Energy Engineering

Tech Science Press

https://doi.org/10.32604/ee.2026.076751

ARTICLE

Data-Driven Voltage Control for Distribution Networks with High Penetration of PVs via Improved DDPG

Xiaolong Xiao[1,*], Linghao Zhang[2], Ziran Guo[1], Xiaoxing Lu[1], Wenqiang Xie[1], Shukang Lv[1] and Peishuai Li[3]

[1]Electric Power Research Institute, State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China
[2]State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China
[3]School of Automation, Nanjing University of Science & Technology, Nanjing, China

*Corresponding Author: Xiaolong Xiao. Email: ethan518@126.com

Received: 26 November 2025; Accepted: 12 February 2026; Published: 12 November 2026

ABSTRACT: The highly increasing integration of distributed photovoltaics (PVs) has brought significant challenges for distribution network voltage control, which struggles with fast and stochastic fluctuations. This paper proposes a data-driven voltage control method utilizing an improved deep deterministic policy gradient (DDPG) to minimize power losses while maintaining voltage within a safe operating range. The data-driven voltage control framework is established first, with the controlling center formulated as the agent. Based on the distribution network system and the voltage control original model, the data-driven voltage control model is formulated via the Markov decision process (MDP). A voltage sensitivity-guided DDPG algorithm, which incorporates the voltage-reactive power injection sensitivity matrix into the actor policy gradient, is developed to embed physical grid knowledge and reduce the actor's reliance on critic action-gradient estimates, thus the practicality and generalization of data-driven voltage control is improved. Numerical simulations on the modified IEEE 33-node test system show the method maintains all nodal voltages within limits and reduces total system losses under varying PV generation, demonstrating its effectiveness and potential for future sustainable power systems.

KEYWORDS: Distributed photovoltaics (PVs); distribution network; voltage control; deep reinforcement learning (DRL); improved DDPG

1 Introduction

The next type of power system with high penetration of renewable energy sources has become an important trend during the global energy transition [1]. As for the distribution network, the penetration of distributed photovoltaics (PVs) has increasing rapidly [2]. While PV integration promotes cleaner energy, it also transforms distribution networks from passive systems with largely unidirectional power flow into active networks with bidirectional flows and stronger coupling between power injections and voltage profiles.

However, the intermittency and stochastic output of PVs, together with the electromechanical differences between grid-connected inverters and traditional rotating machines, pose new challenges to secure operation for the distribution network [3]. In particular, voltage violations and fast voltage fluctuations become more frequent while rapid variations in load and generation keep voltages continuously time-varying. Conventional passive voltage regulation schemes based on fixed capacitor banks and fixed-ratio transformers are often too slow or insufficiently adaptive for these dynamics, and may even lead to

Copyright © 2027 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

undesirable interactions under high PV penetration [4]. Consequently, real-time coordinated voltage control that is scalable and responsive has become a critical requirement for modern distribution networks.

The existing approaches to distribution network voltage control are predominantly categorized as model-based and data-driven methods. The model-based methods, which rely on explicit grid models and employ control theory or optimization to compute setpoints, can be further divided into two branches. The first encompasses local rule-based controllers. Recent enhancements to these controllers incorporate projection or gradient-based updates to improve responsiveness and reduce communication overhead [5], with further refinements using memory structures and historical data for better long-term performance [6,7]. The second branch involves optimization-based coordination methods, while the typical method is the optimal power flow (OPF)-based algorithms for globally coordinated voltage regulation [8–11]. While these model-based approaches benefit from strong theoretical foundations and interpretability, their effectiveness is inherently constrained by dependency on accurate system models and often high computational complexity.

Despite their effectiveness and interpretability, model-based methods face well-known limitations in real-time voltage regulation. Local droop-like schemes respond quickly but are inherently myopic because they rely on local measurements and limited system-wide information, which makes it difficult to achieve globally optimal coordinated performance under strong network coupling [12]. In contrast, optimization-based coordination can yield system-level optimal setpoints under accurate models, yet it often requires reliable parameter information and repeated solution of nonconvex optimization problems under fast-varying operating conditions. Traditional OPF and related static optimization formulations may therefore struggle to react instantaneously to rapid PV fluctuations, falling short of modern real-time control requirements [13]. These issues motivate increasing interest in data-driven closed-loop control strategies that can adapt online with reduced reliance on perfect system models.

Motivated by these limitations and enabled by the proliferation of distribution-level measurements, data-driven control strategies, which learn control policies directly from data, have gained significant traction. Reinforcement learning (RL) provides a natural framework for sequential decision making under uncertainty via a Markov decision process formulation [14]. The distribution networks with high PV penetration are characterized by high-dimensional state spaces and complex nonlinear dynamics, and vanilla RL may suffer from the "curse of dimensionality," leading to slow learning and limited scalability [15]. Deep learning (DL), with its strong nonlinear approximation and representation capability, can extract compact features from high-dimensional measurements and has been widely adopted for prediction and regression tasks [16]. The combination of DL and RL has led to deep reinforcement learning (DRL), initiated by the deep Q-Network (DQN) framework [17] and extended to a broad class of end-to-end perception-and-control algorithms [18]. DRL has since been explored in power systems for security and stability control [19], preventive control strategy generation [20], voltage optimization control [21–23], and energy management and optimal dispatch [24–27].

Beyond the control-centric focus of DRL, other AI paradigms have been extensively applied to power and energy systems, particularly for problems involving strategic interaction. For instance, game-theoretic formulations have been widely adopted to model strategic interactions among aggregators, prosumers, and electric-vehicle fleets in demand response and energy scheduling problems [28], and peer-to-peer energy trading has also been studied through game-theoretic mechanisms [29]. Such games are often solved by distributed optimization or meta-heuristic/swarm-intelligence optimizers, and recent studies have started to explore quantum-inspired swarm intelligence to cope with nonconvex and high-dimensional equilibrium computation [30]. In addition, survey works provide a systematic overview of game-theoretic approaches in modern electricity-market demand response and energy trading [31]. These approaches are particularly

effective for market/scheduling problems with competing objectives, whereas real-time distribution voltage regulation additionally requires fast closed-loop feedback and strict operational constraints.

While DRL-based voltage control has demonstrated promise, existing approaches—particularly multi-agent DRL (MADRL) variants for multi-area or multi-device coordination—often treat the network as a black box and do not fully exploit physical coupling information embedded in the power-flow structure. In addition, actor–critic training in continuous control can be sensitive to value-function approximation and noisy critic action gradients, which may degrade training stability and result in suboptimal control performance, especially under changing operating conditions. Moreover, learned policies may generalize poorly when the operating point, PV patterns, or network configuration shifts, which is a major concern for deployment-oriented distribution control applications. The characteristics of different methods are summarized in Table 1.

Table 1: Comparison of voltage control methods.

