Dependency-Aware GraphSAGE-Based Interpretable FDIA Detection Using BiLSTM With SE-Attention in Smart Grids

Siming Huang, Fengyong Li, Kunzhan Li, Xiangjing Su, *Senior Member, IEEE*,
and Zhao Yang Dong, *Fellow, IEEE*

***Abstract*—False data injection attacks (FDIAs) refer to attackers exploiting vulnerabilities in the detection of bad data in smart grid energy management systems to maliciously manipulate state estimation results in the cyber–physical system, resulting in unstable operation of the power system. Existing deep learning-based detection schemes often fail to capture the spatial topology features and long-term dependencies in power grid data well. Meanwhile, the complexity of deep detection models makes them a "difficult-to-interpret system," reducing the credibility of detection results. To address the aforementioned challenges, this article presents a dependency-aware deep interpretable FDIA detection model. The proposed model first introduces graph sample and aggregate (GraphSAGE) network to extract spatial topological features, which are used to represent the deep spatial topological dependencies of adjacent data nodes. Subsequently, we build a Bidirectional Long Short-Term Memory (BiLSTM) network with a Squeeze-and-Excitation (SE) attention module, which can efficiently aggregate attack characteristics and long-term dependency information by dynamically capturing the potential correlations between FDIAs detection and measurement data. Furthermore, the SHapley Additive exPlanations (SHAP) method is used to demonstrate the interpretability of the model in the spatial–temporal dimensions and then provide the basis for high-precision detection results. A series of extensive experiments are carried out over the IEEE 14-bus and 118-bus test systems. The experimental results demonstrate that the proposed model presents a superior overall performance comparing with several state-of-the-art FDIA detection models, and provides reasonable interpretability from the spatial–temporal dimensions.**

***Index Terms*—Bidirectional Long Short-Term Memory (BiLSTM), FDIA, graph sample and aggregate (GraphSAGE), interpretable deep learning, smart grid.**

I. Introduction

WITH the advancement of smart sensors and communication technologies, the power grid has progressively transformed into an intelligent cyber–physical system (CPS), in which the real-time monitoring, control, and distributed energy management can be achieved well. While this transformation enhances the reliability of power supply, the inherent openness of such a system also makes it susceptible to various cyber-attacks [1]. A notable instance is the large-scale power outage in Ukraine [2], which was caused by a cyber-attack, resulting in significant damage. Currently, the cyber threats in smart grid mainly include distributed denial of service attack [3], data spoofing attack [4], packet analysis attack [5], man-in-the-middle attack [6], and false data injection attack (FDIA) [7]. Among these, FDIA is particularly threatening, as it allows attackers to compromise the Supervisory Control and Data Acquisition (SCADA) system, and it is difficult to detect directly. For instance, attackers typically inject false data into sensors via communication networks [8], leading to discrepancies between the power grid's state estimation and its actual state [9]. Considering the threat of this attack, it is essential to conduct research on FDIA detection to ensure the security and stability of the power grid.

In general, FDIA detection methods can generally be categorized into model-based and data-driven approaches. Model-based detection involves identifying abnormal conditions by constructing a mathematical model of the power system. For example, Qu et al. [10] proposed an FDIA detection method based on the Hellinger distance, which can detect anomalies by tracking the dynamic characteristics of grid measurement data over time. Wei and Zhang [11] utilized an improved unscented Kalman filter (UKF), demonstrating high detection accuracy, particularly when applied to dynamic models of the power grid. Shen and Qin [12] applied random matrix theory to effectively detect FDIA. While model-based methods are capable of detecting FDIA with high accuracy, their performance is heavily dependent on the accuracy of system parameters, even small variations in these parameters may result in a significant decline in detection effectiveness. In contrast, data-driven detection methods do not rely on specific

Received 19 April 2025; accepted 4 June 2025. Date of publication 10 June 2025; date of current version 8 August 2025. This work was supported in part by the Engineering Research Center of Offshore Wind Technology Ministry of Education (Shanghai University of Electric Power); in part by the JC STEM Lab of Future Energy Systems under Grant 2025-0039; in part by the Global STEM Professorship under Grant GSP313; and in part by the Startup Grant of City University of Hong Kong (Data Driven Real Time Smart Energy Management System Supporting Energy Transition). *(Corresponding author: Fengyong Li.)*

Siming Huang and Kunzhan Li are with the College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China (e-mail: fyli@shiep.edu.cn).

Fengyong Li is with the College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China, and also with the Engineering Research Center of Offshore Wind Technology Ministry of Education, Shanghai University of Electric Power, Shanghai 200090, China.

Xiangjing Su is with the College of Electrical Engineering, Shanghai University of Electric Power, Shanghai 201306, China (e-mail: xiangjing_su@126.com).

Zhao Yang Dong is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: zydong@ieee.org).

Digital Object Identifier 10.1109/JIOT.2025.3578369

mathematical models of the system, instead of extracting features from historical power grid data to identify abnormal behavior. These methods can automatically adapt to dynamic changes within the system. For instance, Sah Tyagi et al. [13] proposed a novel detection method that first obtains optimal features of power grid data using cooperative paired swarm optimization, and subsequently performs FDIA detection with relational vector learning techniques, which yields excellent detection efficiency and accuracy. Another approach, Parizad and Hatziadoniu [14] introduced a noise clustering method combined with principal component analysis (PCA). PCA method was first used to reduce the feature space and then introduced a clustering method to group different types of data. While traditional machine learning methods are powerful, they often require complex feature processing. What's more, due to a limited number of trainable parameters, the traditional methods have constrained learning capacity.

In contrast, deep learning network architectures [15] offer robust feature extraction and learning capabilities. For example, Lu et al. [16] proposed a convolutional neural network (CNN) based on representation learning (RL-CNN), which effectively captures local features in grid data and performs well in detecting and locating FDIA. Zhang et al. [17] introduced a Bidirectional Long Short-Term Memory (BiLSTM) network combined with an improved whale optimization algorithm (IWOA) method, enabling the effective extraction of long-term dependencies for FDIA detection. Li et al. [18] further integrated federated learning with Transformer network, achieving strong performance in both FDIA detection and security. To further mine local features and long-term dependencies in power grid data, Ji et al. [19] combined CNNs with gated recurrent units (GRU) for more effective detection. However, these aforementioned approaches always overlook the spatial topological features presented in power grid data, thus limiting their capability to fully enhance FDIA detection. To fill this gap, FDIA detection schemes [20], [21] based on graph convolutional network (GCN) were widely used, which successfully extracted spatial topological features to improve detection. Chen et al. [22] introduced an enhanced Graph Sample and Aggregate (GraphSAGE) network, which aggregates the features of neighboring nodes similar to the target node to identify FDIA. To overcome performance degradation from power grid topology changes, Li et al. [23] developed a gated graph neural network with graph attention (GGNN-GAT) FDIA detection method. Su et al. [24] further proposed a dual-attention multihead graph attention (DAMGAT) network to further improve the model's capability to represent spatial topological features. Although the methods mentioned above have already involved the spatial topology relationship of power sensor nodes in CPSs, they almost do not consider the cascading relationship between different nodes and ignore the spatial cascading topology features of different nodes. For example, when a node is attacked, the state changes in that node and other remote nodes can be linked in complex, nonlinear ways. This naturally hinders further improvement of detection performance.

Additionally, deep learning-based network models are often considered as "difficult-to-interpret system" that may undermine the model's credibility in FDIA detection, as it difficult to interpret their internal decision-making mechanisms. Unfortunately, the related research on the interpretability of FDIA detection models remains limited [25]. For example, in [24], some interpretability for the features and spatial dimensions in the model for FDIA detection was provided by employing feature attention modules and spatial topology attention modules. However, attention-based interpretability primarily relies on the local weight distribution within the model, which not fully capture the contribution of global features to prediction outcomes regarding the entire model. At the same time, this technique lacks a solid theoretical foundation.

Facing the aforementioned problems, we are thus motivated to design an efficient dependency-aware deep interpretable network architecture in the context of FDIA detection, which makes the following novel contributions.

1) We propose an efficient FDIA detection scheme by designing a dependency-aware deep interpretable network architecture, where the spatial cascading topology features and long-term dependencies between nodes can be efficiently captured. The proposed scheme can achieve an effective balance in terms of detection performance, interpretability, and robustness.
2) A dependency-aware deep interpretable network framework is reconstructed by introducing GraphSAGE network, which can effectively represent the deep spatial cascading topological dependencies of adjacent multiple data nodes.
3) We further build a BiLSTM network with a Squeeze-and-Excitation (SE) attention module, which can efficiently aggregate attack characteristics and long-term dependency information by dynamically capturing the potential correlations between FDIAs detection and measurement data.
4) Comprehensive experiments are performed over multiple classical IEEE bus test systems, and the experimental results demonstrate that the proposed model presents a superior overall performance comparing with several state-of-the-art FDIA detection schemes, and show the reasonable interpretability from the spatial–temporal dimensions.

The remainder of this article is organized as follows. Section II reviews the preliminary about state estimation of smart gird and FDIA. In Section III, we describe the proposed dependency-aware deep interpretable FDIA detection framework and present its design details. Comprehensive experiments are performed to evaluate the performance of the proposed scheme. The experimental results and corresponding discussion are sequentially presented in Sections IV and V. Finally, Section VI concludes this article.

