IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2025, accepted 20 February 2025, date of publication 5 March 2025, date of current version 14 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3548517

RESEARCH ARTICLE

SPARQ: A Cyber-Resilient Voltage Regulation Using Soft Q-Learning Approach for Autonomous Grid Operations

, (Member,)
, at Qatar, Doha, Qatar
, College Station, TX 77840, USA

e-mail:

This work was supported in part by under Award DE-CR0000018, in part by under Grant 2220347, and in part by

ABSTRACT The growing integration of distributed energy resources and increased interconnectivity in cyber-physical power systems (CPPSs) have heightened their complexity. This complexity has made voltage stability control more vulnerable, especially under cybersecurity threats. Cybersecurity threats enable the manipulation of critical system states, potentially causing blackouts and cascading failures. This highlights the need for adaptive, efficient, and resilient control mechanisms to ensure CPPS stability. This paper presents a novel Stability and voltage Protection Achieved with Resilient Soft Q-learning (SPARQ). The proposed approach leverages a Soft Q-Learning (SQL) framework to autonomously regulate voltage stability while addressing the impact of cyber attacks. The proposed SQL-based control system incorporates adaptive preprocessing mechanisms to normalize observations and enhance policy robustness. The study evaluates the performance of the SQL agent under both normal and cyber-attacked scenarios, with simulated disturbances such as voltage variability, stochastic load dynamics, and deliberate data injections. Comprehensive experiments on the 14-bus, reduced 118-bus, and 118-bus systems demonstrate the effectiveness of the SQL framework in achieving improved voltage regulation. Additionally, the SQL framework exhibits faster convergence and higher rewards compared to baseline reinforcement learning methods. Moreover, the framework's effectiveness under cyber attack highlights its potential for resilient voltage stability control in modern CPPSs.

INDEX TERMS Autonomous voltage regulation, cyber-physical systems, reinforcement learning, soft Q-learning, resilient control systems, voltage stability control.

NOMENCLATURE

FUNCTIONS

P	Active Power.
Q	Reactive Power.

VARIABLES

α	Learning Rate / Entropy Scaling Parameter.
ϵ	Exploration Rate.
γ	Discount Factor.
$\mathcal{A}$	Action Space.
$\mathcal{S}$	State Space.
$\rho_\pi(s, a)$	Occupancy Measure.
τ	Temperature Parameter.
P_{env}	State Transition Probability.
$Q_\pi(s, a)$	Q-Function (Action-Value Function).
$R(s, a)$	Reward Function.
$T(s, a)$	State Transition Model.
$V(s)$	State-Value Function.

ABBREVIATIONS

ALR	Adaptive Learning Rate.
CPPS	Cyber-Physical Power System.
DDPG	Deep Deterministic Policy Gradient.

The associate editor coordinating the review of this manuscript and approving it for publication was Rossano Musca.

DDQN	Double Deep Q-Network.
DERs	Distributed Energy Resources.
DQN	Deep Q-Network.
DRL	Deep Reinforcement Learning.
EBM	Energy-Based Model.
ESS	Energy Storage System.
FDIA	False Data Injection Attack.
GCN	Graph Convolutional Network.
GEVR	Grid Emergency Voltage Regulation.
LTC	Load Tap Changer.
MADRL	Multi-Agent Deep Reinforcement Learning.
MARL	Multi-Agent Reinforcement Learning.
MDP	Markov Decision Process.
MERL	Maximum Entropy Reinforcement Learning.
PCL	Path Consistency Learning.
PG-MA2TD3	Physics-Guided Multi-Agent Twin Delayed Deep Deterministic Policy Gradient.
SAC	Soft Actor-Critic.
SPARQ	Stability and Voltage Protection Achieved with Resilient Soft Q-learning.
SQL	Soft Q-Learning.
VC	Voltage Control.

I. INTRODUCTION

The increasing integration of distributed energy resources (DERs) and the heightened interconnectivity of modern cyber-physical power systems (CPPSs) have significantly amplified the complexity and challenges of voltage stability control. Coupled with the growing threat of cybersecurity breaches, these developments underscore the urgent need for innovative and adaptive voltage regulation mechanisms [1]. Blackout events caused by voltage instability and cyber-induced disturbances have been on the rise, leading to substantial power losses and operational disruptions [2]. Ensuring voltage stability within safe operational limits is critical to mitigating the risk of cascading failures and maintaining grid resilience. The implementation of grid emergency voltage regulation (GEVR) mechanisms is paramount to minimize the frequency and severity of power disruptions and ensure reliable grid operations [3], [4]. GEVR typically involves corrective actions such as under-voltage load shedding and dynamic adjustments in generation or consumption to stabilize voltage levels and restore nominal grid performance [5]. However, the complex nonlinear dynamics of CPPSs, governed by differential-algebraic equations and influenced by numerous dynamic components, pose significant challenges to effective decision-making in GEVR, particularly under cyber attack scenarios [6].

Cybersecurity threats exacerbate these challenges by introducing adversarial conditions, such as false data injection attacks (FDIAs), that compromise the integrity of grid operations and hinder conventional control mechanisms [7]. Traditional approaches, which rely on rules-based load-shedding relay mechanisms, often lack the adaptability and robustness required to counteract adversarial scenarios and dynamic grid conditions. In response to these limitations, advanced deep reinforcement learning (DRL) techniques have emerged as a promising solution for enhancing voltage stability control. Specifically, Soft Q-Learning (SQL), an entropy-regularized variant of Q-learning, provides a robust framework for managing uncertainties and cyber threats in CPPSs [8]. SQL's self-improving mechanism enables it to learn optimal control policies under diverse grid conditions and adversarial environments [9]. For instance, a study in [4] employed a multi-agent graph-attention-based RL algorithm to enhance grid voltage regulation. However, the method demonstrated instability during training under adversarial conditions [10]. Similarly, a Markov decision process (MDP) approach was proposed in [11] for optimal load tap changer (LTC) settings but faced limitations in real-time applicability due to its dependency on accurate grid topology information.

This paper introduces a Stability and voltage Protection Achieved with Resilient Soft Q-learning (SPARQ), a novel SQL-based voltage regulation framework designed to autonomously manage voltage stability while addressing cyber threats. The SPARQ agent leverages a MDP formulation with adaptive preprocessing mechanisms to normalize observations and enhance policy robustness against adversarial scenarios. The primary contributions of this work are summarized as follows.

- A novel SPARQ-based voltage regulation framework is proposed, integrating adaptive preprocessing mechanisms to enhance the resilience of control policies under adversarial conditions.
- The performance of SPARQ is systematically evaluated against state-of-the-art RL strategies, focusing on its ability to maintain voltage stability, maximize rewards, and achieve robust performance under varying grid and attack scenarios.
- The scalability and effectiveness of SPARQ are demonstrated on the IEEE 14-bus, the reduced IEEE 118-bus, and the standard IEEE 118-bus systems.
- The SPARQ resilience is assessed under cyber attacks to highlight its potential for enhancing the cybersecurity of voltage regulation in modern CPPSs.

The remainder of this paper is organized as follows. Section II reviews prior work, while Section III defines the problem formulation. Section IV describes the proposed methodology, and Section V elaborates on the architecture in detail. Section VI presents experimental results and evaluates SPARQ's performance. Lastly, Section VII concludes the paper.

II. PRIOR WORK

Existing VC approaches in CPPSs have evolved from traditional model-based techniques to more sophisticated

control strategies that incorporate real-time data analytics and adaptive functionalities. Early research efforts primarily focused on linearized models and rules-based load shedding schemes to address transient voltage instabilities, emphasizing strict adherence to safety margins and deterministic control actions [12], [13]. Although these classical methods often provided a degree of stability assurance, they were hampered by their reliance on simplified assumptions and off-line studies, thereby lacking robustness in the face of evolving grid dynamics [14]. Over time, researchers recognized the necessity for dynamic models to capture the nonlinear behavior of loads and generators, prompting the development of extended time-domain simulations that employ differential-algebraic equations to predict voltage responses under diverse operating conditions [1]. However, the intricate coupling between generation, transmission, and distribution systems, alongside growing cybersecurity threats, has underscored the limitations of conventional stability analysis. Significant endeavors in system identification and real-time monitoring have attempted to bridge this gap by integrating phasor measurement units and wide-area monitoring systems [15]. Yet, these sensor-centric solutions are often prone to communication delays and data integrity issues, amplifying the vulnerability of voltage stability controls to adversarial interventions and attack vectors.

