and prevent an important machinery like the power grid to get affected from cyber attacks. False data injection attacks (FDIA) are one of the most common attack strategies where an attacker tries to trick the underlying control system of the grid, by injecting false data in sensor measurements to cause disruptions. Our work has focused towards Least Effort attacks of two types i.e., Random and Target Attacks. Further, we propose a data augmented deep learning based solution to detect such attacks in real time. We aim at generating realistic attack simulations on standard IEEE 14 architectures and train neural networks to detect such attacks.**

***Index Terms*—False Data Injection, State Estimation, Random and Target attacks, Long Short Term Memory (LSTM)**

I. Introduction

The power system is a dynamic and complex system connecting diverse electrical components such as generators, transmission lines and distribution systems. To ensure reliable operation of such a complex system, we require secure system monitoring of systems comprising of synchrophasor, Current Transformers (CTs), Potential Tranformers (PTs), etc. State variables, like voltage and phase angles at each bus are estimated from these measurements, and the system operator controls the estimated state to operate the grid. With the use of a state estimator and its associated analysis for contingency, a system operator can review each critical contingency, in order to determine whether each possible future state is within reliability limits, and make decisions regarding its operation. But with the fusion of advanced cyber infrastructure in the physical domain, the measurements are easily altered by the cyber invaders, affecting the process of state estimation and misleading the power grid control system, resulting in catastrophic consequences.

False Data Injection Attacks (FDIAs) can be introduced in a transmission system to trick the state estimator into predicting wrong states without getting detected [1]. Detection based methods try to find anomalies in the data received through the communication channel. Such methods depend on the real-time correlation between data points or the temporal structure of the data to classify a new set of measurements as anomalous. A significant drawback of this approach is that it does not adapt well to changing patterns in transmission behavior over time. FDIAs are challenging to detect using conventional residue based methods since they do not capture the spatial or temporal structure of the measurement data available. However, with the current case of FDIAs, we can ensure that bad data can be injected even by keeping low residual error. This is a classic contextual anomaly detection problem. Deep learning has shown significant promises in solving complex tasks and has been used in pattern recognition problems like object detection, speech recognition, and anomaly detection.

Deep learning uses a data-driven approach where a function approximator is trained using gradient descent over a given set of data points. The success of deep learning can be attributed to the ability of neural networks to learn complex functions and the availability of massive data-sets. Motivated by its application and success in the field of speech recognition, time-series prediction, and anomaly detection, we explore how recurrent neural networks can be applied to detect false data injection attacks in the electric power grid.

Artificial Neural Networks (ANNs) have shown significant performance in representing complex functions [2]. With the advent of Graphics Processing Units (GPUs) and availability of massive data-sets, neural networks have helped to solve complex problems in the fields of object recognition [3], speech recognition [4] and anomaly detection [5]. Especially in anomaly detection, deep neural networks have been applied in many applications like fraud detection, sensor network anomaly detection, video surveillance, log anomaly detection and Internet of Things (IoT). Deep neural networks have been used in supervised [6], semi-supervised [7] and unsupervised setting [8] in the past for anomaly detection. Specifically for anomaly detection in spatially and temporally correlated data, direct supervision using classification networks and unsupervised methods using auto-encoders have shown impressive results in the past. Generative methods like Generative Adversarial Networks (GANs) [9] and Variational Auto-Encoders (VAEs) [10] have also been explored. Both of them are unsupervised methods where neural networks are trained to learn the latent distribution of non-anomalous data. GAN uses a discriminator to judge whether a new set of data points are different from the old set of data on which it was trained on. VAE uses the reconstruction error of the test set of points to find anomalies.

The results are validated in these dynamic attacks through simulation, utilizing both IEEE 14 and 118 based test system. For the first case, we utilize a modified version of the 14-bus system, while utilizing real time data, for a period of a year study. The simulation results demonstrate that the attacker can

978-1-7281-0407-2/19/$31.00 ©2019 IEEE

systematically and efficiently construct different attack vectors in real time for false data injection attacks at any instance of time, making them impossible to detect by conventional methods. The proposed deep learning based defense technique has not been introduced previously.