II. Preliminary

A. Power Grid State Estimation and Bad Data Detection

State estimation can ensure the safety and stability of CPS by analyzing measurement data from the SCADA system [18]. Specifically, the SCADA system collects data from sensors that typically provide power-related measurements of the grid. This data is then used for state estimation in the control center. Fig. 1 illustrates the complete flowchart of a FDIA on the power system. In such an attack, the attacker manipulates the measurement data in the SCADA system, leading to incorrect state estimation results and evading detection by the bad data detection (BDD) mechanism [16], [20], which can significantly affect power generation scheduling and compromise the stability of the power grid. In general, power grid state estimation focuses primarily on phase angle variations by assuming that the voltage amplitude is 1 and neglecting the impact of resistance and grounding branches. Accordingly, the detailed calculation can be shown as follows:

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

where z represents the measurement data collected from the SCADA system, H denotes the Jacobian matrix that encapsulates the physical topology of the power grid, x is the state vector of the system, and e stands for the measurement error vector. After obtaining the measurement data z, the weighted least squares method for state estimation [16] can be calculated by the following:

$$\min F(x) = (z - Hx)^T R^{-1} (z - Hx) \tag{2}$$

where R denotes the covariance matrix of the measurements.

Subsequently, the objective function in (2) is solved to obtain the estimated value $\hat{x}$ of the system state vector

$$\hat{x} = \left(H^T R^{-1} H\right)^{-1} H^T R^{-1} z. \tag{3}$$

To detect the presence of bad data, the measurement residual r is the ℓ_2 norm of the difference between the measured value z and the estimated value $\hat{z}$

$$r = ||z - \hat{z}||_2 = ||z - H\hat{x}||_2 \leq \tau. \tag{4}$$

If $r \leq \tau$, it indicates that there is no bad data in the SCADA system. Otherwise, it suggests the presence of bad data.

B. False Data Injection Attack

When the Jacobian matrix H of the power grid is obtained by the attacker, they can generate the corresponding attack vectors using H, thereby evading detection from the BDD mechanism [24]. The calculation for the attack vector is as follows:

$$\alpha = Hc \tag{5}$$

where $c = [c_1, c_2, \ldots, c_n]^T$ is the deviation of the state variables before and after the attack.

$$z_\alpha = z + \alpha. \tag{6}$$

Then, the formed attack vectors $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_m]^T$ can be injected into the $z = [z_1, z_2, \ldots, z_m]^T$ to form an attacked measurement z_α

$$\hat{x}_\alpha = x + c. \tag{7}$$

Correspondingly, the value of the state variable x is also changed, and the result is shown in (7)

$$r_\alpha = ||z_\alpha - H\hat{x}_\alpha||_2 = ||z + \alpha - H\big(\hat{x} + c\big)||_2$$
$$= ||z + H\big(\hat{x} + c\big) - H\hat{x} - H\big(\hat{x} + c\big)||_2 = r. \tag{8}$$

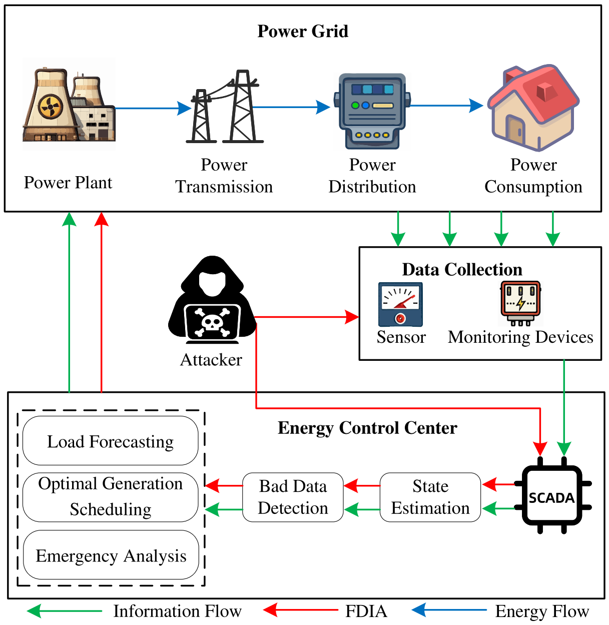


Fig. 1. Standard scenarios of FDIA in smart grid.

Apparently, (8) can conclude that the attack vector α does not change the value of the measurement residual, and can thus efficiently evade detection by the BDD mechanism.

III. Proposed Interpretable Deep Learning Network Architecture

A. Overview of Proposed Interpretable Detection Framework

The proposed detection framework is illustrated in Fig. 2. The data from each node in the power grid is collected through the SCADA system, including the active power P_i, reactive power Q_i, branch active power P_{ij}, and branch reactive power Q_{ij}, forming the measurement vector $z = [P_i, Q_i, P_{ij}, Q_{ij}]$. After obtaining the network topology information, the attacker injects false data to implement a covert attack. If the BDD mechanism fails to detect the false data, the proposed detection model is sequentially employed for further detection. The proposed model first utilizes the GraphSAGE network to effectively extract spatial cascading topological features of different nodes from the grid data. Subsequently, BiLSTM network is applied to capture long-term dependencies across different nodes within the cascaded topological features, and introduces feature-wise SE-Attention mechanism to adaptively adjust feature weights and guide detection models to focus on the compromised features. Finally, the introduction of SHapley Additive exPlanations (SHAP) interpretable techniques [26] further highlights the robustness of the detection model and enhances the credibility of the detection results.

B. Dependency-Aware GraphSAGE

In general, the CPSs in power grid can be conceptualized as a complex graph structure, where each node stands for a distinct power device, such as a generator or load, while the

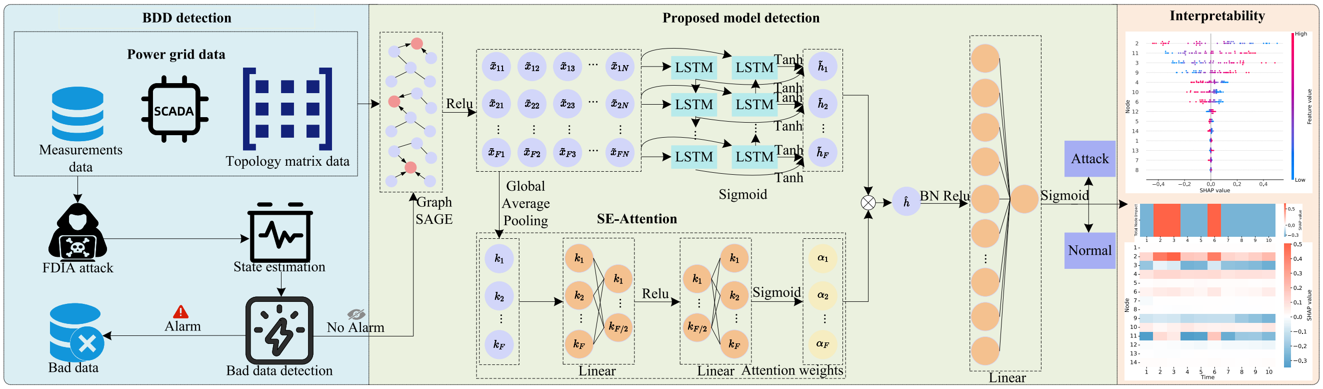


Fig. 2. Proposed deep interpretable FDIA detection model and overall detection architecture.

edges represent the power connections between these devices. Considering that cascading FDIA attacks may affect adjacent nodes by attacking one power node, the edge information in the graph topology can accurately reflect the dependency relationships between different nodes. Accordingly, we introduce a dependency-aware GraphSAGE module, which leverages the connectivity information between these devices to uncover subtle relationships and patterns between different nodes in the CPSs, particularly after encountering FDIA. Different from traditional schemes, the proposed module can learn the entire aggregation function rather than processing each node individually [27], [28], and allows for the sampling and aggregation of neighboring node features.

Specifically, we first use mean aggregation to define the method for aggregating the neighbor features of each node in the lth layer of GraphSAGE

$$o_n^l = \text{mean}\left(\{x_u^{l-1}\ \forall u \in \text{Nb}(n)\}\right) \tag{9}$$

where feature o_n^l represents the averaged features of the neighboring nodes u of node n. $\text{Nb}(n) = \{u\ \forall A_{n,u} = 1\}$ is the neighbor set of node n, where $A \in \hbar^{N \times N}$ denotes the adjacency matrix containing the topological information of the power grid. The specific construction method of A is such that if two nodes in the power grid are directly connected by a transmission line, then $A_{n,u} = 1$; otherwise, $A_{n,u} = 0$.

Subsequently, the features aggregated from the nodes in the lth layer are learned

$$x_n^l = \text{Relu}\left(W_n^l(x_n^{l-1}, o_n^l)\right). \tag{10}$$

During this step, we fuse the feature x_n^{l-1} from the previous layer with the aggregated neighbor feature o_n^l. Correspondingly, the resulting vector is multiplied by the learnable weight matrix W_n^l, and then apply the ReLu activation function to enhance the nonlinear expressive power of the feature.

C. BiLSTM With SE-Attention Mechanism

For the CPS scenario, when a node is performed by FDIA attack, due to the information interaction between the cyber side and the physical side, the state changes between the current attacked node and other remote nodes may become intricately linked. However, traditional standard single-layer LSTM only considers the unidirectional long-term dependency information in a sequence, and does not involve cascading state information in the node sequences, making it difficult for them to capture long-term cross node dependencies between these nodes. Considering the above problem, we introduce the BiLSTM module with the SE attention mechanism, which contains both forward and backward LSTM layers [29], [30]. By integrating bidirectional sequence modeling capabilities, this architecture effectively captures intricate long-term dependency patterns across node data. Meanwhile, SE-Attention mechanism can effectively model the relationships between feature channels, suppresses less important features, and enhances the learning of more relevant features.

