A Reinforcement Learning Approach for Defending Against Multiscenario Load Redistribution Attacks

Jieyu Lei, *Student Member, IEEE*, Shibin Gao, Jian Shi, *Senior Member, IEEE*, Xiaoguang Wei, Ming Dong, *Senior Member, IEEE,* Wenshuang Wang, and Zhu Han, *Fellow, IEEE*

***Abstract*—The accelerated digitalization of today's electric power infrastructure has highlighted the gravity of cyber resilience in the electricity ecosystem. Due to the expanded attack surface inherited in the digitalized operational and information technologies, the electricity sector is facing a continually escalating cyberthreat landscape and must act to prepare for more frequent and sophisticated cyberattacks as the new normal. To aid in this effort, this paper presents a novel approach to identify critical branches to strengthen and by doing so, shield the smart-grid power system from the threat of load redistribution attacks (LRAs) under a wide range of operating scenarios. Compared with conventional critical branch identification approaches, we propose a new concept, namely, chain of defense, that empowers the system operator to incorporate existing cyber protections and develop branch strengthening strategy in a more dynamic, adaptive, and flexible way. We then propose a novel reinforcement learning framework to identify the most effective chain of defense. A cross-updating search strategy is developed to specifically ensure that the identified chain of defense can safeguard the network from the two-fold damaging effects of the LRAs on operation economics and security simultaneously. The effectiveness of the proposed approach is evaluated on the IEEE 14-bus system and the European 89-bus system. Simulation results indicate that the proposed approach provides more comprehensive and effective mitigation against the damaging effects of LRAs compared with the conventional approaches.**

***Index Terms*—False data attack, load redistribution attack, reinforcement learning, critical branch identification.**

Manuscript received 9 August 2021; revised 1 December 2021 and 5 April 2022; accepted 7 May 2022. Date of publication 17 May 2022; date of current version 23 August 2022. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2682021CG005 and Grant 2682021CX036; in part by the Key Projects of National Natural Science Foundation of China under Grant U1734202; in part by NSF under Grant CNS-2128368 and Grant CNS-2107216; in part by Toyota; and in part by Amazon. Paper no. TSG-01224-2021. *(Corresponding author: Xiaoguang Wei.)*

Jieyu Lei, Shibin Gao, and Xiaoguang Wei are with the School of Electrical Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: leijieyu_swjtu@126.com; gao_shi_bin@126.com; wei_xiaoguang@126.com).

Jian Shi is with the Department of Engineering Technology, University of Houston, Houston, TX 77004 USA (e-mail: jshi14@uh.edu).

Ming Dong is with the State Key Laboratory of Power Transmission Equipment and System Security and New Technology, Alberta Electric System Operator, Calgary, AB T2P 0L4, Canada (e-mail: dongming516@gmail.com).

Wenshuang Wang is with the Department of Mathematics, University of Houston, Houston, TX 77004 USA (e-mail: wwang51@uh.edu).

Zhu Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 446-701, South Korea (e-mail: zhan2@uh.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2022.3175470.

Digital Object Identifier 10.1109/TSG.2022.3175470

NOMENCLATURE

Sets

$\boldsymbol{B}$	Set of load buses;
$\boldsymbol{L}$	Set of branches;
$\boldsymbol{D}$	Set of existing/prior branches with enhancement;
$\boldsymbol{B}^+/\boldsymbol{B}^-$	Set of positive/negative nodes;
$\boldsymbol{V}_h$	Set of defense branches at stage h in the Markov decision process;
$\mathcal{S}, \mathcal{A}$	Set of states and actions in the Markov decision process, respectively.

Variables

K	Summation of net volume of false data injected to the system;
$\boldsymbol{A}(K)$	A multi-scenario attack vector whose net volume of false data injection is K;
$\boldsymbol{C}_k$	A defense chain, $\boldsymbol{C}_k = \{\ldots, L_i, \ldots\}, k = 1, \ldots, N_C$;
$\boldsymbol{VC}_o$	Vital defense chain, $o = 1, \ldots, N_v$;
$\boldsymbol{C}^*$	Most vital defense chain;
$\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{C}_k)$	Average damage after $\boldsymbol{C}_k$ is defended under the attack $\boldsymbol{A}(K)$;
s_h, a_h	State and action of the agent at stage h in the Markov decision process;
$r_h(s, a)$	Reward at stage h in the Markov decision process;
$Q(s_h, a_h)$	Q-value function at stage h;
$\boldsymbol{S}^*$	Load shedding;
$\boldsymbol{P}, \boldsymbol{R}$	Branch power flow vector and generator output vector;
$\Delta\boldsymbol{S}$, $\Delta\boldsymbol{P}$	False load and branch power measurement injections;
$\Delta S(B_j)$,	False data injection of load bus B_j;
$\Delta P(L_i)$	False data injection of branch L_i;
$\mu(L_i)$	Indicators. If L_i is defended, $\mu(L_i) = 1$; otherwise, $\mu(L_i) = 0$;
$\varphi_{\boldsymbol{a}_m(K)}$	The amount of operation cost or load shedding after $\boldsymbol{a}_m(K)$ is injected to the system;
$W(\boldsymbol{VC}_i)$	Weight of vital defense chain.

Parameters

$\boldsymbol{SF}$	Shifting factor matrix;
$\boldsymbol{KD}, \boldsymbol{KP}$	Load and generator incidence matrices, respectively;
$\boldsymbol{S}$	Load vector in the normal operation;

$\boldsymbol{c}_G, \boldsymbol{c}_D$	Vectors of generation cost and load shedding cost, respectively;
$\boldsymbol{R}_{\min}, \boldsymbol{R}_{\max}$	Minimum and maximum of generator outputs, respectively;
$\boldsymbol{P}_c$	Line capacity;
τ	Upper bound of false data injection for each load bus;
U	Length of the defense chain;
M	Total number of episodes;
δ, χ	Adjustable parameters, $\chi \in [0, 1], \delta \in [0, 1]$;
ω	A random number with [0,1];
γ	Discount factor in the Markov decision process;
υ	Penalty factor;
α	Learning rate;
ε	A small random number.

I. Introduction

With the increased utilization of cyber technologies in today's power grid, the power sector has become a primary target for cyberattacks in recent years. An expanding variety of threats and exploitations, such as cyber probes, data exfiltration, and malware plantations, have been made to compromise, disrupt, or even destroy the complex electricity transmission and distribution infrastructure [1]. As power systems must operate 24/7 with high reliability and availability, its condition monitoring and state estimation system has become one of the major functions susceptible to cyber-attacks. Cyber intruders can take advantage of the cyber vulnerabilities inherited in the measurement and communication modules, ranging from remote terminal units (RTUs) and smart meters, to the heterogeneous communication networks included in the supervisory control and data acquisition (SCADA) system, and inject malicious data to the authentic measurements of the system states to disturb the outcome of the state estimation. The falsified state estimation outcome would then induce the system operator in the control room to make erroneous or even harmful decisions on power dispatch. In addition to their potential high damaging impacts on the power grid, previous research has also recognized that these cyber-attacks, known as false data injection (FDI) attacks [2]–[4], can effectively bypass the conventional bad data detection (BDD) mechanism in the state estimation through carefully planned and executed cooperative measurements manipulation, making them undetectable [5]. Therefore, FDI attacks have been recognized as a major threat to the proper monitoring, management, and control of the electric grid.

In this paper, we consider a practical FDI attack, namely, *load redistribution attack* (LRA) based on the two practical constraints described in [6]–[8]. First, the attacker can only attack the load-bus and branch/line power measurements; the measurements of generator output, on the other hand, cannot be manipulated due to their advanced security settings. Secondly, the falsified injection at a load bus should stay within a reasonable range from its normal value to not draw attention of the system operator. As pointed out in [7], these two unique features have made LRAs more realistic than general FDI attacks.

To safeguard the system from LRAs and mitigate their adverse effects on system operation and security, it was first proposed in [9] that it is possible to prevent the system from entering the uneconomic and high-risk operation state by strengthening (i.e., defending) the measurement units for a carefully select set of "*critical branches*". Simulation results from [10]–[12] have shown that once the cyber security for these branches is strengthened and their measurement units can no longer be tampered with by the cyber intruder [13], the damaging effect of the LRA on the whole network will be effectively mitigated.

However, the identification of such critical branches is non-trivial. So far, existing literature mainly tackled the critical branch identification (CBI) problem from two different perspectives. The first perspective focuses on identifying a selected set (i.e., minimum set) of essential measurements that yield a unique solution to all state variables in the state estimation to make the power system "observable". To this end, greedy algorithms and strategies were proposed in [14] to select critical measurement units to protect, such that the false data injections cannot compromise the integrity of the state estimation. Similar efforts based on mixed-integer linear programming (MILP) models were presented in [15], [16] to determine the vulnerable measurements to protect under a limited budget. Once these selected measurement units are secured, no undetectable attacks can be formulated [17].