Method	Reference	Global Optimality	Dynamic Response	Physical Network Structure	Reliability
Model-based Local Control	[4] [12]	No	Fast	Considered	High
Model-based Optimization Control	[8] [11]	Yes	Low	Considered	High
Optimized Droop Control	[5] [7]	Moderate	Fast	Considered	High
Tradition DRL-based Control	[28] [29] [30] [31]	Yes	Fast	Not considered	Moderate
Proposed Method	This Paper	Yes	Fast	Considered	High

To address the above challenges, this paper proposes a data-driven voltage control method with sensitivity-guided DDPG to improve the training efficiency with the utilization of physical network information. The data-driven voltage framework based on DRL is established for a distribution network with high penetration of PVs. According to the offline training-online execution structure, the global optimum and dynamic response can be considered simultaneously. A sensitivity-guided DDPG, which integrates the voltage sensitivity matrix into the actor gradient update, is proposed. As the proposed policy gradient is guided by physics-informed sensitivities, that it reduces the heavy relies on critic-provided action-gradient approximations of DDPG, thus to improve the training convergency and efficiency.

The reminder of the paper is organized as follows. Section 2 illustrates the data-driven voltage control framework. In Section 3, the data-driven distribution network voltage control model is established based on a Markov decision process (MDP). Section 4 shows the voltage sensitivity-guided DDPG algorithm and the solution process. In Section 5, the case studies are analyzed, and a conclusion is drawn in Section 6.

2 Data-Driven Voltage Control Framework for Distribution Network

In this paper, a data-driven voltage control framework based on DRL is established to address real-time voltage control in distribution networks with high penetration of PVs, as shown in Fig. 1. The control center is formulated as the agent that determines the PV reactive power generation strategy according to the real-time system operation status fluctuation. This framework is essentially a closed-loop system where the agent continuously interacts with the distribution network simulation environment. At each control time step, the agent relies on the deep neural network to generate the reactive power control strategy and acts on the environment, then the environment simulates the corresponding operations status by performing power flow calculations and feeds the controlling efficiency back to the agent. With continuous interaction between the agent and the environment, the deep neural network can be trained well, and the agent provides the voltage control strategy in real-time. Theoretically, the well-trained deep neural network can react in milliseconds [32]; the detailed action time scale is determined based on the practical data collection, information communication, and hardware, such as PV inverters' action time. Thus, the proposed data-driven voltage control method is in real-time, while the action time scale is determined based on the actual distribution network real-time schedule.

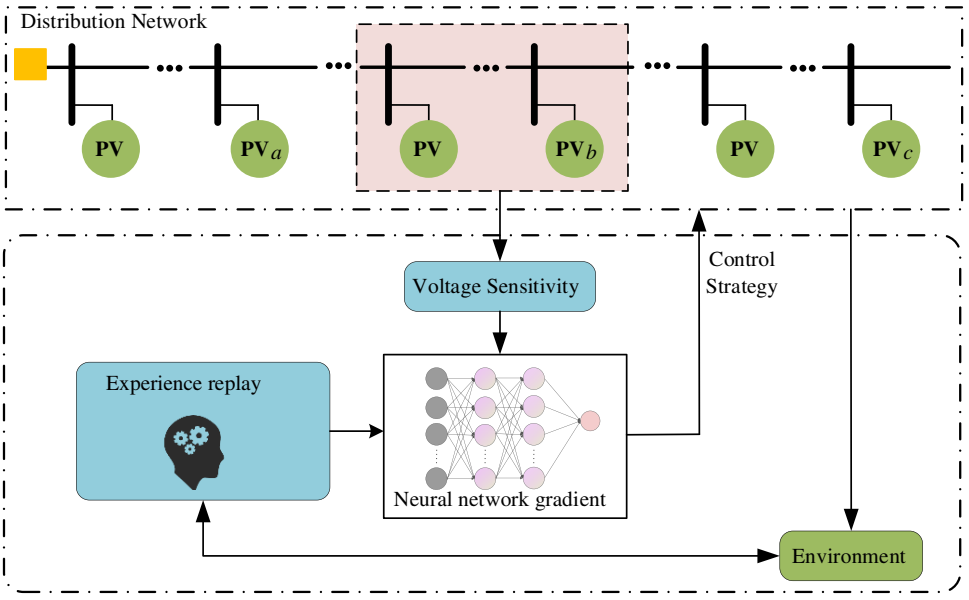


Figure 1: Data-driven voltage control framework for distribution network.

The Markov decision process is utilized to illustrate the data-driven voltage control procedure. With the aim of improving the control efficiency in the actual scenario and enhancing the DRL's generalization, the sensitivity-guided DDPG, which integrates distribution network physical knowledge into the learning process, is proposed within this framework. With this method, the original critic action-gradient estimate is replaced by the gradient derived from the sensitivity between node voltage and reactive power injection, which is more applicable for practical distribution network operation. As for the effective training of the deep neural network, more practically, a comprehensive reward function that considers power losses, voltage quality, and PV inverter operation is established, thus the reactive power output limit violations are strictly restricted. By integrating the physical laws of the distribution network into DRL, the proposed method aims to improve the training stability and convergence speed, thus improving the voltage control economic efficiency response to the real-time distribution network operation status.

3 Data-Driven Distribution Network Voltage Control Model

In the distribution network with high penetration of PVs, the operation status is complex and variable. The traditional central and local control method cannot possess the global coordination of different PVs' reactive power and the dynamic fast response ability. Then, the data-driven voltage control method is proposed by formulating the voltage control problem as an MDP. The voltage control model is established to reduce system power losses and improve voltage quality while strictly adhering to distribution network power flow balance, equipment operating limits, and system safety conditions. The data-driven model is constructed based on the system and traditional voltage control mathematical model shown below.

3.1 Traditional Voltage Control Model for Distribution Network

In distribution network operation, the power loss is an important indicator of system economy. With reactive power compensation, the line current can be reduced, thereby decreasing power losses and improving system operational efficiency. Moreover, the voltage quality is also the key indicator of system operation. With consideration of both these two objectives, the corresponding function is established as follows.

(1) Objective Function

The objective function is expressed as:

$$\min \sum_{t\in T}\left(\Phi_t^{loss}+\boldsymbol{\lambda}_1\Phi_t^{voltage}\right) \tag{1}$$

$$\min f_1^{'}=\min \sum_{t\in \boldsymbol{T}}\Phi_t^{loss}=\min \sum_{t\in \boldsymbol{T}}\sum_{ij\in \mathrm{E}} G_{ij}\left[V_{i,t}^2+V_{j,t}^2-2V_{i,t}V_{j,t}\cos\theta_{ij,t}\right] \tag{2}$$

$$\min f_2^{'}=\min \sum_{t\in \boldsymbol{T}}\Phi_t^{voltage}=\min \sum_{t\in \boldsymbol{T}}\sum_{i\in N}\left|V_{i,t}-V^{ref}\right| \tag{3}$$

where Φ_t^{loss} and $\Phi_t^{voltage}$ quantify the system power loss and voltage quality performance, respectively. The parameter $\boldsymbol{\lambda}_1$ is introduced to balance the contribution of voltage quality in the objective function. The index t indicates the scheduling time step, while $\boldsymbol{T}$ represents the overall dispatch horizon. Sets E and N denote the collections of network branches and nodes, respectively. The notation ij ($ij \in E$) refers to the line connecting node i to node j, $V_{i,t}$ denotes the voltage magnitude at node i at time t. G_{ij} and B_{ij} corresponds to the conductance and susceptance of branch ij, $\theta_{ij,t}$ represents the voltage phase-angle difference between nodes i and node j at time t. V^{ref} denotes the reference voltage value.

