Infinite Impulse Response Graph Neural Networks for Cyberattack Localization in Smart Grids

Osman Boyaci
Electrical Engineering
Texas A&M University
College Station, TX, 77843
osman.boyaci@tamu.edu

M. Rasoul Narimani
College of Engineering
Arkansas State University
Jonesboro, AR, 72404
mnarimani@astate.edu

Katherine Davis
Electrical Engineering
Texas A&M University
College Station, TX, 77843
katedavis@tamu.edu

Erchin Serpedin
Electrical Engineering
Texas A&M University
College Station, TX, 77843
eserpedin@tamu.edu

***Abstract*—This study employs Infinite Impulse Response (IIR) Graph Neural Networks (GNN) to efficiently model the inherent graph network structure of the smart grid data to address the cyberattack localization problem. First, we numerically analyze the empirical frequency response of the Finite Impulse Response (FIR) and IIR graph filters (GFs) to approximate an ideal spectral response. We show that, for the same filter order, IIR GFs provide a better approximation to the desired spectral response and they also present the same level of approximation to a lower order GF due to their rational type filter response. Second, we propose an IIR GNN model to efficiently predict the presence of cyberattacks at the bus level. Finally, we evaluate the model under various cyberattacks at both sample-wise (SW) and bus-wise (BW) level, and compare the results with the existing architectures. It is experimentally verified that the proposed model outperforms the state-of-the-art FIR GNN model by 9.2% and 14% in terms of SW and BW localization, respectively.**

I. Introduction

Graph structural data such as electric grid networks, social networks, sensor networks, and transportation networks cannot be modeled efficiently in the Euclidean space and require graph-type architectures due to their inherent graph-based topologies [1]. Units are ordered and present the same number of neighbors in image/video data, therefore, they can be processed in an Euclidean space. For instance, a sliding kernel can easily capture the spatial correlations of pixels in the Euclidean space. In contrast, neighborhood relationships are unordered and vary from node to node in a graph signal [1]. Thus, graph signals need to be processed in non-Euclidean spaces determined by the underlying graph topology. As a highly complex graph structural data, smart grid signals require graph architectures such as Graph Signal Processing (GSP) or Graph Neural Network (GNN) to exploit the spatial correlations.

To deal with the graph structural data in non-Euclidean spaces, GSP has emerged in the past few years [2]. In GSP, similar to the classical signal processing, a graph signal is first transformed into the spectral domain by Graph Fourier Transform (GFT), then its Fourier coefficients are scaled in the spectral domain, and finally the signal transformed back into the vertex domain by the inverse GFT [3]. To circumvent the computationally complex domain transformation operations, Finite Impulse Response (FIR) graph filters (GFs) are proposed in [4] in which localized filters are learned directly in the vertex domain without any GFT operations [5]. Spectral response of an FIR GF is a K-order polynomial since the output of each vertex v is only dependent on the K-hop neighborhood of v. Yet, to capture the global structure of a graph, FIR GFs may require high degree polynomials since their frequency responses are not "flexible" enough to adapt to sudden changes in the spectral domain [5]. Nevertheless, the interpolation and extrapolation performance of high degree polynomials are not satisfactory [6]. Infinite Impulse Response (IIR) GFs are proposed in [6] to overcome the limitation of FIR GFs. Contrary to FIR GFs, IIR GFs present rational type spectral responses. Thus, IIR GFs can implement more complex responses with low-degree polynomials in the numerator and denominator because rational functions are more flexible than polynomial functions in terms of interpolation and extrapolation capabilities [5], [6].

Information and Communication Technologies (ICT) are integrated into large-scale power networks to increase the efficiency of generation, transmission, and distribution systems [7]. Remote Terminal Units (RTUs) placed in electric power grids acquire the physical measurements and deliver them to the Supervisory Control and Data Acquisition Systems (SCADAs) and the ICT network transfers these measurements to the application level where the power system operators process them. Consequently, the power system's reliability strongly depends on the accuracy of these steps along this cyber-physical pipeline. Power system state estimation (PSSE) modules employ these measurements to estimate the current operating point of the grid [8]. Besides, the accuracy of power system analysis tools such as energy management, contingency and reliability analysis, load and price forecasting, and economic dispatch depend on these measurements. As a direct consequence, metering devices represent highly attractive targets for adversaries that try to obstruct the grid operation by corrupting the measurements.

