2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT) | 979-8-3315-3193-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/CSNT64827.2025.10967688

Detection of False Data Injection Attacks in Power Grids Based on Spatiotemporal Feature Fusion

Bin Li
School of Computer Science
Northeastern Electric Power University
Jilin, China
2202200992@neepu.edu.cn

***Abstract*—With the development of new power systems, fast and accurate detection of False Data Injection Attacks (FDIA) is crucial for the secure operation of power grids. Existing FDIA detection models based on spatiotemporal correlations have poor feature extraction capabilities and suffer from feature shift, facing the issue of disrupted inherent spatiotemporal correlations in measurement data. To address this, we propose a detection model based on adaptive fusion of spatiotemporal features. First, Graph Convolutional Networks (GCN) is used to extract static spatial features from the power grid topology, while Graph Attention Networks (GAT) captures the dynamic spatial features. Next, Long Short-Term Memory (LSTM) is employed to analyze the temporal variations and extract temporal features. Finally, the extracted spatiotemporal information is projected into feature space for alignment, and feature fusion is performed using a bilinear attention mechanism. The fused features are then used for FDIA detection. Experimental results show that, compared to existing detection models, the proposed model performs better in terms of detection accuracy and robustness.**

Keywords-Smart Grid; False Data Injection Attack; Attack Detection; Spatiotemporal Correlation.

I. INTRODUCTION

With the development of new power systems, advanced information and communication technologies have been widely applied, driving the evolution of power systems into tightly coupled cyber-physical systems[1]. However, the communication functions of smart sensors, actuators, and relay nodes expose potential security vulnerabilities, creating opportunities for attackers. As a result, network security on the information side has become increasingly important in the operation of power systems. FDIA is a type of network attack designed for state estimation[2].By injecting carefully crafted false data into the grid through information gathering devices, attackers can bypass Bad Data Detection (BDD), manipulate measurement results, and mislead the power system into making erroneous decisions, thereby threatening the safe operation of the grid[3].

Currently, FDIA detection methods that mine the implicit spatiotemporal correlations in measurement data are widely applied. In reference [4], a graph network model is designed for different topologies, using a multi-graph mechanism and temporal correlation layers to extract relevant features of FDIA data. Reference [5] employs an autoencoder to extract the spatiotemporal correlations of sensor data for FDIA detection. Reference [6] proposes an unsupervised learning model based on a Long Short-Term Memory autoencoder to extract spatial and temporal features from measurement data for FDIA detection, by evaluating the reconstruction residuals of each measurement sample.

Existing FDIA detection models have the following shortcomings: On one hand, models based on power grid topology spatial correlations only allow information to flow along predefined paths, ignoring the dynamic spatial correlations of measurement data. On the other hand, spatiotemporal correlation detection models typically only extract features serially or simply concatenate them, lacking in-depth exploration of the coupling mechanism of spatiotemporal features.

Therefore, this paper proposes a temporal-spatial feature fusion-based FDIA detection model (TSFF) for power grids. The model captures dynamic and static spatial features using GCN and GAT, extracts temporal features with LSTM, and aligns spatiotemporal information before fusing it via bilinear attention to achieve efficient detection. Simulation experiments on IEEE-14 and IEEE-118 bus systems demonstrate that the proposed model outperforms existing methods in terms of detection accuracy and robustness.

II. DESCRIPTION OF FDIA-RELATED ISSUES

A. State Estimation and Bad Data Detection

State estimation is a crucial component of smart grids, as it estimates the state of the power system based on measurement data and power system models to obtain a reliable system state. The system state values obtained from state estimation are typically used in areas such as economic dispatch and optimal power flow with security constraints. In AC systems, the measurement equation for state estimation is as follows:

$$z_t = h(x_t) + e_t \tag{1}$$

In the formula: $z_t \in \mathbb{R}^n$ is the measurement vector composed of aggregated measurement data, n is the number of measurements; $x_t = [\theta^T, V^T]^T \in \mathbb{R}^m$ is the state vector of the power grid at the current time, θ and V are the voltage phase angles and magnitudes of the nodes, respectively,

979-8-3315-3193-5/25/$31.00 ©2025 IEEE
DOI: 10.1109/CSNT.2025.61

where $m = 2N$ and N are the number of nodes in the network; $\boldsymbol{e}_t = [\boldsymbol{e}_1, \boldsymbol{e}_2, ...\boldsymbol{e}_n] \in \mathbb{R}^n$ is the measurement noise; and $h(\cdot)$ is the nonlinear function that relates the measurements to the state vector, determined by the topology and parameters of the specific power grid.