Considering that the two sub-objective functions are described in different units and are not directly comparable, the objective functions $f_1^{'}$ and $f_2^{'}$ are normalized using the min–max normalization method. The normalization process is formulated as follows:

$$f_1=\frac{f_1^{'}-f_1^{\min}}{f_1^{\max}-f_1^{\min}} \tag{4}$$

$$f_2=\frac{f_2^{'}-f_2^{\min}}{f_2^{\max}-f_2^{\min}} \tag{5}$$

where f_1 denotes the normalized sub-objective function of power losses, and f_2 represents the normalized sub-objective function of voltage quality. $f_1^{\max}$ and $f_1^{\min}$ are the maximum and minimum values of the active

power loss in the distribution network, respectively, while ${f_2}^{\max}$ and ${f_2}^{\min}$ correspond to the maximum and minimum values of the voltage deviation.

(2) Power Flow Constraints

The constraints for the distribution network operation include the power flow constraints, network safe operation constraints and PV operations constraints. As in the voltage control model construction, the physical laws whose core is the power flow are strictly satisfied. The power flow constraints can be illustrated by the nonlinear equations between node power injection and node voltage with the line parameters, serving as the theoretical foundation for voltage control decision-making.

$$P_{i,t}^{PV} - P_{i,t}^{PD} = V_{i,t} \sum_{j \in N} V_{j,t} \left(G_{ij} \cos \theta_{ij,t} + B_{ij} \sin \theta_{ij,t} \right) \tag{6}$$

$$Q_{j,t}^{PV} - Q_{j,t}^{QD} = V_{i,t} \sum_{j \in N} V_{j,t} \left(G_{ij} \sin \theta_{ij,t} - B_{ij} \cos \theta_{ij,t} \right) \tag{7}$$

where $P_{i,t}^{PV}$ and $P_{i,t}^{PD}$ denote the active power output of the photovoltaic (PV) unit and the active power demand at node i at time t, respectively, while $Q_{j,t}^{PV}$ and $Q_{j,t}^{QD}$ represent the corresponding reactive power output of the PV unit and the reactive power demand at node i at time t.

(3) PV Operation Constraints

To fully utilize the reactive power regulation capability of distributed PV, only the inverter capacity limitation is considered to determine the reactive power compensation range. Specifically, the PV inverter regulation model is expressed as follows:

$$-Q_{i,t}^{PV,\max} \le Q_{i,t}^{PV} \le Q_{i,t}^{PV,\max} \tag{8}$$

$$Q_{i,t}^{PV,\max} = \sqrt{\left(S_i^{PV}\right)^2 - \left(P_{i,t}^{PV}\right)^2} \tag{9}$$

where $Q_{i,t}^{PV,\max}$ is the maximum adjustable reactive power of the PV connected to node i at time t; S_i^{PV} is the apparent power capacity of the PV connected to node i.

(4) Safety Operation Constraints for Distribution Network

The safety operation constraints for the distribution network include node voltage magnitude constraints and branch current constraints, and the corresponding constraints are established as follows.

$$V_i^{\min} \le V_{i,t} \le V_i^{\max} \tag{10}$$

where $V_i^{\min}$ and $V_i^{\max}$ are the upper and lower limits of voltage magnitude at node i, respectively.

The traditional voltage control mathematical model for the distribution network is constructed as above. The mathematical model comprehensively captures the core elements and physical laws of real-time voltage control in distribution networks with high PV penetration. It provides an interactive and decision-making environment for the DRL agent and also infuses the learning process with clear physical guidance and safety boundaries.

3.2 DRL-Based Data-Driven Voltage Control Formulation

With the established data-driven voltage control framework shown in Fig. 1, the control center is formulated as the agent while the environment is the distribution network. The MDP is utilized to transfer the complex voltage control problem into a structured sequential decision-making task. As in the MDP

formulation, it includes the state space, action space, and the reward function. Based on the local observation obtained from the state space, the agent selects an action according to its policy. Then, the distribution network operation status is updated through a state transition function, which defines the probability distribution of the next state. Subsequently, the agent receives an immediate reward and a new state, with the objective of maximizing its own expected cumulative return over a period of time. In the MDP formulation, state space, action space, and reward function are defined as follows:

(1) *state space:* the state space is designed to capture the real-time operating conditions that drive voltage fluctuations. Specifically, at time step t, the agent observes a state vector $\boldsymbol{S}_t$, which can be expressed as $\boldsymbol{S}_t = \left[\boldsymbol{V}_t, \boldsymbol{P}_t^{PV}, \boldsymbol{P}_t^{PD}, \boldsymbol{Q}_t^{PV}, \boldsymbol{Q}_t^{PD}\right]$, where $\boldsymbol{V}_t$ denotes the voltage magnitude vector, $\boldsymbol{P}_t^{PV}$ and $\boldsymbol{Q}_t^{PV}$ denotes the active and reactive power output vectors of PVs, $\boldsymbol{P}_t^{PD}$ and $\boldsymbol{Q}_t^{PD}$ denote the nodal active and reactive loads vectors. This state representation provides sufficient information for the agent to infer the current voltage security margin and the potential impact of reactive power control.

(2) *action space:* The action space corresponds to real-time reactive power regulation of PV inverters. The action at time t is defined as a continuous vector $\boldsymbol{A}_t$, representing the incremental adjustment of PV reactive power outputs: $\boldsymbol{A}_t = \boldsymbol{\Delta Q_t^{PV}} = \left[\Delta Q_{1,t}^{PV}, \ldots, \Delta Q_{M,t}^{PV}\right]$, where M is the number of controllable PV inverters. The environment updates the inverter reactive power by $Q_{i,t}^{PV} = Q_{i,t}^{PV} + \Delta Q_{i,t}^{PV}$ and enforces inverter capability limits according to the PV operation constraints. Based on the agent's output, the action a_t of the simulation environment can be expressed as:

$$a_t = \Delta Q_t^{PV}\left(s_t \middle| \left(W_i, b_i\right)\right) + \varepsilon_i\left(0, \sigma_t\right) \tag{11}$$

where W_i and b_i represents the weights and biases between the connected layers in the neural network; $\varepsilon_i\left(0, \sigma_t\right)$ represents the noise disturbance experienced by the agent during exploratory action; the noise follows a normal distribution with a mean of 0 and a standard deviation of σ_t. The parameter σ_t decays at a determined rate during the training process.

(3) *reward function:* The reward function is constructed to jointly promote economic operation and secure voltage regulation. After applying action and performing power-flow calculations, the environment returns an immediate reward r^t defined as the negative of the control objective shown as below.

$$R\left(s_t, a_t\right) = r^t = -\sum_{t \in T}\left(\boldsymbol{\Phi}_t^{loss} + \boldsymbol{\lambda}_1 \boldsymbol{\Phi}_t^{voltage} + \boldsymbol{\lambda}_2 \boldsymbol{\Phi}_t^{viol}\right) \tag{12}$$

$$\boldsymbol{\Phi}_t^{viol} = \sum_i \max\left\{\left(V_{i,t} - V_i^{\max}\right), 0, \left(V_i^{\min} - V_{i,t}\right)\right\} \tag{13}$$

where $\boldsymbol{\Phi}_t^{viol}$ is the normalized voltage-limit violations, $\boldsymbol{\lambda}_2$ shows the coefficient of the voltage-limit violations.

