Received 5 April 2024; revised 26 August 2024 and 10 November 2024; accepted 20 December 2024.
Date of publication 6 January 2025; date of current version 24 January 2025.

Digital Object Identifier 10.1109/OAJPE.2024.3524268

Graph Neural Network-Based Approach for Detecting False Data Injection Attacks on Voltage Stability

SHAHRIAR RAHMAN FAHIM[1] (Member, IEEE), RACHAD ATAT[2] (Senior Member, IEEE), CIHAT KECECI[1], ABDULRAHMAN TAKIDDIN[3] (Member, IEEE), MUHAMMAD ISMAIL[4] (Senior Member, IEEE), KATHERINE R. DAVIS[1] (Senior Member, IEEE), AND ERCHIN SERPEDIN[1] (Fellow, IEEE)

[1] Electrical and Computer Engineering Department, Texas A&M University, College Station, TX 77843 USA
[2] Department of Computer Science and Mathematics, Lebanese American University, Beirut 03797751, Lebanon
[3] Department of Electrical and Computer Engineering, Florida State University, Tallahassee, FL 32310 USA
[4] Department of Computer Science, Tennessee Tech University, Cookeville, TN 38505 USA

CORRESPONDING AUTHOR: S. R. FAHIM (sr-fahim@tamu.edu)

This work was supported by the U.S. National Science Foundation, Energy, Power, Control, and Networks (EPCN) under Award 2220346 and Award 2220347.

ABSTRACT The integration of information and communication technologies into modern power systems has contributed to enhanced efficiency, controllability, and voltage regulation. Concurrently, these technologies expose power systems to cyberattacks, which could lead to voltage instability and significant damage. Traditional false data injection attacks (FDIAs) detectors are inadequate in addressing cyberattacks on voltage regulation since a) they overlook such attacks within power grids and b) primarily rely on static thresholds and simple anomaly detection techniques, which cannot capture the complex interplay between voltage stability, cyberattacks, and defensive actions. To address the aforementioned challenges, this paper develops an FDIA detection approach that considers data falsification attacks on voltage regulation and enhances the voltage stability index. A graph autoencoder-based detector that is able to identify cyberattacks targeting voltage regulation is proposed. A bi-level optimization approach is put forward to concurrently optimize the objectives of both attackers and defenders in the context of voltage regulation. The proposed detector underwent rigorous training and testing across different kinds of attacks, demonstrating enhanced generalization performance in all situations. Simulations were performed on the Iberian power system topology, featuring 486 buses. The proposed model achieves 98.11% average detection rate, which represents a significant enhancement of 10-25% compared to the cutting-edge detectors. This provides strong evidence for the effectiveness of proposed strategy in tackling cyberattacks on voltage regulation.

INDEX TERMS Cybersecurity, voltage regulation, graph autoencoder, voltage stability, false data injection attacks, bad data intrusion, machine learning.

I. INTRODUCTION

ELECTRIC power systems are becoming more sophisticated and operating more efficiently, but at the same time, they are frequently closer to their stability limits and exhibit more reduced margins for security [1]. Exceeding these stability limits or neglecting the security margins can lead to significant consequences, including the risk of large-scale blackouts. This makes examining a power system's stability conditions–particularly its voltage stability–a priority [2]. Voltage instability manifests through the system's inability to maintain permissible steady-state voltages across all of its buses during normal operating conditions and/or after physical disturbances. The primary contributing factors to this phenomenon include system overloading, shortage of reactive power, and equipment failures. The major consequences of voltage instability encompass load curtailment, cascade tripping of power components, or even blackout. Voltage instability events have been identified as pivotal factors in several worldwide blackouts, such as the blackout incident in Egypt on April 24, 1990 [3] and the 2012 blackout in India triggered by the overloading of transmission lines [4]. Thus, a fast and precise assessment of voltage stability is necessary to prevent large-scale blackouts.

Within an electrical grid, voltage-regulating equipment such as switched capacitor banks, step voltage regulators, on-load tap changers, static VAR compensators, and smart inverters in photovoltaic (PV) systems function together in a coordinated manner to ensure voltage stability by minimizing voltage fluctuations and system oscillations [5]. In general, grid stabilization is performed by injecting reactive power into the system. Also, a shortage of reactive power may lead to a drop in voltage levels, and the voltage drop may overload other lines, trip transmission lines, or initiate cascading failures. Therefore, regulating devices disconnect the generators to prevent overheating. To counteract the voltage drop, the system responds by injecting reactive power to maintain a stable voltage level. However, this injection leads to a decrease in reactive power, causing further voltage drops. This process results in a progressive and continuous reduction of voltage levels that ultimately leads to a voltage collapse. As part of grid modernization, voltage-regulating devices are now being controlled remotely using a variety of communication technologies. The adoption of communication channels and automation systems has increased the vulnerability of voltage regulation networks to cyberattacks.

In false data injection attacks (FDIAs), malicious actors falsify the voltage readings by making them to seem high (a.k.a. additive attacks), low (a.k.a. deductive attacks), or a a blend of the two (a.k.a. camouflage attacks), to compromise the voltage stability index. In case of additive attacks, attackers manipulate voltage measurements to apparently high voltage levels which give the misleading impression of instability. This may lead further to an overestimation of system robustness and pushing the system to operate closer to its stability margins [6]. Conversely, in deductive attack scenarios, attackers manipulate voltage measurements to make them appear lower than the original values. Combined attacks present a unique challenge, as they involve the manipulation of both high and low voltage values, making them harder to detect. These attacks can introduce fluctuations in the voltage levels which, in turn, disrupt the voltage stability index. Therefore, this study proposes a novel FDIA detector to prevent voltage instability in power grids and evaluates the impact of these attacks on grid's stability.

A. RELATED WORKS

The literature on cyberattack detection in power systems is diverse, encompassing traditional techniques like anomaly detection and signature-based methods, as well as more advanced methods such as deep learning and graph theory-based strategies [7]. The earlier attack detection strategies utilize residuals within the actual and measured data [8]. If the residual exceeds a specific threshold, it raises concerns about the potential presence of bad data. Despite the extensive utilization of these approaches, it has been evidenced that FDIAs can evade these detectors.

With the goal of developing robust FDIA detection strategies, recent approaches employed the Kullback Leibler (KL) distance [9], Bayesian framework [10], and Markov chain framework [11]. However, these approaches often face difficulties to detect FDIAs when it comes to detecting attacks with identical distribution as the historical measurements. They tend to be most effective at detecting attacks that create unusual system conditions.

Recently, machine learning and deep learning-based FDIAs detection mechanisms have attracted attention due to their inherent features learning ability [12]. In this regard, feed-forward neural network (FNN)-based FDIA detectors with obtained accuracy exceeding 90% were reported in [13] and [14]. In [15], the authors introduced a detection scheme using a Support Vector Machine (SVM), which achieved an F1 score of 82%. A strategy that combines the Kalman filter and recurrent neural network (RNN) was proposed in [16], and achieved a 96% detection accuracy. Another hybrid strategy that combines an autoencoder with a generative adversarial network (GAN) reported a detection performance of 96.2% [17]. In search for a reinforcing solution, the study [18] introduced a variational autoencoder for anomaly detection in power grids. A comparative performance analysis conducted in [19] demonstrated that the deep belief network (DBN)-based implementation outperforms extreme learning machine (ELM)-based detectors [20] and residual-based detectors. Furthermore, a detector based on convolutional neural networks (CNNs) achieved 99% detection accuracy, where a CNN and Kalman filter were employed to model temporal and spatial data correlations [21]. One drawback of these methods is that they do not fully extract the spatial relationships inherent in the sensor measurement data, as they overlook the topological features of the power systems [22], [23].