The organization of the rest of the papers sections are as follows: Section II introduces the background on state estimation and FDI attacks. Section III elaborately explains two types of FDI attacks: Random and Target Attacks. Section IV presents the defense algorithm proposed for detecting the FDI attacks. Section V presents the attack and defense case results and analysis on the IEEE 14 bus system. Finally, section VI concludes our paper with scope of future work.

II. Background

The electric power grid uses a set of measuring devices spread across various branches in order to determine the state of the system. These states are then used to take necessary control actions. However, the true state of the system cannot be directly determined from the measuring devices because of induced noise and measurement inaccuracies. Therefore, a Kalman filter is used to determine the correct state of the grid using those measurements. Such an estimation mechanism is described by the equation 1:

$$z = Hx + \epsilon \tag{1}$$

where z denotes the measurement vector, x represents the state vector, H stands for the system characteristic matrix and ϵ is the error in estimation. The objective is to find a state vector x that minimizes the energy (variance) of residue ϵ defined as

$$\min_{x} 1^T (z - Hx)^2 \tag{2}$$

In conventional state estimators, for a new state vector to be considered as a correct state the residue should be below a defined threshold. In a false data injection attack, an adversary aims to hack the readings of multiple measurements to mislead the state estimator to predict incorrect states without affecting the residue. These attacks can be random without any particular motive [1] or targeted to certain state variables with specific intentions [11].

It has been established by some researchers that such attacks can be prevented totally by securing a subset of all measuring devices completely on a encrypted network [12] but as the size of the network increases the number on devices that needs to be secured increases, hence it is not scalable. The basic residual based detection can also be improved by using L_∞-norm instead of L_2-norm [13].In most of the prior work it has been assumed that the attacker has complete knowledge of the system. However, in a practical scenario it can be assumed that an intruder will not be aware of the entire power system and will only have the information of a part of it. Recently, it has been shown that FDIAs can be possible with partial system information as well [14].

The residual based detection systems fail to consider the spatial distribution of the measuring devices and temporal distribution of the measurements. This is an important information that can be used to derive spatio-temporal correlation between measurements which can be used to detect attacks. In a superficial sense the problem can be reduced to detecting anomalies in a dense graph. Inspired by various classical machine learning applications in cyber intrusion, sensor networks and image processing, researchers have tried to apply nearest neighbor classifiers and other statistical classification techniques [15]. However, these methods are slow for huge systems and have a nonlinear run time complexity. In addition, these models do not scale well with respect to the size of the network and cannot be applied effectively to power grids [15]. With the current advancements in deep learning and sequential pattern recognition we propose a deep learning based anomaly detection system to detect and identify various kinds of intrusions.

III. Attack Designs

The basic concept behind FDIAs is very simple, i.e., to generate an attack vector a such that 3:

$$z + a = H(x + c) + \epsilon \tag{3}$$

where c is the change in states induced due to the attack vector. We have experimented on generation of two types of FDIAs.

A. Random Attacks

One of the simplest attacks is least effort random attack where an attacker with access to a fixed set of compromised measuring devices tries to bias random state variables. This attack can be possible only if the attacker has access to all the meters. However, in a real scenario it is not feasible for attacker to get hold of all the measuring devices in a network. As a result we cannot choose any random attack vector. The attacks were generated following the methods described in [1] with some conditions as described in that paper. It can also be logically inferred that the probability of generating a random attack increases if we have access to more meters.