False data injection attacks (FDIAs) constitute a significant portion of the cyber-physical threats to smart grids. FDIAs assumes inoculation of false data to the measurements to mislead the PSSE process. Any action taken by the grid operator based on the false operating point can lead to serious consequences including systematic failures [9].

Numerous methods have been proposed to detect the presence of FDIA cyberattacks without providing information about their location [9]. Cyberattack localization is critical for reliable grid operation and control since preventive actions including isolating the under-attack buses and re-dispatching the system can be taken. For this reason, this paper focuses on cyberattack localization in smart grids.

The current approaches in cyberattack localization in power

This work was supported by NSF under Award Number 1808064.

grids suffer from some limitations because it is a relatively new research topic compared to the detection of these attacks. A multistage localization algorithm based on graph theory results is proposed in [10] to localize the attack at the cluster level. Yet, cluster level algorithms limit the benefits of localization due to their low resolution. A model-driven analytical redundancy approach utilizing Kalman filters is presented in [11] for joint detection and mitigation of cyber-attacks in Automatic Gain Control (AGC) systems. Authors in [11] determine a Mahalanobis norm based threshold of the residuals for the non-attacked case and residues larger than this threshold are regarded as attacked samples. Apart from the manual threshold optimization steps, their detection times are in the range of seconds in their estimation based models. A GSP based approach is developed in [12] to detect and localize the cyberattacks. Nevertheless, the random and easily detectable attack methodologies employed to test their models do not comprehensively assess the actual performance of the models. Authors in [13] propose physics- and learning-based approaches to detect and localize cyberattacks in power systems. They utilize a Long Short Term Memory (LSTM) Neural Network (NN) to generate a model for learning the data pattern. Nonetheless, their results are limited to a 5-bus system and they train an LSTM model for each measurement. The limited number of components deeply restrains the large-scale attributes of their proposed method since training a separate detector for each bus extremely increases the overall model complexity for large systems and reduces its suitability for real world applications.

Cyberattack localization can be a challenging task if an adversary has 'enough' information about the grid [14]. Besides, if the designed GFs do not satisfy the required spectral response, the attacker can craft an attack vector so that a malicious sample can be indistinguishable from an honest sample. Thus, we develop an GNN model by utilizing IIR GFs to fit abrupt changes in the spectral domain of the graph structural data. We utilize existing data-driven techniques in the literature for cyberattack localization to compare our results and tune their hyperparameters using Bayesian optimization technique for a fair comparison.

The contributions of the paper are summarized next: (i) We utilize IIR GFs in smart grids to efficiently capture the spatial correlations of the graph structural data in a non-Euclidean space for cyberattack localization. The proposed model efficiently predicts the presence of the attack at the bus level. (ii) We assess by empirical frequency responses of the GFs on IEEE 300-bus test systems; compared to FIR GFs, IIR GFs better approximate the desired filter for the same order and require lower order filters for the same level of approximation. (iii) We evaluate the localization results both sample-wise and bus-wise by adequately assessing the model performance under various cyberattacks. E.g., sample wise localization could yield fairly high accuracy for the entire system, yet, the same set of buses could be missed or falsely detected for each sample. Thus, the localization results should be evaluated both sample wise and bus (label) wise to reveal the possible weaknesses. The rest of this paper is organized as follows. Section II presents the preliminaries. Proposed approach for cyberattack localization is given in Section III. Section IV presents the numerical experiments. Finally, Section V concludes the paper.

