Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multi-agent Reinforcement Learning

Abianeh, Ali Jafarian ; Wan, Yihao; Ferdowsi, Farzad; Mijatovic, Nenad; Dragicevic, Tomislav

Published in:
IEEE Transactions on Power Electronics

Link to article, DOI:
10.1109/TPEL.2021.3132028

Publication date:
2022

Document Version
Peer reviewed version

Vulnerability Identification and Remediation of FDI Attacks in Islanded DC Microgrids Using Multi-agent Reinforcement Learning

Ali Jafarian Abianeh, *Student Member, IEEE,* Yihao Wan, *Student Member, IEEE,* Farzad Ferdowsi, *Senior Member, IEEE,* Nenad Mijatovic, *Senior Member, IEEE* and Tomislav Dragičević, *Senior Member, IEEE*

***Abstract*—This paper proposes a novel approach to uncover deficiencies of the existing cyber-attack detection schemes and thereby to serve as a foundation for establishing more reliable cybersecure solutions, with particular application in DC microgrids. For this purpose, a multi-agent deep Reinforcement Learning (RL) based algorithm is proposed to automatically discover the vulnerable spots on the conventional index-based cyberattack detection schemes, and automatically generate coordinated stealthy destabilizing False Data Injection (FDI) attacks on cyber-protected islanded DC microgrids. To enable a continuous action space for the trained RL agents and enhance the algorithm's precision and convergence rate, Deep Deterministic Policy Gradient (DDPG) is incorporated. Using this approach, susceptibility of a state-of-the-art detection scheme to several different coordinated FDI attacks on the distributed communication links is identified. The proposed algorithm is also enhanced with a sniffing feature to enable maintaining the stealthy attacks even under the sudden disconnection of any of the compromised links. To address the discovered deficiencies within the index-based detection scheme, a complementary multi-agent RL detection algorithm using Deep Q-Network (DQN) is integrated, which provides a more reliable overall identification performance. Taking into account the communication delays and load changes, the effectiveness of the proposed algorithm is verified by the experimental tests.**

***Index Terms*—Distributed Control, DC Microgrid, Cybersecurity, False Data Injection, Reinforcement Learning.**

I. Introduction

DC microgrids have recently received a wide range of attention and growing popularity in power generation systems, as they provide an efficient way for integration of renewable energy systems, energy storage units and electrical power loads [1]. Using a hierarchical control structure with a combination of both primary and secondary control layers, the voltage regulation at the output terminals and current sharing among generation units are deployed in such systems [2]. Conventional approach for the secondary control schemes were formed on the basis of the centralised control, where a single control block was in charge of receiving secondary signals and dispatching the voltage regulatory terms to all downstream primary control units based on the underlying control objectives. However, this approach makes the system vulnerable to the single point of failure. To overcome this problem, distributed control algorithms have been developed, where the secondary control command signals are generated at the place of each node based on the received distributed signals from the neighboring agents and a consensus rule of operation. Owing to the dense integration of communication links among the neighboring agents and local-to-secondary control layers for each agent, DC microgrids are highly prone to malicious cyber-attacks. Such intrusions can highly deteriorate the system's performance and even result in unstable conditions and protective circuits tripping under severe cases. Different types of cyberattacks on DC microgrids and their detrimental impacts are studied in the literature including False Data Injection (FDI) [3], Denial Of Service (DOS) [4], Hijacking [5] and Man In the Middle (MIM) [6] attacks.

Compared with other forms of cyberattacks, FDI is known as one of the most challenging types for proper detection and it can occur in different forms. Destabilizing FDIs can make the microgrid unstable with only a minimal uncoordinated penetration level. On the other hand, deceptive FDIs can produce deviations from optimal operating points without loss of regulation through more coordinated attacks [7]. The latter can be only generated with a limited set of coordinated intrusions and is effectively detectable by the existing identification algorithms [7], [8]. However, it is highly critical to ensure the reliable performance of the reported detection schemes against all possible forms of destabilizing FDIs, where any detection failure in this regard can result in protective circuit tripping or damage to power converters. In efforts to effectively address the aforementioned destabilizing or deceptive attacks, multiple FDI detection and mitigation algorithms are reported in different research works. Such schemes can be generally categorized into the model-dependent and model-independent methods.

For model-dependent schemes, researchers have incorporated adaptive control concepts [9], or sliding mode observers [10]. However, complexity and precision of such algorithms are highly dependent on the order size of the underlying model. They are also prone to instability under system parameter uncertainties or presence of multiple coordinated cyberattacks. Model-free FDI mitigation methods based on the distributed observers are also reported in the literature, where adaptive distributed terms [11], or sliding mode observer based distributed terms [12] are employed to rectify the FDI adverse

Ali Jafarian Abianeh and Farzad Ferdowsi are with the Electrical and Computer Engineering Department of the University of Louisiana at Lafayette, Louisiana, USA (emails: ali.jafarian-abianeh1@louisiana.edu, farzad.ferdowsi@louisiana.edu).

Yihao Wan, Nenad Mijatovic and Tomislav Dragičević are with the Electrical Engineering Department of Technical University of Denmark, Copenhagen, Denmark (e-mails: wanyh@elektro.dtu.dk, nm@elektro.dtu.dk, tomdr@elektro.dtu.dk).

impacts. However, such schemes are highly reliant on the secure transmission of the extra distributed signals, which can be themselves targeted by the FDIs. In addition, their dynamic performance are significantly deteriorated with the communication delays, and their load switching response is also adversely impacted by the integrated distributed terms.

FDI detection algorithms based on the supervised learning have also been investigated [13], [14], but their performance is greatly impacted by the quality of collected labeled dataset, and they are also prone to the over-fitting phenomenon. To mitigate these challenges, model-free FDI detection schemes, based on the system physical observations, are recently employed by researchers. A signal-temporal-logic approach is used in [15] to detect a sawtooth form of FDI, and an exponential data integrity index on the distributed current signals is employed in [16] to signal out the compromised links. However, the performance of these schemes against deceptive attacks are not studied. A discordant detection algorithm is proposed by [8] for detection of both deceptive and destabilizing FDI attacks on secondary regulation of current signals. This scheme is formed on the basis of monitoring the synchrony of the resultant current references. A similar approach is also utilized by [17] and [18] to develop event-driven cyber-attack detection and mitigation algorithms against FDI attacks. Despite the promising performance of such model-free FDI detection schemes, their performance against more systematic FDIs is still not guaranteed. Thus, it is crucial to explore the susceptibilities within the existing algorithms, and accordingly apply the proper modifications.

Reinforcement Learning (RL) algorithms have recently received an enormous attention in the cyber-physical systems [19], as it is known as the closest form to the human learning compared with other types of intelligent algorithms. However, only very limited number of research works have explored RL application to cyber-security in microgrids and smart grids [20]. Using a combination of SARSA RL on the learning phase and Q-Learning RL for FDI detection are reported in [21] and [22]. However, Q-learning based algorithms do not provide an efficient solution for real-world applications where deep learning based RLs are more desired. Despite some reported RL based FDI detection schemes, the great potential in such learning methods is still not well realized for vulnerability exploration and exploitation in the existing cyberattack detection schemes and developing effective complementary mitigation. In a recent research study [23], the application of a Temporal Difference (TD) RL actor-critic based method is studied for intervening the cost optimization in the tertiary control layer of microgrids.

In this paper, deep RL algorithms are proposed to autonomously discover the vulnerabilities of the index based cyberattack detection methods commonly used in distributed control of DC microgrids, and provide complementary solutions. Using a multi-agent RL approach with Deep Deterministic Policy Gradient (DDPG) agents for exploring the FDI cyberattack continuous space action, a more precise identification of the vulnerable spots is attained. The proposed method explores the detection algorithm susceptibilities against stealthy destabilizing FDIs on distributed links in a way that indices remain minimized to the normal operating condition. Then, a multi-agent RL DQN based scheme is proposed to supplement the identified detection weaknesses and operates in conjunction with it. The performance of the proposed scheme is verified using an experimental testbed against one of the state-of-the-art model-free FDI detection methods [8]. To the best of authors' knowledge, this paper proposes the first multi-agent deep RL based schemes for cybersecurity issues in the secondary control layer of microgrids. Thus, the main contributions of this paper can be summarized as follows:

- A novel approach for automatic discovery of the vulnerabilities within the existing cyberattack detection algorithms is proposed using the reinforcement learning concept. This method enables both wide-range and targeted exploration of the penetrable spots for all the index-based cyberattack detection schemes, and provides foundations for their effective mitigation.
- The proposed deficiency identification scheme is implemented using the multi-agent DDPG RL agents. This multi-agent configuration facilitates its effective integration into the existing distributed cyberattack detection schemes and alleviates the impacts of communication delays on its performance compared with a centralised approach. In addition, utilization of the DPPG agents enables a continuous space action for finer exploration of the susceptibilities to cyberattacks.
- Using the identified deficiencies in the model-free index based detection schemes, a complementary multi-agent RL DQN based cyberattack identification algorithm is proposed, which signals out coordinated attacks undetected by the fundamental scheme.

