ICIC Express Letters
Volume **16**, Number **1**, January **2022**

 ISSN 1881-803X
pp. **43–50**

LOCATIONAL DETECTION OF DATA INTEGRITY ATTACKS WITH MULTI-GATE MIXTURE-OF-EXPERTS IN SMART GRID

Xuzhen Fan[1], Meng Zhang[1,*], Huijie Zeng[2] and Chao Shen[1]

[1]School of Cyber Science and Engineering
Xi'an Jiaotong University
No. 28, West Xianning Road, Xi'an 710049, P. R. China
fxz3783@stu.xjtu.edu.cn; chaoshen@xjtu.edu.cn
*Corresponding author: mengzhang2009@xjtu.edu.cn

[2]College of Equipment Management and UAV Engineering
Air Force Engineering University (AFEU)
Jiazi No. 1, East Changle Road, Xi'an 710038, P. R. China
bonnie1995@126.com

Received May 2021; accepted July 2021

Abstract. *In recent years, data integrity attacks have gradually become a major threat to the security of smart grid. The presence detection and locational detection of data integrity attacks are of vital significance for repairing vulnerable points, but the latter is rarely studied. The multi-task classification problem for locational detection of data integrity attacks is formulated in this paper. Based on the formulation, the locational detection with multi-gate mixture-of-experts (LD-MMoE) scheme is proposed. When reaching similar detection performance, the number of LD-MMoE's trainable parameters is less than that of the state-of-the-art study, which indicates the proposed LD-MMoE scheme is more computational cost-friendly. Simulations conducted on the IEEE 14-bus system and 118-bus system verify the effectiveness of the proposed LD-MMoE scheme.*
Keywords: False data injection attack, Locational detection, Multi-gate mixture-of-experts, Multi-task learning, Smart grid

1. **Introduction.** In recent years, smart grid has become a typical cyber-physical system with high efficiency and high economy. However, the efficient use of information flow also brings new risks to the smart grid [1]. In 2015, three power distribution companies in Ukraine were attacked by cyber attacks, affecting the normal electricity consumption of 225,000 users [2].

In smart grid, system operators make control decisions based on the current system states and formulate the dispatch plan. By using the measurements received by the supervisory control and data acquisition (SCADA) system, system operators can obtain the estimated states, and use the estimation residual for bad data detection (BDD) [3]. Recent research has shown that a type of data integrity attacks named false data injection (FDI) attack can be launched against the state estimation while keeping stealthy to traditional BDD mechanism [4]. Attackers can modify estimated states by compromising measurements in SCADA [5] system, causing severe consequences such as key lines overloading or load shedding [6].

The presence detection or the detection of FDI attacks for whole system is a single-task classification problem, i.e., $\boldsymbol{z}$ is attacked or not [7]. And the locational detection or the detection of FDI attacks for each measurement is a multi-task classification problem, i.e., z_m is attacked or not, $m = 1, \ldots, M$. When attacking a system state, the attacker must tamper with all relevant measurements at the same time [8]. For defender who wants to

DOI: 10.24507/icicel.16.01.43

detect the attack at each measurement, she/he can train a model for each measurement. However, this method cannot effectively use the correlation between different measurements, and requires a great deal of computing resources. In order to improve the detection ability, the defender must take advantage of this correlation. In machine learning, the convolutional neural network (CNN) is designed to capture the correlation between adjacent input dimensions [9]. Based on this, recent research has designed a BDD-CNN architechture to detect the FDI attacks for each measurement, where CNN is used to capture the inconsistency and cooccurrence dependency in the power flow measurements due to potential attacks [7]. However, this architecture is highly dependent on the numbering sequence of measurements, and adjacent measurements should be numbered consecutively. For large-scale power systems, this is difficult to achieve. Meanwhile, the real-time detection is very important for the security of the system. Therefore, the detection scheme should use as few computing resources as possible with guaranteed performance, i.e., have fewer trainable parameters.