The secondary category of CBI approaches formulates an LRA as a max-min bi-level attacker-defender or a tri-level defender-attacker-defender problem, in which an "attacker" seeks to maximize the disruption penalty in the form of unmet demand or load shedding while a defender (i.e., system operator) attempts to react and minimize the disruptive effects of the attack [11], [12], [18]–[22]. The interactions between the attacker and defender can be described in two stages (i.e., attacker disrupts the network and defender reacts by re-dispatch) or three stages (i.e., defender deploys countermeasures considering potential attacking scenarios, attacker disrupts the network, and defender reacts by re-dispatch). This process can then be formulated and solved as a multi-level optimization problem to determine the most vulnerable system components to strengthen.

While the existing CBI models and algorithms offer valuable insights on understanding the attacking patterns of LRAs and providing effective countermeasures, two important challenges remain to be addressed. The first issue lies in the fact that in previous literature, critical measurements are primarily identified based on a particular attack vector that is constructed by the attacker to yield the worst disruption penalty under a single operating scenario. However, in practice, the system's operating status (i.e., generation and loading conditions) varies over time, which suggests that the system operator needs to take into account a variety of attack vectors associated with different system operating conditions to cover the possible grid scenarios under which an LRA may occur. This is especially true for today's power grid that involves a high penetration of renewable generation resources (e.g., wind and solar) [23], active loads, and intelligent demand-side management [24]. These smart features add

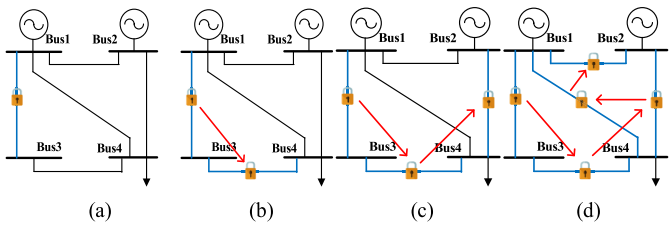

Fig. 1. Concept of chain of defense.

greater uncertainties to the power system operation due to their intermittency and lack of predictability. Therefore, the critical branches should also be identified considering a wide range of stochastic, and to some extent, unpredictable operating scenarios to provide comprehensive and consistent protections against LRAs. To this end, [25] incorporated the time-varying system operating conditions under which LRAs may occur by considering a set of arbitrarily selected representative operating scenarios. However, the inherent operational uncertainty and unpredictability associated with the LRA vectors were not systematically modeled and addressed in [25]. The proposed approach was still solved in a deterministic fashion under each representative operating scenario.

Another major challenge of the existing CBI approaches lies in the fact the critical branch/component identification is performed under two important assumptions. The first assumption is that the system under study has no existing cyber protection in place to safeguard the critical measurements from LRAs. In other words, the existing approaches are mostly developed based on a "greenfield" system model, without incorporating the prior security enhancement efforts. This assumption can be questionable to yield a meaningful solution in practice as the system would commonly have certain cyber security measures in place already. Such a "brownfield" model requires the CBI approach to be more flexible and more dynamic to support a smart grid that needs to be constantly evaluated and upgraded. The second limitation is that the existing CBI approaches require that critical branches, once identified, be strengthened simultaneously to achieve the expected mitigation effect. This assumption can make the existing approaches potentially ineffective when the system operator has limited defensive resources and is thus looking for a multi-stage cyber-security strengthening plan that aims to strengthen a subset of the critical branches at one time. Under such circumstances, the priority of the lines would change, and the existing CBI methods may no longer support the system operator's decision-making process.

To overcome the above limitations of the existing literature, we propose a novel mechanism, namely, *Chain of Defense*, to identify the critical branches in a more flexible and adaptive way. We take an example to explain the concept of Chain of Defense in Fig. 1. As shown in Fig. 1(a), we assume the system already has branches 1-3 strengthened as either a pre-existing condition or an initial part of a long-term infrastructure reinforcement plan, then the operator could establish a chain of defense by ranking the other branches in the system and identifying the most appropriate branch(es) to defend together with branch 1-3. Without loss of generality, we assume that the operator identifies branches 3-4 as the best (i.e., most critical) branch to strengthen as shown in Fig. 1(b). Similarly, if the defensive resource allows, the operator can further identify branch(es) to be strengthened together with branches 1-3 and 3-4, for instance, branch 2-4, and then strengthen it in Fig. 1(c). In this way, we can observe that the operator can establish a "chain" of branches to be strengthened for any arbitrary/given initial system defensive conditions as shown in Fig. 1(d). Fig. 1 demonstrates that different than the conventional CBI approaches, the proposed chain of defense concept allows the system operator to enhance the cyber-security of a power system more flexibly and dynamically.

However, the identification of the proposed chain of defense is not straightforward based on the discussion set forth above. On one hand, the proposed defense chain should be able to tackle the intrinsic uncertainty involved in the operating scenario under which an LRA may occur to determine the appropriate damaging effects of the attack. On the other hand, the proposed chain should support dynamic and flexible decision-making with the pre-existing cyber enhancement conditions taken into account.

To address this challenge, this paper presents a novel approach to identify the most vital chain of defense under *multi-scenario LRAs*. According to the unique characteristics of the problem, we formulate the identification process as a Markov decision process (MDP), which is then solved by using reinforcement learning (RL) [26]. We focus on addressing two of the most common damaging effects of the LRAs, operation economy and security. A cross-updating Q-learning strategy is developed to specifically ensure that the identified chain of defense can safeguard the power system from the two-fold damaging effects of the LRAs simultaneously without introducing additional computational burden. Simulation-based case studies showcase the unique features and strength of the proposed approach in safeguarding the system from multi-scenario LRAs in a more effective, flexible, and adaptive way.

The contributions of this paper are summarized as follows.

1. This paper proposes a novel concept, namely, chain of defense, to overcome two major limitations in the existing CBI literature and yield a more comprehensive, flexible, and adaptive mitigation strategy against LRAs.

2. This paper proposes an RL framework to identify the most effective chain of defense under a wide range of stochastic, and to some extent, unknown operating scenarios under which an LRA may occur. To the best of our knowledge, our work is the first of its kind.

3. Furthermore, this paper develops a Q-learning-based search strategy to simultaneously address two of the major threats posed by the LRAs: economy and security. A parallel updating mechanism is proposed to ensure the computational efficiency of this strategy.

The rest of the paper is organized as follows: we develop the multi-scenarios LRA model in Section II. Section III discusses the formulation and identification procedure of the proposed chain of defense. Case studies are performed in Section IV. Finally, the conclusions are drawn in Section V.

II. Multi-Scenario Attack Vector Model

In this section, we introduce the modeling strategy of the basic LRA, based on which the multi-scenario attack vector can be constructed.

A. LRA Modeling

The attack vector $\boldsymbol{a}$ that an attacker injects to the system measurements can be represented in the general form of $\boldsymbol{a} = [\Delta \boldsymbol{S},\ \Delta \boldsymbol{P}]^T$, where $\Delta \boldsymbol{S}$ and $\Delta \boldsymbol{P}$ denote the vectors of false load and branch power measurement injections, respectively. The cooperative manipulation of $\Delta \boldsymbol{S}$ and $\Delta \boldsymbol{P}$ needs to satisfy the basic power flow equation of:

$$\Delta \boldsymbol{P} = -\boldsymbol{SF} \cdot \boldsymbol{KD} \cdot \Delta \boldsymbol{S} \tag{1a}$$

where $\boldsymbol{SF}$ is the shifting factor matrix and $\boldsymbol{KD}$ is the load incidence matric. Both $\boldsymbol{SF}$ and $\boldsymbol{KD}$ can be derived from the Jacobian Matrix $\boldsymbol{H}$. Note that in cases where the attacker has no direct knowledge of $\boldsymbol{H}$, numerous data-driven approaches existed to estimate the topological information based on measurement data [27], [28]. The approximated $\boldsymbol{H}$ can then be used in (1a) to construct the attack vector. Equation (1a) ensures the false data injections can be hidden from the bad data detection. Since the generation data cannot be manipulated, $\Delta \boldsymbol{S}$ needs to satisfy (1b), i.e., the sum of false load measurements injected to all load buses is equal to zero to ensure the total system load remains unchanged following the attack to hide the attack from being detected:

$$\mathbf{1}^T \Delta \boldsymbol{S} = 0 \tag{1b}$$

Meanwhile, the falsified injection at load bus B_j should stay within a reasonable range from its normal value $S\ (B_j)$ to not draw attention from the system operator. This constraint can be represented as follows:

$$-\tau S\big(B_j\big) \le \Delta S\big(B_j\big) \le \tau S\big(B_j\big), \forall B_j \in \boldsymbol{B} \tag{1c}$$

The goal of the attack is then to construct an attack vector $\boldsymbol{a}$ to mislead the system operator (i.e., defender) and cause maximum disturbances to the optimal operation of the system after the control center implements the SCED solution made based on the contaminated system measurements as [6], [7]:

$$\max_{\Delta \boldsymbol{S}} \ \boldsymbol{c}_{\boldsymbol{G}}^T \boldsymbol{R} + \boldsymbol{c}_{\boldsymbol{D}}^T \boldsymbol{S}^* \tag{1d}$$

