Game Theory Guided Data-Driven Multi-Entity Distribution Network Optimal Strategy

Xingmou Liu, Yuan Zuo, Ning Yang*, Yao Xiao, Ammd Jadoon

***Abstract*—As the penetration of renewable energy sources (RES) continues to increase, more and more microgrids (MG) are interconnected with distribution system operators (DSO). To reduce system operational costs between MGs and DSOs, it is necessary to develop certain optimization strategies. This article proposes an optimized collaborative framework for modeling multi-entity distribution networks. In this model, DSOs are placed at the upper level to formulate policies, while MGs are at the lower level to respond in real-time to these policies. Furthermore, the multi-agent relationships in the model are described using Stackelberg game mechanisms, enhancing economic efficiency through dynamic gaming. Additionally, a data-driven multi-agent twin-delayed deep deterministic policy gradient (MATD3) algorithm is investigated to simulate the gaming process and improve the overall model's non-linear optimization capabilities. Considering that the simulation process can lead to violations of the energy storage system capacity constraints, a physics-based model is designed within the framework to ensure the safety of energy storage systems (ESS). Finally, compared to the MADDPG and penalty function methods, the proposed approach reduces the operational costs by 19.05%.**

***Index Terms*—Microgrids, multi-agent deep reinforcement learning, game theory, distribution network**

I. Introduction

MICROGRIDS, by integrating various distributed energy sources (DERs) and connecting to distribution networks (DNs), form a large-scale energy system [1], [2], [3]. This plays a positive role in increasing the proportion of RES in the power system. However, multi-agent interest relationships exist between MGs and DSOs after connection. MGs operate as third-party operators, and DSOs belong to different interest groups [4], [5]. Both are profit-oriented, with the aim of maximizing their own interests [6]. Additionally, the uncertainty of RES supply [7], [8], the allocation of energy supply and demand, and the control of charging and discharging of the ESS pose significant challenges to the non-linear optimization capacity of the system model [9]. If the system cannot effectively address these challenges, it will affect the stability of the power supply and economic operations [10], [11]. Therefore, an appropriate optimization strategy for the system is needed.

Researchers have proposed centralized optimization and Stackelberg game mechanisms to address the challenges of energy optimization between MGs and DSOs. Yan Du and Zhaoyu Wang formulated DSOs and MGs as a bilevel model, then employed Karush-Kuhn-Tucker (KKT) conditions to transform the model into single-level optimization problems [12], [13]. To address the uncertainty of RES and load profiles, Yongsheng Cao developed a distributed robust optimization algorithm and transformed the bilevel problem into a single-level problem using mathematical programming with equilibrium constraints. However, centralized optimization requires collecting a significant amount of information about the devices and internal details of multiple entities, implying that the privacy of these entities cannot be adequately protected. For multi-entity relationships, the Stackelberg game provides an appropriate mechanism to describe. Qiang Li established a leader-follower relationship between the energy trading system and MG within the Stackelberg game mechanism, designing the corresponding optimization models to encourage more active participation in transactions [14]. Fengzhou Sun modeled DSOs and MGs within the Stackelberg game as an energy optimization model with different priorities and developed an energy trading mechanism based on mathematical programming methods [15]. Juntao Chen used the Stackelberg game to analyze the trading relationships between DERs and MGs and devised a heuristic algorithm to achieve equilibrium [16]. In summary, this article proposes a collaborative optimization model based on multi-entity distribution networks while preserving privacy through a decentralized approach. Additionally, it applies the Stackelberg game mechanism to describe the relationships among different entities. In this mechanism, DSOs and MGs have different priorities, the DSO acts as the leader positioned at the upper level making decisions for MGs, who act as followers at the lower level and provide optimal real-time responses, reducing operational costs via dynamic games.

From previous research, it is evident that most researchers within the framework and Stackelberg game mechanism have primarily employed optimization methods such as model-based mathematical programming or heuristic algorithms. Mathematical programming methods are typically

Manuscript received September 25, 2023; revised January 30, 2024.

Xingmou Liu is a professor of Key Laboratory of Industrial Internet of Things and Networked Control, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400000, China (e-mail: liuxm@cqupt.edu.cn).

Yuan Zuo is a graduate student of Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: 1032732218@qq.com).

Ning Yang is an engineer of State Grid Ningxia Electric Power Co. LTD Yinchuan Power Supply Company Yinchuan City, Ningxia, China (phone:+8618983816609, e-mail:791339327@qq.com).

Yao Xiao is an engineer of Chongqing Hongyu Precision Industry Group Co. LTD, Chongqing, China(e-mail:4711204@qq.com).

Ammad Jadoon is a professor of NFC-Institute of Engineering & Technology, Multan, Pakistan (e-mail: ammad.jadoon@nfciet.edu.pk).

unsuitable for problems with high complexity or fuzzy constraints. Heuristic algorithms often cannot store optimization knowledge and apply it to new tasks, which requires significant time costs to complete search tasks. As power systems move toward greater intelligence, an increasing amount of data needs to be transmitted, processed, and stored. Therefore, it becomes crucial to efficiently harness massive data to respond to complex and variable power demands in real time, quickly, and accurately.

Combining deep neural networks and reinforcement learning, deep reinforcement learning (DRL) has been widely used to solve energy optimization problems to address the limitations of model-based methods. Due to their complex network depth, deep neural networks can extract features from high-dimensional mappings and data. With its unique features of not requiring prior knowledge and model-independent learning, reinforcement learning allows agents to respond flexibly to dynamically changing environments[17]. Therefore, combining these two allows optimization strategies between MGs and DSO to have better adaptability and flexibility[18]. For example, to reduce MG operational costs, Ying Ji used Proximal Policy Optimization (PPO) and Deep Q Network (DQN) to optimize MG operations[19], [20]. Luqin Fan formulated the MG energy scheduling problem as a Markov decision process and then used DRL to optimize for the minimum operating cost of MG[21]. Although these researchers have achieved some success using DRL to tackle energy optimization problems, these achievements have mainly been within the content of single-agent environments.

With the rapid development of the electricity market, it has transitioned into a multi-entity and multi-interest environment[17], [22], [23],. Therefore, when faced with such an environment, it becomes necessary to apply multi-agent deep reinforcement learning (MADRL) to learn and manage the energy scheduling of the system, thereby optimizing operational costs[24]. For example, in early attempts in multi-agent environments, Xiaohan Fang utilized a multi-agent DQN to optimize systems[25]. However, MADQN cannot handle the problem of continuous action space control. Subsequently, Chenyu Guo applied MADDPG to solve the problem, avoiding the drawbacks of MADQN[26]. However, MADDPG suffers from the issue of overestimating Q values when processing data. To cater to the increasingly complex power demands and address the problem of overestimation caused by continuous action space control, this article proposes the multi-agent twin-delayed deep-deterministic policy gradient (MATD3) algorithm for optimizing energy systems. This data-driven approach can significantly reduce the complexity of the problem.

The main contributions of this article are as follows:

(1) We proposed a collaborative optimization framework for modeling multi-entity distribution networks and introduced the Stackelberg game mechanism to describe the relationships. In this framework, DSOs act as leaders positioned at the upper level, while MGs act as followers at the lower level. Through the game between leaders and followers, the operational costs for both parties are reduced;

(2) We introduced a data-driven MATD3 algorithm to simulate the gaming process and enhance the model's nonlinear optimization capabilities;

(3) We proposed a physics-based model to ensure the safety of the ESS and maintain the economic and operational security of the system.

II. Framework for Multi-entity Distribution Networks

The operation of the DN is controlled by the DSO. DN comprises photovoltaic (PV), wind turbine (WT), and ESS. MG1 and MG2 consist of PV, MT, and ESS. MG3 includes WT, MT, and ESS. The specific configuration is illustrated in Figure 1.

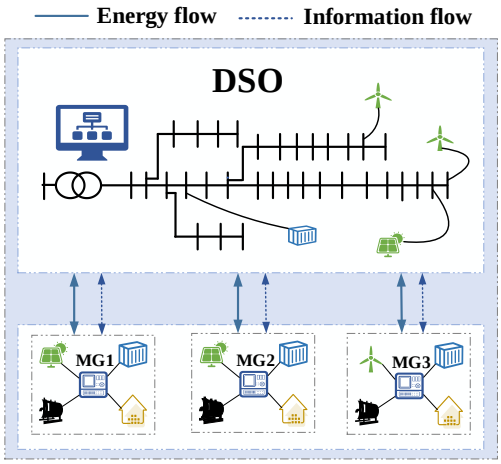


Fig. 1. Proposed Energy System Framework.

MGs coordinate with the DSO by exchanging information and energy within the energy system. The DSO acquires information in a decentralized manner, meaning that it obtains information separately from different MGs and only has access to the power and load information of the MGs, thus preserving the privacy of the MGs through limited information exchange. After obtaining the information, the pricing information is sent to the MGs by DSO. When received, MGs then decide whether to sell or buy electricity from the DSO based on their DER capacity to meet the load demand. If an MG's DER cannot meet its load demand, the MG will purchase electricity from the DSO or dispatch its local MT and ESS to maintain supply-demand balance. Any surplus electricity generated by the MG can be stored in the ESS or sold to the DSO.