Graph-centric attack detection strategies provide a compelling solution to overcome the inherent limitations of traditional deep learning models. One of the key advantages of employing graph-based methods is their ability to extract spatial as well as temporal features from the graph-structured power system data [24]. A study in [22], reported a 4% improvement in the F1 score compared to a GNN-based detector. To detect unobservable attacks, an auto-regressive moving average (ARIMA)-based model was proposed in [25] that helps the detection model to adapt better to sudden variation in the spectral domain. In [26], a revised version of the multi-temporal graph CNN was reported where the training phases of the graph convolutions as well as the multilayer perceptions are blended together to simultaneously illustrate the node features. Such a model achieved a remarkable 96% accuracy across different power system topologies. A hybrid strategy was proposed in [17] where graph CNN was combined with a long short-term memory (LSTM) module in order to achieve 96% detection rate. A graph autoencoder-based implementation was reported in [27] where the detector was tested on unseen topology. An ensemble detector utilizing graph autoencoder (GAE) demonstrated a 12% enhancement in detection performance over shallow detectors [28]. Furthermore, the authors in [29] conducted a comparative analysis of autoencoders equipped with attention mechanism (AAM), simple autoencoders (SAE), and variational autoencoders (VAE) for FDIAs detection. The results showed that AAM enhanced system resilience to cyberattacks and demonstrated superior detection performance. Despite the benefits provided by these detectors based on GNN, it is important to note that they are typically trained and tested on power systems without considering the voltage stability aspect. In general, power systems frequently experience voltage fluctuations and FDIAs can significantly affect the voltage stability index. Therefore, it is crucial to comprehensively explore the consequences of FDIAs on the voltage stability index.

A dedicated FDIA detection algorithm for voltage stability is essential due to the complex nature of maintaining reactive power levels across the grid, ensuring all buses remain within voltage limits. FDIAs on voltage measurements can introduce subtle deviations that traditional detectors might miss. Voltage stability issues often arise from small reactive power imbalances, which, if accumulated, misalign actual and perceived system states, increasing the risk of collapse. These issues worsen under stress, such as peak loads or post-fault scenarios, where the system has minimal error tolerance. A dedicated algorithm would monitor the voltage stability index, providing early warnings to prevent the system from reaching critical points, addressing voltage stability's unique vulnerabilities and associated risks.

B. CONTRIBUTIONS AND ORGANIZATION

The key contributions of this paper are summarized as follows.

- First, we overcome the shortcomings of existing FDIA detectors by introducing a GAE-based detector specifically for cyberattacks on voltage regulation. The proposed detector captures both the temporal and the spatial (topological) relationships inherent in the power grid data by employing Chebyshev convolutional operation.
- Second, we demonstrate that our proposed model effectively identifies FDIAs even when encountering previously unseen topological configurations. Our testing dataset includes scenarios that were not part of the model's training dataset. This corroborates the practical applicability and generalization abilities of our approach to real-world situations.
- Third, we employ a bi-level optimization framework to craft cyberattacks with enhanced effectiveness and stealthiness and to create a more potent threat to the voltage regulation. This enables validation of the detection performance in the presence of more challenging cyberattacks.
- Fourth, to showcase the efficacy of the proposed detector, we undertake comprehensive simulations, subjecting it to a range of power system attacks, including additive, deductive, and camouflage attacks. The simulations encompass two distinct scenarios: i) when the attacker targets random buses, and ii) when the attackers target the most vulnerable buses. The latter case aids in developing strong protection measures by highlighting vulnerabilities within the power grid. An additional contribution involves the development of sophisticated cyberattacks targeting both the distributed generators (DGs) and load buses.

The remainder of the paper is organized as follows. Section II explains voltage stability and its index calculation. Section III presents the bi-level (attacker-defender) optimization problem. The architecture of the proposed GAE model is presented in Section IV. Section V presents the attack modeling, strategies of attacks, and generation of normal and attack data. In section VI the benchmark strategies, their hyperparameters optimization, and performance evaluation metrics are discussed. The experimental results are illustrated in section VII and section IX provides the paper's conclusion.

II. VOLTAGE STABILITY

Voltage stability is achieved through voltage regulation where protective devices work in a coordinated manner to regulate the voltage. When there is a shortage of reactive power, generators equipped with special voltage regulation devices (such as VAR compensators) inject reactive power into the system. This injection counteracts the voltage drop and increases the voltage levels. Conversely, when there is an excess of reactive power in the system, the voltage regulation devices (such as capacitors and reactors) absorb that reactive power to counteract excessive voltage levels. In this paper, we consider voltage compensators that guarantee that the voltage applied to the loads stays within acceptable limits. Our study primarily focuses on steady-state voltage stability. We have chosen this focus because steady-state voltage stability ensures that the power system can maintain acceptable voltage levels under normal operating conditions and small disturbances over extended periods. This aspect of voltage stability is relevant for the cyber-attacks considered in this study. In the case of considered cyber-attacks, attackers aim to alter system parameters stealthily over time without causing immediate, large-scale disruptions.

A. VOLTAGE STABILITY INDEX

The voltage stability index is an indicator of power system health and operational reliability, quantifying the system's ability to maintain steady voltage levels. This index is designed to reach a marginal value as the system reaches close to the instability point. A power system's operational state remains within the specified range by ensuring voltage stability at each bus [30]. To assess the stability of the overall system, we considered both the bus and line voltage stability indices [31] which will be discussed next.

1) BUS VOLTAGE STABILITY INDEX

We express the bus voltage stability index at the i^{th} bus, Δ_{B}^{i}, as

$$\Delta_{\text{B}}^{i} = \left| 1 - \sum_{b=1}^{N_g} \boldsymbol{F}_{ib} \frac{V_b}{V_i} \right|, \tag{1}$$

where V_b denotes the voltage at the b^{th} generator bus, and V_i stands for the voltage at the i^{th} load bus. N_g and N_l indicate the number of generator buses and load buses, respectively. Matrix $\boldsymbol{F}$ is expressed in terms of the sub-matrices Y_{ii} and Y_{ib} of bus admittance matrix $\boldsymbol{Y}$ as follows:

$$[\boldsymbol{F}] = [-Y_{ii}]^{-1}\,[Y_{ib}]\,. \tag{2}$$

Let I_b, I_i, V_b, and V_i represent the current and voltage at the b^{th} generator bus and i^{th} load bus, respectively. The bus admittance matrix is decomposed into 4 sub-matrices as follows:

$$\begin{bmatrix} I_b \\ I_i \end{bmatrix} = \begin{bmatrix} Y_{bb} & Y_{bi} \\ Y_{ib} & Y_{ii} \end{bmatrix} \begin{bmatrix} V_b \\ V_i \end{bmatrix}. \tag{3}$$

The Δ_{B}^{i} index ranges from 0 to 1, with values 0 and 1 indicating low and high voltage instabilities, respectively.

The data concerning the operational status of a power system, such as magnitude of the voltage ($|V|$), is transmitted to the Energy Management Center (EMC) through an extensive communication channel. However, in the context of a communication network, this data is susceptible to malicious cyber-attacks. Such attacks can result in the alteration of the measurement data presented to the EMC operator and provide a false depiction of the voltage stability index. In the event of cyberattacks, the Δ_{B}^{i} index can be falsely altered at various buses. The manipulated Δ_{B}^{i} index, denoted as $\tilde{\Delta}_{\text{B}}^{i}$ at the i^{th} bus is expressed as

$$\tilde{\Delta}_{\text{B}}^{i} = \left| 1 - \sum_{b=1}^{N_g} \boldsymbol{F}_{ib} \frac{V_b}{\tilde{V}_i} \right|, \tag{4}$$

where $\tilde{V}_i$ indicates the false voltage measurement at bus i. Taking the average of Δ_{B}^{i} over all the buses gives the global bus voltage stability index for the whole system.

2) LINE VOLTAGE STABILITY INDEX

The bus voltages V_k and V_j at the ends of the line connecting buses i and j are related by

$$V_k = \sqrt{\left(V_j + \frac{P_{kj}R + Q_{kj}X_R}{V_j}\right)^2 + \left(\frac{P_{kj}X_R - Q_{kj}R}{V_j}\right)^2}, \tag{5}$$

where the active and reactive powers flowing from bus k to bus j are denoted by P_{kj} and Q_{kj}, respectively; R is the equivalent resistance and X_R is the reactance of the branch. The line voltage stability index Δ_{L}^{kj} of the line connecting buses k and j is given by:

$$\Delta_{\text{L}}^{kj} = \frac{V_k}{\sqrt{2V_j^2 + 2\left(P_{kj}R + Q_{kj}X_R\right)}}, \tag{6}$$

and satisfies this condition $0 < \Delta_{\text{L}}^{kj} < 1$. If the system operates close to the stability limit, the index will approach 1. For a power system with N_h branches, the index of the most unstable branch is selected as the indicator of overall line voltage stability, $\Delta_{\text{L}} = \max\{\Delta_{\text{L}}^{1}, \Delta_{\text{L}}^{2}, \ldots, \Delta_{\text{L}}^{N_h}\}$. The final overall voltage stability index Δ_o is determined by taking the maximum between the two indices, i.e., $\Delta_{\text{o}} = \max\{\Delta_{\text{L}}, \Delta_{\text{B}}\}$. By considering the stronger index, this approach acknowledges that the overall system can continue to function effectively as long as one aspect remains uncompromised. This approach will improve the resilience and robustness of the system against cyberattacks.

