


Research Article

A Novel Model-Based Reinforcement Learning for Online Anomaly Detection in Smart Power Grid

Ling Wang ^{1,2}, Yuanzhe Zhu,^{1,2} Wanlin Du,^{1,2} Bo Fu,³ Chuanxu Wang,⁴ and Xin Wang⁵

¹Electric Power Research Institute of Guangdong Power Grid Co., Ltd., Guangzhou, Guangdong 510080, China

²Key Laboratory of Power Quality of Guangdong Power Grid Co., Ltd.,

Electric Power Research Institute of Guangdong Power Grid Co., Ltd., Guangzhou, Guangdong 510080, China

³Guangdong Power Grid Corporation Zhuhai Power Supply Bureau, Zhuhai, Guangdong 519099, China

⁴Guangdong Power Grid Corporation Dongguan Power Supply Bureau, Dongguan, Guangdong 523120, China

⁵CET Shenzhen Electric Technology Inc, Shenzhen, Guangdong 518040, China

Correspondence should be addressed to Ling Wang; wangleng136@gmail.com

Received 7 July 2022; Revised 8 August 2022; Accepted 12 August 2022; Published 28 April 2023

Academic Editor: Martin Calasan

Copyright © 2023 Ling Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Smart grids must detect cyber-attacks early to ensure their safety and reliability. There have been many outlier detection methods presented in the studies, varying from those requiring instance-by-instance decisions t the online diagnosing methods that require the use of accurate models of an attack. This study proposes a novel intelligent online anomaly or attack detection method based on the partially observable Markov decision procedure (POMDP). The proposed model may be categorized as a general detection method according to the reinforcement learning (RL) architecture for POMDP which can help the learning process based on the award concept. The performance of the proposed model is verified using the IEEE test system. Based on numerical results, the suggested RL-based algorithm shows to be very effective in detecting cyber-attacks against the smart grid quickly and accurately.

1. Introduction

The energy grids of the future, the so-called smart grid (SG), rely on enhanced communication and control technology to enhance the quality of the power generation and delivery to the end users. In this way, SGs are vulnerable to cyber-attacks because of these critical cyber infrastructures [1]. Attackers typically aim at damaging or misleading the SG's state estimation (SE) mechanism for generating large-scale energy outages or for manipulating power costs [2]. The most commonly popular kinds of cyber-attacks are denial of service (DoS), jamming, and false data injection (FDI) attacks. In FDI attack (FDIA) meter measurements are tampered with by adding malicious fake data [3, 4], in jamming attacks meter measurements are corrupted by adding additive noise [5], and DoS attacks prevent access of the system to meter measurements [6].

SGs are complex networks and failures or anomalies within them can result in severe damages to the entire

system. A quick and efficient response to cyber-attacks depends on detecting them as soon as possible. As a result, detecting a change as quickly as possible [7, 8] can be extremely beneficial. When quickest change detection is being used, changes in the sensing environment happen at unexpected times, and it aims at detecting the changes as quickly as possible with a minimum of false alarms (FAs) using measurements collected gradually over time. Once the decision-makers have obtained measurements for a particular time interval, they either make a change or wait until the next period to acquire additional measurements. The detection speed will decrease when the optimum detection accuracy improves. Therefore, the stopping time, when a change is declared, should be set such that the detection speed and the accuracy are optimally balanced. As the pre-change state and the post-change state are hidden due to the uncertain change-point, a partially observable Markov decision process (POMDP) problem can be used to model the quickest change detection problem. In the case of online

attacks and anomalies in the SGs, in the pre-change condition, the system has been run within usual situations, and the pre-change metering pdf is defined very precisely utilizing the system model.

The control of unknown environments is effectively possible with reinforcement learning (RL) algorithms. In this case, RL is used to efficiently solve the POMDP problem. One solution implies either learning the model underlying POMDPs and then implementing the model-based RL method for POMDPs [9] or applying a model-free RL (MF-RL) algorithm [10–12] with no learning the model. Due to the computational burden of the model-based method and just an approximate model being able to be learned, using the MF-RL method is preferred.

Outlier detection methods like the Euclidean detector [13] or detector according to the cosine-similarity [14] were general since they need no attacking pattern. Basically, they are computing the dissimilation metric among genuine expected and meter measurements by using the Kalman filter (KF) and if the dissimilarity goes above a particular level, an attack/anomaly is declared. This type of detector, though, does not take into account the temporal relationship among attacked or anomalous measurements and makes decisions on a sample-by-sample basis. As a result, they cannot differentiate immediate great-stage random noise from persistent anomalies, such as those resulting from unfriendly interventions. Accordingly, robust universal attack detection methods are more required than outlier detection methods.

RL methods (single-agent RL) are used to develop a useful detection method in the present study, which is based on the perspective of the defender. It should be noted that the problem could also be viewed from the attacker's side, in which case the goal would be to find the best attack strategy to cause as much damage to the system as possible. An analysis of this kind of problem can be extremely useful in identifying the most severe damage an attacker could inflict on the system and then taking precautions accordingly. Many investigations employ RL to analyze vulnerability, such as for FDI attacks in ref [15] and for sequential topology attacks in ref [16]. It should be noted that the problem could be viewed simultaneously from the perspective of the defender and the perspective of the attacker as well, which can correspond to a game-theoretic setting.

It is the multifactorial RL architecture, which extends standalone RL to multiplex-factors, which includes game theory as agents' optimal policies are driven by their environment as well as the policies of their peers. The stochastic game also extends the Markov decision process to the multiplex-factor status in which the game can be consecutive and includes more than one state, and both the transition from one state to the next as well as the payoffs (reward/cost) are determined by the common functions of whole factors. The solution methods based on RL for stochastic games are studied in ref [17], ref [18]. The partially observable stochastic game is one in which the environment, the functions, and Payments from other factors are observed partially, making identifying solutions increasingly problematic generally.

The goal of this paper is to develop online cyber-attack detection (CAD) method based on MF-RL for POMDP. As the suggested algorithm does not rely on attack models, it is universal and shows a general but robust performance. Consequently, the suggested layout can be broadly used, and it is proactive in that it can detect novel attack types. By following an MF-RL method, the defenders learn by trial-and-error how observations translate into actions (*stop* or *continue*). Although the model can be used to produce observation data under normal operating conditions for the pre-change state, obtaining real attack data can be usually challenging in the training phase. Due to this, a robust detection strategy is adopted that trains the defender with a low-magnitude attack corresponding to the worst cases from the perspective of the defender as detecting these types of attacks are challenging. Once trained, the defenders can identify minor changes from normal meter measurements. The robust detection method also considerably reduces the action space in which an attacker can operate. In other words, in order to avoid detection, attackers could just use small magnitudes of the attack, which are not problematic because of the minimum impact on the grid. To the best of the authors' knowledge, this is the first online CAD work in the SG that uses RL methods.

The model of the system and the SE method are described in Part 2. Part 3 describes the problem formulation and Part 4 proposes a solution. Part 5 demonstrates the effectiveness of the suggested RL-based detection method through a series of simulations. Part 6 concludes the study.