The advent of machine learning and data-driven techniques has reshaped the research landscape, offering new ways to handle the high-dimensional complexities of modern CPPSs. Neural network-based controllers and fuzzy logic systems, for instance, have demonstrated improved adaptability to time-varying load conditions and uncertain renewable generation [16]. Reinforcement learning (RL) approaches such as Q-learning, policy gradient methods, and their variants emerged as promising avenues for real-time decision-making in both centralized and decentralized voltage regulation settings. Table 1 presents a comparison of the prior work with the integration of cutting-edge RL approaches for VC. In particular, an RL-based scheme was introduced in [8] for load tap changer settings, yielding notable improvements in dynamic voltage stability. Although these RL frameworks have enabled self-learning control policies that adapt to fluctuating grid states, they frequently exhibit slow convergence and lack robust performance in the presence of adversarial disturbances. Moreover, the absence of explicit mechanisms to handle cybersecurity breaches and measurement falsifications has limited their efficacy in practical CPPSs.

Building on these foundational works, more recent studies have begun exploring multi-agent reinforcement learning (MARL) and advanced policy regularization techniques to bolster the security and stability of VC. Graph-based MARL strategies, for instance, leverage inter-bus relationships to coordinate local voltage regulation actions while enhancing fault tolerance [4]. Entropy-regularized variants such as Soft Actor-Critic (SAC) and SQL further encourage policy exploration, thus improving convergence rates and resilience to sudden disruptions [9]. Nonetheless, many of these advanced algorithms still struggle with scalability and real-time operational constraints when confronted with malicious data injections and stealth cyber attacks [20]. Efforts to integrate detection mechanisms within the learning loop—such as adversarial training or anomaly detection—have shown promise, but often lead to higher computational overhead and conservative control policies. Thus, despite considerable progress, there is a pressing need for RL frameworks that not only account for the multifaceted nature of power system (PS) dynamics but also explicitly incorporate robust preprocessing and decision-making strategies to mitigate the risks posed by sophisticated cyber threats.

III. PROBLEM FORMULATION

This section covers the problem formulation for VC as well as the underlying RL algorithms.

A. PRELIMINARIES

In the framework of Markov decision processes (MDPs), a dynamical system is represented by a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition kernel $P_{\text{env}}(s_{t+1} \mid s_t, a_t)$, and a scalar reward function $R(s_t, a_t)$. When modeling voltage control (VC), each state s_t encapsulates information such as real-time bus voltages, load demands, and other critical variables influencing power flow. The agent's action a_t may entail adjusting reactive power injections, transformer tap settings, or other system parameters to steer voltages toward desired setpoints. Transitions $P_{\text{env}}(s_{t+1} \mid s_t, a_t)$ are dictated by non-linear PS equations, embodying both deterministic (e.g., network topology, Ohm's law) and stochastic (e.g., load fluctuations, renewable intermittency) effects. The scalar reward $r_t = R(s_t, a_t)$ often penalizes voltage deviations, thus offering a quantitative metric for algorithmic updates. The system evolves over a discrete timeline $t = 0, 1, 2, \ldots$, and the agent's aim is to find an optimal policy $\pi^*(a_t \mid s_t)$ that maximizes the expected cumulative discounted return, $\mathbb{E}_{p_\pi(\tau)}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$. Through repeated exploration, the policy refines its decision rules, enabling robust behavior under varying load profiles and possible cyber-induced disruptions.

A key tool in characterizing policy performance is the action-value function (or Q-function), $Q_\pi(s, a)$, which calculates the expected return starting from state s and action a and thereafter following policy π. This function follows the Bellman equation as

$$Q_\pi(s, a) = \mathbb{E}\Big[r + \gamma Q_\pi(s', a')\Big], \tag{1}$$

where the expectation is taken with respect to $s' \sim P_{\text{env}}(\cdot \mid s, a)$ and $a' \sim \pi(\cdot \mid s')$. For an optimal policy π^*, the corresponding Q-function $Q^*(s, a)$ attains maximal values over all actions. Traditional Q-learning uses iterative updates to approximate Q^* from sampled transitions. However, modern extensions, especially entropy-regularized methods such as SQL, include an entropy term to encourage

TABLE 1. Comparison of voltage control methods.

Ref.	Methodology	Test System	Advantages	Disadvantages
[17]	DA-MADRL with model-free reactive power prediction and optimal inverter placement for day-ahead voltage regulation.	Modified IEEE-33 bus system.	Efficient resource utilization, reduced voltage variation, and power losses.	Assumes a pre-determined network topology; overlooks real-time adaptability.
[18]	PG-MA2TD3 algorithm for intra-day voltage regulation with adversarial learning and Jacobian matrix for neighbor impact analysis.	IEEE 33-bus system with Portuguese power system data.	Enhanced robustness and steady-state reward in trading and regulation.	Requires detailed Jacobian matrix calculations, which may be computationally intensive.
[4]	Graph-Attention-based DRL with GCN for feature representation and attention mechanism for cooperative learning.	IEEE benchmark system.	Data-efficient and scalable with improved decision accuracy.	Scalability may still face challenges with extremely large and complex grids.
[19]	DRL-based scheduling of ESS, addressing high-dimensional state and action spaces for real-time regulation.	Customized 6-bus and modified IEEE 34-bus systems.	Handles high-dimensional problems, providing near-optimal solutions.	Struggles with uncertainties in the model-free DRL setting for long-term predictions.
[5]	Multi-agent DRL with counter-training of local policy and centric critic networks for coordinated PV inverter control.	IEEE 33-bus system.	Enhanced control capability under various conditions.	Lacks explicit detail on handling extreme PV penetration and grid disturbances.

exploration and mitigate convergence issues arising from complex state-action spaces. This augmentation is critical for high-dimensional CPPSs, where balancing exploration and exploitation becomes pivotal amid noisy measurements and potential adversarial data injections.

When applying MDP-based frameworks to VC problems, it is customary to employ a time-discretization strategy that aligns with operational or control intervals in power grids, such as a few seconds to minutes. Each episode spans a fixed horizon T or concludes if certain termination conditions—e.g., system collapse or voltage outside permissible ranges—are met. The occupancy measure $\rho_\pi(s, a)$, given by $\sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}(s_t = s, a_t = a \mid \pi)$, is particularly instructive for quantifying how often different state-action pairs are visited. In practice, ensuring adequate coverage over $\mathcal{S} \times \mathcal{A}$ necessitates both stochastic exploration policies and well-designed reward shaping. Furthermore, stabilizing learning amid non-stationary conditions (e.g., varying loads, ramping renewable generation) often involves online updating schemes or off-policy replay buffers. This ensures that the agent's policy remains responsive to system changes and resilient to adversarial factors, thereby safeguarding voltage stability over extended operational timelines.

B. VOLTAGE CONTROL ENVIRONMENT

VC in modern CPPSs can be rigorously modeled as a Markov Decision Process (MDP), where the goal is to keep voltage levels within specified safety margins despite the stochastic and time-varying nature of system dynamics. Formally, we denote this MDP by the tuple $\langle \mathcal{S}, \mathcal{A}, P_{\text{env}}, R, \gamma \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, $P_{\text{env}}(\cdot \mid s, a)$ the state transition probabilities, $R(s, a)$ the scalar reward function, and γ the discount factor. Each state $s_t \in \mathcal{S}$ provides a snapshot of the power grid at time t, including bus voltage magnitudes, transformer tap positions, and load demands. The decision variables $a_t \in \mathcal{A}$ are the control signals that adjust voltages, such as regulating reactive power or changing transformer taps. The transition dynamics $P_{\text{env}}(s_{t+1} \mid s_t, a_t)$ capture how the PS evolves from one state to another under the chosen action, governed by both deterministic physical laws and stochastic factors such as load uncertainty and renewable generation variability. For every state-action pair, a scalar reward $r_t \in R(s_t, a_t)$ indicates instantaneous performance, often penalizing deviations of bus voltages from a desired reference to promote stable operating conditions. The discount factor $\gamma \in (0, 1)$ balances short-term versus long-term objectives. This environment integrates advanced modeling of stochastic uncertainties and adversarial perturbations, simulating realistic operational scenarios.