The main contributions of this paper are shown as follows.

- The locational detection of FDI attacks is formulated as the form of multi-task classification problem in this paper while most existing results ignore exact attack injection locations.
- Based on the formulation, a locational detection with multi-gate mixture-of-experts (LD-MMoE) scheme is proposed, which does not rely on additional conditions, e.g., the numbering sequence of measurements.
- LD-MMoE can greatly reduce the number of trainable parameters while keeping sufficient effectiveness, which is verified by the simulations conducted on IEEE 14-bus system and 118-bus system.

The rest of this paper is organized as follows. Section 2 introduces some preliminaries. Section 3 designs the LD-MMoE scheme. Section 4 presents simulations on the IEEE test systems and Section 5 concludes this paper.

Table 1. Nomenclature

a	Subscript: the quantity is under FDI attack	$\hat{\boldsymbol{x}}$	The estimated system state
E	Subscript: the quantity for the expert layer	$\boldsymbol{z}$	The measurement vector of system
e	Subscript: index for experts	η	The threshold for BDD
F	Subscript: the quantity for the fully connected layer	σ	The standard deviation of system measurement
i	Subscript: index for buses	b	The biases of neurons
j	Subscript: index for neurons	K	The number of experts
l	Subscript: index for lines	$g(\cdot)$	The gate for each task
m	Subscript: index for measurements	M	The number of measurements
n	Subscript: index for instances	N	The number of system states
T	Subscript: the quantity for the tower layer	o	The outputs of neurons
$\boldsymbol{a}$	The attck vector of the FDI attack	$p(\cdot)$	The probability
$\boldsymbol{c}$	The injected false data	r	The 2-norm residual of state estimation
$\boldsymbol{e}$	The measurement error	w_e	The weights of experts
$\boldsymbol{H}$	The measurement matrix of system	χ^2_α	The upper quantile of chi square distribution
$\boldsymbol{R}$	The covariance matrix of system	y	The groud truth label
$\boldsymbol{w}$	The weights of layers	$\hat{y}$	The predicted probability
$\boldsymbol{x}$	The state of system	σ	The standard deviation of measurements

2. Preliminaries.

2.1. State estimation and bad data detection.

The DC power flow model is used in this paper, where the measurements in system can be expressed as $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{e}$. The objective function of state estimation can be expressed as

$$J(\boldsymbol{x}) = [\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}]^T \boldsymbol{R}^{-1} [\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}]. \tag{1}$$

Based on the measurements in $\boldsymbol{z}$, the result of state estimation is the value of a set of state variables $\hat{\boldsymbol{x}}$ that minimizes the objective function $J(\boldsymbol{x})$,

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{z}. \tag{2}$$

Bad measurements may be introduced due to various reasons, e.g., device failures and malicious cyber attacks. BDD has been developed to protect state estimation. The 2-norm estimation residual $r = \|\boldsymbol{z} - \boldsymbol{H}\hat{\boldsymbol{x}}\|_2$ is used to measure the inconsistency among the normal and the abnormal measurements. For a given threshold η, $r \geqslant \eta$ means there are bad data in the measurements while $r < \eta$ means measurements are normal.

2.2. False data injection attack.

It is assumed that attacker can access the current power system topology and parameters, and can manipulate measurements obtained by the SCADA system. The attack vector $\boldsymbol{a}$ can be constructed based on the above information as $\boldsymbol{a} = \boldsymbol{H}\boldsymbol{c}$. After FDI attacks are launched, the tampered measurements $\boldsymbol{z}_a$ obtained by SCADA are

$$\boldsymbol{z}_a = \boldsymbol{z} + \boldsymbol{a} = \boldsymbol{z} + \boldsymbol{H}\boldsymbol{c}. \tag{3}$$