2. Model of the System and SE

2.1. Model of the System. If K meters exist in the system with $N + 1$ buses, then there should be $K > N$ in order to ensure the required measurement redundancy versus noise [19]. Assume that one of the buses has been taken as the reference bus, and $\mathbf{x}_t = [x_{1,t}, \dots, x_{N,t}]^T$ shows the system state at time t in which $x_{n,t}$ represents the phase angle at the time t at bus. $y_{k,t}$ shows the measurement taken at time t at meter k and $\mathbf{y}_t = [y_{1,t}, \dots, y_{K,t}]^T$ represents the measurement vector. The below state-space equations are used for modeling the SG according to the broadly applied linear DC model [19]:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t, \quad (1)$$

$$\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \mathbf{w}_t, \quad (2)$$

where the system (state transition) matrix is shown by $\mathbf{A} \in R^{N \times N}$, $\mathbf{H} \in R^{K \times N}$ represents the measurement matrix defined according to the topology of the network, the process noise vector is shown by $\mathbf{v}_t = [v_{1,t}, \dots, v_{N,t}]^T$, and the measurement noise vector is represented by $\mathbf{w}_t = [w_{1,t}, \dots, w_{K,t}]^T$. Considering \mathbf{v}_t and \mathbf{w}_t as independent additive white Gaussian random processes in which $\mathbf{v}_t \sim N(0, \sigma_v^2 \mathbf{I}_N)$, $\mathbf{w}_t \sim N(0, \sigma_w^2 \mathbf{I}_K)$, and $\mathbf{I}_K \in R^{K \times K}$ shows an identity matrix. A further assumption is that the network is observable, in other words, the observability matrix has rank N .

$$0 \triangleq \begin{bmatrix} \mathbf{H} \\ \mathbf{HA} \\ \vdots \\ \mathbf{HA}^{N-1} \end{bmatrix}. \quad (3)$$

Equations (1) and (2) give the system model of normal operation. When a cyber-attack occurs, though, the measurement model from equation (2) does not apply. As an example, in the case of a(n):

- (a) The measurement model for an FDI attack launched at time τ is:

$$y_t = \mathbf{H}\mathbf{x}_t + \mathbf{w}_t + \mathbf{b}_t \|\{t \geq \tau\}. \quad (4)$$

Here, an indicator function is shown by $\|\$ and the injected malicious data at time $t \geq \tau$ is represented by $\mathbf{b}_t \triangleq [b_{1,t}, \dots, b_{K,t}]^T$ and the injected false data to the k^{th} meter at time t is shown by $b_{K,t}$.

- (b) The measurement model for a jamming attack with additive noise is as follows:

$$y_t = \mathbf{H}\mathbf{x}_t + \mathbf{w}_t + \mathbf{u}_t \|\{t \geq \tau\}. \quad (5)$$

Here, the random noise realization at time $t \geq \tau$ is shown by $\mathbf{u}_t \triangleq [u_{1,t}, \dots, u_{K,t}]^T$ and the jamming noise corrupting the k^{th} meter at time t is represented by $u_{K,t}$.

- (c) Under an FDIA/jamming hybrid attack [5], the meter measurement appear as follows:

$$y_t = \mathbf{H}\mathbf{x}_t + \mathbf{w}_t + (\mathbf{b}_t + \mathbf{u}_t) \|\{t \geq \tau\}. \quad (6)$$

- (d) When the system controller is under DOS attack, meter measurements cannot partially be available. Therefore, the measurement model is formulated accordingly:

$$y_t = \mathbf{D}_t (\mathbf{H}\mathbf{x}_t + \mathbf{w}_t). \quad (7)$$

Here, a diagonal matrix including 0s and 1s is shown by $\mathbf{D}_t = \text{diag}(d_{1,t}, \dots, d_{K,t})$. In particular, when $y_{k,t}$ exists, afterward, $d_{K,t} = 1$, or else $d_{K,t} = 0$. It should be noted that $\mathbf{D}_t = \mathbf{I}_t$ for $t < \tau$,

- (e) During a system attack, the matrix of measurement alters. \mathbf{H}_t represents the matrix of measurement subjected to topology attacks at time $t \geq \tau$, therefore:

$$y_t = \begin{cases} \mathbf{H}\mathbf{x}_t + \mathbf{w}_t, & \text{if } t < \tau, \\ \overline{\mathbf{H}}\mathbf{x}_t + \mathbf{w}_t, & \text{if } t \geq \tau. \end{cases} \quad (8)$$

- (f) In the case of a blended topology and FDIA/jamming hybrid attack, the measurement layout is:

$$y_t = \begin{cases} \mathbf{H}\mathbf{x}_t + \mathbf{w}_t, & \text{if } t < \tau, \\ \overline{\mathbf{H}}\mathbf{x}_t + \mathbf{w}_t + \mathbf{b}_t + \mathbf{u}_t, & \text{if } t \geq \tau. \end{cases} \quad (9)$$

2.2. SE. As SG regulation relies on the SE system, SE has traditionally been done utilizing static least squares (LS) estimators [3]. As a result of the time-varying load and energy generation in SGs, they are actually very dynamic systems [20]. Additionally, adversaries can design and perform time-varying cyber-attacks. Therefore, dynamic system modeling like in equations (1) and (2) as well as the use of dynamic state estimators could be really beneficial in the development of real-time SG operations and security [4, 5]. when the noise terms are Gaussian in a discrete-time linear dynamic system, the KF can be the best linear forecaster to minimize the average squared SE error [21]. $\hat{\mathbf{x}}_{t|t'}$ represents the state estimates at time t in which $t' = t - 1$ is for the prediction step and $t' = t$ is for measurement update stage, the KF equations at time t is:

Prediction:

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1} &= \mathbf{A}\hat{\mathbf{x}}_{t-1|t-1}, \\ \mathbf{F}_{t|t-1} &= \mathbf{A}\mathbf{F}_{t-1|t-1}\mathbf{A}^T + \sigma_v^2\mathbf{I}_N. \end{aligned} \quad (10)$$

Measurement update:

$$\begin{aligned} \mathbf{G}_t &= \mathbf{F}_{t|t-1}\mathbf{H}^T(\mathbf{H}\mathbf{F}_{t|t-1}\mathbf{H}^T + \sigma_w^2\mathbf{I}_K)^{-1}, \\ \hat{\mathbf{x}}_{t|t} &= \hat{\mathbf{x}}_{t|t-1} + \mathbf{G}_t(y_t - \mathbf{H}\hat{\mathbf{x}}_{t|t-1}), \\ \mathbf{F}_{t|t} &= \mathbf{F}_{t|t-1} - \mathbf{G}_t\mathbf{H}\mathbf{F}_{t|t-1}. \end{aligned} \quad (11)$$

Here, $\mathbf{F}_{t|t-1}$ and $\mathbf{F}_{t|t}$ indicated the approximates of the state covariance matrix according to the measurements up to $t - 1$ and t , respectively. In addition, the Kalman gain matrix at time t is shown by \mathbf{G}_t .

Afterward, an illustrative example is used to illustrate the impact of cyber-attack on the SE method. In the IEEE-14 bus power system with $N = 13$, $K = 23$, and system parameters were selected as $A = I_N$, $\sigma_v^2 = 10^{-4}$, and $\sigma_w^2 = 2 \times 10^{-4}$ is tested with FDI attacks of various magnitudes/intensities, and the average squared SE error of the KF is analyzed. At time $\tau = 100$, attacks will be launched, which means that the system will be operated under normal conditions until time 100 and then under attack thereafter. There are three levels of attack magnitude:

Level1:

$$b_{K,t} \sim u[-0.04, 0.04], \forall k \in \{1, \dots, K\} \text{ and } \forall t \geq \tau.$$

Level2:

$$b_{K,t} \sim u[-0.07, 0.07], \forall k \in \{1, \dots, K\} \text{ and } \forall t \geq \tau.$$

$$\text{Level3: } b_{K,t} \sim u[-0.1, 0.1], \forall k \in \{1, \dots, K\} \text{ and } \forall t \geq \tau.$$

Here, a uniform random variable between $[\zeta_1, \zeta_2]$ is shown by $U[\zeta_1, \zeta_2]$. When cyberattacks occur, the state estimates deviate from the real system states, and the deviation is enhanced by the attack magnitude.

3. Problem Formulation

The following is a description of the POMDP setting prior to introducing the problem formulation. When an agent and an environment are present, the seven-tuple $(S, A, T, R, O, G, \gamma)$

is used to define a discrete-time POMDP in which the group of latent conditions of the environment is shown by S , the group of agent's function is represented by A , the group of contingent transfer probabilities among the conditions is shown by T , $R: S \times A \rightarrow R$ shows the reward function mapping the condition-function pairs to rewards, the group of agent's observations is shown by O , the set of conditional observation probabilities is indicated by G , and a discount factor is shown by $\gamma \in [0, 1]$ indicating how many current rewards have been preferred than subsequent rewards.

