mathematics

MDPI

Article

An Optimal Scheduling Method for Power Grids in Extreme Scenarios Based on an Information-Fusion MADDPG Algorithm

Xun Dou [1], Cheng Li [1], Pengyi Niu [1], Dongmei Sun [1], Quanling Zhang [1,*] and Zhenlan Dou [2]

[1] College of Electrical Engineering and Control Science, Nanjing Tech University, Nanjing 211816, China; dxnjut@njtech.edu.cn (X.D.); w1768763897@163.com (C.L.); niupengyue@foxmail.com (P.N.); sundm75@163.com (D.S.)
[2] State Grid Shanghai Municipal Electric Power Company, Shanghai 200122, China; douzhl@126.com
* Correspondence: zhangquanling@njtech.edu.cn; Tel.: +86-13957137904

Abstract

With the large-scale integration of renewable energy into distribution networks, the intermittency and uncertainty of renewable generation pose significant challenges to the voltage security of the power grid under extreme scenarios. To address this issue, this paper proposes an optimal scheduling method for power grids under extreme scenarios, based on an improved Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm. By simulating potential extreme scenarios in the power system and formulating targeted secure scheduling strategies, the proposed method effectively reduces trial-and-error costs. First, the time series clustering method is used to construct the extreme scene dataset based on the principle of maximizing scene differences. Then, a mathematical model of power grid optimal dispatching is constructed with the objective of ensuring voltage security, with explicit constraints and environmental settings. Then, an interactive scheduling model of distribution network resources is designed based on a multi-agent algorithm, including the construction of an agent state space, an action space, and a reward function. Then, an improved MADDPG multi-agent algorithm based on specific information fusion is proposed, and a hybrid optimization experience sampling strategy is developed to enhance the training efficiency and stability of the model. Finally, the effectiveness of the proposed method is verified by the case studies of the distribution network system.

Keywords: reinforcement learning; distribution network; power grid voltage security; experience fusion strategy

MSC: 62M45

check for updates

Academic Editors: Lefeng Cheng and Wentian Lu

Received: 8 September 2025
Revised: 24 September 2025
Accepted: 28 September 2025
Published: 3 October 2025

Citation: Dou, X.; Li, C.; Niu, P.; Sun, D.; Zhang, Q.; Dou, Z. An Optimal Scheduling Method for Power Grids in Extreme Scenarios Based on an Information-Fusion MADDPG Algorithm. *Mathematics* **2025**, *13*, 3168. https://doi.org/10.3390/math13193168

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the large-scale integration of new energy sources, the penetration of distributed generation (DG) dominated by renewables in distribution networks has steadily increased. High DG penetration aggravates three-phase imbalance and reverse power flow, which significantly raises the difficulty of voltage control in distribution systems [1]. Owing to the intermittency and uncertainty of renewable output, meteorological factors can trigger extreme renewable-energy scenarios characterized by low frequency and high impact [2]. Meanwhile, distribution networks are inherently weak and lack inertia support; under extreme scenarios, voltage instability may further induce cascading failures [3,4], posing severe challenges to network security. To effectively cope with potential extreme scenarios under extreme weather conditions and to reduce trial-and-error costs in actual operation, it

is necessary to provide a secure dispatching scheme that can respond to possible extreme events, thereby ensuring the stable and economic operation of the power system [5,6].

Extensive research on the optimal dispatching of distribution networks has been carried out worldwide, mainly along two lines: optimization-based scheduling and reinforcement learning (RL)-based scheduling.

In optimization-based scheduling studies, two categories prevail: traditional optimization methods that build grid dispatch models using mathematical optimization algorithms such as linear programming (LP), quadratic programming (QP), and mixed-integer linear programming (MILP); and intelligent optimization algorithms that solve complex nonlinear dispatch problems, such as genetic algorithms (GA) and particle swarm optimization (PSO).

For traditional optimization methods, Ref. [7] proposes a two-stage robust optimization method for microgrids based on data-driven uncertainty sets. By using conditional normal Copula and support vector clustering, a compact and outlier-resistant uncertainty set is constructed. A two-stage model and column-and-constraint generation algorithm are employed to optimize the dispatch plan, providing an efficient solution to wind power uncertainty. Ref. [8] constructs a multi-period dispatchable region of a microgrid via recursive robust convex hull fitting and incorporates it into a distribution network optimization model. A data-driven voltage security region enables voltage model-free regulation under incomplete network parameters, and a hierarchical distributed algorithm efficiently solves the multi-level optimal power flow model. Ref. [9] obtains typical scenarios using clustering and scenario reduction techniques, proposes an optimization model that accounts for network loss and its probability, and introduces a network loss service fee. The model is transformed into a convex formulation and solved with a solver. Ref. [10] presents an optimization-based rolling scheduling strategy (ORoHS) with time partitioning, dynamically dividing prediction and execution periods to balance dispatch cost and forecasting accuracy. A robust optimization model is adopted for energy trading scheduling, verifying the superiority of the method in the cost–accuracy trade-off. Ref. [11] employs an MILP approach on the GAMS platform to achieve optimal operational scheduling of wind turbines, solar units, fuel cells, and batteries, and solves it using the CPLEX solver, demonstrating effectiveness in addressing isolated dispatch problems of DC microgrids with renewable energy parks.

For intelligent optimization algorithms, Ref. [12] proposes an energy optimization scheduling method for a traction power supply system with two-phase access of photovoltaic and energy storage. A stochastic optimization model is formulated with the objectives of minimizing daily operating cost and minimizing energy exchange with the external grid; a multi-objective PSO algorithm is used to solve it, improving system economy and energy efficiency. Ref. [13] proposes a multi-objective optimization method for an islanded green energy system, applying a hybrid meta-heuristic algorithm (HFAPSO) to optimize the configuration of a hybrid renewable energy system (HRES). Comparative results show that HFAPSO achieves the best performance in system configuration.

In summary, traditional optimization methods are highly accurate and theoretically sound, making them suitable for small-scale static problems; however, they suffer from high computational complexity and difficulty in handling nonlinear problems. Intelligent optimization algorithms possess strong global search capability and are well-suited to complex nonlinear problems, but they are prone to local optima and require careful parameter tuning. Compared with traditional and intelligent optimization methods, RL-based intelligent scheduling offers stronger dynamic adaptability, the ability to handle complex problems, and self-learning mechanisms—enabling real-time policy adjustment and exploration of unknown spaces.

In RL-based scheduling studies, RL autonomously optimizes strategies via online interaction without relying on precise mathematical models, thereby dynamically adapting to complex environmental changes. It effectively handles high-dimensional state spaces, dynamic disturbances, and multi-objective trade-offs, giving it significant advantages in power system dispatch [14]. RL algorithms can be viewed as a combination of value-based and policy-based methods, with the actor–critic (AC) framework being the most common approach for integrating policy and value functions—for example, deep deterministic policy gradient (DDPG) [15] and soft actor–critic (SAC) [16]. The task in this paper concerns the voltage-security-oriented optimal dispatching of distribution networks: the policy algorithm generates specific continuous actions (e.g., output settings of relevant units), while the value algorithm evaluates the long-term value of these continuous actions. This structure enables the AC framework to excel in handling complex nonlinear problems and multi-objective optimization; thus, it is commonly adopted to solve power system optimal dispatch problems [17].

In the context of preparing secure plans for extreme grid scenarios, substantial computational resources are required. Methods for enhancing model training efficiency while ensuring quality have attracted extensive research. Ref. [18] adopts a centralized training–distributed execution framework combined with a prioritized experience replay mechanism, proposing a multi-agent prioritized twin-delayed DDPG algorithm for optimal dispatch, which improves learning efficiency and training stability. Ref. [19] modifies the proximal policy optimization (PPO) algorithm by introducing a GAN, reducing variance and improving exploration efficiency, and thereby shortening training time and enhancing renewable energy accommodation. Ref. [20] applies an AC framework and introduces a pre-state and environmental feature adaptation mechanism; the pre-state reduces the state space, and the adaptation mechanism improves the agents' adaptability to changes, enhancing convergence speed and stability. Refs. [21,22] design expert policies for secure grid operation and power balance control, and—combined with imitation learning—train a GESIL agent via offline–online imitation learning for grid dispatch decision-making, significantly improving computational speed and decision quality. Although these methods achieve notable advancements in improving training efficiency and optimizing grid dispatch, they each have limitations. The prioritized twin-delayed DDPG algorithm can still exhibit high complexity and slow convergence in large-scale scenarios; the GAN-modified PPO reduces variance and improves exploration, but GAN training is unstable and computationally expensive; the AC framework with pre-state and adaptive mechanisms improves convergence and adaptability, yet initial training performance may be poor and implementation complex; the offline–online imitation learning approach substantially accelerates computation and decision making but relies heavily on high-quality expert policies and offline data [23].

In summary, to develop a feasible safety-preparedness scheme for extreme grid scenarios, this study adopts a MADDPG algorithm within a centralized training/decentralized execution (CTDE) framework. Task-specific information and a decaying Gaussian noise term are injected into the value network to enhance training stability. A hybrid prioritized experience sampling strategy is designed to improve learning efficiency. Incorporating imitation-learning concepts, an extreme scenario dataset and a coach dataset are selected via clustering and a scenario-dissimilarity criterion, providing the data foundation for the safety plan. The reference policy obtained by training on the coach data is then used as experience samples for the model, further accelerating the training process.

The innovative contributions of this study are mainly reflected in the following two aspects:

1. An MADDPG algorithm based on the integration of specific information is proposed, in which both the policy network and the value network adopt lightweight MLP modules. By incorporating structured experiential knowledge into the input of the value network and introducing decaying Gaussian noise to the state and action spaces, the robustness of the value network input is enhanced, thereby improving the stability of model training.
2. A hybrid prioritized experience sampling strategy is designed, which integrates hierarchical sampling to extract experience samples across different time steps, while combining them with the most recent experience samples to construct fused experience data. This ensures that the model can promptly learn from and adapt to recent environmental changes or policy updates. In addition, reference experiences obtained from coach-guided training are inserted into the replay buffer after multiple training iterations, further enhancing the learning efficiency of the model.

2. Framework for Voltage-Security-Oriented Optimal Dispatch of the Power Grid

Figure 1 presents the overall workflow of the proposed method, which mainly comprises the following: the design of the mathematical model for power grid dispatch, the design of the multi-agent interaction model, and improvements to the multi-agent intelligent algorithm.

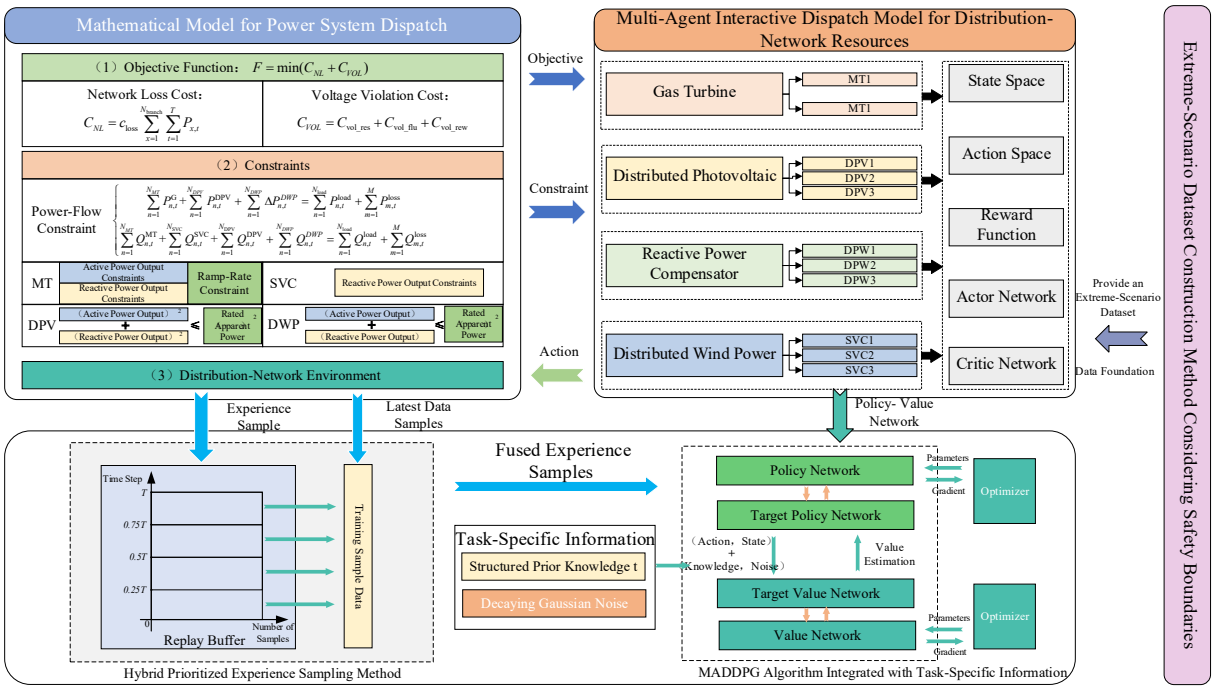


Figure 1. Considering the overall framework of grid voltage security optimization scheduling.