In the forward propagation, the BiLSTM module captures the forward long-term dependencies of sequence data from the front to the back

$$\overrightarrow{h_t} = \overrightarrow{\text{LSTM}}(h_{t-1}, \tilde{x}_t, c_{t-1}), t \in [1, F] \tag{11}$$

where h_{t-1} and c_{t-1} represent the output hidden layer and the saved cell state before position t. The input data is the spatial topological features $\tilde{X} = [X^l]^T = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_F\} \in \hbar^{F \times N}$ that extracted by the graph network. Correspondingly, during the backward propagation, the BiLSTM module effectively captures the backward long-term dependencies of sequence data from the back to the front

$$\overleftarrow{h_t} = \overleftarrow{\text{LSTM}}(h_{t+1}, \tilde{x}_t, c_{t+1}), t \in [1, F] \tag{12}$$

where h_{t+1} and c_{t+1} are the output hidden layer and the saved cell state after position t.

Correspondingly, the forward hidden layer $\overrightarrow{h_t}$ and the backward hidden layer $\overleftarrow{h_t}$ are integrated to build the output hidden layer $\tilde{h}_t$ of BiLSTM at position t

$$\tilde{h}_t = \text{Tanh}\left(W_f \overrightarrow{h_t} + W_b \overleftarrow{h_t}\right) \tag{13}$$

where W_f and W_b are learnable weight variables, and *Tanh* is the activation function that can further increase the nonlinear expression ability of the models.

The subsequently introduced SE-Attention enhances the model's ability to capture complex power grid features through the SE mechanism. In the squeeze step, global average pooling

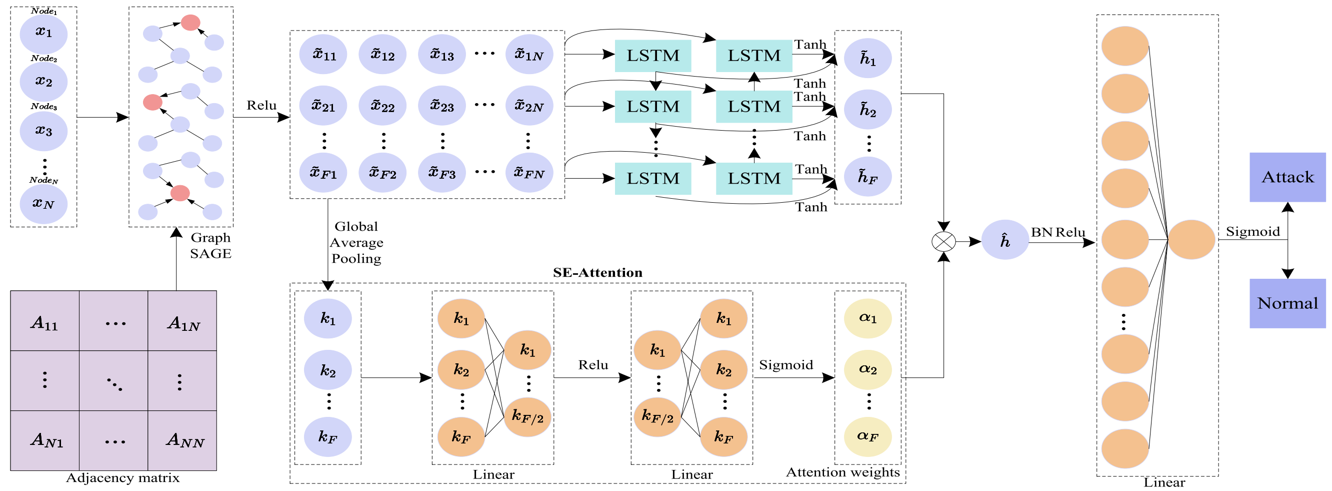


Fig. 3. Detection procedure of the proposed FDIA detection model.

(GAP) compresses node features to generate global channel descriptors, so as to capture the correlation features between cross regional power nodes. In the excitation step, the first linear layer compresses channels to extract key features, while the second linear layer restores dimensions and applies sigmoid to generate channel weights. This enables the model to adaptively highlight channels strongly related to attack behaviors while suppressing those associated with normal conditions or sensor noise, particularly benefiting large-scale power grid. SE-Attention mechanism can be expressed as follows:

$$K = \frac{1}{N}\sum_{i=1}^{N}\tilde{X}(i) \tag{14}$$

where the dimension of $\tilde{X}$ is $F \times N$, and GAP is applied to average the spatial topological features in each of the F channels into a single value. As a result, the dimension of the compressed output $K = \{k_1, k_2, \ldots, k_F\}$ is $F \times 1$

$$\alpha = \text{Sigmoid}(W_2(\text{Relu}(W_1 K + b_1)) + b_2) \tag{15}$$

where α is the generated attention weight, which represents the importance of each corresponding channel. Finally, a weighted summation is performed between the attention weights α and the features $\tilde{h}$ obtained from the BiLSTM module

$$\hat{h} = \text{Relu}\left(Bn(\sum_{t=1}^{F}\alpha_t\tilde{h}_t)\right) \tag{16}$$

where the batch normalization (BN) layer is added to accelerate model training and improve stability, while the Relu activation function is integrated to enhance the nonlinear expression capability of the model. The final output features are processed through a fully connected layer

$$Y = \text{Sigmoid}\left(W\hat{h} + b\right) \tag{17}$$

where W is a learnable parameter, b is the bias, and Y is the prediction result. Notably, the entire model can be optimized and trained using the binary cross-entropy loss function

$$\mathcal{L}_{\text{loss}} = -\sum_{i=1}^{L}(y_i \log(Y_i) + (1-y_i)\log(1-Y_i)) \tag{18}$$

where L denotes the length of the input data, y_i and Y_i represent the ith real label (0 or 1) and the predicted label. The complete detection framework is shown in Fig. 3, and the FDIA detection procedure can be detailedly presented by Algorithm 1.

IV. Experimental Setup

A. Evaluation Metrics

To effectively evaluate the performance of the proposed SAGE-ATT-BiLSTM model, we used four common metrics in the field of machine learning anomaly detection to provide the performance comparison, including Accuracy, Precision, Recall, and F_1 score. Correspondingly, the terms η_{TP}, η_{TN}, η_{FP}, and η_{FN} represent true positives, true negatives, false positives, and false negatives, respectively. These metrics can be calculated as follows:

$$\text{Accuracy} = \frac{\eta_{TP} + \eta_{TN}}{\eta_{TP} + \eta_{FP} + \eta_{TN} + \eta_{FN}}. \tag{19}$$

Accuracy represents the proportion of correctly identified samples among all samples. Apparently, a higher accuracy indicates better model performance

$$\text{Precision} = \frac{\eta_{TP}}{\eta_{TP} + \eta_{FP}}. \tag{20}$$

Precision indicates the proportion of samples classified as attacks that are correctly classified. This metric reflects the model's ability to correctly identify attack samples. A higher precision signifies a lower false alarm rate and better model performance

$$\text{Recall} = \frac{\eta_{TP}}{\eta_{TP} + \eta_{FN}}. \tag{21}$$

Algorithm 1 Deep Interpretable FDIA Detection Model

```
Input: Power grid measurements data $x \in \hbar^{N\times 1}$; Power grid topology matrix data $A \in \hbar^{N\times N}$; Epochs Ep; Batch size Bs;

Output: Model predict $Y$; Trained model $\delta$;
for $i \leftarrow 1$ to Ep do
  for $j \leftarrow 1$ to Bs do
    for $n \leftarrow 1$ to N do
      $o_n^l = mean(\{x_u^{l-1}\ \forall u \in Nb(n)\})$;
      $x_n^l = Relu(W_n^l(x_n^{l-1}, o_n^l))$;
    end for
    $\tilde{X} \in \hbar^{F\times N} \leftarrow [X^l \in \hbar^{N\times F}]^T$;
    for $t \leftarrow 1$ to F do
      $\overrightarrow{h_t} = LSTM(\overrightarrow{h_{t-1}}, \tilde{x}_t, c_{t-1})$;
      $\overleftarrow{h_t} = LSTM(\overleftarrow{h_{t+1}}, \tilde{x}_t, c_{t+1})$;
      $\tilde{h}_t = Tanh(W_f \overrightarrow{h_t} + W_b \overleftarrow{h_t})$;
    end for
    $K = \frac{1}{N}\sum_{i=1}^{N}\tilde{X}(i)$;
    $\alpha = Sigmoid(W_2(Relu(W_1 K + b_1)) + b_2)$;
    $\hat{h} = Relu(Bn(\sum_{t=1}^{F}\alpha_t \tilde{h}_t))$;
  end for
  $Y = Sigmoid(W\hat{h} + b)$;
  $\mathcal{L}_{loss} = -\sum_{i=1}^{B}(y_i \log(Y_i) + (1 - y_i)\log(1 - Y_i))$;
  Output1 $\leftarrow Y$;
end for
Output2 $\leftarrow \delta$;
if $Y = y$ then
  attack;
else
  normal;
end if
Utilizing SHAP interpretable techniques to investigate the decision-making process of the trained model $\delta$.
```

Recall refers to the proportion of actual attack samples correctly identified by the model. A higher recall indicates a lower missed detection rate and better model performance

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{22}$$

F_1 score is the harmonic mean of Precision and Recall, and a higher F_1 score implies better model overall performance.

B. Datasets and Experimental Environment

The data sets in our experiment were generated based on the load data from NYISO [31]. We performed the power flow calculations on the load data to simulate the measurement data of the power grid, and the active power was selected as the measurement data. To enhance the realism of the measurement data, IEEE 14-bus and IEEE 118-bus systems were simulated to obtained the measurement data, and standard Gaussian noise with a mean of 0 and a variance of 0.25 was added to these measurement data. In addition, the MATPOWER toolkit was used to obtain information, such as the power grid topology, measurement data, branch parameters, and the Jacobian matrix H.