At each time step t, the agent observes state s_t, selects an action a_t according to a policy $\pi_\theta(a_t \mid s_t)$ parameterized by θ, and receives a reward r_t based on how well the voltage is regulated. The system transitions to the next state s_{t+1} with probability $P_{\text{env}}(s_{t+1} \mid s_t, a_t)$. Over an episode or trajectory $\tau = (s_1, a_1, s_2, a_2, \dots)$, the cumulative discounted return is defined by $R(\tau) = \sum_{t=1}^{T} \gamma^t \, r_t$, and the probability of observing a particular trajectory τ under π is given by $p_\pi(\tau)$. The objective is to maximize the expected return, which balances immediate and future rewards as follows.

$$\mathbb{E}_{p(\tau)}[R(\tau)] = \int_{\tau} p_\pi(\tau) \, R(\tau) \, d\tau. \tag{2}$$

A fundamental quantity in this framework is the action-value function $Q_\pi(s, a) = \mathbb{E}[R(\tau) \mid s_t = s, a_t = a]$, where the expectation is taken over trajectories starting from state s and action a, with subsequent decisions made by π. Mathematically, the occupancy measure $\rho_\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}(s_t = s, a_t = a \mid \pi)$ represents how often the MDP visits each state–action pair under policy π. This measure is particularly useful in theoretical analysis, as it forms a one-to-one mapping with the policy.

IV. METHODS

In this section, the resilient voltage regulation problem for CPPSs is regarded as a continuous Markov chain (MC) problem, and SQL is used to solve the MDP. Whereupon, the main parameters of the proposed SQL are given in details.

A. NOTATIONS

Q-learning lies at the core of model-free reinforcement learning, aiming to approximate the optimal action-value function $Q^*(s, a)$ by iteratively adjusting parameter estimates through experience in a MDP. Q-learning estimates $Q(s, a)$ via the iterative update as

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\Big], \quad (3)$$

where α is the learning rate and r_t is the reward received at time t. In high-dimensional settings, neural networks are employed to approximate $Q_\theta(s, a)$, updating their weights θ through gradients derived from the temporal-difference error. Modern variants such as Double Q-learning and Dueling DQN aim to correct overestimation biases and isolate the state-value function from the action-value function.

Extending these fundamental ideas, recent work integrates entropy-regularized objectives to encourage broad exploration, thereby mitigating the risk of converging to suboptimal deterministic policies. SQL, for instance, modifies the Bellman backup to include an entropy term that balances exploration against exploitation, often improving convergence properties in non-stationary or adversarial environments. Under this paradigm, the Q-value update rule incorporates a temperature parameter α to scale the entropy term, effectively promoting stochastic behavior in the policy. By merging elements of DQN and SQL techniques—such as the single-stream or dueling architecture for Q-value estimation—contemporary RL systems can achieve robust performance under diverse and uncertain conditions. This synergy is particularly valuable for voltage control and other safety-critical PS applications, where learning policies must adapt swiftly to fluctuating loads, renewable variability, and potential cyber threats without sacrificing numerical stability or long-term performance guarantees.

B. MAXIMUM ENTROPY REINFORCEMENT LEARNING

Maximum Entropy RL (MERL) extends traditional RL by augmenting the objective function with an entropy term, promoting stochastic policies that maximize both cumulative rewards and exploration. The goal is to optimize the maximum entropy objective as

$$\pi^*_{\text{MaxEnt}} = \arg\max_{\pi} \mathbb{E}_{\tau\sim p_\pi}\left[\sum_{t=1}^{T} \left(r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right)\right], \quad (4)$$

where $\mathcal{H}(\pi(\cdot|s_t)) = -\sum_{a\in\mathcal{A}} \pi(a|s_t)\log\pi(a|s_t)$ is the entropy of the policy, $\alpha > 0$ is a temperature parameter that balances reward maximization and exploration, and $\tau = (s_1, a_1, \ldots, s_T, a_T)$ represents a trajectory sampled from the policy π. When $\alpha = 0$, the objective reduces to standard RL.

The entropy term encourages policies that explore multiple actions, leading to improved robustness and the discovery of diverse strategies. Such policies are expressed as

$$\pi^*_{\text{MaxEnt}}(a_t|s_t) = \exp\left(\frac{1}{\alpha}\left(Q^*_{\text{soft}}(s_t, a_t) - V^*_{\text{soft}}(s_t)\right)\right), \quad (5)$$

where $Q^*_{\text{soft}}(s_t, a_t)$ is the soft Q-function, and $V^*_{\text{soft}}(s_t)$ is the soft value function. These are defined as

$$Q^*_{\text{soft}}(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1}\sim p}\left[\gamma\left(r(s_{t+1}, a_{t+1}) + \alpha\mathcal{H}(\pi^*_{\text{MaxEnt}}(\cdot|s_{t+1}))\right)\right], \quad (6)$$

$$V^*_{\text{soft}}(s_t) = \alpha\log\int_{\mathcal{A}} \exp\left(\frac{1}{\alpha}Q^*_{\text{soft}}(s_t, a)\right) da. \quad (7)$$

The optimal soft Q-function, $Q^*_{\text{soft}}(s_t, a_t)$, incorporates the immediate reward $r(s_t, a_t)$, the discounted cumulative rewards, and the entropy bonus. Similarly, the soft value function, $V^*_{\text{soft}}(s_t)$, integrates the entropy-regularized action-value function over all possible actions. Maximum Entropy RL employs an energy-based model (EBM) structure for the policy $\pi^*_{\text{MaxEnt}}(a_t|s_t) \propto \exp\left(\frac{1}{\alpha}Q^*_{\text{soft}}(s_t, a_t)\right)$, where the hyperparameter α determines the degree of randomness in the policy. Lower values of α result in more deterministic policies, while higher values encourage exploratory behaviors. This framework provides multiple benefits, including enhanced exploration, robustness to suboptimal local solutions, and the ability to discover diverse strategies for solving RL tasks. These characteristics make Maximum Entropy RL well-suited for complex, high-dimensional environments where traditional methods struggle to generalize effectively.

C. SOFT Q-LEARNING ARCHITECTURE

The SQL architecture is fundamentally based on entropy-regularized MDPs, where the goal is to maximize the expected cumulative reward as

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right)\right]. \quad (8)$$

The entropy term $\mathcal{H}(\pi(\cdot|s)) = -\sum_{a\in\mathcal{A}(s)} \pi(a|s)\log\pi(a|s)$ encourages stochastic exploration by penalizing deterministic policies. The two primary value functions in this framework are the state value function $V(s)$ and the state-action value function $Q(s, a)$ given by

$$V(s) = \max_{\pi\in\Delta(\mathcal{A}(s))} \mathbb{E}_{a\sim\pi}\left[Q(s, a) - \alpha\log\pi(a|s)\right], \quad (9)$$

$$Q(s, a) = R(s, a) + \gamma\mathbb{E}_{s'\sim T(s,a)}[V(s')], \quad (10)$$

where $R(s, a)$ is the immediate reward, γ is the discount factor, and $T(s, a)$ represents the state transition probability.

The Bellman equations incorporate entropy regularization as

$$Q(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim T(s,a)} \left[\tau \log \sum_{a' \in \mathcal{A}(s')} \frac{\exp(Q(s',a'))}{\tau}) \right], \tag{11}$$

$$V(s) = \tau \log \sum_{a \in \mathcal{A}(s)} \frac{\exp(Q(s,a'))}{\tau}). \tag{12}$$

Here, $\tau > 0$ is the temperature parameter that controls the trade-off between exploration (entropy maximization) and exploitation (reward maximization). The optimal policy $\pi(a|s)$ is derived from the soft Q-function using a Boltzmann distribution as

$$\pi(a|s) = \frac{\frac{\exp(Q(s,a'))}{\tau})}{\sum_{a' \in \mathcal{A}(s)} \frac{\exp(Q(s,a'))}{\tau})}. \tag{13}$$

This ensures that actions with higher $Q(s,a)$ values are selected with higher probabilities, but the entropy term keeps the policy stochastic.