Due to the different objective functions and constraints of the DSO and MG, this paper establishes the DSO and MG as separate hierarchical models within the proposed optimization framework. In the context of the multi-entity relationships within the framework, different hierarchical models are constructed using the Stackelberg game mechanism, with different priorities. In the model, the DSO acts as the upper-level leader, while the MGs act as lower-level followers. Since MGs and DSO represent different stakeholders, the optimization objectives of the upper and lower levels of the models are to minimize operational costs. In practical operations, the DSO sends pricing information to MGs. When this information is received, the MGs decide whether to sell or purchase electricity from the DSO according to local supply-demand conditions. The collaborative optimization framework for

multi-entity distribution networks is illustrated in Figure 2. The mathematical models for the upper and lower levels are shown in Sections II-B and II-A, respectively.

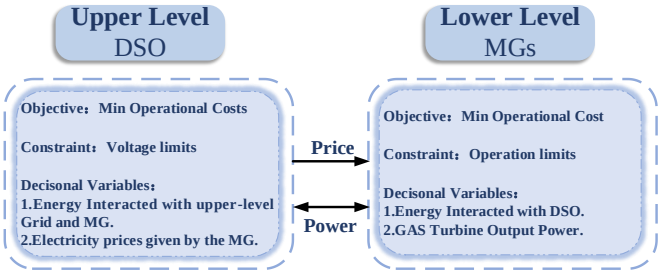


Fig. 2. Coordinated Optimization Framework Based on Multi-Entity.

A. *Lower-level model*

The objective of the lower level is to minimize the operational cost of MG. This cost covers the expenses related to the exchange of electricity with the DSO and the operating cost of MTs. Whether the MG needs to purchase electricity from the DSO is jointly determined by renewable energy sources, dispatchable resources, and electricity load. The model's constraints include power balance at both the supply and demand ends, and operational constraints for the ESS and MT.

1) Objective:

$$F_{i,t}^{MG}=\left[\left(h_{i,t}^{MGbuy}\cdot p_{i,t}^{MGbuy}-b_{i,t}^{MGsell}\cdot p_{i,t}^{MGsell}\right)+\left(c\cdot\left(p_{i,t}^{MT}\right)^{2}+d\cdot p_{i,t}^{MT}\right)\right]\cdot\Delta t \qquad i=\{1,2,3\} \tag{1}$$

In the above, t represents a specific period within a day, while i represents the i-th microgrid. $F_{i,t}^{MG}$ is the total operating cost of MG_i within t. $p_{i,t}^{MGbuy}$ is the electricity purchased by MG_i. $h_{i,t}^{MGbuy}$ is the electricity price at which microgrid i purchases power. $p_{i,t}^{MGsell}$ is the amount of electricity sold by microgrid i. $b_{i,t}^{MGsell}$ is the electricity price at which MGs sell power. $p_{i,t}^{MT}$ is the output power of the MT, c and d are the cost coefficients of the MT.

2) Constraints:

$$p_{i,t}^{MGbuy}+p_{i,t}^{MT}+p_{i,t}^{PV}+p_{i,t}^{dis}-p_{i,t}^{load}-p_{i,t}^{cha}-p_{i,t}^{MGsell}=0 \tag{2}$$

$$p_{\min}^{MT}\le p_{i,t}^{MT}\le p_{\max}^{MT} \tag{3}$$

$$p_{\min}^{ESS}\le p_{i,t}^{cha},p_{i,t}^{dis}\le p_{\max}^{ESS} \tag{4}$$

$$BT_{i,t+1}^{ESS}=BT_{i,t}^{ESS}+\left(P_{i,t}^{cha}-P_{i,t}^{dis}\right)\cdot\Delta t \tag{5}$$

$$SOC_{\min}\le BT_{i,t}^{ESS}/BT_{\max}^{ESS}\le SOC_{\max} \tag{6}$$

$$0\le p_{i,t}^{MGbuy},p_{i,t}^{MGsell}\le p_{\max}^{MG} \tag{7}$$

(2) denotes the power balance constraint. (3) is the operational constraint for the MT. (4) are the charging and discharging constraints of the *ESS*, BT^{ESS} is the capacity of the ESS. (5) represents the relationship between the ESS capacity at time $t+1$ and time t. (6) are the SoC constraints of the ESS, $BT_{\max}^{ESS}$ denotes the maximum capacity of the ESS. (7) represents the constraint on the interaction power between the MGs and the DSO.

B. *Upper-level model*

The primary objective of the upper-level model is to minimize the operational cost of the DSO. This cost includes the electricity prices for interactions with the MGs and the energy purchases from the upper-level grid. The model's constraints include voltage constraints of the distribution network.

1) Objective:

$$F_{t}^{DSO}=\left[e_{t}^{uppergrid}\cdot p_{t}^{uppergrid}-\left(h_{i,t}^{MGbuy}\cdot p_{i,t}^{MGbuy}-b_{i,t}^{MGsell}\cdot p_{i,t}^{MGsell}\right)\right]\cdot\Delta t \qquad i=\{1,2,3\} \tag{8}$$

In the objective function, t represents a specific period within a day and i represents the i-th microgrid. F_{t}^{DSO} denotes the total operational cost within the time step t. $e_{t}^{uppergrid}$ is the cost coefficient for the DSO to purchase electricity. $p_{t}^{uppergrid}$ represents the electricity purchased by the DSO. The meanings of other variables have been elaborated in detail in the lower-level model.

2) Constraints:

$$\begin{aligned}&V_e(i,t)\sum_{j=1}^{N}\left(G(i,j)V_e(j,t)-B(i,j)V_f(j,t)\right)+\\&V_f(i,t)\sum_{j=1}^{N}\left(G(i,j)V_f(j,t)+B(i,j)V_e(j,t)\right)+\\&P(i,t)=0 \quad i\in N\end{aligned} \tag{9}$$

$$\begin{aligned}&V_f(i,t)\sum_{j=1}^{N}\left(G(i,j)V_e(j,t)-B(i,j)V_f(j,t)\right)-\\&V_e(i,t)\sum_{j=1}^{N}\left(G(i,j)V_f(j,t)+B(i,j)V_e(j,t)\right)+\\&Q(i,t)=0 \quad i\in N\end{aligned} \tag{10}$$

In the above formulas, *G(i,j)* is the element of the real part of the complex admittance matrix and *B(i,j)* is the imaginary part. *P(i,t)* and *Q (i,t), respectively*, denote the active power and reactive power injected into bus i during time step t. Equations (9) and (10) are the calculation equations for active and reactive power, respectively.

Bus voltage limit:

$$V_{bus-\min}(i)\le V_{bus}(i,t)\le V_{bus-\max}(i) \tag{11}$$

$V_{bus}(i,t)$ represents the voltage at bus i during timestep t. $V_{bus-\min}(i,t)$ and $V_{bus-\max}(i,t)$ respectively denote the upper and lower voltage limits for bus i.

III. Multi-entity relationships Described by the Stackelberg Game

Since MG and DSO belong to different operators and are independent entities. Therefore, there exists a multi-entity relationship between the grid-connected MG and DSO. Hence, this section introduces the Stackelberg game to analyze the multi-entity relationship between the bilevel energy systems. In the Stackelberg game, DSO and MG have different priorities, where DSO can first pass its strategic information to MG and MG responds upon receiving the information. Based on this, DSO is constructed as the upper-level leader, while MG acts as the follower in the lower-level. The following sections will elaborate on the construction of the game mechanism and the analysis of the game process between DSO and MG, covering these two aspects to explain the multi-entity relationship.

A. Construction of the Optimization Framework Based on the Stackelberg Game Mechanism

The proposed optimization framework in this paper is constructed under the mechanism of the Stackelberg game as a single-leader-multi-follower structure. In this structure, the DSO acts as the upper-level leader, while the microgrids act as lower-level followers. The following is a definition of the mechanism of the game for the model.

$$G=\left\{\Psi^{DSO},\Psi^{MG},\pi_t^{DSO},\left\{\pi_{i,t}^{MG}\right\}_{i\in\Psi^{MG}},C_t^{DSO},\left\{C_{i,t}^{MG}\right\}_{i\in\Psi^{MG}}\right\} \quad (12)$$

In (12), it can be divided into three parts: the participants set (Ψ), the operational strategies (π_t), and the cost functions (C_t).

Let us explain each of these parts:

Participant Set: DSO is represented by Ψ^{DSO}. Multiple microgrids are represented by Ψ^{MG}.

Operational Strategies: The operational strategies for DSO and multiple MGs are represented respectively represented by π_t^{DSO} and $\left\{\pi_{i,t}^{MG}\right\}$. These strategies correspond to the selling prices of electricity that the distribution system operator offers to the microgrids, which are the following:

$$\pi_t^{DSO}=\left\{h_{i,t}^{MGbuy},b_{i,t}^{MGsell}\mid i\in\Psi^{MG}\right\} \quad (13)$$

The strategy of the microgrid not only involves energy exchange with the distribution network but also includes the management of internal controllable resources, which are:

$$\pi_{i,t}^{MG}=\left\{p_{i,t}^{MT},p_{i,t}^{dis},p_{i,t}^{cha},p_{i,t}^{MGsell},p_{i,t}^{MGbuy}\right\} \quad (14)$$

Cost functions: The leader's cost function is $C_t^{DSO}=-F_t^{DSO}$, the follower's cost function is $C_{i,t}^{MG}=-F_{i,t}^{MG}$.