III. BI-LEVEL OPTIMIZATION PROBLEM FORMULATION

In this section, we formulate a bi-level optimization problem for voltage stability in the power systems. This problem involves an attacker and a defender with their distinct roles and objectives. In this context, the attacker aims to maximize the damage by disrupting the voltage stability of the system. On the other hand, the defender seeks to minimize the damage impact by implementing security measures. These conflicting objectives and strategies create a dynamic and adversarial environment where the actions of one party influence the choices made by the other. If an attacker successfully alters the voltage measurement at a bus, the original voltage at that bus will change by $\boldsymbol{\Delta V_a}$. The attacker's manipulations are denoted as $\boldsymbol{\Delta V_a} = [\Delta V_a^1, \Delta V_a^2, \ldots, \Delta V_a^{N_l}]$. Concretely, the defender uses load compensation devices to regulate the voltage by injecting the reactive power at load buses. The defender's action to the attack $\boldsymbol{\Delta V_a}$ is defined as $\boldsymbol{q_d} = [q_d^1, q_d^2, \ldots, q_d^{N_l}]$, where $\boldsymbol{q_d}$ is the reactive power vector that the defender can compensate against attack $\boldsymbol{\Delta V_a}$. Now, given the ΔV_a^i and q_d^i at i^{th} load bus, the utility function of the attacker is defined as

$$U_a(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}) = \sum_{i=1}^{N_l} P_a\left(h_i\right) \Delta_o^i, \tag{7}$$

where $P_a\left(h_i\right)$ is the probability of outcomes of attack at load bus i depending on the binary variable h_i [32]. If the attack on node i is successful, then $h_i = 1$; otherwise, $h_i = 0$. The defender's utility function is defined as

$$U_d(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}) = -U_a(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}). \tag{8}$$

Given the utility functions of both attacker and defender, the bi-level optimization problem is formulated as:

$$g(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}) = \arg\max_{\boldsymbol{\Delta V_a}} U_a(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}) \tag{9}$$

$$f(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}) = \arg\max_{\boldsymbol{q_d}} U_d(g(\boldsymbol{\Delta V_a}, \boldsymbol{q_d}), \boldsymbol{q_d}), \tag{10}$$

$$\text{s.t. } upper - levelconstraints:$$

$$\begin{cases} P_i = \sum\limits_{i',i'\neq i} \alpha_{i,i'} \left(V_i - V_{i'}\right) + \alpha'_{i,i'} \left(\delta_i - \delta_{i'}\right), \forall i \in \{N_l\} \\ Q_i = \sum\limits_{i',i'\neq i} \alpha_{i,i'} \left(V_i - V_{i'}\right) - \alpha'_{i,i'} \left(\delta_i - \delta_{i'}\right), \forall i \in \{N_l\} \\ \sum\limits_i P_i^{\text{inj}} - \sum\limits_i P_i^{\text{con}} = 0 \\ \sum\limits_i Q_i^{\text{inj}} - \sum\limits_i Q_i^{\text{con}} = 0 \\ P_i^{\min} \leq P_i \leq P_i^{\max}, \forall i \in \{N_l\} \\ Q_i^{\min} \leq Q_i \leq Q_i^{\max}, \forall i \in \{N_l\} \\ V_i^{\min} \leq V_i \leq V_i^{\max}, \forall i \in \{N_l\} \\ \sum\limits_{i=1}^{n_a} C_i \leq B_a \end{cases} \tag{11a}$$

$lower - levelconstraints:$

$$\begin{cases} q_d^i \leq q_d^{i,\max}, \forall i \in \{N_l\} \\ \sum\limits_i Q_i^{\text{inj}} - \sum\limits_i Q_i^{\text{con}} = 0 \\ Q_i^{\min} \leq Q_i \leq Q_i^{\max}, \quad \forall i \in \{N_l\} \\ \sum\limits_i L_i^{\text{s}} = 0, \end{cases} \tag{11b}$$

where

- P_i, Q_i: active and reactive power at bus i, resp.
- $P_i^{\text{inj}}, Q_i^{\text{inj}}$: injected active and reactive power at bus i, resp.
- $P_i^{\text{con}}, Q_i^{\text{con}}$: consumed active and reactive power at bus i, resp.
- $P_i^{\min}, P_i^{\max}$: minimum and maximum active power limit at bus i, resp.
- $Q_i^{\min}, Q_i^{\max}$: minimum and maximum reactive power limit at bus i, resp.
- $V_i^{\min}, V_i^{\max}$: minimum and maximum voltage limit at bus i, resp.
- C_i, B_a: cost of attacking bus i and total attacker's budget, resp.
- L_i^{s}: load shedding at bus i.

Eq. (9) represents the upper-level objective function defined from the attacker's perspective that aims to maximize the disruption of the voltage stability index. For defender's action $\boldsymbol{q_d}$, the attacker determines a strategy pair, $g(\boldsymbol{\Delta V_a}, \boldsymbol{q_d})$. The lower-level objective function defined in Eq. (10) represents the defender's perspective, whose aim is to maximize the compensation for the attacker's action.

The constraints of the optimization problem include the upper-level and lower-level constraints defined from the attacker and defender perspectives, respectively. The set of upper-level constraints in Eq. (11a) consists of (in order): the active and reactive power flow from bus i to i', where $\alpha_{i,i'} = R_{i,i'}/R_{i,i'}^2 + X_{i,i'}^2$, $\alpha'_{i,i'} = X_{i,i'}/R_{i,i'}^2 + X_{i,i'}^2$, and δ_i indicates the voltage angle at bus i; active and reactive power balance; the limit of the total active and reactive power demand; the minimum and maximum voltage limit; budget limit as the attacker has a limited budget for creating an attack. An attacker has a limited ability to attack a certain number of nodes, n_a due to budget constraints. The lower-level constraints in Eq. (11b) consist of (in order): maximum reactive power limit of a bus constrained by $q_d^{i,\max}$; reactive power balance; reactive power limit, and load shedding constraints. In the bi-level optimization framework, by iterating between the attacker and defender levels, the optimization framework provides a balance between effective defense and successful attack strategies.

IV. GAE-BASED ATTACK DETECTION SCHEME

The features of GAE-based attack detection framework are next reviewed.

A. COMPONENTS OF GRAPHS

Since an interconnected power system can be portrayed as graph structure, GAE-based methods can be implemented to comprehend its complex behavior and characteristics. In the considered graph representation, the components of the power grid are translated into graph elements with the buses representing the nodes and their connectivity serving as the edges. The modeling of power grids commonly takes the form of undirected interconnected weighted graphs [22], [27]. In this paper, we choose to represent the power system as an undirected graph since the directed graph's asymmetric adjacency matrix constrains the exchange of data within the network. Fig. 1 presents the considered undirected power system model. We define the power system graph as the triplet $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \boldsymbol{W})$. Each element of the set of nodes $\mathcal{N}$ corresponds to a specific bus of the power system. The set of edges $\mathcal{E}$ indicates the physical lines that interconnect the buses, facilitating the power flow throughout the network. The adjacency matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ models the weighted relationships between buses. When buses i and j are connected, the weight $\boldsymbol{W}_{ij}$ is assigned to the edge $e = (i, j)$. This representation facilitates the GAE to capture the system's state by integrating both the topological characteristics (i.e., the position of buses in space and their interconnections) and the temporal attributes (i.e., power flows) of the power grid.