The estimated state under attack is

$$\begin{aligned} \hat{\boldsymbol{x}}_a &= \left(\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{z}_a \\ &= \left(\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} (\boldsymbol{z} + \boldsymbol{a}) \\ &= \hat{\boldsymbol{x}} + \left(\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{a}. \end{aligned} \tag{4}$$

The residual under attack is

$$\begin{aligned} r_a &= \|\boldsymbol{z}_a - \boldsymbol{H}\hat{\boldsymbol{x}}_a\|_2 \\ &= \left\| \boldsymbol{z} + \boldsymbol{H}\boldsymbol{c} - \boldsymbol{H}\left[\hat{\boldsymbol{x}} + \left(\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{a}\right] \right\|_2 \\ &= \left\| \boldsymbol{z} + \boldsymbol{H}\boldsymbol{c} - \boldsymbol{H}\left[\hat{\boldsymbol{x}} + \left(\boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T \boldsymbol{R}^{-1} \boldsymbol{H}\boldsymbol{c}\right] \right\|_2 \\ &= \|\boldsymbol{z} - \boldsymbol{H}\hat{\boldsymbol{x}}\|_2 \\ &= r, \end{aligned} \tag{5}$$

which shows r_a is the same as r; hence the FDI attack with attack vector $\boldsymbol{a} = \boldsymbol{H}\boldsymbol{c}$ will not trigger the BDD.

3. Locational Detection with Multi-Gate Mixture-of-Experts.

In this section, the multi-task classification problem for locational detection of FDI attacks is formulated. Based on this formulation, the LD-MMoE scheme is proposed.

3.1. Multi-task classification problem for locational detection of FDI attacks.

The defender can obtain system measurements as $\boldsymbol{z} = (z_1, \ldots, z_M)$. She/He needs to predict whether each measurement is compromised according to the received measurements. For detection of FDI attacks for measurements, the multi-task classification problem can be formulated as

$$\hat{y}_m = p(y_m = 1 | z_1, \ldots, z_M) \in [0, 1], \quad m = 1, \ldots, M, \tag{6}$$

where $\hat{y}_m$ is the classification probability of measurement m being compromised. The detection of FDI attacks for each measurement is one task, and a single model that learns these multiple tasks simultaneously is proposed in this paper.

3.2. **Multi-gate mixture-of-experts.** The original multi-gate mixture-of-experts (MMoE) model is proposed to balance the task-specific objectives and correlation between tasks in [10]. Based on the mixture-of-experts (MoE) structure, multi tasks are learned by sharing the expert layer, and a gating unit is set up for each task to enhance flexibility.

The expert layer in MMoE is composed of some independent parallel subnets, and these subnets are called experts. In deep neural networks, ensemble subnets have been proven to be able to improve model performance [11], and the ensemble of output of each expert is used as the input of successive layer. The output of each expert can be formulated as

$$o_{E,e,j} = ReLU(\boldsymbol{Input} \times \boldsymbol{w}_{E,e,j} + b_{E,e,j}). \tag{7}$$

The gating unit with softmax function is used to calculate the weights of output of each expert

$$g(\cdot) = softmax(\boldsymbol{Input} \times \boldsymbol{w}). \tag{8}$$

3.3. **Multi-gate mixture-of-experts for locational detection.** In order to effectively extract the strong correlation between all measurements related to the attacked state variables, a shared bottom (fully connected layer) is combined with the original MMoE model in [10]. The LD-MMoE contains an input layer, a fully connected layer, an expert layer, a tower layer and an output layer.