The state estimation vector is typically solved using the weighted least squares method for minimization:

$$\hat{\boldsymbol{x}}_t = \underset{x}{\operatorname{argmin}} [\boldsymbol{z}_t - h(\boldsymbol{x}_t)]^{\mathrm{T}} \boldsymbol{W} [\boldsymbol{z}_t - h(\boldsymbol{x}_t)] \tag{2}$$

In the formula: $\boldsymbol{W}$ is the weight diagonal matrix, with its diagonal elements being the weight coefficients of the corresponding measurements.

Due to random interference and occasional faults during measurement and transmission, bad data inevitably appears in the measurement data. Therefore, a residual-based BDD is needed to remove the bad data. The residual $\boldsymbol{r}_t$ is defined as the $L2$-norm of the difference between the observed measurements $\boldsymbol{z}_t$ and the measurements obtained from the estimated state vector $\hat{\boldsymbol{x}}_t$, represented as:

$$\boldsymbol{r}_t = \left\| \boldsymbol{z}_t - h(\hat{\boldsymbol{x}}_t) \right\|_2 \tag{3}$$

The detection residual $\boldsymbol{r}_t$ is compared with a predefined detection threshold $\boldsymbol{\tau}_0$. If $\boldsymbol{r}_t < \boldsymbol{\tau}_0$, the measurement data is considered normal; otherwise, it is deemed to contain bad measurement data.

B. *Principle of False Data Injection Attacks*

Attackers can tamper with the measurement data collected by terminal instruments and bypass the BDD, leading to incorrect state estimation, which may cause the control center to make wrong decisions.

When false data injection occurs, the measurement equation for state estimation becomes:

$$\boldsymbol{z}_{ta} = \boldsymbol{z}_t + \boldsymbol{a}_t \tag{4}$$

In the formula: $\boldsymbol{a}_t \in \mathbb{R}^n$ is the injected attack vector; $\boldsymbol{z}_{ta} \in \mathbb{R}^n$ is the measurement value observed at the current time, obtained by adding the attack vector $\boldsymbol{a}_t$ to the original measurement vector $\boldsymbol{z}_t$.

Define $\hat{\boldsymbol{x}}_{ta} \in \mathbb{R}^m$ as the state estimated from the measurement vector $\boldsymbol{z}_{ta}$ after false data injection, and $\hat{\boldsymbol{x}}_t \in \mathbb{R}^m$ as the state estimated from the original, unaltered measurement vector $\boldsymbol{z}_{ta}$. Here:

$$\hat{\boldsymbol{x}}_{ta} = \hat{\boldsymbol{x}}_t + \boldsymbol{c}_t \tag{5}$$

In the formula: $\boldsymbol{c}_t \in \mathbb{R}^m$ is the state estimation error caused by the attack vector.

The construction of the attack vector $\boldsymbol{a}_t$ is as follows:

$$\boldsymbol{a}_t = h(\hat{\boldsymbol{x}}_t + \boldsymbol{c}_t) - h(\hat{\boldsymbol{x}}_t) \tag{6}$$

Therefore, the detection residual of the BDD after the attack injection is:

$$\begin{aligned} \boldsymbol{r}_a &= \left\| \boldsymbol{z}_{ta} - h(\hat{\boldsymbol{x}}_{ta}) \right\|_2 \\ &= \left\| \boldsymbol{z}_{ta} + \boldsymbol{a}_t - h(\hat{\boldsymbol{x}}_t + \boldsymbol{c}_t) \right\|_2 \\ &= \left\| \boldsymbol{z}_{ta} + h(\hat{\boldsymbol{x}}_t + \boldsymbol{c}_t) - h(\hat{\boldsymbol{x}}_t) - h(\hat{\boldsymbol{x}}_t + \boldsymbol{c}_t) \right\|_2 \\ &= \boldsymbol{r}_t \end{aligned} \tag{7}$$

In this case, since the residuals of the system before and after the attack injection do not change, the injected false data ideally bypasses the residual detection.