4 Voltage Sensitivity-Guided DDPG Algorithm

For real-time voltage control with PV inverters, the control action is continuous and must be generated within a very short time step. A widely used baseline is the DDPG under the actor–critic framework. However, for distribution networks with high-dimensional and strongly nonlinear power-flow characteristics, the critic may suffer from large approximation errors and gradient variance, which can destabilize training and slow down convergence. To address this issue, this paper develops a voltage sensitivity-guided DDPG algorithm, which embeds physical grid knowledge into the policy-gradient computation and reduces the actor's reliance on critic action-gradient estimates. To be specific, the method adopted in this paper uses

the voltage sensitivity matrix S to analytically derive the parameter update gradient for the agent, which can avoid the approximation errors and learning variance generated by the critic network.

The key idea is to replace the critic-based action-gradient with an analytical gradient derived from the voltage–reactive power sensitivity matrix. Specifically, at each operating point, the AC power flow equations are linearized around the current state. The incremental power mismatch and state increments satisfy:

$$\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} = \begin{bmatrix} J_{P\theta} & J_{PV} \\ J_{Q\theta} & J_{QV} \end{bmatrix} \cdot \begin{bmatrix} \Delta\theta \\ \Delta V \end{bmatrix} \tag{14}$$

where $J_{P\theta}, J_{PV}, J_{Q\theta}, J_{QV}$ are the Jacobian submatrices evaluated at the current operating point, and Δ denotes the increment between two adjacent states. Considering that the control variable is the PV reactive power injection, the voltage sensitivity matrix is defined as:

$$Sen = \frac{\partial V}{\partial Q} = \left(J_{QV} - J_{Q\theta} {J_{P\theta}}^{-1} J_{PV}\right)^{-1} \tag{15}$$

With this sensitivity, the gradient of any voltage-dependent term with respect to the action $\boldsymbol{a}$ can be computed analytically. The physics-informed action gradient is:

$$\nabla_a L_{\text{phys}}(s, a; v) = \nabla_a c(s, a) + v^{\top} \left[\left(\frac{\partial h}{\partial \mathbf{V}} \right)^{\top} Sen \right] \tag{16}$$

where $\nabla_a L_{\text{phys}}$ is a deterministic, low-variance alternative to the critic's estimate $\nabla_a Q(s, a)$. Therefore, the final update for the actor parameters synthesizes the data-driven gradient from the critic and the model-based gradient from the physics:

$$\nabla_\theta J \approx \frac{1}{N} \sum_{n=1}^{N} \nabla_\theta \mu_\theta \left(s^{(n)}\right) \cdot \left[\alpha \cdot \nabla_a Q_w \left(s^{(n)}, a\right) + \beta \cdot \nabla_a L_{\text{phys}} \left(s^{(n)}, a; v\right)\right]|_{a=\mu_\theta\left(s^{(n)}\right)} \tag{17}$$

where α and β are blending coefficients. This hybrid approach stabilizes training by attenuating the high variance from the critic, especially during initial learning, while ensuring control actions are inherently guided by and respect the underlying grid physics and operational limits.

Furthermore, considering that power flow calculations require significant computational resources, the voltage sensitivity matrix is also used to estimate the voltage magnitude $\boldsymbol{V}_{t+1}$ of the next state $\boldsymbol{S}_{t+1}$. For example, the formula for estimating the voltage at node i in state $\boldsymbol{S}_{t+1}$ based on the voltage sensitivity matrix is as follows:

$$V_{i,t+1} = V_{i,t} + \sum_{j=1}^{N_{PV}} Sen_{ij,t} \cdot a_{j,t} \tag{18}$$

where $Sen_{ij,t}$ is the element in the row corresponding to node i and the column corresponding to the j-th PV node in the sensitivity matrix at time t.

During the training process of the agent, the gradient update method for the neural network in a single iteration is expressed as follows:

$$\begin{cases} \nabla_\theta = \frac{1}{m} \sum\limits_{k \in K} \nabla_\theta a_k(s_k) \cdot \nabla_{a_k} R_k(s_k, a_k) \\ \nabla_{a_k} R_k(s_k, a_k) = \frac{\partial \Phi_k^{loss}}{\partial a_k} + \frac{\partial \Phi_k^{voltage}}{\partial a_k} + \frac{\partial \Phi_k^{viol}}{\partial a_k} \\ \theta = \theta' + \eta \nabla_\theta \end{cases} \tag{19}$$

where m represents the number of samples contained in the sample dataset K, ∇_θ represents the gradient to be updated for the neural network parameters of the agent. s_k, a_k and $R_k\left(s_k, a_k\right)$ represent the state, action taken, and reward obtained by the agent in sample data k, respectively. $\nabla_\theta a_k\left(s_k\right)$ represents the gradient of the action a_k taken by the agent with respect to the state s_k in sample data k, $\nabla_{a_k} R_k\left(s_k, a_k\right)$ represents the gradient of the value reward $R_k\left(s_k, a_k\right)$ obtained after the agent takes action a_k with respect to the action a_k in sample data k, P_{loss}^k, V_{vio}^k and Q_{vio}^k represent the active power loss component, node voltage violation component, and reactive power output violation component contained in the value reward $R_k\left(s_k, a_k\right)$ in sample data k, respectively; η represents the learning rate for neural network updates.

The specific calculation formulas for Φ_k^{loss} and Φ_k^{viol} are as follows:

$$\begin{aligned}
\frac{\partial \Phi_k^{loss}}{\partial a_p^k} &= \frac{\partial \Phi_k^{loss}}{\partial V_i^k}\frac{\partial V_i^k}{\partial a_p^k} \\
&= \sum_{l=1}^{N_b} 2g_l\left[\left(V_i^k - V_i^k\cos\theta_{ij}\right)\frac{\partial V_i^k}{\partial a_p^k} + \left(V_j^k - V_j^k\cos\theta_{ij}\right)\frac{\partial V_j^k}{\partial a_p^k}\right] \\
&= \sum_{l=1}^{N_b} 2g_l\left[\left(V_i^k - V_i^k\cos\theta_{ij}\right)Sen_{ip}^k + \left(V_j^k - V_j^k\cos\theta_{ij}\right)Sen_{jp}^k\right]
\end{aligned} \tag{20}$$

$$\frac{\partial \Phi_k^{viol}}{\partial a_p^k} = \frac{1}{N_{vio,\max}^k + N_{vio,\min}^k}\left(\sum_{i=1}^{N_{vio,\max}^k} Sen_{ip}^k - \sum_{i=1}^{N_{vio,\min}^k} Sen_{ip}^k\right) \tag{21}$$

where a_p^k represents the reactive power regulation action taken by the p-th PV in sample data k, V_i^k and V_j^k represent the voltage magnitudes at node i and node j in sample data k, respectively. Sen_{ip}^k represents the voltage sensitivity between node i and the p-th PV access node in sample data k, Sen_{jp}^k represents the voltage sensitivity between node j and the p-th PV access node in sample data k, $N_{vio,\max}^k$ and $N_{vio,\min}^k$ represent the number of nodes violating the upper voltage limit and the lower voltage limit in sample data k, respectively.