B. Targeted Attacks

In a targeted attack the attacker wants to control particular state variables. We denote the states affected by such an attack as follows:

$$I_{\text{states}} = \{i_1, \ldots, i_k\} \text{ where } k < n \tag{4}$$

This set denotes the state variables to be attacked. The objective of the attacker is to inject an attack state vector c such that $\hat{x}_{bad} = \hat{x} + c$ where $c = (\ldots c_1, \ldots, c_k \ldots)^T$. We consider two cases of attacks over here: a constrained case and an unconstrained case. A constrained case is one where we assume that the injected attack does not affect any other state apart from the targets. In the unconstrained case it is assumed that the attacker doesn't care about the impact of his attack on other state variables apart from his targets. The attacks were designed following the algorithm proposed in [1].

IV. Defense Mechanism

As mentioned earlier, the state estimator relies on simple Euclidean distance based anomaly detection mechanisms to recognize incorrect measurements. We have proven that such a system is easy to trick and therefore the spatio-temporal correlation of these measurements need to be accounted in our new FDIA detection system.

In [15], a correlation based FDIA detection mechanism has been proposed, where a semi-supervised structure is employed.

An operator needs to define a correlation sphere for various meters on the network. A single meter might lie in multiple correlation spheres. This approach ensures that the spatio-temporal correlation between the measurements are preserved. At every iteration, correlations within a correlation sphere are calculated and if a huge divergence is found then an anomaly is flagged. This method is highly efficient in terms of run-time complexity but would need humongous effort in designing the correlation spheres manually. In addition this method will not allow adaptive changes to network topologies.

In [16], a few more approaches based on sparse optimization, low rank matrix factorization and nuclear norm minimization have been explained. The assumption here is that the gradually changing state variables will typically lead to a low rank measurement matrix Z_0 and the attack matrix (attack vectors over time) is sparse. Therefore, the problem translates to a matrix separation problem as

$$\min_{\mathbf{Z}_0,\mathbf{A}} \operatorname{Rank}(\mathbf{Z}_0) + \|\mathbf{A}\|_0 \tag{5}$$

s.t. $\mathbf{Z_a} = \mathbf{Z}_0 + \mathbf{A}$

which can be formulated as a convex optimization problem as:

$$\min_{\mathbf{Z}_0,\mathbf{A}} \|\mathbf{Z_o}\|_* + \lambda\|\mathbf{A}\|_1 \tag{6}$$

s.t. $\mathbf{Z_a} = \mathbf{Z}_0 + \mathbf{A}$

$\|\mathbf{Z_0}\|_*$ is the nuclear norm of Z_o, i.e., the sum of singular values of Z_o. This kind of optimization problem has been studied across the domains of compressive sensing and matrix completion and can be solved using off the shelf optimization algorithms. The problem with this approach is the computational complexity because of its iterative nature [16]. This paper also proposes a faster way using low rank matrix factorization where low rank matrix Z_0 is represented as product of two matrices U and V. Even though this approximation helps to remove the expensive SVD step, it is iterative in nature, which is not linear in time.

We propose a deep learning based data driven FDIA detection method which is robust, has an almost linear run-time complexity. Recurrent neural networks are heavily used to capture temporal correlation in data, see e.g., [17], [18]. In addition to addressing variable length sequences they also help to keep the predictor small and are computationally light because of shared parameters.

A. Approach 1

In our first approach we define a recurrent neural network for the entire grid which will take in the actual measurement values in time from all the available meters and combine them into the hidden recurrent state. The last recurrent state is used to determine the status of all the meters. In the actual scenario, we use an advanced version of recurrent neural network called Long Short Term Memory (LSTM) to prevent vanishing and exploding gradients [19]. The output of this network can be represented as:

$$\begin{aligned} y_t &= \sigma(W_o h_t) \\ &= \sigma\left(W_o\left(LSTM\left(W_i z_t, W_h h_{t-1}\right)\right)\right) \end{aligned} \tag{7}$$

