Continuous Voltage Regulation in Active Distribution Networks Using Twin Delayed Deep Deterministic Policy Gradient

Jiangyan Chen[1,2], Long Kou[2,3, *,] Yuhua Wang[1], Yilin Liu[1]
[1]School of Artificial Intelligence and Electrical Engineering, Guangzhou College of Applied Science and Technology, Guangzhou, Guangdong, 511370, China
[2]International College, National Institute of Development Administration, Bangkok, 10240, Thailand
[3]School of Management, Guangzhou Huashang College, Guangzhou, Guangdong, 511300, China
E-mail: 18011832010@163.com
*Corresponding author

Keywords: active distribution networks (ADNs), electric vehicles (EVs), renewable energy sources (RESs), twin delayed deep deterministic policy gradient (TD3), voltage regulation

Received: August 31, 2025

Active Distribution Networks (ADNs), characterized by high levels of penetration by Electric Vehicles (EVs) and renewable energy sources (RES), lead to a high degree of uncertainty and control challenges for many system operators. Most previous studies examined methods for controlling voltage, but the Finite Action Space (AS) set a limit on the control granularity and scalability of the methods investigated. With this in mind, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm is presented, a Continuous-Action (CA) actor-critic reinforcement learning algorithm in which the same system model, reward structure, and EV participant constraints are kept for a direct comparison. The TD3-based controller provided real-valued control actions for distribution system resources, fully approving of the additional flexibility provided by continuous ASs. The issue of value overestimation was overcome by combining TD3 with twin Critic Networks (CNs), while the features of target smoothing and delayed policy updates are also introduced to strengthen the stability and convergence of the learning algorithm. The simulations with the IEEE 33-bus (IEEE-33) and IEEE 123-bus (IEEE-123) systems highlighted improvements in voltage control granularity, convergence speed, and scalability across uncertain State Spaces (SSs).

Povzetek: Abalizirano je uravnavanja napetosti v aktivnih distribucijskih omrežjih z visoko penetracijo EV-jev in obnovljivih virov. Predlaga metodo TD3, ki z zveznimi dejanji, dvojnima kritikoma ter zakasnjenimi posodobitvami politike zmanjšuje precenjevanje Q-vrednosti in izboljša stabilnost. Simulacije (IEEE-33, IEEE-123) pokažejo boljšo točnost, hitrejšo konvergenco in večjo razširljivost kot DQN/DDQN.

1 Introduction

The increasing installation of Distributed Energy Resources (DERs) like Photovoltaics (PVs), along with the surge in EV deployments, is exacerbating the voltage stability challenges of the ADNs [1]. This new situation of uncertainties through bidirectional power flow and dynamic load behavior will require adaptive Voltage Regulation (VR) strategies that are faster and more adaptable than model-based or rule-based control systems. Deep reinforcement learning (DRL) has shown promise for this type of VR; however, discrete-action formulations such as the Averaged Weighted Double Deep Q-Network (AWDDQN) have limits with respect to coarse control resolution and scalability bottlenecks [2]. Therefore, the goal of this study is to provide an alternative to these limitations and offer a CA voltage control technique using the TD3 algorithm.

Conventional VR methods manage EV charging/discharging profiles or utilize multiple grid-side resources to improve voltage profiles. For instance, the model-based approaches leverage convex optimization and heuristic approaches to manage the EV charging actions for cost reduction and voltage support [3], [4]. However, the temporal uncertainty of access to EVs and spatial diversity of the charging needs introduce stochastic schedulable capacity constraints that complicate the control model [5]. To simplify this complexity, hierarchical and distributed frameworks are suggested. Some works model EV aggregators (EVAs) to reduce the complexity of optimizations and support scalable control [5]. Other works applied predictive and Monte Carlo methods to predict EV loads and optimize demand side voltage with multistage or heuristic solvers [6], [7], [8]. Other works use game-theoretical mechanisms [9] and distributed model predictive control (MPC), which is necessary for peer-to-peer coordination [10] to handle decentralized VR with EVA bidding or reactive power optimization.

Despite their effective structures, many of these strategies rely on correct predictions or centralized processing, restricting their versatility when the situations evolve rapidly. To circumnavigate these limitations, model-free reinforcement learning (RL) approaches are being studied for voltage control, as well. RL methods, such as Q-learning, were used to study the action of tap changer and capacitor switching [11], [12]. However, Q-learning has not been practical in ADNs due to the "curse of dimensionality," whereby a large number of controllable nodes and dynamic states have made traditional Q-learning impossible. Therefore, DRL methods such as Deep Q-Networks (DQN) were proposed. Continuous Q-learning and mapping to DQN facilitate the mapping of the state-AS in this new setting using neural networks to act as an interface. DQN has been mainly utilized for coordinating the active and reactive power outputs of energy storage and inverter-based resources for control of voltage [13], [14], [15]. A few studies have used DQN in combination with other DQN models to enable multi-timescale control or DQN as part of an approach with deterministic policy gradients to shape EV charging profiles for stabilizing voltages [16], [17].

research for more resilient and scalable VR methods, especially with high levels of DER and EV penetration. Zhang et al. [18] Recently, we have leveraged DRL via a physics-informed multi-agent approach where each inverter agent learns its control locally, with knowledge of the grid structure. This leads to disturbance-resistant and scalable decentralized VR using PV inverters. Golgol and Pal [19] propose a DRL-based voltage controller that learns solely from feeder-head state estimation to make quick control decisions with minimal communication infrastructure. Shi et al. [20] develop a data-driven affine control method that describes the relationship between each inverter's reactive power adjustment and its active power. In doing this they aim to learn new solutions to deal with solar uncertainty from real time distributed learning and consensus algorithms. Hierarchical schemes have also been further developed in the VR domain. Dutta et al. [21] present a receding horizon control method that operates in three stages: validating smart inverter models, jointly optimizing slow and fast voltage devices, whilst integrating an EV schedule as a grid supporting service. Similarly, Ma et al. [22] develop a hierarchical method using distributed energy storage systems, where the components have been solved using centralized model predictive control (MPC) with a full knowledge of the grid and provides set points on a global scale, whilst the decentralized agents follow certain heuristics to alter their instantaneous output, with limited communication. Zhang et al. [23] represented their work as multi-energy coordination, a strategy for integration through the use of a Power-to-Hydrogen system as a controllable load, absorbing excess generation from PV systems to help support voltage control. Despite advances in voltage control strategies in ADNs using DRL and optimization-based methods, there are two major challenges:

1. Control resolution and scale: Discrete-action DRL implementations such as DQN, DDQN, and AWDDQN require AS discretization, creating exponentially growing as more devices are adopted, resulting in high, prohibitive computational expense, and not a desirable control resolution in systems with continuous dynamics.

2. Learning instability and bias under uncertainty: The tendency to overestimate or underestimate action values is particularly problematic in stochastic settings, like many ADNs with high EV and renewable penetration. While AWDDQN improves DDQN using historical value averaging to some extent, it still relies on discrete action estimators and is sensitive to added noise and variability.

To address these difficulties, this paper offers a TD3-based voltage control framework, where the two major contributions are:

1. Continuous and scalable voltage control: The TD3 controller takes action as a real-valued output for each grid resource, avoiding discretization, and allowing scalable coordination in very high-dimensional ADNs in continuous, precise space.

2. Stability and bias-resistance to learning: Dual critics and target-action delays promote stable policies and reduced estimation bias, so that reliable training can take place in uncertain environments with more variability due to EVs and renewables.

