Physics-guided Deep Learning for Time-Series State Estimation Against False Data Injection Attacks

Lei Wang
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, USA
ray_wang@knights.ucf.edu

Qun Zhou
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, USA
qun.zhou@ucf.edu

***Abstract*—The modern power grid is a cyber-physical system. While the grid is becoming more intelligent with emerging sensing and communication techniques, new vulnerabilities are introduced and cyber security becomes a major concern. One type of cyber attacks – False Data Injection Attacks (FDIAs) – exploits the limitations in traditional power system state estimation, and modifies system states without being detected. In this paper, we propose a physics-guided deep learning (PGDL) approach to defend against FDIAs. The PGDL takes real-time measurements as inputs to neural networks, outputs the estimated states, and reconstructs measurements considering power system physics. A deep recurrent neural network – Long Short Term Memory (LSTM) – is employed to learn the temporal correlations among states. This hybrid learning model leads to a time-series state estimation method to defend against FDIAs. The simulation results using IEEE 14-bus test system demonstrate the accuracy and robustness of the proposed time-series state estimation under FDIAs.**

***Index Terms*—Cyber security, state estimation, physics-guided deep learning, false data injection attacks, temporal correlations**

I. Introduction

The power grid is a critical infrastructure that produces and delivers electricity from generating sources to end consumers. To monitor and control the states of power systems, Energy Management Systems (EMS) are widely adopted. As a core function in EMS, power system state estimation (PSSE) receives raw measurements from Supervisory Control and Data Acquisition (SCADA) system and provide estimates of systems states to be used in other EMS applications [1].

With the advancements of new monitoring technology, more meters and sensors are installed in the power grid, such as Phase Measurement Units (PMUs), Intelligent Electronic Device (IEDs) and smart meters. While the emerging technology helps the traditional electrical system update to a smart grid with intelligent cyber-physical layers, it also raises new security problems [2]. New vulnerabilities are introduced by integrating these smart devices. For instance, False Data Injection Attacks (FDIAs) can bypass the bad data detection and introduce errors to PSSE without being noticed [3]–[6]. Attackers attempt to achieve malicious objectives such as making financial profit or cause system outages, which are extremely harmful to the critical energy infrastructure.

FDIAs stealthily modify the measurement data (PSSE input) so that the estimated states (PSSE output) will change, which further impacts other critical EMS functions. In order to defend against FDIAs, several countermeasures have been proposed in recent studies. In [4], an approach is proposed to filter out the abnormal data based on the consistency of measurements using a subset of PMU measurements. While PMUs can directly provide voltage phasors, it is noticed that PMUs can be manipulated by adversary with GPS spoofing [7]. Bobba *et al.* proposed two ways to defend against the FDIAs: one way is to protect a strategically selected set of sensor measurements, and the other is to independently verify the values of the selected set of state variables [8]. Similarly, the work in [9] proposed countermeasures by protecting a small subset of measurements or deploying secure PMUs based on greedy algorithm. An approach to detect and isolate the data attacks is proposed in [10], it relies on some secure measurements of bus voltage magnitudes.

In this paper, we address the FDIA problem from a different angle. Traditional PSSE is considered as single-snapshot estimation, which takes measurement data at the present time as input and outputs the states for the moment. This single-snapshot estimation is very sensitive to bad measurement data. In this paper, we consider multiple snapshots using the present and past data and propose an time-series PSSE. The proposed time-series PSSE not only looks at the present status of the system, but also exploits the past instances from the system. Therefore, any deviations or stealthy data changes can be easily notified. As a result, the proposed time-series PSSE is less sensitive to the present measurement data and capable of delivering robust state estimates under FDIAs.

In order to take the temporal correlations into account, the proposed PSSE employs the state-of-the-art machine learning method. Specifically, a Long Short Term Memory (LSTM) neural network is adopted to learn the temporal correlations. The time-series PSSE also considers power grid physical aspects and integrate into a physics-guided deep learning approach. The uniqueness of the physics-guided deep learning is that it is not only data-driven, but also based on first principles. In other words, the physics-guided deep learning makes the neural network blackbox explainable in accordance with the physical model.