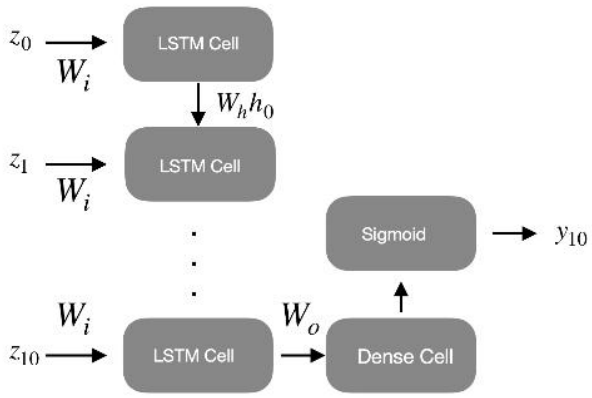

Fig. 1. Architecture of the network.

where $y_t \in R^{m\times 1}$ is the vector of probabilities of m measuring devices to be under attack at time t, $z_t \in R^{m\times 1}$ is the vector of measurement values at time t, σ is the sigmoid activation function, W_o is the weight of the neural network . This forms a recursive chain, where the characteristics of the dynamic process is captured by the weights W_h. W_i captures the mutual information of the nodes or for a graph as in our case it might be thought of as a representation of the adjacency matrix parameterized by the weights of each measurement. In this approach we are enabling the network to learn the dynamics of the process as well as the similarity matrix of the measurements. However, this approach would need the number of input measurement devices for the entire system to be fixed and therefore is not scalable easily.

B. Approach 2

In the second approach we define a similar network as discussed earlier but we do not take all measurements as inputs. Instead we select a set of measuring devices which are connected on the actual power grid graph. In this way we are enforcing the spatial arrangement of the devices on the network. Therefore, the major learning happens in the temporal domain. This is a distributed approach and therefore can be scaled easily. The output of this network can be represented as:

$$\begin{aligned} y_t^i &= \sigma\left(W_o h_t^i\right) \\ &= \sigma\left(W_o\left(LSTM\left(W_i z_t^i, W_h h_{t-1}^i\right)\right)\right) \end{aligned} \tag{8}$$

where $y_t^i \in R^{k\times 1}$ is the vector of probabilities of k measuring devices to be under attack at time t, $z_t^i \in R^{k\times 1}$ is the vector of measurement values at time t for the i^{th} region with k devices in space.

In addition our architecture should be able to detect attacks even when all measurements are not accessible. This might occur due communication network failures and such opportunities can be used by intruders to attack the system. The system should also be able to detect attacks on measurements which have not occurred in the training set.

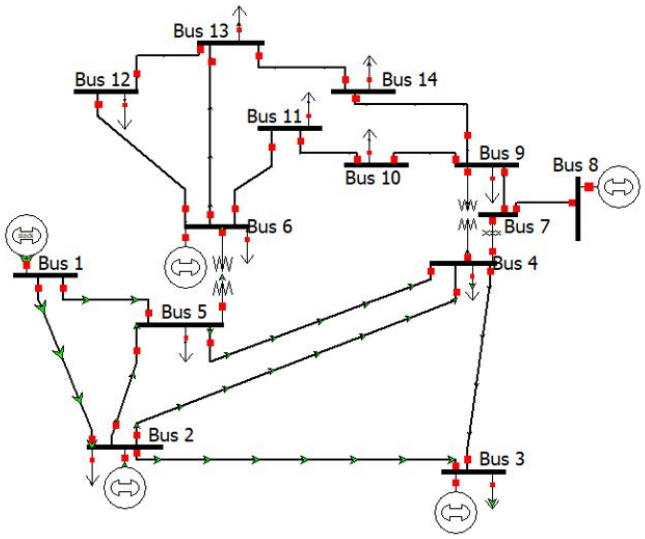

Fig. 2. Topology for the IEEE 14 Bus Case

V. Results and Analysis

A. IEEE 14 Bus Case

The simulation uses real world power consumption data to generate 39 measurements and the intrusion state of these devices at each time step. The methodology for the bus level modeling is based on the synthetic load model proposed in [20]. This model has an hourly basis specification, including residential, commercial and industrial sectors loads and its results were validated utilizing ERCOT Data. The topology for the model developed contains 14 buses, 2 generators, 3 synchronous condensers, and 11 loads. It also contains a three winding transformer equivalent. The final 14 bus topology is presented in Fig. 2.