The Path Consistency Learning (PCL) framework imposes temporal consistency across a trajectory $(s_i, a_i, \ldots, s_j, a_j)$ by satisfying the following condition as

$$V(s_i) + \sum_{t=i}^{j-1} \gamma^{t-i} \left[\tau \log \pi(a_t|s_t) - R(s_t,a_t) \right] = \gamma^{j-i} V(s_j). \tag{14}$$

This condition aligns the policy with the soft value functions over multiple time steps, allowing the agent to optimize long-term rewards while maintaining entropy regularization. The soft value function $V(s)$ can be expressed explicitly in terms of the soft Q-function as

$$V(s) = \tau \log \sum_{a \in \mathcal{A}(s)} \frac{\exp(Q(s,a'))}{\tau}). \tag{15}$$

Using this, the Q-function updates iteratively based on observed transitions as

$$Q(s_t,a_t) \leftarrow (1-\alpha) Q(s_t,a_t) + \alpha \left[R(s_t,a_t) + \gamma \mathbb{E}_{s_{t+1} \sim T(s_t,a_t)} V(s_{t+1}) \right], \tag{16}$$

where α is the learning rate. The gradient of the policy objective with respect to π is:

$$\nabla_\pi J(\pi) = \mathbb{E}_{(s,a) \sim \pi} \left[\nabla_\pi \log \pi(a|s) \left(Q(s,a) - \tau \log \pi(a|s) \right) \right]. \tag{17}$$

This update allows the policy to improve by sampling actions proportionally to their soft Q-values while retaining stochasticity. SQL emphasizes balancing exploration and exploitation by embedding entropy maximization directly into the Bellman equations. Through a combination of soft value functions, stochastic policies, and temporal consistency constraints, the architecture achieves efficient and robust learning in high-dimensional action spaces while encouraging exploration and resilience to suboptimal local solutions.

V. PROPOSED ARCHITECTURE

The SQL framework employs a single-stream architecture for Q-value computation, leveraging a neural network to predict Q-values directly for each action a. Fig. 1 displays the proposed architecture of a SPARQ system operating on a grid environment. Specifically, the Q-values guide the agent's policy using a softmax-based action selection mechanism for improved exploration and numerical stability. The soft Q-value update rule incorporates entropy regularization to balance exploitation and exploration.

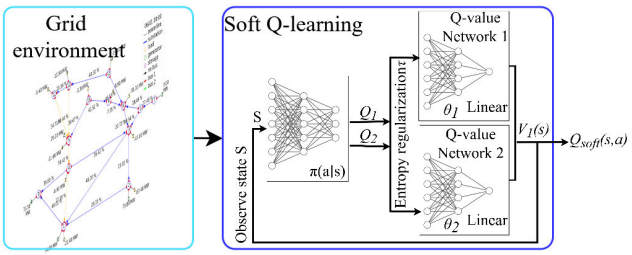

FIGURE 1. SQL architecture illustrating entropy-regularized state-action value functions $Q_{soft}(s,a)$ derived using multiple Q-value networks. The policy $\pi(a|s) = \frac{\exp(Q(s,a)/\tau)}{\sum_{a'} \exp(Q(s,a')/\tau)}$ ensures exploration while balancing reward maximization.

The proposed framework incorporates entropy regularization to balance exploration and exploitation [21]. The Q-values are updated with the soft Bellman operator, integrating robustness against uncertainties as

$$Q_\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P_{\text{env}}} \left[\alpha \log \sum_{a'} \exp\left(\frac{Q(s',a')}{\alpha} \right) \right], \tag{18}$$

where α is the entropy scaling parameter. This formulation enables SQL to handle dynamic grid conditions and adversarial scenarios effectively. Observations are normalized using domain-specific scaling to enhance numerical stability and facilitate learning as

$$s_{\text{norm}} = \frac{\text{prod}_p}{\text{gen}_{\text{pmax}} + \epsilon}, \frac{\text{prod}_v}{200}, \frac{\text{load}_p}{\max(\text{load}_p) + \epsilon}, \frac{\text{load}_q}{\max(\text{load}_q) + \epsilon}, \frac{\text{rho}}{2.0}, \frac{\text{topo}_{\text{vect}}}{4.0}, \tag{19}$$

where ϵ prevents division by zero. Additionally, features such as production voltage prod_v and line loading rho may be subjected to clipping for outlier control. This preprocessing ensures robustness under normal and adversarial conditions.

The agent minimizes an SQL loss function defined as

$$\mathcal{L}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\epsilon_t, 1+\epsilon_t) A_t) - \mathcal{L}_{ALR}(\theta) + \alpha_{exp} H(\pi_\theta) \right], \tag{20}$$

where D is the replay buffer, θ^- are the target network parameters, α is the temperature parameter for entropy regularization, and γ is the discount factor. Experience

replay and periodic updates to the target network stabilize the training process and improve convergence. The detailed pseudocode for the voltage regulation based on the SPARQ algorithm is described in Algorithm 1.

Algorithm 1 SQL for Cyber-Resilient Voltage Regulation

Initialize networks $Q(s, a; \theta)$ with random weights.
Initialize target network $\hat{Q}(s, a; \theta^-)$ with weights $\theta^- = \theta$.
Initialize experience replay memory D and set exploration rate $\epsilon = 1$.
Set softmax temperature parameter τ and reward scaling factor α.
for each episode $n = 1$ to M **do**
  Initialize state s.
  for each step $t = 1$ to T **do**
    Preprocess state s_t to extract normalized features.
    With probability ϵ, select a random action a_t.
    With probability $1 - \epsilon$, compute softmax probabilities:
    $$\pi(a|s_t) = \frac{\exp(Q(s_t,a;\theta)/\tau)}{\sum_{a'}\exp(Q(s_t,a';\theta)/\tau)}.$$
    Sample action $a_t \sim \pi(a|s_t)$.
    Execute action a_t in the environment and observe reward r_t and next state s_{t+1}.
    Store transition (s_t, a_t, r_t, s_{t+1}) in D.
    if D contains sufficient samples **then**
      Sample random mini-batch from D: $(s_j, a_j, r_j, s_{j+1}, d_j)$.
      Compute target values for each sampled transition:
      $$y_j = r_j + (1 - d_j) \cdot \gamma \cdot \alpha \log \textstyle\sum_{a'} \exp(\hat{Q}(s_{j+1}, a'; \theta^-)/\alpha).$$
      Perform a gradient descent step on the loss:
      $$\mathcal{L}(\theta) = \mathbb{E}_j\left[(y_j - Q(s_j, a_j; \theta))^2\right].$$
    end if
    if t mod $K == 0$ **then**
      Update target network weights: $\theta^- \leftarrow \theta$.
    end if
    Decay exploration rate ϵ.
  end for
end for

VI. SIMULATION RESULTS

To assess the performance of Algorithm 1 across different grid sizes and complexities, the simulation results employ three test systems: the 14-bus power grid, the reduced 118-bus system, and the larger IEEE-118 bus network. The SPARQ model is compared to Dueling DQN model, Duel Q Simple (DuelQSimple), Deep Q Network (DQN), Double Deep Q Network (DDQN), and Deep Deterministic Policy Gradient (DDPG). The main Python packages used in this research are Pandapower [22], Grid2op [23], and Tensorflow [24].

A. DATA DESCRIPTION

The implementation utilizes a carefully constructed state space representation derived from the Grid2Op environment. The simulation results incorporate multiple PS parameters that are crucial for voltage regulation. The state space consists of six main feature groups that are preprocessed and normalized to ensure stable learning: power generation ($prod_p$), voltage levels ($prod_v$), active load ($load_p$), reactive power ($load_q$), line Loading (rho), and topology vector ($topo_{vect}$). The feature preprocessing pipeline incorporates normalization for value bounding, outlier handling through clipping, flattening of features into a continuous input vector, and float 32 type conversion for efficient computation. This streamlined process ensures robust and computationally effective input preparation for the neural network.

The state space features undergo comprehensive normalization to ensure effective learning: generator active power ($prod_p$) is normalized by maximum capacity with epsilon protection, bounding it between $[-1, 1]$; voltage outputs ($prod_v$) are scaled by 200 to maintain consistent ranges; load power ($load_p$) and reactive power ($load_q$) are normalized by their respective maximum values to represent relative distributions; line loading ratios (ρ) are divided by 2 and clipped to $[0, 1]$ for security constraint monitoring; and topology vectors ($topo_{vect}$) are scaled by 4.0 to represent circuit breaker states. This normalization strategy ensures consistent feature scaling while preserving critical relationships between PS variables, enabling the model to learn system-wide patterns independent of absolute values.