B. The Stackelberg game process

The bilevel game process is described below. The leader formulates strategies based on the internal environment and subsequently transmits these strategies to lower-level followers. Once each follower receives the strategy information, they respond to the leader's strategy based on their internal environment. The leader in the upper layer then iterates the strategy information based on the feedback received and the internal conditions, transmitting it again to the followers in the lower layer. Subsequently, the followers respond once more to the leader's strategy. The exchange of strategy formulation and response between the leader and followers in this process is considered a game of energy optimization. This dynamic game between the leader and the followers enhances the economic performance of the energy system's operation. Throughout the game process, only strategy information is transmitted among both parties.

IV. MADRL

The operation of MGs and DSO is a complex problem due to the need to consider various energy sources, loads, and market conditions. These factors are often non-linear and dynamic, making them difficult to precisely describe using traditional mathematical optimization methods. Additionally, MGs typically incorporate many sensors and monitoring devices. As MGs become more intelligent, there is an increasing volume of data that must be transmitted, computed, and stored. Therefore, it becomes crucial to effectively harness large amounts of data to address complex and variable load demands.

To address the issues and challenges mentioned above, this section proposes the use of a data-driven MATD3 for energy optimization between MGs and DSO. MATD3, characterized by its complex network depth and model-free learning, can handle intricate nonlinear dynamic systems and adaptively adjust strategies. Moreover, it can learn and make optimization decisions from massive real-time data without relying on predesigned rules. Furthermore, MATD3 can help establish game mechanisms within the model and simulate the game process, thereby reducing the operational costs of the energy system through dynamic simulation. In the game mechanism, the leader and followers are constructed as agents within the algorithm, while the mathematical model proposed in Section II is mapped to a partially observable Markov game (POMG).

A. Mapping the Mathematical Model to a POMG

The mathematical model presented in Section Two is mapped as a POMG. POMG is a modeling tool for addressing multi-agent decision-making problems, where the agents' states are partially observable. Symbolically represented as $POMG\left(M,O_{global},\left\{a_i\right\},\left\{O_i\right\},P,\left\{r_i\right\},\gamma\right)$, where M represents the number of interacting agents, O_{global} is the set of global observations that all agents observe. $\left\{a_i\right\}$ is the set of agent's actions.

P is the state transition probability. $\left\{r_i\right\}$ is the reward function for agents in the interactive environment. γ denotes the discount factor. Given the presence of heterogeneous agents in the environment, Equation (15) expresses the maximum efficiency of the expected reward return for each agent in time T.

$$R_i = \sum_{n=1}^{T} \gamma^t r_{i,t} \tag{15}$$

1) DSO agents: Observation space:

$$O_t^{DSO} = \left\{ V_t^{bus\max}, V_t^{bus\min}, P_t^{MG_i}, BT_t^{DSO}, P_t^{DSOload}, P_t^{DSO-RES}, e_t^{uppergrid}, p_t^{uppergrid} \right\} \tag{16}$$

$V_t^{bus\max}$ and $V_t^{bus\min}$ are the maximum and minimum bus voltage of DN. $P_t^{MG_i}$ is the power interaction between DN and MG. BT_t^{DN} is the capacity of ESS. P_t^{DNload} and P_t^{DN-RES} are the local load and renewable energy generation in DN. $P_t^{uppergrid}$ and $e_t^{uppergrid}$ are the power purchased by DN and the selling price of electricity to the DN.

Action space:

The DN's action space is determined by the capacity of the ESS and the prices of buying and selling electricity.

$$a_t^{DSO} = \left\{ BT_t^{DN}, a_{i,t}^{MGbuy}, b_{i,t}^{MGsell} \right\} \tag{17}$$

Reward function:

Considering that the DN bus voltage must remain within a safe range, a penalty function is added to the reward function of the distribution network for bus voltage violations. This is used to limit deviations from the desired bus voltage levels, as shown in Equation (18).

$$penalty = \lambda \cdot \left(\sum_{n=1}^{33} \left| V_n^{bus} - \alpha \right| + \sum_{n=1}^{33} \left| V_n^{bus} - \beta \right| \right) \tag{18}$$

V_n^{bus} is the bus voltage, in per-unit (p.u) values, where n represents the *n*-th node. α and β are the upper and lower bounds for the system node voltage. λ is a coefficient determined by the system operator.

$$r_t^{DSO} = \left(-penalty - F_t^{DSO} \right) \tag{19}$$

The reward function for the DN consists of the bus voltage penalty function and the operating cost of the DN.

2) MG agents: In the optimization framework, involves multiple MGs. In each time step *t*, the observation space, the action space, and the reward function for each MG are defined as follows.

observation space:

$$O_{i,t}^{MG} = \left\{ V^{bus\max}, V^{bus\min}, P_{i,t}^{MGload}, P_{i,t}^{MGRES}, a_{i,t}^{MGsell}, b_{i,t}^{MGbuy}, BT_{i,t}^{MG}, P_t^{MG_i} \right\} \tag{20}$$

$P_{i,t}^{MGload}$ is the local load of microgrids, $P_{i,t}^{MGRES}$ representing the power generation from RES within the MGs.

Action space:

$$a_{i,t}^{MG} = \left\{ P_{i,t}^{MG-MT}, BT_{i,t}^{MG} \right\} \tag{21}$$

$P_{i,t}^{MG-MT}$ is the power generation of an internal microturbine in the MG. $BT_{i,t}^{MG}$ is the capacity state of the internal ESS in the MG.

Reward function:

$$r_{i,t}^{MG} = -F_{i,t}^{MG} \tag{22}$$

The reward function of the MG is the operational costs.

B. TD3

The Twin-Delayed Deep Deterministic Policy Gradient (TD3) is a model-free actor-critic algorithm that falls into the category of off-policy and deterministic policy gradient methods[27]. TD3 was introduced to address the problem of overestimation bias that exists in Deep Deterministic Policy Gradient (DDPG) algorithms. The critic network consists of two Q-networks q (q_{ω_1} and q_{ω_2}), along with two target Q-networks q' (q'_{ω_1} and q'_{ω_2}). The actor network includes a deterministic policy network (μ_θ) , and an target policy network (μ'_θ). The existence of objective networks enhances the stability of neural network learning.

For network updates, the q in the critic network is updated using Equation (23). Here, s_t is the state space, a_t is the action space, and y_t is the target value. In the actor, μ_θ is updated via Equation (24).

Due to the initial differences in the network parameters, the calculated target values may differ. Therefore, the critic estimates the Q value by taking the minimum of the two target Q networks by Equation (25), to avoid overestimation of the Q values. γ is the discount factor, and r_t is the reward function. Regarding policy updates, TD3 updates the μ_θ and three target networks every few updates of the Q-networks. This delayed update strategy significantly reduces the probability of misusing Q values. To enhance the robustness of the network, clip Gaussian noise is added to the actor's updates, as depicted in Equation (26), where c represents the bounds for the noise.

$$L_\omega = \mathrm{E}\left[\left(q_\omega(s_t, a_t) - y_t \right)^2 \right] \tag{23}$$

$$\nabla_\theta g \approx \mathrm{E}\left[\nabla_a q_\omega(s,a)|_{s=s_t, a=\mu_\theta(s_t)} \nabla_\theta \mu_\theta(s)|_{s=s_t} \right] \tag{24}$$

$$y_t = r_t + \gamma \min_{j=1,2} q'_{\omega_j}\left(s_{t+1}, \mu'_{\theta,t+1} \right) \tag{25}$$

$$a'_{t+1} = \mu'_\theta(s_{t+1}) + \xi, \xi \sim CN(N(0, \sigma^2), -c, c) \tag{26}$$

C. MATD3

However, the framework proposed in this article is a multi-agent environment. For the multi-agent learning strategy, it is not feasible to simply apply the TD3 algorithm

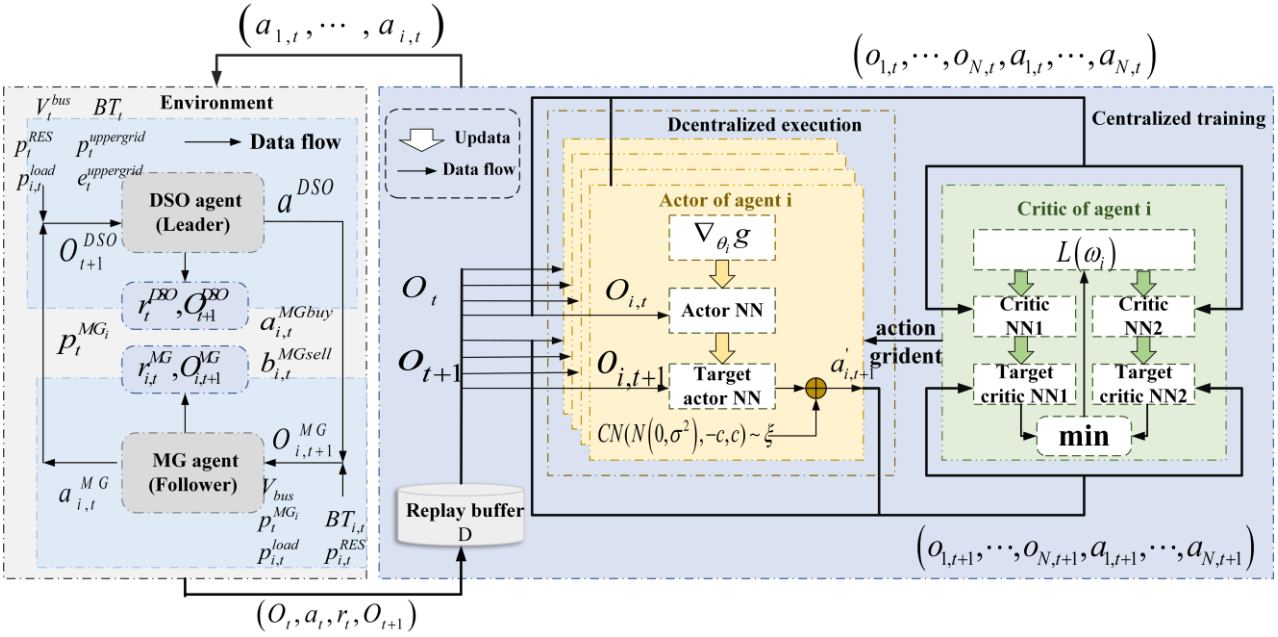


Fig. 3. MATD3 Training and Execution Framework.