- The fully connected layer:

$$o_{F,j} = ReLU(\boldsymbol{z} \times \boldsymbol{w}_{F,j} + b_{F,j}). \tag{9}$$

- The expert layer:

$$o_{E,e,j} = ReLU(\boldsymbol{o}_F \times \boldsymbol{w}_{E,e,j} + b_{E,e,j}). \tag{10}$$

- The tower layer:

$$o_{T,m,j} = ReLU\left(\left(\sum_{e=1}^{K} w_e \boldsymbol{o}_{E,e}\right) \times \boldsymbol{w}_{T,m,j} + b_{T,m,j}\right), \tag{11}$$

  where the weights of experts w_e are calculated by the gating unit,

$$(w_1, \ldots, w_K)^T = softmax(\boldsymbol{z} \times \boldsymbol{w}_{T,m}). \tag{12}$$

- The output layer:

$$\hat{y}_m = sigmoid(\boldsymbol{o}_{T,m} \times \boldsymbol{w}_{O,m} + b_{O,m}). \tag{13}$$

The input layer has M neurons which are corresponding to M measurements. The fully connected layer crosses and combines the input measurements to extract the relationship between them. The different experts in expert layer can learn different knowledge, which is important for model to extract independent features for different measurements. The tower layer is composed of M towers which are corresponding to M measurements. Towers are also independent parallel subnets, and the input of each tower is weighted sum of experts. It is worth noting that $\boldsymbol{w}_{T,m}$ in gating unit is a trainable matrix, which is used to balance the correlation and difference between the detection of FDI attacks for different measurements. The output layer has M neurons which are corresponding to M measurements, and each neuron is connected to only one tower. The classification probability of each measurement is calculated by the sigmoid function.

Besides, the rectified-linear unit (ReLU) activation function is used for nonlinear transformation. The *logloss* is used to evaluate the difference between the ground truth labels

and the output of network. In this paper, each batch contains 512 instances of data and the *logloss* of a batch is formulated as

$$logloss = -\frac{1}{512}\sum_{n=1}^{512}\sum_{m=1}^{M}\left[y_{n,m}\log\left(\hat{y}_{n,m}\right) + \left(1-y_{n,m}\right)\log\left(1-\hat{y}_{n,m}\right)\right]. \tag{14}$$

3.4. **Performance metrics for LD-MMoE.** The detection scheme should improve the detection accuracy and avoid the false alarms as much as possible [12], so it needs comprehensive performance evaluation metrics. In order to comprehensively evaluate the performance of the proposed scheme, *Precision*, *Recall*, *F1-score* and the area under the ROC curve (AUC) are used as evaluation metrics for LD-MMoE. In order to calculate these metrics, the true positive rate (TPR), false positive rate (FPR), false negative rate (FNR), true negative rate (TNR) are defined in Table 2.

Table 2. Definitions of some performance metrics

	Compromised	Uncompromised
Classified as compromised	TPR	FPR
Classified as uncompromised	FNR	TNR

Based on the above definition, the following metrics are presented.

- The *Precision* is defined as $Precision = TPR/(TPR + FPR)$.
- The *Recall* is defined as $Recall = TPR/(TPR + FNR)$.
- The *F1-score* is defined as $F1\text{-}score = 2\times(Precision\times Recall)/(Precision+Recall)$.
- The AUC is the area under the ROC curve. The ROC curve is drawn with TPR as ordinate and FPR as abscissa.

4. **Case Study.** In this section, the case study is conducted based on IEEE 14-bus system and 118-bus system in MATPOWER [13].

4.1. **Dataset.** The dataset in [7] is used in this section. The loads and noises are sampled from the normal distribution, and the min-cut FDI method is used to generate attacks [14]. In the simulation, the 2-norm of $\boldsymbol{c}$ is 2, and the σ of noise is 0.2.

4.2. **Simulation results for IEEE 14-bus system.** Table 3 shows the architecture of LD-MMoE for IEEE 14-bus system. The trainable parameters of the proposed scheme are 16,555, which is about 6.71% of that in [7]. Therefore, the proposed model needs fewer computing resources and the detection is faster. Figure 1 shows the learning curve of LD-MMoE for IEEE 14-bus system. The curves of training loss and validation loss are very smooth, which means that the training process of LD-MMoE is very stable for IEEE 14-bus system.