At every time t , the zone is in a certain latent condition $s_t \in S$. An observation $o_t \in O$ is obtained according to the present zone condition with the probability $G(o_t|s_t)$, the agent can take an action $a_t \in A$ and receive a reward $r_t = R(s_t, a_t)$ from the zone according to the function and the present condition of the area. In parallel, the zone can make the transmission to the subsequent condition s_{t+1} with the probability $T(s_{t+1}|s_t, a_t)$. Repetition of the procedure has been required till the final condition has been achieved. In the method, the factor aims at determining the best policy $\pi: O \rightarrow A$, which can map observations to functions and maximize the anticipated factor overall discounted rewards, that is, $E[\sum_{t=0}^{\infty} \gamma^t r_t]$. The objective would be to reduce the expected overall discounted cost for an agent that gets costs rather than rewards from the environment. If the latter is taken into account, the POMDP problem is:

$$\min_{\pi: O \rightarrow A} E \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]. \quad (12)$$

Afterward, a POMDP setting is used to define the online CAD issue. The assumption is that at the unspecified time τ , the cyber-attacks have been started against the network, and it aims at detecting the attack soon once it has occurred, without knowing the attacker's capabilities or strategies. Here is the definition of the quickest change detection problem, which aims at minimizing the average detection delay and also the FA rate (FAR). It is possible to express the problem as a POMDP problem (according to Figure 1). There are two hidden states because of the unspecified launch time of the attack τ : *post-attack* & *pre-attack*. Every time $[t]$, the agent (defender) has two options following receiving the measurement vector \mathbf{y}_t : *stop* and express the attack or *go ahead* to make more measurements. When the action *stop* has been selected, the system can move into a *terminal* state and stay there permanently.

In order to reduce both FAs and detection delays, both FA and diagnosing delay occurrences must be accompanied by several costs. $c > 0$ is the relevant cost of the diagnosing delay in comparison with a FA. As a result, when the true basic condition is *pre-attack* and the action *stop* has been selected, there is a FA, and the defender can receive a cost of 1. However, when the underlying state is *post-attack* and the action *continue* has been selected, so the defender can receive a cost of c because of the detection delay. The remaining (hidden) state-action pairs are supposed to have zero costs. Furthermore, if the action *stop* has been selected, the defender does not achieve any more costs as long as staying in the *final* status. The defender aims at minimizing its expected

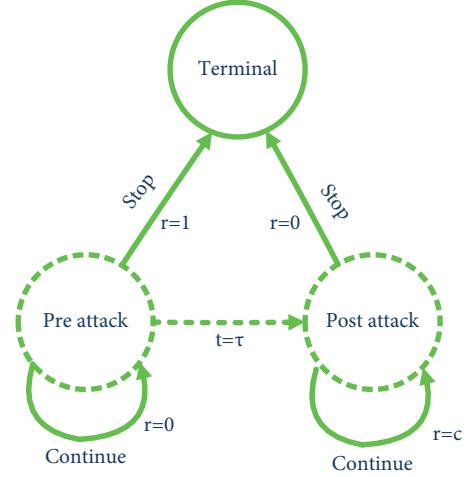


FIGURE 1: Diagram of state-machine to investigate POMDP adjustment.

overall cost by carefully selecting the functions. In particular, the defender must define the stopping time when an attack has been declared according to its observations.

The stopping time selected via the defender is shown by Γ . In addition, the probability measure is shown by P_k when the attack has been launched at time k , that is $\tau = k$, and the related expectation is shown by E_k . It should be noted that as the attacking strategies are unknown, P_k has been supposed to be unknown. The expected overall discounted cost is calculated for the proposed online CAD issue in the following way:

$$\begin{aligned} E \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] &= E_{\tau} \left[\mathbb{1}\{t \geq \tau\} + \sum_{t=\tau}^{\Gamma} c \right] \\ &= E_{\tau} [\mathbb{1}\{t \geq \tau\} + c(\Gamma - \tau)^+] \\ &= P_{\tau} (\{\Gamma < \tau\}) + cE_{\tau} [(\Gamma - \tau)^+]. \end{aligned} \quad (13)$$

Here, $\gamma = 1$ has been selected as the current and subsequent costs have been weighted equally in the subject, $\{\Gamma < \tau\}$ shows the FA occurrence, which has been penalized with the cost of zero, and $E_{\tau} [(\Gamma - \tau)^+]$ shows the mean diagnosing lag in which every detection lag has been penalized with the cost of c and $(\cdot)^+ = \max(\cdot, 0)$.

According to equations (12) and (13), the online attack detection problem is:

$$\min_{\Gamma} P_{\tau} (\{\Gamma < \tau\}) + cE_{\tau} [(\Gamma - \tau)^+]. \quad (14)$$

As c represents the relevant cost among the FA and the detection lag occurrences, the transaction curve among mean detection lag and FAR is determined via changing c and solving the related problem in equation (14). Furthermore, $c < 1$ is selected for avoiding frequent FAs.

The MF-RL method obtains a solution to equation (14) because the actual POMDP layout is uncertain because of the uncertain attack start time τ and attack strategy, and the RL algorithms have been proved to perform well under uncertain conditions. There is therefore a necessity to learn a direct mapping from observations to functions, that is, the time of stopping $[\Gamma]$.

Moreover, generally, similar observations can be obtained in both *pre-attack* & *post-attack* statuses. It is known as conceptual harmony and avoids well inferences about the underlying status from being made via just watching the observation one time. In addition, it should be noted that in the problem the decision to attack is only according to a single observation, which is equivalent to an outlier detection layout with better detectors that require no learning, refer to [13, 14]. The purpose of this research is to detect sudden and persistent attacks or anomalies caused by a hostile intervention in the system, instead of random disturbances caused by high-level system noise.

The measurements $\{y_t\}$ have been collected via intelligent meters and analysis to gain $o_t = f(\{y_t\})$. The defender observes $f(\{y_t\})$ at every time t and has decided on the CAD statement time $[\Gamma]$.

$f(\cdot)$ shows the function, which can process a measurement's limit history and produce the observation signal, therefore, $o_t = f(\{y_t\})$ shows the observation signal at time t . Afterward, every time, the defender can observe $f(\{y_t\})$ and decide on the stopping time Γ , according to Figure 2. The defender aims at solving equation (14) via applying an RL algorithm. The following part describes this in more detail.

4. Solution Method

First, the methodology is explained for obtaining the observation signal $o_t = f(\{y_t\})$. The state estimates derived from the KF and the baseline measurement model in equation (2) are used to infer the meter measurements PDF in the *pre-attack* condition. Particularly, it is possible to estimate the measurements PDF within usual operating statuses according to the following:

$$y_t \sim \mathcal{N}(H\hat{x}_{t|t}, \sigma_w^2 I_K), \quad (15)$$

$L(y_t)$ is the likelihood of the measurement according to the estimation of base density:

$$\begin{aligned} L(y_t) &= (2\pi\sigma_w^2)^{-K/2} \exp\left(\frac{-1}{2\sigma_w^2}(y_t - H\hat{x}_{t|t})^T (y_t - H\hat{x}_{t|t})\right) \\ &= (2\pi\sigma_w^2)^{-K/2} \exp\left(\frac{-1}{2\sigma_w^2}\eta_t\right), \end{aligned} \quad (16)$$

where

$$\eta_t \triangleq (y_t - H\hat{x}_{t|t})^T (y_t - H\hat{x}_{t|t}). \quad (17)$$

Within normal operating situations, it has been anticipated that $L(y_t)$ will be high. If η_t is small (near zero), the system is operating normally. The likelihood $L(y_t)$, however, is anticipated to drop in the cases where the systems deviate from normal operating conditions as a result of an attack or anomaly. When high η_t values persist over time, there may be an attack or anomaly present. As such, η_t might contribute to reducing the uncertainty of the fundamental status in some cases.



FIGURE 2: An explanation of the online CAD issue in the SG.

Due to the fact that η_t could have any positive amount, the observation area has been continued, making the mapping from every observation to function mathematically impossible. It is possible to decrease the computing burden for these continuous spaces by quantizing the observations. After partitioning the observation area into I disjoint and exclusive reciprocal distances utilizing $\beta_0 = 0 < \beta_1 < \dots < \beta_{I-1} < \beta_I = \infty$ quantization thresholds, the observation at time t will be described as θ_i if $\beta_{i-1} \leq \eta_t < \beta_i$, $i \in 1, \dots, I$ is met. Next, $\theta_1, \dots, \theta_I$ indicate possible observations for any particular moment. θ_i 's represent the quantization levels; therefore, every θ_i has to have a diverse value.