To address the optimal dispatch problem of power systems under extreme scenarios, this paper proposes a construction method for an extreme scenario dataset that explicitly considers safety boundaries, thereby providing the data foundation for subsequent optimal dispatch. For the voltage-security-oriented optimal dispatch task, we first formulate a mathematical model of grid dispatch, including: an objective function that minimizes network losses while ensuring voltage security; power flow constraints; active/reactive power output constraints for various resources; and the configuration of the distribution network environment.

Next, based on the distribution network setting, we design a multi-agent interactive dispatch model for network resources, specifying the state space, action space, and reward function for different types of agents, as well as their corresponding policy and value networks. The action space is restricted according to the output constraints defined in the mathematical model.

Finally, we develop a multi-agent algorithm that integrates task-specific information. By introducing structured prior knowledge and a decaying Gaussian noise term, we enhance the stability of model training. Moreover, by combining stratified random sampling with prioritized experience replay in the replay buffer, we design a hybrid prioritized experience sampling strategy to further improve training efficiency.

3. Method for Constructing an Extreme Scenario Dataset with Safety Boundaries Considered

To tackle the optimal dispatch problem of power systems under extreme scenarios, it is essential to supply a sufficiently large corpus of extreme data. If only "typical" extreme scenarios are considered, the preparatory schemes designed to handle such events may lack an adequate safety margin and thus fail to meet supply–demand challenges under truly extreme conditions.

Drawing on the approach in Ref. [24] to generate seasonal extreme scenario data as the baseline, this paper proposes an extreme scenario dataset construction method that maximizes scenario dissimilarity while explicitly considering safety boundaries. First, the data are partitioned into four seasons. For the extreme scenario data of each season, a time series K-means clustering method is employed. Interval statistics and the silhouette coefficient are used for cross-validation to determine the optimal number of cluster centers for each season. Second, based on the identified number of cluster centers, seasonal data are clustered. For each cluster, the N most dissimilar extreme scenarios are selected according to a scenario-dissimilarity maximization principle. Finally, these maximally dissimilar extreme scenarios are combined with the corresponding cluster centers to form a dataset capable of covering extreme patterns. The cluster center data of each season are further organized into a "coach" dataset, providing the data foundation for subsequent optimization dispatch tasks under extreme scenarios.

The overall process of the proposed extreme scenario dataset construction method with safety boundaries is illustrated in Appendix A, Figure A1.

4. Power Grid Optimal Dispatch Model Based on the Improved MADDPG Algorithm

4.1. Mathematical Model of Resource-Interactive Dispatch for Voltage-Security-Oriented Optimal Scheduling

4.1.1. Objective Function

With the large-scale integration and high penetration of renewable energy, power systems are increasingly exposed to extreme-weather-induced demand fluctuations, making such extreme scenarios more frequent. Against this backdrop, prioritizing voltage security in dispatch is crucial if we are to avoid stability issues such as voltage collapse, thereby protecting grid equipment and improving power supply quality. Focusing on distribution network operations, this study leverages the coordinated control of distributed generators (DGs) and static var compensators (SVCs) to improve power distribution and system stability, reduce network losses, and ensure the safe and stable operation of the distribution network under extreme scenarios. Accordingly, taking conventional resources within the distribution network as the control objects, the objective is to minimize the

optimal dispatch cost over the operating horizon, T, comprising the network loss cost C_{NL} and the voltage violation cost C_{VOL}. The objective function is given as follows:

$$F = \min(C_{NL} + C_{VOL}) \tag{1}$$

The comprehensive dispatch cost of the distribution network system is composed as follows:

(1) Network loss cost:

$$C_{NL} = c_{loss} \sum_{x=1}^{N_{branch}} \sum_{t=1}^{T} P_{x,t} \tag{2}$$

$$P_{x,t} = \sum_{i=1}^{I} \sum_{j=1}^{J} U_t^i U_t^j \left(\frac{\cos\delta_t^i - \cos\delta_t^j}{R_x} + \frac{\sin\delta_t^i - \sin\delta_t^j}{X_x}\right) \tag{3}$$

where c_{loss} is the unit cost coefficient of network losses in the distribution network; N_{branch} is the total number of branches; U_t^i and U_t^j are the voltage magnitudes at buses i and j at time t; δ_t^i and δ_t^j are the corresponding voltage phase angles at buses i and j at time t; and R_x and X_x denote the line resistance and reactance, respectively, of branch x between buses i and j.

(2) Voltage violation cost:

To satisfy system voltage stability and quality requirements, a virtual penalty cost is introduced for voltages, comprising: the over-limit voltage penalty cost C_{vol_res}, the voltage fluctuation penalty cost C_{vol_flu}, and the normal-voltage reward C_{vol_rew}.

$$C_{VOL} = C_{vol_res} + C_{vol_flu} + C_{vol_rew} \tag{4}$$

$$C_{vol_res} = c_{vol_res} \sum_{t=1}^{T} \sum_{n=1}^{N_{node}} \left(\frac{\Delta U_{n,t}}{U_n^{\max} - U_n^{\min}}\right) \tag{5}$$

$$\Delta U_{n,t} = \begin{cases} U_n^{\min} - U_{n,t} & U_{n,t} < U_n^{\min} \\ 0 & U_n^{\min} \le U_{n,t} \le U_n^{\max} \\ U_{n,t} - U_n^{\max} & U_{n,t} > U_n^{\max} \end{cases} \tag{6}$$

$$C_{vol_flu} = c_{vol_flu} \sum_{n=1}^{N_{node}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (U_{n,t} - \overline{U_n})^2} \tag{7}$$

$$C_{vol_rew} = c_{vol_rew} \sum_{t=1}^{T} \sum_{n=1}^{N_{node}} \left(1 - |U_{n,t} - 1| - 0.5(U_n^{\max} - U_n^{\min})^2\right) \tag{8}$$

where C_{VOL} defines the overall cost related to voltage quality, This structure provides a more nuanced objective than a simple limit check; C_{vol_res} is the primary penalty function, which activates only when a bus voltage goes outside the predefined safe range. Its purpose is to strongly discourage dispatch actions that lead to direct violations of operational security standards. The fluctuation penalty fluctuation, C_{vol_flu}, penalizes the deviation of the current voltage from the average voltage over the entire scheduling horizon. The goal here is to promote a smoother voltage profile over time, reducing the stress on voltage regulation equipment and improving power quality, even if the voltage is technically within the safe limits; C_{vol_rew} provides a continuous positive incentive for keeping the voltage within the safe range, encouraging the agent to actively seek out and maintain ideal voltage levels rather than just avoiding the boundaries; $c_{\text{vol_res}}$ and $c_{\text{vol_flu}}$ are the unit cost coefficients for over-limit and fluctuation penalties, respectively; $c_{\text{vol_rew}}$ is the reward coefficient for maintaining normal voltage; $\Delta U_{n,t}$ denotes the voltage deviation; $U_n^{\max}$ and $U_n^{\min}$ are the

prescribed minimum and maximum voltage limits; $U_{n,t}$ represents the voltage magnitude at bus n at time t; and $\overline{U_n}$ is the average voltage at bus n over one scheduling horizon, T.

4.1.2. Constraints

In this study, DG resources in the distribution network are categorized into three types: microturbines (MT), distributed photovoltaics (DPV), and distributed wind power (DWP). MTs participate in voltage regulation directly via reactive power control, whereas DPV and DWP units are scheduled through power electronic converters. To ensure operational stability and voltage security under extreme scenarios, the following settings are adopted: DPVs operate in maximum power point tracking (MPPT) mode under irradiance, and switch to a static synchronous compensator (STATCOM) mode otherwise; DWPs operate in maximum power tracking mode within the rated wind-speed range, and switch to STATCOM mode when wind speed exceeds the limit or when grid faults occur. In STATCOM mode, both types of renewable units contribute to dispatch solely through their reactive power capacity.

(1) Power Balance Constraints:

$$\begin{cases} \sum_{n=1}^{N_{MT}} P_{n,t}^{G} + \sum_{n=1}^{N_{DPV}} P_{n,t}^{DPV} + \sum_{n=1}^{N_{DWP}} \Delta P_{n,t}^{DWP} = \sum_{n=1}^{N_{load}} P_{n,t}^{load} + \sum_{m=1}^{M} P_{m,t}^{loss} \\ \sum_{n=1}^{N_{MT}} Q_{n,t}^{MT} + \sum_{n=1}^{N_{SVC}} Q_{n,t}^{SVC} + \sum_{n=1}^{N_{DPV}} Q_{n,t}^{DPV} + \sum_{n=1}^{N_{DWP}} Q_{n,t}^{DWP} = \sum_{n=1}^{N_{load}} Q_{n,t}^{load} + \sum_{m=1}^{M} Q_{m,t}^{loss} \end{cases} \tag{9}$$

where $P_{n,t}^{MT}$ and $Q_{n,t}^{MT}$ denote the active and reactive power outputs of the MT at time t; $P_{n,t}^{DPV}$ and $Q_{n,t}^{DPV}$ are the active and reactive power outputs of the DPV at time t; $P_{n,t}^{DWP}$ and $Q_{n,t}^{DWP}$ are the active and reactive power outputs of the DWP at time t; $P_{n,t}^{load}$ and $Q_{n,t}^{load}$ represent the active and reactive loads at bus n at time t; $Q_{n,t}^{SVC}$ is the reactive power output of the SVC at time t; $P_{m,t}^{loss}$ and $Q_{m,t}^{loss}$ are the active and reactive power losses of line mmm at time t.

(2) DPV Power Constraints

$$\left(P_{n,t}^{DPV}\right)^2 + \left(Q_{n,t}^{DPV}\right)^2 \le \left(S_n^{DPV}\right)^2 \tag{10}$$

where S_n^{DPV} denotes the rated apparent power of the DPV.

(3) DWP Power Constraints

$$\left(P_{n,t}^{DWP}\right)^2 + \left(Q_{n,t}^{DWP}\right)^2 \le \left(S_n^{DWP}\right)^2 \tag{11}$$

where S_n^{DWP} denotes the rated apparent power of the DWP.

(4) MT Power Constraints

$$P_{n,t}^{MT_{\min}} \le P_{n,t}^{MT} \le P_{n,t}^{MT_{\max}} \tag{12}$$

$$-Q_{n,t}^{MT\max} \le Q_{n,t}^{MT} \le Q_{n,t}^{MT\max} \tag{13}$$

$$(P_{n,t}^{MT})^2 + (Q_{n,t}^{MT})^2 \le (S_{n,\max}^{MT})^2 \tag{14}$$

$$\Delta P_n^{down} \le P_{n,t}^{MT} - P_{n,t-1}^{MT} \le \Delta P_n^{up} \tag{15}$$

where $P_{n,t}^{MT_{\min}}$ and $P_{n,t}^{MT_{\max}}$ are the maximum and minimum active power outputs of the MT at bus n, respectively; $Q_{n,t}^{MT}$ is the maximum reactive power of the MT at bus n; $S_{n,\max}^{MT}$ is the rated capacity of the MT at bus n; ΔP_n^{up} and ΔP_n^{down} are the maximum ramp-up and ramp-down capabilities of the MT at bus n, respectively.

(5) SVC Power Constraints

$$-Q_{n,t}^{SVC_{\max}} \le Q_{n,t}^{SVC} \le Q_{n,t}^{SVC_{\max}} \tag{16}$$

where $Q_{n,t}^{SVC_{\max}}$ denotes the maximum reactive power output of the SVC at bus n.