Meanwhile, the defender, represented by the SCED function, aims to minimize the damaging effects of the LRA subject to a set of constraints as follows:

$$\min_{\Delta \boldsymbol{S}} \ \boldsymbol{c}_{\boldsymbol{G}}^T \boldsymbol{R} + \boldsymbol{c}_{\boldsymbol{D}}^T \boldsymbol{S}^* \tag{2a}$$

$$\min_{\Delta \boldsymbol{S}} \ 1^T \boldsymbol{S}^* \tag{2b}$$

$$s.t. \ \ 1^T \boldsymbol{R} = 1^T \big(\boldsymbol{S} - \boldsymbol{S}^*\big) \tag{2c}$$

$$\boldsymbol{P} = \boldsymbol{SF} \cdot \boldsymbol{KP} \cdot \boldsymbol{R} - \boldsymbol{SF} \cdot \boldsymbol{KD} \cdot \big(\boldsymbol{S} + \Delta \boldsymbol{S} - \boldsymbol{S}^*\big) \tag{2d}$$

$$\boldsymbol{R}_{\min} \le \boldsymbol{R} \le \boldsymbol{R}_{\max} \tag{2e}$$

$$-\boldsymbol{P}_c \le \boldsymbol{P} \le \boldsymbol{P}_c \tag{2f}$$

$$0 \le \boldsymbol{S}^* \le \boldsymbol{S} + \Delta \boldsymbol{S} \tag{2g}$$

In the above SCED function, the objective function is to minimize the operation cost (OC) in (2a) or load shedding (LS) amount in (2b) based on the manipulated data $\boldsymbol{S} + \Delta \boldsymbol{S}$, subject to the power balance constraints (2c), branch flow equation (2d), generator output limits (2e), branch flow limits (2f), and load shedding limits (2g) [6], [7].

In this way, an LRA can be formulated as a bilevel linear programming model in which the upper-level model (1a)-(1d) represents the attacker, and the lower-level model (2a)-(2g) represents the defender. Note that the above LRA modeling strategy has been well studied and rigorously examined in previous literature, such as in [11], [12], [18]–[22].

B. Multi-Scenario Attack Vector Construction

To set the stage for the rest of the discussion, we introduce the following naming convention: for load-bus B_j, if the false data injection $\Delta S(B_j) > 0$, we name B_j a "*positive node*", denoted by B_j^+. If $\Delta S(B_j) < 0$, we name load-bus B_j a "*negative node*", denoted by B_j^-. According to (2a), the sum of the false data injections at all positive and negative nodes needs to satisfy:

$$\sum_{j_1=1}^{N^+} \Delta S\Big(B_{j_1}^+\Big) = -\sum_{j_2=1}^{N^-} \Delta S\Big(B_{j_2}^-\Big) = K \tag{3}$$

where N^+ and N^- are the number of positive and negative nodes, respectively. K represents the summation of the "net" volume of false data injected into the system in an attack. To cover all the different attacking scenarios under different system operating conditions, we employ a random method to determine the value of K. First, we randomly divide the set of load-buses $\boldsymbol{B}$ into the set of positive nodes $\boldsymbol{B}^+ (\boldsymbol{B}^+ \subseteq \boldsymbol{B})$ and the set of negative nodes $\boldsymbol{B}^- (\boldsymbol{B}^- \subseteq \boldsymbol{B})$. Then, we can have the following:

$$\Delta S_{\max}^+\big(\boldsymbol{B}^+\big) = \tau \sum_{j_1=1}^{N^+} S\Big(B_{j_1}^+\Big), \forall B_{j_1}^+ \in \boldsymbol{B}^+ \tag{4a}$$

$$\Delta S_{\max}^-\big(\boldsymbol{B}^-\big) = -\tau \sum_{j_2=1}^{N^-} S\Big(B_{j_2}^-\Big), \forall B_{j_1}^- \in \boldsymbol{B}^- \tag{4b}$$

where (4a) and (4b) denote the maximum volume of false data injections at all positive nodes and all negative nodes, respectively. Then, the maximum of K for the sets $\{\boldsymbol{B}^+, \boldsymbol{B}^-\}$ can be determined as:

$$K_{\max} = \min\big\{\Delta \boldsymbol{S}_{\max}^-, \Delta \boldsymbol{S}_{\max}^+\big\} \tag{4c}$$

To assure the difference between $\Delta S_{\max}^-$ and $\Delta S_{\max}^+$ is reasonable to yield an appropriate $K_{\max}$, we introduce the following constraint:

$$K_{\max} \ge \frac{1}{2}\big(\Delta \boldsymbol{S}_{\max}^+ + \Delta \boldsymbol{S}_{\max}^-\big)\delta \tag{4d}$$

where δ is an arbitrary number and $\delta \in [0, 1]$. (4d) ensures that if the derived $K_{\max}$ is lower than a certain threshold, a new $K_{\max}$ can be generated based on (4a)-(4c). This threshold is controlled by the selection of δ. The larger δ is, the more false data can be injected into the system, which represents a

potentially more severe attack. We then employ (5) to select the value of K.

$$K = K_{\max}(\chi + (1-\chi)\omega) \tag{5}$$

where $\omega \in [0,1]$ is a random number, $\chi \in [0,1]$is an adjustable parameter that ensures that K is greater than or equal to $K_{\max}\chi$. Therefore, χ in (5) controls the lower limit of the net volume of false data injection in each attack.

After determining the value of K, we can construct a random attack vector by solving a linear programming problem with a dummy objective function as shown below:

$$\min_{\Delta S\left(B_{j1}^{+}\right),\Delta S\left(B_{j2}^{-}\right)} \quad 1 \tag{6a}$$

$$\text{subject to:} \quad (3)-(5) \tag{6b}$$

Constraints (3)-(5) ensures that the attack vector satisfies the practical constraints and will result in a net volume of false data injection of K. In this way, the multi-scenario attack vector, denoted by $\boldsymbol{A}(K)$, can be formulated as:

$$\boldsymbol{A}(K) = \{\boldsymbol{a}_m(K) | m = 1, 2, \ldots, M(K)\} \tag{7}$$

where $\boldsymbol{a}_m(K)$ represents the m-th attack vector whose net volume of false data injection is K, and $M(K)$ is the number of attack vectors.

III. Chain of Defense Formulation and Identification

A. Incorporation of the Defense Strategy

One of the most effective defensive strategies against LRAs is to limit the ability of the attacker to inject false data into critical measurements and by doing so, shield the system from the adverse effects of LRAs. More specifically, the defender can deploy three categories of enhancements to improve the cyber security of the measurement units and their communication links involved in the heterogeneous SCADA network [9], [10]: 1) enhance the measurement device with tamper-resistant hardware and tamper alarms to prevent the potential physical manipulations; 2) deploy advanced encryption, authentication, and validation techniques to protect the confidentiality and integrity of the measurement data; and 3) adopt intrusion detection and prevention modules to detect and preventively react to the malicious data operations. In the following analysis, we assume that once a branch is strengthened, its measurement data is sufficiently secured and can no longer be accessed /manipulated by the attacker. We further assume that the system operator has limited defensive resources and would thus only be able to strengthen a selected number of branches that forms a defense chain, denoted by $\boldsymbol{C}_k$ ($\boldsymbol{C}_k \subseteq \boldsymbol{L}$, where $\boldsymbol{L}$ is the set of all branches in the network), to protect the network from the LRA.

Based on the discussion above, if a branch L_i ($L_i \in \boldsymbol{C}_k$) is strengthened, the attacker will no longer be able to inject any false data into its measurement as shown in (8):

$$|\Delta P(L_i)| = 0 (\forall L_i \in \boldsymbol{C}_k) \tag{8}$$

Equation (8) can be seen as the *defensive action* taken by the system operator in the face of an LRA. Incorporating (8) into the attack model in (6), we can now evaluate the impacts of the LRA model with the defender's strategy taken into account in the form of:

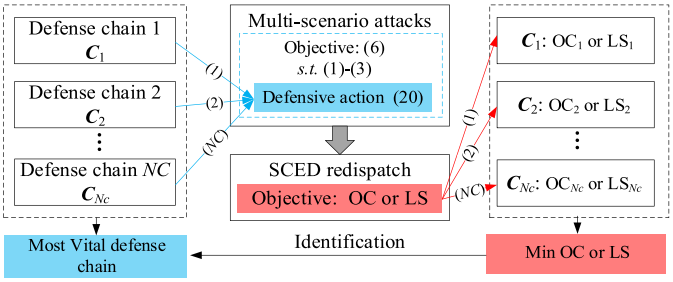

Fig. 2. Chain of defense identification under multi-scenario LRAs.

$$\text{objective:} \quad (6a) \tag{9a}$$

$$\text{subject to:} \quad (3)-(5) \text{ and } (8) \tag{9b}$$

It is evident that the defense effectiveness of strengthening a particular chain $\boldsymbol{C}_k$ can be evaluated by solving the optimization model described by (9). The resulted multi-scenario attack vectors $\boldsymbol{A}(K)$ can be fed into a control center model represented by SCED in (2c)-(2g) to obtain the specific response from the system operator after each attack vector contained in $\boldsymbol{A}(K)$ is injected to the system, respectively. Then, we can examine whether the misled response can lead the system into a non-optimal/insecure operating state.