The rest of this paper is organized as follows; in Section II, distributed control for DC microgrids and discordant detection algorithm are discussed. The proposed multi-agent RL based algorithms are also presented in Section III. Experimental results are provided in Section IV, and Section V concludes this research study.

II. Distributed control of DC microgrids with Discordant Cyberattack Detection

An autonomous DC microgrid with the topology shown in Fig. 1 is considered. In this system, each DC source is connected to the common DC bus through a DC-DC converter, which is regulated with cascaded voltage and current controllers at the primary layer. Distributed secondary regulators are also integrated with the primary controllers to enable transmission and sharing the distributed terms of $\phi_n = \{\bar{V}_{dc_n}, I_{dc_n}\}$ for n distributed agents based on the underlying communication topology. In this case, $\bar{V}_{dc_n}$ denotes the estimated average voltage, and I_{dc_n} is the per unit value for the output current. For the adjacency matrix $\boldsymbol{A} = [a_{ij}]$ with the dimension of $n \times n$, the resultant consensus secondary term (u_i) at the place of node i can be then represented by:

$$u_i = \sum_{j=1}^{n} a_{ij}(\phi_j - \phi_i), \quad where \;\; a_{ij} = \begin{cases} > 0, \; if(x_i, x_j) \, \epsilon \, \boldsymbol{G} \\ 0, \; else \end{cases} \tag{1}$$

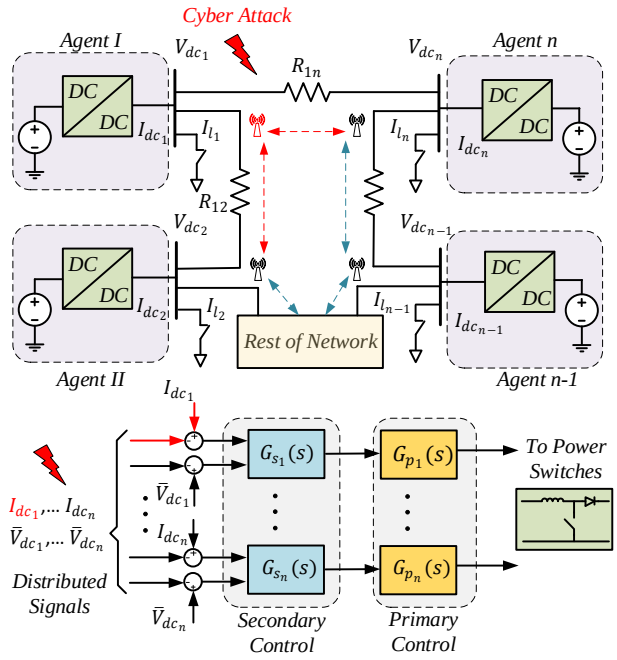


Fig. 1. General block diagram for distributed control of DC microgrid with N agents under cyber-attacks.

where a_{ij} denotes the interconnection between all nodes, $\boldsymbol{G}$ represents the existing communication topology, and x_i and x_j are the secondary signals for local and neighboring nodes.

To implement secondary voltage and current sharing between the neighboring agents, it is required to modify the primary voltage setpoint for node i as follows:

$$V^*_{dc_i} = V_{dc_{ref}} + \triangle V_{1_i} + \triangle V_{2_i} \tag{2}$$

where $V_{dc_{ref}}$ is the global reference voltage for all agents and $V^*_{dc_i}$ is the voltage reference to the primary controller at node i. $\triangle V_{1_i}$ and $\triangle V_{2_i}$ also represent the resultant regulatory terms from the secondary voltage and current controllers at node i, respectively, and can be formulated with:

$$\triangle V_{1_i} = (K^V_p + \frac{K^V_i}{s}) \cdot (V_{dc_{ref}} - u^V_i) \tag{3}$$

$$\triangle V_{2_i} = (K^I_p + \frac{K^I_i}{s}) \cdot (I_{dc_{ref}} - u^I_i) \tag{4}$$

where $I_{dc_{ref}}$ is the global current setpoint, u^V_i and u^I_i are the consensus terms for voltage and current, and K^V_p, K^V_i, K^I_p, K^I_i represent the proportional and integral gains for voltage and current PI controllers, respectively. For a proportionate current sharing among the neighboring agents, $I_{dc_{ref}}$ is set by zero.

Due to the droop concept for the interconnected DC power sources, the secondary sharing algorithms are usually applied to either the voltage or current distributed terms. Since this paper is focused on discovering the vulnerabilities in the index based detection algorithms applied on the current sharing [8], only FDI attacks on secondary current control are investigated. For secondary current regulation schemes, FDI cyberattacks can be applied through an offsetting term either on sensor or neighboring communication links, as formulated by (5):

$$\boldsymbol{X^{FDI}} = \begin{cases} x^{FDI}_i = x_i + k_i x_f, \ for\ sensor\ attack \\ x^{FDI}_{ij} = x_{ij} + k_{ij} x_f, for\ link\ attack \end{cases} \tag{5}$$

where $\boldsymbol{X^{FDI}}$ denotes the matrices of compromised communication signals, x^{FDI}_i and x^{FDI}_{ij} are the FDI manipulated signals for sensor and neighboring links, k_i and k_{ij} are the FDI scaling factors for sensor and link communications, respectively, while x_f is the fundamental FDI intrusion term.

The FDI cyberattacks can be generated in the form of destabilizing or deceptive attacks through coordinated or uncoordinated intrusions. As a result, a discordance effect is experienced on the input current signals, which forms the basis for the discordant FDI detection algorithm [8], as formulated in (6). In this algorithm, the resultant impacts from any forms of FDIs including attacks on sensors, communication links or concurrent ones are monitored through the deviations introduced on the input current signal references for the neighboring agents. As also represented by (7), the presence of cyberattack on the distributed control block for node i is detected when the positive term DE_i has a value greater than its minimum threshold DE_{min}. This value is chosen by considering the resultant DE_i values under normal operating conditions in the presence of other possible contributing factors such as underlying controllers' performance and existing communication delays. In addition, a time delay triggering process is employed before applying the counteracting measure in order to account for load disturbances. However, for the normal operating conditions, the discordant value should never consistently retain a value greater than its lower threshold.

$$DE_i = M_i \cdot [\sum_{j \epsilon G_i} (I^*_{j_{in}} - I^*_{i_{in}})][\sum_{j \epsilon G_i} (I^*_{j_{in}} + I^*_{i_{in}})] \tag{6}$$

$$DE_i = \begin{cases} < DE_{min} : & for\ k_i\ \&\ k_{ij} = 0 \\ > DE_{min} : & for\ k_i \| k_{ij} \neq 0 \end{cases} \tag{7}$$

where DE_i is the discordant term at node i, $DE_{min} > 0$ is its minimum threshold, $I^*_{i_{in}}$ and $I^*_{j_{in}}$ are the reference input current values at local node i, and neighboring nodes j, M_i is the scaling factor for discordant term, and G_i represents the communication graph to the neighboring agents.