When cyber attacks are considered, the implementation includes additional processing where $prod_v$ and rho can be perturbed by an attack magnitude parameter. This perturbation is applied probabilistically based on an attack probability parameter, simulating realistic scenarios where measurements might be compromised. The total observation space dimension is dynamically calculated based on the sum of individual feature dimensions, allowing the implementation to adapt to different PS sizes (14)-bus, reduced 118-bus, and full 118-bus systems). This flexible state space representation enables the SPARQ agent to learn voltage regulation policies that are both effective and resilient to cyber-attacks while maintaining awareness of critical system constraints and operational conditions. The feature engineering process ensures that the agent receives a consistent, normalized view of the PS state, facilitating stable learning while preserving the essential relationships between different PS variables. This comprehensive state representation allows the SPARQ agent to learn complex voltage regulation policies that consider both the immediate system conditions and potential security threats. To provide a comprehensive understanding of the simulation results, it's essential to delve into the specific configurations and parameters utilized in the study. The used bus system specifications are reported in Table 2.

TABLE 2. Comparison of the different IEEE bus systems.

Bus system	14-bus system	118-bus system	Reduced 118-bus system
System Size	14 buses	118 buses	36 buses
Power Lines	20 lines	186 lines	59 lines
Generators	6 generators	62 generators	22 generators
Load Points	11 load points	99 load points	37 load points
Network Complexity	Low (simple radial)	High (meshed topology)	Medium-High (simplified mesh)
Chronics Data	1 month of data	50 years of data	240 equivalent years of data

B. PERFORMANCE IN ABSENCE OF ATTACK

The proposed SPARQ algorithm's hyperparameters and counterparts are optimized using the Grid Search method. Table 3 summarizes the SPARQ hyperparameters. Notably, the algorithm employs a learning rate of 0.0003 and a discount factor gamma of 0.99 for value estimation. The temperature parameter is set to 1, balancing stochastic exploration and exploitation. The replay buffer size is configured to 8192, and the batch size is set to 64 for efficient training. Target network updates are performed every 10 steps, with one gradient step taken per training iteration. Additionally, reward scaling is applied with a factor of 0.1, and the minimum value clipping is set at $-1e{+}06$ to stabilize learning.

TABLE 3. SQL hyperparameters used in our experiments.

Hyperparameter	Value
learning_rate	0.0003
gamma	0.99
alpha	0.2
temperature	1
buffer_size	8192
batch_size	64
target_update_freq	10
gradient_steps	1
reward_scaling	0.1
min_value	-1e+06

1) IEEE 14-BUS SYSTEM: FEASIBILITY STUDY

Numerical simulations are conducted using the IEEE-14 bus system. The modified IEEE 14-bus system represents a portion of the real American Electric PS as seen in Fig. 2 [25].

The average rewards and computation time comparison between different agents is provided in Table 4.

According to Table 4, SPARQ achieves the highest average reward of 320.25 while also demonstrating the lowest computation time of 10.19 seconds, indicating its superior efficiency and performance. Dueling DQN follows with an average reward of 61.1466 and a computation time of 58.1978 seconds, showing a balance between reward and time. DDQN performs slightly better than DQN in terms

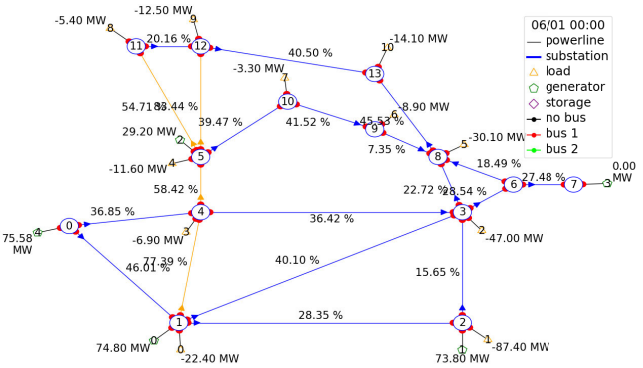

FIGURE 2. Modified IEEE 14-bus system used in the evaluation.

TABLE 4. Comparison of agent performance for the IEEE 14-bus system.

Agent	Avg. Reward	Computation Time (s)
DQN	58.3413	47.5151
DDQN	60.5717	62.6129
Dueling DQN	61.1466	58.1978
DDPG	35.6587	30.6819
DuelQSimple	41.6011	16.275
SPARQ	320.25	10.19

of reward, achieving 60.5717 compared to DQN's 58.3413, but requires more computation time at 62.6129 seconds compared to DQN's 47.5151 seconds. The DuelQSimple agent demonstrates moderate performance with a reward of 41.6011 and a significantly reduced computation time of 16.275 seconds. DDPG exhibits the lowest average reward of 35.6587 and a computation time of 30.6819 seconds, indicating its relatively weaker performance compared to the other agents. This comparison highlights the trade-offs between reward maximization and computational efficiency among the agents.

Fig. 3 illustrates the number of trainable parameters for five reinforcement learning agents: DQN, DDQN, Dueling DQN, DuelQSimple, and SPARQ. DQN and DDQN exhibit the highest model complexity, each having approximately 23,581 trainable parameters, closely followed by Dueling DQN with 22,750 parameters. On the other hand, DuelQSimple and SPARQ are less complex, with 16,733 and 15,581 parameters, respectively. This reduction in parameters for the latter two methods reflects their simpler architectures, which likely contribute to their lower computational and memory requirements.

Table 5 compares five agents—DQN, DDQN, Dueling DQN, DuelQSimple, and SPARQ—based on memory consumption, CPU utilization, and the number of parameters. DQN has the highest memory usage at 93.08 MB, which is substantially higher than the other methods, while SPARQ and DDQN demonstrate much lower memory footprints of 3.69 MB and 4.87 MB, respectively. The CPU utilization values range between 85.50% for DQN and 100.55% for DuelQSimple, with most methods utilizing over 96% of

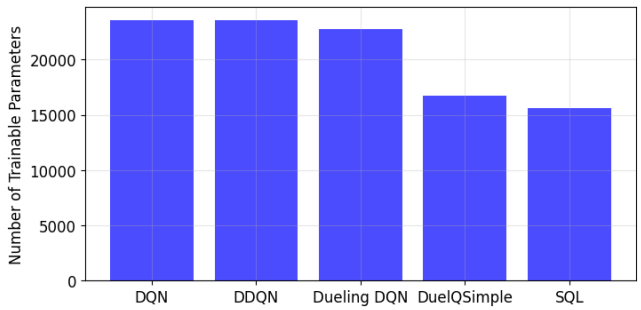

FIGURE 3. Number of trainable parameters.

CPU capacity. Regarding model complexity, measured by the number of parameters, DQN and DDQN have the highest values at 23,581 parameters, followed by Dueling DQN with 22,750 parameters. DuelQSimple and SPARQ have fewer parameters at 16,733 and 15,581, respectively, indicating lower computational complexity. Notably, DuelQSimple has an anomalous memory usage value of -0.00 MB, which may be an error or an artifact of measurement. Overall, the table highlights the trade-offs among the agents in terms of resource efficiency and computational demands, with SPARQ offering a good balance of low memory usage and moderate CPU demand, while DQN stands out for its higher memory and lower CPU utilization.

TABLE 5. Comparison of agents based on memory, CPU utilization, and number of parameters.

Agent	Memory	CPU	# Params
DQN	93.08 MB	85.50%	23,581
DDQN	4.87 MB	97.50%	23,581
Dueling DQN	6.15 MB	98.40%	22,750
DuelQSimple	N/A	100.55%	16,733
SPARQ	3.69 MB	96.75%	15,581

2) IEEE 118-BUS SYSTEM: SCALABILITY STUDY

Comparative tests are implemented on an IEEE 118-bus system to assess the scalability of our proposed method. Table 6 compares the performance of various DRL methods in terms of average reward and computation time. SPARQ achieved the highest average reward of 433.99 while maintaining a computation time of 9.1 seconds, demonstrating a strong balance between performance and efficiency. DuelQSimple followed closely with an average reward of 381.12 but required the highest computation time of 16.08 seconds. DDQN also performed well with an average reward of 376.56 and a computation time of 14.57 seconds. In contrast, DQN recorded a moderate reward of 171.44 with a computation time of 10.19 seconds. Dueling DQN showed the lowest average reward of 85.3 but had one of the shortest computation times at 8.59 seconds. Overall, SPARQ emerges as the most effective method in this comparison, offering superior rewards with reasonable computational demands.