4.2. Multi-Agent Interactive Dispatch Model for Distribution Network Resources

This study reformulates the voltage-security-oriented optimal dispatch task as a cooperative Markov decision process (MDP)-based multi-agent optimization problem. By assigning MT, DPV, DPW, and SVC resources as independent agents, we construct a decision model with a joint state space (grid parameters/load demand) and a joint action space (equipment control commands), aiming to minimize the global dispatch cost.

The system adopts a two-layer independent-decision-centralized interaction architecture: each agent achieves local optimal control via deep reinforcement learning, while a communication network enables sharing of state information to form global observations. A bidirectional environment feedback mechanism is employed to adjust policies in real time [24]. This design combines dynamic responsiveness with distributed fault tolerance: when a single agent malfunctions, the remaining agents can autonomously update their policies based on environmental changes. Through cooperative exploration, the multi-agent scheme attains progressive global optimization, and its scalability enables the rapid integration of new resources, markedly enhancing adaptability in complex distribution network environments.

The state space, action space, and reward function of the proposed interactive dispatch model are defined as follows.

(1) State Space

The state set, S, of the interactive dispatch model is jointly composed of the states of the MT, DPV, DPW, and SVC agents. Since the bus voltages are directly affected by the injected power at each bus, the state space of each agent primarily includes the following: the power information of the agent's own bus and its observed buses, the current time step, and the agent's active and reactive power outputs at the previous time step. Taking the microturbine (MT) as an example, the state set of this agent at time step t is:

$$S_t^{MT} = \left\{ t, P_{n,t-1}^{MT}, Q_{n,t-1}^{MT}, P_{n,t}^{load}, Q_{n,t}^{load}, P_{n,t}^{k}, Q_{n,t}^{k} \right\} \tag{17}$$

where n denotes the bus at which the agent is located; $P_{n,t-1}^{MT}$ and $Q_{n,t-1}^{MT}$ are the active and reactive power outputs of the MT agent at time step $t-1$; $P_{n,t}^{load}$ and $Q_{n,t}^{load}$ denote the active and reactive loads at the agent's own bus; k represents the set of buses within the agent's observation region; $P_{n,t}^{k}$ and $Q_{n,t}^{\mathrm{k}}$ denote the active and reactive loads of the buses observed by the MT agent.

(2) Action Space

Each agent's action consists of the active and reactive power outputs of its corresponding resource, and all agent actions are continuous. Taking the MT agent as an example, its action space at time step t can be expressed as:

$$S_t^{MT} = \left\{ t, P_{n,t-1}^{MT}, Q_{n,t-1}^{MT}, P_{n,t}^{load}, Q_{n,t}^{load}, P_{n,t}^{k}, Q_{n,t}^{k} \right\} \tag{18}$$

where $P_{n,t}^{MT}$ and $Q_{n,t}^{MT}$ denote the active and reactive power outputs of the MT agent at time step t, respectively.

(3) Reward Function

In this interactive dispatch model, each agent's reward, R, is composed of three parts: the global network loss of the distribution system, the global voltage violation cost, and the voltage violation cost within the agent's observation region. This design encourages agents to place greater emphasis on voltage security and stability constraints during policy updates. Taking the MT agent as an example, its reward function over the entire time horizon is given by:

$$R_n^{MT} = C_{NL} + C_{VOL} + C_{VOL}^k \tag{19}$$

$$C_{VOL}^k = C_{vol_res}^k + C_{vol_flu}^k \tag{20}$$

$$C_{vol_res}^k = c_{vol_res} \sum_{t=1}^{T} \sum_{n \in k}^{k} \left(\frac{\Delta U_{n,t}}{U_n^{\max} - U_n^{\min}} \right) \tag{21}$$

$$C_{vol_flu}^k = c_{vol_flu} \sum_{n \in k}^{k} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (U_{n,t} - \overline{U_n})^2} \tag{22}$$

where $C_{VOL}^{MT,k}$ denotes the voltage over-limit penalty for buses within the agent's observation region, and $C_{vol_res}^k$ and $C_{vol_flu}^k$ are the corresponding penalty costs for voltage limit violations and voltage fluctuations, respectively. To compute the agent's reward, the Newton–Raphson method is used for power flow calculation. The distribution network is operated to minimize network losses while satisfying power flow constraints and maintaining voltage security and stability. Accordingly, each agent's objective is to maximize R_n.

4.3. Hybrid Prioritized Experience Sampling Strategy

In MADDPG, both the policy and value networks are updated using off-policy data sampled from a replay buffer. In conventional off-policy algorithms, experiences collected over multiple iterations are stored in the replay buffer; once the buffer size meets the batch requirement, a mini-batch is randomly sampled to train the networks. This paper proposes a hybrid prioritized experience sampling method to improve experience utilization. Specifically, the replay buffer is stratified for random sampling; a portion of the most recent experiences is preferentially selected; and the randomly sampled data are then fused with the newest samples. The implementation process is shown in Figure A2.

By combining the immediacy of recent samples with the diversity of randomly selected historical samples, the hybrid sampling strategy aims to enhance learning efficiency and the model's generalization capability. Stratified random sampling ensures uniform sampling across different time periods, preventing over- or under-sampling of certain intervals and guaranteeing that the training set contains diverse conditions. This helps the model learn broader environmental characteristics and dynamics, thereby avoiding overfitting to the most recent experiences. Introducing a controlled proportion of the latest samples ensures that the model can promptly learn and adapt to recent environmental changes or policy updates. Thus, the proposed sampling method strikes a balance between stability (learning from historical data) and adaptability (rapidly adjusting via recent data).

4.4. Information-Fusion MADDPG Algorithm

Building on MADDPG, this study develops a multi-agent algorithm that integrates task-specific information. The algorithmic structure is shown in Figure 2, which illustrates the update process of each agent in MADDPG. For every agent, the improved MADDPG employs a neural architecture consisting of a policy network, a target policy network, a value network, and a target value network. The policy network receives the agent's state as input and outputs the agent policy $\mu_j(s_j^{(k)}, \theta_j)$, thereby approximating the agent's policy

π_j. In the conventional MADDPG framework, during agent–environment interactions, the value network estimates the expected return based on the states and actions of all agents. In the proposed improvement, structured prior knowledge t_j (the agent type) is injected into the value network's input. In addition, gradually decaying Gaussian noise (ϵ_s, ϵ_a) is added to the agents' states and actions to enhance the robustness of the value network inputs, thereby improving the stability of model training.

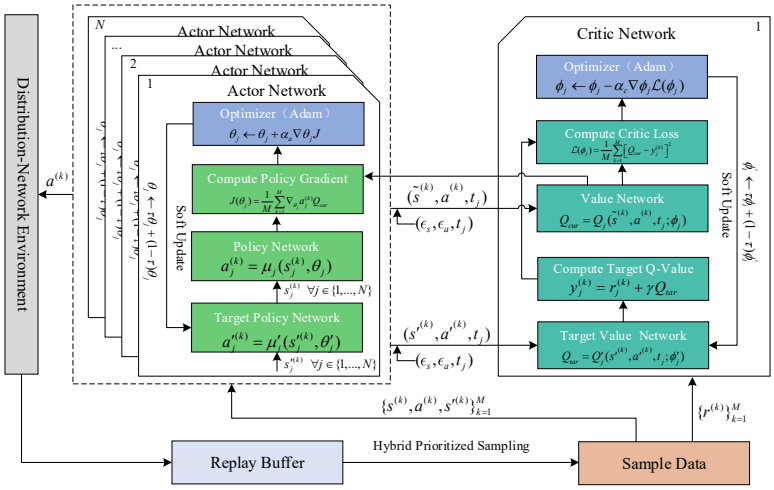


Figure 2. MADDPG algorithm based on the combination of specific information.

The agent interacts with the environment by outputting an action $a^{(k)}$ and obtains an experience sample, which is stored in the replay buffer. When the replay buffer satisfies the mini-batch size requirement M, M groups of experience samples $\left\{s^{(k)}, a^{(k)}, r^{(k)}, s'^{(k)}\right\}_{k=1}^{M}$ are drawn from the buffer at each training iteration to update the model. Here, $s^{(k)}$ denotes the joint state of all agents; $s'^{(k)}$ is the next state of the agents after interacting with the distribution network environment; $r^{(k)}$ is the action output by the agents; $a^{(k)}$ is the reward obtained from the interaction.

The training process for each agent consists mainly of two steps: updating the value network and updating the policy (actor) network. First, the target policy networks and policy networks of all agents take as inputs the next-state samples $s'^{(k)}$ and the current-state samples $s^{(k)}$, respectively, producing the next-state actions $a'^{(k)}$ and the current-state actions $a^{(k)}$. The computation for each agent is given by the following expressions.

$$a'^{(k)}_j = \mu'_j(s'^{(k)}_j, \theta'_j) \tag{23}$$

$$a^{(k)}_j = \mu_j(s^{(k)}_j, \theta_j) \tag{24}$$

$$a^{(k)} = \left\{a^{(k)}_1, \ldots, a^{(k)}_N\right\} \tag{25}$$

$$a'^{(k)} = \left\{a'^{(k)}_1, \ldots, a'^{(k)}_N\right\} \tag{26}$$

where $s^{(k)}_j$ and $s'^{(k)}_j$ are the states of agent j; θ_j and θ'_j are the parameters of agent j's policy network and target policy network, respectively.

In the conventional MADDPG algorithm, the value network takes as input the joint states and actions of all agents $(s^{(k)}, a^{(k)})$, while the target value network uses the joint next states and next actions $(s'^{(k)}, a'^{(k)})$ to obtain the current and future value estimates. In the improved MADDPG proposed here, structured prior knowledge t_j and Gaussian noise

(ϵ_s, ϵ_a) are injected into the inputs of the value network. The corresponding computation is given by:

$$\widetilde{s}^{(k)} = s^{(k)} + \epsilon_s,\ \epsilon_s \sim \mathcal{N}(0, \sigma_s^2 I) \tag{27}$$

$$\widetilde{a}^{(k)} = a^{(k)} + \epsilon_a,\ \epsilon_a \sim \mathcal{N}(0, \sigma_a^2 I) \tag{28}$$

$$\widetilde{s'}^{(k)} = s'^{(k)} + \epsilon_s,\ \epsilon_s \sim \mathcal{N}(0, \sigma_s^2 I) \tag{29}$$

$$\widetilde{a'}^{(k)} = a'^{(k)} + \epsilon_a,\ \epsilon_a \sim \mathcal{N}(0, \sigma_a^2 I) \tag{30}$$

where $\widetilde{a}^{(k)}$ and $\widetilde{a'}^{(k)}$ denote the current and next action sets after adding decaying noise, respectively, and $\widetilde{s}^{(k)}$ and $\widetilde{s'}^{(k)}$ denote the current and next state sets after adding decaying noise, respectively.

The combinations of $(\widetilde{s}^{(k)}, \widetilde{a}^{(k)})$ and $(\widetilde{s'}^{(k)}, \widetilde{a'}^{(k)})$ with the structured prior knowledge GGG are then used as the inputs to the value network and the target value network, respectively, to compute the current value estimate and the target value estimate. The corresponding equations are as follows:

$$Q_{cur} = Q_j(\widetilde{s}^{(k)}, \widetilde{a}^{(k)}, t_j; \phi_j) \tag{31}$$

$$Q_{tar} = {Q'}_j(\widetilde{s'}^{(k)}, \widetilde{a'}^{(k)}, t_j; {\phi'}_j) \tag{32}$$

$$y_j^{(k)} = r_j^{(k)} + \gamma Q_{tar} \tag{33}$$

where Q_{cur} and Q_{tar} are the current value estimate and the target value estimate, respectively; ϕ_j and ${\phi'}_j$ are the parameters of agent j's value network and target value network; $y_j^{(k)}$ is the computed target return; and γ is the discount factor.