III. Proposed Multi-Agent RL Schemes to Unveil Susceptibilities and Complement Detection

A. Multi-Agent Deep Reinforcement Learning

For development of multi-agent RL algorithms in an observable environment, the problem is defined as a Markov Decision Process (MDP) characterised with the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma \rangle$ for each agent in the agent set $\mathcal{N} = \{N_1, .., N_m\}$. Where $\mathcal{S} \epsilon \mathbb{R}^n$ represents the finite set of states, $\mathcal{A} \epsilon \mathbb{R}^m$ denotes the finite set of actions, $\mathcal{T}$ is the state transition function that represents the probability of state transition $s^t \rightarrow s^{t+1}$ by taking the action a^t and receiving the immediate reward of r^t, r is the reward function, and $\gamma \ \epsilon \ [0, 1]$ is the discount factor. In each time step, the RL agents observe the current system state s^t and take the action a^t based on the selected policy $\pi(a^t|s^t)$. This

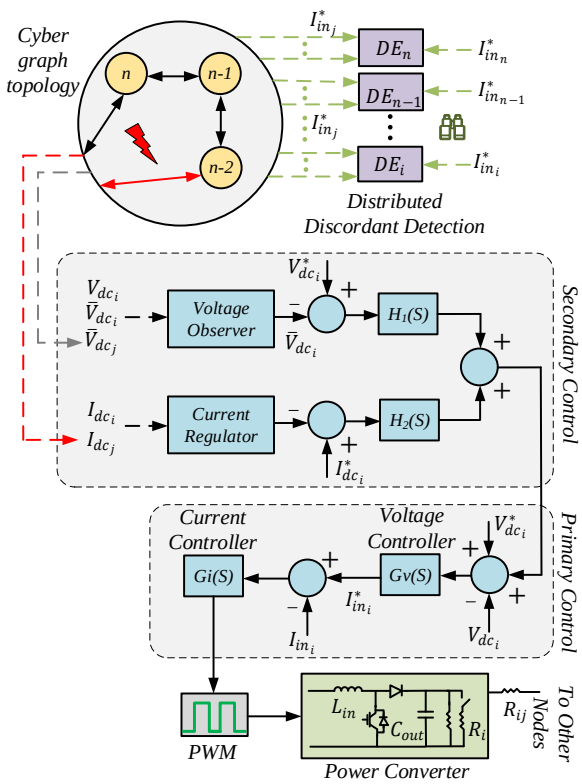


Fig. 2. Block diagram for discordant cyberattack detection and the hierarchical control structure in DC microgrid.

taken action results in receiving an immediate reward r^t, and its transition into the new state s^{t+1}. The term accumulative reward over an infinite time horizon for each RL agent i can be also represented by:

$$\Gamma_i^t = \sum_{n=0}^{\infty} \gamma_i^n r_i^{t+n+1} \tag{8}$$

In order to maximize the accumulative reward in (8), different recursive training algorithms can be applied. In the off-policy based algorithms such as Q-learning, the Bellman iterative equation in (9) with the learning rate α_i is employed to estimate the action-value function Q^π:

$$\begin{aligned} Q^\pi(s_i^t,a_i^t) \leftarrow\ & Q^\pi(s_i^t,a_i^t)+ \\ & \alpha_i[r_i^{t+1} + \gamma_i max Q^\pi(s_i^{t+1},a_i^{t+1}) - Q^\pi(s_i^t,a_i^t)] \end{aligned} \tag{9}$$

However, this iterative approach does not provide a feasible performance in a high dimensional real-world application. To enable a more precise prediction of the action-value function Q_i^π for each pair of state-action, DQN is employed for the RL agents [24]. In DQN algorithm, first a random mini-batch of $\mathcal{S}$ samples (s^j, a^j, r^j, s'^j) from the replay buffer $\mathcal{D}$ is chosen for each agent i. Then, the critic network is adjusted by trying to predict the return value with minimizing the following loss function:

$$L(\theta_i) = \frac{1}{S}\sum_j (y^j - Q_i^\pi(s^j, a_1^j, ..., a_m^j))^2 \tag{10}$$

where y^j is set with:

$$y^j = r_i^j + \gamma_i Q_i^{\pi'}(s'^j, a'_1, ..., a'_m)|_{a'_i=\pi'_i(s^j)} \tag{11}$$

While DQN RL agents can efficiently meet the algorithm objectives in some applications where the limited set of discrete actions are adequate to interact with the environment, RL agents with the capability of continuous space action using actor-critic networks such as DDPG, are unavoidable for more complex environments [25]. Similar to the DQN, the critic network is adjusted by minimizing the loss function in (10), but the actions are decided based on the adjusted actor network with minimizing the loss function in (12) to acquire the optimal policy parameter θ:

$$\nabla_{\theta_i} J = \frac{1}{S}\sum_j \nabla_{\theta_i}\pi_{\theta i}(a^j|s^j)\nabla_{a_i} Q_i^\pi(s^j, a_1^j, ..., a_m^j)|_{a_i=\pi_{\theta i}(s^j)} \tag{12}$$

Then the target network parameters for both actor and critic networks are updated with (13):

$$\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i \tag{13}$$

where $\tau \ll 1$.

B. Multi-Agent RL DDPG to Uncover Cyberattack Detection Deficiencies

In order to automatically discover the vulnerabilities within an index-based cyberattack detection scheme, the problem of coordinated cyberattack generation can be formulated as a Markov Decision Process (MDP). With specifying a continuous action space for cyberattack exploration using the DDPG RL agent, the task of generating stealthy FDI attacks can be accomplished by properly rewarding the low detection indices in presence of intrusions. The application of this approach to the discordant detection scheme, as an existing well-established identification algorithm, is explained in this section, while the similar approach can be applied to all other index based detection schemes. Due to the lower vulnerabilities and exposure of node links to cyberattacks, the proposed algorithm only explores the undetectable FDI attacks on the neighboring links. However, it can be easily reconfigured for nodes and combinative attacks as well. In this section, a multi-agent DDPG reinforcement learning based FDI attack generation scheme is proposed. This approach enables modular integration of the proposed distributed RL cyberattack on the more densely connected or expanded networks, lessens the impact of transmission delays from distributed agents to a centralised attacking unit, and ensures optimal complexity level for each trained agent in terms of possible output actions and convergence effort.

In a DC microgrid with m incoming distributed communication links, the agents list is set with $\boldsymbol{N} = \{CA_1, ..., CA_{m+1}\}$ in which CA denotes the RL based cyberattack generation agents. Considering the existing inter-dependency of the neighboring discordant terms, the observations list for each agent at the instance t is defined as $\boldsymbol{O_{CA}^t} = \{\boldsymbol{O_1^t}, \boldsymbol{O_2^t}\}$, where $\boldsymbol{O_1^t} = \{DE_1^t, ..., DE_{m+1}^t\}$ is devoted to the neighboring

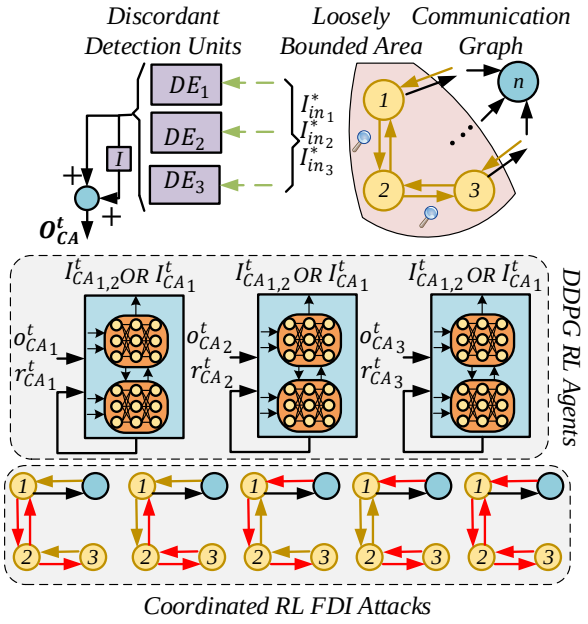


Fig. 3. Proposed multi-agent DDPG RL-based FDI attacks against DC microgrid equipped with discordant detection algorithm- Sample network.

discordant terms, and $\boldsymbol{O_2^t} = \{\int DE_1^t, ..., \int DE_{m+1}^t\}$ represents their associated integration. With respect to the received observations, the action set $\boldsymbol{A_{CA}^t} = \{I_{CA_1}^t, ..., I_{CA_m}^t\}$ is generated at instance t, where I_{CA}^t is the FDI intrusion term applied on the incoming communication link m at moment t, and $\boldsymbol{A_{CA}^t} \epsilon \boldsymbol{A_{CA}}$ where $\boldsymbol{A_{CA}}$ represents the finite set of available actions to CA agent. The corresponding algorithm steps are also presented by Algorithm 1. The reward function for each agent at time t is also defined by (14), which is characterised with both continuous and discrete reward terms:

$$r_{CA}^t = -\left(k_{DE}\sum_{i=1}^{m+1}(DE_i^t)^2 + k_{\dot{DE}}\sum_{i=1}^{m+1}(\dot{DE}_i^t)^2 + k_{\dot{I}_{CA}}\sum_{j=1}^{m}(\dot{I}_{CA_j}^{t-1}))^2 + r_{d_p}^t - r_{d_n}^t\right) \tag{14}$$

where k_{DE}, $k_{\dot{DE}}$ and $k_{\dot{I}_{CA}}$ are the reward coefficients for summation of neighboring discordant terms, their corresponding derivatives, and derivatives of cyberattack actions taken in $t-1$, respectively. $\dot{DE}_i^t$ and $\dot{I}_{CA_j}^{t-1}$ also denote the derivatives for discordant terms at time t and cyberattack action signals at time $t-1$, respectively. $r_{d_p}^t$ and $r_{d_n}^t$ also represent the positive and negative discrete reward terms considered to ensure effective training for desired destabilizing conditions, penalizing the excessive DE observation values, while rewarding the stealthy destabilizing attacks. These discrete reward terms also facilitate the convergence process during the training stage.