The data generated has a 5 minute resolution data, as a basis design and specification, for an entire year, which gives 105397 time steps of SCADA data for the test case. For each timestep, a state estimation model is solved utilizing the power flow equations. In order to obtain a flawless resolution of 5 minutes in our test case, a piecewise polynomial algorithm based on the cubic spline extrapolation methodology is utilized to correct the missing measurements in the grid.

1) Attack generation: We design random and targeted attacks on these measurements to affect the state variables. The attack data is stored along with the state variables under attack and the devices compromised. This is treated as the training data for our deep learning models.

For the random FDI attacks, the attacker is trying to observe its influence on the probability of finding an attack vector by modifying the number of meters . The IEEE 14 bus system with 39 measurements and 26 state variables to estimate is considered for the case study. The experiment is repeated for 105397 timestamps. As it can be observed from Fig. 3, as the number of meters increased, the probability of finding an attack vector increases. At n = 2, it can be seen that it has a probability of attack at around 0.27, while at n = 11, it is possible to have a higher probability of almost 1. From the equation, k $\geq$ m-n+1, it is observed that if k = 11, then

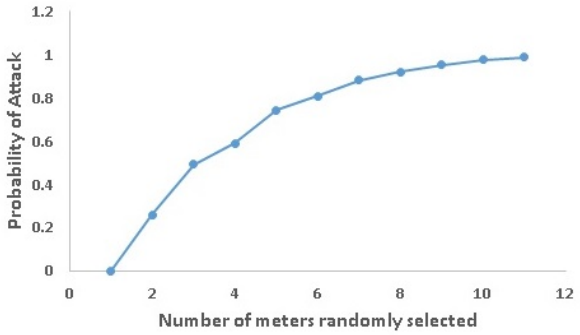

Fig. 3. Probability of finding an attack vector with varying number of meters compromised

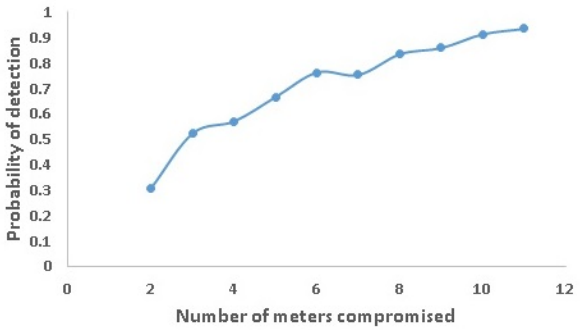

Fig. 4. Probability of detection using conventional Chi square Test with level of significance α = 0.005

to perform a successful attack with 26 state variables, 36 measurements are needed at least.

To check the efficiency of the Random attack, the Chi-square test is implemented, which is a conventional bad data measurement detection algorithm. As per the chi-square test, the minimum threshold to prevent the attack detection is 29.8, for degree of freedom 13, which is the difference between the number of meters and the state variables, with level of significance α set to 0.005. While keeping this threshold, the probability of detection is evaluated, for varying number of meters compromised by the attacker as shown in Fig. 4.

For the targeted FDI attack, the objective of the attacker is to attack a specific state variable. So the targeted attack is performed, with different numbers of target state variables. It is observed that the probability of attack required at least 15 meters to be compromised, to perform attack on 3 state variables. For 31 meters, the probability of attack is 1, if the attacker is aiming at 3 target variables. As the number of target state variables is reduced, the number of meters required to be compromised reduces. For example, as the target variables become 1, the distribution of the meters ranged primarily from 2 to 7, as shown in Fig. 5.