Using the computed target value $y_j^{(k)}$ and the current value estimate Q_{cur}, the loss function of the value network is obtained. Based on the value estimate Q_{cur} from the value network and the action $a_j^{(k)}$ produced by the policy network, the loss function of the policy network is then calculated. The corresponding equations are as follows:

$$\mathcal{L}(\phi_j) = \frac{1}{M}\sum_{k=1}^{M}\left[Q_{cur} - y_j^{(k)}\right]^2 + \lambda//\epsilon//^2 \tag{34}$$

$$J(\theta_j) = \frac{1}{M}\sum_{k=1}^{M}\nabla_{\theta_j} a_j^{(k)} Q_{cur} \tag{35}$$

where ϵ is the noise term, ϵ_s and ϵ_a are noise vectors, and $//*//$ denotes the squared norm operation of a vector.

During backpropagation, gradient information of the loss functions is propagated layer by layer. The policy network is updated via gradient ascent, using an optimizer to adjust its parameters θ_j. The value network is updated via gradient descent, with an optimizer adjusting its parameters ϕ_j, thereby improving the accuracy of value estimation. The corresponding equations are as follows:

$$J(\theta_j) = \frac{1}{M}\sum_{k=1}^{M}\nabla_{\theta_j} a_j^{(k)} Q_{cur} \tag{36}$$

$$J(\theta_j) = \frac{1}{M}\sum_{k=1}^{M}\nabla_{\theta_j} a_j^{(k)} Q_{cur} \tag{37}$$

where ∇ is the computed model gradient, α_a and α_c are the updated parameters of the policy (actor) network and the value network, respectively.

After all agents' policy and value networks have been updated, the target policy and target value networks for each agent are soft-updated to enhance the accuracy of their evaluations and improve their training stability. The update equations are given as follows:

$$J(\theta_j) = \frac{1}{M}\sum_{k=1}^{M} \nabla_{\theta_j} a_j^{(k)} Q_{cur} \tag{38}$$

$$J(\theta_j) = \frac{1}{M}\sum_{k=1}^{M} \nabla_{\theta_j} a_j^{(k)} Q_{cur} \tag{39}$$

where τ is the soft update coefficient.

The overall procedure of the improved MADDPG algorithm is illustrated in Figure 3.

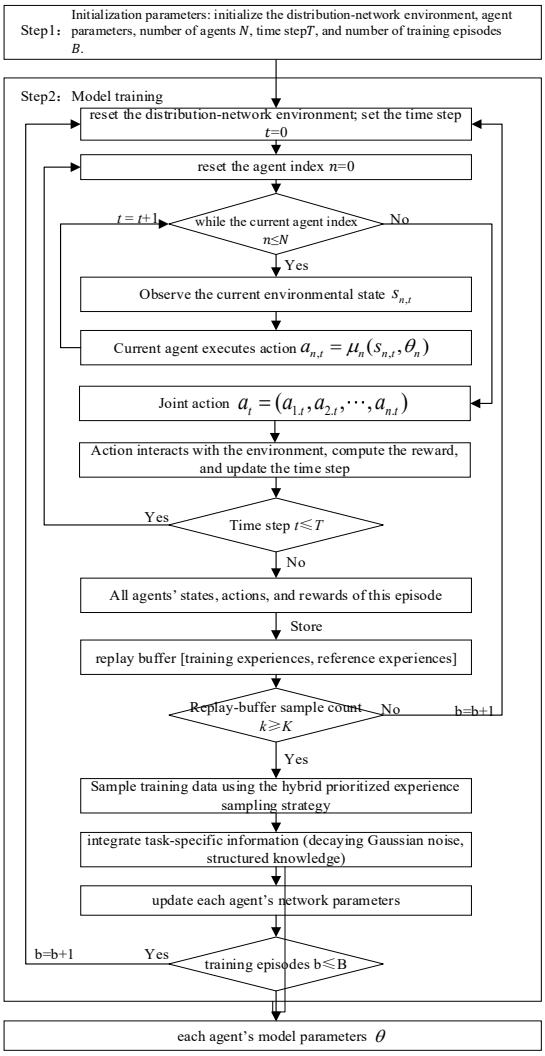


Figure 3. Improved MADDPG algorithm training flowchart.

5. Case Study

5.1. Data Basis and Algorithm Parameters

To verify the effectiveness of the proposed improved MADDPG algorithm for the interactive dispatch of distribution network resources, a simulation environment of the IEEE 33-bus distribution system was established. A multi-agent environment based on the

improved MADDPG algorithm was implemented using the PyTorch (version 1.12.0) deep learning framework. The system topology is shown in Figure A3. The base voltage is 10 kV, and the safe per-unit voltage range is set to [0.94,1.06] p.u. To emulate the dispatch process, the optimization horizon is set to 24 h with a time step of 1 h.

To simulate extreme scenario data in the distribution network, this study follows the method proposed in Ref. [25] to generate seasonal extreme scenario data. Using a scenario dissimilarity maximization principle and time series clustering, an extreme scenario dataset is constructed. Bus loads are allocated according to the standard IEEE 33-bus case. Taking the summer scenario as an example, seven extreme scenarios are obtained via clustering, and their cluster center data are shown in Figure A4. The dissimilarity of the constructed scenario dataset is visualized in ridge plots, as illustrated in Appendix A, Figures A5–A11.

Within the distribution network, microturbines, distributed photovoltaics, distributed wind power, and static var compensators are configured to achieve active–reactive coordinated dispatch for voltage stability and security. The system includes two MT units, three DPV units, three DWP units, and three SVC units. The specific grid-connection locations and capacity configurations of these resources are listed in Table 1, and the corresponding unit cost coefficients for resource regulation are given in Table 2.

Table 1. Resource distribution location and capacity configuration parameters.

Controllable Resource Type	Location	Capacity Configuration
MT	Bus 13	200 MW
	Bus 21	200 MW
DPV	Bus 15	100 MW
	Bus 19	100 MW
	Bus 29	100 MW
DWP	Bus 5	100 MW
	Bus 17	100 MW
	Bus 26	100 MW
SVC	Bus 8	100 Mvar
	Bus 18	100 Mvar
	Bus 19	100 Mvar

Table 2. Distribution network unit resource-related parameters.

Parameter Type	Value
c_{loss}	44
c_{vol_res}	1000
c_{vol_flu}	500
c_{vol_rew}	5

Based on the constructed 7-day extreme scenario dataset, six days were randomly selected as the training set (after appropriate preprocessing), and the remaining one day was used as the test set. The improved MADDPG algorithm proposed in this paper was implemented in Python 3.9 and executed on a workstation equipped with an AMD Ryzen 7 5800H CPU, an NVIDIA GeForce RTX 3060 Laptop GPU (6.0 GB), and 16 GB of RAM, using PyTorch 2.1.1. The relevant hyperparameters of the improved MADDPG are listed in Table 3.

Table 3. Algorithm-related parameters.

Parameter Type	Actor Network	Critic Network
Learning Rate	0.0001	0.0001
Number of Network Layers	3	3
Neurons per Layer	256	256
Soft Update Coefficient for Target Network	0.15	0.15
Discount Factor	0.95	0.95
Number of Agents	11	
Batch Size	1024	
Training Steps	144	
Initial Sampling Ratio	0.99	
Sampling Ratio Decay Rate	0.5	

5.2. Model Training Comparison

To validate the effectiveness of the proposed method, five algorithms are designed to solve the power grid dispatch problem for comparative analysis. Drawing on relevant literature and tailoring them to the task at hand, the comparison set includes the MASAC algorithm [16], the GRPO algorithm [26], the DDPG algorithm [15], the MADDPG algorithm, and the improved MADDPG algorithm proposed in this paper. The structural diagrams of these algorithms are shown in Appendix A, Figures A12–A15.

Using the summer cluster center extreme scenario set as an example, the variations in network loss cost and global voltage penalty during training for the above RL algorithms are depicted in Figures 4 and 5.

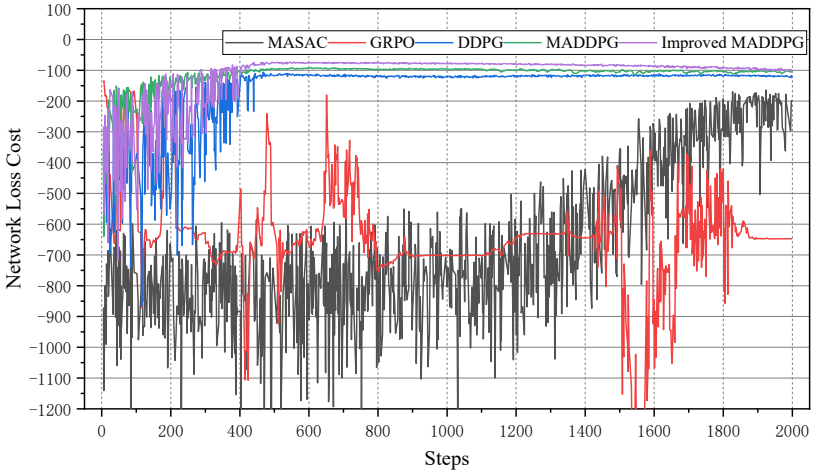


Figure 4. Comparison of network loss cost training results of different algorithms.

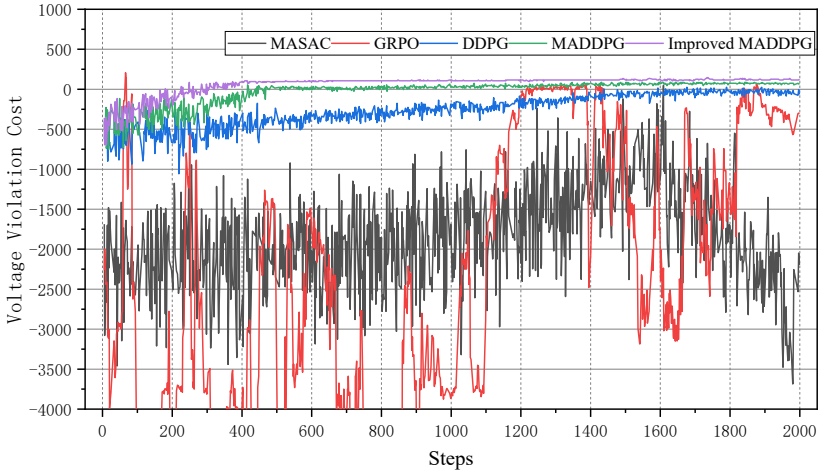


Figure 5. Comparison of voltage penalty cost training results of different algorithms.

As shown in Figures 4 and 5, the proposed improved MADDPG algorithm exhibits considerable fluctuations in network loss cost and voltage penalty cost before 400 steps, primarily because the model lacks sufficient data to learn and adapt during the early training phase. As training proceeds and more experience is accumulated, the model's performance gradually stabilizes and improves. Around step 450, the improved MADDPG begins to converge, with the global network loss cost stabilizing at approximately −95 and the global voltage penalty cost stabilizing at about 125. By contrast, the conventional MADDPG converges around step 500; the DDPG algorithm's network loss cost converges near 500, and its voltage violation cost converges around 1600; MASAC converges at roughly 1900 steps; and GRPO first converges after 1200 steps but then exhibits large oscillations before gradually converging again around 1900 steps. The voltage penalty costs of conventional MADDPG, DDPG, GRPO, and MASAC stabilize at roughly −65, −260, and −200, respectively, while their final global network loss costs settle near −105, −125, −650, and −180. The improved MADDPG accelerates convergence by at least 6.25% (compared with conventional MADDPG and DDPG). It also achieves markedly better performance stability: the voltage penalty cost is the lowest among all methods, with an overall improvement of about 5%; in terms of network loss cost, the overall improvement is about 9.5%. During subsequent training, the improved MADDPG is able to account for both local information of individual agents and global environmental information, thereby yielding superior dispatch decisions that balance the global network loss cost and voltage stability of the distribution system.

The evolution of each agent's voltage penalty during training is shown in Appendix A, Figures A16–A19.

5.3. Analysis of Model Testing Results

Figure 6 depicts the active-power outputs of all distributed energy resources (DERs) in the test distribution network, showing bar charts for two MT units, three DPV units, and three DWP units at each time step. As can be observed, the wind–solar units supply almost no active power throughout the day, contributing only a small amount at 12:00. Nearly the entire active-power demand is therefore met by the two MT units, with the MT located at bus 21 shouldering most of the supply. Peak MT outputs occur mainly between 11:00–13:00 and 16:00–18:00.