III. Grid FDIA Detection Model Based on Temporal-Spatial Feature Fusion

The overall structure of the proposed detection model is shown in Figure 1. The model's input is the measurement dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, which contains n samples with d-dimensional features. The model's output is the label set $\boldsymbol{Y} \in \mathbb{R}^n$, where each label is represented by a binary value: 0 indicates the sample is in a normal state, and 1 indicates the sample is under attack.

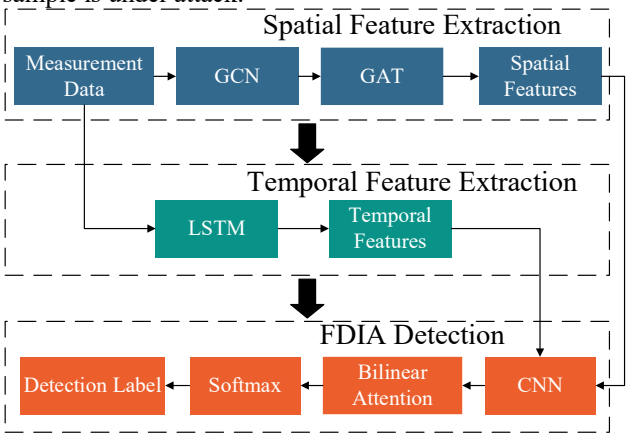

Figure 1. Structure of the proposed detection model

A. *Spatial Feature Extraction Based on GCN-GAT*

Measurement data is influenced not only by the static power grid topology but also by dynamic associations between measurements at different nodes over time, which are constrained by information transmission paths. To accurately analyze the spatial relationships between nodes in the power grid, a spatial feature extraction method based on GCN-GAT is proposed.

Given the input measurement dataset $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ to be detected, where $n = N \times D$, N is the number of nodes in the input graph, and D is the feature dimension of each node.

Firstly, GCN is used to extract static spatial features from the power grid topology, with the adjacency matrix as input to reflect the topological correlation between different measurement devices. Each hidden layer of the GCN is represented by the following nonlinear function:

$$H^{(l+1)} = f(H^{(l)}, \boldsymbol{A}) \tag{8}$$

In the formula: $H^{(0)} = \boldsymbol{X}$ is the input layer; $H^{(L)} = \boldsymbol{Z} \in \mathbb{R}^{n \times d}$ is the output layer; l represents the number of layers; and $\boldsymbol{A}$ is the adjacency matrix of the power grid for the measurement data to be detected.

Where $f(\cdot,\cdot)$ specifically refers to:

$$f(H^{(l)}, \boldsymbol{A}) = \text{Leaky ReLU}(\hat{\boldsymbol{D}}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\hat{\boldsymbol{D}}^{-\frac{1}{2}}H^{(l)}\boldsymbol{W}_S^{(l)}) \tag{9}$$

In the formula: $\boldsymbol{W}_S^{(l)}$ is the weight matrix for the linear transformation; Leaky ReLU is a activation function.; $\boldsymbol{I}$ is the identity matrix; $\hat{\boldsymbol{D}}$ is the degree matrix of $\hat{\boldsymbol{A}}$.

Subsequently, GAT is introduced to model the dynamic spatial correlations of real-time measurement data. GAT effectively captures dynamic correlations in the topological structure by adaptively adjusting the information aggregation between nodes and their neighbors. Taking adjacent nodes *i* and *j* as an example, where $\boldsymbol{z}_i, \boldsymbol{z}_j \subseteq \boldsymbol{Z}$, the attention of node *j* to node *i* and the normalized attention α_{ij} are computed as follows:

$$\begin{cases} e(z_i, z_j) = \text{Leaky ReLU}(u^{\mathrm{T}}[Qz_i \,\|\, Qz_j]) \\ \alpha_{ij} = \sigma(e(z_i, z_j)) = \dfrac{\exp(e(z_i, z_j))}{\sum_{j' \in \Omega_i} \exp(e(z_i, z_{j'}))} \end{cases} \tag{10}$$

In the formula: $\|$ represents the concatenation of node features; u and Q are learnable weight vectors; $\sigma(\cdot)$ is the Sigmoid activation function; and Ω_i is the set of neighboring nodes connected to node *i*.

The output of each node is:

$$x_{i,out} = \sigma(\sum_{j \in \Omega_i} \alpha_{ij} \boldsymbol{Q} z_j) \tag{11}$$

In the formula: $x_{i,out}$ is the output of node i, which has the same dimensionality as the input.