For the continuous reward terms, k_{DE} is tuned to ensure minimized discordant observations, $k_{\dot{DE}}$ is applied to reflect the impact of excessive transient modes from disturbances such as load changes or cyberattack variations for enhanced stealthy performance, and $k_{\dot{I}_{CA}}$ is in charge of minimizing variations on the generated cyberattack terms especially while the desired objectives are met. However, if more dynamic cyberattack steps are desired, this coefficient can be adjusted with lower values. The expansion of discrete reward terms are also represented by (15)-(21):

$$r_{d_p}^t = K_{d_p}(\mathcal{E}^t \,\&\, \mathcal{P}^t) \tag{15}$$

$$r_{d_n}^t = K_{d_{n1}}\left(\overline{\mathcal{E}^t} \,\&\, (\mathcal{G}^t|\mathcal{H}^t)\right) + K_{d_{n2}}\mathcal{F}^t \tag{16}$$

$$\mathcal{P}^t = (|\sum_{j=1}^{m}(I^{t-1}{}_{CA_j} - I^{t-1}{}_{CA_i})| > I_{CA_{min}}) \tag{17}$$

$$\mathcal{E}^t = \left((DE_1^t \,\&\, ... \,\&\, DE_i^t) < DE_{min}^t\right)\Big|_{i=1}^{m+1} \tag{18}$$

$$\mathcal{F}^t = \left((DE_1^t|...|DE_i^t) > DE_{max}^t\right)\Big|_{i=1}^{m+1} \tag{19}$$

$$\mathcal{G}^t = (I_{1_{in}}^*|...|I_{i_{in}}^* > I_{max}^*)\Big|_{i=1}^{m+1} \tag{20}$$

$$\mathcal{H}^t = (I_{1_{in}}^*|...|I_{i_{in}}^* < I_{min}^*)\Big|_{i=1}^{m+1} \tag{21}$$

where K_{d_p} is the positive discrete reward coefficient for stealthy destabilizing condition, $K_{d_{n1}}$ and $K_{d_{n2}}$ are the negative discrete reward coefficients for non-stealthy destabilizing condition and excessive discordant term detection, respectively. $\mathcal{P}^t$ denotes the presence of significant non-canceling intrusion terms on the overall action outputs from time $t-1$. $\mathcal{E}^t$ and $\mathcal{F}^t$ also represent the acceptable and excessive discordant term detection, and $\mathcal{G}^t$ and $\mathcal{H}^t$ also represent occurrence of outbounded $I_{i_{in}}^*$ when it hits the upper and lower limits, respectively. In terms of the specified threshold values, $I_{CA_{min}}$ is the minimum overall intrusion action term, DE_{max}^t and DE_{min}^t are the upper and lower thresholds for discordant terms, I_{max}^* and I_{min}^* are the upper and lower bounds to the observed neighboring terms I_{in}^*, respectively.

In terms of the threshold value selection for these discrete reward terms, DE_{min}^t is selected based on the normal operation condition and common system disturbances, $I_{CA_{min}}$ is chosen with a trade-off between the desired ramp up/down slope for the destabilizing phenomenon, and the incorporated minimum discordant threshold. While I_{max}^* is chosen with respect to the protective circuit tripping thresholds, I_{min}^* is adjusted with zero for a dc microgrid with purely resistive loads, or with a safe value for a system which has constant power loads. In terms of negative discrete reward coefficients, $K_{d_{n2}} > K_{d_{n1}}$ is chosen to further penalize occurrence of non-stealthy attack conditions. In addition, the positive discrete reward coefficient of K_{d_p} is adjusted with a value with respect to the observed convergence performance.

In order to minimize the intrusion level requirements for the proposed algorithm, it can target the node with the weakest bonding level to its neighboring agents and also have the m action signals for each agent combined to a single action. The only drawback to merging action signals is that it becomes more vulnerable to detection if any of the targeted communication links is disconnected. To ensure a dynamic stealthy destabilizing FDI attack performance under

this circumstance, the algorithm can be enhanced with a sniffer on the compromised link data transmissions. Using this sniffing feature, any disconnection on the compromised links can be detected and surpassed by switching to other operational incoming communication links connected to the same node.

Algorithm 1 Multi-Agent DDPG to Unveil Cyberattack Susceptibilities

Initialize weights of actor and critic networks, replay buffer $\mathcal{D}$, and target networks.
for episode = 1 to M **do**
  Receive initial process observation at state s_1^t.
  for iteration = 1 to T **do**
    For each agent i, select and execute action a_i^t with respect to policy $\pi(s_i^t|a_i^t)$, receive the reward r_i^t calculated with (14) and transition into state s_i^{t+1}.
    Store tuple $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ in the $\mathcal{D}$.
    $s_i^t \leftarrow s_i^{t+1}$
    for each agent i = 1 to m **do**
      Randomly select the mini-batch $\mathcal{S}$ from $\mathcal{D}$.
      Set y^j according to (11).
      Update the actor and critic networks with (12) and (10), respectively.
    end for
    Update the target network using (13).
  end for
end for

C. Complementary Cyberattack Detection with Multi-Agent RL DQN

In order to overcome the inefficacy of the discordant scheme on detecting a group of coordinated link FDI attacks, as discovered by the proposed multi-agent DDPG algorithm, a complementary multi-agent DQN FDI detection algorithm is proposed. This algorithm is trained with the recorded dataset from undetected FDI intrusion vectors. This complementary feature is activated if the discordant scheme does not reflect any irregularities and its associated DE terms remain minimized. In this case, the corresponding list of RL agents are set with $\boldsymbol{N} = \{RD_1, ..., RD_n\}$, where RD represents the RL based detection agent operated in parallel with the distributed discordant unit for each of the n nodes. The list of observation signals for each agent is also defined as $\boldsymbol{O_D^t} = \{IE_s^t, \dot{IE}_s^t\}$, where $IE_s^t = \sum_{j=1}^{m} |I_j - I_i|$ is the accumulative distributed current error term and $\dot{IE}_s^t$ denotes its corresponding derivatives for m incoming distributed communication links connected to the node i. The list of discrete actions for each agent is also selected with $\boldsymbol{R_D^t} = \{0, 1\}$, where 1 represents detection of FDI intrusion presence and 0 is signaled out under normal operating condition. It should be also noted that the observations to the DQN agents are only enabled if the corresponding discordant terms are within the normal operating range with some delay to avoid overlapped or false detection.

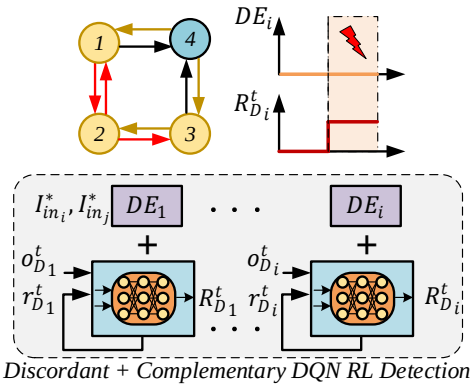


Fig. 4. Proposed complementary multi-agent DQN RL detection scheme to mitigate the detection failure on discordant algorithm.