independently to the network training of each agent. This is because in a multi-agent environment, each agent operates in a non-stationary environment, and the actions taken by one agent can interfere with the convergence of the others' strategies. Therefore, this article adopts the principles of centralized training and distributed execution to assist multiple agents in learning the optimal strategy within the environment. Centralized training refers to the critic network considering the observation space and action space of all agents during training. Distributed execution means that the actor network uses only local information to update during training. The update formula for the centralized Q-value network is as follows: (27)-(28). The gradient update formula for the policy network is as follows: (30).

$$L(\theta_i)=\mathrm{E}_{o_t,a_t,r_t,o_{t+1}}\left[\left(q_{\omega_i}\left(o_{1,t},\cdots,o_{N,t},a_{1,t},\cdots,a_{N,t}\right)-y_{i,t}\right)^2\right] \quad (27)$$

$$y_{i,t}=r_{i,t}+\gamma\min_{j=1,2}q'_{\omega_{ij}}\left(o_{1,t+1},\cdots,o_{N,t+1},a_{1,t+1},\cdots,a_{N,t+1}\right) \quad (28)$$

$$a'_{i,t+1}=\mu'_{\theta_i}\left(s_{i,t+1}\right)+\xi_i,\xi_i\sim CN(N\left(0,\sigma_i^2\right),-c_i,c_i) \quad (29)$$

$$\begin{aligned}&\nabla_{\theta_i}g\\&\approx\mathrm{E}_{o_i,a_i}\left[\nabla_{a_i}q_{\omega_i}\left(o_{1,t},\cdots,o_{N,t},a_{1,t},\cdots,a_{N,t}\right)|_{o=o_{i,t},a=\mu_{\theta_i}(o_{i,t})}\nabla_{\theta_i}\mu_{\theta_i}\left(o_i\right)|_{o=o_i}\right]\end{aligned} \quad (30)$$

D. Centralized training and distributed execution of MATD3

The distributed execution and centralized training framework of the algorithm is illustrated in Figure 3. The algorithm training process will be explained in the following sections, focusing on centralized training and distributed execution.

1) Centralized Training Stage:

Algorithm 1

1. Initialization: ω_1, ω_2, γ, θ_1 and Replay buffer
2. For episodes 1 to *M*,
3. For time slots (t) 1 to *T* do
4. According to the actor network, each agent output action $a_t=\mu_\theta(O_t)$;
5. Agents execute action a_t, and return r_t and O_{t+1} to the environment;
6. Save element matrix (O_t,a_t,r_t,O_{t+1}) to Replay buffer;
7. If reach pre-set element matrix amount;
8. Sampling from the replay buffer according to the pre-configured batch size;
9. Import O_t and O_{t+1} to actor network. Next, integrate the network's outputs a_t, a_{t+1} with the observation space. Then, import $(O_{1,t},\cdots,O_{i,t},a_{1,t},\cdots,a_{i,t})$ and $(O_{1,t+1},\cdots,O_{i,t+1},a_{1,t+1},\cdots,a_{i,t+1})$ to critic network;
10. Update the parameters of the critic network's parameters via $L(\theta)$;
11. Update actor network's parameters via $\nabla_\theta g$;
12. Update target critic network's parameters.
13. End
14. End
15. End

2) Distributed Execution Process:

Algorithm 2

1. For episodes 1 to *M* do
2. Obtaining the observation space for all agents $(O_{1,t},\cdots,O_{i,t})$;
3. For time slot(t) 1 to *T* do
4. Utilizing the trained actor network to generate action spaces for all agents;
6. All agents in the environment execute distributed actions $(a_{1,t},\cdots,a_{i,t})$, and return r_t, O_{t+1} to the environment.
7. End
8. End

E. Model Based on Physical Information

The ESS capacities in MGs and DN are elements within the action set of agents. To find the optimal strategy, agents generally aim to expand their range of exploration as much as possible when exploring the environment. Therefore, this exploration process can potentially lead the ESS to exceed its

safety limits. To ensure that the capacity of the ESS remains within the safe range, this paper introduces a model based on physical information as an intermediate layer between agents and the environment. This model is designed to prevent risky behavior from agents.

$$BT_{\mathrm{mod}} = \min\left(BT_{action}, \left(BT_{\max} - BT_{initial}\right)\right)\left(BT_{action} \geq 0\right) \quad (31)$$

$$BT_{\mathrm{mod}} = \max\left(BT_{action}, \left(BT_{\min} - BT_{initial}\right)\right)\left(BT_{action} < 0\right) \quad (32)$$

BT_{mod} is the value modified by formulas, BT_{action} is the value assigned by the agents during the exploration process. BT_{max} and BT_{min} are the maximum and minimum values of the ESS, respectively. $BT_{initial}$ is the initial value of the ESS.

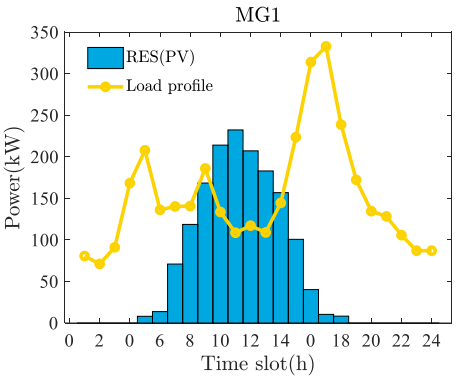
V. Case Study

The objective of this section is to validate the effectiveness of the framework, algorithm, and model proposed in this paper within the IEEE-33 system.

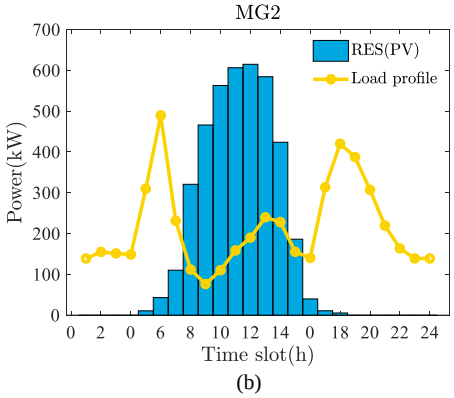
A. Experimental Setup and Parameter Configuration

The tests were carried out within the IEEE-33 system. In the test system, MGs are connected to nodes 30, 21, and 13. Node 4 is connected to an ESS; Nodes 32 and 18 are each connected to a WT, and node 16 has a photovoltaic system.

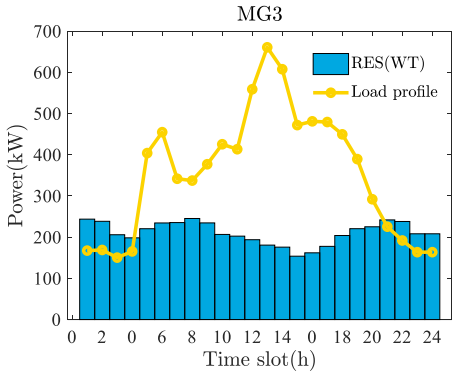
MG1 and MG2 consist of a photovoltaic system, an ESS and a microturbine (MT). MG3 includes a WT, an ESS and a MT. The DN contains two WTs, one photovoltaic system, and an ESS. Figure 4 illustrates the distribution of renewable energy and the loads for the MGs over a 24-hour period, while Figure 5 shows the distribution of renewable energy in the DN, with the curve representing the trend of the total network load (in percentage).



(a)



(b)



(c)

Fig. 4. Distribution of the RES and load profile in the MG.

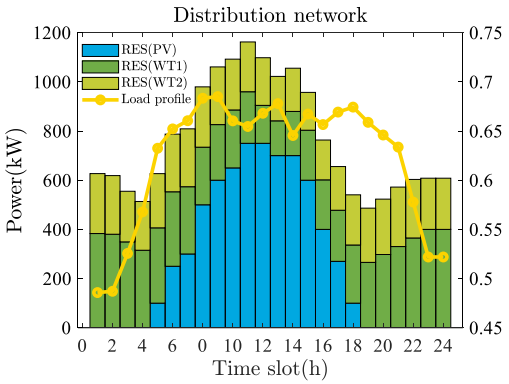


Fig. 5. Distribution of RES and load profile in DN.

The system parameters are described in Table I. The capacity of ESS is 300 kW ħ, with a charging and discharging power range of 0-100 kW. The SoC remains within 0.1 to 0.9. The power generation for the MT is 0-220 kW. The cost coefficients of MT, denoted as *c* and *d*, are 0.27 \$/(kW)^2 and 0.14 \$/kW, respectively. The voltage reference for the DN is 12.66 kV, with an allowable voltage deviation range of 0.90-1.10. The initial capacities of the ESS for MGs and DN are 40 kW ħ and 200 kW ħ, respectively.