The paper is organized as follows. Section II reviews FDIAs and its impact on state estimation. Section III presents the time-series PSSE using a physics-guided deep learning method against FDIAs. Section IV presents simulation results and the estimation performance in the IEEE 14-bus system. Section V provides concluding remarks.

II. False Data Injection Attacks

The bad data detection algorithms in traditional PSSE are residual-based. FDIAs exploits this feature and constructs attack vectors that do not change residuals and hence bypass the bad data detection algorithm.

Specifically, the relationship between system measurement and states can be represented by:

$$\mathbf{Z} = \mathbf{h}(\mathbf{X}) + \mathbf{e} \tag{1}$$

where $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_m]^T$ is the measurement vector, consisting of voltage V_i at bus i , real power injection P_i, reactive power injection Q_i, real power flow P_{ij} and reactive power flow Q_{ij} between bus i and bus j. The state vector is $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$ consisting of voltage magnitudes $|V_i|$ and phase angles θ_i. The measurement error is represented by $\mathbf{e}$, assumed to follow Gaussian distribution, i.e., $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$ where Σ is the error covariance matrix [11]. The number of measurements is m and the number of states is n. Usually measurement redundancy ensures that $m \geq n$.

State estimation is to find an estimated state vector $\hat{\mathbf{X}}$ that could best fit the available measurements. The state estimation problem can be expressed by:

$$\hat{\mathbf{X}} = \arg\min_{\hat{\mathbf{X}}} [\mathbf{Z} - \mathbf{h}(\hat{\mathbf{X}})]^{\mathbf{T}} \mathbf{W} [\mathbf{Z} - \mathbf{h}(\hat{\mathbf{X}})] \tag{2}$$

where $\mathbf{W}$ denotes the weight matrix. The state estimation problem can be solved by Weighted Least Square (WLS) criterion.

Measurement bad data exists due to device misconfiguration, fluctuating noises, or malicious attacks. The bad data can deviate the estimated states from true states. Thus, in order to obtain good estimates, these bad data need to be detected, identified and removed in time. A commonly used criterion to detect bad data is based on measurement residuals:

$$r = \|\mathbf{Z} - \mathbf{h}(\hat{\mathbf{X}})\| > \tau \tag{3}$$

where $\hat{\mathbf{X}}$ is the estimated state vector and τ is the prescribed threshold. If inequality (3) holds, there are bad data in the measurement vector.

FDIAs exploits this criterion and modifies measurement data to spoof the bad data detection mechanism in the PSSE. The FDIAs are based on measurement residual criterion and DC power flow model, where reactive power is neglected and voltage magnitudes of all buses are known as 1 per unit. As a simplified model, the DC state estimation is expressed by:

$$\hat{\mathbf{X}} = \arg\min_{\hat{\mathbf{X}}} [\mathbf{Z} - \mathbf{H}\hat{\mathbf{X}}]^{\mathbf{T}} \mathbf{W} [\mathbf{Z} - \mathbf{H}\hat{\mathbf{X}}] \tag{4}$$

where $\mathbf{H}$ is a sensitivity matrix from DC power flow equations.

The closed-loop state estimate $\hat{\mathbf{X}}$ is given by:

$$\hat{\mathbf{X}} = (\mathbf{H}^{\mathbf{T}}\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{W}\mathbf{Z} \tag{5}$$

FDIAs try to modify the measurement data without changing the residual r. Let r_a denote the measurement residual after the attack:

$$\begin{aligned} r_a &= \|\mathbf{z_a} - \widehat{\mathbf{z}}_\mathbf{a}\| \\ &= \|(\mathbf{z} + \mathbf{a}) - \mathbf{H}(\mathbf{H}^{\mathbf{T}}\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{W}(\mathbf{z} + \mathbf{a})\| \\ &= \|(\mathbf{z} - \hat{\mathbf{z}}) + (\mathbf{a} - \mathbf{H}(\mathbf{H}^{\mathbf{T}}\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{W}\mathbf{a})\| \end{aligned} \tag{6}$$

If the attack vector is carefully structured, it may bypass the bad data detection without being detected. As shown in [3], the residual r_a is equal to original residual r when the attack vector is a linear combination of the column vectors of $\mathbf{H}$,

$$a = \mathbf{H}\mathbf{c} \tag{7}$$

where c is a non-zero arbitrary vector.