The subsequent sections of this paper are structured in the following manner. Section II elucidates the theoretical basis of the TD3 algorithm, including its architecture and update rules, and its appropriateness for power systems control within the uninterrupted action area. Section III gives the TD3-based voltage control framework, including the system model, definition of the states and actions, reward design, and the hierarchical coordination framework for the EVA. Section IV details the simulation setup and examines the effectiveness of the suggested technique using altered versions of the IEEE-33 and IEEE-123 test systems, in terms of simulation training behavior, convergence behavior, and real-time VR ability. In conclusion, Section V recaps the key results to bring the study to a close and offers recommendations for subsequent research.

2 Principle of TD3 algorithm

2.1 DDPG algorithm

TD3 uses the Deep Deterministic Policy Gradient (DDPG) algorithm - as an actor-critic method for continuous ASs - as its baseline. The agent in RL is presented a state s, chooses an action a according to its policy, receives a reward r, and transitions to a next state s'. The process can be outlined as a Markov Decision Process (MDP) with a SS S, AS A, reward function $r(s,a)$, transition probability $P(s' \mid s,a)$, and discount factor γ. The objective of the agent is to determine a superior action-value function $Q(s,a)$ that fulfills the Bellman optimality equation.

$$Q^*(s,a) = \mathbb{E}_{s'}\left[r(s,a) + \gamma \max_{a'} Q^*(s',a')\right] \qquad (1)$$

Where the expectation is taken over the next state s' In DDPG (which was created for continuous actions), the maximization over a' is not handled by brute force search, but instead by a differentiable policy (actor) network $\mu(s)$. DDPG trains both a CN $Q(s, a \mid \theta^Q)$ which is to approximate the action-value function, and an Actor Network (AN) $\mu(s \mid \theta^\mu)$ which is to approximate the optimal policy. The critic takes in a state-action combination and forecasts the Q-value, whereas the actor converts states into continuous actions. Furthermore, DDPG uses experience replay and Target Networks (TNs), like DQN, to make training more stable. DDPG uses slowly moving target parameters θ^Q_{targ} and θ^μ_{targ} which tracks the learned parameters θ^Q, θ^μ (that are updated via polyak averaging) in order to give consistency to the targets. With the actor μ being used for the next action, the target value from the critic can be written:

$$y_i^{DDPG} = r_i + \gamma Q(\, s'_i, \mu(s'_i \mid \theta^\mu_{\text{targ}}) \mid \theta^Q_{\text{targ}}\,), \tag{2}$$

Which indicates the anticipated return from state s'. The CN is trained to minimize the mean-squared Bellman error between this target. The critic loses at iteration i is:

$$L(\theta_i^Q) = \mathbb{E}_{(s,a,r,s')}\left[Q(s, a \mid \theta_i^Q) - y_i^{DDPG}\right]^2, \tag{3}$$

The parameters θ^Q of the critic are then updated by using gradient descent to reduce the critic loss while the AN will seek to maximize Q as estimated by the critic. When optimizing the actor policy, θ^μ is adjusted in the direction that improves $Q(s, \mu(s \mid \theta^\mu))$. In practice, the gradient of the actor is computed using deterministic policy gradient theory, which provides it.

$$\nabla_{\theta^\mu} J \approx \mathbb{E}_{s\sim\mathcal{D}}\left[\nabla_a Q(s, a \mid \theta^Q)|_{a=\mu(s|\theta^\mu)} \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\right] \tag{4}$$

Shifting the policy toward actions that have higher Q values. While DDPG will continually update the critic at each time step, it will also continue to update both the AN and TN. An unstable behavior can sometimes be observed in a naive implementation of DDPG, with a failure mode often occurring early during training, as the overestimation of values by the critic Q function begins, which incentivizes the actor to exploit this incorrect Q value, resulting in poor performance.

2.2 Overestimation and double-Q learning

Overestimation bias occurs because of the nature of function approximation and the max operator, which allows Q-values to be overestimated in a manner that is time invariant. In the DDPG context, the actor is always trying to maximize the output of the critic, so it is crucial that the overestimation of the Q value does not lead to the actor learning a divergent or suboptimal policy. A known solution (see Double Q-learning for discrete domains) is to use separate CNs, yielding two estimates $Q_1\,(s\,, a)$ and $Q_2\,(s, a)$. Each critic is separately initialized and trained. To combat overestimation, the minimum of the 2 critic values is selected as the target estimate, such that the clipped Double-Q technique restricts the impact of an overestimated Q-value on the target. Formally, let Q_1 and Q_2 be defined as the two critics with corresponding TN, and the target is then modified as follows:

$$y_i^{Double\text{-}Q} = r_i + \gamma \min_{j=1,2} Q_j \left(s'_i, \mu(s'_i \mid \theta^\mu_{targ}) \mid \theta^{Qj}_{targ}\right). \tag{5}$$

By using the minimum of Q_1, Q_2 for the bootstrap target, this method encourages underestimation (if a network overestimates, the minimum still predicts an estimate closer to the truth) and therefore avoids the compounding of bias. Each CN is trained using regression to the shared target y above, which minimizes $\left(Q_j(s, a \mid \theta^{Q_j}) - y\right)^2$. This dual-'critic' approach is similar to the DDQN approach in discrete RL, but instead is designed for continuous actor-critic methods. It has been observed that simply applying Double DQN to an actor-critic architecture is ineffective, as the slow-changing policy maintains the bias correlation. While having two independent critics (similar to Double-Q-learning) works better, even unbiased estimates with high variance will destabilize training. Thus, additional techniques are needed to stabilize the training process further.

2.3 TD3 algorithm improvements

The Twin Delayed DDPG (TD3) algorithm provides three main advancements to DDPG to combat overestimation and instability. The first advancement is the "Twin" critics with clipped double-Q as described above. Using 2 critics ensures TD3 reduces overestimation bias because it takes the smallest Q-value into account when computing the target value. The second advancement is the "Delayed" policy update. TD3 keeps the actor (policy network) update consistent with the critic delay. Although TD3 updates the AN (and TN) after every two critic updates (first the target critic optimizations followed by the Target Actor (TA) updates or just one normal critic update), this delay allows TD3 to update the actor only after allowing the critic optimization steps to converge closer to the correct Q-value function. This helps keep the actor from immediately chasing noisy Q-value estimates provided directly by a critic. The third and final improvement is the smoothing of the targeted policy. TD3 incorporates a minor degree of random noise to actions chosen by the target policy (AN) before being queried at the target CNs to find the target action value. In simpler mechanics, TD3 instead of getting from the TA its output $a' = \mu(s' \mid \theta^\mu_{\text{targ}})$ directly, it gets a sampled action:

$$a'(s') = clip\left(\mu(s' \mid \theta^\mu_{targ}) + \epsilon, a_{Low}, a_{High}\right), \epsilon \sim \mathcal{N}(0, \sigma), \epsilon \ clipped\ to\ [-c, c], \tag{6}$$

Where a_{Low} and a_{High} are the action bounds. This noisy action $a'(s')$ is then used in the target Q computation:

$$y_i^{TD3} = r_i + \gamma \min_{j=1,2} Q_j \left(s'_i, a'(s') \mid \theta^{Qj}_{targ}\right). \tag{7}$$

The added noise (with small variance σ) averages the metric Q-values over a neighborhood of actions, and therefore prevents the policy from exploiting sharp peaks or inconsistencies in the Q-function approximator over the policy. This target policy smoothing is a regularizer that makes the critic less vulnerable to erroneous high-value spikes. The policy will not tend to a small number of actions that may be overestimated, as the target value will be based on slightly perturbed actions, as well.