TABLE I
System Parameters

ESS		MT		DN	
SOC	**Max Capacity**	**Power Range**	**Cost factor**	**Vref**	**Voltage. Deviation**
0.1~0.9	300 kW	0~220 kW	*c*=0.27 \$/kW^2 *d*=0.14 \$/kW	12.66 kV	0.9~1.1

The hyperparameters for MATD3 and MADDPG, the training iterations, and the applied environment are consistent. The training iterations are set to 3000 episodes. The batch size for each time step is 128. The discount factor is 0.99, and the target smoothing coefficient is 0.005. The target update frequency is 2. The size of the experience pool is 1x10^5. The learning rates for the actor and critic networks are set at 0.005.

The simulation is carried out in MATLAB/Simulink 2022b. The laptop includes a Core i7-6700H, 16GB RAM, and a NVIDIA GeForce MTX 950m GPU.

B. Optimization results analysis of optimization results based on DRL

Figure 6 presents the scheduling results for the three MGs under TD3 over a 24-hour period. Each figure includes the

charging and discharging power of the ESS, RES generation, MT generation, power exchange with the DSO, and load profile.

From the figures, it can be observed that the MT actively participates in the energy scheduling at every time step. During the 0-7 and 16-24 time periods (Figures 6. (a) and (b)), when PV cannot generate power or its power generation is insufficient to meet load demand, MT is scheduled to compensate for the RES shortage and maintain the internal supply-demand balance. In the period of maximum solar intensity from 8 to 15 (Figures 6.(a) and (b)), photovoltaic generation can provide sufficient power to the system. During this time, MT scheduling can be used for ESS energy storage or energy trading with the DSO, increasing revenue for the MG operator. In the 1-4 and 21-24 time periods in Figure 6. (c), the WT's power generation can meet the load demand, and the MT is scheduled to work in coordination with the WT for energy storage of the ESS or energy trading with the DSO. During the 5-20 time periods (Figure 6. (c)), the load gradually reaches its peak and RES cannot meet the load demand. At this point, the MT is scheduled to participate in response to internal demand, reducing supply pressure.

ESS charging actions are performed mainly when the generation of RES is sufficient to meet the internal load demand. This is evident in time intervals such as 8-10 in Figure 6. (a), 11-13 in Figure 6. (b) and 2-3 in Figure 6.(c). When RES generation cannot meet the load demand or is not available, the ESS is scheduled to work in coordination with the MT and the RES, releasing stored energy to provide more electricity to the demand side.

Regarding energy trading, it can be observed from the figures that when the internal load in the MGs is not at its peak, RES, together with the MT and ESS, can sell electricity to the DSO. A typical example is the time range of 8-15 in Figure 6. (a). If the internal supply-demand balance cannot be maintained, the MGs can purchase electricity from the DSO to alleviate the supply pressure during peak load periods. A typical example is the time range of 9-19 in Figure 6. (c).

In summary, with the support of MATD3, MGs can effectively utilize RES, dispatchable resources, and respond flexibly to DSO trading signals to maintain supply-demand balance. The learning results based on the MADDPG are shown in Figure 7, and the analysis process is similar to that described above.

Figures 6 and Figure 7 provide detailed results of the testing for both algorithms in relation to RES, dispatchable resources, and energy trading. Figure 8, on the other hand, presents the results from a cost perspective after both algorithms have undergone the same learning process. It is evident from the graph that the cost trends based on the two algorithms remain largely consistent over the 24-time steps. Table II displays the operational costs for both the DSO and the MGs according to the two algorithms. It can be observed that within the 24-hour period, the total cost of MG1 and MG3 according to MATD3 is higher than that according to MADDPG. However, from the total cost figures in the table, it is apparent that the total cost based on MADDPG is $68.667 higher than that based on MATD3. Therefore, under the conditions of consistent hyperparameters and learning iterations, the MATD3 algorithm appears to be more cost-effective for the current system compared to MADDPG.

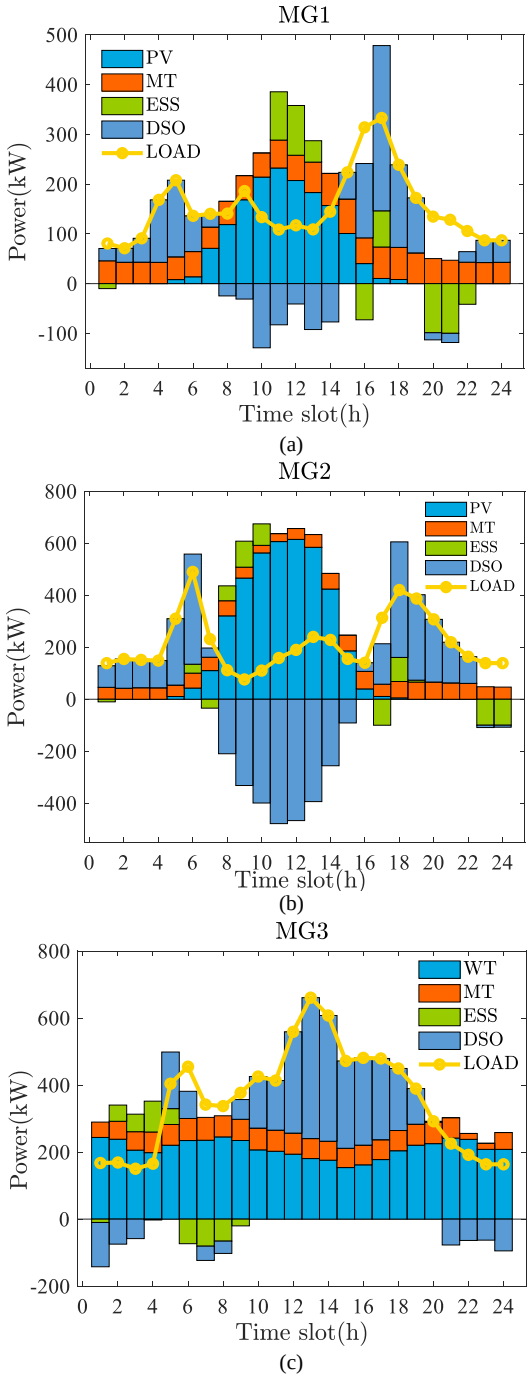
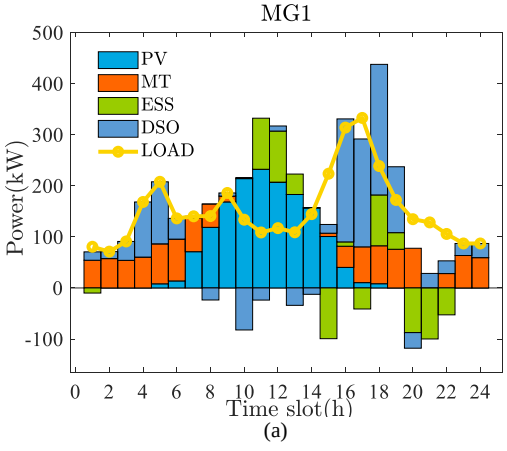
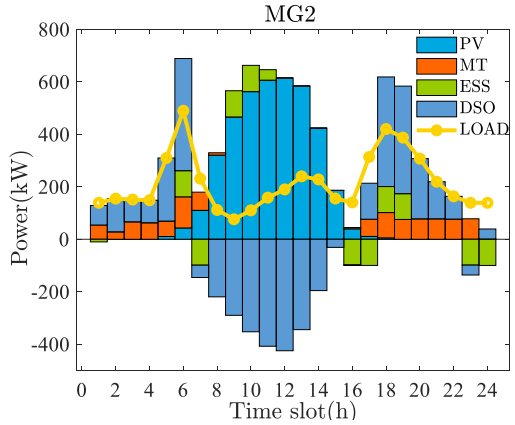
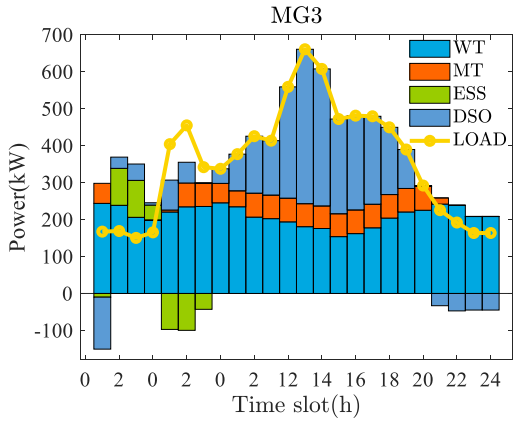


Fig. 6. MG's Operation Results Based on MATD3.





(b)



(c)

Fig. 7. MG's operation results based on MADDPG.

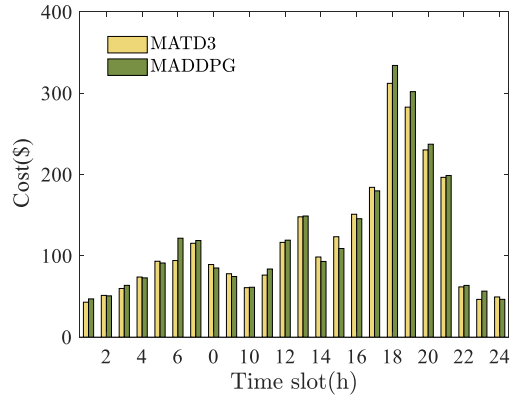


Fig. 8. Distribution of system total costs for MATD3-based and MADDPG-based systems over 24h.