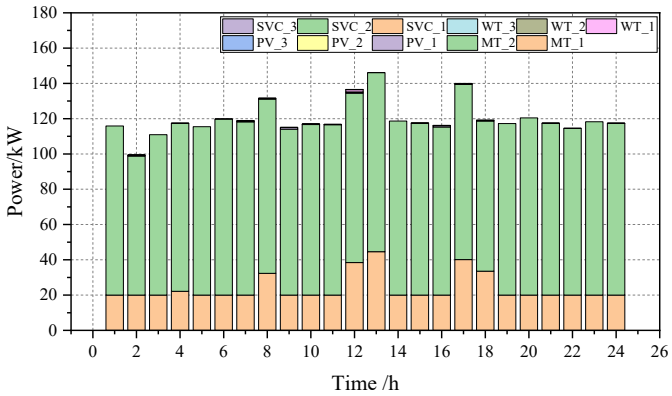


Figure 6. Active power output of each unit in distribution network.

To verify the effectiveness of the improved MADDPG-based optimal dispatch method proposed in this chapter, the trained MADDPG model was employed for testing. Test data were selected from the summer cluster center extreme scenario set, and the active and reactive power outputs of each unit for that day were obtained, as shown in Figures 7 and 8.

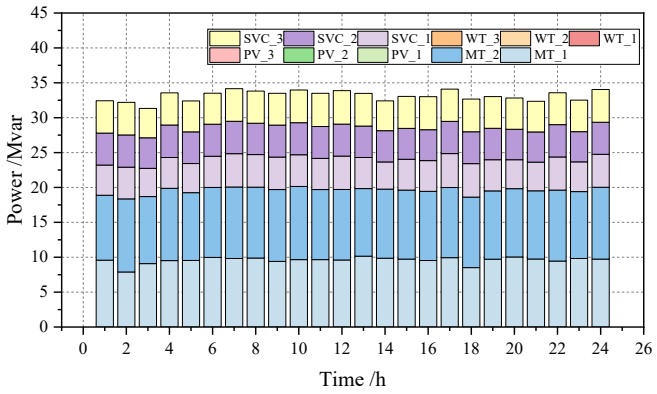


Figure 7. Reactive power output of each unit in distribution network.

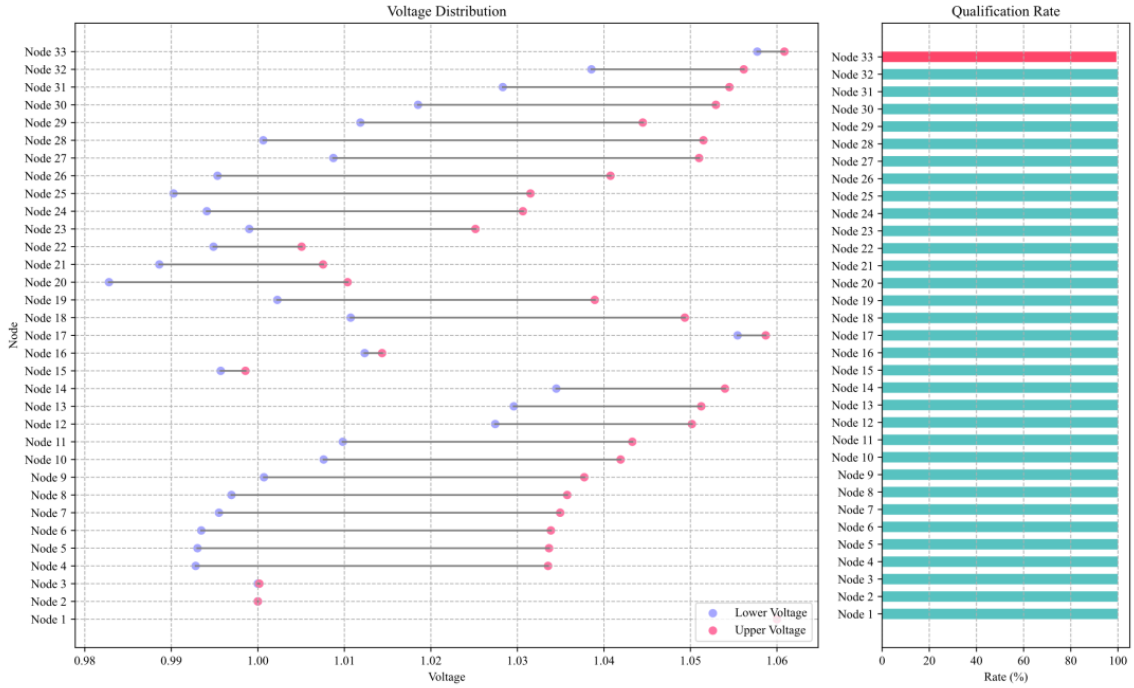


Figure 8. The voltage distribution results of the improved MADDPG algorithm under the test set.

Figure 7 illustrates the reactive-power outputs of the DERs. On the test day, all reactive-power dispatch is provided jointly by the MT units and the SVC units, each accounting for roughly 50% of the total. This is because the operating cost of SVCs is generally lower than the generation loss incurred by curtailing wind–solar units. Hence, SVCs are used preferentially for preventive control, while reserving the regulation capability of wind–solar units as a backup for voltage contingencies. This strategy maintains normal system operation and voltage stability while enhancing overall economic efficiency, fully complying with the requirements of the safety plan.

To further confirm the robustness of the proposed improved MADDPG model for voltage-security-oriented optimal dispatch, test data from the extreme scenario sets of all four seasons were fed into the trained model. The evaluation metrics include the following: average network loss, average voltage deviation, and voltage compliance rate.

The average network loss is defined as the mean value, over all time intervals of a day, of the power-loss cost across all branches.

$$C_{avg} = \frac{1}{T} c_{\text{loss}} \sum_{x=1}^{N_{branch}} \sum_{t=1}^{T} P_{x,t} \tag{40}$$

The average voltage deviation is the mean, over all time intervals in a day, of each bus's voltage deviation from the reference voltage.

$$U_{RMS} = \sum_{t=1}^{T} \sum_{n=1}^{N_{node}} \left| U_{n,t} - U_{ref} \right| \tag{41}$$

where U_{ref} is the reference voltage.

The voltage compliance rate is the proportion of time intervals in a day during which bus voltages remain within the prescribed safety limits.

$$U_{PR} = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_{node}} 1\left(\forall i, U_{i,t} \in [U_n^{\min}, U_n^{\max}]\right)}{T \times N_{node}} \tag{42}$$

In addition to the five RL-based algorithms discussed above, an optimization-based approach was also employed. The dispatch model was formulated and solved using a mathematical solver. The test results for the five RL algorithms and the optimization algorithm are summarized in Table 4, while the voltage distribution outcomes on the test set are shown in Figures 8 and A20–A23.

Table 4. Performance test results of different algorithms.

Algorithm	Average Network Loss	Average Voltage Deviation/p.u	Voltage Compliance Rate/%
Optimization Algorithm	0.6945	0.0266	100%
MASAC	1.2855	0.0367	89.27%
GRPO	1.3487	0.0387	81.31%
DDPG	0.7517	0.0270	99.24%
MADDPG	0.9085	0.0276	98.36%
Proposed	0.6813	0.0268	99.98%

As indicated in Table 4, the optimization algorithm achieves the best performance in terms of voltage compliance rate and average voltage deviation—100% and 0.0266, respectively. Compared with RL-based optimal dispatch, the optimization method has a marked advantage in result rationality, primarily because the computed solutions strictly satisfy all model constraints, with no limit violations. However, its relatively low computational efficiency and high resource demand make it difficult to scale to large datasets.

The improved MADDPG algorithm proposed in this paper performs well in terms of voltage compliance rate and average voltage deviation, achieving 99.98% and 0.0268, respectively. Compared with the optimization algorithm, its performance is nearly equivalent, while surpassing it in average network loss with a 1.9% improvement. After training on the training set to obtain the optimal model, the average execution time on the test set is about 5 s. Although the training time of this intelligent algorithm is comparable to the computation time of the traditional optimization method, its stronger adaptability makes it better suited to the task of safety-oriented optimal dispatch planning. Relative to the other optimization algorithms, the proposed improved MADDPG achieves at least 1.62%, 2.9%, and 9.3% improvements in voltage compliance rate, average voltage deviation, and average network loss, respectively.

From the analysis of Figures 8 and A20–A23, although RL-based 4 voltage dispatch methods exhibit occasional voltage violations at certain buses, the overall voltage distribution is more compact, with smaller fluctuations. The improved MADDPG proposed in this paper demonstrates superior performance: only bus 33 experiences a voltage violation, and its voltage compliance rate still reaches 99.34%. The case study results indicate that, compared with traditional optimization methods, RL-based approaches can provide

more stable voltages at specific buses—particularly buses 2, 15, 16, and 17—whereas the traditional method yields a more uniform voltage distribution across all buses.

6. Conclusions

To address the optimal dispatch under extreme power system scenarios, this paper proposes an improved MADDPG-based dispatch model that provides a secure scheduling scheme for such conditions. Simulation analyses verify the effectiveness of the proposed model and method, leading to the following conclusions:

1. Guided by a scenario dissimilarity maximization principle, extreme scenarios with the greatest dissimilarity within each cluster are selected and combined with the cluster center data to construct an extreme scenario dataset suitable for optimal dispatch. The datasets thus obtained maintain similar overall trends while exhibiting pronounced differences, thereby supplying maximized incremental information for subsequent optimization tasks.
2. The proposed improved MADDPG algorithm surpasses conventional MADDPG, GRPO, and other counterparts in training stability and efficiency. Compared with benchmark algorithms, it converges at least 50 steps earlier and achieves an overall performance gain of approximately 5%, thereby minimizing training resource consumption and demonstrating better suitability for large-scale optimization dispatch problems.
3. In terms of voltage compliance rate, average voltage deviation, and average network loss, the improved MADDPG outperforms traditional MADDPG, GRPO, and other RL algorithms; it also surpasses the traditional optimization method on average network loss and approaches it on average voltage deviation and network loss. Moreover, the combined time for model training and decision making is far lower than that of the optimization approach, and the voltage distribution obtained by the improved MADDPG is more compact. These results indicate that the improved MADDPG possesses significant advantages for voltage-security-oriented optimal dispatch, combining the adaptability of reinforcement learning with the accuracy of optimization methods.

Author Contributions: Conceptualization, X.D., C.L. and Q.Z.; methodology, C.L.; software, X.D., C.L. and Q.Z.; validation, P.N., D.S. and Z.D.; formal analysis, X.D. and C.L.; investigation, Z.D., D.S. and P.N.; data curation, X.D. and C.L.; writing—original draft preparation, X.D. and C.L.; writing—review and editing, P.N., D.S. and Q.Z.; visualization, X.D., C.L. and Q.Z.; project administration, X.D. and Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the Science and Technology Projects from the State Grid Corporation of China (Research and application on multi-temporal adjustable load of normal marketization participating in grid interactive technology, No.: 5400-202317586A-3-2-ZN).

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to the State Grid Corporation of China.