II. Preliminaries

The PSSE iteratively solves the following minimization:

$$\hat{\boldsymbol{x}} = \arg\min_{x} ||\boldsymbol{z} - h(\boldsymbol{x})||_2^2 \tag{1}$$

to estimate the system state $\boldsymbol{x}$ (voltage and magnitude of each bus) by using complex power measurements $\boldsymbol{z}$ (active/reactive bus power injections $\boldsymbol{P_i}$, $\boldsymbol{Q_i}$ at each bus i and active/reactive branch power flows $\boldsymbol{Q_{ij}}$, $\boldsymbol{P_{ij}}$ between bus i and j) where h represents a nonlinear measurement function which relates $\boldsymbol{x}$ to $\boldsymbol{z}$. In FDIA, if an intruder attacks the measurements $\boldsymbol{z}$ and can craft an attack vector $\boldsymbol{a} = \boldsymbol{z_a} - \boldsymbol{z} = h(\boldsymbol{x} + \boldsymbol{c}) - h(\boldsymbol{x})$, then s/he can alter the system state from $\boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{c}$ and inject his/her false data into the system.

In FDIA, an intruder targets a particular region of the grid represented by $\mathcal{A}$ and design the attack vector $\boldsymbol{a}$ by altering $\boldsymbol{z}$ to spoil $\boldsymbol{x}$ in the targeted region. In contrast, the grid operator tries to detect these attempts and localize the under attack grid area. Thus, the localization of the cyberattack problem can be formulated as a *multi-label* classification task in which each bus is equipped with a binary flag to indicate the attack presence. The proposed formulation is visualized in Fig. 1 by depicting the actual and predicted targeted buses for an example attack on the IEEE 14-bus test system.

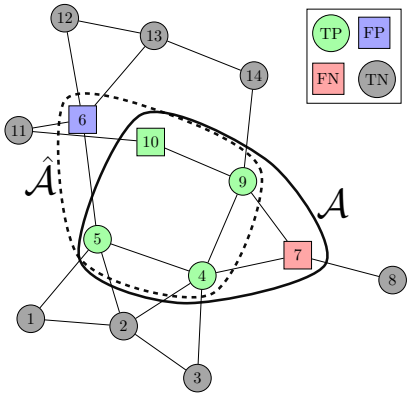


Fig. 1. An example attack $\mathcal{A} = \{4, 5, 7, 9, 10\}$ and its prediction $\hat{\mathcal{A}} = \{4, 5, 6, 9, 10\}$ on the IEEE-14 bus system. In this example, bus 6 is falsely alarmed (false positive) and attack to the bus 7 is missed (false negative).

III. IIR Graph Neural Networks for Cyberattack Localization

Weighted, undirected, and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$ is used to represent the topology of a smart grid, where buses are mapped to vertices $\mathcal{V}$, branches and transformers are denoted by edges $\mathcal{E}$, and line admittances are captured by weighted adjacency matrix $\boldsymbol{W}$. Thus, a signal or function f in $\mathcal{G}$ is represented by a vector $\boldsymbol{f} \in \mathbb{R}^n$, where the element i of the vector corresponds to a scalar at the vertex $i \in \mathcal{V}$.

A. Spectral Graph Filters

Normalized Laplacian $\boldsymbol{L} = \boldsymbol{I_n} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \in \mathbb{R}^{n \times n}$ is a fundamental operator in spectral graph theory. Matrices $\boldsymbol{D} \in \mathbb{R}^{n \times n}$, $\boldsymbol{\Lambda} = \text{diag}([\lambda_1, \ldots, \lambda_n]) \in \mathbb{R}^{n \times n}$ and $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n] \in \mathbb{R}^{n \times n}$ captures the degrees, the

n eigenvalues associated with the graph Fourier frequencies and the n orthonormal eigenvectors $\boldsymbol{u_i}$ representing the graph Fourier basis [2], respectively. A vertex (spectral) domain signal is transformed into the spectral (vertex) domain by the forward (inverse) Graph Fourier Transform defined by $\tilde{\boldsymbol{x}} = \boldsymbol{U}^\top \boldsymbol{x}$ ($\boldsymbol{x} = \boldsymbol{U}\tilde{\boldsymbol{x}}$) where $\boldsymbol{x}$ ($\tilde{\boldsymbol{x}}$) $\in \mathbb{R}^n$ denote the vertex (spectral) domain signal [2]. A GF h is convolved with $\boldsymbol{x}$:

$$\boldsymbol{Y} = h * \boldsymbol{X} = h(\boldsymbol{L})\boldsymbol{X} = \boldsymbol{U}h(\boldsymbol{\Lambda})\boldsymbol{U}^\top \boldsymbol{X} \tag{2}$$

by first transforming $\boldsymbol{x}$ into the spectral domain using the forward GFT, then multiplying the Fourier components with $h(\boldsymbol{\Lambda}) = \text{diag}[h(\lambda_1), \ldots, h(\lambda_n)]$, and finally inverting it back to the vertex domain by the inverse GFT [2]. However, since each λ_i is processed for each node, spectral filters are not spatially localized. In addition, due to eigenvalue decomposition (EVD) of $\boldsymbol{L}$ and the matrix multiplications with $\boldsymbol{U}$ and $\boldsymbol{U}^\top$, they are computationally complex.