TABLE II
STATISTICAL RESULTS OF TOTAL COSTS FOR MATD3 AND MADDPG FOR 24H

	Operational costs ($)				
	MG1	**MG2**	**MG3**	**DSO**	**Total**
MATD3	498.788	643.162	927.302	767.044	2836.296
MADDPG	372.351	664.475	926.420	941.717	2904.963

C. *Scenario comparison*

Due to the potential for random actions by agents that could cause the ESS to exceed its capacity limits, this paper introduced a model based on physical information in the environment to protect the ESS. However, some works (such as [7], [9]) constructed penalty functions in the reward function to penalize actions that cause capacity violations. However, the construction of penalty functions can limit agents' exploration of the environment because agents, in their exploration, prioritize avoiding penalties to achieve higher reward values. This can lead to neglect of other important aspects of the reward function, such as operational costs.

To highlight the advantages of this model, this section compares the MATD3 learning performance in the two scenarios mentioned above. Assesses the timing of battery charging and discharge, power distribution, and operating costs.

In Figure 9, a power greater than 0 indicates that the ESS is charging, while a power less than 0 indicates that it is discharging. For MG1, combined with Figure 6. (a), we can understand that MG1's ESS chooses the timing of charging and discharging based on internal supply-demand balance and the selling price of electricity set by the DSO. For example, in the time range of 11-13 in Figure 6. (a), MG1's supply power is significantly greater than the load demand. During this time, the ESS chooses to charge, thus storing energy for future use during periods of high demand. Consequently, in the time range of 20-22 in Figure 6. (a), MG1's supply power is notably lower than the load power, and the ESS chooses to discharge to alleviate the internal supply pressure. Additionally, from Figures 9 and Figure 6. (a), we can observe that the magnitude of MG1's charging and discharging power adjusts dynamically based on the difference between supply and load power. Furthermore, it is evident from the curves in Figure 6 that the SoC remains within the range of 0.1-0.9, consistent with the settings of the ESS parameters, thus ensuring that the ESS operates safely. For MG2 and MG3, a similar analysis can be performed by combining Figure 6. (a), (b) and Figure 9. The analysis process for MG2 and MG3 is analogous to that for MG1.

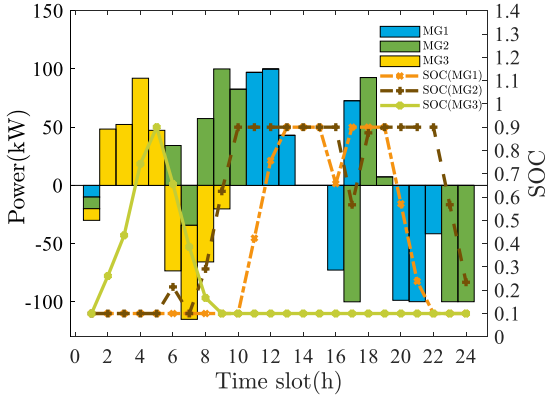


Fig. 9. Charging and Discharging Power and SoC Trends of the Physically Informed Model for ESS.

In Figure 10, it is apparent that the charging and discharging actions of the ESS in the three MGs follow a repetitive pattern over the 24 time steps, with both the charging and discharging powers maintained at 100 kW. Although this regularity ensures the voltage constraints of the system, it does not allow the system to explore economically optimal choices. Furthermore, this pattern can increase the operational burden of MGs. For example, when an MG experiences a peak load period, the ESS chooses to charge at that moment. If the internal supply power is not significantly greater than the load power, this charging behavior forces the MG to purchase more electricity from the DSO to meet the

ESS's charging requirements, consequently increasing the operational costs of the microgrid.

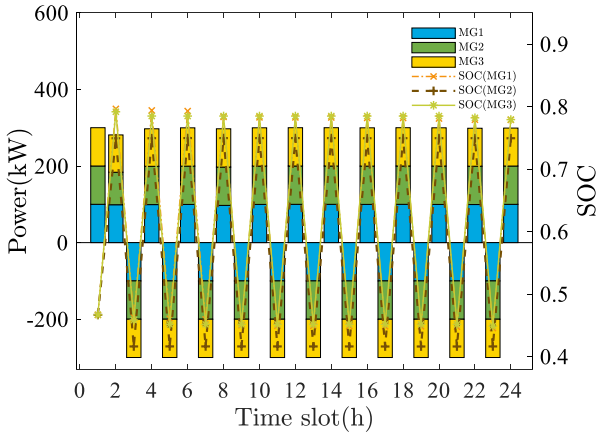


Fig. 10. Charging and Discharging Power and SoC Trends of the Penalty-Function-Based Model for ESS.

The comparison above has been made on the basis of the timing of the charging and discharging choices and the distribution of the charging and discharging power, analyzing the merits and drawbacks of the two methods. Now, let us further compare the two methods from the perspective of system operational costs. Figure 11 presents the total cost of the system for both methods at each time step, and Table III provides the specific values of Figure 11. It is evident from the table that only at time steps 11, 13, and 15, the cost based on this paper's model is higher than the cost based on the penalty function approach. Furthermore, when considering the total cost, the cost of the penalty-based system is $1266.2789 higher than the cost of the model proposed in this paper in the 24-time steps.

TABLE III
STATISTICAL RESULTS OF SYSTEM TOTAL COSTS OVER 24H FOR THE PROPOSED MODEL AND THE PENALTY FUNCTION-BASED MODEL

	Time	Proposed	Penalty Function-Based
Total cost of the system at each timestep ($)	1	42.929	265.481
	2	51.302	58.323
	3	59.814	98.668
	4	73.893	96.130
	5	93.296	196.745
	6	94.185	158.662
	7	115.418	189.511
	8	89.257	168.710
	9	77.923	95.004
	10	60.808	83.885
	11	76.300	52.042
	12	116.535	141.845
	13	147.960	112.286
	14	98.539	119.012
	15	123.466	104.998
	16	151.111	174.827
	17	184.205	289.461
	18	312.115	398.710
	19	282.875	347.957
	20	230.253	277.246
	21	196.513	275.719
	22	61.768	92.904
	23	46.416	170.025
	24	49.415	134.415
Total cost ($)		2836.296	4102.566

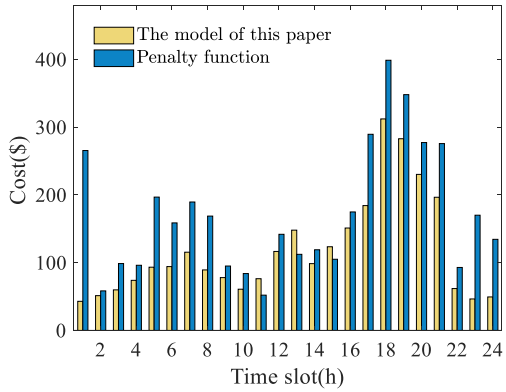


Fig. 11. Distribution of total costs over 24 hours for the proposed model and the penal function-based model.

In summary, compared to the penalty-based approach, the model proposed in this paper can save costs for the system and reduce operational burdens.

Figure 12 shows the voltage variation over 24 time steps in a bus voltage scenario based on the physical information model. Table IV provides the maximum and minimum values of each bus voltage in 24-time steps. The table reveals that among the 33 bus voltages, the maximum value is 1.035 and the minimum value is 0.945. Importantly, the fluctuation range of all these values falls within the set parameters for voltage deviation (0.90 to 1.10). This indicates that the physical information-based model proposed in this paper can ensure voltage safety during the system's operational process.

TABLE IV
MAXIMUM AND MINIMUM VALUES OF EACH BUS VOLTAGE OVER 24 HOURS

Bus	Vmax (.pu)	Vmin (.pu)
1	1	1
2	0.999	0.997
3	0.997	0.994
4	0.996	0.988
5	0.994	0.983
6	0.992	0.969
7	0.992	0.967
8	0.994	0.964
9	0.999	0.959
10	1.004	0.955
11	1.005	0.954
12	1.007	0.953
13	1.017	0.949
14	1.022	0.949
15	1.028	0.948
16	1.035	0.948
17	1.035	0.948
18	1.035	0.949
19	0.999	0.996
20	1.003	0.986
21	1.005	0.983
22	1.004	0.983
23	0.995	0.991
24	0.991	0.985
25	0.989	0.982
26	0.991	0.968
27	0.990	0.965
28	0.990	0.956
29	0.989	0.949
30	0.989	0.945
31	0.993	0.945
32	0.994	0.946
33	0.994	0.945

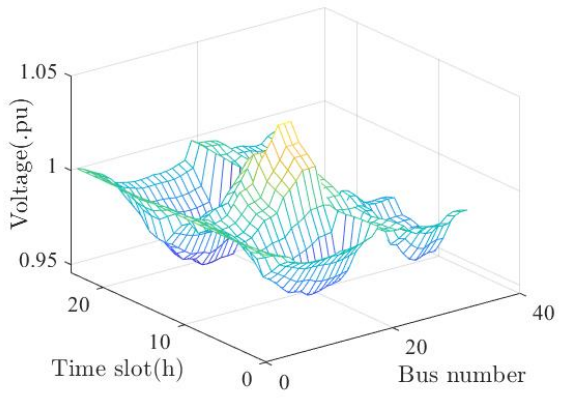


Fig. 12. Bus Voltage over 24 Hrs.