2) Preprocessing the data for training: The attack data needs to be formatted for the training process. This step splits the data into sequences over a rolling window for training. The data is already in per unit scale therefore, normalization is not necessary. This split in sequences is important because we are using a recurrent framework which needs features at

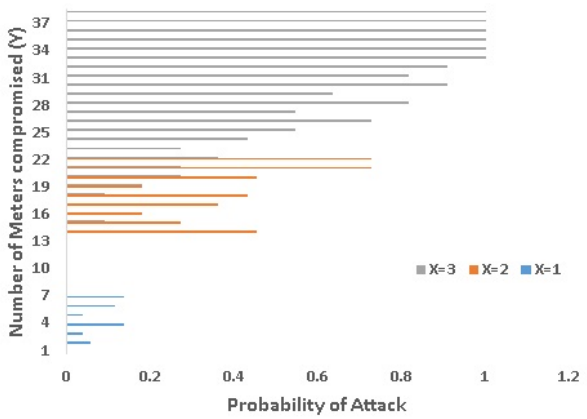

Fig. 5. Probability of a successful attack when Y meters are considered to target X number of state variables

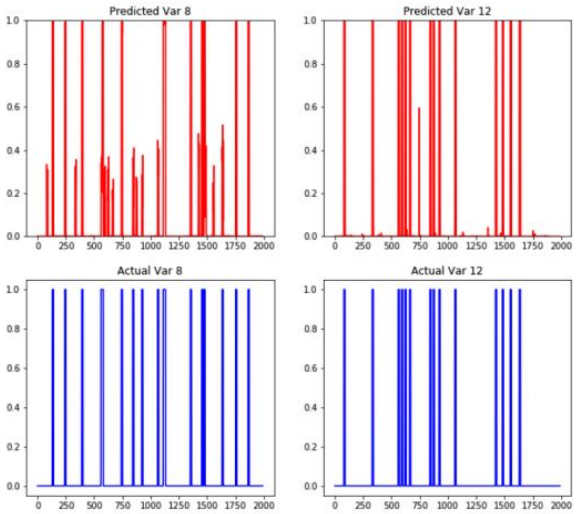

Fig. 6. Predicted vs Actual data for detection using Approach 1

every time step over a fixed sequence length for training.

3) Training: We trained two neural networks using each of the approaches described in Section IV on the attack data generated. Training is done using binary cross entropy loss on an ADAM optimizer. The performance of both approaches is described below:

In approach 1, we trained one network on all 39 measurements to predict the state of all 39 devices for attacked and non attacked conditions. The results are shown in Fig. 6. We categorized our test conditions into two segments: normal and unseen conditions. In unseen conditions we ensure that the training data never encounters an instance of an attack on a particular bus and the test data has instances of attack on the same bus. To check for addressing unseen attacks, in approach1, we trained a neural network for scenarios where devices 19 and 24 are un-attacked. For this we fit our network on attacks of all other devices. Finally during testing we have tested our model on the rows where attacks on device 19 and 24 are present and the results are shown in Fig. 7.

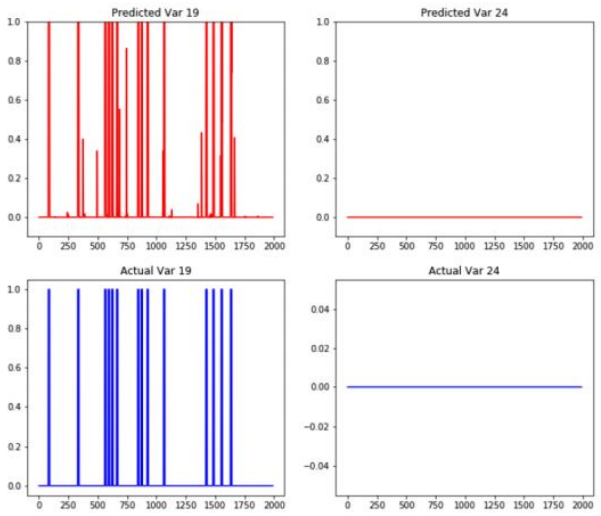

Fig. 7. Predicted vs Actual data for detection of unseen attacks using Approach 1