3 Voltage control method based on TD3 algorithm

3.1 Voltage control model for ADNs

The intention of VR in the ADN remains to minimize the deviation of the node voltages from their nominal values over a scheduling horizon. The optimization challenge can be stated as follows:

$$min\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{n=1}^{N} ts\cdot\left(U_{n,t}-U_{base}\right)^2\right) \tag{8}$$

The system must satisfy operational constraints on voltage magnitude and resource capacities:

$$\begin{aligned} U_{min} &\le U_{n,t} \le U_{max}\\ Q_{j,min} &\le Q_{j,t} \le Q_{j,max}\\ P_{j,t,min} &\le P_{j,t} \le P_{j,t,max}\end{aligned} \tag{9}$$

Where $P_{j,t}$ and $Q_{j,t}$ denotes the active and reactive power of the j-th controllable resource at time t, subject to upper and lower bounds.

3.2 Schedulable capacity of EVAs

Every Electric Vehicle Aggregator (EVA) is inherently a controllable active power source with time-varying characteristics based on the schedulable charging and discharging power capabilities of its fleet of EVs. For each EV, the schedulable charging capability (SCC), discharging capacity (SDC), and power bounds (SCP, SDP) are calculated as follows:

$$\begin{aligned}\mathrm{SCC}_{d,t} = C_d\left(\mathrm{SOC}_{d,s}-\mathrm{SOC}_{d,t}\right)\\ +\eta_c\left(t+ts-t_{d,s}\right)P_{d,c,\max}\end{aligned} \tag{10}$$

$$\mathrm{SCP}_{d,t} = \min\left(\frac{\mathrm{SCC}_{d,t}}{ts\cdot\eta_c}, P_{d,c,\max}\right) \tag{11}$$

$$\begin{aligned}\mathrm{SDC}_{d,t} = C_d\left(\mathrm{SOC}_{d,t}-\mathrm{SOC}_{d,\min}\right)\\ +\frac{\left(t_{d,e}-t-ts\right)P_{d,d,\max}}{\eta_d}\end{aligned} \tag{12}$$

$$\mathrm{SDP}_{d,t} = \min\left(\frac{\mathrm{SDC}_{d,t}\cdot\eta_d}{ts}, P_{d,d,\max}\right) \tag{13}$$

Aggregating over all EVs d in EVA l, the total capacity metrics are:

$$SCP_{l,t}=\sum_{d=1}^{N_{l,t}} SCP_{d,t}, SDP_{l,t}=\sum_{d=1}^{N_{l,t}} SDP_{d,t} \tag{14}$$

3.3 MDP formulation for TD3

The RL formulation follows an MDP framework with components defined as:

- SS $\mathcal{S}$:

$$s_t=\left\{U_{i,t}, SCP_{l,t}, SDP_{l,t}, P_{j,t}, Q_{j,t}\right\} \tag{15}$$

This includes voltage magnitudes, schedulable capacities, and current output of resources. Total state dimension is $N+2L+2J$.

- AS $\mathcal{A}$:

In contrast to AWDDQN, the TD3 algorithm outputs continuous-valued control actions for each device directly:

$$a_t=\{\hat{P}_{j,t}, \hat{Q}_{j,t}\}, j=1,\dots,J \tag{16}$$

These actions are constrained within their operational bounds:

$$\hat{P}_{j,t}\in\left[P_{j,t,min}, P_{j,t,max}\right], \hat{Q}_{j,t}\in\left[Q_{j,min}, Q_{j,max}\right] \tag{17}$$

3.4 Reward function design

The reward function within the TD3-based voltage control model reflects the objective of minimizing voltage deviations from the base value while applying soft penalties for violations. It originated from the objective function (1) in the first part of the paper, and was rewritten as:

$$r_t=-\frac{1}{N}\sum_{i=1}^{N}\lambda_i\cdot ts\cdot\left(U_{i,t+1}-U_{base}\right)^2 \tag{18}$$

Where λ_i is a penalty factor that is determined dynamically based on the size of the deviation at node i, and ts is the time step. The penalty factor is selected as:

$$\lambda_i=\begin{cases}1 & 0<\left|U_{i,t+1}-U_{base}\right|\le 0.01\\ 5 & 0.01<\left|U_{i,t+1}-U_{base}\right|\le 0.03\\ 10 & 0.03<\left|U_{i,t+1}-U_{base}\right|\le 0.05\\ 50 & \left|U_{i,t+1}-U_{base}\right|>0.05\end{cases} \tag{19}$$

In this way, the reward structure will allow the TD3 agent to reduce large voltage violations, while somewhat it will train on the best way to allocate resources towards the most important nodes within constraints, and account for operational priorities and tolerance in VR.

3.5 TD3 network architecture for voltage control

The TD3 algorithm deploys an actor-critic architecture comprising:

- AN $\mu(s\mid\theta^{\mu})$: produces continuous-valued actions (i.e., active and reactive setpoints for each device) given the system state.
- Two CNs $Q_1(s,a\mid\theta^{Q_1})$, $Q_2(s,a\mid\theta^{Q_2})$: estimate the expected cumulative reward from a given state-action pair. To lessen the tendency to overestimate, the lower of the two Q-values is utilized when updating policies.

The critics are trained by lessening the subsequent deficit:

$$L(\theta^{Q_i})=\mathbb{E}_{s,a,r,s'}[(Q_i(s,a)-y)^2], i=1,2 \tag{20}$$

Where the target Q-value, represented by y, is calculated as follows:

$$\begin{aligned} y\\ = r+\gamma\\ \cdot min\left(Q_1'\left(s', \mu'(s')\right), Q_2'\left(s', \mu'(s')\right)\right)\end{aligned} \tag{21}$$

Here, μ' and Q_i' are the TA and CNs, and γ is the discount factor. The actor is upgraded utilizing the deterministic policy gradient:

$$\begin{aligned}\nabla_{\theta^{\mu}}J \approx \mathbb{E}_s\left[\nabla_a Q_1(s,a\mid\theta^{Q_1})|_{a=\mu(s)}\right.\\ \left.\cdot\nabla_{\theta^{\mu}}\mu(s\mid\theta^{\mu})\right]\end{aligned} \tag{22}$$

To stabilize training and prevent premature convergence, TD3 incorporates:

- Delayed policy updates: the AN and TN are updated less frequently than the critics.

- Target policy smoothing: small noise $\mathcal{N}(0, \sigma^2)$ is added to the target action to regularize Q -values:

$$a' = \mu'(s') + clip\ (\mathcal{N}(0,\ \sigma^2), -c, c) \tag{23}$$

This strategy improves robustness against overfitting and variance from stochastic environments.

3.6 Integration of TD3 into the voltage control loop

The trained TD3 agent is implemented into a hierarchical voltage controller framework while preserving the system architecture from the AWDDQN model. Given the absolute number of decision steps t, the agent observes the complete system state s_t (refer to Equation 15), and returns a vector of continuous might control actions (where a_t is a vector of dimension 2 m – i.e., two actions for every m controllable device):

$$a_t = \mu(s_t \mid \theta^{\mu}) = [P_{1,t}, Q_{1,t}, \dots, P_{J,t}, Q_{J,t}] \tag{24}$$

These outputs are directly interpreted as active and reactive power setpoints for every controllable device, although these actions must be clipped to ensure bounds of Equations (9)-(11) are not violated. The control signals to the vector a_t are saved to either the ADN simulator or the advanced real-time controller, which calculates the next state s_{t+1} based upon the UCT. The start of the next cycle is completed by the agent observing the reward r_t (refer to Equation 18), and storing the experience (s_t, a_t, r_t, s_{t+1}) in the replay buffer, while updating the online network and TN per the TD3 rules - refer to Equations (20)-(23).