The training flowchart for the real-time distribution network voltage control agent adopted in this paper is shown in Fig. 2.

As shown in Fig. 2, the training process for the real-time distribution network voltage control agent is as follows:

Step 1: Randomly initialize the agent's neural network parameters.

Step 2: Perform power flow calculation based on the PV dispatch schedule obtained from the distribution network voltage optimization control strategy to generate the initial state s_0.

Step 3: Obtain the sample dataset K by running a 24-h real-time voltage control simulation based on the agent.

Step 4: Randomly sample m data samples from the sample dataset K as training data.

Step 5: Based on the m data samples, calculate the agent's neural network update gradient ∇_θ according to Eqs. (14)–(19).

Step 6: Update the neural network based on the agent's neural network update gradient ∇_θ.

Step 7: Check if ∇_θ is less than a predetermined very small value. If the condition is met, end the training process; if not, proceed to Step 8.

Step 8: Check whether the iteration residual meets the set condition. If it meets the condition, the training process ends; if not, a new iteration begins and step two is restarted.

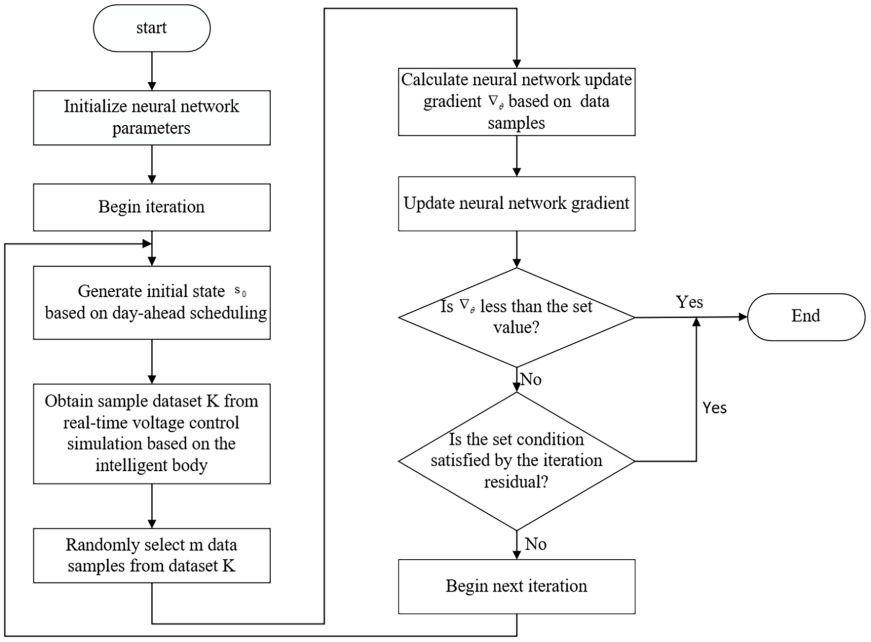


Figure 2: Training flowchart of the real-time distribution network voltage control agent.

5 Case Study

5.1 Test System Description

The numerical simulations are conducted on a computer with a 3.50 GHz CPU and 16 GB RAM to verify the effectiveness of the proposed method. The model is programmed in Visual Studio Code, and the neural network design and training process are completed using Pytorch 2.6 framework while network parameters are updated by the Adam optimizer. The power flow simulation is conducted via Matpower. A modified IEEE 33 bus test system is employed, and the topology is shown in Fig. 3. The voltage security limits for all other nodes are set according to grid operational safety standards, ranging from 0.95 to 1.05 p.u. The capacities, operating parameters, and installation node locations of the PVs are shown in Table 2. The trading and testing database is established, while the electricity consumption data is derived from Portugal, while the active power data is derived from the operation data provided by the Belgian grid operator Elia Group [32].

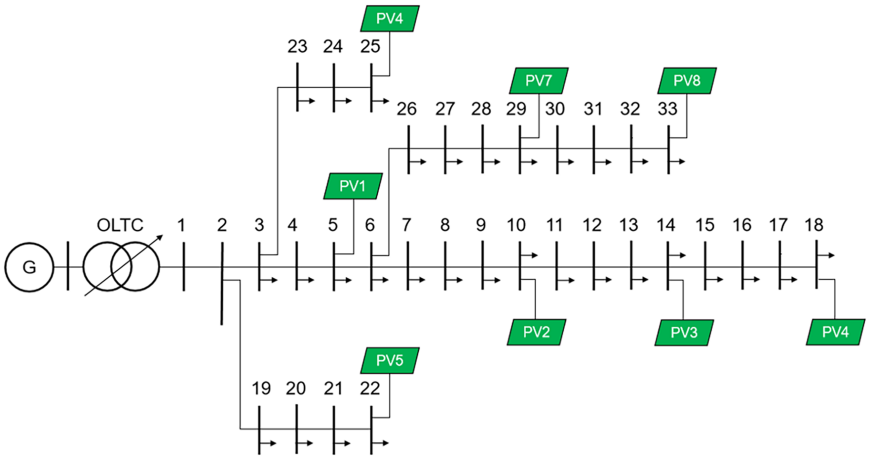


Figure 3: The topology of IEEE 33 bus test system.

Table 2: The capacities and node locations of PVs.

PVs	Capacity	Node
PV1	400 kVA	5
PV2	400 kVA	10
PV3	400 kVA	14
PV4	400 kVA	18
PV5	500 kVA	22
PV6	400 kVA	25
PV7	400 kVA	29
PV8	400 kVA	33

As for the deep neural network employed in this paper, it has 2 hidden layers, with the first layer containing 800 neurons and the second layer containing 400 neurons. The sigmoid function is employed as the activation function between the input layer and hidden layer, while the tanh function is utilized as the activation function for the hidden layers, as well as between the hidden layer and the output layer. The learning rate for the deep neural network gradient update strategy is set as $\eta = 0.0001$. The penalty coefficient for the voltage violation is set as $\lambda_2 = 10$ to guarantee the superiority of safety operation. The penalty factors optimization, which can avoid the influence of manual choice, is not the main research scope of this paper, and the details can be found in ref. [33].

The algorithm overall uses the Adam optimizer to update the network parameters, and the upper limit for the number of training iterations is set to 1000. Specific settings for the deep neural network structure parameters are shown in Table 3.

Table 3: Deep neural network parameters.

Input Layer	Activation Function	Hidden Layer 1	Activation Function	Hidden Layer 2	Activation Function	Output Layer
52	sigmoid	800	tanh	400	tanh	8

5.2 Data-Driven Voltage Control Offline Training Performance

With the data-driven voltage control model and deep neural network setting, the offline training is conducted, while the action-value curve during the training process for the agent of a high PV penetrated distribution network voltage control is shown in Fig. 4. It can be seen that the action value of the agent rises rapidly in the early stage of training and stabilizes after 400 episodes. During the early stage, the reward increases rapidly as the agent adjusts its action space during exploration, quickly resulting from the large penalty coefficient for voltage deviation beyond the permissible range. Then, the voltage safety constraints can be satisfied with the controlling decisions determined by the agent. In the middle and later stages, after the 400 episodes, the reward values stabilize generally, and it still shows slight oscillations. Moreover, the oscillation amplitude shows a converging trend. During this period, the voltage safety constraints can be guaranteed, then the voltage violation value in the action-value function is 0 while the active power losses are non-zero. Then, the agent continuously fine-tunes its control decisions during exploration to further optimize the active power loss during control.