Moreover, as discussed previously, even though η_t can be used for inferring the underlying state at time T , similar observations can be obtained in the *pre-attack* & *post-attack* statuses. Therefore, a finite history of observations is proposed. M is the sliding observation window (SOW) size, therefore, there are I^M feasible observation windows that exist and the sliding window at time $[t]$ includes the quantized versions of $\{\eta_j; t - M + 1 \leq j \leq t\}$. An observation o is, therefore, a window, meaning that an observation space O includes all possible windows. As an example, when $I = M = 2$, afterward, $O = \{[\theta_1, \theta_1], [\theta_1, \theta_2], [\theta_2, \theta_1], [\theta_2, \theta_2]\}$.

RL algorithm is used for learning a $Q(o, a)$ value, that is, the expected future cost for every observation-action pair (o, a) , in which all $Q(o, a)$ values have been saved in the Q -table of size $I^M \times 2$. Following the Q -table's learning, the defender's policy is to choose the function a with the minimal $Q(o, a)$ for every observation o . Generally, as I and M increase the learning efficiency enhances and simultaneously causes in a bigger Q table requiring to enhance in the training episodes number and therefore the calculation burden of the learning step. Therefore, I and M must be selected regarding the anticipated exchange among efficiency and calculation burden.

The learning step and online CAD step are included in the suggested RL-based detection method. SARSA, which is a MF-RL control layout [22], performed better than the model-free POMDP settings [12]. The SARSA algorithm is used in order to train the defender on numerous episodes of experience, and the defender learns a Q -table during the learning phase. According to Figure 3, the simulation environment has been produced for training during which the defender has taken an action according to its observations and received a cost from the simulation in return. On the basis of this experience, a Q -table is updated and learned by the defender. Afterward, according to the observations, in the online CAD stage, the previously learned Q -table is used to choose the action with the minimum anticipated future cost (Q amount) every time. Once the defender selects the action *stop*, the online detection phase ends. An attack has been declared when the *stop* has been selected, and the procedure has been stopped.

In the event of an attack declaration, the online detection phase may be restarted any time the system has recovered and is back to normal operating conditions. After a defender has been trained, additional training is not required.

Every iteration of RL (learning episode) involves repeating the same actions. An RL algorithm's time complexity would then be regarded as a single iteration's time complexity [23]. SARSA updates the Q -table one at a time, and the maximum learning episode time is T , so the time complexity is $O(T)$. Furthermore, $O(TE)$ shows the total complexity of the learning process, since E indicates the number of learning episodes. It should be noted that the space of action and observation does not affect the time complexity. Increasing I or/and M , in contrast, requires learning a more complex Q -table, for which one needs to enhance E . Additionally, the space complexity (memory cost) is $M + 2I^M$ since the SOW is M and the Q -table is $I^M \times 2$. It should be noted that space complexity remains constant through time. With an SG model and several attack models, the measurement data is obtained online throughout the learning process and the defender has been trained using the observed data streams. Due to this, storing enormous amounts of training data for the learning phase is not necessary since the size of SOW (M) has been saved at every stop.

A distributed SG system is implemented using the suggested solution layout, in which learning and CAD tasks have been handled at a single center while meter measurements have been collected on a distributed basis. This setup is shortly described below.

- (i) SGs have multiple local control centers as well as a global control center in the large-scale monitoring model. Local centers collect and process measurements from smart meters in their neighborhoods, and they communicate with global centers as well as neighboring local centers.
- (ii) A distributed KF, such as the one developed for large-scale SGs in [4] is used to estimate the system state.
- (iii) In the measurement matrix, $h_k^T \in R^N$ is the k th row, that is, $H^T = [h_1, \dots, h_k]$. A negative log-scaled likelihood estimate, η_t , is given by the following (refer to equation (17)):

$$\eta_t = \sum_{k=1}^K \left(y_{k,t} - h_k^T \hat{x}_{t|t} \right)^2. \quad (18)$$

The local centers have the capability of estimating the system state via utilizing the distributed KF for every time t . Afterward, the local centers could calculate the term $(y_{k,t} - h_k^T \hat{x}_{t|t})^2$ for their neighborhood meters. R shows the local centers number and S_r denotes the group of meters in the neighbors of the r^{th} local center. Therefore, η_t in equation (18) is:

$$\eta_t = \sum_{r=1}^R \sum_{k \in S_r} \left(y_{k,t} - h_k^T \hat{x}_{t|t} \right)^2 = \sum_{r=1}^R \eta_{t,r}. \quad (19)$$

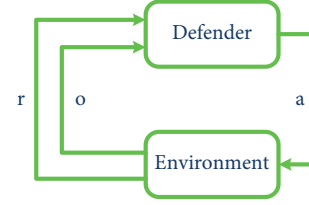


FIGURE 3: Interaction between the environment and defender within the learning procedure.

- (i) A distributed implementation allows every local center to calculate $\eta_{t,r}$ and send it to the global center for summing $\{\eta_{t,r}, r = 1, 2, \dots, R\}$ and calculating η_t ,
- (ii) Learning and detection tasks have been carried out at the global center on the same basis as previously described.

5. Simulation Outcomes

5.1. Simulation Setup and Parameters. The IEEE-14 bus electrical network with $[N + 1 = 14]$ buses and $[K = 23]$ intelligent meters is used to perform the simulation. In MATPOWER [24], the DC optimal power flow algorithm is used to determine the initial state variables (phase angles). System matrix A has been selected as the measurement and identification matrixes H based on the IEEE-14 electrical grid. $\sigma_v^2 = 10^{-4}$ and $\sigma_w^2 = 2 \times 10^{-4}$ have been selected as the noise variances for the usual operation of the system. As part of the suggested online CAD layout on the basis of RL, $I = 4$ quantization levels are selected and thresholds $\beta_1 = 0.95 \times 10^{-2}$, $\beta_2 = 1.05 \times 10^{-2}$, and $\beta_3 = 1.15 \times 10^{-2}$ are selected using an offline simulation based on monitoring $\{\eta_t\}$ throughout normal operation. The observation window includes 4 entries, thus $M = 4$. Additionally, $\alpha = 0.1$ and $\epsilon = 0.1$ have been selected as the learning parameters, and $T = 200$ has been selected as the episode length. During the learning stage, the defender has been first trained more than 4×10^5 episodes with the attack start time off $\tau = 100$ and next, more than 4×10^5 episodes with $\tau = 1$ for ensuring that the defender can properly explore the observation space within usual operating situations and also during an attack. As a learning episode ends when the action *stop* has been selected and observations are available to the defender just for $\geq \tau$, $\tau = 1$ has been selected during the half of the learning episodes to ensure that the defender has been adequately trained within the post-attack regime.

The suggested algorithm has been trained for both $c = 0.02$ and $c = 0.2$, for illustrating the trade-off between mean CAD lag and FA probability. It is necessary to train defenders with very low-magnitude attacks associated with small deviations from the baseline in order to achieve a detector, which can be robust and useful versus tiny deviations from the usual exploitation of the system. Several known low-magnitude attack kinds have been applied in this case. One-half of the learning episodes use random FDIAs with attack extents equal to uniform random realization parameter $\pm U[0.02, 0.06]$, i.e., $b_{k,t} \sim U[0.02, 0.06]$ is the injected false data to the k^{th} meter at time $t \geq \tau$,

$\forall k \in \{1, \dots, K\}$. The other ones use random hybrid FDI/jamming attacks with $b_{k,t} \sim U[0.02, 0.06]$, $u_{k,t} \sim N(0, \sigma_{k,t})$, and $\sigma_{k,t} \sim U[2 \times 10^{-4}, 4 \times 10^{-4}]$, $\forall k \in \{1, \dots, K\}$ and $\forall t \geq \tau$. The overall training time costs have been computed about as [5018sec] and [5106sec] for $c = 0.2$ and $c = 0.02$, respectively.