3.7 Offline training and online deployment

The TD3-learned voltage control system has two key phases:

1. Offline Training phase: The agent explores the environment using a Gaussian noise process over actions and learns the optimal voltage control policies, collecting them through experience replay until convergence of the average episode reward and decreasing the frequency of voltage violations.
2. Online Control phase: After training is complete, the TD3 agent can determine actions in real time without exploration noise. The AN outputs deterministic control actions that are determined only from observations of the system state. The outputs are continuous and bounded, so they can be used immediately with no extra decoding or quantization required for grid/backup-unstable devices.
3.

3.8 Real-time feasibility and constraint enforcement

To maintain operational safety, control outputs from the TD3 actor are bounded at runtime by device-specific constraints:

$$a_{j,t} \in [P_{j,t}^{min}, P_{j,t}^{max}]\ or\ [Q_{j,t}^{min}, Q_{j,t}^{max}] \tag{25}$$

These bounds reflect:

- Schedulable EV capacities as defined by Equations (10)-(14),
- Static VAR limits of reactive units.

These constraints are enforced by clipping the outputs during both action selection and training updates. Additionally, by using a design with penalty-based reward, agents learn to give voltage sensitive nodes priority adaptively.

3.9 Hierarchical EVA coordination

Charging for EVs is controlled in a two-level hierarchy:

- Upper Level: The TD3 agent figures out optimal total power for each EVA (aggregator) by taking into consideration the system state and schedulable availability.
- Lower Level: Each EVA takes the total power and distributes the power to the individual EVs based on SOC, arrival/departure windows, and ability to charge, with a weighted distribution.

$$P_{d,t} = \begin{cases} \dfrac{O_{j,t}}{SCP_{P,t}} \cdot P_{j,t}^{max}, & if\, O_{j,t} \geq 0 \\ \dfrac{O_{j,t}}{SDP_{d,t}} \cdot P_{j,t}^{min}, & if\, O_{j,t} < 0 \end{cases} \tag{26}$$

This design allows fulfillment of the initial scheduling guarantees while still ensuring feasibility in implementation and benefiting from TD3's fine-grained continuous control.

4 Simulation and result analysis

4.1 Test system and parameter settings

To assess how well the novel TD3-based voltage control method works, simulations are performed with two modified benchmark distribution networks, namely the IEEE-33 and IEEE-123 benchmark systems. Both systems have DERs, EVA, and reactive compensation units. The revised connection points with power rating settings are shown in Table I. Relative to the baseline condition identified for each system, the assumption of PV and WT generator connected at buses 10, 26 (IEEE-33) and buses 17, 63 (IEEE-123) with total capacity of 2 MW for PV and 1.2 MW for WT, while the EVA units are established at buses 21, 29 (IEEE-33) and buses 55, 102 (IEEE-123) with reactive units at buses 14 and 31 in both systems. The reactive power resources can operate within limits of [-0.6 Mvar, 0.6 Mvar].

Updated output profiles for the PV and WT units with ±15% random variations to mimic uncertainty are shown in Figure 1(a). The base load values are 4 MW and 6 MW for the IEEE-33 and IEEE-123 systems, respectively, along with ±10% load perturbations, as illustrated in Figure 1(b). The EVSC profiles are given in Figure 1(c), accounting for 120 EVs per aggregator, with 45kWh batteries, and charging/discharging power limited to 11kW. Charging efficiencies are kept at 0.97, and it is assumed that excited about charging EVs typically charge during the afternoon period at a workplace and continue charging again at their residential points during the night if not fulfilled.

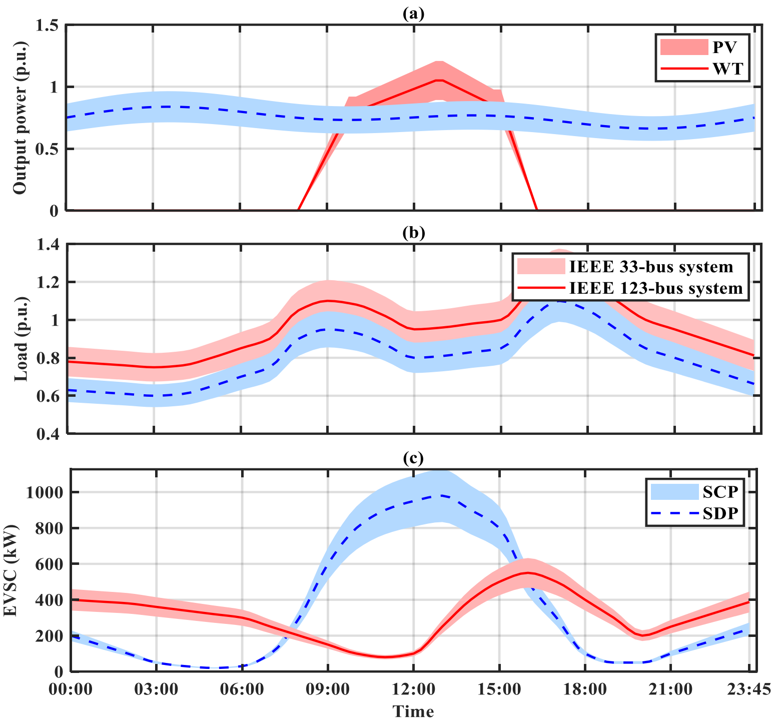

Figure 1: (a) Output profiles of PV and WT units. (b) load profiles with ±10% uncertainty. (c) EVSC schedules

Table 1: Settings of DERs and adjustable resources

Testing System	Connected Node	Resource Type	Power Setting
IEEE-33	10, 26	PV Generators	2.0 MW
	15	WT Units	1.2 MW
	21, 29	EV Aggregators (EVAs)	Shown in Figure 1(c)
	14, 31	Reactive Power Resources	[-0.6 Mvar, 0.6 Mvar]
IEEE-123	17, 63	PV Generators	2.0 MW
	30	WT Units	1.5 MW
	55, 102	EV Aggregators (EVAs)	Shown in Figure 1(c)
	45, 88	Reactive Power Resources	[-0.6 Mvar, 0.6 Mvar]

In the IEEE-33 system, the SS was defined as 33 node voltages plus 2 EVAs with 4 variables each (SCP, SDP, SCC, SDC), and 3 adjustable units, resulting in 44 input dimensions. The TD3 agent takes continuous ASs, which are bounded by the limits of the devices. However, a discretized mapping is kept for interpretability purposes to maintain consistent representations. Each actuator could be represented with 10 levels of granularity; in effect, 1000 equivalent action vectors can be visualized. With the same method, the IEEE-123 system is considered to consist of 123 voltage nodes, 4 EVAs, and 6 adjustable resources, which leads to a 151-dimensional state input to the neural networks.