According to the FDIA construction principles presented in Sections II-A and II-B, coordinated attacks [32] were executed through the Jacobian matrix to inject false data into the power grid's measurement data. Correspondingly, the generated FDIA samples based on the aforementioned attack methods were designed to evade detection by the BDD mechanism as much as possible, and the ratio of the average injection power deviation to the actual measurement should be controlled to be within 30%. Moreover, a total of 15 000 measurement samples were generated for each bus systems, including 7500 FDIA attack samples and 7500 normal samples, respectively. The FDIA attack samples were labeled as 1, while the normal samples were labeled as 0. Subsequently, the dataset was partitioned into training set, validation set, and test set with the ratio 6:2:2, while ensuring that the label categories were balanced across each set. All experiments were repeated five times, and the reported results were averaged to ensure the reliability of the results.

In addition, the experimental platform was conducted with an Intel Core i7-8750H CPU, 16 GB of RAM, and an NVIDIA GeForce GTX 1060 GPU. All experiments were implemented over the PyTorch 2.0 framework, where the power flow calculations and state estimation were performed using MATPOWER in MATLAB platform.

V. Experimental Results and Discussion

In this section, a series of experiments were carried out to evaluate our proposed network model. First, the overall model performance was analyzed between our scheme and several state-of-the-art schemes in Section V-A. Subsequently, ablation study was implemented to further disclose the benefits of different module of our scheme in Section V-B. Meanwhile, the robustness of our model under different noise conditions was thoroughly investigated in Section V-C. Furthermore, we presented the interpretability of the proposed model in Section V-D to provide a meaningful explanation. Last but not least, we discussed the computational complexity in Section V-E.

A. Comparison With the State of the Arts

In order to show the advantages of our proposed SAGE-ATT-BiLSTM model, we first compared it with several state-of-the-art FDIA detection methods in smart grid, including CNN [16], Transformer [18], CNN-GRU [19], GCN [20], SAGE [22], DAMGAT [24], TSGCN [21], and GGNN-GAT [23], where DAMGAT, TSGCN, and GGNN-GAT were three recent excellent FDIA detection schemes. Notably, the architecture of the proposed model mainly consists of two SAGE layers (each with 16 hidden units) and one ATT-BiLSTM layer (containing 64 BiLSTM hidden units), and BN momentum was set to 0.99 for optimal comprehensive performance. In the following experiments, the proposed model was evaluated based on the aforementioned parameter. In order to verify the effectiveness and reliability of the proposed FDIAs detection method, four different evaluation metrics, Accuracy, Precision, Recall, F_1 score, were sequentially tested. All experiments were conducted over two datasets

TABLE I

COMPARISON OF DIFFERENT DETECTION MODELS OVER IEEE 14-BUS SYSTEM AND IEEE 118-BUS SYSTEM UNDER TWO ATTACK INTENSITIES, SINGLE-NODE ATTACK AND MULTINODE ATTACK. EIGHT EXISTING CLASSICAL DETECTION MODELS AND OUR PROPOSED MODEL WERE TESTED TO GIVE THE COMPARISON RESULTS

Bus systems	Models	Single-node attack				Multi-node attack			
		Accuracy	Precision	Recall	F_1 score	Accuracy	Precision	Recall	F_1 score
IEEE 14-bus system	SAGE-ATT-BiLSTM	**94.20**%	**92.74**%	**95.81**%	**94.25**%	**98.31**%	**98.25**%	**98.34**%	**98.30**%
	GGNN-GAT [23]	92.99%	92.84%	93.05%	92.94%	97.62%	97.28%	97.93%	97.60%
	TSGCN [21]	93.04%	93.25%	92.67%	92.96%	97.49%	97.16%	97.81%	97.48%
	DAMGAT [24]	92.64%	92.14%	93.12%	92.62%	97.08%	97.06%	97.05%	97.05%
	SAGE [22]	92.87%	92.90%	92.70%	92.80%	97.29%	96.91%	97.64%	97.28%
	GCN [20]	92.74%	92.71%	92.64%	92.68%	96.71%	95.98%	97.44%	96.70%
	CNN-GRU [19]	90.02%	89.91%	89.97%	89.94%	95.89%	96.21%	95.48%	95.84%
	Transformer [18]	91.64%	92.22%	90.79%	91.50%	97.03%	97.22%	96.78%	97.00%
	CNN [16]	85.61%	84.97%	86.25%	85.60%	92.78%	92.00%	93.59%	92.78%
IEEE 118-bus system	SAGE-ATT-BiLSTM	**91.88**%	**91.50**%	**92.22**%	**91.85**%	**96.47**%	**95.86**%	**97.07**%	**96.46**%
	GGNN-GAT [23]	78.19%	78.08%	78.07%	78.03%	92.36%	93.07%	91.42%	92.23%
	TSGCN [21]	72.08%	73.01%	69.70%	71.25%	89.54%	88.96%	90.10%	89.52%
	DAMGAT [24]	79.54%	80.41%	77.68%	79.00%	91.69%	91.04%	92.34%	91.68%
	SAGE [22]	77.46%	78.17%	75.71%	76.90%	90.34%	90.15%	90.44%	90.28%
	GCN [20]	72.34%	74.19%	68.21%	70.96%	87.38%	86.94%	87.77%	87.34%
	CNN-GRU [19]	57.11%	57.45%	56.99%	56.89%	88.08%	88.42%	87.42%	87.91%
	Transformer [18]	66.56%	66.44%	66.10%	66.24%	85.19%	84.98%	85.19%	85.08%
	CNN [16]	56.78%	46.90%	45.64%	45.97%	82.64%	81.80%	83.64%	82.70%

IEEE 14-bus system and IEEE 118-bus system. In each experiment, our model was trained for 50 iterations with a batch size of 128, an initial learning rate of 0.001, and the Adam optimizer, incorporating cosine annealing for learning rate adjustment. Notably, all models were evaluated under the same experimental conditions to get a fair comparison. Additionally, the generated FDIAs samples were categorized into two attack types [24] to evaluate the model's robustness: 1) single-node attack and 2) multinode attack.

The corresponding experimental results were shown in Table I. We can observe from these results that our scheme can achieve the best overall detection performance comparing with other state-of-the-art schemes, no matter which dataset is used. Specifically, for the IEEE 14-bus system under single-node attack, our proposed method achieved significant performance improvements compared to state-of-the-art approaches, e.g., DAMGAT, TSGCN, GGNN-GAT. In terms of the F_1 score, our model demonstrated the performance gains of 1.63%, 1.29%, and 1.31%, respectively. Similarly, for the large-scale dataset from the IEEE 118-bus system, an approximate performance advantage can be also obtained, e.g., under single-node attack, the performance gains of F_1 score can reach by 12.85%, 20.6%, 13.78%, respectively. In addition, we also tested the overall performance of different detection models under multinode attack, and it can be still observed an obvious advantages comparing with other schemes. The corresponding results verified that our scheme can consistently achieve a superior performance, whichever small-scale dataset and large-scale dataset was used.

To thoroughly assess the model's robustness against FDIA, we designed three attack scenarios: 1) coordinated attack (Jacobian matrix) [32]; 2) stealth attack (GAN-generated) [33]; and 3) random attack (Laplace noise) [34]. The attack intensity was uniformly set to tampering all-node measurements across all 7500 generated attack samples (1:1 ratio with normal samples). The corresponding detection results were shown in Table II. For the IEEE 14-bus system, our proposed model demonstrated superior performance compared to three state-of-the-art models GGNN-GAT, TSGCN, DAMGAT across three attack scenarios, achieving F_1 score improvements of 0.61%, 0.66%, and 0.81% in coordinated attacks,11.62%, 2.13%, and 9.25% in stealth attacks, and particularly notable performance gains of 34.28%, 14.02%, and 30.94% in the more challenging random attack scenario, respectively.

TABLE II

DETECTION PERFORMANCE COMPARISON OF DIFFERENT DETECTION MODELS UNDER ALL-NODE MEASUREMENT TAMPERING BASED ON DIFFERENT ATTACK STRATEGIES IN IEEE 14-BUS SYSTEM

Attack strategies	**Models**	**Accuracy**	**Precision**	**Recall**	F_1 **score**
Coordinated attacks	SAGE-ATT-BiLSTM	**99.08**%	**98.72**%	**99.42**%	**99.07**%
	GGNN-GAT [23]	98.47%	98.37%	98.55%	98.46%
	TSGCN [21]	98.42%	98.46%	98.36%	98.41%
	DAMGAT [24]	98.27%	98.21%	98.32%	98.26%
	SAGE [22]	98.34%	98.30%	98.34%	98.32%
	GCN [20]	97.74%	97.35%	98.12%	97.73%
Stealth attacks	SAGE-ATT-BiLSTM	**90.13**%	**98.13**%	**81.64**%	**89.13**%
	GGNN-GAT [23]	80.72%	89.68%	68.55%	77.51%
	TSGCN [21]	88.17%	95.62%	79.81%	87.00%
	DAMGAT [24]	82.22%	91.05%	71.16%	79.88%
	SAGE [22]	84.00%	92.71%	73.55%	82.01%
	GCN [20]	54.27%	67.55%	19.51%	28.68%
Random attacks	SAGE-ATT-BiLSTM	**90.00**%	**93.40**%	**85.96**%	**89.50**%
	GGNN-GAT [23]	61.23%	64.77%	48.25%	55.22%
	TSGCN [21]	77.87%	83.73%	68.73%	75.48%
	DAMGAT [24]	64.59%	69.79%	50.51%	58.56%
	SAGE [22]	68.37%	74.29%	55.49%	63.49%
	GCN [20]	49.86%	52.07%	22.33%	25.03%

In fact, this phenomenon can be easily explained by the following two reasons. First, based on the GraphSAGE, the deep spatial topological dependencies of complex network nodes can be introduced to further optimize the whole network structure so that it can be better applied to power grid structures of different scales. Accordingly, the proposed network model can effectively capture the differences between subtle changes in the power grid and actual FDIA attacks, thereby enhancing the detection effectiveness for the potential FDIA attacks. Second, BiLSTM network with SE-Attention module