TABLE 6. Performance comparison of DRL methods on the IEEE 118-bus system.

Method	Average Reward	Time (s)
DQN	171.44	10.19
DDQN	376.56	14.57
Dueling DQN	85.3	8.59
DuelQSimple	381.12	16.08
SPARQ	433.99	9.1

Fig. 4 compares the computation times across different methods or agents for the IEEE 118-bus system. The bar chart shows that SPARQ has the shortest computation time at approximately 5 seconds, while DuelQSimple requires the longest at around 21 seconds. The DQN, DDQN, and Dueling DQN - fall between these extremes, with computation times ranging from roughly 13 to 19 seconds. This comparison suggests that traditional SPARQ-based approaches may be more computationally efficient than various deep Q-learning methods for resilient PS operations.

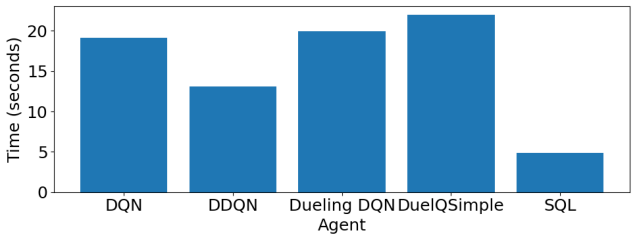

FIGURE 4. Computation time comparison on the IEEE 118-bus system.

Fig. 5 presents a boxplot comparison of computation times across different agent types in an IEEE 118-bus system. From the visualization, the DQN agent shows moderate computation time with a median around 17-18 units. The DDQN demonstrates notably lower computation time, with its median approximately at 12 units, suggesting improved efficiency. The Dueling DQN and DuelQSimple agents exhibit higher computation times with medians around 20-22 units and also show greater variability in their performance, as indicated by their larger box sizes and whiskers. Finally, the SPARQ agent appears to have the most consistent performance, though its exact values are not clearly visible in the plot.

3) REDUCED IEEE 118-BUS SYSTEM: ADAPTABILITY STUDY

Further tests are conducted to evaluate the performance of the proposed method using the reduced 118-bus system. Table 7 presents the comparison of agent performance metrics for the reduced IEEE 118 bus system. The comparison of agent performance metrics for the reduced IEEE 118 bus system reveals distinct variations in both average reward and computation time across different agents. The SQL agent achieved the highest average reward of 483.14, significantly outperforming all others, while also demonstrating the shortest computation time of 21.59 seconds. The Dueling

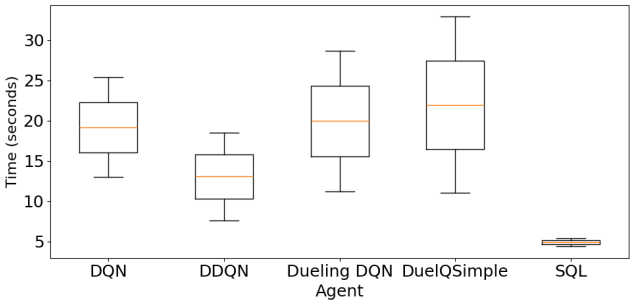

FIGURE 5. Computation time boxplot comparison for the IEEE 118-bus system.

DQN agent followed with an average reward of 54.95 and a moderate computation time of 110.50 seconds. The DDPG agent achieved an average reward of 51.3629 with a relatively efficient computation time of 49.31 seconds. The DQN and DDQN agents recorded average rewards of 48.34 and 45.09, respectively, with computation times of 157.66 and 62.08 seconds. Finally, the DuelQSimple agent demonstrated the lowest average reward of 41.51 but achieved a computation time of 34.48 seconds, placing it among the more computationally efficient agents. These results highlight the trade-offs between reward optimization and computation efficiency across the tested agents.

TABLE 7. Comparison of agent performance metrics on the reduced IEEE 118-bus system.

Agent	Average Reward	Computation Time (s)
DQN	48.3496	157.66
DDQN	45.0998	62.08
Dueling DQN	54.9550	110.50
DDPG	51.3629	49.31
DuelQSimple	41.5143	34.48
SPARQ	483.140	21.59

The SPARQ performance across the three bus systems for ten trails is presented in Table 8. The 14-bus system demonstrates lower rewards and shorter computation times, indicating its suitability for quick simulations with less complexity. The reduced 118-bus system shows significantly higher rewards in some trials but at the cost of increased computation time and longer episode lengths. The 118-bus system balances complexity and performance, with rewards and computation times falling between those of the other two systems.

Fig. 6 illustrates the comparison of reward values across different bus systems over 10 trials. The graph reveals significant performance variations across trials, with the standard 118-bus system demonstrating the highest volatility, reaching peak rewards of around 750 units in trial 3 and experiencing another notable spike near 650 units in trial 8. The 14-bus system and modified 118-bus system show relatively more stable performance patterns, generally maintaining reward values between 200 and 400 units throughout the trials. The modified 118-bus system appears to have slightly more consistent performance compared to the original 14-bus system, suggesting that the modifications may have contributed to improved stability in the reward outcomes. By trial 10, all three systems converge with rewards ranging between 250 and 450 units, though the standard 118-bus system maintains a slightly higher final value.

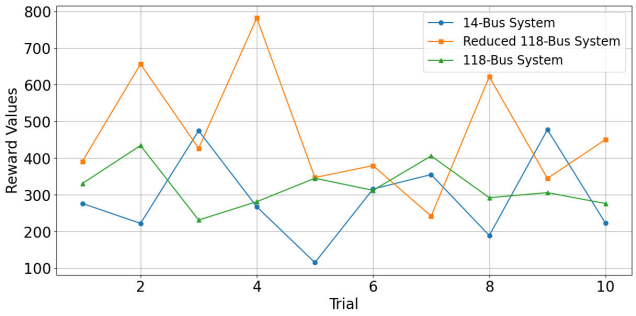

FIGURE 6. Reward values across different bus systems.

C. PERFORMANCE IN PRESENCE OF ATTACK

While performance in standard scenarios is important, resilience against adversarial conditions is a critical aspect of evaluating the robustness of any control method. This subsection presents the efficiency metrics summary for the SPARQ agent under two scenarios: Normal (no attacks) and Attacked (with a 30% probability of cyber-attacks). Inspired by energy-based policy gradients, SPARQ introduces an entropy term in the objective function to balance exploitation with exploration, reducing the risk of local optima in high-dimensional control spaces. To make the learning agent resilient to malicious data injections, the architecture incorporates random perturbations into measurements with controllable probability and magnitude. During the training phase, the agent encounters these possible corruptions and learns adaptive responses to maintain robust performance.

Fig. 7 illustrates the evolution of average episode rewards during training for both normal and attacked scenarios, with shaded regions representing standard deviations. Despite potential reward volatility in both cases, the higher overall amplitude of the normal curve suggests better stability and performance relative to the attacked variant. In contrast, the attacked mean reward curve exhibits dampened rewards and larger oscillations in certain intervals, indicative of the agent's struggle when data integrity is compromised. Nonetheless, SPARQ's entropy-based exploration helps the agent adapt to the presence of uncertainties and maintains meaningful policy improvements over multiple episodes. This ability to adjust under attack conditions demonstrates the effectiveness of SPARQ's design in enhancing power grid resilience.

Table 11 presents a comparison of the performance of the SPARQ agent on different IEEE bus systems. By analyzing how different grid configurations influence the model's response, this comparison provides insight into scalability

TABLE 8. SPARQ performance across different bus systems.