This non-detectable attack directly impact the traditional single-snapshot state estimation, but the adverse impact is lessened with the proposed time-series PSSE in Section **??**.

III. Physics-guided Deep Learning Against FDIAs

To defend the undetectable FDIAs, we propose a hybrid machine learning model inspired by the emerging autoencoder in the Artificial Intelligent (AI) field. As shown in Fig. 1, an autoencoder is a neural network that is trained to copy its input to its output [12]. Internally, it has a hidden layer that divides the neural network into two parts: encoder and decoder. The autoencoder is initially designed for dimension reduction so that the number of features can be reduced to represent a system. To apply autoencoders in state estimation, the measurement $\mathbf{z}$ is fed into the encoder, and the hidden layer output is the estimated system states $\hat{\mathbf{x}}$, which can be considered as features to represent the power grid. The estimated states $\hat{\mathbf{x}}$ then go through the decoder to output the reconstructed measurement data $\hat{\mathbf{z}}$.

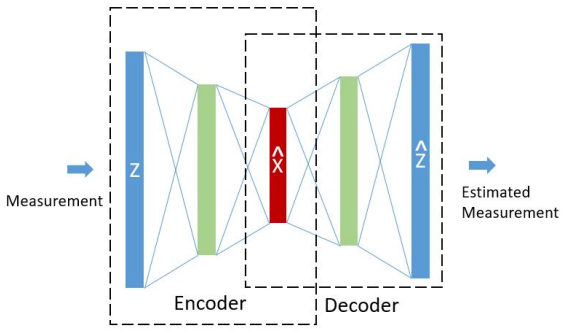

Fig. 1: Autoencoder for state estimation

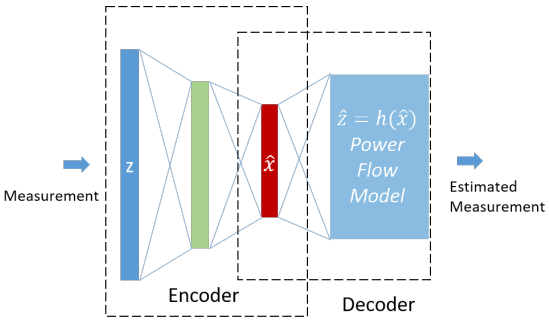

Fig. 2: Architecture of the PGDL state estimator

A. Physics-guided Deep Learning

With the domain knowledge of power systems, we propose to improve the autoencoder incorporating the first principles in the power grid. The proposed PGDL model is both data-driven and first-principle-based. Fig. 2 shows the structure of the PGDL for state estimation.

Here the first part of the autoencoder is maintained with a deep neural network, while the second part is replaced by power system physics with $h(\hat{X})$ – the power flow equations in (8).

$$
\begin{aligned}
P_i &= \sum_{j=1}^{N} V_i V_j (G_{ij}\cos(\theta_i-\theta_j) + B_{ij}\sin(\theta_i-\theta_j)) \\
Q_i &= \sum_{j=1}^{N} V_i V_j (G_{ij}\sin(\theta_i-\theta_j) + B_{ij}\cos(\theta_i-\theta_j)) \\
P_{ij} &= -{V_i}^2 G_{ij} + V_i V_j (G_{ij}\cos(\theta_i-\theta_j) + B_{ij}\sin(\theta_i-\theta_j)) \\
Q_{ij} &= -{V_i}^2 B_{ij} + V_i V_j (G_{ij}\sin(\theta_i-\theta_j) - B_{ij}\cos(\theta_i-\theta_j))
\end{aligned}
\tag{8}
$$

where:
V_i is the voltage magnitude at bus i
P_i is the real power injection at bus i
Q_i is the reactive power injection at the i_{th} node
P_{ij} is the real power flow from bus i to bus j
Q_{ij} is the reactive power flow from bus i to bus j
G_{ij} is the real part in admittance matrix
B_{ij} is the imaginary part in admittance matrix

The proposed PGDL method is able to produce state estimates using the neural network, and the deviations between reconstructed measurements and actual measurements are used to train the neural network.