B. Markov Decision Process Model

Based on the SCED model in (2c)-(2g), it is evident that we can evaluate the damaging effects of the multi-scenario LRA from two perspectives: economy, i.e., minimizing OC using (2a) and security, i.e., minimizing LS using (2b). We employ $\bar{\varphi}_{\boldsymbol{A}(K)}$ in (10) to represent the average damage of multi-scenario LRAs in terms of OC and LS:

$$\bar{\varphi}_{\boldsymbol{A}(K)} = \frac{1}{M(K)} \sum_{m=1}^{M(K)} \varphi_{\boldsymbol{a}_m(K)}, \forall \boldsymbol{a}_m(K) \in \boldsymbol{A}(K) \tag{10}$$

where $\varphi_{\boldsymbol{a}_m(K)}$ denotes the amount of OC or LS obtained after $\boldsymbol{a}_m(K)$ is injected into the control center modeled by the SCED function. A lower $\bar{\varphi}_{\boldsymbol{A}(K)}$ indicates overall better defense effectiveness. To capture the most *vital chain* $\boldsymbol{C}^*$ that leads to the best defense effectiveness among all possible defense chains, the selection of $\boldsymbol{C}^*$ needs to satisfy (11):

$$\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{C}^*) = \underset{\boldsymbol{C}_1,\ldots,\boldsymbol{C}_{N_C}}{\arg\min} \left\{\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{C}_1), \ldots, \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{C}_{N_C})\right\} \tag{11}$$

where N_c is the total number of chains that could be potentially formulated. $\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{C}_k)$ ($k = 1, \ldots, N_c$) represents the average value of OC or LS after $\boldsymbol{C}_k$ is defended under multi-scenario LRAs. Incorporating the aforementioned process, the identification process of $\boldsymbol{C}^*$ is shown in Fig. 2.

While the enumeration method can be applied to determine N_c and solve (11) and identify the most vital chain $\boldsymbol{C}^*$, it is based on brutal force and may lead to a tremendous computational burden. For example, for an electrical network consisting of 20 branches, to enumerate all possible defense chains with a length of 5, we need to formulate and examine a

total number of $N_c = 15504$ chains (i.e., C_{20}^5). This is clearly infeasible for a practical scale power network.

To overcome this challenge, we propose to formulate the defense chain identification problem as a Markov decision process (MDP). As illustrated below, we intend to employ MDP to formulate a number of vital chains, i.e., $\boldsymbol{VC}_1, \ldots, \boldsymbol{VC}_{Nv}$, where Nv is the number of vital chains, and then identify the most vital chain $\boldsymbol{C}^*$ as an approximation to the solution to (11). Under this setting, the system operator can be seen as a virtual *agent* who aims to enhance the system's cyber security through a series of branch strengthening actions. Assume that the agent has limited defensive resource and is capable of defending a vital defense chain that consists of U branches, the agent then need to interact with the *environment* sequentially U times and select a branch from the set of candidate branches at each stage $h (h = 1, \ldots, U)$ of the interaction to formulate the chain. More specifically, we denote the agent's decision-making process in a four-tuple as follows:

State $\mathcal{S}$: From the agent's perspective, the state of an electrical network can be viewed as a combination of its branches' security enhancement status. Hence, we denote the state of a branch by a binary variable: *defended* or *undefended*. In this way, the state of the entire network can be described by a series of binary indicators. Specifically, we define state $s_h \in \mathcal{S}$ at stage h, for $h = 1, \ldots, U$ as follows:

$$s_h = \{\mu(L_1), \ldots, \mu(L_i), \ldots, \mu(L_U)\}|_h \tag{12}$$

where indicator $\mu(L_i) = 1$ if branch L_i is defended; otherwise, $\mu(L_i) = 0$. It is evident based on (12) that the state space $\mathcal{S}$ is finite.

Action $\mathcal{A}$: At each stage, the agent selects a branch and adds it to the defense chain. More specifically, if an undefended branch L_i is selected by an action a_h of the agent, where $a_h \in \mathcal{A}$, $\mu(L_i)$ becomes 1, and branch L_i will be added as the h-th component of the defense chain. As a branch cannot be defended twice, once a branch is selected, it is marked as *unavailable* and cannot be selected again by future actions. It is thus evident that the action space $\mathcal{A}$ is also finite.

Transition probability: The dynamic of the system is governed by the transition probability $\Pr(s_{h+1}|a_h, s_h)$, which represents the state transition from the current state s_h to the next state s_{h+1} when an action a_h is taken at stage h. As defined in (13), during the identification process, the action a_h has a probability $\gamma (\gamma \leq 1)$, namely, *discounted factor*, of transitioning the system from state s_h to a new state s_{h+1}. The system also has a $(1-\gamma)$ probability to enter an "*absorbing*" state with no available actions, which suggests: 1) we have reached the length of the vital chain; or 2) no valid attack vector can be constructed following action a_h.

$$\begin{cases} \Pr(s_{h+1}|a_h, s_h) = \gamma \\ \Pr(s_f|a_h, s_h) = 1 - \gamma. \end{cases} \tag{13}$$

Reward: As described in (13), a deterministic reward $r_h(s, a)$ is associated with the state-action pair (s, a) at stage h:

$$r_h(s,a) = \begin{cases} (\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_{h-1}) - \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h))e^{-h}, \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_{h-1}) \geq \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h) \\ \upsilon(\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_{h-1}) - \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h))e^{h-C}, \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_{h-1}) \leq \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h) \end{cases} \tag{14}$$

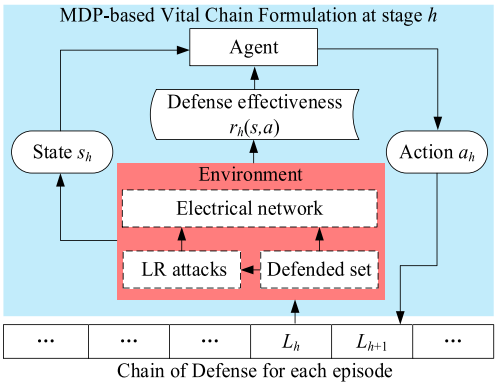

Fig. 3. MDP-based chain of defense identification.

where the term $\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_{h-1}) - \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h)$ represents the difference in damaging effect after a branch is added to the current chain. Therefore, if $\bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_{h-1}) \geq \bar{\varphi}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h)$, i.e., the damaging effect of the LRA is weakened after defending this particular branch, then $r_h(s, a)$ is a positive reward. On the other hand, if $r_h(s, a)$ is negative, it suggests that the damaging effect would intensify after defending this branch. In addition, we set up a coefficient e^{-h} to provide a higher positive reward if an effective defense action can be incorporated earlier in the identification process; on the contrary, if the reward is negative, we set up the term e^{h-C} to punish the agent if an incorrect action is selected at the late stage of the identification process. Note that in (13), υ ($\upsilon \geq 1$) is a *penalty factor* to adjust the severity of the negative reward. A greater υ would make the negative reward linked with an undesired action more significant, so that this particular action can be chosen with a lower probability.

According to the above description, the MDP-based vital chain formulation process, as depicted in Fig. 3, can be specified as follows: at each stage $h (h = 1, \ldots, U)$ of the interaction, the agent selects an action (defending a certain branch) a_h from the set of candidate actions and then calculate the defense effectiveness (i.e., reward) $r_h(s, a)$ based on the current state s_h to measure how effective the action is. Once an action is selected and added to the chain, the system state will be updated, and we will move on to the next stage $h + 1$. This process is repeated until the last stage U, which we call an *episode*. We aim to maximize the following *value function* as the accumulated expected reward during the entire episode as shown in (15):

$$\max \mathrm{E}\left(\sum_{h=1}^{C} \gamma^h r_h(s,a)\right) \tag{15}$$

where E(.) denotes the expected value. Note that in (15), with h increasing, i.e., more branches added to the chain, $\gamma^h r_h(s, a)$ gets smaller as the setting of γ^h helps the agent identify branches with better defense effectiveness at earlier stages. It also facilitates the convergence of (15). In this way, (15) ensures that the system operator selects actions that lead to the most effective mitigation against the damaging effects of the multi-scenario LRA when formulating a vital chain.

C. RL-Based Defense Chain Formulation

While (14) can be solved by enumerating approaches, such as dynamic programming, it can be computationally difficult and even intractable considering the number of branches and states involved. Therefore, we propose a reinforcement learning (RL)-approach, based on the classic Q-learning algorithm, to approximate the solution to (15) iteratively and incrementally [29]. The RL algorithm aims to find an optimal policy for the agent by updating the *Q-value functions* based on the Bellman equation [30] as follows:

$$Q(s_h, a_h) \leftarrow (1-\alpha)Q(s_h, a_h) + \alpha\left(r(s_h, a_h) + \gamma \max_{a_{h+1}\in\mathcal{A}} Q(s_{h+1}, a_{h+1})\right) \quad (16)$$

where $Q(s_h, a_h)$ is the Q-value function for the state-action pair (s,a) at stage h. In (16), α represents the learning rate that controls the aggressiveness of learning. A rate of 0 makes the agent extremely conservative who sticks to its initial estimate and learns nothing from its actions, thus the Q-value never updates, while a rate of 1 means the agent considers only the most recent information, which leads to fast learning but loses the previously learned knowledge. It is evident that (16) is a self-updating process and can be solved iteratively.