B. FIR Graph Filters

To localize and reduce the spectral GFs' complexity, FIR graph filters were proposed in [4]. FIR GFs are K-localized and their spectral responses assume the form $h_{FIR}(\lambda) = \sum_{k=0}^{K-1} a_k \lambda^k$, where only K-hop neighbors of a vertex v are considered to calculate the filter response at $v \in \mathcal{V}$. Nevertheless, FIR GFs require high-degree polynomials to capture the graph's global structure, and due to the poor interpolation and extrapolation capabilities of high degree polynomials, their ability to capture sharp transitions in the frequency response is limited [15].

C. IIR Graph Filters

To circumvent this limitation, researchers in [6], [15] proposed IIR GFs. A potential building block of a K-order IIR GF is the first-order recursive filter:

$$\boldsymbol{Y}^{t+1} = a\tilde{\boldsymbol{L}}\boldsymbol{Y}^t + b\boldsymbol{X}, \tag{3}$$

where $\boldsymbol{X}$ denotes the input, $\boldsymbol{Y}^t$ is the output at iteration t, a and b are arbitrary coefficients, and $\tilde{\boldsymbol{L}} = \frac{\lambda_{max} - \lambda_{min}}{2}\boldsymbol{I_n} - \boldsymbol{L}$ represents the modified Laplacian. According to Theorem 1 in [16], eq. (3) converges regardless of $\boldsymbol{Y}^0$ and $\boldsymbol{L}$ and its frequency response is given by $h_{ARMA_1}(\tilde{\lambda}_n) = \frac{b}{1 - a\tilde{\lambda}_n}$. It can be implemented as a NN layer if we unroll the recursion into T fixed iterations:

$$\boldsymbol{Y}^{t+1} = \tilde{\boldsymbol{L}}\boldsymbol{Y}^t\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\theta}, \tag{4}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{c_{out} \times c_{out}}$, $\boldsymbol{\beta} \in \mathbb{R}^{c_{in} \times c_{out}}$, and $\boldsymbol{\theta} \in \mathbb{R}^{c_{out}}$ are trainable weights, and c_{in} and c_{out} denote the number of channels in the input and output tensors, respectively. Since $0 \leq \lambda_{min} \leq \lambda_{max} \leq 2$, the modified Laplacian can be simplified to $\tilde{\boldsymbol{L}} = \boldsymbol{I_n} - \boldsymbol{L}$ for $\lambda_{min} = 0$, and $\lambda_{max} = 2$ [5]. NN realization of the IIR_1 block implementing eq. (4) in T fixed iterations is depicted in Fig. 2. IIR_K GFs can be implemented by averaging K parallel IIR_1 filters: $\boldsymbol{Y} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{Y}_K^\top$, which leads to a rational frequency response $h_{IIR_K}(\tilde{\lambda}_n) = \sum_{k=1}^{K} \frac{b_k}{1 - a_k\tilde{\lambda}_n}$, with $K-1$ and K order polynomials at its numerator and denominator, respectively. Motivated readers are directed to [5], [6], [15]–[17] for detailed analysis.

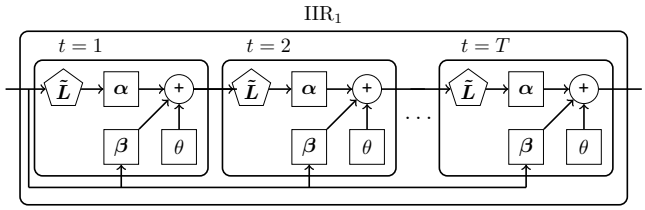


Fig. 2. NN implementation of IIR_1 GF as a building block of IIR_K GF. In T fixed iterations, an IIR_1 block realizes eq. (4).