Conflicts of Interest: Authors Zhenlan Dou was employed by the State Grid Shanghai Municipal Electric Power Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A

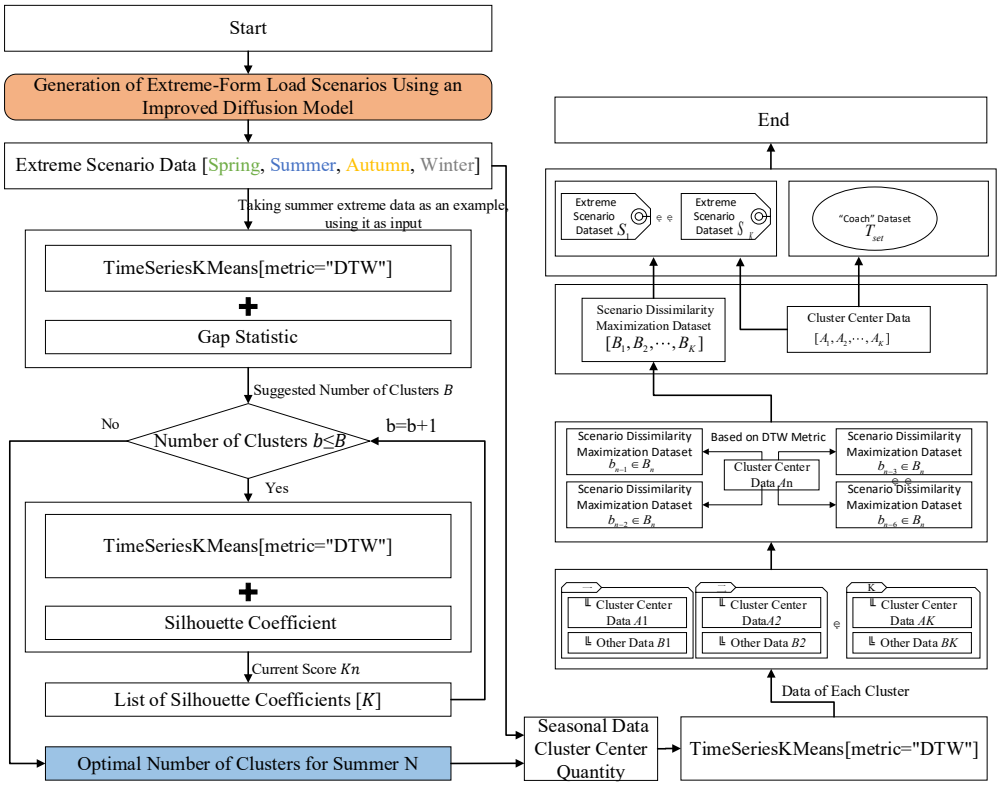


Figure A1. Extreme scene dataset construction flowchart.

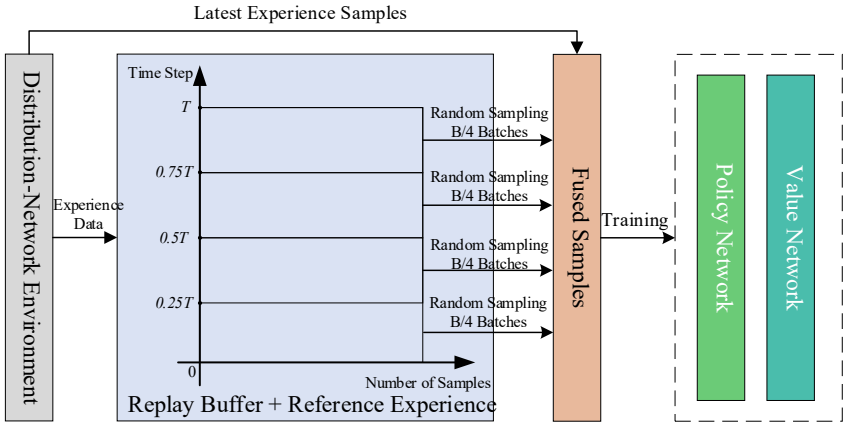


Figure A2. Mixed-priority empirical sampling steps.

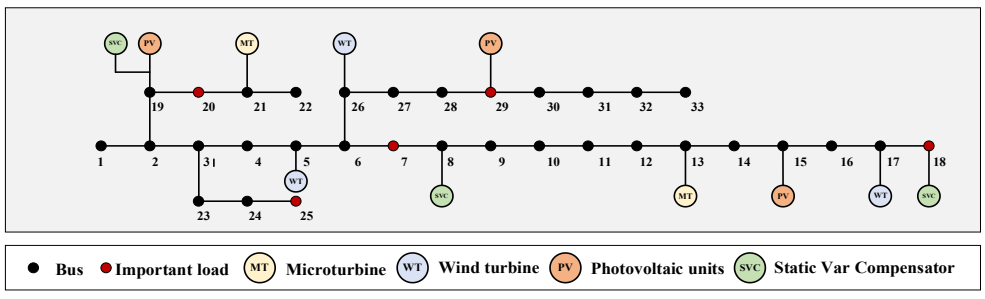


Figure A3. IEEE 33-bus distribution network system topology diagram.

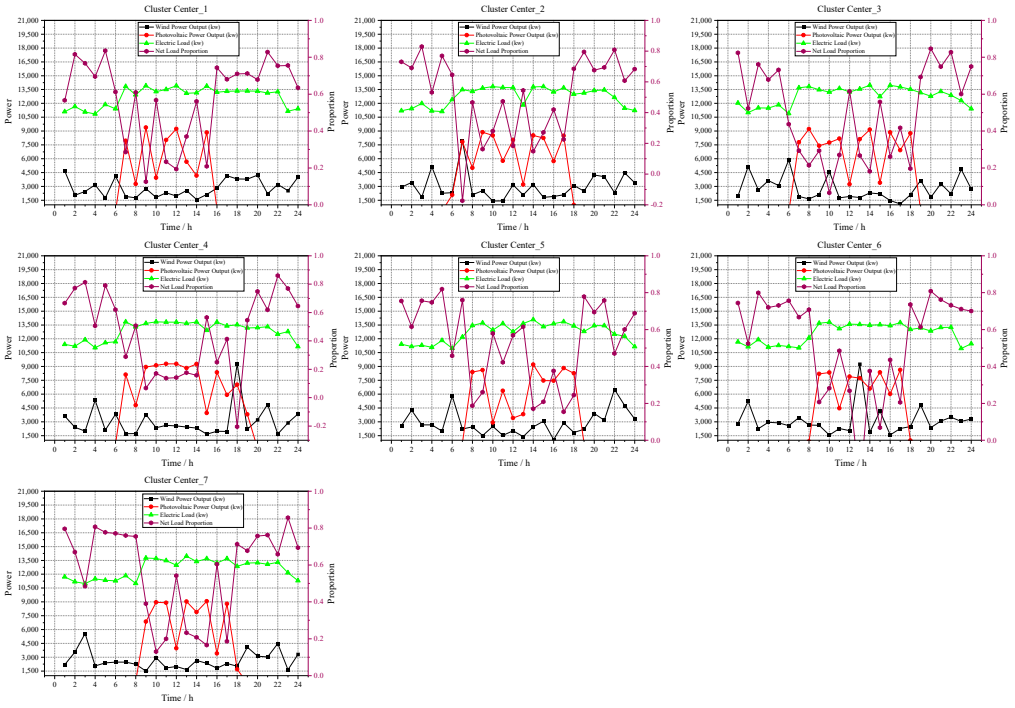


Figure A4. Summer clustering center data results.

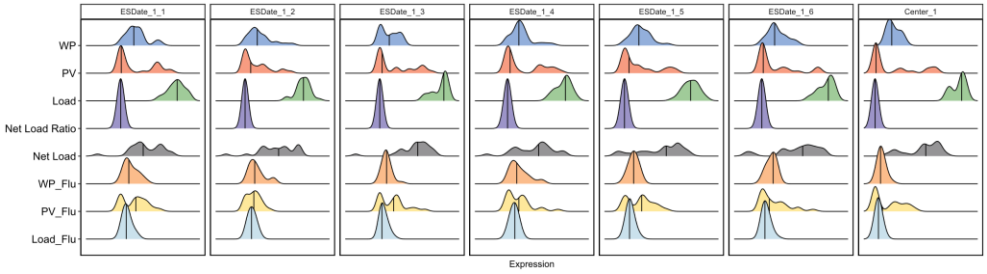


Figure A5. Summer clustering center 1 extreme scene set distribution characteristics.

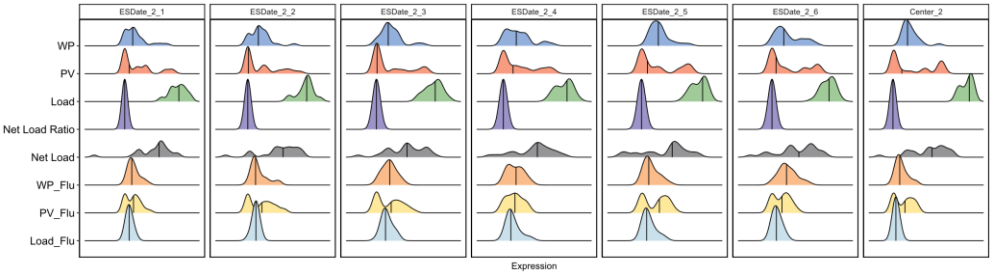


Figure A6. Summer clustering center 2 extreme scene set distribution characteristics.

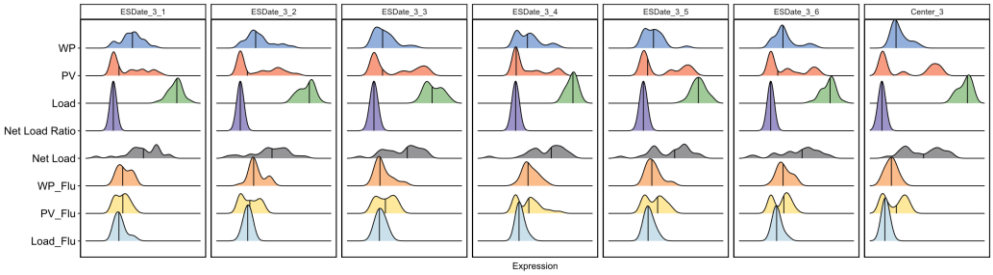


Figure A7. Summer clustering center 3 extreme scene set distribution characteristics.

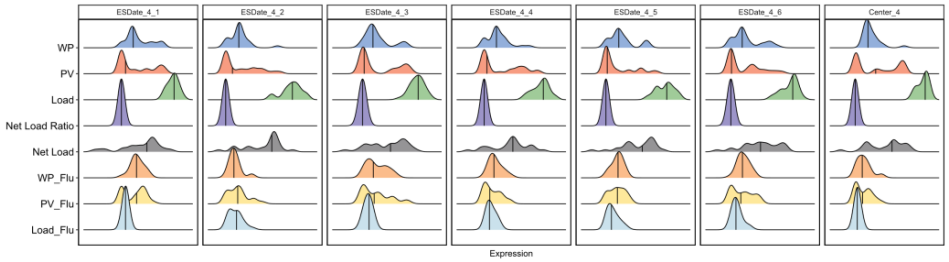


Figure A8. Summer clustering center 4 extreme scene set distribution characteristics.

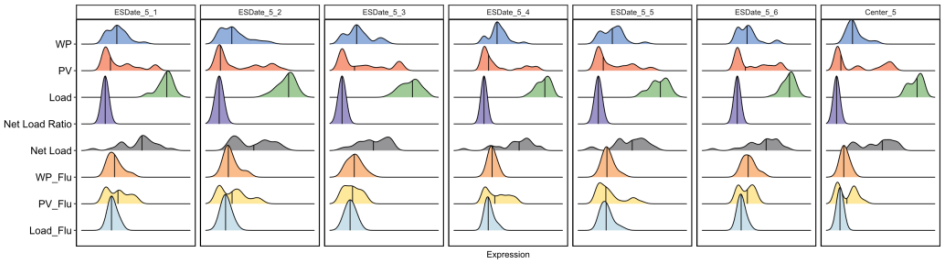


Figure A9. Summer clustering center 5 extreme scene set distribution characteristics.

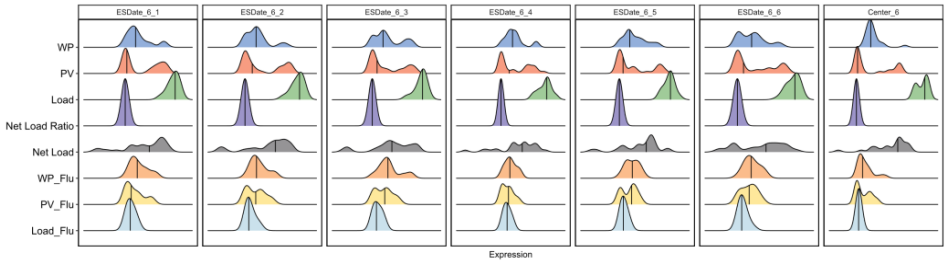


Figure A10. Summer clustering center 6 extreme scene set distribution characteristics.