Trial	14-bus system			Reduced 118-bus system			118-bus system		
	Reward	Avg Episode Length	Time (s)	Reward	Avg Episode Length	Time (s)	Reward	Avg Episode Length	Time (s)
1	276.44	4.29	9.15	391.04	5.89	10.5	330.87	6.43	8.46
2	222.12	3.64	8.98	656.39	47	59.45	434.34	46.5	46.34
3	474.65	6.29	8.36	426.71	44.94	60.36	231.3	4.71	6.78
4	267.35	34.4	67.06	781.86	11	10.61	281.32	5.55	11.46
5	115.29	2.62	4.24	347.13	5.22	9.09	345.17	7	7.63
6	315.76	22.62	32.51	379.62	27.85	31.4	312.23	42.92	42.2
7	355.2	5.22	7.91	242.05	4.12	7.31	405.94	47	45.83
8	189.17	3.36	7.31	621.81	36.1	32.37	292.16	21.58	25.41
9	477.58	6.4	11.71	344.68	5.33	9.13	305.93	13.83	18.64
10	222.43	3.56	6.14	450.86	6.12	9.31	276.25	5.29	7.41

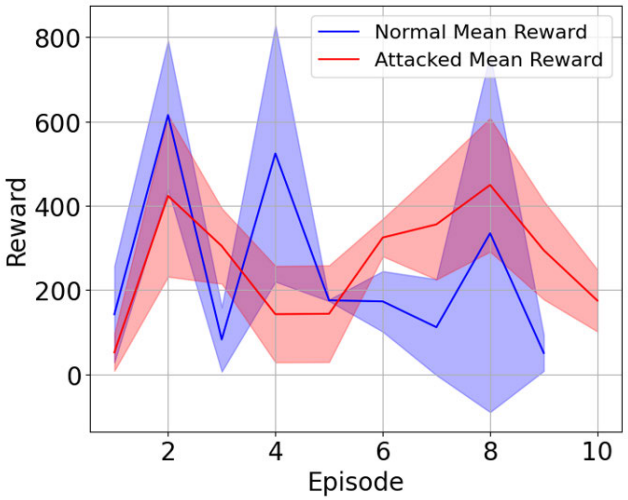

FIGURE 7. Episode rewards over training for SPARQ.

and generalizability. The average reward achieved in the 118-bus system is significantly higher at 384.10, compared to 267.02 in the 14-bus system, indicating that the agent performed better on the larger system. This could be attributed to the increased opportunities for optimization or different environmental dynamics in the larger grid. The average entropy values are nearly identical, with 5.02 for the 14-bus system and 5.03 for the 118-bus system, suggesting that the agent maintained a consistent balance between exploration and exploitation across both systems. Interestingly, the computation time for the 118-bus system is lower, taking 8.74 seconds compared to 11.36 seconds for the 14-bus system, which may be due to shorter episodes or more efficient learning dynamics in the larger environment.

Table 9 shows two key scenarios - normal and attacked - for the SPARQ agent for different IEEE bus systems.

Based on Table 9, the performance analysis reveals distinct patterns across different IEEE bus systems under normal and attacked scenarios. In the IEEE 14-bus system, the normal scenario outperforms the attacked scenario with a higher average reward (396.07 vs 361.55), though the attacked

TABLE 9. Performance metrics for various IEEE systems under normal and attacked scenarios.

Scenario	Average Reward	Time (s)
IEEE 14-bus system		
Normal Scenario	396.07	8.80
Attacked Scenario	361.55	6.50
Reduced 118-bus system		
Normal Scenario	336.06	8.40
Attacked Scenario	401.39	25.67
IEEE 118-bus system		
Normal Scenario	365.19	9.74
Attacked Scenario	234.90	9.95

scenario processes slightly faster. For the reduced 118-bus system, interestingly, the attacked scenario shows better performance with an average reward of 401.39 compared to 336.06 in the normal scenario but requires significantly more processing time (25.67s vs 8.4s). The full IEEE 118-bus system demonstrates the most pronounced impact of attacks, where the normal scenario maintains a solid average reward of 365.19 while the attacked scenario's performance drops substantially to 234.90, with both scenarios requiring similar processing times around 9.8 seconds. Table 10 provides key performance indicators, including training time, memory usage, CPU utilization, the number of trainable parameters, average step time, and the final reward.

TABLE 10. Efficiency metrics for SPARQ model using the IEEE 14-bus system.

Agent	Time (s)	Memory (MB)	CPU (%)	Param.	Step (s)
Normal	6.86	11.53	101.3	32,79	0.167
Attacked	6.43	8.10	99.3	32,79	0.172

The results in Table 10 provide insights into the performance and efficiency of the SPARQ agent under the two scenarios. The normal scenario takes slightly longer to train (6.86 seconds) compared to the attacked scenario (6.43 seconds), which may be attributed to differences in the number of episodes completed or the complexity

introduced by attacks affecting convergence rates. Memory usage is higher in the Normal scenario (11.53 MB) than in the attacked scenario (8.10 MB), potentially due to differences in replay buffer states or additional computations in the presence of attacks. CPU utilization is slightly greater in the Normal scenario (101.3%) compared to the attacked scenario (99.3%), possibly reflecting the absence of attack-related processing in the Normal case. Both scenarios involve the same number of trainable parameters (32,797), as the neural network architecture remains consistent. The average step time is marginally higher in the attacked scenario (0.1723 seconds) than in the Normal scenario (0.1673 seconds), suggesting that handling attack-related computations slightly impacts efficiency. In summary, the presence of cyber-attacks results in marginally lower memory and CPU usage while slightly increasing step time, indicating subtle differences in computational demands across the two scenarios.

Table 11 highlights the performance of the SPARQ model across three different IEEE bus systems. From Table 11, the model achieves the highest reward of 464.22 on the reduced 118-bus system, indicating its ability to perform well in moderately complex environments. The entropy values increase progressively from the 14-bus system to the 118-bus system, reflecting greater uncertainty or exploration in larger grid environments. Computation time also rises with system complexity, peaking at 23.95 seconds for the reduced 118-bus system. These results demonstrate the model's adaptability and computational demands across varying grid complexities, with the reduced 118-bus system presenting the most challenging yet rewarding environment for the SPARQ model.

TABLE 11. Performance comparison of SPARQ across different bus systems.

System	Reward	Entropy	Time (s)
14-bus system	291.60	5.03	16.33
Reduced 118-bus system	464.22	6.19	23.95
118-bus system	321.55	7.30	22.02

D. LIMITATIONS

The proposed approach may have some potential limitations. First, the integration of SPARQ with traditional grid control systems requires careful coordination of different operational timescales, where fast-acting local controls handle immediate voltage deviations. A hierarchical control structure enables seamless interaction between conventional methods (droop control, tap changers) and the RL agent, with proper state estimation bridging supervisory control and data acquisition (SCADA) measurements and RL requirements. The system complements existing methods by optimizing their parameters while keeping traditional control logic as a safety backbone. Success relies on robust technical infrastructure, comprehensive operator training, and extensive validation through simulation and pilot deployments. The future work of this study aims to integrate domain-adaptive anomaly detection modules and hierarchical RL to enhance dynamic adaptability.

Second, the proposed SPARQ method may hamper the comprehensive detection and mitigation of multi-stage infiltration and stealth attacks in modern power grids. The reliance on fixed security measures can hinder responsiveness to advanced or novel attacks. Although standard intrusion detection and anomaly-based safeguards can identify some threats, advanced persistent threats continue to evolve, exploiting zero-day vulnerabilities and camouflage strategies. In future work, we aim to integrate robust defense-in-depth strategies by combining distributed monitoring with dynamic risk assessment, guided by real-time telemetry. This approach will not only enhance the coverage of potential attack vectors but also ensure continuous adaptation of countermeasures as threat patterns shift. Such synergy fosters the evolution of detection and protection layers. Additionally, employing federated learning mechanisms could enable collaborative intelligence across multiple substations without exposing sensitive data, following guidelines such as NIST SP 800-82 [26]. We also intend to leverage the IEEE 1547-2018 standard to harmonize cybersecurity measures with distributed energy resource operations [27].

VII. CONCLUSION

This paper introduced SPARQ for resilient voltage regulation in modern power grids. By integrating entropy-regularized exploration, SPARQ balanced exploration and exploitation, allowing it to adapt to dynamic grid conditions and effectively mitigate false data injection attacks. The proposed preprocessing pipelines and off-policy learning techniques further enhanced its computational efficiency and convergence stability. Empirical results highlighted SPARQ's resilience to adversarial scenarios, demonstrating its ability to maintain robust performance under cybersecurity breaches. Evaluated on the IEEE 14-bus, reduced 118-bus system, and IEEE 118-bus systems, SPARQ outperformed traditional RL methods such as DQN, DDQN, Dueling DQN, DuelQSimple, and DDPG. The proposed approach achieved higher rewards, faster convergence, high privacy-preserving aspects, and optimum computational demands. Moving forward, this research will explore extending SPARQ to multi-agent architectures for coordinated control, integrating domain-specific anomaly detection systems, and leveraging transfer learning for abrupt topology changes. Additionally, Future research focuses on extending SPARQ's capabilities to address more sophisticated and practical cyberattacks, such as coordinated multi-point attacks and stealthy data manipulation techniques, to enhance its applicability in real-world scenarios.