TABLE III
ABLATION STUDY OF DIFFERENT MODULES FOR OUR PROPOSED SAGE-ATT-BILSTM MODEL OVER IEEE 14-BUS SYSTEM AND IEEE 118-BUS SYSTEM UNDER DIFFERENT ATTACK INTENSITIES

Bus systems	Models	Single-node attack				Multi-node attack			
		Accuracy	Precision	Recall	F_1 score	Accuracy	Precision	Recall	F_1 score
IEEE 14-bus system	SAGE-ATT-BiLSTM	94.20%	92.74%	**95.81%**	94.25%	**98.31%**	**98.25%**	**98.34%**	**98.30%**
	SAGE-NORATT-BiLSTM	**94.42%**	**94.49%**	94.23%	**94.36%**	98.23%	98.10%	98.34%	98.22%
	SAGE-BiLSTM	93.41%	93.05%	93.73%	93.38%	98.04%	97.99%	98.05%	98.02%
	SAGE [22]	92.87%	92.90%	92.70%	92.80%	97.29%	96.91%	97.64%	97.28%
	BiLSTM [29]	92.34%	92.05%	92.56%	92.30%	96.20%	96.24%	96.10%	96.17%
IEEE 118-bus system	SAGE-ATT-BiLSTM	**91.88%**	**91.50%**	**92.22%**	**91.85%**	**96.47%**	**95.86%**	**97.07%**	**96.46%**
	SAGE-NORATT-BiLSTM	74.63%	74.77%	74.00%	74.38%	93.05%	92.80%	93.23%	93.00%
	SAGE-BiLSTM	85.43%	86.04%	84.27%	85.14%	94.84%	94.29%	95.36%	94.82%
	SAGE [22]	77.46%	78.17%	75.71%	76.90%	90.34%	90.15%	90.44%	90.28%
	BiLSTM [29]	57.13%	57.51%	52.64%	54.67%	85.33%	85.03%	85.47%	85.24%

was introduced to optimize the long-term feature construction, which can efficiently aggregate attack characteristics and long-term dependency information by dynamically capturing the potential correlations between FDIAs detection and measurement data. Also, the SE-Attention module can guide the detection features to focus on the differences between FDIA attack data and normal data, resulting in a significant improvement of detection accuracy.

TABLE IV
DETECTION PERFORMANCE COMPARISON OF DIFFERENT DETECTION MODELS OVER IEEE 14-BUS SYSTEM AND IEEE 118-BUS SYSTEM UNDER MULTINODE ATTACK WITH TOPOLOGICAL CHANGES

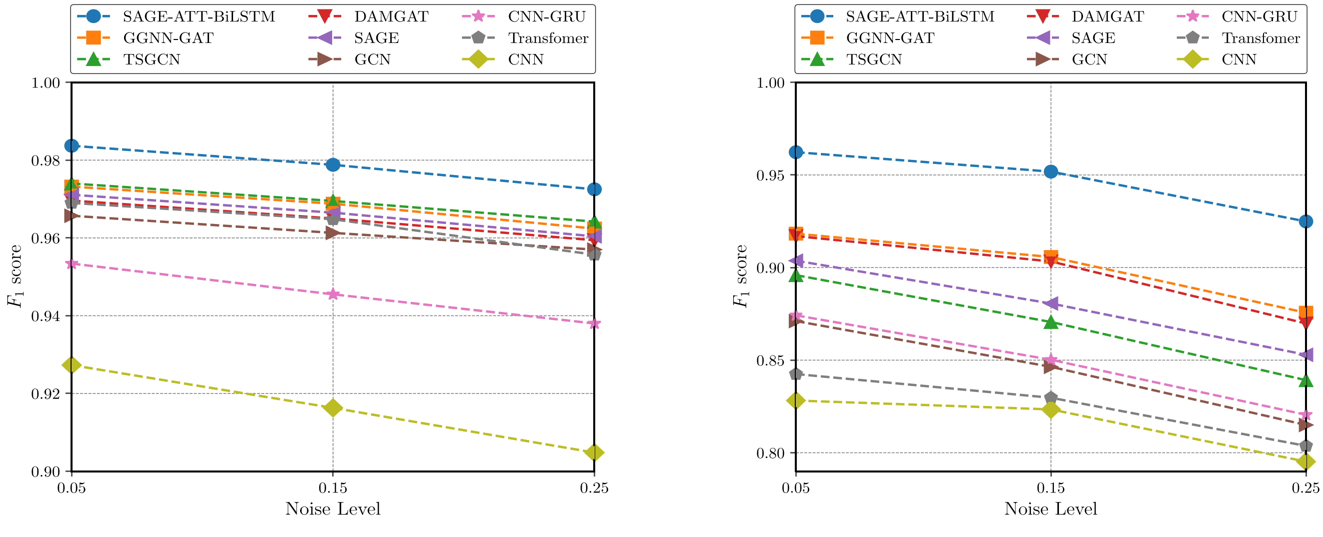
Bus systems	**Models**	**Accuracy**	**Precision**	**Recall**	F_1 **score**
IEEE 14-bus system	SAGE-ATT-BiLSTM	**98.28%**	**98.66%**	**97.86%**	**98.26%**
	GGNN-GAT [23]	96.81%	96.55%	97.04%	96.80%
	TSGCN [21]	97.47%	97.36%	97.55%	97.45%
	DAMGAT [24]	95.89%	95.65%	96.08%	95.87%
	SAGE [22]	96.79%	96.26%	97.32%	96.78%
	GCN [20]	94.82%	94.72%	94.84%	94.78%
IEEE 118-bus system	SAGE-ATT-BiLSTM	**96.45%**	**95.58%**	**97.37%**	**96.46%**
	GGNN-GAT [23]	89.52%	88.94%	90.08%	89.50%
	TSGCN [21]	88.86%	87.89%	89.92%	88.89%
	DAMGAT [24]	88.29%	87.63%	88.96%	88.29%
	SAGE [22]	88.29%	88.09%	88.33%	88.21%
	GCN [20]	82.19%	81.43%	83.03%	82.21%

B. Ablation Study

Our proposed network architecture combines GraphSAGE with BiLSTM, while introducing a SE-Attention module. In this section, we discussed the effectiveness of different network modules through a series of ablation experiments. Similar to previous experiments, two different-scale datasets, IEEE 14-bus system and IEEE 118-bus system, were tested to provide the detection results for single-node attack and multinode attack, respectively.

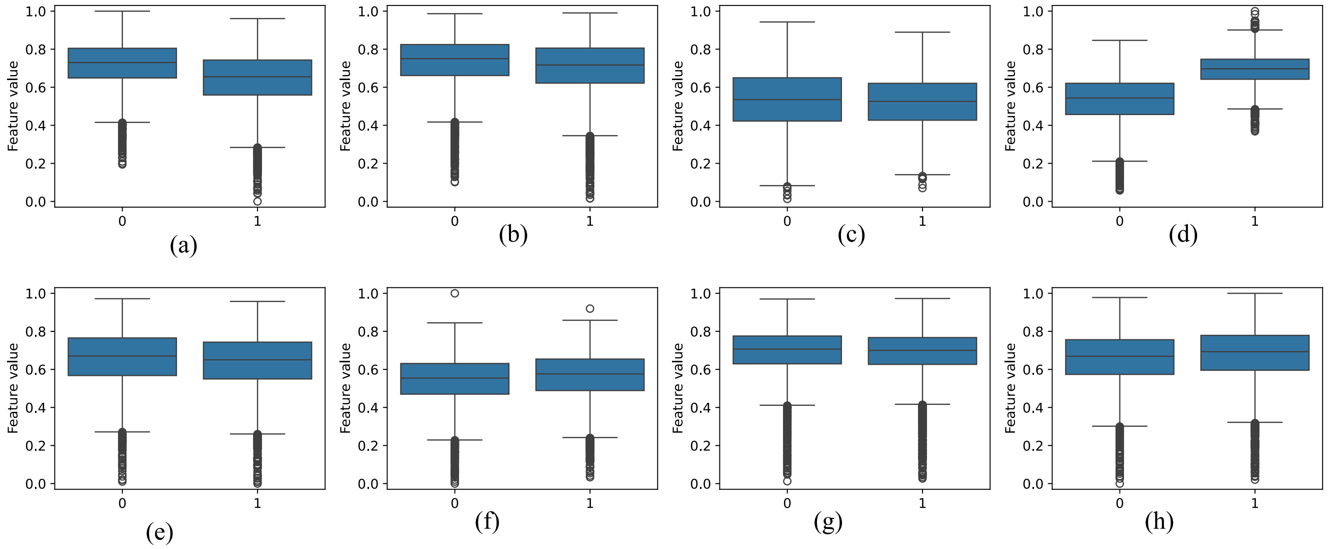
We first sequentially tested five ablation experiment combinations, SAGE-ATT-BiLSTM, SAGE-NORATT-BiLSTM, SAGE-BiLSTM, SAGE, BiLSTM, where SAGE-ATT-BiLSTM is the proposed complete network model. Similar to previous experiments, we compared the F_1 score to evaluate the overall performance. The corresponding detection results were shown in Table III. From this table, it can be seen that for F_1 score, under single-node attack, SAGE-BiLSTM outperformed the individual SAGE and BiLSTM modules in IEEE 14-bus system, and achieved the improvements of 0.58% and 1.08%, respectively. When the SE-Attention module was added to SAGE-BiLSTM (corresponding to the SAGE-ATT-BiLSTM model), it could yield a further performance gain of 0.87%, while its performance was basically on par with the SAGE-BiLSTM with the standard attention module (corresponding to the SAGE-NORATT-BiLSTM model), which demonstrated that our proposed SAGE-ATT-BiLSTM model significantly improved FDIA detection performance by integrating the SAGE, BiLSTM, and SE-Attention mechanism. Actually, this phenomenon can be explained easily. Since the SAGE network can capture the spatial dependencies of complex adjacent nodes, while the BiLSTM module can effectively enhance the model's capability to learn long-term dependencies between different nodes. Accordingly, the combination of these two network models significantly improved the network model's capability that perceives the subtle changes in attack data based on spatial topological and long-term dependency features. Meanwhile, the SE-Attention module can continuously monitor the subtle changes in node data, while filtering out irrelevant information, guiding the network model to effectively distinguish between attack data and normal data, thereby gradually improving the detection performance of the proposed network model.