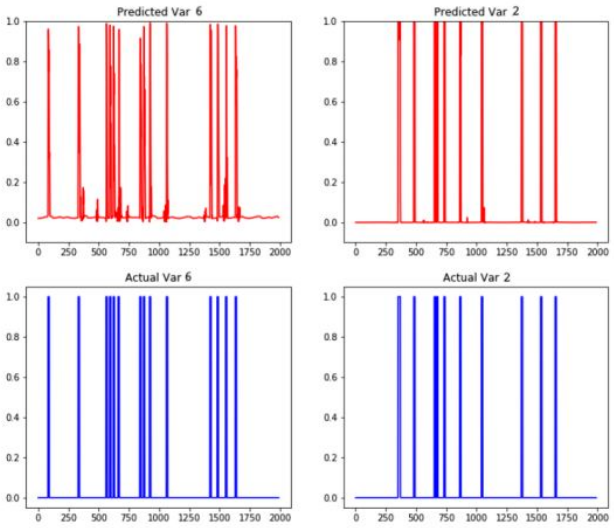

Fig. 8. Predicted vs Actual data for detection attacks using Approach 2

Similarly, for approach 2, we trained a network on two measuring devices which are spatially located nearby to predict the state of their state. The output is shown in Fig. 8. After that to make it robust for other devices we have fine-tuned this network over measurements from devices 8 and 25 which are not under-attack during training. Finally we test this network under attacked conditions and the results are shown in Fig. 9.

TABLE I
COMPARISON OF PERFORMANCE OF TWO APPROACHES

Approach	Normal Attacks	Unseen Attacks
Approach 1	**0.999**	**0.923**
Approach 2	**0.949**	**0.534**

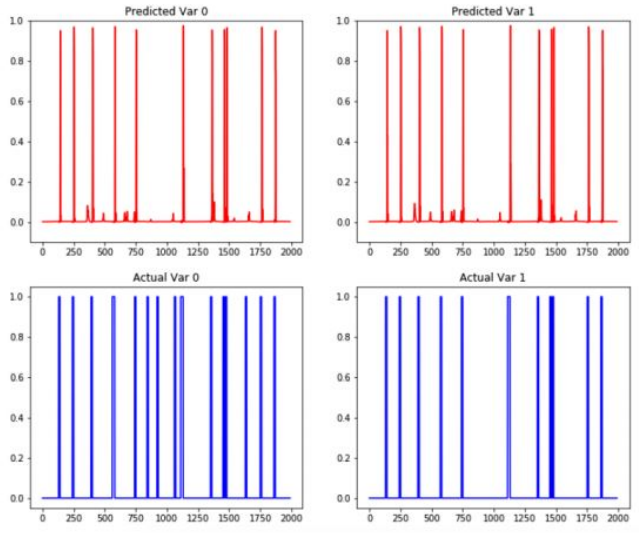

Fig. 9. Predicted vs Actual data for detection of unseen attacks using Approach 2

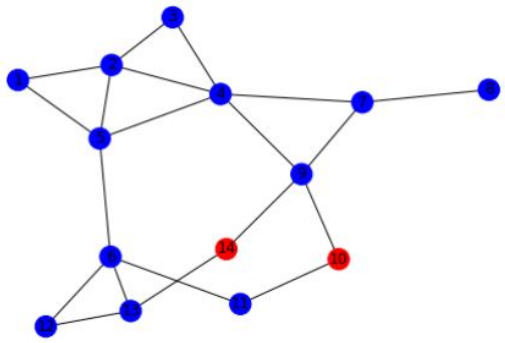

Fig. 10. Red nodes indicate compromised meters on bus 10 and 14

The performance of both these approaches are compared in Table I. Since this is a classification problem we use ROC-AUC (Receiver Operating Characteristics-Area Under The Curve) to measure the performance of our approaches. We also generate a visual of the network graph to denote the buses under attack based on the devices we detect to be compromised as shown in Fig. 10.