B. LEARNING OBJECTIVE

We would like to classify input samples $\boldsymbol{X}$ into the two distinct categories, corresponding to the presence and absence of cyberattacks, respectively. The input samples contain the temporal measurement data for active and reactive power values, $\left[\boldsymbol{P}_t, \boldsymbol{Q}_t\right] \in \mathbb{R}^{n\times 2}$ at the t^{th} timestamp. Fig. 2 depicts the model's architecture. The input data is fed to the graph encoder layers l_E, which output the latent representation at latent layer l_H. The graph decoding layer l_D follows. The aim is to learn the data patterns from benign input samples so that at the time of testing any deviation from the normal operating state can be flagged. Such a deviation is measured by the reconstruction error η while reconstructing. The graph encoder and graph decoder functions are denoted by $E_{\mathcal{G}} = f_E\left(\boldsymbol{X}\right)$ and $D_{\mathcal{G}} = f_D\left(\boldsymbol{X}\right)$ respectively. The objective function

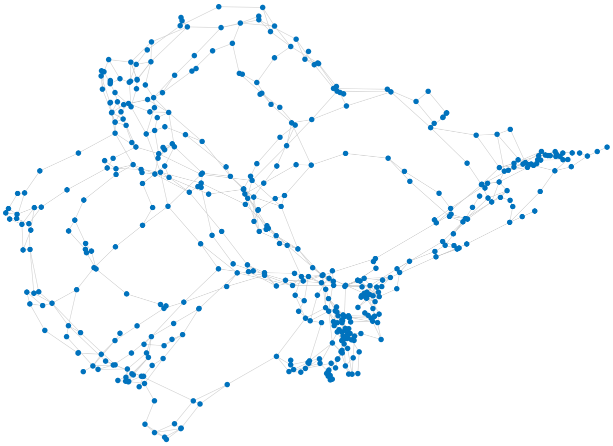

FIGURE 1. **Graph representation of the Iberian power system.**

of the proposed model is defined as

$$\min_{\{\mu\}} \mathcal{C}\left(\boldsymbol{X}, f_D\left(f_E(\boldsymbol{X})\right)\right), \tag{12}$$

where $\{\mu\}$ stands for the set of training parameters. The cost function $\mathcal{C}\,(.)$ represents the mean squared error and measures the dissimilarity between $f_D\left(f_E(\boldsymbol{X})\right)$ and $\boldsymbol{X}$.

C. CHEBYSHEV CONVOLUTION OPERATION

During each training stage, the spectral graph convolution of signal $\boldsymbol{\sigma} \in \boldsymbol{X}$ is performed as $\boldsymbol{U}\boldsymbol{\psi}_\theta\boldsymbol{U}^T\boldsymbol{\sigma}$. Matrix $\boldsymbol{U}$ incorporates the eigenvectors of normalized Laplacian $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Omega}\boldsymbol{U}^T$, spectral filter $\boldsymbol{\psi}_\theta = \text{diagonal}\,(\theta)$, and parameter vector is denoted as $\theta \in \mathbb{R}^n$ in the Fourier domain. The diagonal matrix $\boldsymbol{\Omega}$ captures the non-negative eigenvalues λ of $\boldsymbol{L}$. The Fourier transformation of $\boldsymbol{\sigma}$ is conducted through $\boldsymbol{U}^T\boldsymbol{\sigma}$. A major limitation of this filtering operation is that it is not always guaranteed to be spatially localized. Spatially localized filters are important as they can extract features from particular regions of interest within the input data, rather than performing filtering operations over the entire input sequence. This concern can be addressed by employing a polynomial approximation, $H_{\boldsymbol{\gamma}}(\boldsymbol{\Omega}) = \sum_{k=0}^{m} \gamma_k \boldsymbol{\Omega}^k$ where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)$ represents the vector of coefficients that the model seeks to learn for an m-order polynomial. To improve the training stability of these polynomial filters, a truncated Chebyshev polynomial expansion of $H_{\boldsymbol{\gamma}}(\boldsymbol{\Omega})$ is introduced in [33]. Concretely, the enhancement of $H_{\boldsymbol{\gamma}}(\boldsymbol{\Omega})$ using Chebyshev polynomials $N_m(\tilde{\boldsymbol{\Omega}})$ of order m is defined as

$$H_{\boldsymbol{\gamma}}(\boldsymbol{\Omega}) = \sum_{k=0}^{m} \gamma_k N_k(\tilde{\boldsymbol{\Omega}}) \tag{13}$$

where $\tilde{\Omega} = 2\boldsymbol{\Omega}/\lambda - \mathbf{I}$. The Chebysheb polynomial-based recursive formulation takes the form $N_m(p) = 2pN_{m-1}(p) - N_{m-2}(p)$ with $N_0 = 1$ and $N_1 = p$. The filtering process is expressed as

$$\boldsymbol{U}H_{\boldsymbol{\gamma}}(\boldsymbol{\Omega})\boldsymbol{U}^\top\boldsymbol{\sigma} = H_{\boldsymbol{\gamma}}(\boldsymbol{L})\boldsymbol{\sigma} = \sum_{k=0}^{m} \gamma_k N_k(\tilde{\boldsymbol{L}})\boldsymbol{\sigma}, \tag{14}$$

where $\tilde{\mathbf{L}} = 2\boldsymbol{L}/\lambda - \mathbf{I}$. The computational complexity of the filtering operation is $\mathcal{O}(m|\mathcal{E}|)$. Moreover, as the Chebyshev polynomials are limited to the m^{th} order, the filter operation is limited to m hops. This enables the implementation of localized Chebyshev convolutions.

D. GAE ARCHITECTURE

The GAE architecture consists of three layers: 1) encoder $E_{\mathcal{G}}$ that maps the input graph data into a lower-dimensional latent space; 2) latent layer $l_H(X)$; and 3) decoder $D_{\mathcal{G}}$ that reconstructs the original graph data from the latent space.

1) GRAPH ENCODER $E_{\mathcal{G}}$

The graph encoder has l_E Chebyshev graph convolutional layers. The inputs to the graph convolutional layers or the number of channels in a hidden encoding layer l_E is indicated by N_c. These layers extract the spatial characteristics from the network via graph convolution operations, bias addition, and the application of the ReLU activation function. The result is the output tensor denoted as

$$\boldsymbol{X}^{l_E} = \text{ReLU}\left(\gamma_m *_{\mathcal{G}} \boldsymbol{X}^{l_E-1} + \boldsymbol{b}^{l_E}\right). \tag{15}$$

Here $\boldsymbol{b}^{l_E}$ denotes the bias of layer l_E and $*_{\mathcal{G}}$ represents the graph convolutional operator. The bias in the ReLU activation function helps to capture non-linearities.

To extract the temporal relationships from the time-series signal, we incorporate an LSTM unit that facilitates the modeling of recurrent information flows. For each node, the output of the previous graph encoder is a vector that serves as the input to the LSTM layer. As the information flows through the LSTM, it maintains a memory of past information, allowing it to capture temporal dependencies and patterns over time. The LSTM layer adapts to handle sequential data and is ideal for time series processing. The memory module mitigates issues such as vanishing or exploding gradients that often arise during the learning process. An LSTM cell consists of the input $i_{l_E}^t$, output $o_{l_E}^t$, and forget gate $f_{l_E}^t$.

Inside an LSTM unit, there exist two distinct states: i) the cell state $\boldsymbol{C}_{l_E}^t$, which retains information for an extended period, and ii) the LSTM output or hidden state $\boldsymbol{H}_{l_E}^t$. The two states are expressed as:

- $\boldsymbol{C}_{l_E}^t = f_{l_E}^t \boldsymbol{C}_{l_E}^{t-1} + i_{l_E}^t \tanh\left(\boldsymbol{W}_{l_E}\boldsymbol{X}_{l_E}^t + \boldsymbol{U}_{l_E}\boldsymbol{H}_{l_E}^{t-1} + \boldsymbol{b}_{l_E}\right)$
- $\boldsymbol{H}_{l_E}^t = o_{l_E}^t \tanh\left(\boldsymbol{C}_{l_E}^t\right)$.

$\boldsymbol{C}_{l_E}^{t-1}$ and $\boldsymbol{H}_{l_E}^{t-1}$ represent the previous cell and hidden states, respectively; $\boldsymbol{W}_{l_E}$ and $\boldsymbol{U}_{l_E}$ refers to the learning weights and $\boldsymbol{b}_{l_E}$ is the bias; and $\varphi\,(\cdot)$ signifies the non-linear activation function.