Double Q-value functions: Since the damaging effect of the LRAs are essentially two-fold (OC and LS), it is natural for us to define two sets of r and Q which would allow us to formulate two vital chains from the perspectives of economics and security, respectively. The proposed Q-value functions and the associated updating strategies are as follows:

(a) The OC-based Q-value function:

$$Q_1(s_h, a_h) \leftarrow (1-\alpha)Q_1(s_h, a_h) + \alpha\begin{pmatrix} r_{h,1}(s,a) + \gamma_2 Q_2(s_{h+1}, a_1) \\ +\gamma_1 \max_{a\in\mathcal{A}} Q_1(s_{h+1}, a_1)\end{pmatrix} \quad (17a)$$

$$a_1 = \underset{a\in\mathcal{A}}{\operatorname{argmax}} Q_1(s_{h+1}, a). \quad (17b)$$

(b) The LS-based Q-value function:

$$Q_2(s_h, a_h) \leftarrow (1-\alpha)Q_2(s_h, a_h) + \alpha\begin{pmatrix} r_{h,2}(s,a) + \gamma_2 Q_1(s_{h+1}, a_2) \\ +\gamma_1 \max_{a\in\mathcal{A}} Q_2(s_{h+1}, a_2)\end{pmatrix} \quad (18a)$$

$$a_2 = \underset{a\in\mathcal{A}}{\operatorname{argmax}} Q_2(s_{h+1}, a) \quad (18b)$$

where Q_1 and Q_2 represent the OC-based and LS-based Q-value functions, respectively. This "cross-updating" process ensures that the agent will not blindly pursue economy or security at the expense of the other.

As revealed in (17) and (18), the agent needs to take two different actions to update the OC- and LS-based Q-value functions at each stage separately, resulting in higher computational complexity. In fact, as shown in the SCED model (2a) and (2b), the damaging effect of LS is already taken into consideration when calculating OC. This inspires us to develop a parallel strategy to simultaneously calculate the OC- and LS-based rewards of an action. More specially,

(a) When updating OC-based Q-value function (17) to formulate the OC-based chain, the agent can update the LS-based Q-value function simultaneously as follows:

$$Q_2(s_h, a_h) \leftarrow (1-\alpha)Q_2(s_h, a_h) + \alpha\begin{pmatrix} r_{h,2}(s,a) + \gamma_1 Q_2(s_{h+1}, a_1) \\ +\gamma_2 \max_{a\in\mathcal{A}} Q_1(s_{h+1}, a)\end{pmatrix} \quad (19a)$$

$$a_1 = \underset{a\in\mathcal{A}}{\operatorname{argmax}} Q_1(s_{h+1}, a). \quad (19b)$$

(b) Similarly, when updating LS-based Q-value function (18) to formulate the LS-based chain, the agent can update the OC-based Q-value function simultaneously as follows:

$$Q_1(s_h, a_h) \leftarrow (1-\alpha)Q_1(s_h, a_h) + \alpha\begin{pmatrix} r_{h,1}(s,a) + \gamma_1 Q_1(s_{h+1}, a_2) \\ +\gamma_2 \max_{a\in\mathcal{A}} Q_2(s_{h+1}, a)\end{pmatrix} \quad (20a)$$

$$a_2 = \underset{a\in\mathcal{A}}{\operatorname{argmax}} Q_2(s_{h+1}, a) \quad (20b)$$

By combining (17) and (18) with (19) and (20), we can formulate the OC- and LS-based vital chains simultaneously. Note that while the proposed parallel updating strategy involves the interactions of two Q-value functions, it is still based on Q-learning algorithm.

Action selection: When taking actions, if the agent only chooses the action to optimize the next immediate reward, then the identification process may get stuck in a local optimum. To avoid such circumstances, the algorithm must carefully balance *exploitation* and *exploration*. More specifically, the agent performs exploration of unexplored search space looking for solutions potentially beneficial in the long run whereas the exploitation is done by performing greedy actions to get the most reward with a larger Q-value in the current state. In this paper, we adopt the ε-greedy method to balance exploration and exploitation, where ε refers to a small probability of the agent choosing to explore new non-optimal actions. The probability of action a taken by the agent in state s_h should satisfy:

$$\Pr(a) = \begin{cases} 1-\varepsilon, a = \operatorname{argmax}_{a\in\mathcal{A}} Q(s_{h+1}, a) \\ \varepsilon/(|\mathcal{A}|-1), a \neq \operatorname{argmax}_{a\in\mathcal{A}} Q(s_{h+1}, a)\end{cases} \quad (21)$$

Then the ε-greedy method can be implemented as follows: we first generate a random number ε'. If $\varepsilon' \in [\varepsilon, 1]$, the agent selects the best action according to Q-value at the next state, i.e., $a = \operatorname{argmax}_{a\in\mathcal{A}} Q(s_{h+1}, a)$; if $\varepsilon' \in [0, \varepsilon)$, the agent selects the action randomly with a probability of $\varepsilon/(|\mathcal{A}|-1)$.

Based on the discussion above, the proposed RL-based vital chain formulation process is described in Algorithm 1.

Furthermore, to improve the search efficiency of the agent, we propose the following two acceleration techniques.

(a) When the agent is taking action a_t in state s_t, if the average OC reward $\bar{\varphi}^{OC}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h)$ and the average LS reward $\bar{\varphi}^{LS}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h)$ satisfy (22), it indicates that the defense effectiveness is ideal. Therefore, the agent terminates the current episode and moves to the next:

$$\left|\bar{\varphi}^{OC}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h) - \varphi_0^{OC}\right| \le \varepsilon_1 \text{ and } \left|\bar{\varphi}^{LS}_{\boldsymbol{A}(K)}(\boldsymbol{V}_h) - \varphi_0^{LS}\right| \le \varepsilon_2 \quad (22)$$

where φ_0^{OC}andφ_0^{LS} denote the OC and LS amounts at the normal operating state, respectively. ε_1 and ε_2 are small numbers.

Algorithm 1 RL-Based Vital Chain Formulation

Input: Initialize the OC-based Q-matrix $\boldsymbol{Q}_1 = \boldsymbol{0}$, LS-based Q-matrix $\boldsymbol{Q}_2 = \boldsymbol{0}$ and the number of episodes M

Output: $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$

While $e \leq M$ (for each episode)
  Calculate the summation of injection load K;
  Select an initial state s_0 randomly;
  For $h = 1 : U$ (for each step in the e-th episode)
    Calculate $\boldsymbol{A}(K)$ according to (6) and (7)
    Formulate the OC-based vital chain
    Choose an OC-based action a_h in state s_h using (21)
    Calculate $r_{h,1}(s, a)$ and $r_{h,2}(s, a)$ based on SCED using (14)
    Update $Q_1(s_h, a_h)$ and $Q_2(s_h, a_h)$ using (17) and (19)
    Go to the next stage s_{h+1};
    Formulate the LS-based vital chain
    Choose an LS-based action a_h in state s_h using (21)
    Calculate $r_{h,1}(s, a)$ and $r_{h,2}(s, a)$ using SCED using (14)
    Update $Q_1(s_h, a_h)$ and $Q_2(s_h, a_h)$ using (18) and (20)
    Go to the next stage s_{h+1};
  End For
End While

(b) When the agent is taking action a_t in state s_t, if no valid attack vector can be constructed according to (6), i.e., $\boldsymbol{A}(K) = \emptyset$, the agent terminates the current episode and move to the next episode.

D. Optimal Chain of Defense Identification

According to Algorithm 1, we can obtain a comprehensive collection of Q-value functions after M episodes, based on which we can identify the most vital chain to defend. More specifically, given any arbitrary branch L_i in the system, we can use it as the initial branch and select the best branch to defend sequentially with L_i, namely, L_{i+1}, that results in the maximum Q-value, i.e., maximum defensive effectiveness as shown in (23):

$$L_{i+1} = \arg\max_{L_{i+1} \in \boldsymbol{L}} Q(L_i, L_{i+1}) \tag{23}$$

By extending (23), we can identify components for the rest of the vital chain $\boldsymbol{VC}_i$, i.e., $L_{i+2}, \ldots, L_{i+U-1}$, for the initial branch L_i.

Furthermore, based on to the Q-values, we define the weight $W(\boldsymbol{VC}_i)$ of vital chain $\boldsymbol{VC}_i$ as the sum of the Q-values for all the chain components as shown in (24). A higher weight indicates that the associated chain has better performances in terms of defense effectiveness.

$$W(\boldsymbol{VC}_i) = \sum_{i=1}^{U} Q(L_i, L_{i+1}) \tag{24}$$

Based on (24), we can make the following two important rules regarding the identification of the most vital chain.

Rule 1: When the system has no existing/prior branch enhancement: for a network consisting of N_L branches, we can take each branch as the initial branch to find N_L vital chain $\boldsymbol{VC}_1, \ldots, \boldsymbol{VC}_{NL}$. We can then count how often an ordered pair of branches $\{L_i, L_j\}$, i.e., a sequence $L_i \rightarrow L_j$, shows up in all the formulated vital chains. A higher frequency indicates the ordered pair of $L_i \rightarrow L_j$ is more significant in terms of establishing the defense effectiveness despite the initial conditions. Hence, we can select these branches to construct the most vital chain $\boldsymbol{C}^*$.