B. Time-Series PSSE using Deep Neural Nets

The PSSE structure in Fig. 2 can be for single-snapshot estimation similar to traditional PSSE. To take advantage of historical measurements and detect any temporal deviations from normal states, we propose the time-series PSSE. The times-series PSSE takes into account temporal correlations among states, which indeed exist in real-world power systems. This special consideration allows the estimated states converge to the true states more accurately than that in traditional PSSE.

In order to learn the dynamics between states at different times, a deep recurrent neural network – LSTM – is chosen. LSTM networks can be considered as one type of Recurrent Neural Networks (RNNs). But unlike conventional RNNs, LSTM networks can learn the correlations for arbitrarily long time. This attribute is due to its innovative gating mechanism. An LSTM block with the gating mechanism is described in Fig. 3. The inputs of this LSTM are $\boldsymbol{x_t}$ and $\boldsymbol{h_{t-1}}$ while the output is $\boldsymbol{h_t}$. There are three gates, input gate $\boldsymbol{i}$, forget gate $\boldsymbol{f}$, and output gate $\boldsymbol{o}$. The LSTM unit will learn the dependencies between the data in the input sequence. The input gate manages the values flowing into the memory cell from the inputs. The forget gate determines which part is passed to the next step. As for the output, it is a product of the result of output gate and the activation of the memory cell. The mathematical expressions of LSTM are given as follows:

$$
\begin{aligned}
\boldsymbol{f_t} &= \sigma(\boldsymbol{W_f}\cdot[\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_f}) \\
\boldsymbol{i_t} &= \sigma(\boldsymbol{W_i}\cdot[\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_i}) \\
\boldsymbol{C_t} &= \boldsymbol{f_t} * \boldsymbol{C_{t-1}} + \boldsymbol{i_t} * (tanh(\boldsymbol{W_C}\cdot[\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_C})) \\
\boldsymbol{o_t} &= \sigma(\boldsymbol{W_o}\cdot[\boldsymbol{h_{t-1}}, \boldsymbol{x_t}] + \boldsymbol{b_o}) \\
\boldsymbol{h_t} &= \boldsymbol{o_t} * tanh(\boldsymbol{C_t})
\end{aligned}
$$

where $\boldsymbol{W_f}$, $\boldsymbol{W_i}$, $\boldsymbol{W_C}$, $\boldsymbol{W_o}$ are weights that connect different layers, while $\boldsymbol{b_f}$, $\boldsymbol{b_i}$, $\boldsymbol{b_C}$, $\boldsymbol{b_o}$ are the biases with each individual gate. $\sigma(\cdot)$ represents the $sigmoid$ activation function, and $tanh$ is the hyperbolic tangent function.

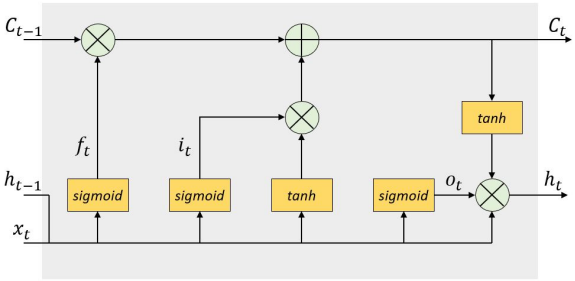

Fig. 3: LSTM block

Unlike the common supervised training model, we extend the proposed state estimator by the part of power flow model and use the difference between Z and $\hat{Z}$ to update the weights and biases of the neural network. This physics-guided deep learning model can be used in an online environment. Compared with the traditional snapshot-based WLS estimation, this model outperforms in accuracy and robustness against FDIAs, which will be shown in Section IV.