5.2. Efficiency Assessment. This part evaluates the efficiency of the suggested CAD method on the basis of RL and compares it with several current detector methods [25]. First, $E_\infty[_]$ is reported as the mean FA cycle of the suggested CAD method, that is, the 1^{th} time on the mean the suggested detector has given an alarm, however, no anomaly/attack occurs at a whole ($\tau = \infty$). The mean FA period for $c = 0.2$ is about $E_\infty[\Gamma] = 9.4696 \times 10^5$ and it is about $E_\infty[\Gamma] = 7.921 \times 10^6$ for $c = 0.02$. It is anticipated that the FAR of the suggested detector decrease by increasing the relevant cost of the FA occurrence, $1/c$.

According to the optimization problem in equation (14), the efficiency factors include the probability of FA, that is, $P_\tau(\{\Gamma < \tau\})$, and the average detection delay, that is, $E_\tau[(\Gamma - \tau)^+]$. It should be noted that the unknown attack launch time τ affects both efficiency factors. Therefore, generally, the efficiency factors must be computed for every possible τ . To illustrate efficiency, τ as the numeral random parameter is selected with variable ρ so, $P(\tau = k) = \rho(1 - \rho)^{k-1}$, $k = 1, 2, 3, \dots$ in which $\rho \sim U[10^{-4}, 10^{-3}]$ shows a uniform random variable.

Monte Carlo simulations over 10000 trials are used to calculate the average detection delay and the probability of FA of the suggested detector, the Euclidean detector [13], and the cosine-similarity factor on the basis of the detector [14]. The thresholds of the benchmark tests are changed as well as c for the suggested algorithm is changed in order to determine the efficiency curves. $c = 0.02$ and $c = 0.2$. I are used for evaluating the suggested algorithm [26]. In addition, the F-score, recall, and precision for whole simulation scenarios are reported. The bound as ten-time units is selected. Afterward, the F-score, recall, and precision out of 1×10^4 tests are calculated in the following way:

$$\begin{aligned} \text{Precision} &= \frac{\#\text{trials}(\tau \leq \Gamma \leq \tau + 10)}{\#\text{trials}(\tau \leq \Gamma \leq \tau + 10) + \#\text{trials}(\Gamma < \tau)} \\ \text{Recall} &= \frac{\#\text{trials}(\tau \leq \Gamma \leq \tau + 10)}{\#\text{trials}(\tau \leq \Gamma \leq \tau + 10) + \#\text{trials}(\Gamma < \tau + 10)} \\ \text{F-score} &= 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \quad (20)$$

Here, “# trials” shows “the number of tests with.” The suggested and the benchmark detectors are evaluated within the below attack case studies:

- (1) First, the detectors versus the random FDIA are evaluated in which $b_{k,t} \sim U[-0.07, 0.07]$, $\forall k \in \{1, \dots, K\}$ and $\forall t \geq \tau$. Figure 4 shows the related tradeoff curves.

- (2) Second, the detectors versus a structured FDI attack are evaluated [3], in which the injected data b_t is located on the column space of the measurement matrix H . $b_t = Hg_t$ is selected in which $g_t \triangleq [g_{1,t}, \dots, g_{N,t}]^T$ and $g_{n,t} \sim U[0.08, 0.12]$, $\forall n \in \{1, \dots, N\}$ and $\forall t \geq \tau$. Figure 5 shows the related efficiency curves.
- (3) Afterward, the detectors are evaluated when the jamming attack occurs with zero-mean AWGN in which $u_{k,t} \sim N(0, \sigma_{k,t})$ and $\sigma_{k,t} \sim u(10^{-3}, 2e - 3)$, $\forall k \in \{1, \dots, K\}$ and $\forall t \geq \tau$. Figure 6 shows the related tradeoff curves.
- (4) The detectors are evaluated when a jamming attack occurs with jamming noise related over the meters in which $u_t \sim N(0, U_t)$, $U_t = \sum_t \Sigma_t^T$, and Σ_t shows a random Gaussian matrix with its entry at the i^{th} row and the j^{th} column can be $\sum_{t,i,j} \sim N(0, 8 \times 10^{-5})$. Figure 7 shows the related efficiency curves.
- (4) In addition, the detectors are evaluated in the case of the hybrid FDIA or jamming attack in which $b_{k,t} \sim U[-5, 5] \times 10^{-2}$, $u_{k,t} \sim N(0, \sigma_{k,t})$, and $\sigma_{k,t} \sim U[5 \times 10^{-4}, 10^{-3}]$, $\forall k \in \{1, \dots, K\}$ and $\forall t \geq \tau$. Figure 8 shows the related tradeoff curves.
- (5) Next, the detectors are evaluated when a random DoS attack occurs in which the measurement of every smart meter is not available for the controller at every time with probability of 0.2. It means that for every meter k , $d_{k,t}$ can be zero with probability $2e - 1$ and one with probability $8e - 1$ at every time $t \geq \tau$. Figure 9 shows the efficiency curves versus the DoS attack.
- (6) In addition, a network topology attack is considered in which the lines among the buses (9, 10) and (12, 13) break down. So, the measurement matrix, H_t for $t \geq \tau$ has been obtained. Figure 10 shows the related tradeoff curves.
- (7) Finally, a combined technique and hybrid FDIA or jamming attack are considered, in which the lines among buses 9–10 and 12–13 break down for $t \geq \tau$ and so, $b_{k,t} \sim U[-0.05, 0.05]$, $u_{k,t} \sim N(0, \sigma_{k,t})$, and $\sigma_{k,t} \sim U[5 \times 10^{-4}, 10^{-3}]$, $\forall k \in \{1, \dots, K\}$ and $\forall t \geq \tau$. Figure 11 shows the related efficiency curves.

The F-score, recall, and precision for the suggested detector on the basis of RL for $c = 2e - 1$ and $c = 2e - 2$ are summarized in Table 1 and 2, respectively versus whole the proposed simulation case studies earlier. In addition, in the case of the random FDI attack, the precision against recall curves for the suggested and benchmark detectors is illustrated Figure 12. Because meter measurements are not partially available in DoS attacks, therefore, the system significantly strays from the usual operation, whole detectors are capable of detecting DoS attacks with nearly zero mean detection lags (refer to Figure 9).

Eventually, the impact of the window size, M , is evaluated on the efficiency of the detector on the basis of RL (trained for $c = 2e - 1$) versus random FDIAs with changing

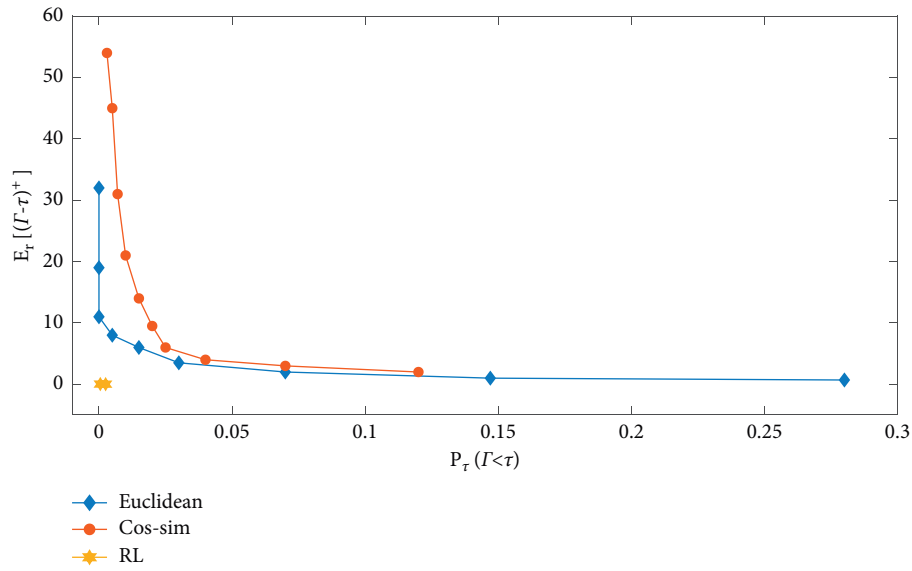


FIGURE 4: Mean CAD lag versus probability of FA curves for the suggested method and the benchmark trails in case of the random FDIA.