4.2 Training process analysis

The TD3 agent was trained with synthetic data that was generated with Monte Carlo Sampling for 500 days (with 15% randomness applied to the DER and load values. Each episode was at a 15-minute granularity for a 24-hour time frame consisting of 96-time steps. The unseen data sets are important to validate generalization. The AC and twin CNs consist of {128, 128, 64} neurons for the IEEE-33 system and a deeper architecture of {256, 256, 128} neurons for the IEEE-123 system. Important TD3 hyperparameters for this training run were a policy update delay of 2 steps, target smoothing coefficient for τ = 0.005, discounting factor for γ = 0.98, and action noise standard deviation of σ = 0.1 (refer to Table II).

Table 2: TD3 training hyperparameters for the simulation environment

Parameter	IEEE-33 system	IEEE-123 system
Activation function	ReLU	ReLU
Discount factor γ	0.98	0.98
Learning rate η	0.0005	0.0005
Policy updates delayed	2	2

Replay buffer size M	50,000	100,000
Mini-batch size B	256	256
Target smoothing factor τ	0.005	0.005
Number of episodes I	2500	7000
Exploration noise std. σ	0.1	0.1
Hidden layers (Actor/Critic)	{128, 128, 64}	{256, 256, 128}

The training process demonstrates the superior ability of the TD3 algorithm, compared to the DDPG and DQN-based baselines, to converge more consistently throughout training. Observed episode rewards started to consistently rise and stabilize at approximately 2500 episodes in the case of the IEEE-33 system and 7000 episodes in the IEEE-123 setup. Figure 2 depicts the convergence process, and it demonstrates how the TD3 agent received the highest cumulative rewards across nearly all episodes. In the early parts of training, there was a lot of variation in the rewards received by the TD3 agent. This is because of the exploratory noise that stimulates exploration, and this variation in rewards decreases throughout training as policy variation decreases. The double critic mechanism provided an additional layer of stabilisation to the learning, and the delayed updates of the actor were other contributing factors that aided learning stability and with alleviating the ability to overestimate value.

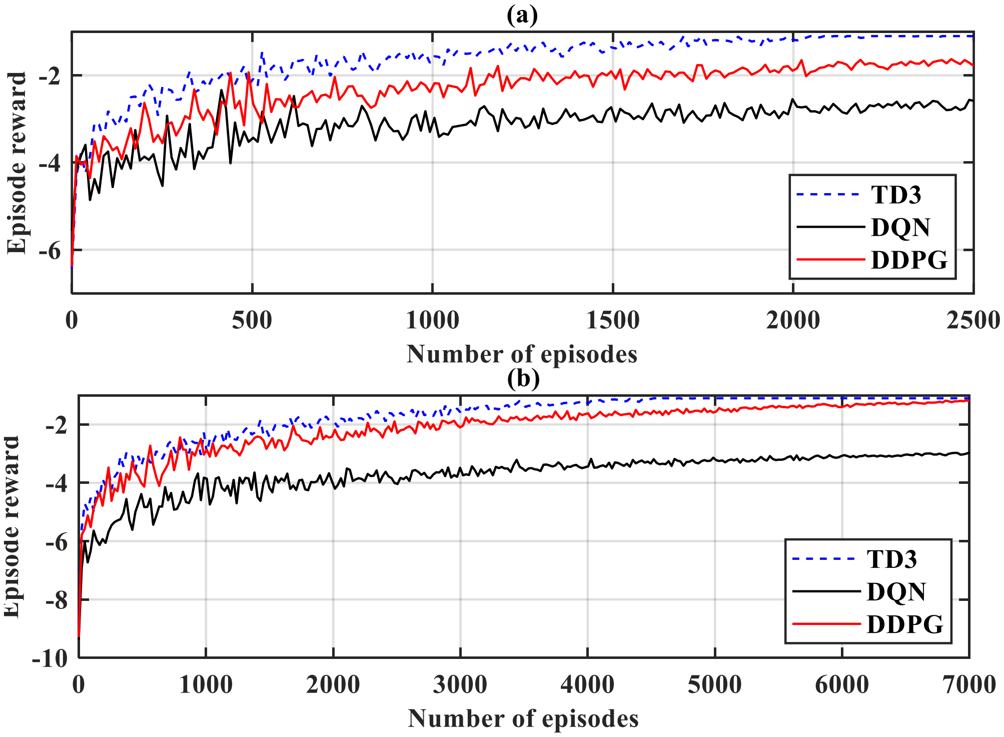

Figure 2: Episode reward progression for TD3, DDPG, and DQN-based agents. a) IEEE-33, b) IEEE-123

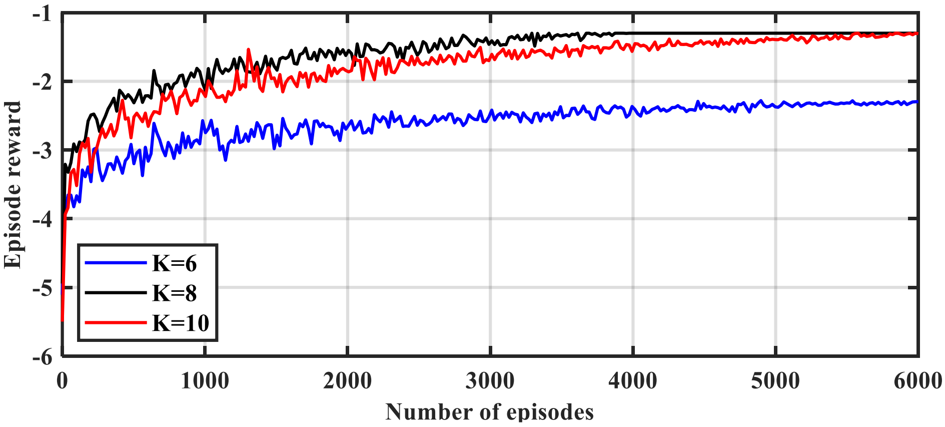

Figure 3: Training reward curves for TD3 under different action discretization resolutions

4.3 Effect of K value on training

While TD3 uses a continuous AS, the discretization granularity is evaluated to provide a comparative assessment and visualization. The parameter represented by K indicates the number of discrete control levels per actuator (or action resolution). For K = {6,8,10}, this translates to approximately action dimensions of 125, 1000, and 3375 for the IEEE-33 system. Figure 3 illustrates the training convergence for the different K's; smaller K values have higher reward convergence, suggested by quicker learning, which is a result of less action exploration complexity; however, smaller K also causes suboptimal final rewards as controls are too coarse. K = 8 allowed for convergence towards the end of the training period, where the time and the accuracy were balanced at the episode number of about 2500. Even though K = 10, when K increased, there was a better reward, but this was at the expense of learning speed and greater variance. K = 8 will be adopted for final deployment as this value represents the best tradeoff.