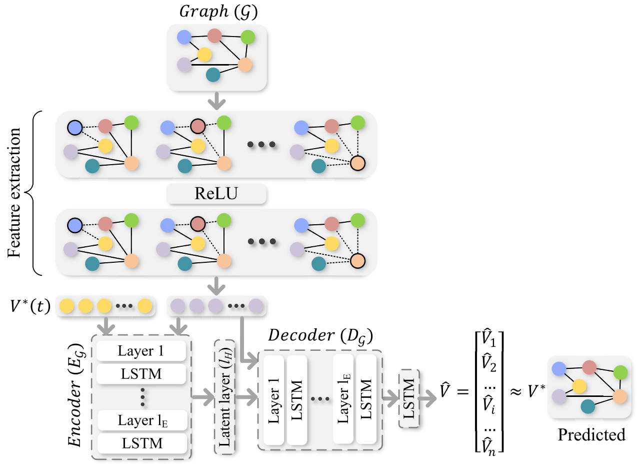
Graph $(\mathcal{G})$
Feature extraction
ReLU
$V^*(t)$
Encoder $(E_{\mathcal{G}})$
Layer 1
LSTM
Layer l_E
LSTM
Latent layer (l_H)
Decoder $(D_{\mathcal{G}})$
Layer 1
LSTM
Layer l_E
LSTM
LSTM
$\hat{V} = \begin{bmatrix} \hat{V}_1 \\ \hat{V}_2 \\ \vdots \\ \hat{V}_i \\ \vdots \\ \hat{V}_n \end{bmatrix} \approx V^*$
Predicted

FIGURE 2. **Architecture of the proposed GAE.**

2) LATENT LAYER l_H

l_H enables a compressed representation of the input information. The latent layer holds the compact representation of data which is then concatenated with $\boldsymbol{X}^{l_E}$ and conveyed to the graph decoder.

3) GRAPH DECODER $D_{\mathcal{G}}$

The main aim of the graph decoder is to produce an output $\boldsymbol{V}^*$ that closely resembles the input $\boldsymbol{X}$. The reconstruction error is measured via

$$\eta = \left\| \boldsymbol{V}^* - \boldsymbol{X} \right\|^2 . \tag{16}$$

In the same vein as the graph encoder, the outputs of the graph decoder are sequentially fed to the LSTM that processes time-evolving graph features. The LSTM updates its current hidden state $\boldsymbol{H}_{l_D}^t$ based on the current input from the graph decoder layer and the previous hidden state $\boldsymbol{H}_{l_D}^{t-1}$ seamlessly. The update mechanism allows LSTM to retain in memory the previous graph states and the capturing of temporal dependencies across graph states. The cell state of the graph decoder-LSTM is regulated by $i_{l_D}^t$, $o_{l_D}^t$, and $f_{l_D}^t$, which stand for the input, output, and forget gates, respectively. The decoder cell and hidden state are given by:

- $\boldsymbol{C}_{l_D}^t = f_{l_D}^t \boldsymbol{C}_{l_D}^{t-1} + i_{l_D}^t \tanh\left(\boldsymbol{W}_{l_D}^C \boldsymbol{X}_{l_D}^t + \boldsymbol{U}_{l_D}^C \boldsymbol{H}_{l_D}^{t-1} + \boldsymbol{b}_{l_D}^C \right)$.
- $\boldsymbol{H}_{l_D}^t = o_{l_D}^t \tanh\left(\boldsymbol{C}_{l_D}^t \right)$.

E. TRAINING AND TESTING

At each timestamp t, each node (i.e. bus) of the constructed graph contains a voltage signal, $\boldsymbol{V}(v_i, t)$, used as an input to the GAE model. As discussed in Section IV-B, the objective is to reconstruct the voltage tensor, $\boldsymbol{V}^*(t) =< v_1^*(t), v_2^*(t), \ldots, v_n^*(t) >$ from the input $\boldsymbol{V} =< v_1(t), v_2(t), \ldots, v_n(t) >$. Given a training dataset that contains $|n_{\mathrm{T}}|$ training samples, our task involves training GAE with the aim of maximizing the likelihood of $\boldsymbol{V}^*$. The GAE model is designed to distinguish benign samples from malicious samples based on the reconstruction error (12). During training, the model extracts the features from the normal condition data. At the time of testing, it can identify abnormalities by performing a comparison between the reconstructed samples and the actual samples. To expedite the training process, we partitioned the training data into batches of equal size, $\boldsymbol{X}_n \in \boldsymbol{X}$, which are processed through the model for 132 epochs with a specific learning rate ζ.

V. THREAT MODELING AND DATA GENERATION

A. THREAT MODEL

If the voltage measurement at bus i and timestamp t is denoted as V_i^t, then the true voltage measurement, $V_{\mathrm{true},i}^t$ should align with the measured voltage, $V_{\mathrm{m},i}^t$ at control end (i.e., $V_{\mathrm{true},i}^t = V_{\mathrm{m},i}^t$). The tampered voltage measurement may contain false data values. The attack functions during different attack scenarios are represented as

$$\mathit{Attackfunctions} \begin{cases} V_{\mathrm{false},i}^t = V_{\mathrm{true},i}^t + \Delta V_i^t \\ V_{\mathrm{false},i}^t = V_{\mathrm{true},i}^t - \Delta V_i^t \\ V_{\mathrm{false},i}^t = V_{\mathrm{true},i}^t + e \cdot \Delta V_i^t - (1-e) \cdot \Delta V_i^t, \end{cases} \tag{17a}$$

where ΔV_i^t denotes the maliciously inserted data by the adversary, and e denotes a binary variable with a value of 1 indicating an additive attack and 0 representing a deductive attack. The attack functions in Eq. (17a) incorporate additive, deductive, and combined attacks. During the additive attacks, the manipulation intends to keep the voltage measurements within the permissible range, while in reality, the actual voltage readings might fall below the considered threshold. Additive attacks can create a situation where an unusually high voltage level is introduced into the system. The excessive voltage, if not promptly detected and mitigated, may overload the protective elements and safety mechanisms. On the other hand, deductive attacks create the false impression of lower power levels than what truly exists in reality. The reduction in voltage level may lead to a false sense of security, potentially causing the voltage regulation elements to underreact or remain passive. Lastly, in the combined attack scenarios, the attacker creates a more complex attack pattern containing both additive and deductive attacks. The attacker may target some nodes with additive attacks and other nodes with deductive attacks, potentially leading to system inefficiencies and malfunctions.

B. IMPACT OF ATTACK SCENARIOS ON SYSTEM BEHAVIOR

In this section, we elaborate on how these attacks influence the system's dynamics and the overall reliability of the system.

1) ADDITIVE ATTACKS

Additive attacks result in apparently higher voltage readings, potentially causing the system to experience overvoltage. Due to this false sense of security, the system might not increase reactive power support or may delay shedding load to relieve stress on the system.

2) DEDUCTIVE ATTACKS

In contrast, deductive attacks decrease the voltage readings, potentially leading to undervoltage conditions. This false perception of instability may cause the system to overcompensate by injecting excessive reactive power. Such overcompensation can destabilize the system, leading to inefficient operations and potential voltage oscillations.

3) COMBINED ATTACKS

Combined attacks involve a combination of both apparently higher and lower voltage readings across different parts of the system. These fluctuating voltage readings can cause significant confusion for system operators, as some parts of the grid may appear to be stable while others seem unstable. This inconsistent data can lead to both under and over-compensation, where operators might simultaneously reduce reactive power in some areas while unnecessarily injecting it in others.

C. STRATEGIES FOR ATTACKS

1) RANDOM NODE ATTACKS

These attacks involve randomly selecting power buses as targets. In essence, they consist of randomly selecting r buses from a total of N_l buses. This random selection offers a multitude of possible subsets $N_l!/(r!(N_l - r)!)$ and their impact can cause severe voltage instability, particularly if the affected buses are not quickly returned to normal operation.

2) VULNERABLE NODES ATTACKS

Vulnerability concerns the likelihood of a power node acting as a possible weak spot in the system. A cyberattack on such a sensitive bus could potentially inflict significant voltage instability on the entire system. The objective of vulnerability evaluation is to assign vulnerability scores to individual nodes. We examined an extensive set of vulnerability metrics, containing electrical as well as topological metrics. The electrical vulnerability metrics comprise i) load shedding, which quantifies the total apparent power in the aftermath of a disruption, ii) betweenness centrality that evaluates the degree to which a bus is positioned along paths connecting two other buses, iii) effective graph resistance provides a comprehensive measure of the cost associated with power transfer between two buses of the network, and iv) degree centrality, representing the count of direct power flows affecting a power bus. The metrics for assessing topological vulnerability encompass i) degree centrality, representing the count of nodes and power lines that directly impact a node, ii) connectivity impact determines the count of nodes that stay interconnected following a disruption, iii) clustering coefficient, modeling the tendency of buses to form clusters, iv) connectivity loss, which measures the average reduction in the count of generation units following a disruption, v) betweenness centrality, indicating a bus's role in the smallest paths that connect two other buses.