In addition, we can observe an interesting phenomenon that in IEEE 118-bus system, the performance improvements were more pronounced comparing with small-scale IEEE 14-bus system. For example, under single-node attack scenario, the performance gains of SAGE-BiLSTM can get 8.65% and 29.6% over the individual SAGE and BiLSTM modules, while SAGE-ATT-BiLSTM can achieved 6.71% and 17.47% performance improvement compared to SAGE-BiLSTM and SAGE-NORATT-BiLSTM. This is mainly because in large-scale power grids, the system topology and internode dependencies are more intricate, individual SAGE network can easily overlook the complex dependency relationships between different nodes, leading to changes in cascading attack data being overwhelmed by the node relationships of complex graph network structures. Meanwhile, in large-scale bus systems, data diversity and network noise are also stronger, and seriously interfere with the feature extraction and detection capabilities of individual detection networks in complex environments, e.g., individual SAGE network or BiLSTM network. Furthermore, due to the standard attention lack of



(a) (b)

Fig. 4. Detection performance comparison of different detection models over the IEEE 14-bus system and IEEE 118-bus system under multinode attack with different levels of noise interference. (a) Detection for IEEE 14-bus system. (b) Detection for IEEE 118-bus system.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. FDIA node feature distribution in each category under multinode attack for IEEE 14-bus system and IEEE 118-bus system. (a) IEEE 14-Node2. (b) IEEE 14-Node6. (c) IEEE 14-Node10. (d) IEEE 14-Node11. (e) IEEE 118-Node1. (f) IEEE 118-Node11. (g) IEEE 118-Node14. (h) IEEE 118-Node15.

SE operations compared to SE-Attention, which means that the global information of the power grid data is not explicitly utilized, and interference information in the power grid data cannot be effectively suppressed, leading to the poor model detection performance. While our proposed detection network can integrate the topology change perception capability of complex networks and the long-term dependency relationship of node data changes, obtaining further improvement in overall performance. This conclusion can be experimentally verified in Section V-A.

The topology structure of the power grid changes due to the failure or damage of sensor nodes, and accordingly, the performance of detection models based on the graph topology structure of the power grid may significantly decrease. To evaluate robustness under such scenarios, we designed detection experiments simulating transmission line disconnections in the power grid, where the training set, validation set, and test set each contained 10% of data reflecting topological changes [23]. The detection results were shown in Table IV. As can be observed that our model significantly outperformed other state-of-the-art graph detection models over both IEEE 14-bus and IEEE 118-bus systems. Particularly compared to the GGNN-GAT model specifically designed for topological changes scenario detection, our model can achieve F_1 score improvements of 1.46% and 7.14% over both IEEE 14-bus and IEEE 118-bus systems, respectively. This is mainly because in our model, the aggregation mechanism of GraphSAGE effectively captures the dynamic topological dependencies between adjacent nodes, and can robustly identify cascading attack patterns even as the topology structure of the power grid evolves. Moreover, SE-Attention dynamically weights the key attack features in the measurement data sequence, enhancing the traditional BiLSTM network and effectively mining long-term dependencies in measurement data.

C. Discussion for Noise Interference

In real power grid environment, measurement data contains various noise, it is essential to evaluate model detection performance under different noise conditions. Therefore, we conducted a series of detection experiments over the IEEE 14-bus and IEEE 118-bus systems using standard Gaussian noise with three variance levels: 1) 0.05; 2) 0.15; and 3) 0.25. As shown in Fig. 4, although the detection performance of all models declined with increasing noise levels, our proposed model still outperformed all other state-of-the-art detection methods across different environments, whichever dataset was used. The reasons can be explained easily. First, our proposed model can explicitly characterize the deep spatial cascaded topological dependencies between adjacent data nodes by introducing GraphSAGE network. This structural characteristic enables the detection model to effectively identify distorted topological association patterns even under noise interference, avoiding local feature misjudgment caused by noise. Second, the BiLSTM layer can capture long-term dependencies in measurement data, while the SE-attention mechanism adaptively strengthens feature channels with high correlation with FDIA detection through attention, suppressing noise dominated interference dimensions. Accordingly, the BiLSTM network integrated with SE-attention mechanism further enhances the detection capability of the model through the dual advantages of bidirectional sequence modeling and dynamic weight allocation.

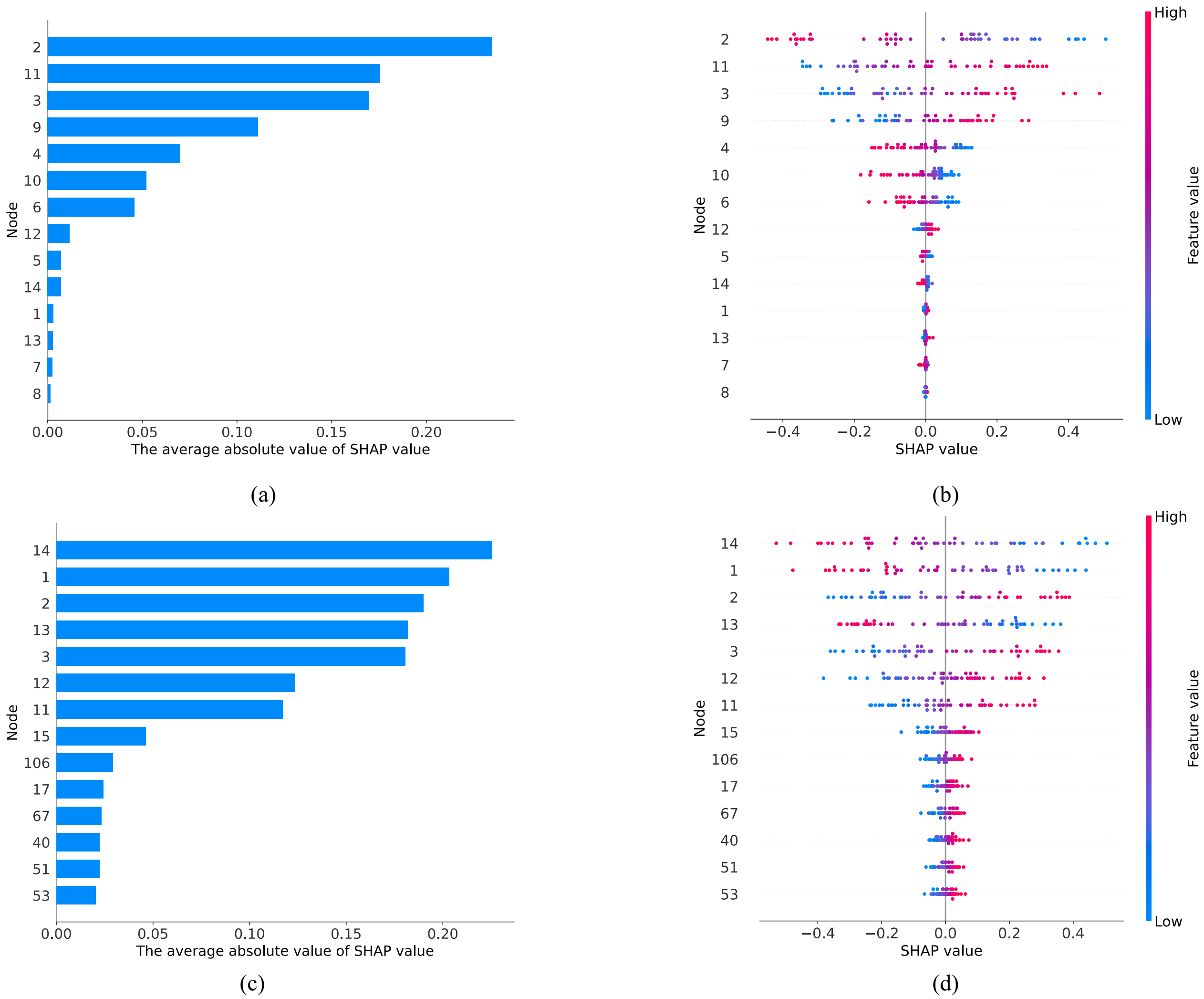
D. Interpretability Analysis

To gain more insights for the attack identification procedure of detection model, we further verified the interpretability of the SAGE-ATT-BiLSTM model. SHAP interpretability technology was used to analyze the impact of each node feature on the model's detection and integrated FDIA domain knowledge to provide a meaningful explanation.