IV. Simulation Results

In this section, the performance of time-series PSSE using the physics-guided deep learning is demonstrated in IEEE 14-bus system shown in Fig. 4. As seen in Table I, the test system consists of 32 measurements and 27 states. Different FDI attack scenarios are investigated and the results are discussed in detail.

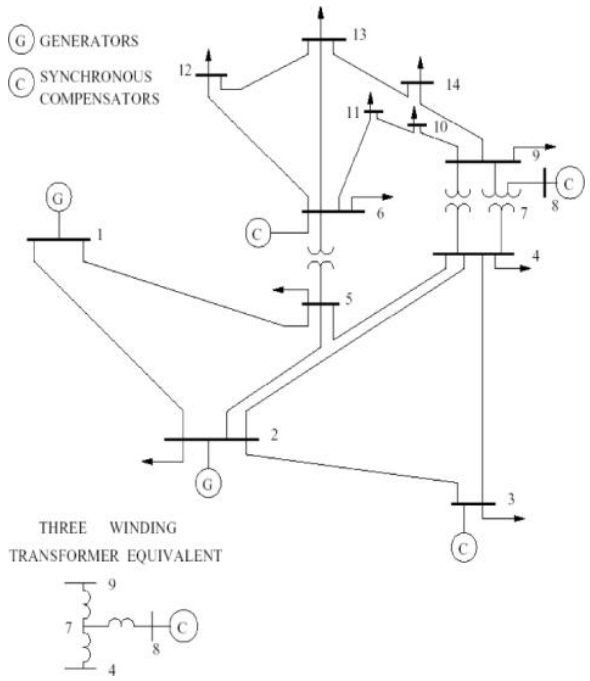

Fig. 4: IEEE 14-bus system [13]

TABLE I: Measurement vectors of targeted system

measurements	IEEE14
V_i	1
P_i	8
Q_i	8
P_ij	9
Q_ij	6

A. Simulation Setup

1) Measurement Data: Simulation is designed in MATLAB environment with the MATPOWER package [14], [15]. To simulate the power network in a more practical way, the load profile from NYISO is adopted. We first fit the normalized load profile into the IEEE 14-bus case file. Then, the power flow is run to generate the true states and measurements. After obtaining the true measurements, the zero-mean White Gaussian Noise (WGN) is added to each measurement according to their standard deviation following equation (11). Here A_i is the actual value of measurement, σ_i is the standard deviation of measurement.

$$Z_i = A_i + randn * \sigma_i \tag{9}$$

Specifically, the load profile from NYISO is recorded at a 5 minutes interval. The load data of 6 months are generated for training, validation and testing. The data of the last day in our selected time horizon is used as the testing data (288 data points).

2) Attack Data: To launch FDIAs, attackers are assumed to have the knowledge of the targeted system and the access to meters. In this paper, we assume attackers have limited access to meters and can only modify the measurement data for those meters. Let $P = H(H^TH)^{-1}H^T$, the attack vector can be obtained by solving (10).

$$\begin{aligned} a = Hc \Leftrightarrow Pa = a \Leftrightarrow Pa - a = 0 \\ \Leftrightarrow (P - I)a = 0 \Leftrightarrow Ba = 0 \end{aligned} \tag{10}$$

Let ℓ_{meter} denotes the set of indices of meters that can be accessed by attacker. For unaccessible meters $i \notin \ell_{meter}$, the element in the attack vector $a_i = 0$. Then the modified measurement set is:

$$Z_a = Z + a \tag{11}$$

As shown in [3], if the number of accessible meters is larger than $\boldsymbol{m-n}$, the attackers are always able to find attack vectors without being detected using a DC power flow model. In our test case given in Table I, the total number of real power injection and real power flow is 17, which is enough for attacker to design attacks given the accessibility of 5 meters. One instance is shown in Table II.