Table 3. Architecture of LD-MMoE for IEEE 14-bus system

Stage	Type	Neurons	Output size	Trainable parameters
0	Input	19	19×1	0
1	FullyConn	128	128×1	2,560
2	Experts	4×8	4×8	4,128
3	Towers	19×8	19×8	9,867
4	Sigmoid	19	19×1	0

Table 4 shows the metrics of the proposed scheme for each measurement in IEEE 14-bus system. The 9-th measurment has never been compromised in the dataset, so the related metrics are none. Compared with the model with 4 convolution layers and multi hidden layers in [7], the proposed scheme only needs 3 hidden layers to achieve similar

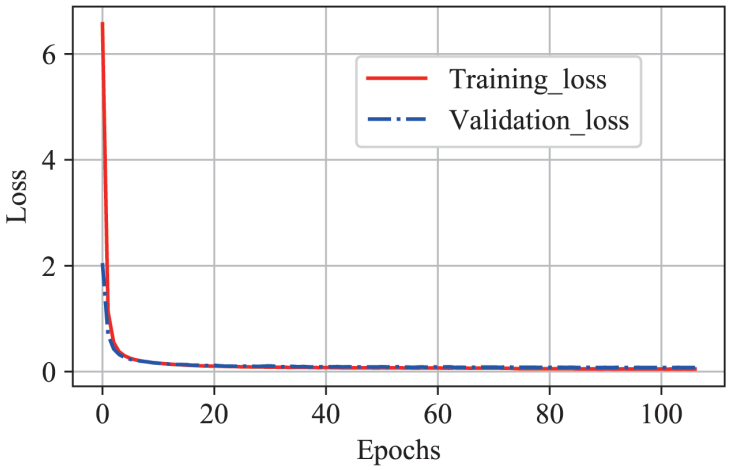

FIGURE 1. The learning curve of LD-MMoE for IEEE 14-bus system

TABLE 4. Metrics for each measurement in IEEE 14-bus system

Measurement	*Precision*	*Recall*	*F1-score*	*AUC*
1	1.0000	1.0000	1.0000	1.0000
2	0.9882	0.9850	0.9866	0.9992
3	1.0000	1.0000	1.0000	1.0000
4	1.0000	1.0000	1.0000	1.0000
5	1.0000	1.0000	1.0000	1.0000
6	0.9968	0.9958	0.9963	0.9997
7	1.0000	1.0000	1.0000	1.0000
8	0.9968	0.9939	0.9953	0.9996
9	–	–	–	–
10	0.9869	0.9843	0.9856	0.9990
11	1.0000	1.0000	1.0000	1.0000
12	1.0000	1.0000	1.0000	1.0000
13	0.9996	1.0000	0.9998	1.0000
14	1.0000	1.0000	1.0000	1.0000
15	1.0000	1.0000	1.0000	1.0000
16	1.0000	0.9996	0.9998	0.9998
17	1.0000	1.0000	1.0000	1.0000
18	1.0000	1.0000	1.0000	1.0000
19	0.9997	0.9889	0.9943	0.9878

performance. Meanwhile, the proposed scheme is more cost-friendly and has lower risk of over fitting due to its simple structure.

4.3. **Simulation results for IEEE 118-bus system.** Table 5 shows the architecture of LD-MMoE used in this subsection. The number of trainable parameters is 121,044, which is about 7.31 times of that in Subsection 4.2. It is worth noting that the number of measurements in IEEE 118-bus system is 180, which is 9.47 times of 19 in IEEE 14-bus system. The growth rate of the trainable parameters of the proposed scheme is slower than that of the system scale, so LD-MMoE is suitable for large-scale smart grid in reality.