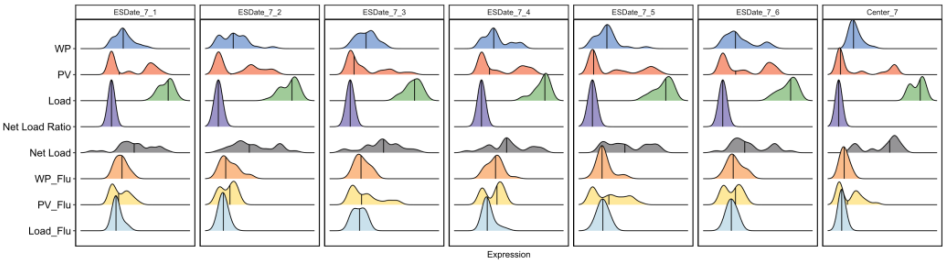


Figure A11. Summer clustering center 7 extreme scene set distribution characteristics.

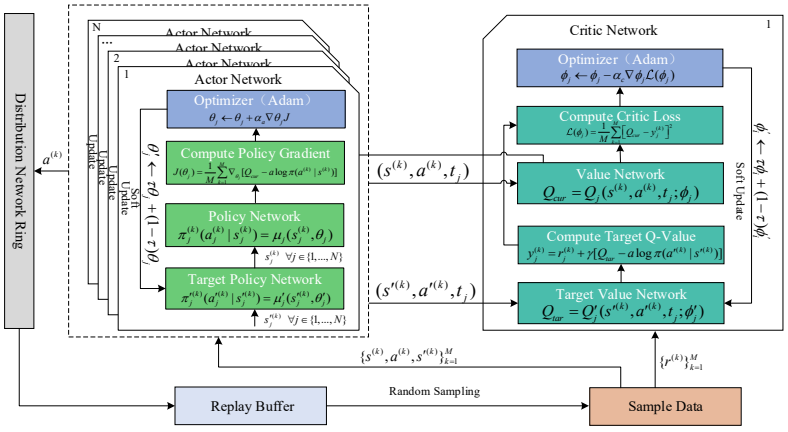


Figure A12. Comparison algorithm—MASAC algorithm structure.

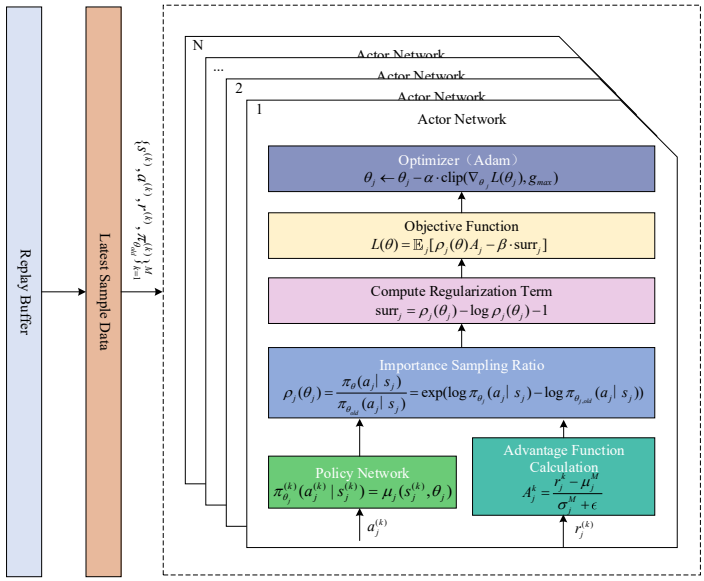


Figure A13. Comparison algorithm—GRPO algorithm structure.

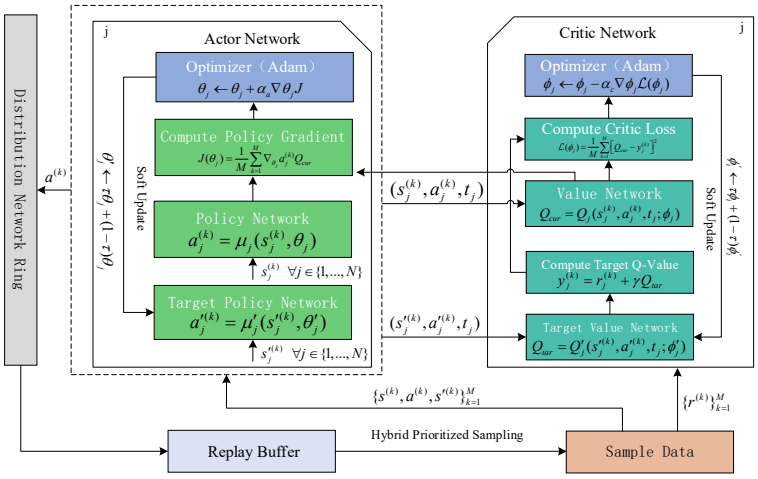


Figure A14. Comparison algorithm—DDPG algorithm structure.

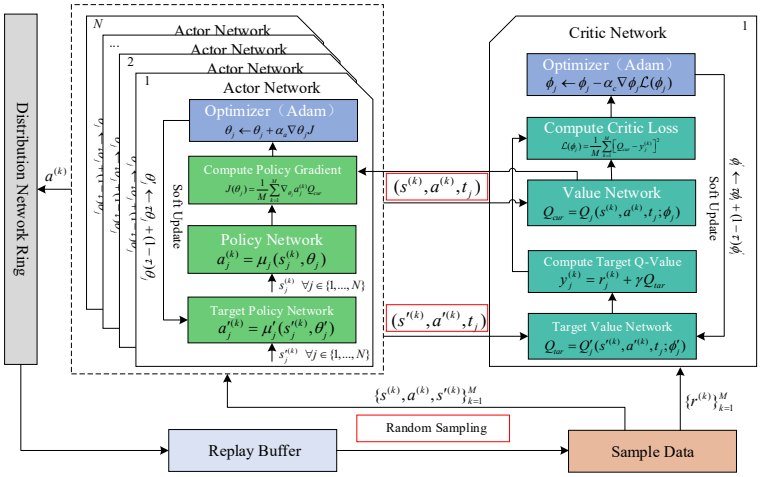


Figure A15. Comparison algorithm—traditional MADDPG algorithm structure.

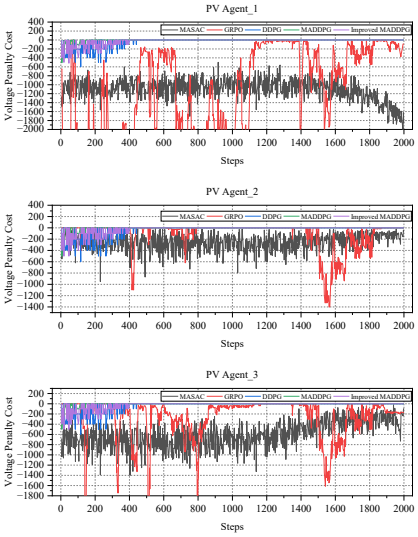


Figure A16. Comparison of PV agent training results in different algorithms.

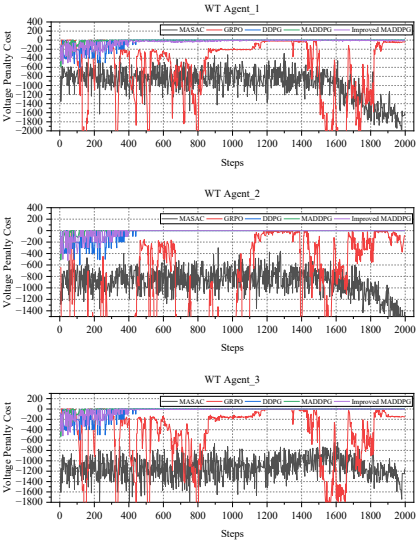


Figure A17. Comparison of WT agent training results in different algorithms.

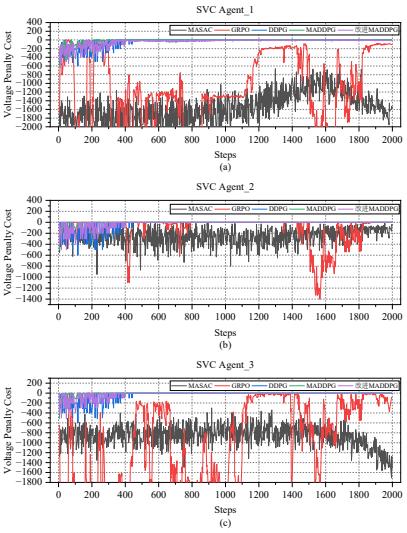


Figure A18. Comparison of SVC agent training results in different algorithms.

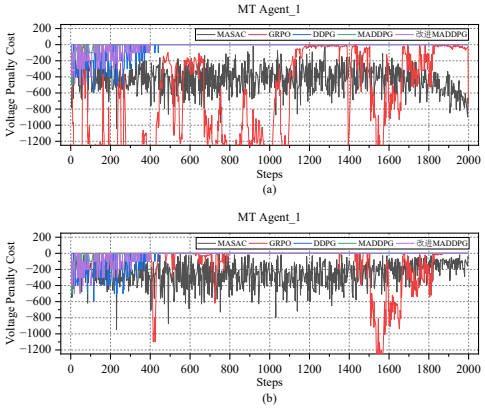


Figure A19. Comparison of MT agent training results in different algorithms.

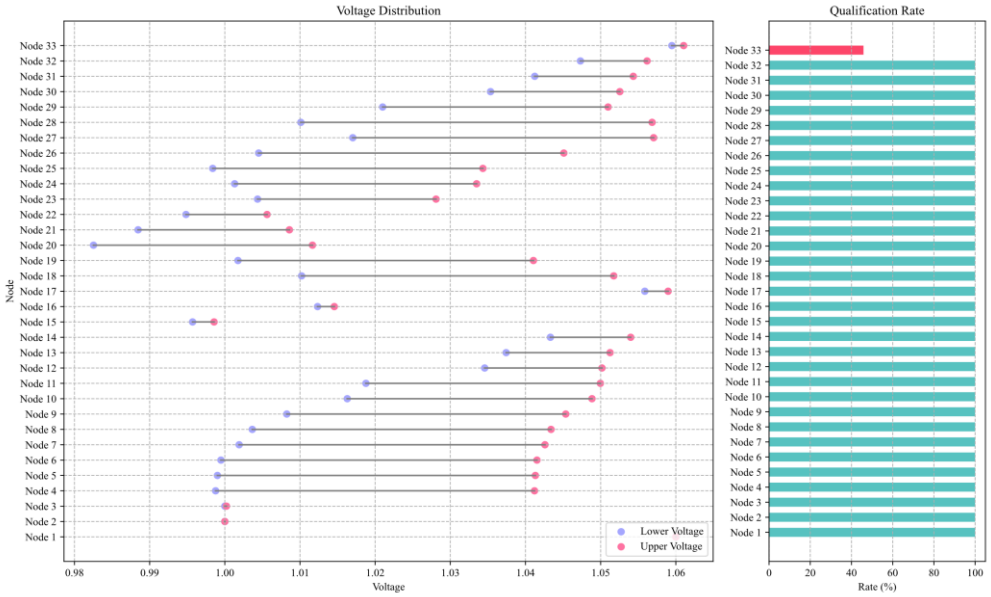


Figure A20. The voltage distribution results of the traditional MADDPG under the test set.

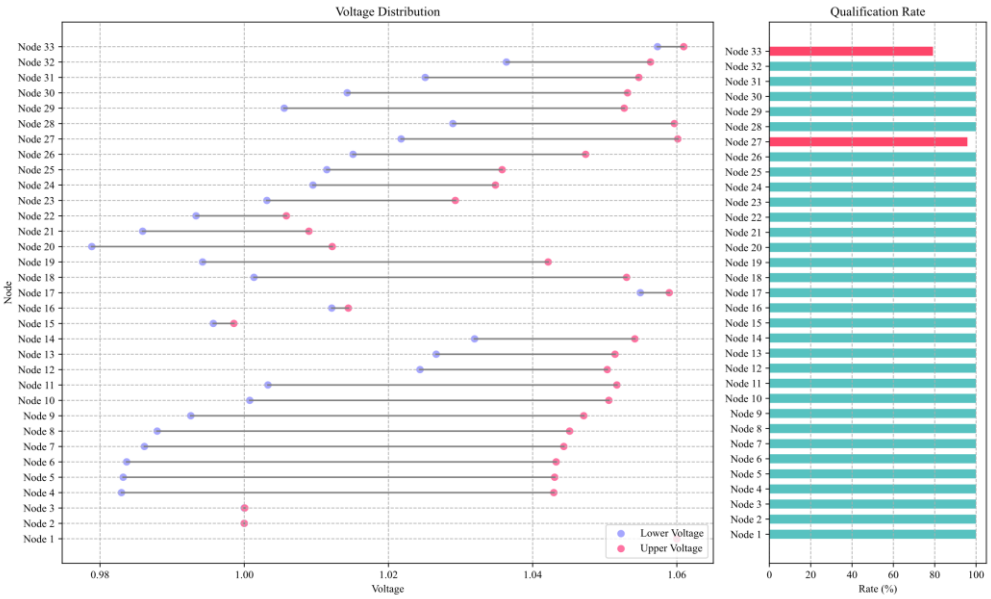


Figure A21. The voltage distribution results of the traditional DDPG under the test set.