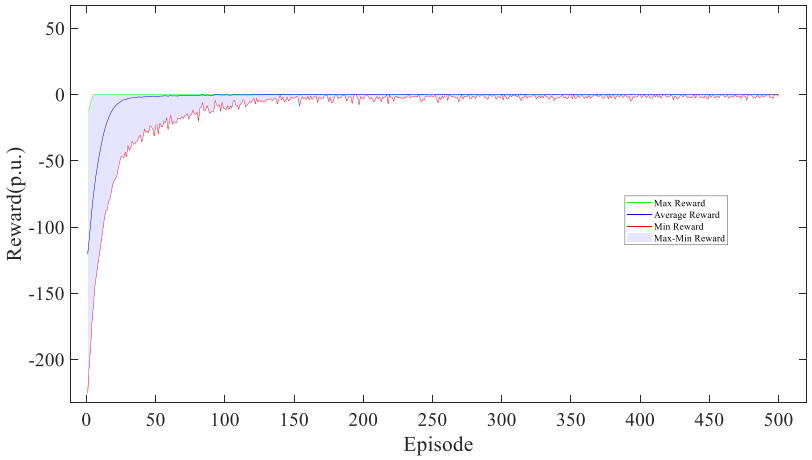


Figure 4: The rewards vale during training procedure.

5.3 Data-Driven Voltage Control Testing Performance

In order to validate the effectiveness of the proposed method, the dynamic and fluctuation characteristic of PV active power is accounted for and a 24-h test scenario using real data from a single day is established. Then, the simulation results of the real-time reactive power output of the PV inverters are shown in Fig. 5. It should be noted that the day-ahead reactive power optimization is scheduled firstly, then the data-driven voltage control is implemented to adjust the reactive power output of PVs response to the real-time PV active power fluctuation.

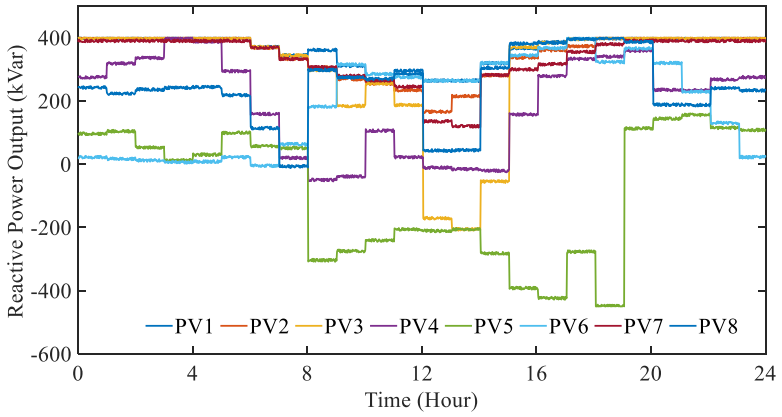
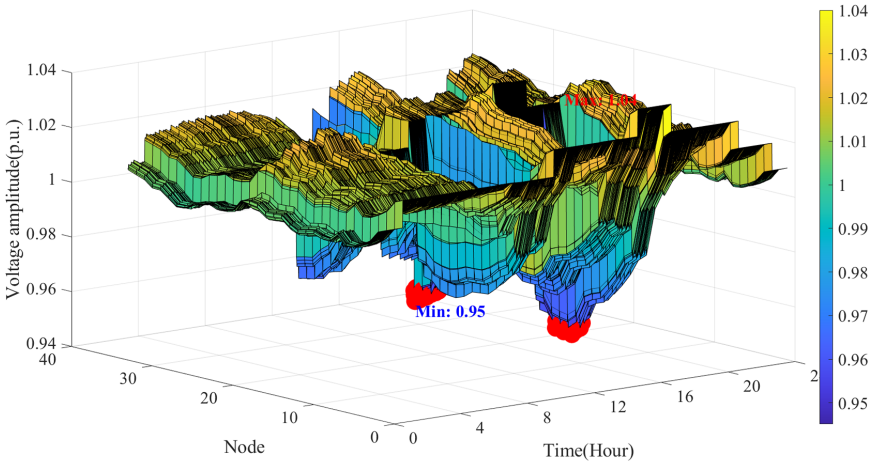


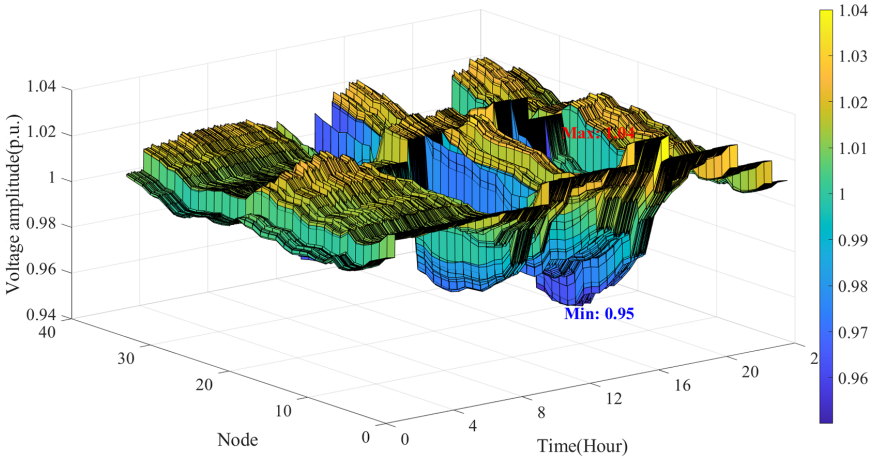
Figure 5: The real-time reactive power output of PVs.

It can be seen that the reactive power outputs are adjusted according to the real-time fluctuations of load demands and PV active power supplies. During the controlling procedure, the trained policy responds to voltage deviations online and adjusts PV inverter reactive power outputs in real time, while strictly respecting inverter capability constraints. In the recorded results, the reactive power outputs always remain within the feasible range from −450 to 400, and no reactive power constraint violations occur during the entire 24-h operation. Only a small fraction of PV–time points reach the capability boundaries. About 7.84 percent of the points reach the upper reactive power limit and about 0.19 percent reach the lower reactive power limit, which indicates that saturation is mainly triggered under stressed operating conditions. Among these PVs, the PV1, PV2, and PV7 exhibit pronounced fluctuations in reactive power output. This is primarily due

to the significant variability in the PV active power supply and reactive load, which leads to more evident voltage violations. In contrast, PV3, PV4, PV5, and PV6 show relatively stable output. Moreover, it can be seen that during the time period of 10:00–14:00 when the PV provides large amount of active power, most of the PVs provide low reactive power while some only absorb reactive power from the distribution network. This is because the voltage can be effectively lowered to prevent it from exceeding the safe limits, by decreasing reactive power injection or drawing reactive power from the grid. Then, the comparison between the proposed method and the traditional model-based two stage voltage control [8] method is shown in Fig. 6.



(a) The voltage profile with traditional voltage control



(b) The voltage profile with proposed data-driven voltage control

Figure 6: The comparison of voltage profiles.