We determine the weight factors for each vulnerability metric using the Analytical Hierarchical Process (AHP) that performs a pairwise comparison to establish their relative importance. We then calculate the topological vulnerability score by weighting and summing the topological metrics, and perform similar operations for the electrical vulnerability score. Finally, through a second AHP analysis, we attain the weights for both the electrical and topological vulnerability scores to calculate the overall vulnerability score.

D. DATA GENERATION

To generate the normal time-series voltage data, we perform power flow analysis using Newton's method in the MATLAB MATPOWER toolbox. This toolbox facilitates the calculation of system voltages, currents, and both real and reactive power flows. Initially, a scalar vector, F is created by normalizing the voltage data from the considered power system. This scaling factor is then applied to the voltage data from the preceding timestamp, using a normal distribution with a mean of $1+0.025F$ and a standard deviation of 0.01. This operation increases the dynamic range of the data, generating dynamic changes in the time-series voltage datasets.

This approach generates extensive spatiotemporal datasets for power systems in both standard and attack conditions, essential for developing detection mechanisms against voltage stability attacks. For the chosen system, 288 daily power dynamics snapshots (every five minutes) are recorded. Following the specified attack strategy, bad data are injected, and the resulting datasets, containing node voltage features and edge power flow features, are used to train and test the model. Input features provide a full view of system operations, while binary output labels classify samples as ''normal'' or ''anomalous.'' This structured data supports model training to enhance power system security against cyber threats, focusing on voltage stability.

VI. EXPERIMENTAL SETUP

In this section, we present the benchmark detectors and the process of hyperparameter selection. Choosing the optimal hyperparameters allows the detectors to achieve their best performance and provides a fair and balanced comparison within the detectors. Subsequently, in this section, we provide the definitions for the performance evaluation metrics.

A. PROPOSED MODEL SIMULATION

The developed GCNN-LSTM prediction algorithm takes the data sample, learning rate, regularization weight, the number of LSTM hidden layers as well as the number of graph convolution structures as input. The model is configured with 64 neurons per layer and incorporates a dropout rate of 0.2 to prevent overfitting. The Adam optimizer is used for efficient training, with an order of neighborhood K $=$ 5 to capture the local structure of the graph. The model is processed by first calculating each sample's adjacency and

Laplacian matrices. The encoder transforms these matrices into a latent representation, from which the decoder reconstructs the graph. The model's performance is optimized by minimizing the reconstruction loss through backpropagation, iterating until a predefined error threshold is reached. The training of the model, performed on an NVIDIA GeForce RTX 3080, takes approximately two to four hours to achieve optimal results. This GAE model is a robust tool for encoding and reconstructing graph structures with high accuracy.

B. BENCHMARK DETECTION STRATEGIES

We herein assess and compare the FDIAs detection performance of the GAE-based approach against two primary classes of detectors: the detectors based on graph theory and the traditional machine learning-based detectors. The graph-based detector is considered based on the formulation in [22]. On the other hand, the traditional machine-learning detectors include: 1) ARIMA model: a shallow unsupervised learning model that forecasts future patterns by reducing the mean squared error (MSE); 2) LSTM: a type of recurrent neural network (RNN) that equipped with specialized memory cells that can capture long-range dependencies and handle vanishing gradient problems; 3) feedforward neural network (FNN): a supervised architecture that extracts features by using a stack of hidden layers consisting of fully connected neurons; 4) CNN, which employs convolutional operation to learn the features adaptively; and 5) SVM: a supervised machine learning algorithm that works by finding a hyperplane separating different classes.

C. HYPERPARAMETER OPTIMIZATION

To optimize the detection performance, we utilize a sequential grid search algorithm to optimize the hyperparameters of both the proposed and benchmark detectors. The optimal hyperparameter, $\mathcal{H} = \{\text{number of layer, number of neurons in each layer, dropout rate, optimizer, activation function, order of neighborhood}\}$ for CNN, FNN, LSTM, GCNN and GNN are (in order): $\mathcal{H}_{\text{CNN}} = \{4, 32, 0.4, \text{Rmsprop}, 5, \text{Relu}\}$, $\mathcal{H}_{\text{FNN}} = \{4, 32, 0, \text{Adam}, \text{N/A}, \text{Relu}\}$, $\mathcal{H}_{\text{LSTM}} = \{3, 32, 0.2, \text{Adam}, \text{N/A}, \text{Relu}, \}$ $\mathcal{H}_{\text{GCNN}} = \{5, 32, 0.2, \text{Rmsprop}, 4, \text{RelU}\}$, and $\mathcal{H}_{\text{GAN}} = \{6, 64, 0.2, \text{Adam}, 5, \text{Relu}\}$. For ARIMA model, we conducted an exploration of the search space from the set $\{0, 1, 2, 3\}$. Ultimately, we determined that the optimal values for the differencing degree and moving average were 1 and 0, correspondingly. For the SVM model, we have chosen the optimal settings for the gamma, kernel, and regularization parameters as: auto, sigmoid, and 1.

D. PERFORMANCE EVALUATION METRICS

The performance metrics to evaluate the detection performance of the proposed FDIA detector are next discussed.

- Detection rate, $\text{DR} = \frac{\text{TP}}{\text{TP+FN}}$, measuring the ability to identify actual poisonous samples.
- False alarm rate, $\text{FAR} = \frac{\text{FP}}{\text{FP+TN}}$, indicating how frequently non-malicious samples are incorrectly flagged as threats.
- Accuracy, $\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, offering a comprehensive assessment of the detector's performance in detecting both malicious and benign samples.

In this context, TP, TN, FP, and FN signify the number of true positives, true negatives, false positives, and false negatives, respectively.

E. AHP OUTPUTS

We adopted the AHP methodology to determine the vulnerability score of each node. The AHP outputs for topological metrics are as follows: CSs neighborhood density: 0.3637, connectivity impact: 0.1612, connectivity loss: 0.2458, and betweenness centrality: 0.2294. For the electrical metrics, the AHP outputs are: load shedding: 0.7555, effective graph resistance: 0.0300, and electrical degree centrality: 0.2145. After obtaining the weights through AHP analysis, each normalized metric is multiplied by its respective weight. The weighted metric scores are then summed to calculate the vulnerability score of each node. Following this, we performed AHP analysis again to determine the overall weights for the topological and electrical metrics, yielding 0.8713 for topological and 0.1287 for electrical metrics.

VII. EXPERIMENTAL EVALUATIONS

This section presents the overall FDIA detection performance of the proposed model across different attack scenarios and injection levels.

A. OVERALL PERFORMANCE ANALYSIS

In this study, three different attack types are considered: additive, deductive, and camouflage attacks. For the latter attack strategy, both additive and deductive attacks are chosen in equal proportions. For each attack scenario, 5, 10, 15, and 20% attack injection levels are chosen. On average the proposed model achieves 98.11% accuracy (ACC), 98.76% detection rate (DR), and 8.13% false alarm rate (FAR). The model performs with relatively higher accuracy for additive and deductive attacks compared to the combined attacks. Moreover, the detection performance of the proposed approach is further evaluated with the F1-β score which allows to control the balance between false positives and false negatives using the β parameter. The F1-β score is defined as

$$\text{F1-}\beta = \frac{\left(1+\beta^2\right) \times \left(P_\beta \times R_\beta\right)}{\left(\beta^2 \times P_\beta + R_\beta\right)}, \tag{18}$$

where precision P_β indicates the accuracy of positive predictions and is defined as $P_\beta = \frac{\text{TP}}{\text{TP+FP}}$. Recall R_β assesses the model's ability to identify all relevant instances in the dataset and is defined as $R_\beta = \frac{\text{TP}}{\text{TP+FN}}$.