D. Architecture of the Proposed Model

The proposed model consists of one input layer to represent complex bus power injections, $L-1$ hidden IIR_K layers to extract spatial features, one dense layer to distribute the node features, and one output layer to predict the attack at each node. Fig. 3 illustrates the proposed model's architecture for $L = 3$ with a small graph having $n = 5$.

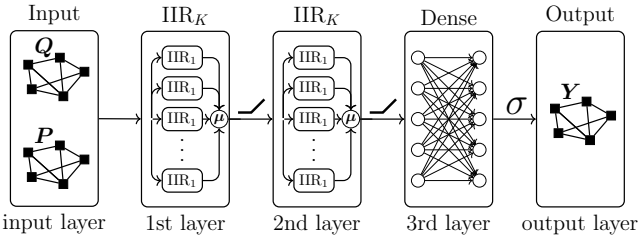


Fig. 3. Architecture of the proposed model with three hidden layers where each IIR_K layer consists of K parallel IIR_1.

In the multi-layered architecture, $\boldsymbol{X^0} \in \mathbb{R}^{n \times 2}$ represents the input tensor $[\boldsymbol{P}, \boldsymbol{Q}]$, $\boldsymbol{X^l} \in \mathbb{R}^{n \times c_l}$ defines hidden layer l's output tensor, $\boldsymbol{Y} \in \mathbb{R}^n$ denotes the model outputs as the location of the attack, and c_l identifies the number of channels in layer l for $1 \leq l \leq L$. Specifically, an IIR_K layer takes $\boldsymbol{X}^{l-1} \in \mathbb{R}^{n \times c_{l-1}}$ as its input and produces $\boldsymbol{X}^l \in \mathbb{R}^{n \times c_l}$ as its output in layer l, the dense layer propagates the information to the whole graph and the output layer returns the probability of the attack at the node level via $\boldsymbol{Y} \in \mathbb{R}^n$. While the ReLU activation is used at the end of each IIR_K layer to increase the model's nonlinear modeling ability, the sigmoid is employed to transform the outputs into probabilities.

IV. Numerical Experiments

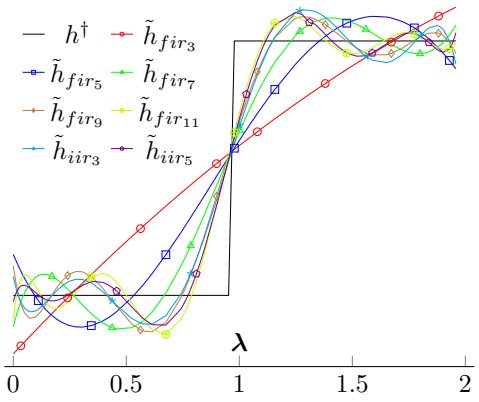
A. Frequency Response of FIR and IIR GFs

To demonstrate that IIR GFs better fit the sharp changes in frequency response compared to FIR GFs, we design an ideal highpass GF $h^\dagger$ for IEEE 300-bus test system:

$$h^\dagger(\lambda) = \begin{cases} 1, & \lambda > \frac{\lambda_{max}}{2} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Note that similar results can be obtained by any other filter or test cases [6]. Let $\boldsymbol{x}, \boldsymbol{y} \in R^n$ denote the input and output of a GF $h(\lambda)$, respectively. The empirical frequency response $\tilde{h}$ can be expressed as $\tilde{h}(\lambda_i) = \frac{\boldsymbol{u}_i^\top \boldsymbol{y}}{\boldsymbol{u}_i^\top \boldsymbol{x}}$ [15]. Each $\tilde{h}(\lambda_i)$ shows how $\boldsymbol{u}_i$, corresponding to λ_i, "scales" $\boldsymbol{x}$ to obtain $\boldsymbol{y}$.