TABLE II: Attack vectors of false data injection

measurements	Measured Value (pu)	Malicious Value (pu)	Attack Vector (pu)
P_2	0.1888	0.1937	-0.0049
P_{10}	-0.1156	-0.0557	-0.0599
P_{12}	0.2428	-0.0576	0.3004
P_{14}	0.2027	-0.1083	0.3110
$P_{4,5}$	-0.5030	-0.4885	-0.0144

B. Result Analysis

In this paper, we consider two attack scenarios: random and consecutive attacks. For the random false data injection attacks (FDIAs), attackers will manipulate the accessible measurements at random time points. In our test case, 30 time points are randomly selected to inject false data in the testing period. While regarding the consecutive FDIAs, we assume the attacks will last for one hour with a random beginning point during the testing period, that is, 12 consecutive time points are under attack.

1) Temporal Comparison: Figs. 5 and 6 give the RMSEs between estimated states and true states of the time points under FDIAs. It is obvious that our proposed PGDL model outperforms WLS in both scenarios in terms of FDIAs. With the random FDIAs, PGDL model still can keep a more accurate state estimation in most cases compared to WLS, as illustrated in Fig 5. When there are consecutive FDIAs, the perofmance of PGDL model is still better than WLS. While compared with random FDIAs, the PGDL model has a larger RMSE facing continuous attacks. The reason is that the dynamic temporal correlation learned by PGDL model can suppress the bad data considering multiple data points. This also can be verified by the result shown in 6. When FDIAs begin at point 1, the state RMSE is still small due to the suppression of normal condition at point 0. In addition, when the FDIAs stop at point 12, the

RMSE of the normal condition at point 13 decreases but still a little large. The mean and standard deviation of RMSEs are described in Table III. The lower mean value and stand deivation of PGDL model indicates the stable performance of our proposed method against FDIAs.

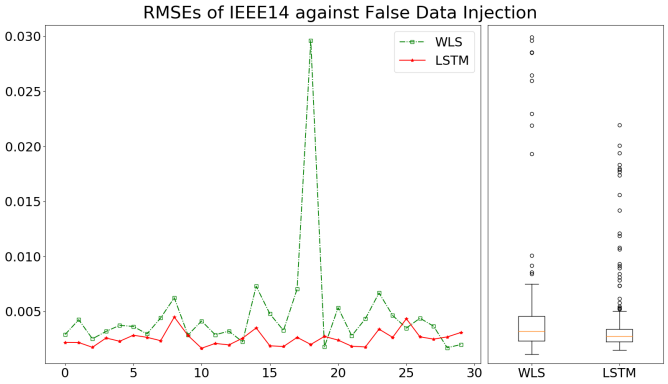

Fig. 5: State RMSE with random FDIAs

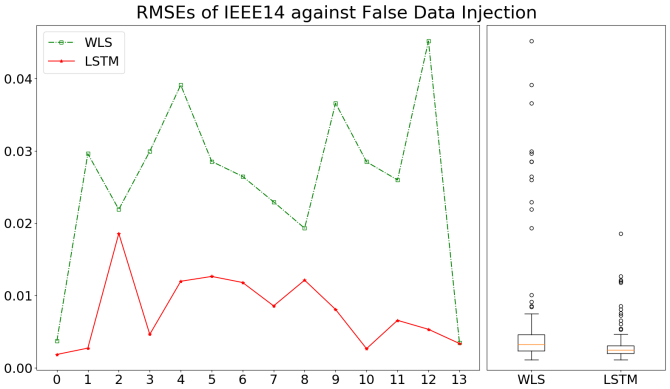

Fig. 6: State RMSE with consecutive FDIAs

TABLE III: RMSEs under FDI attacks

	Random FDIAs		Consecutive FDIAs	
	Mean(RMSE)	SD(RMSE)	Mean(RMSE)	SD(RMSE)
WLS	0.0261	0.0090	0.0046	0.0056
LSTM	0.0028	0.0008	0.0028	0.0018

2) Spatial Comparison: Figs. 7 and 8 indicate the estimated states at the time points when there are malicious false data injections. They show the largest RMSE in random and consecutive FDIAs respectively. It can be seen that the results of PGDL model is much close to the true states when there are random FDI attacks. While consecutive attacks happen, PGDL shows a higher deviation at voltage angles but much closer to true voltage magnitudes at the selected time point. Overall, PGDL has lower RMSE indicated in Fig. 6.