To make the GAT learning process more stable, multiple layers of graph attention are used to effectively extract spatial features from the measurement data, and the information from different layers is averaged for integration. Finally, the extracted spatial features are output as $\boldsymbol{Fs} \in \mathbb{R}^{n \times F}$, where F is the dimensionality of the output features.

B. Temporal Feature Extraction Based on LSTM

When processing power grid measurement time series data, the LSTM model is used for extraction to avoid the loss of temporal information. LSTM can improve the detection of long-term dynamic changes caused by attacks. Given the input as the measurement data set $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, consisting of d time series of measurements with length n, the output is the extracted temporal features $\boldsymbol{Ft} \in \mathbb{R}^{n \times F}$, where F is the dimensionality of the output features.

LSTM includes four main gating mechanisms: the input gate i_t receives external data $\boldsymbol{x}_t \subseteq \boldsymbol{X}$ and decides which information to store in memory; the forget gate f_t selects which memory from the previous time step to discard; the input modulation gate $\tilde{C}_t$ generates new candidate memory and updates the state; and the output gate o_t decides which memories affect the current output.

The gating mechanisms effectively store and access the time-varying information in sequential data, addressing issues such as gradient explosion and vanishing gradients in Recurrent Neural Networks. The computation formulas are as follows:

$$\begin{cases} i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, \boldsymbol{x}_t] + b_i) \\ f_t = \sigma(\boldsymbol{W}_f \cdot [h_{t-1}, \boldsymbol{x}_t] + b_f) \\ \tilde{C}_t = \tanh(\boldsymbol{W}_C \cdot [h_{t-1}, \boldsymbol{x}_t] + b_C) \\ o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, \boldsymbol{x}_t] + b_o) \end{cases} \tag{12}$$

In the formula: $\boldsymbol{W}_i$, $\boldsymbol{W}_f$, $\boldsymbol{W}_C$ and $\boldsymbol{W}_o$ are the weight matrices for the input gate, forget gate, input modulation gate, and output gate, respectively; b_i, b_f, b_C and b_o are the bias terms for the input gate, forget gate, input modulation gate, and output gate, respectively; and tanh is the hyperbolic tangent activation function.

The state update formula for the memory cell is:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{13}$$

In the formula: $\odot$ represents element-wise multiplication. The output h_t is calculated by the following formula:

$$h_t = o_t \odot \tanh(C_t) \tag{14}$$

C. FDIA Detection Based on Adaptive Alignment and Fusion of Spatiotemporal Features

Temporal and spatial features are often interrelated, and simply stacking them may cause information loss, hindering the full utilization of their relationships. Therefore, a spatiotemporal feature alignment fusion method is proposed to optimize the use of extracted features.

First, the temporal feature $\boldsymbol{Ft}$ and spatial feature $\boldsymbol{Fs}$ are concatenated, and a convolutional neural network is used to project them into the spatiotemporal feature space. Then, the multi-head attention mechanism is applied to map the original spatial features and temporal features into the spatiotemporal feature space separately.

Next, the mapped features are concatenated with the original features and processed through two output networks to obtain the aligned spatiotemporal output features:

$$\begin{cases} \boldsymbol{F}_S = Conv(\text{MultiHead}(\boldsymbol{Ft}, \Omega) \,\|\, \boldsymbol{Fs}) \\ \boldsymbol{F}_T = Conv(\text{MultiHead}(\boldsymbol{Fs}, \Omega) \,\|\, \boldsymbol{Ft}) \end{cases} \tag{15}$$

In the formula: $\|$ is the concatenation of features; $\text{MultiHead}(\cdot,\cdot)$ is the function used to compute the spatiotemporal feature space mapping with the multi-head attention mechanism, ensuring element alignment; Ω is the spatiotemporal feature space obtained from the convolutional neural network; and $Conv(\cdot)$ represents the convolution.

Then, the spatiotemporal feature weights are calculated using the Softmax and bilinear functions:

$$\begin{cases} \boldsymbol{w}_S = \text{Softmax}(\boldsymbol{F}_S{}^{\mathrm{T}} \boldsymbol{W}_S (\boldsymbol{F}_S + \boldsymbol{F}_T)) \\ \boldsymbol{w}_T = \text{Softmax}(\boldsymbol{F}_T{}^{\mathrm{T}} \boldsymbol{W}_T (\boldsymbol{F}_S + \boldsymbol{F}_T)) \end{cases} \tag{16}$$

In the formula: $\boldsymbol{W}_S$ and $\boldsymbol{W}_T$ are learnable matrices representing the bilinear relationship between input vectors.