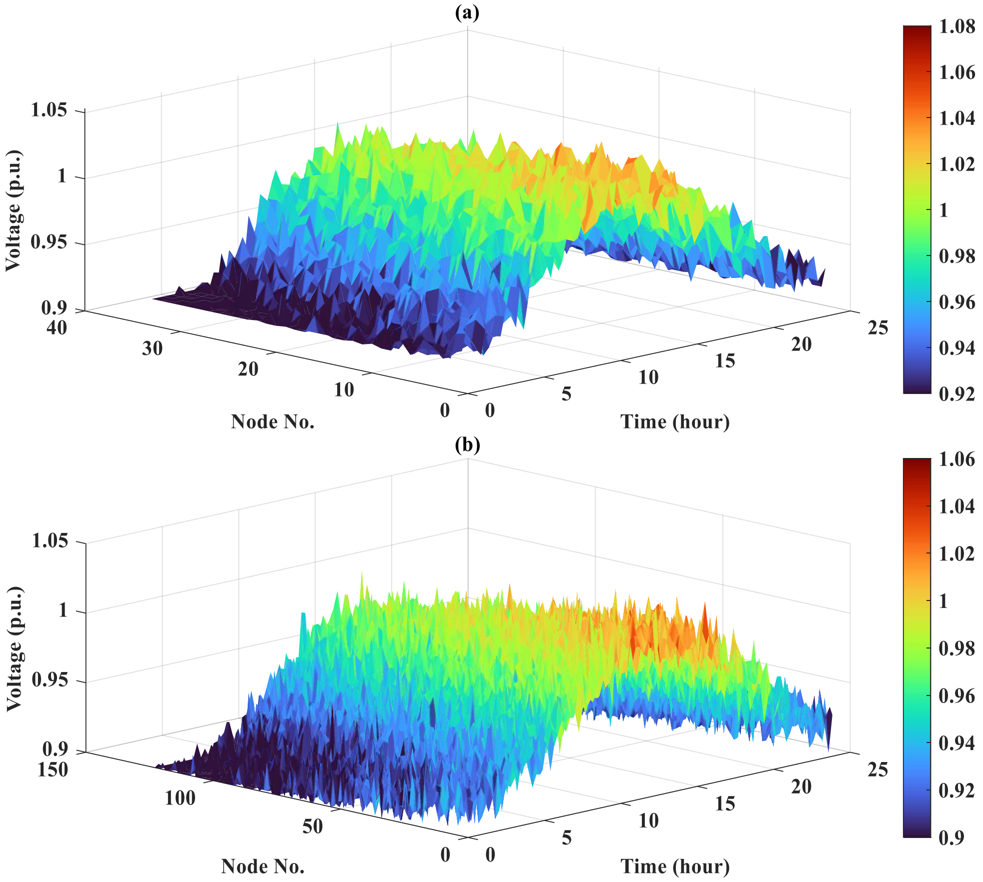

Figure 4: Uncontrolled voltage profiles across nodes over 24 hours in both test systems. a) IEEE-33, b) IEEE-123

4.4 Simulation results and discussions

Simulation results of the trained TD3 voltage control agent in control simulations show it is performing better in both test systems. Uncontrolled voltage profiles under the three test scenarios are provided in Figure 4. Clear signs of overvoltage during midday (with the peaks of PV) and undervoltage during the night (due to peak demand and lack of generation) are seen from the uncontrolled voltage profiles. For example, Node 26 in the IEEE-33 test system manages to exceed 1.07 p.u. at around midday. In the same test system, Node 33 demonstrates a loss of generation at considerably too low a voltage, with a measured voltage of around 0.93 p.u.

The controlled voltage results for traditional MINLP, DQN, and TD3 are presented in Figure 5. The TD3 agent accomplishes all nodes staying within [0.95, 1.05] p.u., in addition to achieving smoother voltage profiles. In the case of the IEEE-123 network, TD3 achieves a maximum deviation of ±0.025 p.u., which is outstanding, and a great improvement from DDQN.

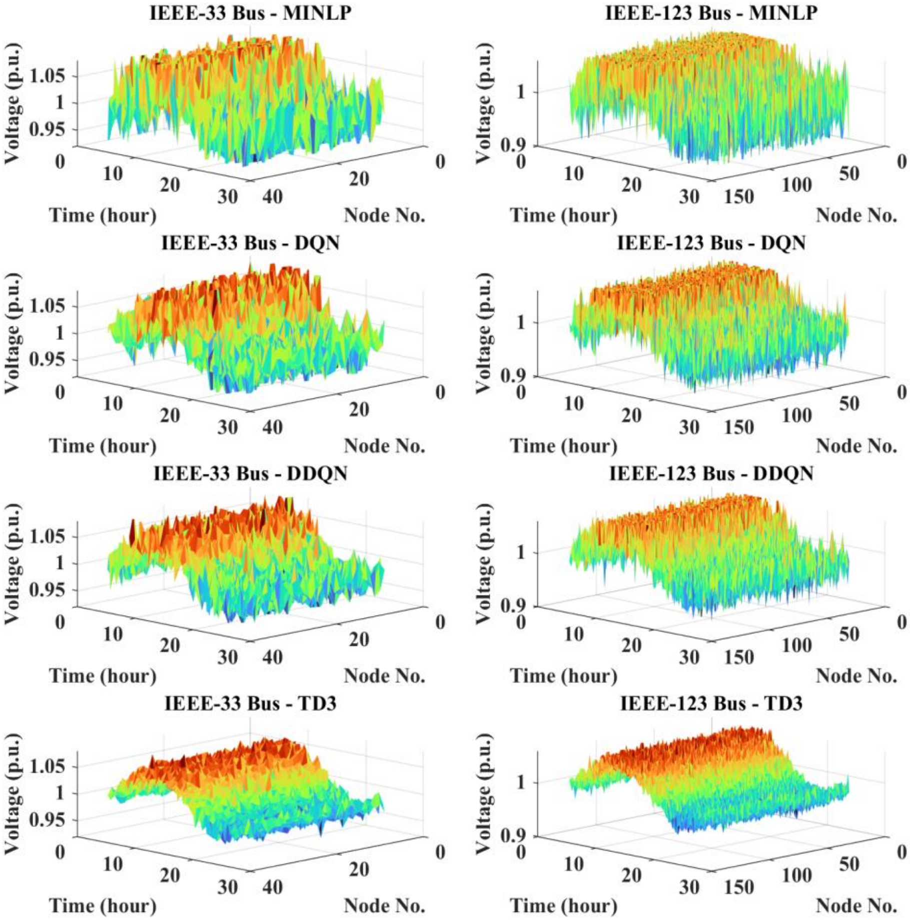

Figure 5: Controlled voltage profiles of IEEE-33 and IEEE-123 under MINLP, DQN, DDQN, and TD3 algorithms

In-depth actuator outputs (see Figure 6) depict that TD3 uses reactive power units at mid-day to achieve overvoltage mitigation effectively, while at night it employs appropriate control of EV discharging to facilitate recovery to a usable voltage. Table III summarizes the objective values, voltage ranges, and computation times across all control schemes. In both systems, TD3 yields the lowest voltage variance with only 0.1% of the computation times of MINLP, making it appropriate for real-time implementations.