The F1-beta scores of the proposed model versus the β values in the interval [0.6 − 1.5] are depicted in Fig. 3. The scores approaching 1 indicate higher performance and

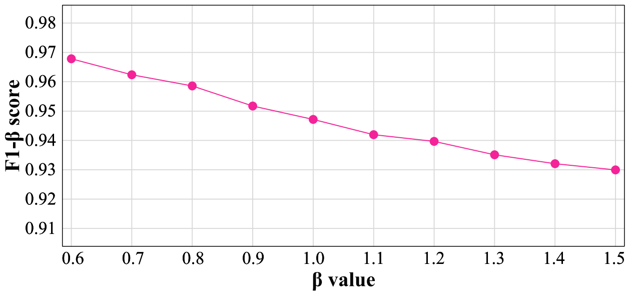

FIGURE 3. F1-beta score across different β values.

TABLE 1. Performance against random node attacks.

Attack type	Performance Metric	Injection levels			
		5%	10%	15%	20%
Additive	DR	99.18	99.07	98.17	96.40
	FAR	6.48	8.28	9.97	10.79
	ACC	98.87	98.77	97.88	95.80
Deductive	DR	98.10	97.33	96.48	94.47
	FAR	8.24	9.63	10.99	12.98
	ACC	97.03	97.09	96.50	94.61
Combined	DR	97.15	95.58	94.97	92.90
	FAR	10.38	10.93	12.80	13.72
	ACC	95.22	95.58	95.38	93.54

the scores approaching 0 indicate lower performance. The results shown in Fig. 3 confirm that the model achieves a balanced performance between false positives and false negatives.

B. PERFORMANCE AGAINST RANDOM ATTACKS ON BUSES

The proposed model's detection performance against random buses attacks is depicted in Table 1. The results reveal that as the injection level of the attack increases, the effectiveness of the detection decreases. This drop in performance can be attributed to the corresponding rise in false positives detection. In the event of additive attacks, the model attains 98.85-95.78% in ACC, 6.50-10.83% in FAR, and 99.20-96.42% in DR. In the event of deductive attack scenarios, the model shows almost similar performance. Compared to the additive and deductive attacks, the combined attack reports a 3-5% drop in detection performance. This performance drop may be due to the increased likelihood of false positives.

C. PERFORMANCE AGAINST ATTACKS ON VULNERABLE BUSES

During the most vulnerable bus attack scenario, an attacker targets the most susceptible power buses to maximize the damage. The performance of the proposed model for the mentioned attack strategy is presented in Table 2. From the table, it is observed that for each test case, the model reports relatively lower accuracy compared to the random node attacks. However, the model achieves more than 93% accuracy across the attack scenarios.

TABLE 2. Performance against vulnerable node attacks.

Attack type	Performance Metric	Injection levels			
		5%	10%	15%	20%
Additive	DR	98.31	97.78	97.01	95.88
	FAR	6.51	8.27	10.11	10.83
	ACC	98.07	97.31	96.49	94.73
Deductive	DR	98.22	97.67	96.99	95.75
	FAR	6.58	8.37	9.95	11.33
	ACC	98.00	97.17	96.37	94.56
Combined	DR	97.50	97.03	95.91	94.64
	FAR	9.56	10.22	11.13	12.66
	ACC	97.25	96.41	95.12	93.98

TABLE 3. Performance against DG node attacks.

Attack type	Performance Metric	Injection levels			
		5%	10%	15%	20%
Additive	DR	97.91	97.16	96.01	94.63
	FAR	6.59	8.57	10.99	11.26
	ACC	97.99	97.05	96.19	95.36
Deductive	DR	97.85	97.00	95.59	94.21
	FAR	6.64	8.71	11.07	11.89
	ACC	97.93	97.10	96.11	95.03
Combined	DR	97.30	96.82	95.22	94.00
	FAR	6.95	8.75	11.13	11.98
	ACC	97.76	96.90	95.49	94.94

D. DETECTION PERFORMANCE AGAINST DGs ATTACKS

In this section, we assess the performance of the proposed model in detecting cyberattacks specifically on DGs. Detecting attacks on DGs presents unique challenges due to the distributed and dynamic nature of these resources. In this study, we considered non-dispatchable DGs as they are more susceptible to the impacts of cyberattacks. Their output cannot be easily adjusted to counteract the effects of such attacks. If an attacker manipulates the voltage data associated with these generators, the grid operators have limited options to compensate for the discrepancy, potentially leading to stability issues. By focusing on non-dispatchable DGs, the study reflects real-world challenges and vulnerabilities in managing these types of energy sources, making the research more relevant to current and future power systems. Table 3 reports the model's performance in the presence of DGs attacks. Overall, from Table 3, we can conclude that the model achieves more than 87% accuracy over the test scenarios. Similar to the random and most vulnerable buses attack strategy, the performance for DGs attacks decreases with the increase of the attack injection levels.

E. COMPARATIVE DETECTION PERFORMANCE

In Fig. 4, the detection performance of the proposed GAE model is compared with other state-of-the-art detectors. To obtain a consistent comparison, all models are tested with optimal hyperparameters, as described in Section VI. The attack injection level for all the models is kept constant at 20%. From Fig. 4, it is observed that the proposed GAE-based detector achieves the highest DR and ACC and the lowest FAR. The graph CNN (GCNN)-based detector performs closer to the proposed GAE-based detector, yet the proposed model holds superiority over the GCNN model.

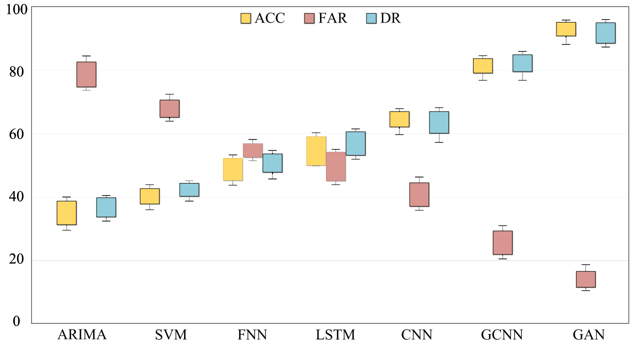

FIGURE 4. **Detection comparison with benchmark detectors.**

ARIMA model-based detector exhibits the poorest performance in terms of detection capabilities. This indicates that, in the context of the application, ARIMA's ability to detect specific criteria or anomalies falls behind that of the other models considered. The aforementioned relative performance analysis confirms the superiority of the proposed model.

F. PERFORMANCE AGAINST NOISE

In this section we introduced synthetic noise into the original data to mimic the real-world scenarios where data can be inherently tampered, noisy, or subject to uncertainties. By incorporating noise into our modeling process, we aim to create a more realistic representation of the system's behavior and topology. In a realistic system setting, the elements of power systems are exposed to interfering signals, for instance, corona noise, jet flyovers, insulator noise, wind-induced noise, or noise due to human intervention. This indicates that the measured signals from the power systems present irregular and changing properties, and important fault-related information may get hidden amidst strong noise. To mimic such a scenario, we included Additive White Gaussian Noise (AWGN) into the original data with a signal-to-noise ratio (SNR) ranging from 10 to 20 dB (as per [34]). We then tested our proposed system using this noisy data. The results presented in Fig. 5 reveal that within the considered noise range, the proposed model maintains a good detection performance. Specifically, only a 2% drop in detection performance is observed when compared to the low noise condition at 10 dB SNR. During the robustness analysis, we kept the training data free from noise and tested the system with noise-injected data it had not been exposed to before. This analysis confirms the noise-immune performance of the proposed method.

VIII. PRACTICAL IMPLEMENTATION

Implementing the research on detecting cyber attacks on voltage stability in a practical system setting involves several key steps. First, the GAE model developed in this study can be integrated into the power grid's existing monitoring and control infrastructure. This integration can occur at centralized control centers where real-time data from various nodes in the power grid are continuously collected and analyzed.

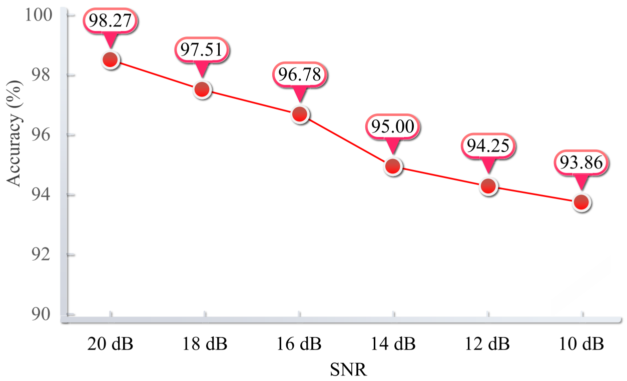

FIGURE 5. **Performance against different levels of noise.**

The model will process incoming voltage data, which could come from SCADA systems. The practical deployment may also include setting up automatic response mechanisms based on the model's detections. For instance, the system could isolate affected parts of the grid, adjust reactive power support, or initiate load shedding to prevent a voltage collapse. Additionally, the proposed model is trained on the year-round data, it generally does not require frequent retraining unless there are significant changes in the power system's topology or operational conditions. Moreover, periodic retraining could also be considered a best practice, especially in systems where operational patterns evolve over time due to factors like changing demand profiles and the increasing penetration of renewable energy sources.