Finally, the fused final features are obtained by weighted summation with residual connections:

$$\boldsymbol{F}_{fusion} = \boldsymbol{F}_S \boldsymbol{w}_S + \boldsymbol{F}_T \boldsymbol{w}_T + \boldsymbol{F}_S + \boldsymbol{F}_T \tag{17}$$

The fused features are then input into a fully connected layer to extract higher-level features. Subsequently, the output is transformed into a class probability distribution through the Softmax layer, completing the detection task. To address class imbalance during training, a weighted binary cross-entropy loss assigns greater focus to minority class samples, enhancing detection performance.

IV. Experimental Validation and Analysis

A. Dataset Introduction

Real-time load data from the New York Independent System Operator is used to simulate the operating state. The normalized load data is injected into the system, and power flow is calculated using the Newton-Raphson method to obtain measurements like node injection power, branch power, and node voltage magnitudes.

Assuming the attacker has full information about the IEEE-14 node system and partial information about the IEEE-118 node system, FDIA is implemented using the attack models from [7] (partial information) and [8] (full information). The normal-to-attack sample ratio is 10:1, and the task is framed as a binary classification with labels "0" for normal and "1" for attack. The measurement noise follows a Gaussian distribution with a mean of 0 and a standard deviation of 0.02.

B. Experimental Results Analysis

TSFF is compared with existing FDIA detection models based on the temporal-spatial correlation of measurement data, including ST-Transformer[9], LSTM-GCN[10], and LSTM-AE[6].

TABLE I. Comparison of Detection Models in the IEEE-14 Node System

Model	Precision	Recall	F1-score	Accuracy
ST-Transformer	0.8367	0.9664	0.8969	0.9778
LSTM-GCN	0.9341	0.9838	0.9583	0.9914
LSTM-AE	0.9575	0.9907	0.9738	0.9947
TSFF	0.9685	0.9925	0.9804	0.9951

TABLE II. Comparison of Detection Models in the IEEE-118 Node System

Model	Precision	Recall	F1-score	Accuracy
ST-Transformer	0.8298	0.9595	0.8901	0.9763
LSTM-GCN	0.8557	0.9884	0.9173	0.9822
LSTM-AE	0.8756	0.9942	0.9312	0.9853
TSFF	0.9227	0.9723	0.9398	0.9869

Table I compares various detection models in the IEEE-14 node system, where TSFF demonstrates the best detection accuracy. TSFF also outperforms other models in other metrics, particularly in Recall, which reached 0.9925.

As shown in Table II, the performance of various detection models in the IEEE-118 node system shows significant fluctuations. This phenomenon is due to the more complex data dimensions and attack methods in the IEEE-118 bus system, which place higher demands on the model's feature learning ability. Nevertheless, the proposed model still demonstrates excellent overall performance.

The Receiver Operating Characteristic (ROC) curve shows the relationship between the true positive rate and false positive rate. A larger area under the curve (AUC) indicates better model performance. Figure 2 presents the FDIA detection ROC curves for the proposed model in the IEEE-14 and IEEE-118 node systems, with AUC values highlighting its excellent detection performance.

The robustness of the proposed detection model under different noise intensities (0.02, 0.04, 0.06, 0.08, and 0.10) is validated in the IEEE-14 bus system and the IEEE-118 node system, as shown in Figure 3.