From Fig. 6a, it can be seen that with the traditional voltage control method, it occurs 66 times undervoltage violations with voltage lower than 0.95. As for the entire dispatch time period, there were more than 2.15% of instances where the voltage exceeded the lower limit. The bus 13, bus 14, bus 15, bus 16, bus 31, bus 32 and bus 33 are the main buses affected much. The worst-case voltage reaches 0.9451 p.u. at bus 33 around 17:06. At peak moments such as 17:19 and 17:33, up to 6 buses violate the lower voltage limit simultaneously. These results confirm that during the peak-load and high-fluctuation period, the traditional voltage control method cannot ensure the distribution network safe operation. As seen in Fig. 6b, the voltage

profile of all buses during the whole dispatch period is regulated within the safe scope. Comparing with the traditional voltage control method, the voltage exceeding the lower limit is avoided and the power quality is improved. Corresponding with the voltage profile, the power losses with the proposed method and the traditional method is shown in Fig. 6.

From Fig. 7, it can be seen that during 0:00–7:59 and 19:00–23:59, when the distribution network load is low and fluctuations are small, the proposed voltage control method reduces the power losses significantly comparing with the traditional method. During 8:00–9:59, 12:00–14:59, and 18:00–18:59, when the load is low and PVs' active power fluctuations are large, the power losses reduction is less, this is because reactive power distribution is optimized to satisfy the voltage safety constraints which has higher superiority. More prominent, during 10:00–12:59 and 15:00–17:59, when the load is high and PVs' active power fluctuations are significant, the agent determines PV reactive power decisions to maintain the voltage safety firstly, which leads to an increase in power losses.

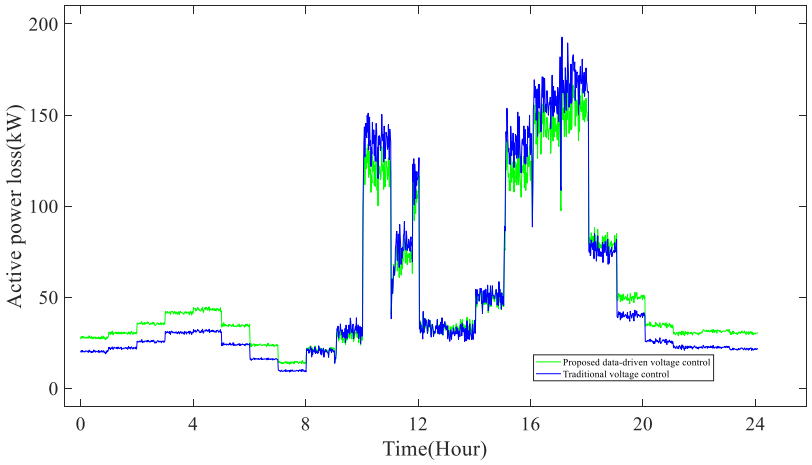


Figure 7: The power losses of distribution network for the dispatch time period.

With the proposed method, the voltage profiles are regulated within the safe scope under continuously fluctuating load and PV generation conditions, while also guaranteeing inverter reactive power feasibility. The quantitative results further confirm that the method not only eliminates voltage violations in the tested 24-h scenario, but also utilizes inverter reactive capability in a constrained and purposeful manner, with saturation occurring primarily during stressed operating intervals. Furthermore, the power losses are also reduced significantly with the proposed method.

To validate the effectiveness of the proposed data-driven voltage control method, the comparisons between the OPF-based voltage control (OPF-VC) [8], the conventional droop-based voltage control (CD-VC) [5], and the standard DDPG algorithm are conducted. To ensure a comprehensive and reliable evaluation, the testing phase was designed by randomly selecting 48,000 scenarios from the complete set of 17,520 historical scenarios, thereby constructing a set of independent samples for the systematic evaluation of the trained data-driven voltage control policy. The detailed comparisons are shown in Table 4.

Table 4: The comparisons between four voltage control methods.

	Proposed Method	OPF-VC	CD-VC	DDPG
Average voltage deviation (p.u.)	0.0146	0.0171	0.0202	0.0155
Voltage limit violations Percentage	0	7.52%	21.76%	0.21%
Average power losses (kW)	51.13 kW	58.33 kW	69.53 kW	53.47 kW

The comparative results in Table 4 demonstrate the superior performance of the proposed sensitivity-guided DDPG method. Notably, it achieves zero voltage limit violations, outperforming both the model-based OPF-VC (7.52% violations) and the conventional droop-based CD-VC (21.76% violations), while also reducing the average voltage deviation and power losses to the lowest levels among all methods. This validates the core claim that embedding physical voltage sensitivity into the policy gradient significantly enhances control safety and economic efficiency. Although the standard DDPG also performs well, its slight violation percentage as 0.21% and marginally higher losses highlight the limitations of a purely data-driven approach. Moreover, the proposed method also provides the lowest power losses, which shows the effectiveness of voltage control for distribution network.

6 Conclusion

This paper proposes a data-driven voltage control method with a voltage sensitivity guided DDPG algorithm for a distribution network with high penetration of PVs. The traditional voltage control methods are ineffective against the rapid and intense disturbance of distribution network operation status, while the traditional DDPG suffers from slow convergence and training instability. The node voltage-reactive power injection sensitivity matrix is integrated into the policy gradient calculation, and this physics-informed guidance replaces the Critic's approximate gradient. Consequently, the approximation errors and variance are reduced, and the training stability and convergence speed are significantly enhanced. Simulations on the modified IEEE 33-node system demonstrate that the trained agent fully eliminates daily voltage violations and reduces power losses. In conclusion, embedding physical knowledge into a DRL framework enables fast, stable voltage control for distribution systems. The future work will focus on the data-driven voltage control in more complex scenarios, such as variable topologies, communication failures, and explore the optimization of the network architecture for efficiency and its formal sensitivity analysis.

Although the proposed method performs effectively in simulation, several factors may impact real-world deployment. The distribution-level data acquisition and quality can be limited, including incomplete measurements, missing or delayed telemetry. The communication infrastructure constraints may cause imperfect or intermittent observations and control signal delivery, which can degrade closed-loop performance if not explicitly addressed. Meanwhile, distribution network digitalization—such as increased deployment of sensing, feeder automation, edge computing, and improved communication networks—is expected to progressively mitigate these barriers and provide a stronger foundation for practical closed-loop data-driven voltage control.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the State Grid Jiangsu Electric Power Co., Ltd. Science and Technology Project (Grant No. J2024148).

Author Contributions: The authors confirm contribution to the paper as follows: paper structure organization and writing, Xiaolong Xiao and Linghao Zhang; methodology and supervision, Ziran Guo and Xiaoxing Lu; programming

and editing, Wenqiang Xie, Shukang Lv and Peishuai Li. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article. And the additional data that support the findings of this study are available on request from the corresponding author.