D. Comparison of Dynamic Electricity Pricing Scenarios

This section compares two scenarios: dynamic electricity pricing and time-of-use pricing. Time-of-use pricing involves dividing the day into several periods, such as peak, off-peak, and valley, and setting different electricity prices for each period. Table V shows the selling and buying electricity prices for DSO in different time periods over 24 hours under the time-of-use pricing scenario. The buying price in the table is the price at which MGs purchase electricity from the DSO, while the selling price is the price at which the DSO sells electricity to MGs. The table shows that the buying prices are divided into six periods over 24 hours. The first period is from 1 to 7 a.m., with a price of $0.4. The second period is from 8 to 11 a.m., priced at $0.75. The third period is from 12 to 2 p.m., with a price of $1.2. The fourth period is from 3 to 6 p.m., priced at $0.75. The fifth period is from 7 to 10 p.m., with a price of $1.2. The sixth period is from 11 p.m. to midnight, priced at $0.4. The selling price is set uniformly at $0.2 for each period.

TABLE VI
COSTS OF MGS AND DSO AT EACH TIME STEP IN THE TIME-OF-USE PRICING SCENARIO

	Time (h)	MG1	MG2	MG3	DSO
Time-of-use pricing scenario, the cost at each time step ($)	1	23.273	46.562	-19.582	-15.038
	2	13.769	63.877	-17.549	-3.741
	3	25.275	32.323	-14.785	25.550
	4	49.608	42.632	-10.050	-1.143
	5	79.049	99.277	55.869	-138.251
	6	31.651	165.793	92.301	-107.255
	7	11.475	64.424	74.280	66.503
	8	0.881	-46.725	27.230	111.904
	9	0.507	-60.069	79.846	63.132
	10	-19.917	-72.603	155.848	15.461
	11	-26.259	-83.357	94.115	146.839
	12	-3.019	-86.936	159.359	92.188
	13	-0.017	-71.088	296.306	-27.708
	14	1.088	-42.076	223.518	-25.642
	15	31.177	-10.429	168.819	8.761
	16	72.585	24.090	169.528	19.418
	17	97.073	157.093	156.497	-46.841
	18	78.279	217.925	135.724	15.156
	19	73.206	185.077	67.946	78.088
	20	37.385	160.960	13.068	110.049
	21	37.265	103.412	-8.180	155.009
	22	23.720	69.370	-14.148	67.834
	23	17.483	55.572	-13.676	-3.090
	24	18.088	61.505	-13.126	7.919
Total Cost ($)		671.449	1076.609	927.359	615.102

Table VI shows the operational costs of MGs and DSO in 24-time steps in the time-of-use pricing scenario. Analyzing the data in this table, we find that MG1 is profitable during the 10 to 13 hour period but incurs losses in other time slots, resulting in a total operational cost of $671.449. MG2 shows profits from 8 to 15 hours and losses during other periods, leading to a total operational cost of $1076.609. MG3 is profitable in the 1 to 4 and 21 to 24 hour segments, while experiencing losses at other times, culminating in a total operational cost of $927.359. The DSO makes profits during 1 to 2, 4 to 6, 13 to 14, and 23 hours, but faces losses in the remaining periods, bringing its total operational cost to $615.102.

However, with the continuous development and improvement of smart grids, advanced artificial intelligence algorithms and communication technologies have made real-time interaction between electricity companies and consumers possible. Building on this, dynamic electricity pricing has been proposed to further shorten the time periods, dividing the electricity scheduling cycle into several intervals, with each interval potentially as short as 1 hour or even less, and establishing different electricity prices for each interval.

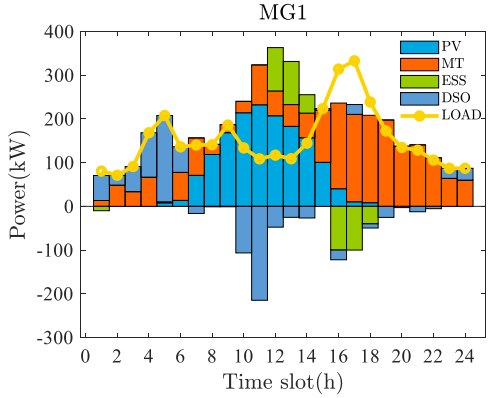
Table VII presents the operational costs of the MGs and the DSO in 24-time steps under the dynamic electricity pricing scenario. The table reveals that MG1 is profitable at 10, 11, and 13 hours, but incurs losses during other times, resulting in a total operational cost of $498.851. MG2 shows profits from 8 to 15 hours and losses in the remaining periods, leading to a total operational cost of $643.16. MG3 is profitable between 1 to 2 and 21 to 24 hours, while experiencing losses at other times, resulting in a total operational cost of $927.539. The DSO makes profits during 4 to 6 and 16 to 17 hours, but faces losses in the other periods, bringing its total operational cost to $767.1.

TABLE VII
COSTS OF MGS AND DSO AT EACH TIME STEP IN THE DYNAMIC ELECTRICITY PRICING SCENARIO

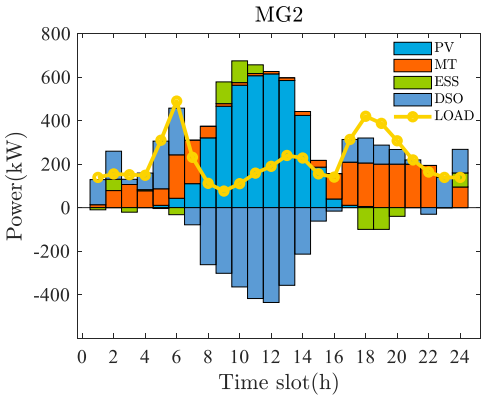
	Time (h)	MG1	MG2	MG3	DSO
Dynamic electricity pricing scenario, the cost at each time step ($).	1	11.619	29.132	-8.974	11.161
	2	12.174	37.396	-1.694	3.425
	3	18.137	36.054	0.282	5.338
	4	41.365	35.310	7.398	-10.118
	5	50.410	80.499	58.406	-96.016
	6	26.723	133.716	33.071	-99.325
	7	11.662	15.969	5.106	82.680
	8	1.951	-14.171	4.339	97.136
	9	1.660	-29.634	25.743	80.163
	10	-8.140	-38.187	54.684	52.449
	11	-1.930	-45.936	52.393	71.771
	12	1.121	-43.137	98.667	59.881
	13	-1.671	-34.489	133.305	50.816
	14	0.712	-18.294	119.029	-2.909
	15	25.760	-1.794	84.863	14.639
	16	50.190	19.110	85.079	-3.273
	17	107.631	51.140	79.838	-54.402
	18	58.152	141.318	62.899	49.746
	19	40.845	107.394	39.787	94.848
	20	3.679	81.161	9.082	136.330
	21	2.611	54.901	-0.283	139.282
	22	10.101	38.284	-5.734	19.118
	23	17.040	3.768	-5.598	31.204
	24	16.981	3.650	-4.392	33.156
Total Cost ($)		498.851	643.160	927.359	767.100

When comparing the operational cost results from both tables, it is evident that the operational costs of MGs and DSO in the time-of-use pricing scenario are consistently higher than those under the dynamic electricity pricing scenario. The reason lies in the fact that dynamic pricing links user-side prices with the electricity market in real time, reflecting the changing costs of power supply during different periods. Furthermore, the data-driven dynamic pricing scenario effectively balances supply and demand in the electricity market, significantly protecting the interests of both electricity companies and consumers. This approach fully leverages the economic leverage of electricity pricing.

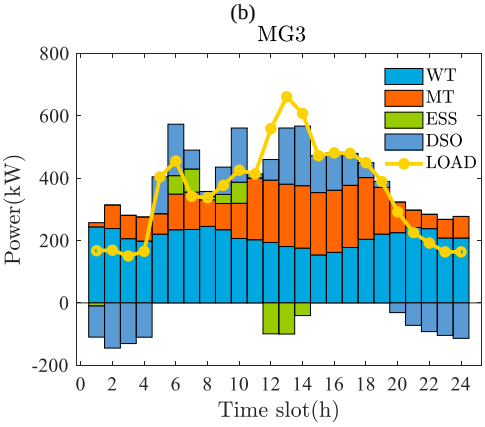
Figure 13 presents the dispatch results of three MGs under TOU pricing scenario. From Figures 13 (a) and (b), it can be observed that, within the initial time steps, MG1 and MG2 engaged in energy transactions with the DSO due to weak or no sunlight and a demand exceeding the MGs' own energy supply capability. Consequently, MG1 and MG2 purchased energy from the DSO to meet their load demands. Subsequently, from 7-15h, the intensity of sunlight gradually increased, and the PV systems began generating a surplus of energy. At this point, MG1 and MG2 traded the excess energy with the DSO while charging their ESS. Later, as sunlight diminished and the capability of the PV systems decreased, the load demand sharply increased. At this juncture, MT and ESS were utilized to provide energy to users. During the periods of 1-4h and 21-24h in Figure 13 (c), the WT generated enough power to meet the full load demand, allowing the operator of MG3 to sell the surplus energy to the DSO, thus increasing MG3's revenue. In the period of 5-20h, as the load gradually reached its peak, the generation power of MT was insufficient to meet the demand. Hence, the operator of MG3 dispatched MT and engaged in energy transactions in the electricity market to participate in the internal demand response, alleviating the pressure on supply. However, the dispatch results for MG3 indicate that the ESS was set to charge during 6-10h, an action that would exacerbate the operational costs and reduce economic efficiency for MG3. The above is an analysis of the dispatch results of the three MGs under the TOU pricing scenario. However, compared to the dynamic pricing dispatch results, the degree of MT participation in energy supply under the TOU scenario was greater for the three MGs, and since the operational cost of MT is a component of the MGs' objective function, it undoubtedly increases the operational costs and reduces economic efficiency.