V. Conclusion

Cyber security becomes a major concern in the modern power grid. To defend against FDIAs, a physics-guided deep learning model is proposed to develop the time-series PSSE. In contrast to traditional state estimation, the time-series PSSE uses LSMT to learn the temporal correlations among states at different times. The learned temporal correlations strengthen the state estimation under FDIAs. The proposed time-series PSSE take a sequential measurements as inputs, and output the estimated states. The adverse impact of any random attacks contaminating measurements at times is mitigated by incorporating measurements at different times. The proposed PGDL approach is demonstrated in IEEE 14-bus system, and the time-series estimator can better withstand FDIAs.

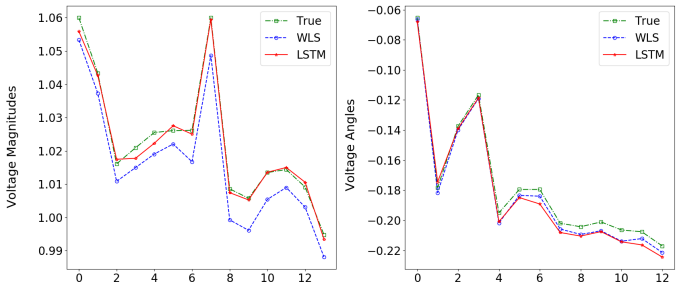

Fig. 7: Estimated states with random FDIAs

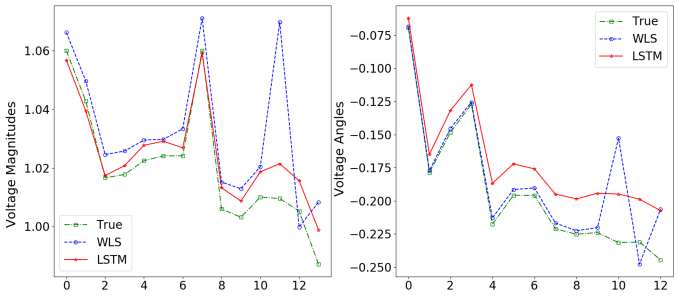

Fig. 8: Estimated states with consecutive FDIAs

References

[1] A. Abur and A. Expósito, *Power System State Estimation: Theory and Implementation*, ser. Power Engineering (Willis). CRC Press, 2004.

[2] E. Knapp and J. Langill, *Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems*. Elsevier Science, 2011.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 13:1–13:33, June 2011.

[4] J. Zhao, L. Mili, and M. Wang, "A generalized false data injection attacks against power system nonlinear state estimator and countermeasures," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 4868–4877, Sept 2018.

[5] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks against nonlinear state estimation in smart power grids," in *2013 IEEE Power Energy Society General Meeting*, July 2013, pp. 1–5.

[6] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec 2011.

[7] Z. Zhang, S. Gong, A. D. Dimitrovski, and H. Li, "Time synchronization attack in smart grid: Impact and analysis," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 87–98, March 2013.

[8] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on dc state estimation," 2010.

[9] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 326–333, June 2011.
[10] K. C. Sou, H. Sandberg, and K. H. Johansson, "Data attack isolation in power networks using secure voltage magnitude measurements," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 14–28, Jan 2014.
[11] F. C. Schweppe and J. Wildes, "Power system static-state estimation, part 1,2,3," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-89, no. 1, pp. 120–135, Jan 1970.
[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[13] N. Mithulananthan, C. A. C. nizares, and J. Reeve, "Indices to detect hopf bifurcation in power systems," in *In Proc. of NAPS-2000*, 2000, pp. 15–18.
[14] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, "Matpower's extensible optimal power flow architecture," in *2009 IEEE Power Energy Society General Meeting*, July 2009, pp. 1–7.
[15] Matpower. [Online]. Available: http://www.pserc.cornell.edu/matpower