REFERENCES

[1] L. Xie, T. Huang, P. R. Kumar, A. A. Thatte, and S. K. Mitter, "On an information and control architecture for future electric energy systems," *Proc. IEEE*, vol. 110, no. 12, pp. 1940–1962, Dec. 2022.

[2] M. E. Eddin, M. , H. Abu-Rub, M. Shadmand, and M. Abdallah, "Novel functional community detection in networked smart grid systems-based improved Louvain algorithm," in *Proc. IEEE Texas Power Energy Conf. (TPEC)*, Feb. 2023, pp. 1–6.
[3] L. Xue, T. Niu, H. Ge, J. Zhang, Y. Xue, S. Fang, G. Chen, and Z. Wang, "A joint distributed optimization framework for voltage control and emergency energy storage vehicle scheduling in community distribution networks," *IEEE Trans. Ind. Appl.*, vol. 60, no. 4, pp. 5317–5330, Aug. 2024.
[4] Y. Zhang, M. Yue, J. Wang, and S. Yoo, "Multi-agent graph-attention deep reinforcement learning for post-contingency grid emergency voltage control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 3, pp. 3340–3350, Mar. 2024.
[5] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 4120–4123, Sep. 2020.
[6] M. , S. S. Refaat, A. Ghrayeb, and H. Abu-Rub, "Short-term dynamic voltage stability status estimation using multilayer neural networks," in *Proc. IEEE Texas Power Energy Conf. (TPEC)*, Feb. 2023, pp. 1–6.
[7] F. S. Al-Ismail, "A critical review on DC microgrids voltage control and power management," *IEEE Access*, vol. 12, pp. 30345–30361, 2024.
[8] A. U. Rehman, Z. Ullah, H. S. Qazi, H. M. Hasanien, and H. M. Khalid, "Reinforcement learning-driven proximal policy optimization-based voltage control for PV and WT integrated power system," *Renew. Energy*, vol. 227, Jun. 2024, Art. no. 120590.
[9] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive power system emergency control using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1171–1182, Mar. 2020.
[10] P. Li, J. Hao, H. Tang, X. Fu, Y. Zhen, and K. Tang, "Bridging evolutionary algorithms and reinforcement learning: A comprehensive survey on hybrid algorithms," *IEEE Trans. Evol. Comput.*, early access, Aug. 14, 2024, doi: 10.1109/TEVC.2024.3443913.
[11] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Trans. Power Syst.*, vol. 35, no. 3, pp. 1990–2001, May 2020.
[12] K. E. Antoniadou-Plytaria, I. N. Kouveliotis-Lysikatos, P. S. Georgilakis, and N. D. Hatziargyriou, "Distributed and decentralized voltage control of smart distribution networks: Models, methods, and future research," *IEEE Trans. Smart Grid*, vol. 8, no. 6, pp. 2999–3008, Nov. 2017.
[13] N. Duan, C. Huang, C.-C. Sun, and L. Min, "Smart meters enabling voltage monitoring and control: The last-mile voltage stability issue," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 677–687, Jan. 2022.
[14] Y. Li, C. Yu, M. Shahidehpour, T. Yang, Z. Zeng, and T. Chai, "Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects," *Proc. IEEE*, vol. 111, no. 9, pp. 1055–1096, Sep. 2023.
[15] S. Liu, S. Liu, X. Li, Y. Gu, Z. Li, C. Yang, Y. Zhang, J. Hu, R. Tan, and C. Liu, "Neyman-Pearson umbrella algorithm-based static voltage stability assessment with misclassification restriction: An integrated data-driven scheme," *IEEE Trans. Ind. Informat.*, vol. 19, no. 12, pp. 11391–11402, Dec. 2023.
[16] M. , H. Abu-Rub, S. S. Refaat, I. Chihi, and F. S. Oueslati, "Deep learning in smart grid technology: A review of recent advancements and future prospects," *IEEE Access*, vol. 9, pp. 54558–54578, 2021.
[17] A. Ali, C. Li, and B. Hredzak, "Dynamic voltage regulation in active distribution networks using day-ahead multi-agent deep reinforcement learning," *IEEE Trans. Power Del.*, vol. 39, no. 2, pp. 1186–1197, Apr. 2024.
[18] P. Chen, S. Liu, X. Wang, and I. Kamwa, "Physics-guided multi-agent adversarial reinforcement learning for robust active voltage control with peer-to-peer (P2P) energy trading," *IEEE Trans. Power Syst.*, vol. 39, no. 6, pp. 7089–7101, Nov. 2024.
[19] S. Wang, L. Du, X. Fan, and Q. Huang, "Deep reinforcement scheduling of energy storage systems for real-time voltage regulation in unbalanced LV networks with high PV penetration," *IEEE Trans. Sustain. Energy*, vol. 12, no. 4, pp. 2342–2352, Oct. 2021.
[20] O. Zandi and J. Poshtan, "Voltage control of a quasi Z-source converter under constant power load condition using reinforcement learning," *Control Eng. Pract.*, vol. 135, Jun. 2023, Art. no. 105499.
[21] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2017, pp. 1352–1361.
[22] R. Bolgaryn, G. Banerjee, S. Meinecke, H. Maschke, F. Marten, M. Richter, Z. Liu, P. Lytaev, B. Alfakhouri, J. M. Kisse, and D. Lohmeier, "Open source simulation software pandapower and pandapipes: Recent developments," in *2023 Open Source Modelling and Simulation of Energy Systems (OSMSES)*. Piscataway, NJ, USA: IEEE, 2023, pp. 1–8.
[23] B. Donnot, "Grid2Op-a testbed platform to model sequential decision making in power systems," GitHub Repository, San Francisco, CA, USA, Tech. Rep., 2020.
[24] B. Pang, E. Nijkamp, and Y. Wu, "Deep learning with TensorFlow: A review," *J. Educ. Behav. Statist.*, vol. 45, no. 2, pp. 227–248, Sep. 2019.
[25] I. Damjanović, I. Pavić, M. Brčić, and R. Jerčić, "High performance computing reinforcement learning framework for power system control," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT)*, Jan. 2023, pp. 1–5.
[26] K. Stouffer, K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, and A. Sherule, "Guide to operational technology (OT) security," U.S. Dept. Commerce, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NIST SP 800 82r3, 2023.
[27] E. Desardén-Carrero, R. Darbali-Zamora, and E. E. Aponte-Bezares, "Analysis of commonly used local anti-islanding protection methods in photovoltaic systems in light of the new IEEE 1547–2018 standard requirements," in *Proc. IEEE 46th Photovoltaic Spec. Conf. (PVSC)*, Jun. 2019, pp. 2962–2969.

(Member, IEEE) received the M.Eng. degree in energy engineering from the National Engineering School of Monastir (ENIM), University of Monastir, Tunisia, in 2018, the Ph.D. degree in electronics engineering from the National Institute of Applied Sciences and Technology (INSAT), University of Carthage, Tunisia, in 2021, and the second Ph.D. degree in electrical and computer engineering from Texas A&M University (TAMU), College Station, TX, USA, in 2024.

He completed preparatory studies in mathematics and physics from the Preparatory Institute for Engineering Studies of Nabeul (IPEIN), University of Tunis El Manar, Tunisia, in 2015. He has eight years of hands-on experience in applying deep learning and machine learning strategies to tackle real-world problems. During his work at Texas A&M University at Qatar, he is the lead author of more than 50 peer-reviewed journal articles and conference publications and three book chapters. His H-index is 15 and his work has been cited more than 1100 times. His research interests include AI applications for demand flexibility, cybersecurity in smart grids, and innovative prediction models. He was a recipient of the Outstanding Student Research Excellence Award, in 2021, the Thomas W. Powell'62 and Powell Industries Inc., Fellowship Award, in 2024, the Best Paper Presentation Recognition from the IECON 2024, and the Richard E. Ewing Award for Excellence, in 2024, for his research contributions.