Rule 2: When a system has a set of existing/prior branches $\boldsymbol{D} = \{L_1, L_2, \ldots, L_v\}$ with enhancement, where v denotes the length of $\boldsymbol{D}$, we propose the following two-step procedure: first, we can take each branch in $\boldsymbol{D}$ as the initial branch and rank the other branches included in $\boldsymbol{D}$ according to their Q values to formulate v preliminary vital chains with an initial length of v. According to (24), we can rank and select the preliminary chain with the maximum weight. For instance, a system has three branches L_j, L_k, L_m with existing enhancement, i.e., $\boldsymbol{D} = \{L_j, L_k, L_m\}$, and their Q-values are $Q(L_j, L_k) = 10$, $Q(L_j, L_m) = 15$, $Q(L_k, L_j) = 12$, $Q(L_k, L_m) = 21$, $Q(L_m, L_j) = 16$ and $Q(L_m, L_k) = 6$. We first take L_j as the initial branch to construct a preliminary vital chain. As $Q(L_j, L_m)$ is larger than $Q(L_j, L_k)$, we will select L_m as the second component of the preliminary vital chain. Therefore, we can obtain the preliminary vital chain as $L_j \rightarrow L_m \rightarrow L_k$ with a total weight of 21. Similarly, we can take branch L_m and L_k as the initial branch respectively and formulate the following two preliminary vital chains: $L_k \rightarrow L_m \rightarrow L_j$ with a total weight of 37 and $L_m \rightarrow L_j \rightarrow L_k$ with a total weight of 26. According to their weights, we can select $L_k \rightarrow L_m \rightarrow L_j$ as the most vital preliminary chain with all L_j, L_k, and L_m in it. In the following step, based on the available defensive resource, the defender can add new branches into the most vital preliminary chain identified in the previous step according to (24).

IV. Case Study

In this section, we evaluate the performance of the proposed approach based on the IEEE 14-bus system and the European 89-bus system provided in MatPower [31], respectively. To simulate the large variety of operating conditions that an LRA may occur under, we consider a load variation range of $-30\% \sim 30\%$ based on the default load profile. The parameters used in the case studies are set up as follows: $\tau = 0.5$, $\delta = 0.8$, $\chi = 0.8$, $\gamma_1 = 0.8$, $\gamma_2 = 0.3$, $\alpha = 0.5$, and $\upsilon = 3$. The following simulation results were obtained in MATLAB on a laptop, which was equipped with an Intel i5-7200U CPU @2.50 GHz and 8.00 GB RAM. For the IEEE 14-bus system, it takes roughly 5 hours to complete the training. For the European 89-bus system, the training time is roughly 12 hours. Note that while the computation time increases as the size of the system grow, since the proposed approach is performed in an off-line fashion, the computational overhead is not considered a primary performance indicator of the proposed approach.

A. Training Curve of the Agent

To train the agent, we use 2000 episodes for the IEEE 14-bus system and 8000 episodes for the European 89-bus

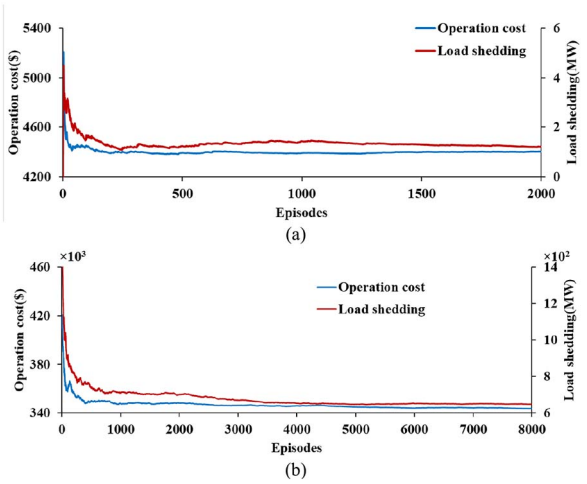

Fig. 4. Training curve of the proposed Q-learning algorithm for (a) IEEE 14-bus system; (b) European 89-bus system.

TABLE I
OC-BASED VITAL CHAINS IN THE IEEE 14-BUS SYSTEM

Start	Vital chains	Weights	Start	Vital chains	Weights
9→	13→7→11→14	75.59	7→	11→14→6→16	69.05
13→	7→11→14→6	74.43	4→	17→16→10→3	67.98
20→	11→14→6→16	74.02	17→	16→16→10→3	67.76
3→	5→11→14→6	70.39	10→	3→5→11→14	67.75
2→	3→5→11→14	70.06	5→	11→14→6→16	65.58

system, respectively. The obtained training curve in terms of average operation cost and load shedding for these two systems are shown in Fig. 4. We can clearly obverse that the agent is able to gradually improve its policy in terms of mitigating the damaging effects of the multi-scenario LRAs with increasing episodes of learning.

B. Vital Defense Chain for the IEEE 14-Bus System

We first formulate the OC- and LS-based vital chains for the IEEE 14-bus system. Without loss of generality, we assume that the defender has the defensive resources to defend 5 branches, i.e., the length of the vital defense chain is 5. The results of this analysis are shown in Tables I and II. Note that for the sake of brevity, only the top 10 vital chains with the highest weights are presented. It can be observed that as discussed in the previous section, certain sequences, such as $11 \rightarrow 14$, $14 \rightarrow 6$, $6 \rightarrow 16$, $16 \rightarrow 10$, $10 \rightarrow 3$, and $3 \rightarrow 5$, are more frequently included in the identified OC-based vital chains. Similarly, sequences such as $17 \rightarrow 12$, $12 \rightarrow 16$, $9 \rightarrow 14$, and $16 \rightarrow 10$ are more frequently included in the LS-based vital chains. We can also observe that the weights of the top LS-based vital chains are greater than the top OC-based vital chains. Therefore, we can estimate that LS-based vital chains have better defensive effectiveness, i.e., the system operator can better mitigate the load shedding effect of the multi-scenario LRAs via strengthening the identified vital chains.

Under the assumption that there is no existing branch enhancement in the system, we can record the frequency of each sequence included in the formulated vital chains.

TABLE II
LS-BASED VITAL CHAINS IN THE IEEE 14-BUS SYSTEM

Start	Vital chains	Weights	Start	Vital chains	Weights
4→	3→13→11→9	145.66	8→	7→12→16→9	108.54
20→	17→12→16→9	126.05	7→	12→16→9→14	106.52
5→	17→12→16→9	119.61	3→	13→11→9→14	104.49
17→	12→16→9→14	114.32	12→	16→9→14→13	103.36
19→	12→16→9→14	110.46	1→	3→13→11→9	103.33

TABLE III
MOST VITAL CHAIN IDENTIFIED BASED ON DIFFERENT INITIAL ENHANCEMENT CONDITIONS

Branches with existing enhancement	OC	LS
	Most vital chain	Most vital chain
16,17	17→16→10→3→5	17→16→9→14→13
12,13	12→13→7→11→14	12→13→11→9→14
8,15	8→15→15→11→14	8→15→16→9→14
11,20,14	20→11→14→6→16	20→14→11→9→14
3,5,6	3→5→6→16→10	6→5→3→13→11

For OC-based vital chains, sequence $11 \rightarrow 14$ is included 11 times (i.e., in 11 formulated vital chains), while sequences $10 \rightarrow 3$ and $16 \rightarrow 10$ are included 10 times. Sequences $14 \rightarrow 6$ and $6 \rightarrow 16$ are included 9 and 8 times, respectively. Therefore, we can construct the most vital OC-based chain as $11 \rightarrow 14 \rightarrow 6 \rightarrow 16 \rightarrow 10$. The system operator can select these branches for sequential strengthening to reduce the OC loss under multi-scenario LRAs. A close alternative OC-based defense chain is $14 \rightarrow 6 \rightarrow 16 \rightarrow 10 \rightarrow 3$ if branch 11 is not available for cyber enhancement. Meanwhile, for LS-based vital chains, $9 \rightarrow 14$ is included 14 times, $14 \rightarrow 13$ and $16 \rightarrow 9$ are included 10 times, and $13 \rightarrow 11$ is included 9 times. The most vital chain can thus be formulated as $16 \rightarrow 9 \rightarrow 14 \rightarrow 13 \rightarrow 11$.

On the other hand, if the system has existing branch enhancement(s), we can identify the most vital chain with the highest weight and enhance each component of the chain sequentially. In Table III, we provide the identified most vital chain based on a representative set of initial conditions. It can be observed that given the same existing enhancement, OC- and LS-based most vital chains can be different due to their distinct design objectives. Hence, the system operator needs to carefully evaluate the enhancement strategy for overall best defensive effectiveness.

The results presented in Table III further confirm that the proposed algorithm allows the system operator to enhance the cyber-security of the system in a flexible and adaptive fashion as existing enhancement was not supported by the conventional critical branch identification approaches in previous literature.