The reward function r_D^t for each agent is also formulated by (22). Since the proposed RL DQN scheme is designed to operate as a FDI detection scheme, similar to the discordant method, and generate only discrete action signals, the reward function only includes the discrete reward terms. It is basically formed with two negative and one positive reward terms using the detection action signal from the previous time step $R_{D_i}^{t-1}$, FDI presence on any of the incoming distributed signals $F_{D_i}^t$, and their corresponding derivatives as represented by $\dot{R}_{D_i}^{t-1}$ and $\dot{F}_{D_i}^t$, respectively. The first negative reward term ensures minimal delay on intrusion state detection, and the second negative discrete reward term is in charge of minimizing excessive action signal alterations. The third term is devoted to rewarding proper detection of FDIs with minimal delay over one complete episode period. The inclusion of $F_{D_i}^t$ as a part of reward equation not only enables more effective rewarding formulation but also ensures the improved algorithm convergence. It should be also noted that after the training process in offline mode, only the observation signals are fed into the RL agents. The algorithm steps for the proposed complementary multi-agent RL DQN detection scheme are also presented by Algorithm 2.

$$\begin{aligned} r_D^t = &- K_{d_{n1}}(R_{D_i}^{t-1} \oplus F_{D_i}^t) - K_{d_{n2}}|\dot{R}_{D_i}^{t-1}| \\ &+ K_{d_p}(0 \leq \mathcal{B}^t \leq \mathcal{B}_d)u(t - T_p + T_s) \end{aligned} \tag{22}$$

where $K_{d_{n1}}$ and $K_{d_{n2}}$ are the negative reward coefficients for proper detection and action variations, respectively, and K_{d_p} is the positive reward coefficient for desired detection performance over an episode time period of T_p and with the step time of T_s. Considering the impacts associated with each of these reward terms, the reward coefficients should be adjusted in a way that condition $K_{d_{n2}} < K_{d_{n1}} \ll K_{d_p}$ is satisfied. Also, $u(.)$ is the step function and the term $\mathcal{B}^t$ is represented by (23), which implies a desired detection performance over the complete episode period with p steps. $\mathcal{B}_d$ is also an integer constant value chosen based on the desired delay performance on detection and with respect to the observation input delays.

$$\mathcal{B}^t = (\sum_{k=1}^{p}(R_{D_{ik}}^{t-1} \oplus F_{D_{ik}}^t) - \sum_{k=1}^{p} \dot{F}_{D_{ik}}^t) \tag{23}$$

Algorithm 2 Multi-Agent DQN to Complement Cyberattack Detection

Initialize replay buffer $\mathcal{D}$, and action-value function Q with random weights.
for episode = 1 to M **do**
  Receive initial process observation at state s_1^t.
  for iteration = 1 to T **do**
    For each agent i, select and execute action a_i^t with respect to policy $\pi(s_i^t|a_i^t)$, receive the reward r_i^t calculated with (22) and transition into state s_i^{t+1}.
    Store tuple $(s_i^t, a_i^t, r_i^t, s_i^{t+1})$ in the $\mathcal{D}$.
    $s_i^t \leftarrow s_i^{t+1}$
    for each agent i = 1 to m **do**
      Randomly select the mini-batch $\mathcal{S}$ from $\mathcal{D}$.
      Set y^j according to (11).
      Perform gradient descent on (10) with (11).
    end for
    Update the target network using (13).
  end for
end for

IV. Experimental Results

In order to verify the performance of the proposed multi-agent DDPG RL-based system on discovering the vulnerabilities in the cyberattack detection scheme, and generating stealthy destabilizing FDI intrusions against the discordant algorithm, an autonomous DC microgrid configuration, as previously depicted in Fig. 1, with $n = 4$ power generation units is considered. The system electrical and control parameters for both primary and secondary control layers are also presented in Table. I. The effectiveness of the proposed multi-agent RL DQN on complementing the discordant detection algorithm and mitigating its vulnerability to coordinated FDIs is also experimented. The training process is carried out in the Matlab/Simulink environment, and the algorithm verification is performed using the experimental setup shown in Fig. 5 by means of dSPACE MicroLabBox DS1202, where only control parameters are slightly modified to ensure similar controller performance under both conditions.

TABLE I. Experimental Testbed Parameters

Parameter Sets	Parameter Values
Plant	$R_{12} = R_{23} = R_{34} = 0.5\ \Omega, R_{14} = 0\ \Omega$
Converter	$L_{in} = 0.86\ mH, C_{out} = 1.1\ mF, f_s = 10\ khz, I_{rated} = 32\ A$
Controller	$V_{in} = 48\ V, V_{dc_{ref}} = 60\ V, I_{dc_{ref}} = 0, M_i = 2$ $G_p(s): K_{pV} = 1, K_{iV} = 20, K_{pI} = 2.4, K_{iI} = 10$ $G_s(s): K_p^I = 0.15, K_i^I = 0.06$
Load	$R_1 = R_2 = R_3 = R_4 = 30.6\ \Omega, R_1 = 30.6 \rightarrow 65.7\ \Omega$

Three DDPG RL agents are structured similar to Fig. 3, to enable proper exploration of cyberattacks on the neighboring communication links and generate a group of different coordinated FDI attacks between the agents on nodes 1, 2, and 4. Moreover, for complementing discordant detection units, a DQN RL agent is integrated for each node. The parameters

TABLE II. Hyperparamaters for DDPG RL Agents and DQN RL Agents

Hyperparameters	RL DDPG	RL DQN
Batch Size	512	64
Discount Factor	Actor: 0.9995 Critic: 0.9995	0.99
Learning Rates	Actor: 10^{-4} Critic: $5x10^{-4}$	$1x10^{-3}$
Hidden Layers/Nodes	Actor: 2/2048 Critic: 2/1024	2/512

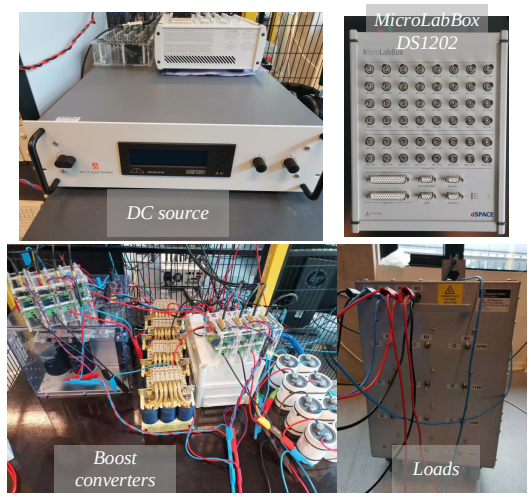


Fig. 5. Experimental Setup

for DDPG RL agents as well as DQN RL agents are also provided in Table II. Using the previously discussed reward function and observation vector for each agent, the multi-agent FDI generation unit and complementary detection unit are trained through the simulation model for a run-time period of 6 seconds for each training episode. Both primary and secondary controllers as well as MA RL units are implemented inside the dSPACE controller with the sampling time of 100 μs. In order to account for the communication delays, the primary to secondary delay term t_{p-s}, and cyberattack output delay term t_{CA} values are set with 5 ms and 40 ms, respectively, and 70-100 ms delay on the secondary to secondary communication link delays t_{s-s}, are considered in the test cases.

Scenario A: The performance of the discordant cyber-attack detection algorithm under load switchings and both conventional deceptive and destabilizing FDI attacks on the distributed current signals is shown in Fig. 6, where the associated output voltages, distributed currents and discordant terms are displayed. For this experiment, $t_{s-s} = 100\ ms$ is set and the default delay value for t_{p-s} is maintained. In this case, subsequent load step-up and step-down are first applied at $t = 21\ s$ and $t = 41\ s$, respectively, where a proper convergence among the distributed current signals to $I = 1.55\ A$ and $I = 1.25\ A$ within less than 8 seconds are resulted, as shown in Fig. 6b. A slight dwell on the discordant terms under the load transient conditions in Fig. 6c is also noteworthy, where their convergence rates and peak values are a function of underlying current controllers performance. At

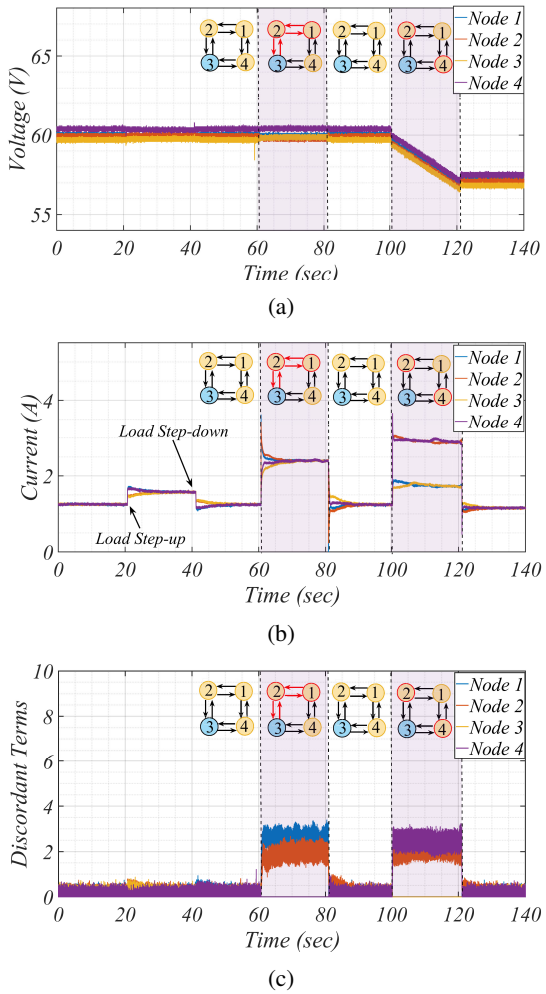


Fig. 6. Performance of the discordant cyberattack detection algorithm under load steps and both conventional deceptive and destabilizing FDI attacks with $t_{s-s} = 100\ ms$.