First, we investigated thoroughly the interpretability of our proposed network model on the IEEE 14-bus system, including 14 nodes. In this test, we sequentially performed FDIA attacks on the nodes 2, 3, 4, 6, 9, 10, 11. We calculated the average of the absolute SHAP contribution values of all observed measurements. The corresponding results were shown in Fig. 6(a), which presented the contribution of 14 node features when the model performs FDIA detection. As can be seen from this figure, it is evident that the model primarily detects FDIA by focusing on the features of the attacked nodes. Especially for the top seven nodes, these node features can contribute a total detection performance of 96.02%. In contrast, the model may pay less attention to the features of nonattacked nodes. In addition, with the interpretability effects, our proposed network model can not only detect the FDIA attacks, but also can locate the exact nodes that are under attack. Fig. 6(b) presented a more detailed view of feature contributions, where the data points in this figure represent the values of each node feature, with red indicating higher values and blue indicating lower values. On the horizontal axis, data points to the right indicate a positive impact on the model's classification as an attack, while data points to the left stands for a negative impact. Combined with Fig. 5, it can be seen that in IEEE 14-bus system, the numerical range of the attack category was lower than that of the normal category for nodes 2, 6, and 10, while the opposite was true for node 11. Apparently, once the detection model has learned this feature, as shown in Fig. 6(b), the lower values in nodes 2, 6, and 10 can help the model identify the attack category, while the higher values in node 11 also have the similar effect. The identical interpretability effects can be also obtained for other FDIA nodes. With the help of these interpretable results, we can experimentally demonstrate that the proposed network model can effectively uncover the attack patterns within the features, resulting in outstanding detection performance.

Furthermore, by arranging the SHAP contribution values of each observed measurement data in chronological order, a heatmap, e.g., Fig. 7(a), can be generated to visualize how feature importance evolves temporally during the model's detection process of the IEEE 14-bus system. Apparently, our proposed model mainly achieved the attack detection by focusing on the features of the FDIA attack nodes. At the same time, when the model encounters FDIA attack time points 2, 3, and 6, the overall feature contribution values shift toward the positive values (indicated by a redder color), guiding the model to recognize these time points as the attacked nodes. In contrast, for other normal time points, the feature contribution values shift toward negative values (indicated by a bluer color). By summing the feature contribution values of all nodes at each time point, the resulting total node contribution heatmap indicates that the variations correspond precisely to the FDIA attack time points, further confirming the rationality and interpretation of the model's detection.

To provide more insights, we also implemented a series of experiments over the IEEE 118-bus system, where the nodes 1, 2, 3, 11, 12, 13, 14, 15, were sequentially performed with FDIA attacks. Notably, due to the large scale of the IEEE 118-bus system, we only presented the interpretability analysis of the top 14 nodes with the highest feature contributions. The corresponding experimental results were shown in Figs. 5 and 6(c) and (d). As can be observed from these figures, the interpretability effects of the features of several typical FDIA nodes in the IEEE 118-bus system were consistent with the IEEE 14-bus system, e.g., Fig. 6(c), the feature contribution of the top eight nodes in the IEEE 118-bus system accounts for 61.2% of the total. This means that our proposed network model can obtain the similar interpretability effects, whichever large-scale or small-scale datasets are used. That being said, we can also find an interesting phenomenon that comparing with the total detection contribution of 96.02% for the top seven nodes in the IEEE 14-bus system, the proposed network model may get an inferior detection performance and a weaker interpretability. This is mainly because in large-scale power grids, the system topology and node dependencies



(a)

(b)

(c)

(d)

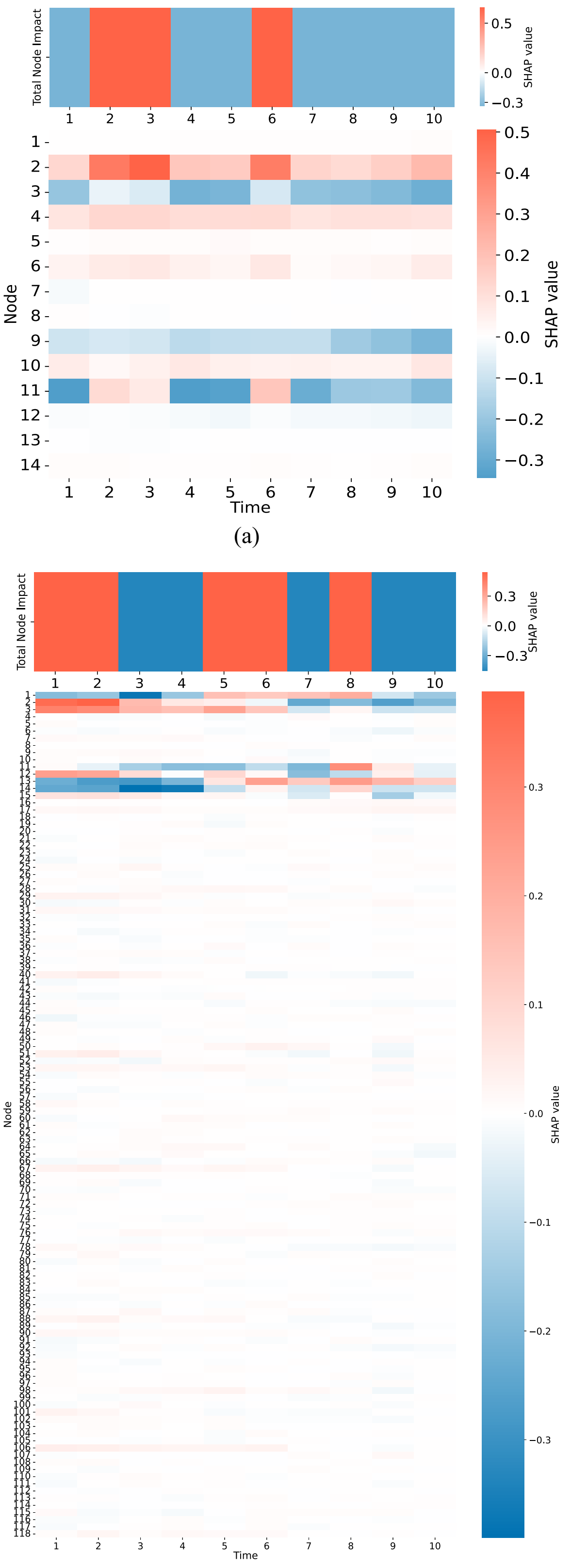
Fig. 6. SHAP feature contribution plots on the IEEE 14-bus system and IEEE 118-bus system for multinode attack. (a) SHAP bar for IEEE 14-bus system. (b) SHAP beeswarm for IEEE 14-bus system. (c) SHAP bar for IEEE 118-bus system. (d) SHAP beeswarm for IEEE 118-bus system.

are more complex. Correspondingly, the information diversity and power grid noise can easily mask the feature changes caused by real attacks, leading to detection models easily confusing real attack data and noisy data, further lowering the detection performance of identifying the attacked nodes. In addition, according to the time-domain analysis in Fig. 7(b), our proposed model can be also able to accurately identify the attack time points at 1, 2, 5, 6, and 8. This observation further confirmed the following conclusion that the proposed network model can effectively extract key spatial topological and long-term dependency features of different nodes to detect cascaded FDIA attacks. We believe that this benefit is mainly from the integration both dependency-aware GraphSAGE and BiLSTM with SE-Attention mechanism, where the spatial cascading topology features and long-term dependencies between nodes can be efficiently captured.

E. Computational Complexity Analysis

To further demonstrate the superiority of the proposed method, we conducted a series of experiments to demonstrate the efficiency of our proposed model. Four performance metrics, model training time, detection time, number of parameters, and flops (floating point operations), were sequentially tested. In order to give a fair comparison, we considered training time as the duration taken for the model to complete 50 epochs of training, detection time as the time taken for the model to detect testing dataset, and flops as the computational costs for processing 1000 samples. Notably, the computational complexity of our proposed model was measured under a lightweight configuration, i.e., reducing the hidden units in BiLSTM from 64 to 16 to ensure low cost, while maintaining competitive detection performance.

Table V showed the actually testing results. As can be seen from this table, the proposed network model, SAGE-ATT-BiLSTM, can consistently obtain a relative superior values for four metrics, whichever bus system is used. In particular, compared to the DAMGAT model in the IEEE 14-bus system, the training and detection times of our model reduced by 62.8 and 0.07 s, respectively. The number of parameters reduced by 0.3K, while the flops only slightly increased by 0.01G. Similarly, in the IEEE 118-bus system, the advantages



(a)

(b)

Fig. 7. SHAP feature contribution heat map in the time dimension on the (a) IEEE 14-bus system and IEEE 118-bus system under multinode attack.

TABLE V
COMPUTATIONAL COMPLEXITY COMPARISONS OF DIFFERENT DETECTION MODELS ON THE IEEE 14-BUS SYSTEM AND IEEE 118-BUS SYSTEM

Bus systems	**Models**	**Train(s)**	**Detect(s)**	**Parameter(K)**	**Flops(G)**
IEEE 14-bus system	SAGE-ATT-BiLSTM	49.79	0.13	5.03	0.08
	GGNN-GAT [23]	47.28	0.14	3.97	0.08
	TSGCN [21]	39.72	0.16	55.66	0.06
	DAMGAT [24]	112.59	0.20	5.33	0.07
	SAGE [22]	36.59	0.14	0.80	0.01
	GCN [20]	38.80	0.14	53.98	0.05
	CNN-GRU [19]	41.06	0.11	116.74	0.93
	Transfomer [18]	70.88	0.23	135.59	1.84
	CNN [16]	35.86	0.07	34.18	0.46
IEEE 118-bus system	SAGE-ATT-BiLSTM	51.20	0.21	18.34	0.35
	GGNN-GAT [23]	111.49	0.29	5.63	0.66
	TSGCN [21]	111.05	0.24	3905.33	3.90
	DAMGAT [24]	453.18	1.70	20.83	2.22
	SAGE [22]	37.49	0.14	2.46	0.07
	GCN [20]	53.93	0.17	3792.99	3.79
	CNN-GRU [19]	89.31	0.24	123.39	7.85
	Transfomer [18]	267.86	0.50	137.25	15.50
	CNN [16]	52.09	0.13	40.83	3.89

of our model were further expanded that the training and detection times reduced by 401.98 and 1.49 s, respectively. The number of parameters reduced by 2.49K, and the flops lowered by 1.87G. In fact, the lower computational overhead of our proposed model can be attributed to the introduction of the GraphSAGE model due to the simple and efficient aggregation mechanism, because compared to GCN model and TSGCN model, GraphSAGE can avoid the complex Laplacian matrix computations, while compared to DAMGAT and GGNN-GAT, it can eliminate the complex multihead attention mechanism, thereby further lowered the computational costs.