As a first step, we randomly generate 2^{16} $\boldsymbol{x}$s for the aforementioned system from the normal distribution and filter them by $h^\dagger$ using eq. (2) to obtain $\boldsymbol{y}$s. Next, we train a layer of FIR and IIR GNN models in mini batches with 2^6 samples of $\boldsymbol{x}$ and $\boldsymbol{y}$ as the input and output values of the models until there is no further improvement. Then, we calculate and plot $\tilde{h}(\lambda_i)$ values for each $\boldsymbol{x}, \boldsymbol{y}$ tuple [18]. Fig. 4 demonstrates that IIR GFs are more flexible to fit sudden changes for a fixed K compared to FIR GFs. It is the main motivation of this paper to select IIR GFs for localizing cyberattacks in smart grids.



(a) an ideal highpass $h^\dagger$ for IEEE-300.

Fig. 4. Empirical frequency response of FIR and IIR GFs when approximating an ideal filter $h^\dagger$ applied on IEEE 300-bus test system. Compared to FIR GF, IIR GF better approximates the desired filter for the same K (e.g., $\tilde{h}_{cheb_3}$ vs $\tilde{h}_{arma_3}$) and it requires a lower K for the same level of approximation (e.g., $\tilde{h}_{cheb_{11}}$ vs $\tilde{h}_{arma_5}$).

B. Data Generation

Since there is no publicly available dataset in the task of cyberattack localization, we generate a synthetic dataset using historical load profiles. First, we download 5-minute intervals of the actual load profile of NYISO for July 2021 and interpolate them to increase the resolution to 1-minute. Next, we generate a realistic dataset for IEEE 300-bus test system using 1-minute interval load profile. Specifically, the load values are distributed and scaled proportional to their initial values, AC power flow algorithms are executed, and 1% noisy power measurements are saved for each timestamp.

To simulate the cyberattacks, we implement some of the frequently used FDIA generation algorithms in the literature, such as data replay attacks (A_r) [19], data scale attacks (A_s) [13], distribution-based (A_d) attacks [20], and optimization based attacks (A_o) [21], [22]. A measurement z_o^i is replaced with one of its previous values in A_r, it is multiplied with a number sampled from a uniform distribution ($\mathcal{U}$) between 0.9 and 1.1 in A_s, and it is changed with a value drawn from the Gaussian distribution satisfying the same mean and variance with it in A_d. A_o solves a constrained optimization problem to maximize the state variables' deviation while minimizing the attack power on measurements.

We shuffle the whole data for seasonality elimination, scale it with the normal scaler for faster training, and split it into three sections: 4/6 for training, 1/6 for validation, and 1/6 for testing the proposed models. We arbitrarily select A_o and A_d and include them in the training and validation splits to evaluate the performance of our method under unseen attack types. Test split, on the contrary, includes all four cyberattacks. To have a balanced classification problem, the number of attacked samples are kept equal with the number of unattacked samples in each split. The final dataset has 34560 samples, where each sample consist of complex power measurements and n binary labels for the attack presence at each bus.

C. Features, Training, and Metrics

Since $\boldsymbol{P_i} + j\boldsymbol{Q_i} = \sum_{k\in\Omega_i} \boldsymbol{P_{ik}} + j\boldsymbol{Q_{ik}}$, node features can represent branch features as summation in their corresponding set of buses Ω_i connected to bus i. Therefore, we only feed $\boldsymbol{P}$ and $\boldsymbol{Q}$ values to the model as seen from the input layer of Fig. 3. In addition, we select $\boldsymbol{W} = |\boldsymbol{Y_{bus}}|$ to calculate $\tilde{\boldsymbol{L}}$ and feed the IIR_K layers where $\boldsymbol{Y_{bus}} \in \mathbb{R}^{n\times n}$ represents the nodal admittance matrix of the power grid.

We use $F1$ score $F1 = \frac{2*TP}{2TP+FP+FN}$ to evaluate model performances, where TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively. In addition, to overcome the division by zero problem we assume $F1 = 1$ when there is no attack at all and all labels are correctly predicted as not attacked. Otherwise, we assign $F1 = 0$ even if there is one mismatch.