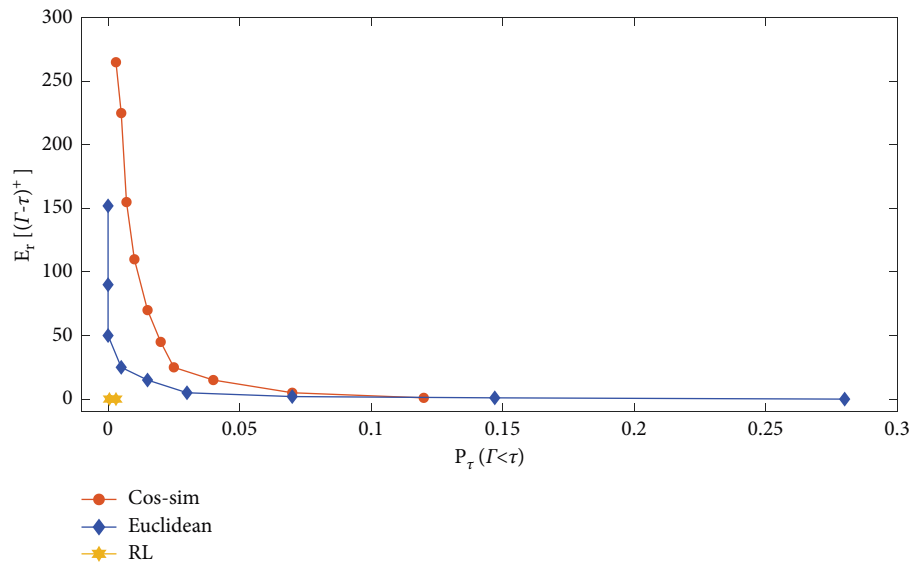


FIGURE 5: Proficiency curves for the suggested method and the benchmark trails in the case of the structured FDIA.

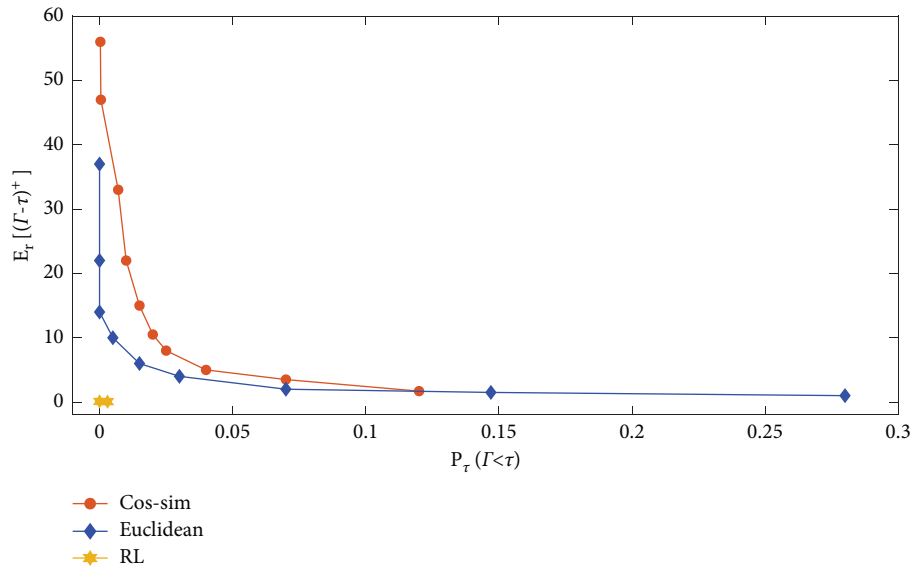


FIGURE 6: Proficiency curves for the suggested method and the benchmark trails in case of the jamming attack with AWGN.

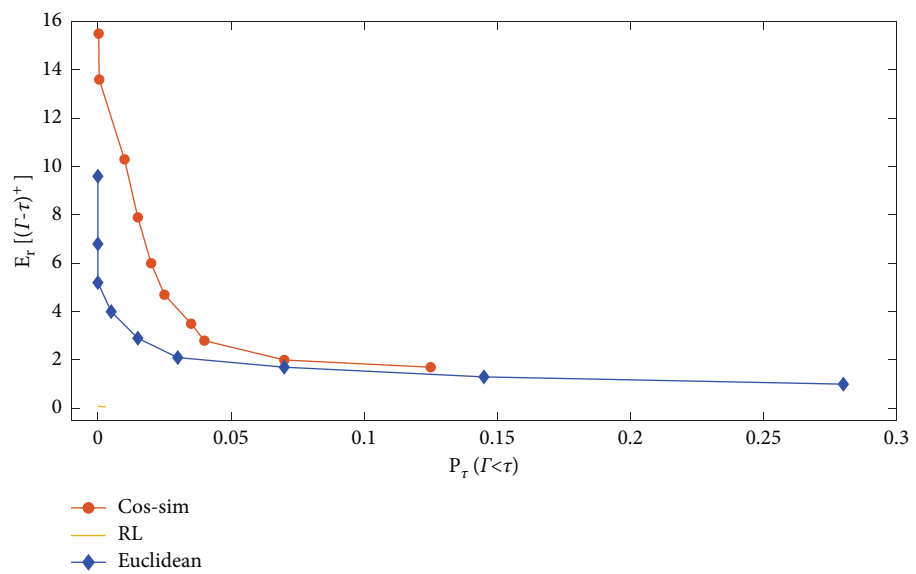


FIGURE 7: Proficiency curves for the suggested method and the benchmark trails in case of a jamming attack with jamming noise associated with the area.

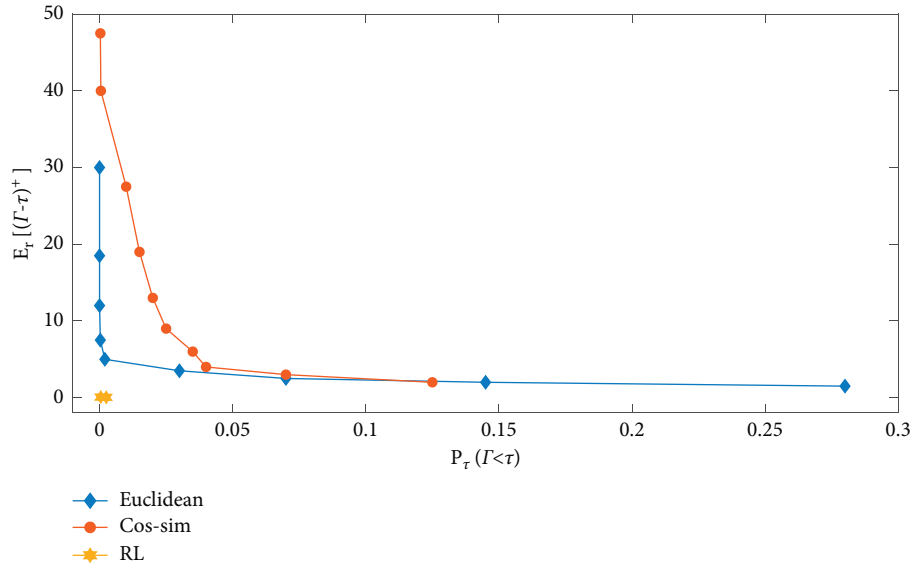


FIGURE 8: Proficiency curves for the suggested method and the benchmark trails in case of a hybrid FDI/jamming attack.