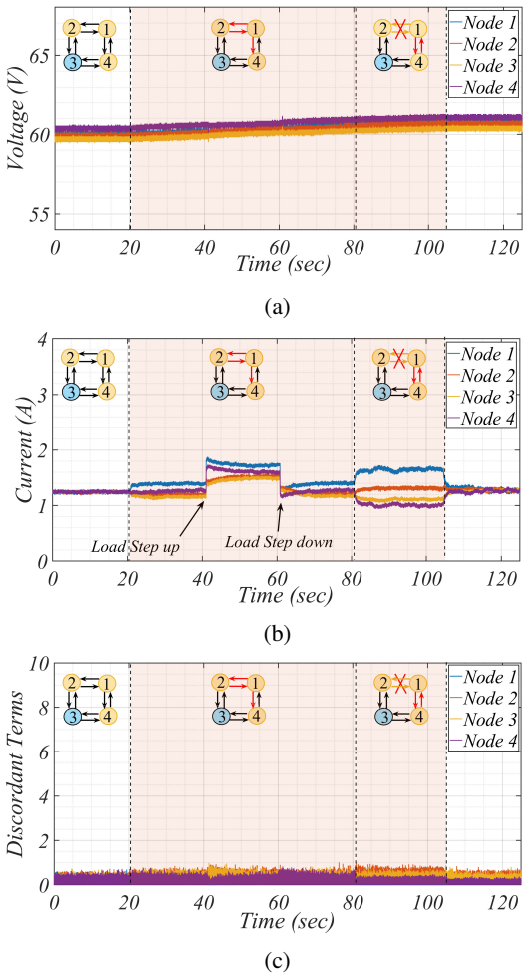


Fig. 7. Performance of the proposed Multi-agent DDPG RL unit on generation of stealthy destabilizing FDI attacks under load steps and sudden compromised link disconnection- two compromised links and $t_{s-s} = 100\ ms$.

$t = 62\ s$, conventional deceptive FDI attacks are introduced into the nodes 1 and 2 with 2.2 A and 2 A, respectively, which target the associated sensors and outgoing distributed terms. The impact of such an attack is reflected in the similar manner to the load step-ups as all current signals converge to a value of 2.5 A, and this attack is properly detected with the significant increase on the associated discordant terms. Effectiveness of the discordant algorithm on identifying the destabilizing attack is also verified where it detects the intrusion vector 2.2 A and 2 A on sensors for agent 2 and 4, which initiated at $t = 100\ s$, with discordant values greater than 2 which is significantly distinctive from its normal operating condition. Destabilizing phenomenon is also further evident for that period with the voltage ramp down to a value of about 57 V, as shown in Fig. 6a ,where it is stopped after removal of attack at $t = 120\ s$.

Scenario B: The effectiveness of the proposed multi-agent RL algorithm on deceiving discordant detection algorithm and generating stealthy destabilizing FDI attacks on this DC microgrid is tested using a specific attack configuration where only two communication links are compromised, as shown in Fig. 7. In this case, $t_{s-s} = 100\ ms$ is applied to the distributed terms and the default delay values for t_{p-s} and t_{CA} are used. While the same initial loading condition is maintained, starting at $t = 20\ s$, the connection link between nodes 1 and 2 is fully compromised and only the incoming signal to node 3 is manipulated. It is observed that despite about 0.35 A deviations on node 1 from the original consensus setpoint and some deviations on current signals for nodes 2 and 3, as shown in Fig. 7b, the associated discordant terms depicted in Fig. 7c fail to properly detect such intrusions. By applying subsequent load step-up and step-down at about $t = 40\ s$ and $t = 60\ s$, respectively, similar convergence performance to the normal operating condition is observed, and the coordinated stealthy attack is remained hidden to the identification algorithm. While sniffing tool is utilized to monitor the availability of connection links between the compromised agents, a sudden communication link disconnection is applied at about $t = 80\ s$. A reaction delay of 60 ms is then considered for re-configuring the intelligent FDI attack to the link compromise between nodes 1 and 3. It is observed that significant deviations up to $0.5A$ are introduced between agents 1, 3, and 4 current signals. However, discordant algorithm is not able to signal

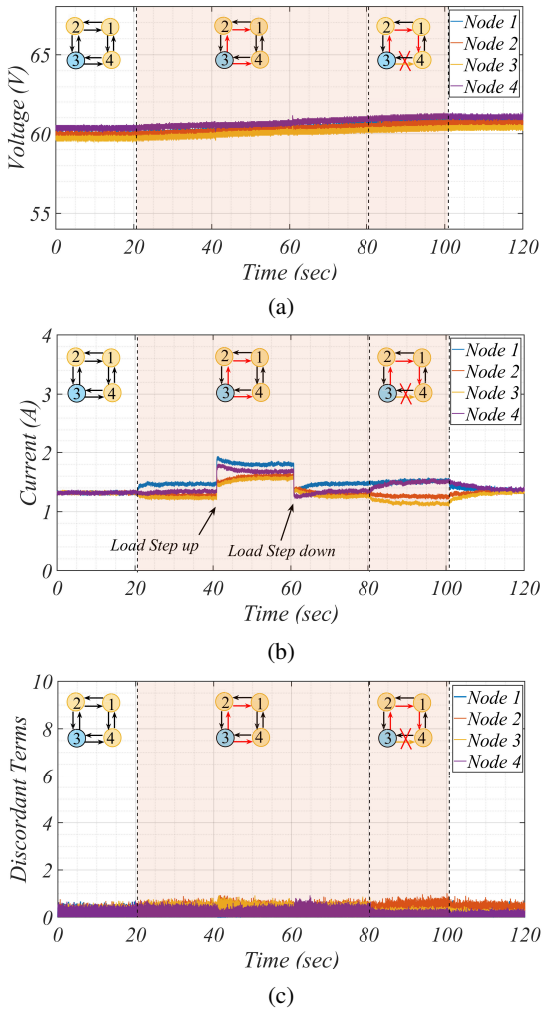


Fig. 8. Performance of the proposed Multi-agent DDPG RL unit on generation of stealthy destabilizing FDI attacks under load steps and sudden compromised link disconnection- only one transmission pathway is impacted on each of the three compromised links and $t_{s-s} = 70\ ms$.

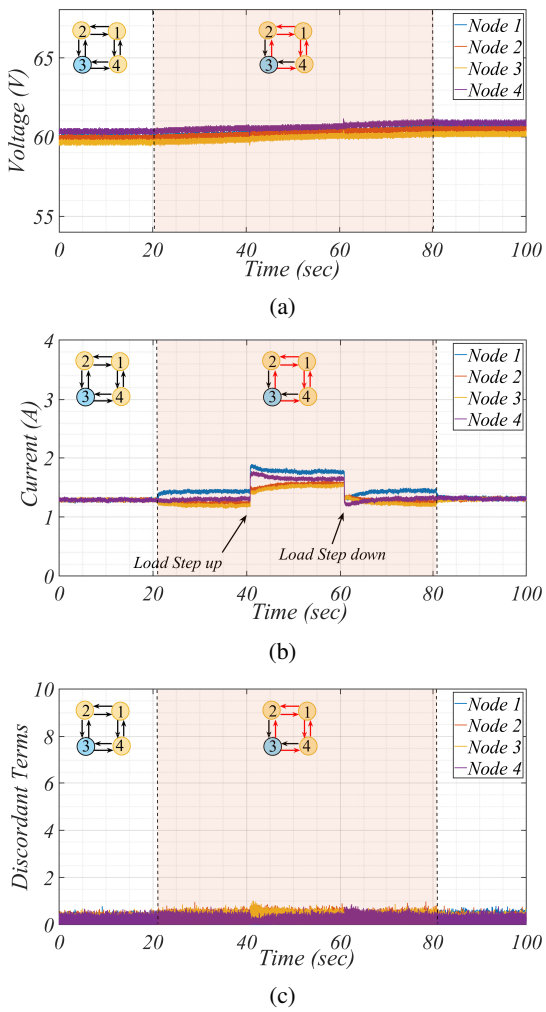


Fig. 9. Performance of the proposed Multi-agent DDPG RL unit on generation of stealthy destabilizing FDI attacks under load steps - two transmission pathways are impacted on each of the three compromised links and $t_{s-s} = 90\ ms$.

out the compromised agents. By keeping all discordant terms to their minimal level values, a voltage ramp-up as a function of interlinking impedances between the nodes is resulted and maintained over the whole RL FDI attack duration, where in this case introduced about $1\ V$ increment on all agents, as depicted in Fig. 7a. This destabilizing condition can deteriorate the regulation performance, and even lead to protective circuits tripping, and erroneous communication link disconnections, especially if it remains over a longer time period.