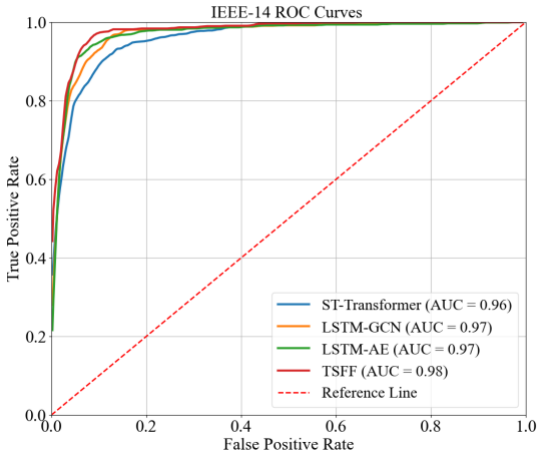

Figure 2. ROC curves of detection models in different systems

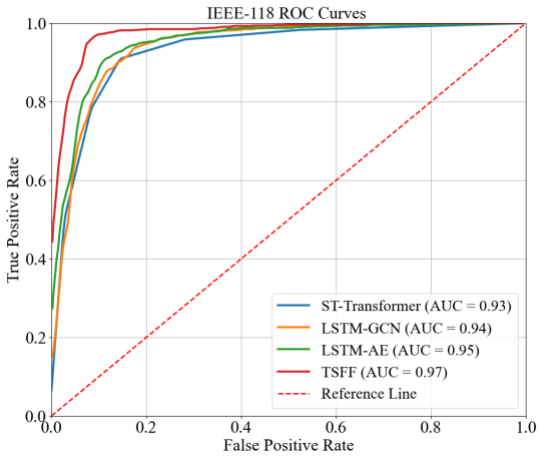

Figure 3. F1-Score of detection models under varying noise levels in different systems

As shown in Figure 3, with the increase in the standard deviation of measurement noise, the F1-score of the TSFF in the IEEE-14 node system shows a gradual decline, while the F1 scores of other detection methods drop more sharply. This indicates that the TSFF model has stronger noise resistance and more robust detection performance. As the noise level increases, the proposed method effectively mitigates the negative impact of noise through spatiotemporal feature alignment, maintaining better detection performance. It can be seen that with the increase in detection complexity, the F1 scores of all models decrease, but relatively speaking, the TSFF model demonstrates better noise resistance.

V. CONCLUSION

This paper proposes a spatiotemporal fusion-based power grid FDIA detection model to address the poor feature extraction and feature shift in existing models. The model captures dynamic and static spatial features and temporal features from power grid measurement data, comprehensively capturing complex spatiotemporal correlations. It then adaptively aligns and fuses these features for FDIA detection. Experimental results demonstrate the model's superior detection accuracy and robustness.

Although the model excels in detection accuracy and robustness, the increased time for graph structure generation with more nodes limits real-time detection. Future research will focus on FDIA detection in complex networks, including analyzing network structures to classify threat levels and prioritize vulnerable node detection.

REFERENCES

[1] Wang Y, Zhang Z, Ma J, et al. KFRNN: An effective false data injection attack detection in smart grid based on Kalman filter and recurrent neural network[J]. IEEE Internet of Things Journal, 2021, 9(9): 6893-6904.

[2] Wang S, Bi S, Zhang Y J A. Locational detection of the false data injection attack in a smart grid: A multilabel classification approach[J]. IEEE Internet of Things Journal, 2020, 7(9): 8218-8227

[3] Han Y, Feng H, Li K, et al. False data injection attacks detection with modified temporal multi-graph convolutional network in smart grids[J]. Computers & Security, 2023, 124: 103016

[4] Boyaci O, Narimani M R, Davis K R, et al. Joint detection and localization of stealth false data injection attacks in smart grids using graph neural networks[J]. IEEE Transactions on Smart Grid, 2021, 13(1): 807-819.

[5] Aboelwafa M M N, Seddik K G, Eldefrawy M H, et al. A machine-learning-based technique for false data injection attacks detection in industrial IoT[J]. IEEE Internet of Things Journal, 2020, 7(9): 8462-8471.

[6] Musleh A S, Chen G, Dong Z Y, et al. Attack detection in automatic generation control systems using LSTM-based stacked autoencoders[J]. IEEE Transactions on Industrial Informatics, 2022, 19(1): 153-165.

[7] Liu X, Li Z. False data attacks against AC state estimation with incomplete network information[J]. IEEE Transactions on smart grid, 2016, 8(5): 2239-2248.

[8] Liu X, Li Z. Local load redistribution attacks in power systems with incomplete network information[J]. IEEE Transactions on Smart Grid, 2014, 5(4): 1665-1676.

[9] Li X, Hu L, Lu Z. Detection of false data injection attack in power grid based on spatial-temporal transformer network[J]. Expert Systems with Applications 2024; 238: 121706.

[10] Li H, Dou C, Yue D, et al. End-edge-cloud collaboration based false data injection attack detection in distribution networks[J]. IEEE Transactions on Industrial Informatics 2024; 20(2): 1786-1797.