Figure 2 shows the learning curve of LD-MMoE for IEEE 118-bus system. The curves of training loss and validation loss are also very smooth, which proves that LD-MMoE is suitable for large-scale smart grid. Figure 3 shows the histogram of *Precision*, *Recall*, *F1-score*, *AUC* of LD-MMoE for IEEE 118-bus system, respectively. The four metrics of the proposed scheme are higher than those in [7], even slightly higher than those for

TABLE 5. Architecture of LD-MMoE for IEEE 118-bus system

Stage	Type	Neurons	Output size	Trainable parameters
0	Input	180	180×1	0
1	FullyConn	128	128×1	23,168
2	Experts	4×8	4×8	4,128
3	Towers	180×8	180×8	93,748
4	Sigmoid	180	180×1	0

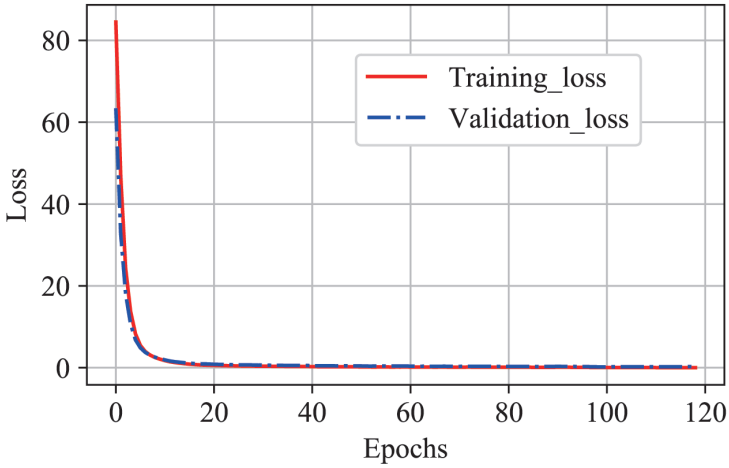

FIGURE 2. The learning curve of LD-MMoE for IEEE 118-bus system

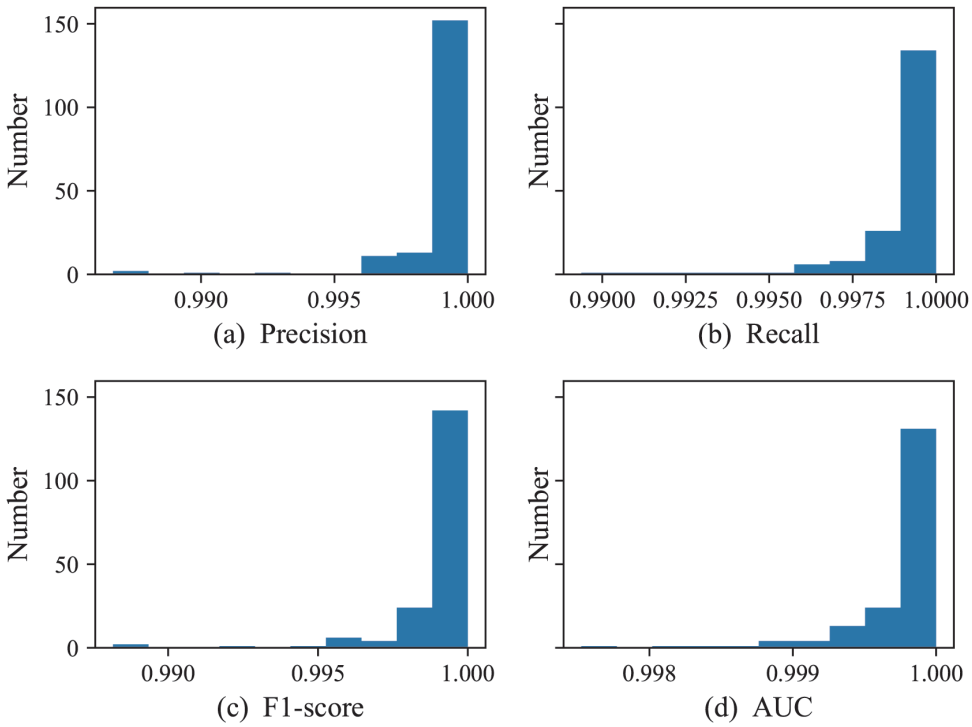

FIGURE 3. The histograms of metrics for LD-MMoE in IEEE 118-bus system

IEEE 14-bus system in Subsection 4.2. Since larger systems have more measurements and more information can be used to train the model, above simulations show the LD-MMoE performs better in larger system.