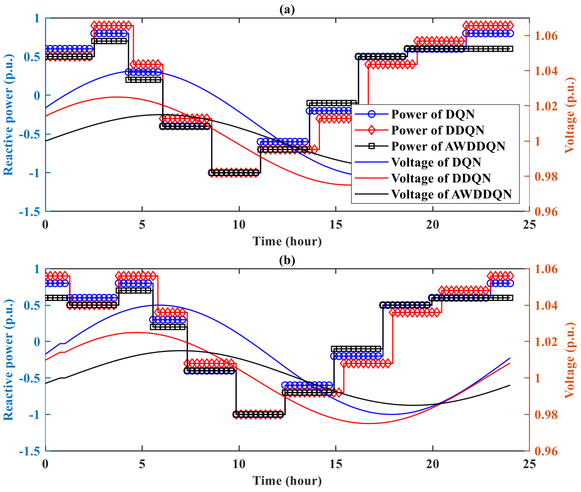

Figure 6: Real-time output actions of reactive units and voltage tracking under TD3 and baseline methods

Table 3: Voltage control performance comparison

Testing System	Case	Objective Function Value	Voltage Range (p.u.)	Calculation Time (s)
IEEE-33	Initial state	0.0256	[0.928, 1.071]	–
	MINLP	0.0089	[0.960, 1.050]	215.63
	DQN	0.0105	[0.947, 1.058]	0.18
	DDQN	0.0081	[0.950, 1.049]	0.24
	TD3 (Proposed)	0.0067	[0.957, 1.018]	0.26
IEEE-123	Initial state	0.0349	[0.920, 1.056]	–
	MINLP	0.0045	[0.964, 1.044]	1030.12
	DQN	0.0059	[0.945, 1.047]	0.60
	DDQN	0.0031	[0.973, 1.043]	0.94
	TD3 (Proposed)	0.0023	[0.976, 1.038]	0.99

The SOC curves for EVA 1 in Figure 7 demonstrate that the TD3 agent coordinated both system-level voltage requirements on the overarching transmission network and EV charging requirements (individually). In respect to the requirements of charging and discharging within the given time steps, nearly all EVs have met their SOC targets, indicating the successful coordinated control of the TD3 agent despite actively participating in grid level, active actions.

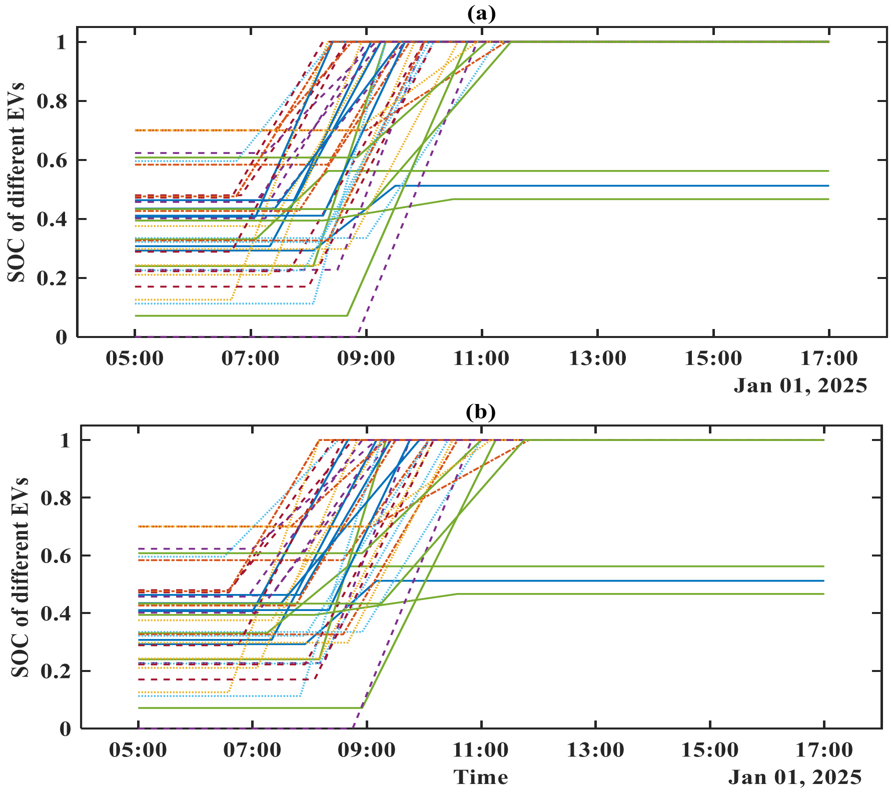

Figure 7: EV SOC trajectories post-dispatch for EVA 1, validating consumer satisfaction and grid compatibility

5 Conclusion

This research proposed a TD3-based smart voltage control framework for ADNs that included CA DRL coupled with system-aware constraints, like EV charging flexibility and renewable energy variability. By utilizing the TD3 algorithm, the proposed controller produces continuous output of power setpoints of adjustable resources, like EVAs and reactive power units. The controller is capable of smooth, fine-grained VR at nodes throughout the network since the voltage controller does not require discretized action sets. Simulations conducted on modified IEEE-33 and IEEE-123 test systems showed that TD3-based controllers were capable of maintaining node voltages within operational limits, despite variation in renewable generation and load. Convergence during training of the algorithm exhibited stability, and it was able to train at an accelerated manner, surpassing previously established RL model convergence rates. The algorithm was robust to random stochastic disturbances in the environment and resulted in a high level of control accuracy while representing a low computational effort. The TD3 algorithm also preserved a user-side EV constraint, leaving enough power capacity to meet EV charging demands while still supporting the distribution grid control needs. The method proposed fulfills the main research aim of establishing a scalable, stable, and high-resolution voltage control strategy that is implementable in modern distributions with high DERs and EVs. It also exhibited congruence with recent developments of CA RL and confirmed that deterministic actor–critic approaches are relevant for real-time grid applications. The implementation and performance of the method are contingent on proper parameter tuning and good simulation data; and while the simulation results offered improvements, operational deployment must involve realistically promoting constraints of sensors that are noisy, communication that is delayed, and hardware interfaces that cause bottlenecks. Future investigations should look at extending this framework with adaptive exploration, decentralized multi-agent coordination, and hybrid model-based components to enhance explainability and safety.

Nomenclature

Symbol	Definition	Symbol	Definition
s_t	State vector at time t	$P_{d,c,\max}$	Maximum charging power of EV d
a_t	Action vector at time t	$P_{d,d,\max}$	Maximum discharging power of EV d
r_t	Reward at time t	η_c	Charging efficiency of EV d
$U_{n,t}$	Voltage at node n at time t	η_d	Discharging efficiency of EV d
U_{base}	Base/reference voltage (1.0 p.u.)	$\text{SCC}_{d,t}$	Schedulable charging capacity of EV d at time t
	Time step duration	$\text{SCP}_{d,t}$	Schedulable charging power of EV d at time t
T	Dispatch time horizon	$\text{SDC}_{d,t}$	Schedulable discharging capacity of EVd at time t
N	Total number of nodes	$\text{SDP}_{d,t}$	Schedulable discharging power of EV d at time t
$Q_{j,t}$	Reactive power of resource j at time t	$\text{SCP}_{l,t}$	Total schedulable charging power of EVA l at time t
$P_{j,t}$	Active power of resource j at time t	$\text{SDP}_{l,t}$	Total schedulable discharging power of EVA l at time t
$Q_{j,\min}$	Minimum reactive power limit of resource j	$U_{i,t}$	Voltage at node i at time t
$Q_{j,\max}$	Maximum reactive power limit of resource j	λ_i	Penalty factor for node i
$P_{j,t,\min}$	Minimum active power limit of resource j at time t	γ	Discount factor in TD3
$P_{j,t,\max}$	Maximum active power limit of resource j at time t	$\mu(s \mid \theta^{\mu})$	An actor gives an action
C_d	Battery capacity of EV d	$Q_i(s,a \mid \theta^{Q_i})$	Critic estimates the value of that action.
$\text{SOC}_{d,t}$	State of charge of EV d at time t	y	Target Q-value
$\text{SOC}_{d,s}$	Initial SOC of EV d upon arrival	$L(\theta^{Q_i})$	Critic loss function
$\text{SOC}_{d,\min}$	Minimum required SOC of EV d upon departure	θ^{μ}	Parameters of the AN
$t_{d,s}$	Expected arrival time of EV d	θ^{Q_i}	Parameters of the CN
$t_{d,e}$	Expected departure time of EV d		

Acknowledgements

Interim Research Results of the Electrical Control System Design Practice Base under the 2024 Teaching Quality and Teaching Reform Project of Guangzhou College of Applied Science and Technology (Project No.: 2024ZG009)
Interim Research Results of the 2024 Higher Education Research Project under the 14th Five-Year Plan of Guangdong Higher Education Society (Project No.: 24GYB96)
Development of New Energy-Metal Air Battery Emergency Power Supply for Zhaoqing City, Guangdong Province, Science and Technology Innovation Guidance Category Project (Project No. 24122519009103)

Authorship contribution statement

Long Kou: Supervision, Conceptualization, Project administration, Writing-Original draft preparation.
Jiangyan Chen: Software, Methodology.
Yuhua Wang: Validation.
Yilin Liu: Language review.