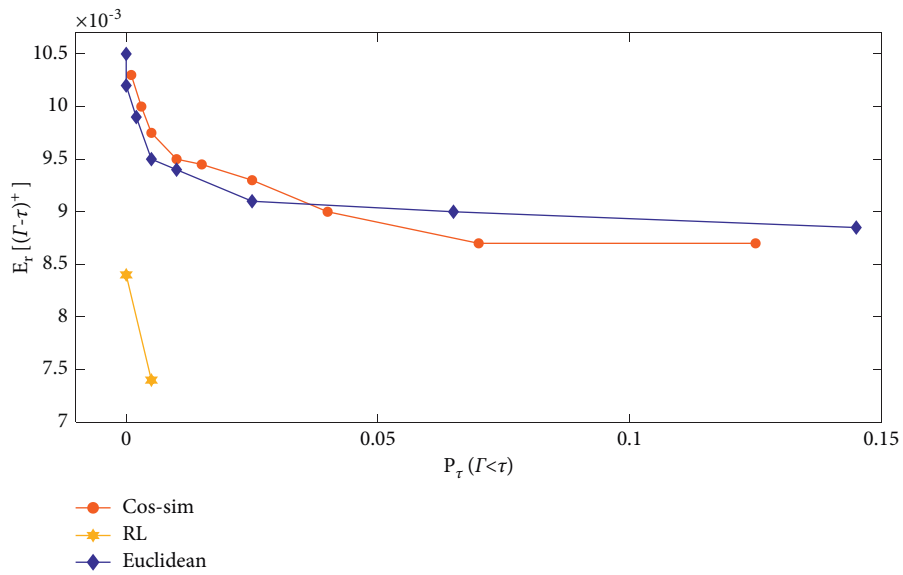


FIGURE 9: Efficiency curves for the suggested method and the benchmark trails when the DoS attack occurs.

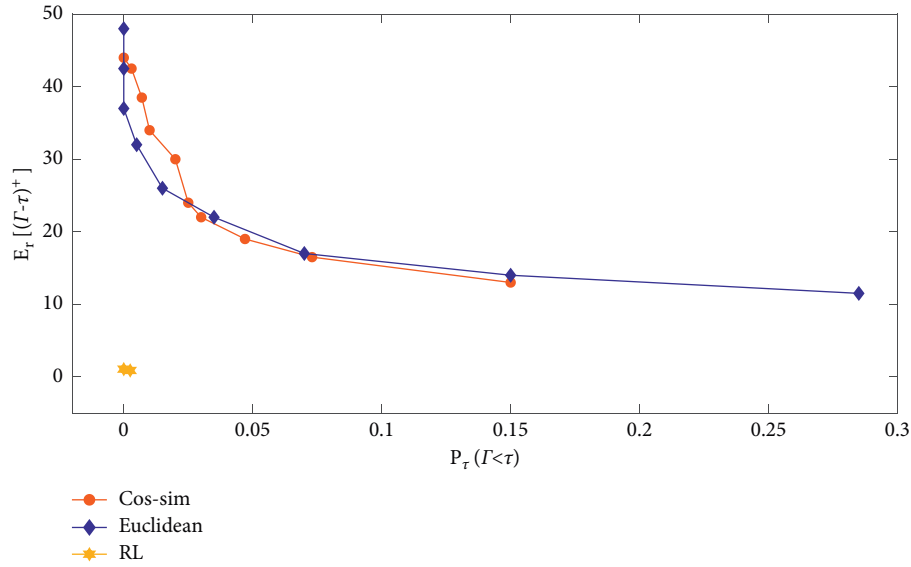


FIGURE 10: Efficiency curves for the suggested method and the benchmark trails under the network topology CAD.

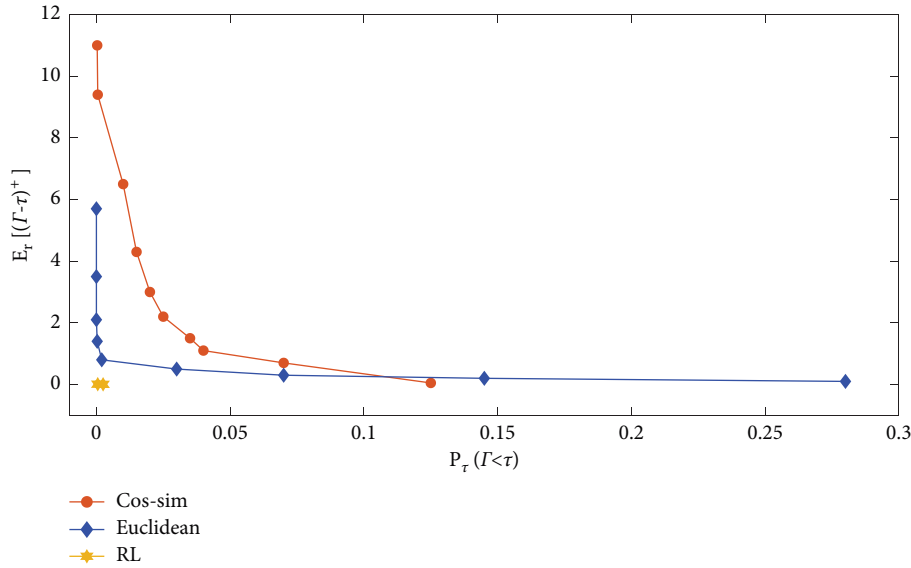


FIGURE 11: Efficiency curves for the suggested method and the benchmark trails under a mixed system topology and hybrid FDIA or jamming attack.

TABLE 1: F -score, recall, and precision for the suggested detection method ($c = 0.2$) in different kinds of cyber-attacks.

Measure	F -score	Recall	Precision
Structured FDI	0.9860	0.9755	0.9967
Corr. Jamm	0.9983	1	0.9967
DOS	0.9987	1	
Jamming	0.9986	1	0.9974
Hybrid	0.9985	1	0.9972
FDI	0.9987	1	0.9976
Topology	0.9889	0.9807	0.9971
Mixes	0.9985	1	0.9972

TABLE 2: F -score, recall, and precision for the suggested detection method ($c = 0.02$) in different kinds of cyber-attacks.

Measure	F -score	Recall	Precision
Structured FDI	0.9712	0.9448	0.9992
Corr. Jamm	0.9998	1	0.9997
DOS	0.9996	1	0.9994
Jamming	0.9996	1	0.9993
Hybrid	0.9997	1	0.9996
FDI	0.9998	1	0.9997
Topology	0.9890	0.9784	0.9998
Mixes	0.9996	1	0.9994

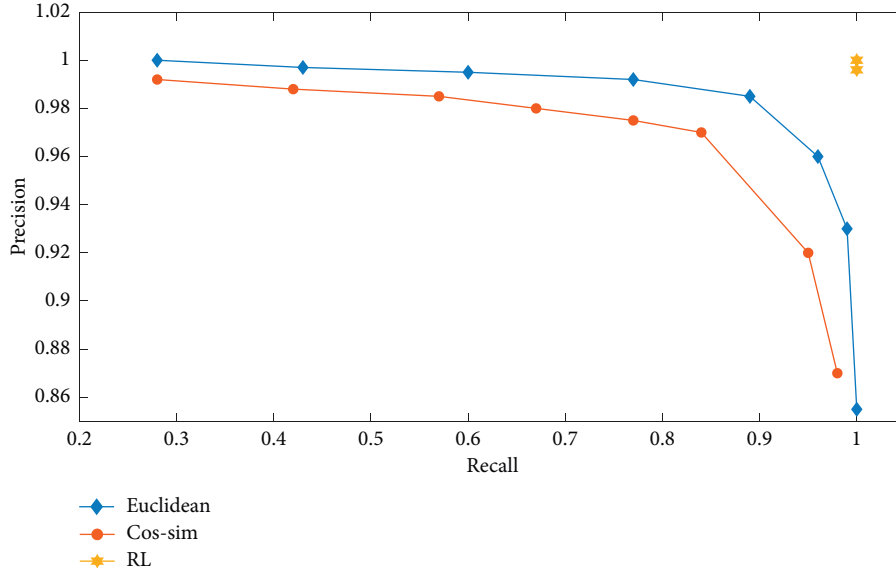


FIGURE 12: Precision, recall for the suggested and the benchmark CAD methods versus the random FDIA.

TABLE 3: The window size effect (M) on the efficiency of the detector on the basis of RL for $c = 2e - 1$ within random FDIA with diverse amount stages every entrance in the table displays the outcomes according to the following expression: $P_{\tau}(\{\Gamma < \tau\})/E_{\tau}[(\Gamma - \tau)^+]/F\text{-score}/\text{recall}/\text{precision}$.