Moreover, it can be observed from the Table V that our model brought a slight higher computation complexity comparing with the individual SAGE model, e.g., the computational costs increased by 13.2 s of training time and 4.23K parameters for the IEEE 14-bus system, 13.71 s of training time and 15.88K parameters for the IEEE 118-bus system. Nevertheless, this complexity limitation should not be a concern at all for current hardware devices, because a higher detection performance is the focus of current FDIA detection schemes, yet the complexity limitation can be fully compensated by improving hardware performance. In other words, low-complexity is a nice-to-have but absolutely not a must-have feature for most FDIA detection models.

VI. CONCLUSION

In this article, we proposed an efficient FDIA detection scheme by designing a dependency-aware deep interpretable network architecture. We first constructed a dependency-aware deep interpretable network framework by introducing GraphSAGE network, which can effectively represent the deep spatial cascading topological dependencies of adjacent multiple data nodes. Furthermore, a BiLSTM network with a SE attention module was designed to efficiently aggregate attack characteristics and long-term dependency information by dynamically capturing the potential correlations between FDIAs detection and measurement data. The proposed scheme can achieve an efficient balance in terms of detection performance, interpretability, and robustness. Extensive experimental results demonstrated that the proposed model obtained a superior overall performance comparing with several state-of-the-art FDIA detection schemes, and showed the reasonable interpretability from the spatial–temporal dimensions.

While our method can get a better FDIA detection performance, it should be noted that our model still contains two slight limitations. First, our model has not yet taken into account the topological changes in the graph caused by node loss or damage in the power grids. Second, our network model assumes that the attacker can obtain complete topology information of the power grid system. In fact, this is also an open challenge for the current FDIA detection. Therefore, we plan to further improve our work aiming at the above two issues in the future.

REFERENCES

[1] N. Abdi, A. Albaseer, and M. Abdallah, "The role of deep learning in advancing proactive cybersecurity measures for smart grid networks: A survey," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16398–16421, May 2024.

[2] M. N. Nafees, N. Saxena, A. Cardenas, S. Grijalva, and P. Burnap, "Smart grid cyber–physical situational awareness of complex operational technology attacks: A review," *ACM Comput. Surveys*, vol. 55, no. 10, pp. 1–36, 2023.

[3] S. Y. Diaba and M. Elmusrati, "Proposed algorithm for smart grid DDoS detection based on deep learning," *Neural Netw.*, vol. 159, pp. 175–184, Feb. 2023.

[4] F. Hua, W. Gao, Y. Li, X. Guo, and P. Hu, "Joint detection and state estimation based on GPS spoofing attack in smart grids," *Int. J. Elect. Power Energy Syst.*, vol. 161, Oct. 2024, Art. no. 110151.

[5] R. Ren, Y. Li, Q. Sun, S. Zhang, D. W. Gao, and S. Maharjan, "Switched surplus-based distributed security dispatch for smart grid with persistent packet loss," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6185–6198, Feb. 2024.

[6] M. F. Elrawy, L. Hadjidemetriou, C. Laoudias, and M. K. Michael, "Detecting and classifying man-in-the-middle attacks in the private area network of smart grids," *Sustain. Energy, Grids Netw.*, vol. 36, Dec. 2023, Art. no. 101167.

[7] H. Yang and Z. Wang, "A false data injection attack approach without knowledge of system parameters considering measurement noise," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1452–1464, Jan. 2024.

[8] Z. Zhang et al., "Limitation of reactance perturbation strategy against false data injection attacks on IoT-based smart grid," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 11619–11631, Apr. 2024.

[9] W. Xia, D. He, and L. Yu, "Locational detection of false data injection attacks in smart grids: A graph convolutional attention network approach," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 9324–9337, Mar. 2024.

[10] Z. Qu, J. Yang, Y. Wang, and P. M. Georgievitch, "Detection of false data injection attack in power system based on Hellinger distance," *IEEE Trans. Ind. Informat.*, vol. 20, no. 2, pp. 2119–2128, Feb. 2024.

[11] L. Wei and Q. Zhang, "Detection of false data injection attack in smart grid based on improved UKF," *J. Syst. Simul.*, vol. 35, no. 7, pp. 1508–1516, 2023.

[12] Y. Shen and Z. Qin, "Detection, differentiation and localization of replay attack and false data injection attack based on random matrix," *Sci. Rep.*, vol. 14, no. 1, p. 2758, 2024.

[13] S. K. Sah Tyagi, R. Yadav, D. K. Jain, Y. Tu, and W. Zhang, "Paired swarm Optimized relational vector learning for FDI attack detection in IoT-aided smart grid," *IEEE Internet Things J.*, vol. 10, no. 21, pp. 18708–18717, Nov. 2023.

[14] A. Parizad and C. J. Hatziadoniu, "Cyber-attack detection using principal component analysis and noisy clustering algorithms: A collaborative machine learning-based framework," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4848–4861, Nov. 2022.

[15] J. Tian et al., "LESSON: Multi-label adversarial false data injection attack for deep learning locational detection," *IEEE Trans. Depend. Secure Comput.*, vol. 21, no. 5, pp. 4418–4432, Sep./Oct. 2024.

[16] K.-D. Lu, L. Zhou, and Z.-G. Wu, "Representation-learning-based CNN for intelligent attack localization and recovery of cyber–physical power systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 6145–6155, May 2024.

[17] Z. Zhang, J. Hu, J. Lu, J. Yu, J. Cao, and A. Kashkynbayev, "Detection and defense method against false data injection attacks for distributed load frequency control system in microgrid," *J. Modern Power Syst. Clean Energy*, vol. 12, no. 3, pp. 913–924, 2024.

[18] Y. Li, X. Wei, Y. Li, Z. Dong, and M. Shahidehpour, "Detection of false data injection attacks in smart grid: A secure federated deep learning approach," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4862–4872, Nov. 2022.

[19] J. Ji, Y. Liu, J. Chen, Z. Yao, M. Zhang, and Y. Gong, "False data injection attack detection method based on deep learning with multi-scale feature fusion," *IEEE Access*, vol. 12, pp. 89262–89274, 2024.

[20] E. Vincent, M. Korki, M. Seyedmahmoudian, A. Stojcevski, and S. Mekhilef, "Reinforcement learning-empowered graph convolutional network framework for data integrity attack detection in cyber–physical systems," *CSEE J. Power Energy Syst.*, vol. 10, no. 2, pp. 797–806, 2024.

[21] H. Li et al., "End-edge-cloud collaboration-based false data injection attack detection in distribution networks," *IEEE Trans. Ind. Informat.*, vol. 20, no. 2, pp. 1786–1797, Feb. 2024.

[22] C. Chen, Q. Li, L. Chen, Y. Liang, and H. Huang, "An improved GraphSAGE to detect power system anomaly based on time-neighbor feature," *Energy Rep.*, vol. 9, pp. 930–937, Mar. 2023.

[23] X. Li, Y. Wang, and Z. Lu, "Graph-based detection for false data injection attacks in power grid," *Energy*, vol. 263, Jan. 2023, Art. no. 125865.

[24] X. Su et al., "DAMGAT-based interpretable detection of false data injection attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 15, no. 4, pp. 4182–4195, Jul. 2024.

[25] Z. Zhang et al., "Vulnerability of machine learning approaches applied in IoT-based smart grid: A review," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 18951–18975, Jun. 2024.

[26] R. Kalakoti, H. Bahsi, and S. Nõmm, "Improving IoT security with explainable AI: Quantitative evaluation of explainability for IoT Botnet detection," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 18237–18254, May 2024.

[27] T. Liu, A. Jiang, J. Zhou, M. Li, and H. K. Kwan, "GraphSAGE-based dynamic spatial–temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 11210–11224, Oct. 2023.

[28] L. Wang, F. Xie, X. Zhang, L. Jiang, and B. Huang, "spatial–temporal graph feature learning driven by time–frequency similarity assessment for robust fault diagnosis of rotating machinery," *Adv. Eng. Inform.*, vol. 62, Oct. 2024, Art. no. 102711.

[29] S. Peng et al., "Prediction of wind and PV power by fusing the multi-stage feature extraction and a PSO-BiLSTM model," *Energy*, vol. 298, Jul. 2024, Art. no. 131345.

[30] S. Zhu, X. Xu, J. Zhao, and F. Xiao, "LKD-STNN: A lightweight malicious traffic detection method for Internet of Things based on knowledge distillation," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6438–6453, Feb. 2024.

[31] Y. Yu, C. Liu, L. Xiong, Y. Tang, and F. Qian, "Localization of false data injection attacks in smart grids with renewable energy integration via spatiotemporal network," *IEEE Internet Things J.*, vol. 11, no. 23, pp. 37571–37581, Dec. 2024.

[32] S. Basumallik, S. Eftekharnejad, and B. K. Johnson, "The impact of false data injection attacks against remedial action schemes," *Int. J. Elect. Power Energy Syst.*, vol. 123, Dec. 2020, Art. no. 106225.

[33] Z. Liu, Q. Wang, Y. Ye, and Y. Tang, "A GAN-based data injection attack method on data-driven strategies in power systems," *IEEE Trans. Smart Grid*, vol. 13, no. 4, pp. 3203–3213, Jul. 2022.

[34] Y. Wu, T. Zu, N. Guo, Z. Zhu, and F. Li, "Laplace-domain hybrid distribution model based FDIA attack sample generation in smart grids," *Symmetry*, vol. 15, no. 9, p. 1669, 2023.