Data availability

Available upon request.

Conflicts of interest

The authors affirm that they have no competing interests related to the posting of this document.

Author statement

All authors have reviewed and endorsed the manuscript, confirming adherence to the stated authorship criteria. Each author also attests to the integrity and authenticity of the work presented.

Ethical approval

Each author was directly and significantly involved in the work that led to this paper and will publicly stand behind its content.

References

[1] S. Nematshahi, D. Shi, F. Wang, B. Yan, and A. Nair, “Deep reinforcement learning based voltage control revisited,” *IET Generation, Transmission & Distribution*, 17(21): 4826–4835, 2023. https://doi.org/10.1049/gtd2.13001

[2] S. M. Abdelkader *et al.*, “Advancements in data-driven voltage control in active distribution networks: A Comprehensive review,” *Results in Engineering*, 23: 102741, 2024. https://doi.org/10.1016/j.rineng.2024.102741

[3] Y. Zheng, Y. Song, D. J. Hill, and K. Meng, “Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems,” *IEEE Trans Industr Inform*, 15(2): 638–649, 2018. DOI:10.1109/TII.2018.2812755

[4] M. Mazumder and S. Debbarma, “EV charging stations with a provision of V2G and voltage support in a distribution network,” *IEEE Syst J*, 15(1): 662–671, 2020. https://doi.org/10.1016/j.ecmx.2025.101138

[5] A. Ahmadian, B. Mohammadi-Ivatloo, and A. Elkamel, “A review on plug-in electric vehicles: Introduction, current status, and load modeling techniques,” *Journal of Modern Power Systems and Clean Energy*, 8(3): 412–425, 2020. DOI:10.35833/MPCE.2018.000802

[6] H. Patil and V. N. Kalkhambkar, “Grid integration of electric vehicles for economic benefits: A review,” *Journal of Modern Power Systems and Clean Energy*, 9(1): 13–26, 2020. DOI:10.35833/MPCE.2019.000326

[7] X. Sun and J. Qiu, “Hierarchical voltage control strategy in distribution networks considering customized charging navigation of electric vehicles,” *IEEE Trans Smart Grid*, 12(6): 4752–4764, 2021. DOI:10.1109/TSG.2021.3094891

[8] Y. Wang, T. John, and B. Xiong, “A two-level coordinated voltage control scheme of electric vehicle chargers in low-voltage distribution networks,” *Electric Power Systems Research*, 168: 218–227, 2019. https://doi.org/10.1016/j.epsr.2018.12.005

[9] Y. Liu and H. Liang, “A discounted stochastic multiplayer game approach for vehicle-to-grid voltage regulation,” *IEEE Trans Veh Technol*, 68(10): 9647–9659, 2019. https://doi.org/10.1016/j.rineng.2025.106813

[10] J. Hu, C. Ye, Y. Ding, J. Tang, and S. Liu, “A distributed MPC to exploit reactive power V2G for real-time voltage regulation in distribution networks,” *IEEE Trans Smart Grid*, 13(1): 576–588, 2021. DOI: 10.1109/TSG.2021.3109453

[11] D. Haoran, Y. Ming, C. Fang, and S. Guozhong, “Reactive power and voltage optimization control approach of the regional power grid based on reinforcement learning theory,” *Transactions of China Electrotechnical Society*, 30(12): 408–414, 2015. https://doi.org/10.3390/en17246454

[12] D. Cao *et al.*, “Reinforcement learning and its applications in modern power and energy systems: A review,” *Journal of modern power systems and clean energy*, 8(6): 1029–1042, 2020. DOI:10.35833/MPCE.2020.000552

[13] V. Mnih *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, 518(7540): 529–533, 2015. DOI:10.1038/nature14236

[14] J. Shi, W. Zhou, N. Zhang, Q. Chen, J. Liu, and Z. Cao, “Deep reinforcement learning algorithm of voltage regulation in distribution network with energy storage system,” *Electric Power Construction*, 41(03): 71–78, 2020. https://doi.org/10.1016/j.apenergy.2022.120510

[15] R. Diao, Z. Wang, D. Shi, Q. Chang, J. Duan, and X. Zhang, “Autonomous voltage control for grid operation using deep reinforcement learning,” in *2019 IEEE power & energy society general meeting (PESGM)*, IEEE, 2019, 1–5. DOI:10.48550/arXiv.1904.10597

[16] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, “Two-timescale voltage control in distribution grids using deep reinforcement learning,” *IEEE Trans Smart Grid*, 11(3): 2313–2323, 2019. DOI:10.1109/TSG.2019.2951769

[17] X. Sun and J. Qiu, “A customized voltage control strategy for electric vehicles in distribution networks with reinforcement learning method,” *IEEE Trans Industr Inform*, 17(10): 6852–6863, 2021. DOI: 10.1109/TII.2021.3050039

[18] B. Zhang, D. Cao, W. Hu, A. M. Y. M. Ghias, and Z. Chen, “Physics-Informed Multi-Agent deep reinforcement learning enabled distributed voltage control for active distribution network using PV inverters,” *International Journal of Electrical Power & Energy Systems*, 155: 109641, 2024. https://doi.org/10.1016/j.ijepes.2023.109641

[19] M. Golgol and A. Pal, “High-speed voltage control in active distribution systems with smart inverter coordination and DRL,” in *2024 IEEE Power & Energy Society General Meeting*

(PESGM), IEEE, 2024, 1–5. https://doi.org/10.1016/j.epsr.2024.110528

[20] N. Shi, R. Cheng, L. Liu, Z. Wang, Q. Zhang, and M. J. Reno, "Data-driven affinely adjustable robust Volt/VAr control," *IEEE Trans Smart Grid*, 15(1): 247–259, 2023. DOI: 10.1109/TSG.2023.3270112

[21] A. Dutta, S. Ganguly, and C. Kumar, "Three-stage receding horizon-based voltage control and electric vehicle charge scheduling of active distribution networks," *IET Renewable Power Generation*, 18, 4308–4317, 2024. https://doi.org/10.1049/rpg2.12977

[22] C. Ma, W. Xiong, Z. Tang, Z. Li, Y. Xiong, and Q. Wang, "A Hierarchical Voltage Control Strategy for Distribution Networks Using Distributed Energy Storage," *Electronics (Basel)*, 14(9): 1888, 2025. https://doi.org/10.3390/electronics14091888

[23] Y. Zhang, J. Chen, H. Zhao, W. Zhang, W. Jiao, and W. Dai, "Coordinated voltage control of active distribution networks with photovoltaic and power to hydrogen," *IET Energy Systems Integration*, 5(3): 245–260, 2023. https://doi.org/10.1049/esi2.12096