We employ multilabel supervised training using the binary cross-entropy loss to compute unknown trainable parameters of the model. Training samples are fed into the model as mini batches of 256 samples with 256 maximum number of epochs. Moreover, we utilize early stopping criteria where 16 epochs are tolerated without any improvement in the validation set's cross entropy loss. Implementations are carried out in Python 3.8 on Intel i9-8950 HK CPU 2.90GHz with NVIDIA GeForce RTX 2070 GPU.

D. Localization Results

We implement other data-driven methods from the literature to compare them with our method. To the best of our knowledge, [13] is the only data-driven approach in the literature in which LSTM based model is used for cyberattack localization. Thus, we train an LSTM based localizer with our dataset to compare model performances. Moreover, despite the fact that they are proposed for cyberattak detection, we implement other data-driven methods from the literature that are suitable for the multi-label classification task such as Fully Connected Network (FCN) [23], Convolutional Neural Network (CNN) [24], and FIR-GNN [21], [22]. We train, validate and test these models similarly to the proposed IIR-GNN model using the generated dataset. Models are trained on the training set and their hyper-parameters are optimized on the validation set by Bayesian optimization techniques for each model in 250 trials.

Cyberattack localization is a multi-label classification problem, therefore, it can be evaluated in two possible ways: (i) bus-wise (BW) evaluation where each bus is evaluated separately along the samples, and (ii) sample-wise (SW) evaluation where each sample at a fixed time-step is treated individually along the buse. Thus, for detailed assessment, we analyze the distributions of BW and SW localization results in $F1$ percentages by the ratio of items satisfying some predetermined thresholds which provides quantifiable metrics to assess model performance. For instance, the percentage of samples (buses) having $F1 \geq 90\%$ in SW (BW) evaluation

are used to measure the ratio of "acceptable" samples (buses) in the distributions.

Localization results are plotted in Fig. 5. Only IIR-GNN model reaches 93% $F1$ level in both SW and BW evaluation. To be specific, in 92.95% of the samples, the localization ratio is greater than 90% in the IIR-GNN model. Similarly, in 93.33% of the buses, which corresponds to 279 buses for IEEE 300-bus test system, attack localization success is greater than 90% $F1$ level. Its "acceptable" ($F1 \geq 90\%$) percentages are 9.2%, and 14% greater than the second best model FIR-GNN in SW and BW localization, respectively.

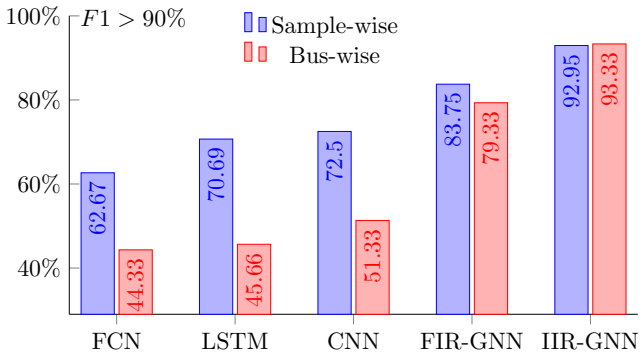


Fig. 5. Ratio of $F1$ scores greater than 90% for sample-wise and bus-wise evaluation of localization.

For the target test system having 300 buses, the localization delays are measured as 1.50, 99.78, 2.73, 2.71, 2.94 milliseconds for the FCN, LSTM, CNN, FIR-GNN, and IIR-GNN models, respectively. So, except for the LSTM architecture, all models can be considered real-time compatible.

TABLE I
JOINT DETECTION AND LOCALIZATION TIMES IN MILLISECONDS.

FCN	**LSTM**	**CNN**	**FIR-GNN**	**IIR-GNN**
1.50	99.78	2.73	2.71	2.94

V. CONCLUSION

To effectively model the smart grid data's implicit graph structure, we utilized IIR GNN for cyberattack localization purposes. As a first step, by comparing their empirical frequency responses we verified that IIR GFs better approximate the desired filter response compared to the FIR GFs. Next, we present a multilayered IIR GNN model to localize the cyberattack at the bus level resolution. Then, we validate the proposed model along with the existing architectures under distinct cyberattack models using both sample-wise (SW) and bus-wise (BW) evaluations. It is numerically shown that the proposed model surpasses the state-of-the-art FIR GNN model by 9.2% and 14% in SW and BW localization, respectively.