IX. CONCLUSION

This study introduced a GAE-based FDIA detection framework in the context of voltage regulation. The proposed detector combines the autoencoder with Chebyshev graph convolution recurrent layers to effectively capture both spatial and temporal correlations in measurement data. A bi-level optimization framework is proposed to design cyberattacks and enhance the model's adaptability and resilience in the face of dynamic network changes and errors. In addition, we considered attacks on DGs that contribute to grid resilience and renewable energy integration. The extensive simulation studies conducted in this paper report that the proposed model achieves accuracy levels as high as 93.80%. A comparative performance analysis against the benchmark detectors shows an average of 20% improvement in ACC. The development of a generalized cyber attack detection scheme remains open for future investigations.

REFERENCES

[1] T. Hathiyaldeniye, U. D. Annakkage, N. Pahalawaththa, and C. Karawita, "A comparison of inverter control modes for maintaining voltage stability during system contingencies," *IEEE Open Access J. Power Energy*, vol. 9, pp. 55–65, 2022.

[2] P. Kessel and H. Glavitsch, "Estimating the voltage stability of a power system," *IEEE Trans. Power Del.*, vol. PD-1, no. 3, pp. 346–354, Jul. 1986.

[3] S. A. Adegoke and Y. Sun, "Power system optimization approach to mitigate voltage instability issues: A review," *Cogent Eng.*, vol. 10, no. 1, Dec. 2023, Art. no. 2153416.

[4] J. Sia, E. Jonckheere, L. Shalalfeh, and P. Bogdan, "Phasor measurement unit change-point detection of frequency Hurst exponent anomaly with time-to-event," *IEEE Trans. Depend. Secure Comput.*, vol. 21, no. 2, pp. 1–10, Mar./Apr. 2023.

[5] Y. Hua, Q. Xie, H. Hui, Y. Ding, J. Cui, and L. Shao, "Use of inverter-based air conditioners to provide voltage regulation services in unbalanced distribution networks," *IEEE Trans. Power Del.*, vol. 38, no. 3, pp. 1–10, Jun. 2022.

[6] P. Zhao, C. Gu, Y. Ding, H. Liu, Y. Bian, and S. Li, "Cyber-resilience enhancement and protection for uneconomic power dispatch under cyber-attacks," *IEEE Trans. Power Del.*, vol. 36, no. 4, pp. 2253–2263, Aug. 2021.

[7] J. Appiah-Kubi and C.-C. Liu, "Cyberattack correlation and mitigation for distribution systems via machine learning," *IEEE Open Access J. Power Energy*, vol. 10, pp. 128–140, 2023.

[8] A. Sayghe et al., "Survey of machine learning methods for detecting false data injection attacks in power systems," *IET Smart Grid*, vol. 3, no. 5, pp. 581–595, Oct. 2020.

[9] R. Moslemi, A. Mesbahi, and J. M. Velni, "A fast, decentralized covariance selection-based approach to detect cyber attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4930–4941, Sep. 2018.

[10] M. Zhang et al., "False data injection attacks against smart gird state estimation: Construction, detection and defense," *Sci. China Technol. Sci.*, vol. 62, no. 12, pp. 2077–2087, Dec. 2019.

[11] G. Chaojun, P. Jirutitijaroen, and M. Motani, "Detecting false data injection attacks in AC state estimation," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2476–2483, Sep. 2015.

[12] Y. Chen, A. Nath, C. Peng, and A. Kuhnle, "Discretely beyond $1/e$: Guided combinatorial algorithms for submodular maximization," 2024, *arXiv:2405.05202*.

[13] D. Xue, X. Jing, and H. Liu, "Detection of false data injection attacks in smart grid utilizing ELM-based OCON framework," *IEEE Access*, vol. 7, pp. 31762–31773, 2019.

[14] E. M. Ferragut, J. Laska, M. M. Olama, and O. Ozmen, "Real-time cyber-physical false data attack detection in smart grids using neural networks," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2017, pp. 1–6.

[15] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Syst. J.*, vol. 11, no. 3, pp. 1644–1652, Sep. 2017.

[16] Y. Wang, Z. Zhang, J. Ma, and Q. Jin, "KFRNN: An effective false data injection attack detection in smart grid based on Kalman filter and recurrent neural network," *IEEE Internet Things J.*, vol. 9, no. 9, pp. 6893–6904, May 2022.

[17] Y. Zhang, J. Wang, and B. Chen, "Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 623–634, Jan. 2021.

[18] R. Zheng, J. Gu, Z. Jin, H. Peng, and Y. Zhu, "Load forecasting under data corruption based on anomaly detection and combined robust regression," *Int. Trans. Electr. Energy Syst.*, vol. 30, no. 7, Jul. 2020, Art. no. e12103.

[19] Y. Li and Y. Wang, "False data injection attacks with incomplete network topology information in smart grid," *IEEE Access*, vol. 7, pp. 3656–3664, 2019.

[20] L. Yang, Y. Li, and Z. Li, "Improved-ELM method for detecting false data attack in smart grid," *Int. J. Electr. Power Energy Syst.*, vol. 91, pp. 183–191, Oct. 2017.

[21] G. Zhang, J. Li, O. Bamisile, D. Cai, W. Hu, and Q. Huang, "Spatio-temporal correlation-based false data injection attack detection using deep convolutional neural network," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 750–761, Jan. 2022.

[22] O. Boyaci et al., "Graph neural networks based detection of stealth false data injection attacks in smart grids," *IEEE Syst. J.*, vol. 16, no. 2, pp. 2946–2957, Jun. 2022.

[23] A. S. Musleh, G. Chen, and Z. Y. Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2218–2234, May 2020.

[24] B. L. H. Nguyen, T. V. Vu, T.-T. Nguyen, M. Panwar, and R. Hovsapian, "Spatial–temporal recurrent graph neural networks for fault diagnostics (don't short) in power distribution systems," *IEEE Access*, vol. 11, pp. 46039–46050, 2023.

[25] O. Boyaci, M. R. Narimani, K. R. Davis, M. Ismail, T. J. Overbye, and E. Serpedin, "Joint detection and localization of stealth false data injection attacks in smart grids using graph neural networks," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 807–819, Jan. 2022.

[26] Y. Han, H. Feng, K. Li, and Q. Zhao, "False data injection attacks detection with modified temporal multi-graph convolutional network in smart grids," *Comput. Secur.*, vol. 124, Jan. 2023, Art. no. 103016.

[27] A. Takiddin, R. Atat, M. Ismail, O. Boyaci, K. R. Davis, and E. Serpedin, "Generalized graph neural network-based detection of false data injection attacks in smart grids," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 618–630, Jun. 2023.

[28] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Robust electricity theft detection against data poisoning attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2675–2684, May 2021.

[29] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids," *IEEE Syst. J.*, vol. 16, no. 3, pp. 4106–4117, Sep. 2022.

[30] A. Agrawal, D. M. Momin, D. Syndor, and S. Affijulla, "Impact analysis of cyber attack under stable state of power system: Voltage stability," in *Proc. IEEE Region Symp. (TENSYMP)*, Jun. 2020, pp. 402–405.

[31] X. Ancheng et al., "On-line voltage stability index based on the voltage equation of transmission lines," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 14, pp. 3441–3448, Nov. 2016.

[32] L. An, A. Chakrabortty, and A. Duel-Hallen, "A Stackelberg security investment game for voltage stability of power systems," in *Proc. 59th IEEE Conf. Decis. Control (CDC)*, Dec. 2020, pp. 3359–3364.

[33] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html

[34] K. Chen, J. Hu, and J. He, "Detection and classification of transmission line faults based on unsupervised feature learning and convolutional sparse autoencoder," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 1748–1758, May 2018.