VI. Conclusion and Future Works

We have explored a few ways to detect attacks on the power grid in real-time using deep learning. The proposed approach is highly scalable and runs in linear time. We have demonstrated strictly supervised and semi supervised approaches towards attack detection for targeted and random attacks. It can be seen that the performance of the network which trains on all measurements is better than that of the one which trains on localized measurements. Especially the difference is noticeable for the cases of unseen attacks. This can be because an attack does not just affect states at local clusters but also some far off connections. The correlations between these connections are learnt in the first approach whereas we fail to encode them in the second approach. Moreover, we need to find a better decentralized approach for intrusion detection so that it is scalable and also performs well on unseen attacks. We have tested our framework on a 14-bus case for random and targeted attacks. This system can be scaled to larger scenarios and other kinds of attacks like LMP attacks, voltage over loading attacks, load redistribution attacks etc.

Acknowledgment

The material presented in this paper is based upon work supported by the NSF division of Electrical, Communication and Cyber System under Award Number 1808064.

References

[1] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 21–32. [Online]. Available: http://doi.acm.org/10.1145/1653662.1653666

[2] K. Mehrotra, C. K. Mohan, and S. Ranka., *Elements of artificial neural networks.*, 1997.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. [Online]. Available: http://doi.acm.org/10.1145/3065386

[4] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," 2019.

[5] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019.

[6] R. Chalapathy, E. Z. Borzeshi, and M. Piccardi, "An investigation of recurrent neural architectures for drug name recognition," 2016.

[7] D. Wulsin, J. Blanco, R. Mani, and B. Litt, "Semi-supervised anomaly detection for eeg waveforms using deep belief nets," in *2010 Ninth International Conference on Machine Learning and Applications*, Dec 2010, pp. 436–441.

[8] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, "Deep learning for unsupervised insider threat detection in structured cybersecurity data streams," 2017.

[9] T. Schlegl, P. Seebck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," 2017.

[10] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," 2015.

[11] Q. Yang, J. Yang, W. Yu, D. An, N. Zhang, and W. Zhao, "On false data-injection attacks against power system state estimation: Modeling and countermeasures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 717–729, March 2014.

[12] S. Bi and Y. J. Zhang, "Defending mechanisms against false-data injection attacks in the power system state estimation," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, Dec 2011, pp. 1162–1167.

[13] O. Kosut, Liyan Jia, R. J. Thomas, and Lang Tong, "Limiting false data attacks on power system state estimation," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, March 2010.

[14] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks with incomplete information against smart power grids," in *2012 IEEE Global Communications Conference (GLOBECOM)*, Dec 2012, pp. 3153–3158.

[15] P. Chen, S. Yang, J. A. McCann, J. Lin, and X. Yang, "Detection of false data injection attacks in smart-grid systems," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 206–213, Feb 2015.

[16] L. Liu, M. Esmalifalak, Q. Ding, V. A. Emesih, and Z. Han, "Detecting false data injection attacks on power grid by sparse optimization," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 612–621, March 2014.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams., pp. 696–699.

[18] H.-j. Kim and K.-s. Shin, "A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets," *Appl. Soft Comput.*, vol. 7, no. 2, pp. 569–576, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2006.03.004

[19] S. Hochreiter and J. Schmidhuber., pp. 1735–1780.

[20] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *2018 IEEE Texas Power and Energy Conference (TPEC)*, Feb 2018, pp. 1–6.