REFERENCES

[1] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[2] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[3] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE signal processing magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[4] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[5] Filippo Maria Bianchi, Daniele Grattarola, Lorenzo Livi, and Cesare Alippi, "Graph neural networks with convolutional arma filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] Xuesong Shi, Hui Feng, Muyuan Zhai, Tao Yang, and Bo Hu, "Infinite impulse response graph filters in wireless sensor networks," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1113–1117, 2015.

[7] Xinghuo Yu and Yusheng Xue, "Smart grids: A cyber–physical systems perspective," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1058–1070, 2016.

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

[9] Ahmed S Musleh, Guo Chen, and Zhao Yang Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2218–2234, 2019.

[10] Thomas R Nudell, Seyedbehzad Nabavi, and Aranya Chakrabortty, "A real-time attack localization algorithm for large power system networks using graph-theoretic techniques," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2551–2559, 2015.

[11] Mohsen Khalaf, Amr Youssef, and Ehab El-Saadany, "Joint detection and mitigation of false data injection attacks in agc systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 4985–4995, 2018.

[12] Md Abul Hasnat and Mahshid Rahnamay-Naeini, "Detection and locating cyber and physical stresses in smart grids using graph signal processing," *arXiv preprint arXiv:2006.06095*, 2020.

[13] Ana Jevtic, Fengli Zhang, Qinghua Li, and Marija Ilic, "Physics- and learning-based detection and localization of false data injections in automatic generation control," *IFAC-PapersOnLine*, vol. 51, no. 28, pp. 702–707, 2018.

[14] Yao Liu, Peng Ning, and Michael K Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, pp. 1–33, 2011.

[15] Andreas Loukas, Andrea Simonetto, and Geert Leus, "Distributed autoregressive moving average graph filters," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1931–1935, 2015.

[16] Andreas Loukas, Marco Zuniga, Matthias Woehrle, Marco Cattani, and Koen Langendoen, "Think globally, act locally: On the reshaping of information landscapes," in *Proceedings of the 12th international conference on Information processing in sensor networks*, 2013, pp. 265–276.

[17] Elvin Isufi, Andreas Loukas, Andrea Simonetto, and Geert Leus, "Autoregressive moving average graph filtering," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 274–288, 2016.

[18] Osman Boyaci, Mohammad Rasoul Narimani, Katherine Davis, Muhammad Ismail, Thomas J Overbye, and Erchin Serpedin, "Joint detection and localization of stealth false data injection attacks in smart grids using graph neural networks," *IEEE Transactions on Smart Grid*, pp. 1–1, 2021.

[19] Gu Chaojun, Panida Jirutitijaroen, and Mehul Motani, "Detecting false data injection attacks in ac state estimation," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2476–2483, 2015.

[20] Mete Ozay, Inaki Esnaola, Fatos Tunay Yarman Vural, Sanjeev R Kulkarni, and H Vincent Poor, "Machine learning methods for attack detection in the smart grid," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 8, pp. 1773–1786, 2015.

[21] Osman Boyaci, Amarachi Umunnakwe, Abhijeet Sahu, Mohammad Rasoul Narimani, Muhammad Ismail, Katherine R. Davis, and Erchin Serpedin, "Graph neural networks based detection of stealth false data injection attacks in smart grids," *IEEE Systems Journal*, pp. 1–12, 2021.

[22] Osman Boyaci, M Rasoul Narimani, Katherine Davis, and Erchin Serpedin, "Cyberattack detection in large-scale smart grids using chebyshev graph convolutional networks," in *2022 9th International Conference on Electrical and Electronics Engineering (ICEEE)*. IEEE, 2022, pp. 217–221.

[23] Yi Wang, Mahmoud M Amin, Jian Fu, and Heba B Moussa, "A novel data analytical approach for false data injection cyber-physical attack mitigation in smart grids," *IEEE Access*, vol. 5, pp. 26022–26033, 2017.

[24] Defu Wang, Xiaojuan Wang, Yong Zhang, and Lei Jin, "Detection of power grid disturbances and cyber-attacks based on machine learning," *Journal of Information Security and Applications*, vol. 46, pp. 42–52, 2019.