Scenario C: With a modified RL FDI cyberattack configuration in the distributed control layer with lower transmission delay of $t_{s-s} = 70\ ms$, the attacks are concentrated on three distinctive communication links between the neighboring agents where for each only one communication pathway is impacted, as shown in Fig. 8. It is observed that after applying such a coordinated intrusion at $t = 20\ s$, a similar destabilizing phenomenon occurs where distributed current signals experience deviations from the desired setpoint value, as shown in Fig. 8b, and this effect is maintained after consequent load stepping incidents. At $t = 80\ s$, a sudden link disconnection between nodes 2 and 4 is introduced, where its detection with sniffing tool and re-configuring the attack to the alternative incoming link is applied with the delay of $50\ ms$. Compared with the former attack configuration, lower deviation on distributed term from node 1 is observed which is attributed to its stronger communication bonding under the existing condition. In this case, the other major difference is the negative error introduced on agent 4 with respect to significant positive error in the former test scenario. As a result of introduced deviations, a similar destabilizing voltage ramp-up by about $1V$ over the intrusion period is resulted, as shown in Fig. 8a. From the discordant signals in Fig. 8c, it is also evident that the attacks remained undetected as their values resemble the normal operating conditions.

Scenario D: In this scenario, a more widespread coordinated attack is launched against the three neighboring nodes which impacts three out of four available communication links,

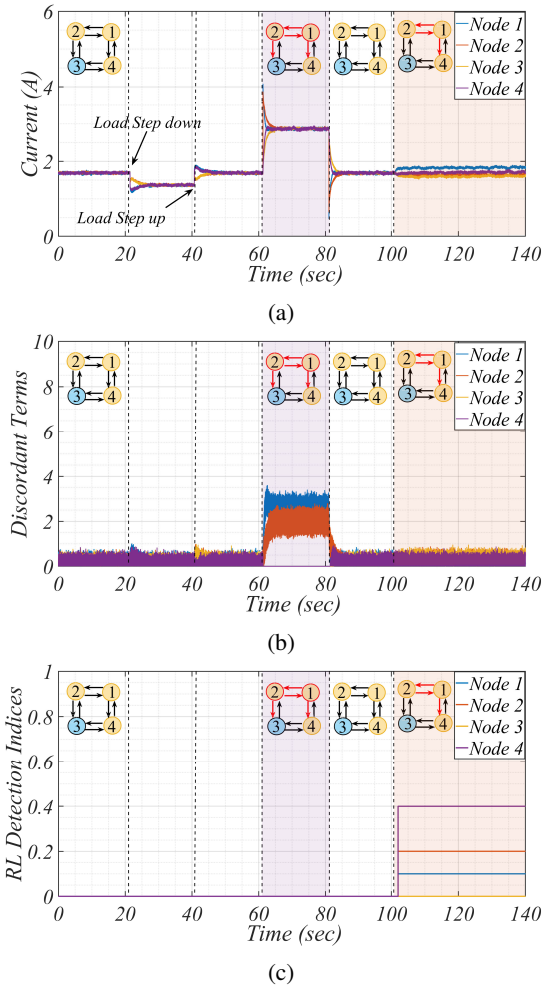


Fig. 10. Performance of the proposed Multi-agent DDPG RL cyberattack generation and multi-agent RL DQN complementary detection units under load steps, $t_{s-s} = 50\ ms$ and $t_{CA} = 80\ ms$.

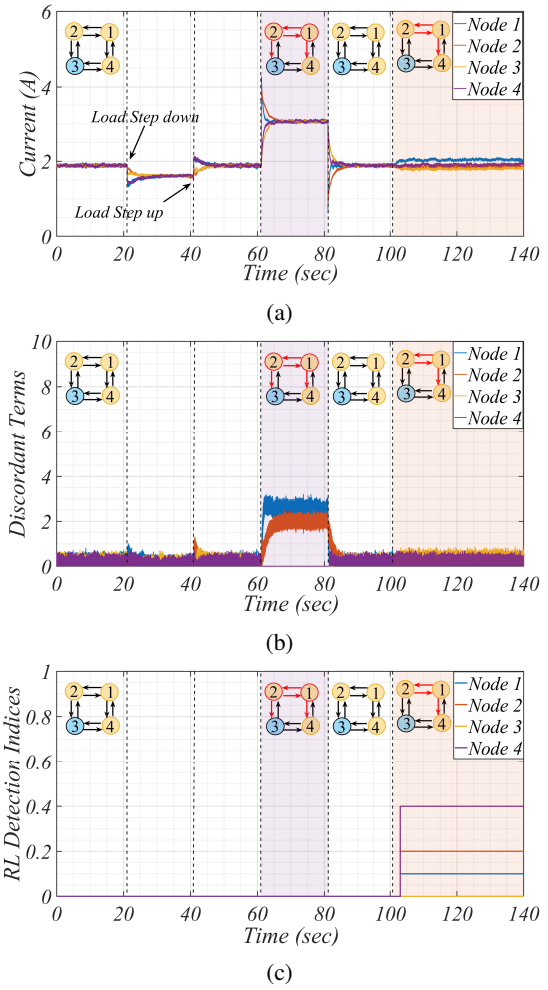


Fig. 11. Performance of the proposed Multi-agent DDPG RL cyberattack generation and multi-agent RL DQN complementary detection units under load steps, $t_{s-s} = 150\ ms$ and $t_{CA} = 80\ ms$.

as depicted in Fig. 9. Unlike the previous coordinated RL attacks which used the merged action signals from the DDPG RL agents, both output actions are incorporated to generate such a stealthy destabilizing attack against the DC microgrid protected with discordant algorithm. It is observed that despite the high level of penetration by the attacker and its persistence for about 60 seconds, the intrusion still remains stealthy as discordant terms does not reflect any distinctive value than their minimal values on the normal operating condition, as depicted in Fig. 9c. It is also observed that such attacks can produce the similar destabilizing impact if it is used either in the merged or independent mode to target at least three incoming communication links for three neighboring agents. In this case, slightly lower voltage ramp up, by about $0.2\ V$, for the destabilizing duration is resulted in Fig. 9a, which is mainly attributed to lower duration of the applied coordinated attacks.

Scenario E: In this case, the performance of the overall combined detection scheme is verified under load steps, conventional deceptive FDI attacks and DDPG RL based stealthy destabilizing FDI attacks. This test scenario is carried out for the cyberattack delay of $80\ ms$ and two different distributed communication delays of $50\ ms$ and $150\ ms$, as depicted in Fig. 10 and Fig. 11, respectively. From the obtained results, it is evident that the discordant method is only capable of detecting the conventional FDI attacks, as applied during 60-80 s, and the corresponding indices remain minimized under load switching and DDPG RL FDIs for both distributed delay conditions. However, the proposed RL DQN detection algorithm properly signals out the non-cooperative nodes within 2-3 seconds after launching stealthy RL attacks. This delay on detection is mainly attributed to the filtered observation signals, communication delays, as well as the chosen $1\ s$ timestep for DQN agents. For enhanced visibility of indices, they are scaled with the corresponding node index and 0.1 scaling factor. Also, it is noteworthy that the performance of the proposed RL DDPG agents are not impacted by the distributed communication delays, where permissible low levels on discordant terms are well maintained during the interval 100-140 for both delay conditions.

V. Conclusions

This paper proposes a multi-agent deep reinforcement learning based algorithm to exploit the vulnerabilities in the existing cyberattack detection methods, which basically provides the foundations for their effective mitigation. The effectiveness of the proposed algorithm is verified by locating the penetrable spots on a sample cyberattack detection algorithm. Using this approach, stealthy destabilizing cyberattacks are launched on the distributed control layer in a DC microgrid protected with the discordant detection algorithm. It is observed that despite the effectiveness of the discordant scheme on detection of the conventional deceptive and destabilizing FDI attacks, it fails to identify more coordinated FDI attacks generated by the proposed scheme. Using the proposed reward function, the training algorithm is reinforced to introduce distributed destabilizing terms into the neighboring communication links in a way that remains hidden to the discordant observers. To overcome the discordant method failure on proper detection of such coordinated stealthy FDIs, a complementary RL DQN detection algorithm is proposed. This hybrid detection approach enables enhancing the reliability of all such index based detection algorithms against the autonomously detected FDI susceptibilities with the aim of reaching a comprehensive cybersecure solution.