5. **Conclusions.** This paper has formulated the locational detection of FDI attacks as a multi-task classification problem and designed an LD-MMoE scheme as a multi-task classifier. In order to effectively extract the strong correlation between all measurements related to the attacked states, a shared bottom is combined with the original MMoE model. Simulations in the IEEE 14-bus system and 118-bus system show that the training process of LD-MMoE in large-scale smart grid are very stable and the growth rate of the model's trainable parameters is slower than that of the system scale, which prove that LD-MMoE performs well for large-scale smart grid. In future work, the topology information of the system will be used to efficiently extract information for locational detection of data integrity attacks.

Acknowledgment. This work is supported in part by the National Natural Science Foundation of China under Grant 61903292, the Natural Science Foundation of Shaanxi Province under Grant 2019JQ-084, the China Post-Doctoral Science Foundation under Grant 2020T130515, and the Fundamental Research Funds for the Central Universities under Grant xzy012019040.

REFERENCES

[1] Y. Sun, X. Zhou and G. Yang, Location sensitive multi-task oriented service composition for cyber physical systems, *International Journal of Innovative Computing, Information and Control*, vol.14, no.3, pp.1057-1077, 2018.

[2] G. Liang, S. R. Weller, J. Zhao, F. Luo and Z. Y. Dong, The 2015 Ukraine blackout: Implications for false data injection attacks, *IEEE Trans. Power Systems*, vol.32, no.4, pp.3317-3318, 2017.

[3] A. Monticelli, *State Estimation in Electric Power Systems, A Generalized Approach*, Springer, Boston, MA, 1999.

[4] G. Hug and J. Giampapa, Vulnerability assessment of AC state estimation with respect to false data injection cyber-attacks, *IEEE Trans. Smart Grid*, vol.3, no.3, pp.1362-1370, 2012.

[5] A. Sanjab and W. Saad, Data injection attacks on smart grids with multiple adversaries: A game-theoretic perspective, *IEEE Trans. Smart Grid*, vol.7, no.4, pp.2038-2049, 2016.

[6] Y. Yuan, Z. Li and K. Ren, Modeling load redistribution attacks in power systems, *IEEE Trans. Smart Grid*, vol.2, no.2, pp.382-390, 2011.

[7] S. Wang, S. Bi and Y.-J. A. Zhang, Locational detection of the false data injection attack in a smart grid: A multilabel classification approach, *IEEE Internet of Things Journal*, vol.7, no.9, pp.8218-8227, 2020.

[8] Y. Liu, P. Ning and M. K. Reiter, False data injection attacks against state estimation in electric power grids, *ACM Trans. Information and System Security*, vol.14, no.1, p.13, 2011.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1-9, 2015.

[10] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong and E. H. Chi, Modeling task relationships in multi-task learning with multi-gate mixture-of-experts, *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.1930-1939, 2018.

[11] G. Hinton, O. Vinyals and J. Dean, Distilling the knowledge in a neural network, *arXiv Preprint*, arXiv: 1503.02531, 2015.

[12] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni and H. V. Poor, Machine learning methods for attack detection in the smart grid, *IEEE Trans. Neural Networks and Learning Systems*, vol.27, no.8, pp.1773-1786, 2016.

[13] R. D. Zimmerman, C. E. Murillo-Sánchez and R. J. Thomas, MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education, *IEEE Trans. Power Systems*, vol.26, no.1, pp.12-19, 2011.

[14] S. Bi and Y. J. Zhang, Using covert topological information for defense against malicious attacks on dc state estimation, *IEEE Journal on Selected Areas in Communications*, vol.32, no.7, pp.1471-1485, 2014.