C. Vital Defense Chain for the European 89-Bus System

Furthermore, we investigate the vital chains in the European 89-bus system. For the sake of brevity, the top 6 OC- and LS-based vital chains according to the ranking of their weights are shown in Tables IV and V, respectively. We can observe that different than the IEEE 14-bus system, the weights of the top OC-based vital chains are significantly greater than that of the top LS-based vital chains. This suggests that for the European

TABLE IV
OC-BASED VITAL CHAINS IN 89-BUS SYSTEM

Start	Vital chains	Weights
159→	98→66→146→166→124→170→199→123→180	14694.50
68→	196→189→66→146→166→124→170→199→123	13942.75
196→	189→66→146→166→124→170→199→123→180	13352.15
73→	135→54→3→85→199→123→180→112→74	12618.10
95→	98→66→146→166→124→170→199→123→180	12382.09
14→	95→98→66→146→166→124→170→199→123	12312.75

TABLE V
LS-BASED VITAL CHAINS IN 89-BUS SYSTEM

Start	Vital chains	Weights
161→	181→199→135→109→141→23→198→106→69	8823.48
206→	128→102→53→140→110→33→145→198→106	8602.12
181→	199→135→109→141→23→198→106→69→6	8227.58
128→	102→53→140→110→33→145→198→106→69	8019.41
101→	210→126→43→113→96→132→32→156→184	7653.66
71→	177→126→43→113→96→132→32→156→184	7572.56

89-bus system, reducing the LS losses may be relatively easier than reducing the OC losses via branch strengthening under multi-scenario LRAs.

Based on the formulated vital chains, we can further identify the optimal chain if the system has no existing branch enhancement. We can record the frequency of each sequence included in all the formulated vital chains. Specifically, for the OC-based vital chains, sequences 141 → 100, 66 → 146, 100 → 66, 170 → 199, 146 → 166, 129 → 104, 104 → 141, 166 → 124 appear more frequently compared to the other sequences. According to their frequency of being included in the vital chains, we can formulate the OC-based most vital chain for the European 89-bus system as 141 → 100 → 66 → 146 → 166 → 124 → 170 → 199 → 123 → 180. Meanwhile, based on the LS-based vital chains identified, we can find that sequences 69 → 6, 6 → 103, 103 → 140, 110 → 33, 140 → 110, 112 → 69, 33 → 145, and 145 → 198 appear more frequently, based on which we can formulate the LS-based most vital chain as 69 → 6 → 103 → 140 → 110 → 3 → 145 → 198 → 106 → 177. We can observe that the identified most vital chain forms a closed loop, which suggests that defending branches included in this chain sequentially would provide the system operator a self-consistent solution to mitigate the load shedding effects of multiple-scenario LRAs.

D. Performance Evaluation: No Existing Branch Enhancement

Once the most vital chains are identified, we can evaluate the defensive effectiveness of strengthening these branches by measuring the damaging effects of the multi-scenario LRAs following the strengthening. We first compare the proposed approach with the conventional bi-level optimization model [6] (hereinafter referred to as BOM) and tri-level optimization model [7] (hereinafter referred to as TOM) that identify branches to strengthen based on a single, specific operating condition. Specifically, BOM and TOM work as follows: based on equations (1) and (2) and a given representative system operating condition, we can obtain the false load and

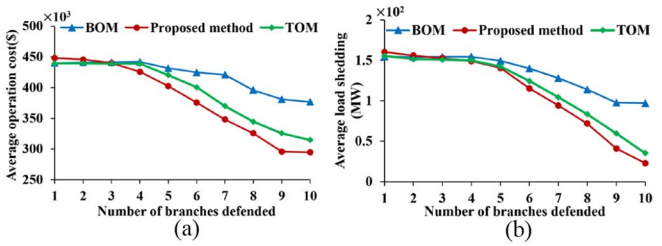

Fig. 5. Operation cost (a) and load shedding (b) following the strengthening of critical branches.

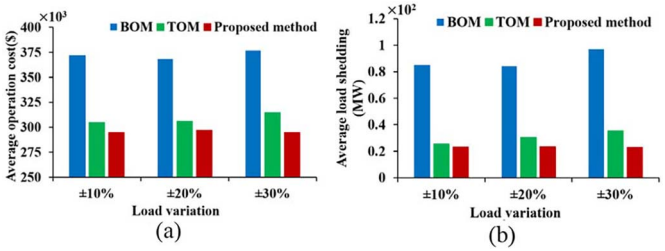

Fig. 6. Operation cost (a) and load shedding (b) following the strengthening of critical branches.

branch power measurement injections $\Delta \boldsymbol{S}$ and $\Delta \boldsymbol{P}$ that the attacker needs to construct to maximize the damaging effect of the LRA, as well as the branch power flow $\boldsymbol{P}$ following the attack. We can then rank the branches according to their power flow to determine the most critical branches and strengthen them accordingly. As mentioned in the previous discussion, conventional approaches, such as BOM and TOM, are incapable of incorporating existing branches; hence we assume that the system has no existing branch enhancement.

To compare the effectiveness of BOM, TOM, and the proposed approach in handling multi-scenario LRAs, we select 2500 randomly generated operating scenarios of the European 89-bus system. Then, in each operating scenario, 20 attack vectors are generated (2500×20 attack vectors in total) based on equation (9) and injected into the system. Fig. 5 depicts the average OC and LS of the European 89-bus system following the attack. It can be observed that when the number of strengthened branches is low, BOM and TOM slightly outperform the proposed approach under multi-scenario LRAs. However, with the number of strengthened branches increasing, which is a more realistic setting for the European 89-bus system that has 210 branches, it can be seen that the proposed approach consistently outperforms BOM and TOM significantly both in terms of OC and LS.

Furthermore, we specify three types of load ranges (−10%∼10%, −20%∼20%, and −30%∼30%) in 2500 randomly generated operating scenarios to investigate the performance of the proposed approach under different levels of operational uncertainties. Without loss of generality, it is assumed that the system operator has defensive resources to strengthen 10 branches. The results of this analysis are shown in Fig. 6. We can observe a clear trend that the proposed approach is more effective in countering the damaging effects of the multi-scenario LRAs than BOM/TOM, which identify their critical branches based on a particular LRA profile.

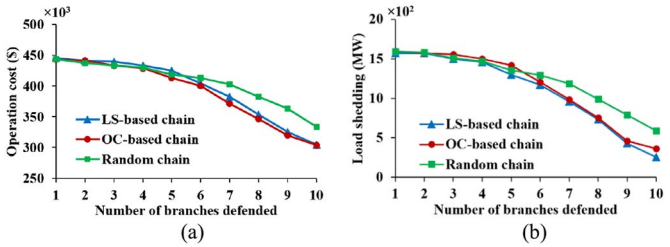

Fig. 7. Operation cost (a) and load shedding (b) with existing branches 9 and 204 enhancement following sequential strengthening.

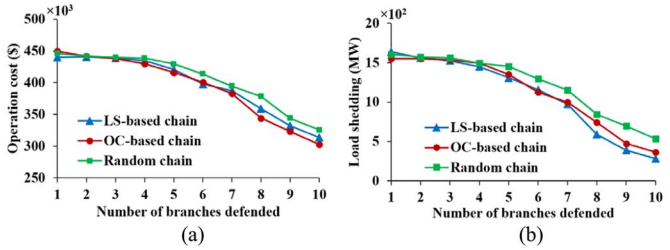

Fig. 8. Operation cost (a) and load shedding (b) with existing branches 100, 137 and 141 enhancement following sequential strengthening.

E. Performance Evaluation: With Existing Branch Enhancement

To demonstrate the performance of the proposed approach with existing branch enhancement taken into account for the European 89-bus system, we consider the following two sets of existing branches: $\{9, 204\}$ and $\{100, 137, 141\}$ as examples. We first construct the OC- and LS-based most vital chains for these two initial conditions. Specifically, for $\{9, 204\}$, we can obtain the OC-based most vital chain as: $9 \rightarrow 204 \rightarrow 129 \rightarrow 104 \rightarrow 141 \rightarrow 100 \rightarrow 66 \rightarrow 146 \rightarrow 166 \rightarrow 124$ and the LS-based most vital chain as: $9 \rightarrow 204 \rightarrow 138 \rightarrow 150 \rightarrow 205 \rightarrow 143 \rightarrow 78 \rightarrow 120 \rightarrow 10 \rightarrow 192$. Similarly, for existing branches $\{100, 137, 141\}$, the OC- based most vital chain can be identified as: $137 \rightarrow 141 \rightarrow 100 \rightarrow 66 \rightarrow 146 \rightarrow 166 \rightarrow 124 \rightarrow 170 \rightarrow 199 \rightarrow 123$ and the LS- based most vital chain can be identified as: $100 \rightarrow 137 \rightarrow 141 \rightarrow 23 \rightarrow 198 \rightarrow 106 \rightarrow 69 \rightarrow 6 \rightarrow 103 \rightarrow 104$.

We can now verify the derived chain of defense by comparing its defensive effectiveness with a randomly selected chain based on the same initial condition. Note that here we use a randomly selected chain for performance verification as there is no existing critical branch identification approach that incorporates existing branch enhancements in the literature. The performance comparisons are given in Figs. 7 and 8, respectively. We can clearly observe that for both initial conditions, strengthening the OC-based most vital chains provides the best defensive effectiveness against the multi-scenario LRAs in terms of operation economics. Meanwhile, strengthening the LS-based most vital chains provides the best defensive effectiveness against the load shedding effect of the multi-scenario LRAs. It is thus evident that the proposed approach is capable of addressing the two-fold damaging effects of multi-scenario LRAs in a brownfield setting where existing enhancements are in place.