Ethics Approval: This study did not involve any human or animal subjects.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. National Development and Reform Commission (NDRC). Action plan for carbon dioxide peaking before 2030. 2021 Oct 27 [cited 2023 Jan 27]. Available from: https://en.ndrc.gov.cn/policies/202110/t20211027_1301020.html.
2. Ratnam KS, Palanisamy K, Yang G. Future low-inertia power systems: requirements, issues, and solutions—a review. Renew Sustain Energy Rev. 2020;124:109773. doi:10.1016/j.rser.2020.109773.
3. Farrokhabadi M, Cañizares CA, Simpson-Porco JW, Nasr E, Fan L, Mendoza-Araya PA, et al. Microgrid stability definitions, analysis, and examples. IEEE Trans Power Syst. 2020;35(1):13–29. doi:10.1109/TPWRS.2019.2925703.
4. Bolognani S, Zampieri S. A distributed control strategy for reactive power compensation in smart microgrids. IEEE Trans Autom Control. 2013;58(11):2818–33. doi:10.1109/TAC.2013.2270317.
5. Zhu H, Liu H. Fast local voltage control under limited reactive power: optimality and stability analysis. IEEE Trans Power Syst. 2016;31(5):3794–803. doi:10.1109/TPWRS.2015.2504419.
6. Kekatos V, Zhang L, Giannakis GB, Baldick R. Voltage regulation algorithms for multiphase power distribution grids. IEEE Trans Power Syst. 2015;31(5):3913–23. doi:10.1109/TPWRS.2015.2493520.
7. Weckx S, Driesen J. Optimal local reactive power control by PV inverters. IEEE Trans Sustain Energy. 2016;7(4):1624–33. doi:10.1109/TSTE.2016.2572162.
8. Li P, Zhang C, Wu Z, Xu Y, Hu M, Dong Z. Distributed adaptive robust voltage/var control with network partition in active distribution networks. IEEE Trans Smart Grid. 2020;11(3):2245–56. doi:10.1109/tsg.2019.2950120.
9. Yang L, He M, Vittal V, Zhang J. Stochastic optimization-based economic dispatch and interruptible load management with increased wind penetration. IEEE Trans Smart Grid. 2016;7(2):730–9. doi:10.1109/TSG.2015.2444813.
10. Growe-Kuska N, Heitsch H, Romisch W. Scenario reduction and scenario tree construction for power management problems. In: Proceedings of the 2003 IEEE Bologna Power Tech Conference; 2003 Jun 23–26; Bologna, Italy. Vol. 3, 7 p. doi:10.1109/PTC.2003.1304379.
11. Zhang B, Lam AYS, Domínguez-García AD, Tse D. An optimal and distributed method for voltage regulation in power distribution systems. IEEE Trans Power Syst. 2014;30(4):1714–26. doi:10.1109/TPWRS.2014.2347281.
12. Singhal A, Ajjarapu V, Fuller J, Hansen J. Real-time local volt/var control under external disturbances with high PV penetration. IEEE Trans Smart Grid. 2019;10(4):3849–59. doi:10.1109/TSG.2018.2840965.
13. Tang Y, Dvijotham K, Low S. Real-time optimal power flow. IEEE Trans Smart Grid. 2017;8(6):2963–73. doi:10.1109/TSG.2017.2704922.
14. Sutton RS, Barto AG. Reinforcement learning: an introduction. 1st ed. Cambridge, MA, USA: MIT Press; 1998. p. 445–551.
15. Busoniu L, Babuska R, De Schutter B, Ernst D. Reinforcement learning and dynamic programming using function approximators. 1st ed. Boca Raton, FL, USA: CRC Press; 2010. doi:10.1201/9781439821091.
16. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA, USA: MIT Press; 2016.
17. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. arXiv:1312.5602. 2013.
18. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33. doi:10.1038/nature14236.

19. Huang Q, Huang R, Hao W, Tan J, Fan R, Huang Z. Adaptive power system emergency control using deep reinforcement learning. IEEE Trans Smart Grid. 2020;11(2):1171–82. doi:10.1109/TSG.2019.2933191.
20. Zeng H, Zhou Y, Guo Q, Cai Z, Sun H. Distributed deep reinforcement learning-based approach for fast preventive control considering transient stability constraints. CSEE J Power Energy Syst. 2023;9(1):197–208. doi:10.17775/CSEEJPES.2020.04610.
21. Liu H, Wu W. Online multi-agent reinforcement learning for decentralized inverter-based Volt-VAR control. IEEE Trans Smart Grid. 2021;12(4):2980–90. doi:10.1109/TSG.2021.3060027.
22. Cao D, Zhao J, Hu W, Yu N, Ding F, Huang Q, et al. Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems. IEEE Trans Smart Grid. 2022;13(1):149–65. doi:10.1109/TSG.2021.3113085.
23. Peng H, Liao K, Yang J, Pang B, He Z. Deep reinforcement learning based multi-timescale Volt/Var control in distribution networks considering network reconfiguration. IEEE Trans Sustain Energy. 2025;16(4):2948–58. doi:10.1109/TSTE.2025.3574806.
24. Dorokhova M, Martinson Y, Ballif C, Wyrsch N. Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation. Appl Energy. 2021;301(18):117504. doi:10.1016/j.apenergy.2021.117504.
25. Park K, Moon I. Multi-agent deep reinforcement learning approach for EV charging scheduling in a smart grid. Appl Energy. 2022;328:120111. doi:10.1016/j.apenergy.2022.120111.
26. Wang X, Zhong H, Zhang G, Ruan G, He Y, Yu Z. Adaptive look-ahead economic dispatch based on deep reinforcement learning. Appl Energy. 2024;353:122121. doi:10.1016/j.apenergy.2023.122121.
27. Wang Y, Qiu D, Sun M, Strbac G, Gao Z. Secure energy management of multi-energy microgrid: a physical-informed safe reinforcement learning approach. Appl Energy. 2023;335(3):120759. doi:10.1016/j.apenergy.2023.120759.
28. Wu C, Mohsenian-Rad H, Huang J. Vehicle-to-aggregator interaction game. IEEE Trans Smart Grid. 2012;3(1):434–42. doi:10.1109/TSG.2011.2166414.
29. Tushar W, Yuen C, Mohsenian-Rad H, Saha T, Poor HV, Wood KL. Transforming energy networks via peer-to-peer energy trading: the potential of game-theoretic approaches. IEEE Signal Process Mag. 2018;35(4):90–111. doi:10.1109/MSP.2018.2818327.
30. Xiao D, Deng W, Liu B, Huang W, Zhu Z. Game-theoretic energy scheduling for community multi-energy system with electric vehicle aggregator leveraging quantum swarm intelligence. Int J Electr Power Energy Syst. 2025;172(8):111200. doi:10.1016/j.ijepes.2025.111200.
31. Cheng L, Yu T. Game-theoretic approaches applied to transactions in the open and ever-growing electricity markets from the perspective of power demand response: an overview. IEEE Access. 2019;7:25727–62. doi:10.1109/ACCESS.2019.2900356.
32. Li P, Shen J, Wu Z, Yin M, Dong Y, Han J. Optimal real-time voltage/var control for distribution network: droop-control based multi-agent deep reinforcement learning. Int J Electr Power Energy Syst. 2023;153(1):109370. doi:10.1016/j.ijepes.2023.109370.
33. Li H, Wan Z, He H. Constrained EV charging scheduling based on safe deep reinforcement learning. IEEE Trans Smart Grid. 2020;11(3):2427–39. doi:10.1109/TSG.2019.2955437.