Measure	M			
	1	2	4	6
$\varphi = 0.03$	0.0187/8.4208/0.7119/0.8226	0.0021/15.9532/0.9957/0.4872/ 0.6543	0.0021/14.8548/0.9959/0.5077/ 0.6725	0.0021/14.2506/0.996/0.6841
$\varphi = 0.04$	0.0187/1.2219/0.9813/0.9995/ 0.9903	0.0021/1.9046/0.9979/0.9923/ 0.9951	0.0021/1.8437/0.9979/0.9943/ 0.9961	0.0021/1.8207/0.9979/0.9955/ 0.9967
$\varphi = 0.05$	0.0187/0.2606/0.9813/1/ 0.9906	0.0021/0.4049/0.9979/1/0.9989	0.0021/0.4016/0.9979/1/0.9989	0.0021/0.4016/0.9979/1/ 0.9989

extends. Table 3 shows the outcomes for $M = 1$, $M = 2$, $M = 4$, and $M = 6$ in which $b_{k,t} \sim U[-\varphi, \varphi]$, $\forall k \in \{1, \dots, K\}$, $\forall t \geq \tau$ and φ has the amounts of $[3, 4, \text{ and } 5] \times 10^{-2}$.

6. Conclusion

The present study formulates an online CAD structure as the POMDP subject and proposes a solution on the basis of MF-RL for POMDPs. In the numerical tests, the suggested detection layout proves to be efficient, reliable, and quick in CADs that target the SG. In addition, RL algorithms have been shown to have a strong potential for solving difficult cyber-security problems. It is possible to greatly improve the algorithm suggested in this study by utilizing additional enhanced techniques. This study is concluded by considering a single-agent RL setting to optimize the defender's policy, such that the attacking methods, like the attack kinds, magnitudes, set of attack meters, and so on, do not affect the defender's optimal policy. The optimal policy for the defender after launching an attack is to *stop* and declare an attack.

Data Availability

All data are available in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Science and technology project support of the China Southern Power Grid Corporation (No. GDKJXM20210170 (036100KK52210070)).

References

- [1] O. Lukačević, A. Almalaq, K. Alqunun et al., "Optimal CONOPT solver-based coordination of bi-directional converters and energy storage systems for regulation of active and reactive power injection in modern power networks," *Ain Shams Engineering Journal*, vol. 13, no. 6, Article ID 101803, 2022.
- [2] Y. Li, B. Wang, H. Wang et al., "Importance assessment of communication equipment in cyber-physical coupled distribution network based on dynamic node failure mechanism," *Frontiers in Energy Research*, p. 654, 2022.
- [3] M. Dehghani, T. Niknam, M. Ghiasi, P. Siano, H. Haes Alhelou, and A. Al-Hinai, "Fourier singular values-based false data injection attack detection in AC smart-grids," *Applied Sciences*, vol. 11, no. 12, p. 5706, 2021.

- [4] J. Chen, M. A. Mohamed, U. Dampage et al., "A multi-layer security scheme for mitigating smart grid vulnerability against faults and cyber-attacks," *Applied Sciences*, vol. 11, no. 21, p. 9972, 2021.
- [5] K. Alnowibet, A. Annuk, U. Dampage, and M. A. Mohamed, "Effective energy management via false data detection scheme for the interconnected smart energy hub-microgrid system under stochastic framework," *Sustainability*, vol. 13, no. 21, Article ID 11836, 2021.
- [6] W. Xu, J. Li, M. Dehghani, and M. GhasemiGarpachi, "Blockchain-based secure energy policy and management of renewable-based smart microgrids," *Sustainable Cities and Society*, vol. 72, Article ID 103010, 2021.
- [7] G. Rovatsos, G. V. Moustakides, and V. V. Veeravalli, "Quickest detection of moving anomalies in sensor networks," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 762–773, 2021.
- [8] U. Bermejo, A. Almeida, A. Bilbao-Jayo, and G. Azkune, "Embedding-based real-time change point detection with application to activity segmentation in smart home time series data," *Expert Systems with Applications*, vol. 185, Article ID 115641, 2021.
- [9] B. Wu and Y. Feng, "Policy reuse for learning and planning in partially observable Markov decision processes," in *Proceedings of the 2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 549–552, IEEE, Beijing China, 2017 July.
- [10] T. P. Le, N. A. Vien, and T. Chung, "A deep hierarchical reinforcement learning algorithm in partially observable Markov decision processes," *IEEE Access*, vol. 6, pp. 49089–49102, 2018.
- [11] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson, "Deep variational reinforcement learning for POMDPs," in *International Conference on Machine Learning*, vol. 3, pp. 2117–2126, 2018.
- [12] J. Zhang, L. Tai, M. Liu, J. Boedecker, and W. Burgard, "Neural slam: learning to explore with external memory," 2017, <https://arxiv.org/abs/1706.09520>.
- [13] M. Khalaf, A. Youssef, and E. El-Saadany, "Detection of false data injection in automatic generation control systems using kalman filter," in *Proceedings of the 2017 IEEE Electrical Power and Energy Conference (EPEC)*, pp. 1–6, IEEE, Saskatoon, Canada, 2017 October.
- [14] X. Niu, J. Li, J. Sun, and K. Tomsovic, "Dynamic detection of false data injection attack in smart grid using deep learning," in *Proceedings of the 2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1–6, IEEE, Washington, DC, USA, 2019 February.
- [15] A. Almalaq, S. Albadran, and M. A. Mohamed, "Deep machine learning model-based cyber-attacks detection in smart power systems," *Mathematics*, vol. 10, p. 2574, 2022.
- [16] H. Jiang, Z. Wang, and H. He, "An evolutionary computation approach for smart grid cascading failure vulnerability analysis," in *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 332–338, IEEE, Xiamen, China, 2019 December.
- [17] C. Xu, S. Liu, C. Zhang, Y. Huang, Z. Lu, and L. Yang, "Multi-agent reinforcement learning based distributed transmission in collaborative cloud-edge systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1658–1672, 2021.
- [18] A. Marinescu, I. Dusparic, and S. Clarke, "Prediction-based multi-agent reinforcement learning in inherently non-stationary environments," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 12, no. 2, pp. 1–23, 2017.
- [19] J. Zhao, J. Qi, Z. Huang et al., "Power system dynamic state estimation: motivations, definitions, methodologies, and future work," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3188–3198, 2019.
- [20] M. Dehghani, M. Ghiasi, T. Niknam et al., "Cyber-attack detection based on wavelet singular entropy in AC smart islands: false data injection attack," *IEEE Access*, vol. 9, pp. 16488–16507, 2021.
- [21] M. J. Grimble, "Polynomial systems approach to optimal linear filtering and prediction," *International Journal of Control*, vol. 41, no. 6, pp. 1545–1564, 1985.
- [22] P. R. Montague, "Reinforcement learning: an introduction, by sutton, RS and Barto, AG," *Trends in Cognitive Sciences*, vol. 3, no. 9, p. 360, 1999.
- [23] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman, "Quantifying generalization in reinforcement learning," 2019, <https://arxiv.org/abs/1812.02341>.
- [24] S. Chen, Z. Wei, G. Sun, D. Wang, and H. Zang, "Steady state and transient simulation for electricity-gas integrated energy systems by using convex optimisation," *IET Generation, Transmission & Distribution*, vol. 12, no. 9, pp. 2199–2206, 2018.
- [25] A. Almalaq, S. Albadran, A. Alghadhban, T. Jin, and M. A. Mohamed, "An effective hybrid-energy framework for grid vulnerability alleviation under cyber-stealthy intrusions," *Mathematics*, vol. 10, p. 2510, 2022.
- [26] M. Calasan, A. F. Zobaa, H. M. Hasanien, S. H. Abdel Aleem, and Z. M. Ali, "Towards accurate calculation of supercapacitor electrical variables in constant power applications using new analytical closed-form expressions," *Journal of Energy Storage*, vol. 42, Article ID 102998, 2021.