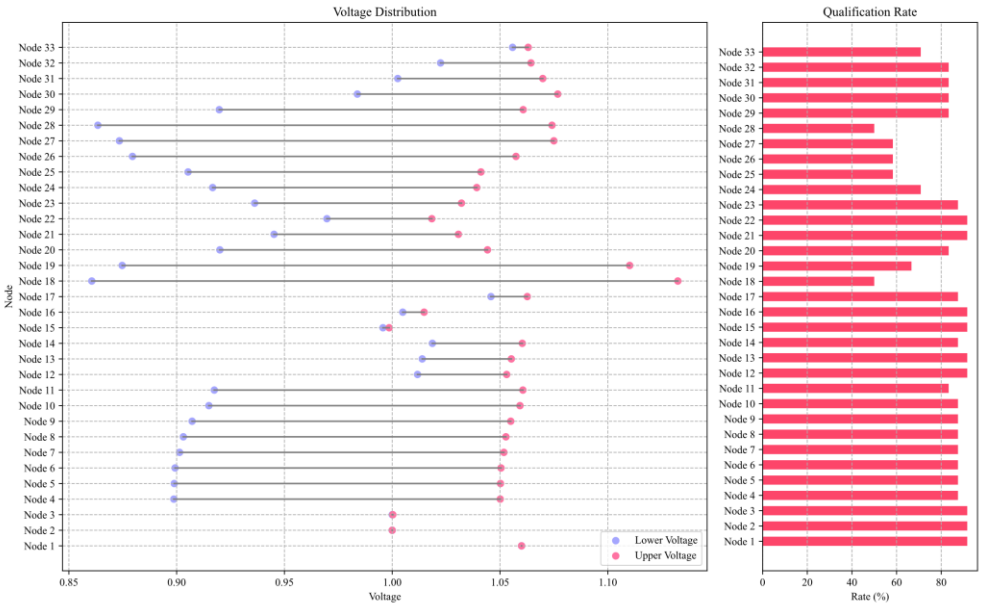


Figure A22. The voltage distribution results of the traditional GRPO under the test set.

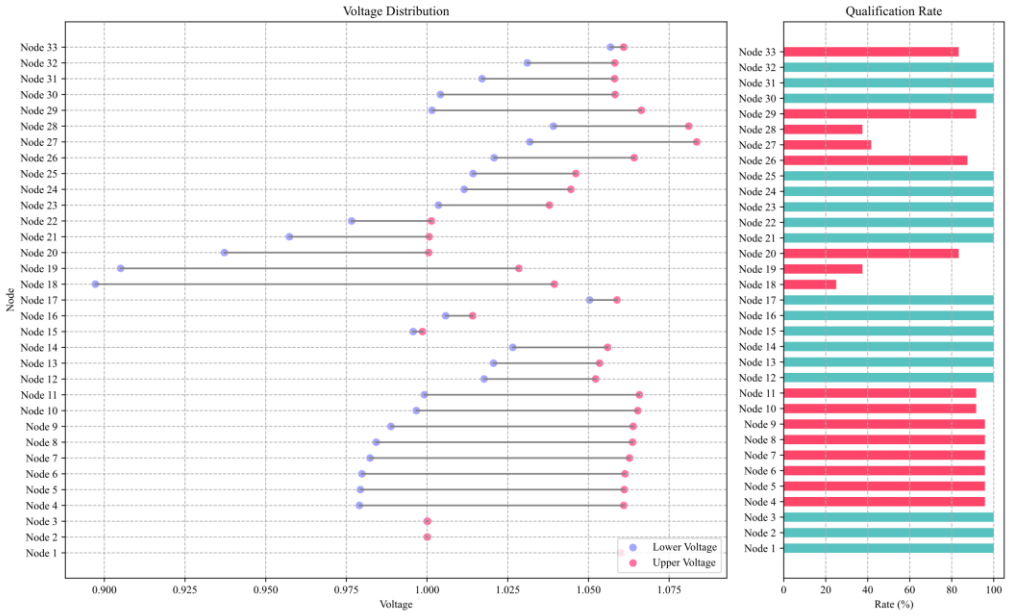


Figure A23. The voltage distribution results of the traditional MASAC under the test set.

References

1. Wu, L.; Chen, C.; Hu, J.; Wang, C.; Tong, Y. User-Side Resource Applications and Key Technologies Supporting the Flexibility Requirements of New-Energy Power Systems. *Power Syst. Technol.* **2024**, *48*, 1435–1450.
2. Bie, C.; Li, G. Risk Assessment and Resilience Enhancement of New-Type Power Systems under Extreme Weather Conditions. *Glob. Energy Interconnect.* **2024**, *7*, 1–2. [CrossRef]
3. Zhou, Y.; Zhao, Y.; Ma, Z. Resilience analysis and improvement strategy of microgrid system considering new energy connection. *PLoS ONE* **2024**, *19*, e0301910.
4. Liu, W.; Liu, J.; Wan, H.; Wang, Y.; Zhang, S.; Feng, W.; Yang, T. Full-Scenario Risk Assessment for New Power System Planning Scheme Facing Multiple Types of Extreme Weather. *Autom. Electr. Power Syst.* **2025**, *49*, 65–78.
5. Cheng, L.; Peng, P.; Huang, P.; Zhang, M.; Meng, X.; Lu, W. Leveraging evolutionary game theory for cleaner production: Strategic insights for sustainable energy markets, electric vehicles, and carbon trading. *J. Clean. Prod.* **2025**, *512*, 145682. [CrossRef]
6. Cheng, L.; Yu, F.; Huang, P.; Liu, G.; Zhang, M.; Sun, R. Game-theoretic evolution in renewable energy systems: Advancing sustainable energy management and decision optimization in decentralized power markets. *Renew. Sustain. Energy Rev.* **2025**, *217*, 115776. [CrossRef]

7. Wei, B.; Qiao, S.; Meng, R.; Li, J. Two-Stage Robust Optimization Dispatch of Microgrids Based on Data-Driven Uncertainty Sets. *High Volt. Eng.* **2025**, *51*, 852–863. [CrossRef]
8. Yang, M.; Liu, Y.; Guo, L.; Zhang, Y.; Wang, Z.; Wang, C. Data-Driven Security Region-Based Multi-Level Distributed Economic Dispatch Method for Transmission-Distribution-Microgrid. *Chinese Society for Electrical Engineering*, 30 March 2025; 1–17.
9. Huang, J.; Luo, Z.; Li, X.; Zhou, R. System Optimization Dispatch Considering Internal Operating Scenarios and Network Losses of Virtual Power Plants. *Power System Technology*, 30 March 2025; 1–8.
10. Aguilar, D.; Quinones, J.J.; Pineda, L.R.; Ostanek, J.; Castillo, L. Optimal scheduling of renewable energy microgrids: A robust multi-objective approach with machine learning-based probabilistic forecasting. *Appl. Energy* **2024**, *369*, 123548. [CrossRef]
11. Morais, H.; Kádár, P.; Faria, P.; Vale, Z.A.; Khodr, H.M. Optimal scheduling of a renewable micro-grid in an isolated load area using mixed-integer linear programming. *Renew. Energy* **2010**, *35*, 151–156. [CrossRef]
12. Li, X.; Ma, X.; Zhao, T. Energy Optimization Dispatch of Photovoltaic and Energy Storage Two-Phase Access Traction Power Supply System Considering Regenerative Braking Energy Uncertainty. *High Voltage Engineering*, 30 March 2025; 1–13.
13. Güven, A.F.; Yörükeren, N.U.R.A.N.; Tag-Eldin, E.; Samy, M.M. Multi-objective optimization of an islanded green energy system utilizing sophisticated hybrid metaheuristic approach. *IEEE Access* **2023**, *11*, 103044–103068. [CrossRef]
14. Xu, Z.; Gong, Y.; Zhou, Y.; Bao, Q.; Qian, W. Enhancing kubernetes automated scheduling with deep learning and reinforcement techniques for large-scale cloud computing optimization. In Proceedings of the Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024), Changchun, China, 16 October 2024; Volume 13291, pp. 1595–1600.
15. Cheng, L.; Wei, X.; Li, M.; Tan, C.; Yin, M.; Shen, T.; Zou, T. Integrating Evolutionary Game-Theoretical Methods and Deep Reinforcement Learning for Adaptive Strategy Optimization in User-Side Electricity Markets: A Comprehensive Review. *Mathematics* **2024**, *12*, 3241. [CrossRef]
16. Zhu, Z.; Zhang, X.; Chen, H. Voltage Control Method for Distribution Networks with Intelligent Soft Switches Based on Deep Reinforcement Learning. *High Volt. Eng.* **2024**, *50*, 1214–1224.
17. Feng, C.; Tang, F.; Wang, G.; Wen, F.; Zhang, Y. Voltage Control of Distribution Networks Based on Fusion Experience Safety Reinforcement Learning. *Automation of Electric Power Systems*, 30 March 2025; 1–12.
18. Gao, F.; Yao, H.; Gao, Q.; Ying, L.; Cai, Y.; Jin, Y.; Pan, Y. Deep Reinforcement Learning-Based Two-Stage Distributed Power Source Optimization Considering Parameter Sharing. *Chinese Society for Electrical Engineering*, 30 March 2025; 1–18.
19. Hua, X.; Kai, L.; Jingbiao, Q.; Zhang, P.; Wang, Z.; Lu, X. Source-Grid-Load-Storage Collaborative Optimization Dispatch Based on Generative Adversarial Network Modification. *Proc. Chin. Soc. Electr. Eng.* **2025**, *45*, 1668–1680.
20. Yang, Y.; Lu, X.; Zhang, L.; Zhou, S.; Pei, W. Reinforcement Learning-Based Power Grid Dispatch Method Considering Agent Pre-States and Environmental Feature Adaptation Mechanism. *High Volt. Eng.* **2024**, *50*, 3497–3509.
21. Zhu, J.; Xu, S.; Li, B.; Wang, Y.; Wang, Y.; Yu, L.; Xiong, X.; Wang, C. Real-Time Dispatch of New Power Systems Based on Power Grid Expert Strategy Imitation Learning. *Power Syst. Technol.* **2023**, *47*, 517–530.
22. Xu, S.; Zhu, J.; Li, B.; Yu, L.; Zhu, X.; Jia, H.; Chung, C.Y.; Booth, C.D.; Terzija, V. Real-time power system dispatch scheme using grid expert strategy-based imitation learning. *Int. J. Electr. Power Energy Syst.* **2024**, *161*, 110148. [CrossRef]
23. Wang, K.; Wan, X.; Wang, J.; Li, Y.; Xu, Y.; ASAD WAQAR. A Review and Outlook on the Model-Data-Knowledge Fusion Method for Power Grid Optimization Dispatch. *Proc. Chin. Soc. Electr. Eng.* **2024**, *44* (Suppl. S1), 131–145.
24. Dou, X.; Deng, Y.; Wang, S.; Chu, T.; Li, J.; Zhou, J.; Li, C. Reactive Power Coordination Optimization Dispatch of Distribution Networks Based on an Improved Multi-Agent Deep Deterministic Policy Gradient Algorithm. *Electric Power Automation Equipment*, 10 June 2025; 1–23.
25. Dou, X.; Niu, P.; Zheng, Y.; Feng, S.; Shi, F.; Yang, H. Method for Generating Extreme-Form Load Scenarios Based on an Improved Diffusion Model. *Chinese Society for Electrical Engineering*, 9 April 2025; 1–18.
26. Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.K.; Wu, Y.; et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv* **2024**, arXiv:2402.03300.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.