Acknowledgment

This work was partially supported by Louisiana Board of Regents under grant number: LEQSF(2021-24)-RD-B-06.

References

[1] A. Jafarian Abianeh and F. Ferdowsi, "Sliding mode control enabled hybrid energy storage system integrated into islanded dc microgrids with pulsing loads," *Sustainable Cities and Society*, 2021.

[2] T. Dragičević, X. Lu, J. C. Vasquez, and J. M. Guerrero, "Dc microgrids—part i: A review of control strategies and stabilization techniques," *IEEE Transactions on Power Electronics*, vol. 31, no. 7, pp. 4876–4891, 2015.

[3] O. A. Beg, T. T. Johnson, and A. Davoudi, "Detection of false-data injection attacks in cyber-physical dc microgrids," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2693–2703, 2017.

[4] J. Liu, X. Lu, and J. Wang, "Resilience analysis of dc microgrids under denial of service threats," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3199–3208, 2019.

[5] S. Sahoo, J. C.-H. Peng, S. Mishra, and T. Dragičević, "Distributed screening of hijacking attacks in dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 7, pp. 7574–7582, 2019.

[6] S. Sahoo, T. Dragičević, and F. Blaabjerg, "Multilayer resilience paradigm against cyber attacks in dc microgrids," *IEEE Transactions on Power Electronics*, vol. 36, no. 3, pp. 2522–2532, 2020.

[7] D. Shi, P. Lin, Y. Wang, C.-C. Chu, Y. Xu, and P. Wang, "Deception attack detection of isolated dc microgrids under consensus-based distributed voltage control architecture," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 1, pp. 155–167, 2021.

[8] S. Sahoo, J. C.-H. Peng, A. Devakumar, S. Mishra, and T. Dragičević, "On detection of false data in cooperative dc microgrids—a discordant element approach," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6562–6571, 2019.

[9] X.-K. Liu, C. Wen, Q. Xu, and Y.-W. Wang, "Resilient control and analysis for dc microgrid system under dos and impulsive fdi attacks," *IEEE Transactions on Smart Grid*, 2021.

[10] A. Cecilia, S. Sahoo, T. Dragičević, R. Costa-Castelló, and F. Blaabjerg, "Detection and mitigation of false data in cooperative dc microgrids with unknown constant power loads," *IEEE Transactions on Power Electronics*, vol. 36, no. 8, pp. 9565–9577, 2021.

[11] M. S. Sadabadi, S. Sahoo, and F. Blaabjerg, "Stability oriented design of cyber attack resilient controllers for cooperative dc microgrids," *IEEE Transactions on Power Electronics*, 2021.

[12] Y. Jiang, Y. Yang, S.-C. Tan, and S. Y. Hui, "Distributed sliding mode observer-based secondary control for dc microgrids under cyber-attacks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 11, no. 1, pp. 144–154, 2020.

[13] M. R. Habibi, H. R. Baghaee, T. Dragičević, F. Blaabjerg, *et al.*, "Detection of false data injection cyber-attacks in dc microgrids based on recurrent neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2020.

[14] M. R. Habibi, S. Sahoo, S. Rivera, T. Dragičević, and F. Blaabjerg, "Decentralized coordinated cyber-attack detection and mitigation strategy in dc microgrids based on artificial neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 2021.

[15] O. A. Beg, L. V. Nguyen, T. T. Johnson, and A. Davoudi, "Signal temporal logic-based attack detection in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3585–3595, 2018.

[16] S. Jena, N. P. Padhy, and J. M. Guerrero, "Cyber-resilient cooperative control of dc microgrid clusters," *IEEE Systems Journal*, 2021.

[17] S. Sahoo, T. Dragičević, and F. Blaabjerg, "Resilient operation of heterogeneous sources in cooperative dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 12 601–12 605, 2020.

[18] S. Sahoo, T. Dragicevic, and F. Blaabjerg, "An event-driven resilient control strategy for dc microgrids," *IEEE Transactions on Power Electronics*, vol. 35, no. 12, pp. 13 714–13 724, 2020.

[19] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *arXiv preprint arXiv:1906.05799*, 2019.

[20] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, 2018.

[21] M. N. Kurt, O. Ogundijo, C. Li, and X. Wang, "Online cyber-attack detection in smart grid: A reinforcement learning approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5174–5185, 2018.

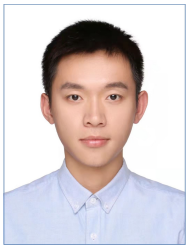
[22] W. Jiang, W. Yang, J. Zhou, W. Ding, Y. Luo, and Y. Liu, "Reinforcement learning based detection for state estimation under false data injection," *IEEE Access*, 2021.

[23] C. Neal, H. Dagdougui, A. Lodi, and J. M. Fernandez, "Reinforcement learning based penetration testing of a microgrid control algorithm," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2021, pp. 0038–0044.

[24] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, *et al.*, "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.

[25] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *arXiv preprint arXiv:1706.02275*, 2017.

Ali Jafarian Abianeh (S'19) received his M. Eng. degree in Electrical Engineering from University of Malaya, Kuala Lumpur, Malaysia, in 2010. He is currently working toward his Ph.D. degree at Department of Electrical and Computer Engineering at University of Louisiana at Lafayette, USA. He developed some solid professional expertise through several years of working in the industry as a power electronics engineer with the main focus on electric motor drives, and grid-tied power converters. His current research interests include application of advanced control algorithms and machine learning techniques to AC/DC microgrids, power converters, motor drive control, distributed control, fault tolerant control algorithms and cybersecurity.



Yihao Wan (S'21) received the B.S. degree in electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2017 and the M.S. degree in electrical engineering from Chongqing University, Chongqing, China, in 2020. He is currently pursuing the Ph.D. degree in electrical engineering with Technical University of Denmark. His current research interests include application of artificial intelligence in power electronics and power systems.

Farzad Ferdowsi (S'13–M'17-SM'20) is an Assistant Professor at University of Louisiana at Lafayette. He is with the Electrical and Computer Engineering Dept. Prior to joining UL Lafayette, Farzad worked as a research associate at the Center for Energy Studies at Louisiana State University. He received his Ph.D. from Florida State University in 2016. His research interests include power system stability and control and application of power electronic-based components in power systems.

Nenad Mijatovic after obtaining his Dipl.Ing. education in Electrical Power Engineering at University of Belgrade, Serbia in 2007, was enrolled as a doctoral candidate at Technical University of Denmark. He received his Ph.D. degree from Technical University of Denmark for his work on technical feasibility of novel machines and drives for wind industry. Upon completion of his PhD, he continued work within the field of wind turbine direct-drive concepts as an Industrial PostDoc. Dr. N. Mijatovic currently holds position of Associate Professor at Technical University of Denmark where he is in charge of managing research projects and education related to the field of electrical machines and drives, power electronic converters, motion control, application of energy storage and general applications of low frequency electromagnetism and large scale application of superconductivity with main focus on emerging eMobility and renewable energy generation.

He is a member of IEEE since 2008 and senior member of IEEE since 2018 and his field of interest and research includes novel electrical machine drives/actuator designs, operation, control and diagnostic of electromagnetic assemblies, advance control of drives and grid connected power electronics, energy storage and eMobility.

Tomislav Dragičević (S'09-M'13-SM'17) received the M.Sc. and the industrial Ph.D. degrees in Electrical Engineering from the Faculty of Electrical Engineering, University of Zagreb, Croatia, in 2009 and 2013, respectively. From 2013 until 2016 he has been a Postdoctoral researcher at Aalborg University, Denmark. From 2016 until 2020 he was an Associate Professor at Aalborg University, Denmark. From 2020 he is a Professor at the Technical University of Denmark. He made a guest professor stay at Nottingham University, UK during spring/summer of 2018. His research interest is application of advanced control, optimization and artificial intelligence inspired techniques to provide innovative and effective solutions to emerging challenges in design, control and cyber-security of power electronics intensive electrical distributions systems and microgrids. He has authored and co-authored more than 250 technical publications (more than 120 of them are published in international journals, mostly in IEEE), 8 book chapters and a book in the field.

He serves as an Associate Editor in the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, in IEEE TRANSACTIONS ON POWER ELECTRONICS, in IEEE Emerging and Selected Topics in Power Electronics and in IEEE Industrial Electronics Magazine. Dr. Dragičević is a recipient of the Končar prize for the best industrial PhD thesis in Croatia, a Robert Mayer Energy Conservation award, and he is a winner of an Alexander von Humboldt fellowship for experienced researchers.