V. Conclusion

This paper presents a novel critical branch identification approach to strengthen the cyber security of the power grid and protect against two-fold disruptions resulting from LRAs. We introduced a new concept, chain of defense, to overcome two major limitations in the existing LRA literature. We then proposed to model the construction of the defense chain as a Markov decision process, which is then solved by a Q-learning-based algorithm under an RL framework. Simulation results showed that the proposed approach provides more comprehensive and effective protection against the damaging effects of the multi-scenario LRAs. The proposed approach also enables the incorporation of existing/multi-stage branch enhancement for the system operator to design a more flexible, adaptive, and comprehensive cyber enhancement strategy. To the best of our knowledge, our work is the first of its kind. The proposed RL framework also establishes the foundation and opens up the opportunity for future work to deploy other more advanced RL algorithms to identify critical branches and mitigate the adverse effects of LRAs and potentially other FDAs.

To extend the proposed research, one can enhance the Q-learning-based search strategy adopted in this paper to improve the training process of the agent. Future work can also include the systematic handling of topology variations and dynamic changes in the network configuration that might affect the branch defense priorities.

References

[1] J. Hull, H. Khurana, T. Markham, and K. Staggs, "Staying in control: Cybersecurity and the modern electric grid," *IEEE Power Energy Mag.*, vol. 10, no. 1, pp. 41–48, Jan./Feb. 2012.

[2] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong, "A review of false data injection attacks against modern power systems," *IEEE Trans. Smart Grid,* vol. 8, no. 4, pp. 1630–1638, Jul. 2017.

[3] L. Yao, N. Peng, and K. R. Michael, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Info. Syst. Sec.*, vol. 14, no. 1, p. 13, May 2011.

[4] P. Zhao, C. Gu, and D. Huo, "Coordinated risk mitigation strategy for integrated energy systems under cyber-attacks," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 4014–4025, Sep. 2020.

[5] S. Wang, S. Bi, and Y.-J. A. Zhang, "Locational detection of the false data injection attack in a smart grid: A multilabel classification approach," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8218–8227, Sep. 2020.

[6] Y. Yuan, Z. Li, and K. Ren, "Modeling load redistribution attacks in power systems," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 382–390, Jun. 2011.

[7] Y. Yuan, Z. Li, and K. Ren, "Quantitative analysis of load redistribution attacks in power systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1731–1738, Sep. 2012.

[8] R. Jiao, G. Xun, X. Liu, and G. Yan, "A new AC false data injection attack method without network information," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5280–5289, Nov. 2021.

[9] T. T. Kim and H. V. Poor, "Strategic protection against data injection attacks on power grids," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 326–333, Jun. 2011.

[10] R. Deng, G. Xiao, and R. Lu, "Defending against false data injection attacks on power system state estimation," *IEEE Trans. Ind. Informat.*, vol. 13, no. 1, pp. 198–207, Feb. 2017.

[11] X. Liu, Z. Li, and Z. Li, "Optimal protection strategy against false data injection attacks in power systems," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1802–1810, Jul. 2017.

[12] A. Abusorrah, A. Alabdulwahab, Z. Li, and M. Shahidehpour, "Minimax-regret robust defensive strategy against false data injection attacks," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2068–2079, Mar. 2019.
[13] T. S. Sreeram and S. Krishna, "Protection against false data injection attacks considering degrees of freedom in attack vectors," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5258–5267, Nov. 2021.
[14] Y. Xiang, Z. Ding, Y. Zhang, and L. Wang, "Power system reliability evaluation considering load redistribution attacks," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 889–901, Mar. 2017.
[15] C. Pei, Y. Xiao, W. Liang, and X. Han, "PMU placement protection against coordinated false data injection attacks in smart grid," *IEEE Trans. Ind. Appl.*, vol. 56, no. 4, pp. 4381–4393, Jul./Aug. 2020.
[16] S. Bi and Y. J. Zhang, "Graphical methods for defense against false-data injection attacks on power system state estimation," *IEEE Trans. Smart Grid*, vol. 5, no. 3, pp. 1216–1227, May 2014.
[17] Y. Liu, S. Gao, J. Shi, X. Wei, Z. Han, and T. Huang, "Pre-overload-graph-based vulnerable correlation identification under load redistribution attacks," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5216–5226, Nov. 2020.
[18] Z. Li, M. Shahidehpour, A. Alabdulwahab, and A. Abusorrah, "Bilevel model for analyzing coordinated cyber-physical attacks on power systems," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2260–2272, Sep. 2016.
[19] X. Liu and Z. Li, "Trilevel modeling of cyber attacks on transmission lines," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 720–729, Mar. 2017.
[20] L. Che, X. Liu, and Z. Li, "Fast screening of high-risk lines under false data injection attacks," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4003–4014, Jul. 2019.
[21] L. Che, X. Liu, Z. Shuai, Z. Li, and Y. Wen, "Cyber cascades screening considering the impacts of false data injection attacks," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6545–6556, Nov. 2018.
[22] L. Che, X. Liu, and Z. Li, "Mitigating false data attacks induced overloads using a corrective dispatch scheme," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3081–3091, May 2019.
[23] C. Yan, Y. Tang, J. Dai, C. Wang, and S. Wu, "Uncertainty modeling of wind power frequency regulation potential considering distributed characteristics of forecast errors," *Prot. Control Mod. Power Syst.*, vol. 6, no. 3, pp. 276–288, Nov. 2021.
[24] L. Chen, L. Ma, N. Liu, L. Wang, and Z. Liu, "Parameter tampering cyberattack and event-trigger detection in game-based interactive demand response," *Int. J. Electr. Power*, vol. 35, Feb. 2022, Art. no. 107550.
[25] Y. Liu, S. Gao, J. Shi, X. Wei, and Z. Han, "Sequential-mining-based vulnerable branches identification for the transmission network under continuous load redistribution attacks," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5151–5160, Nov. 2020.
[26] C. Chen, M. Cui, F. Li, S. Yin, and X. Wang, "Model-free emergency frequency control based on reinforcement learning," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2336–2346, Apr. 2021.
[27] M. Esmalifalak, H. Nguyen, R. Zheng, L. Xie, L. Song, and Z. Han, "A stealthy attack against electricity market using independent component analysis," *IEEE Syst. J.*, vol. 12, no. 1, pp. 297–307, Mar. 2018.
[28] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Syst. J.*, vol. 11, no. 3, pp. 1644–1652, Sep. 2017.
[29] Z. Zhang, S. Huang, Y. Chen, S. Mei, and K. Sun, "An online search method for representative risky fault chains based on reinforcement learning and knowledge transfer," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1856–1867, May 2020.
[30] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Wiley, 1958.
[31] "MATPOWER V.7.0. Download." [Online]. Available: www.pserc.cornell.edu/matpower/ (Accessed: 2022).

Jieyu Lei (Student Member, IEEE) is currently pursuing the Ph.D. degree in electrical engineering with Southwest Jiaotong University, Chengdu, China. His research interests include power market and cyber–physical power systems.

Shibin Gao received the Ph.D. degree from Southwest Jiaotong University, Chengdu, China, in 2004, where he is currently a Professor with the School of Electrical Engineering. His research interests include power system protection and automation, online monitoring of electrical equipment, traction power supplies, railway electrification, and power system security.

Jian Shi (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Mississippi State University in 2014. He is currently an Assistant Professor with the Engineering Technology Department, University of Houston. His research interests include microgrid modeling and control, transportation electrification, and cyber–physical power systems.

Xiaoguang Wei received the Ph.D. degree in electrical engineering from Southwest Jiaotong University, Chengdu, China, in 2019, where he is currently an Assistant Professor with the School of Electrical Engineering. His research interests include power market and energy system security.

Ming Dong (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Alberta. Since graduation, he has been working with ISO and major utilities in Canada as a Senior Engineer Specialist and various other roles for nine years. His research interests include applications of artificial intelligence and big data technologies to power system planning, operation and asset management, power quality, and smart energy management. He is the recipient of the Certificate of Data Science and Big Data Analytics from Massachusetts Institute of Technology. He is currently an Associate Editor of IEEE TRANSACTIONS ON POWER DELIVERY, IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, and *CSEE Journal of Power and Energy Systems*.

Wenshuang Wang received the Ph.D. degree in mathematical sciences from Mississippi State University in 2017. She is currently a ConocoPhillips Instructional Assistant Professor of Data Science with the Department of Mathematics, University of Houston (UH). Prior to joining UH, she was a Statistician with Texas Medical Center. Her research interest includes data science in healthcare and energy for better decision making and quality improvement.

Zhu Han (Fellow, IEEE) received the B.S. degree in electronic engineering from Tsinghua University in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland at College Park in 1999 and 2003, respectively. From 2000 to 2002, he was a Research and Development Engineer with JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate with the University of Maryland at College Park. From 2006 to 2008, he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department and the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, the IEEE Leonard G. Abraham Prize in the field of Communications Systems (Best Paper Award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. He has been a 1% Highly Cited Researcher since 2017 according to Web of Science. He is also the winner of the 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks." He was an IEEE Communications Society Distinguished Lecturer from 2015 to 2018 and has been an AAAS Fellow since 2019 and an ACM Distinguished Member since 2019.