(a)



(b)



(c)

Fig. 13. MG's dispatch results under TOU pricing scenario.

Figure 14 displays the charging and discharging power and the SoC trends of ESS under the TOU pricing scenario. The diagram clearly shows the 24-hour charging and discharging states and SoC trends of the three MGs' ESS. For MG1, the ESS chose to charge during 12-14h and to discharge during 16-18h. The ESS of MG2 selected charging times at 2, 9-11h, and 24h, and discharging periods at 3-6h and 18-20h. The ESS of MG3 opted to charge between 6-10h and discharge between 12-14h. Additionally, it can be observed that the SoC of each ESS varied within the range of 0.1 to 0.9.

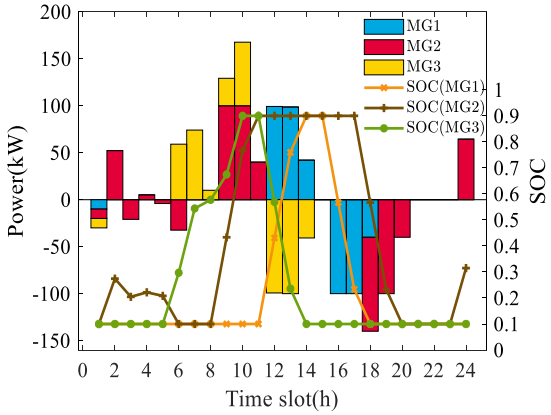


Fig. 14. ESS's char/discharging power and SoC trends under TOU pricing scenario.

E. Comparison of Energy Trading

This section compares the results of the internal energy trading in the system under different algorithms. Figure 15 illustrates the distribution of energy exchange between MGs and DSO under the MATD3 algorithm. In the figure, the positions above zero represent energy purchases by MGs from DSO, while the positions below represent energy sales

by MGs to DSO. This figure, in conjunction with Figure 16, indicates that MG1 produced a significant amount of energy from photovoltaic sources between 8 to 15 hours when the load was at its lowest. In contrast, during other periods, the intensity of sunlight decreased and the load was higher. Therefore, MG1 purchased energy from DSO during 1 to 7 hours and 16 to 22 hours and sold energy to DSO during 8 to 15 hours and 23 to 24 hours. Similarly, as shown in Figure 6, MG2 produced high power from photovoltaic sources between 8 to 15 hours when the load was low. Thus, MG2 sold energy to DSO during 8 to 15 hours and purchased energy during other periods. MG3, due to the wind power generation capacity fully covering the load between 1 to 3 hours and 21 to 24 hours, sold energy to DSO during these periods. However, at other times when it could not meet the demand, MG3 had to purchase energy from DSO to meet the load requirement. Figure 16 shows the distribution of energy transactions between MGs and DSO under the MADDPG algorithm, and the analysis process is similar to that above.

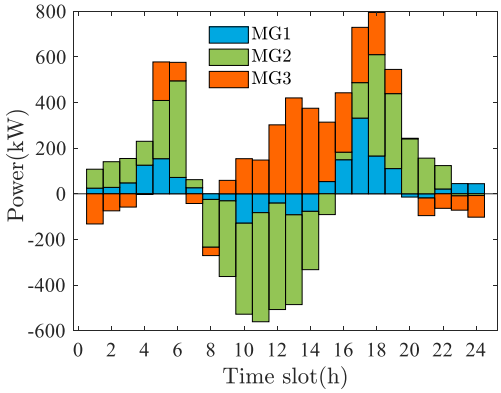


Fig. 15. Energy Exchange Results within the System Under the MATD3 Algorithm.

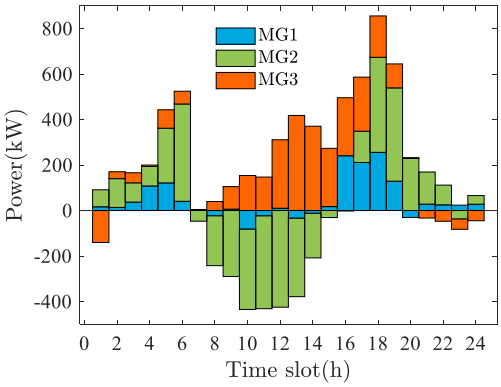


Fig. 16. Energy Exchange Results within the System Under the MADDPG Algorithm.

Table VIII provides specific figures for energy transactions between MGs and DSO under MATD3 and MADDPG optimization. From these values, it is evident that under MATD3, the volume of energy transactions between MGs and DSO is higher than that under MADDPG. This indicates that under MATD3, there is more extensive energy trading within the system, which greatly helps to enhance the vitality of the electricity market. Furthermore, frequent transactions between DSO and MGs in the electricity market can increase the penetration of renewable energy in the grid, reduce reliance on fossil fuels, and positively impact energy security and the promotion of a low-carbon economy.

TABLE VIII
ENERGY TRADING RESULTS IN THE ELECTRICITY MARKET

	MG1(kW)	MG2(kW)	MG3(kW)	Total(kW)
MATD3	1910.526	5226.161	3411.617	10548.304
MADDPG	1520.682	4851.472	3096.674	9468.828

VI. CONCLUSIONS

This article, through relevant test cases, has drawn the following conclusions regarding the proposed framework, algorithms, and safety model.

(1) The distributed bilevel optimization framework based on Stackelberg games can protect the privacy information of participants with limited information exchange. It can also coordinate the multi-entity relationships within the framework through Stackelberg game mechanisms. (2) Data-driven MATD3 is capable of saving costs for the system and produces optimized strategies by learning from the system's data. Compared to MADDPG, MATD3 has advantages in terms of cost savings for the system. The model based on physical information ensures that the capacity of the ESS and node voltage remain within safe ranges. (3) Compared to the penalty-based approach, the model proposed in this document allows agents to make optimal economic choices during the environmental exploration process, resulting in cost savings for the system. It also relieves the operational pressure of the MGs.

In summary, the framework, algorithms, and models proposed in this paper provide a secure and stable environment for interconnected systems while significantly improving the economic efficiency of the system. Future research can further optimize the market mechanisms between MGs and DSO in grid-connected modes and consider interactions among multiple microgrids.

REFERENCES

[1] X. Chen, W. Dong, and Q. Yang, "Robust optimal capacity planning of grid-connected microgrid considering energy management under multidimensional uncertainties," *Applied Energy*, vol. 323, no. 4, 2022.

[2] H. H. Goh, S. Shi, X. Liang, D. Zhang, W. Dai, H. Liu, S. Y. Wong, T. A. Kurniawan, K. C. Goh, and C. L. Cham, "Optimal energy scheduling of grid-connected microgrids with demand side response considering uncertainty," *Applied Energy*, vol. 327, no. 2, 2022.

[3] CHEN G, WANG X, MO S, et al. Multi-objective Power Flow Optimization Based on Improved Hybrid Crow Search Algorithm: A Novel Approach [J]. Engineering Letters, vol. 30, no. 4, 2022.

[4] J. Zhao, W. Wang, and C. Guo, "Hierarchical optimal configuration of multi-energy microgrids system considering energy management in electricity market environment," *International Journal of Electrical Power & Energy Systems*, vol. 144, no.4, 2023.

[5] Z. Wang, B. Chen, J. Wang et al., "Decentralized energy management system for networked microgrids in grid-connected and islanded modes," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 1097–1105, 2016.

[6] Y. Du, Z. Wang, G. Liu, X. Chen, H. Yuan, Y. Wei, and F. Li, "A cooperative game approach for coordinating multi-microgrid operation within distribution systems," *Applied Energy*, vol. 222, no. 3, pp. 383–395, 2018.

[7] P. Xie, Y. Jia, H. Chen, J. Wu, and Z. Cai, "Mixed-stage energy management for decentralized microgrid cluster based on enhanced tube model predictive control," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 3780–3792, 2021.

[8] CHEN G, LI J, XU Y, et al. Optimal Configuration of Renewable Energy DGs Based on Improved Northern Goshawk Optimization

Algorithm Considering Load and Generation Uncertainties [J]. Engineering Letters, vol. 31, no. 2, pp. 511-530, 2023.

[9] M. Yousif, Q. Ai, Y. Gao, W. A. Wattoo, Z. Jiang, and R. Hao, "An optimal dispatch strategy for distributed microgrids using PSO," *CSEE Journal of Power and Energy Systems*, vol. 19, no. 6, pp. 724–734, 2020.

[10] M. Khosravi, H. Azarinfar, and S. A. Nejati, "Microgrids energy management in automated distribution networks by considering consumers' comfort index," *International Journal of Electrical Power & Energy Systems*, vol. 139, no. 3, 2022.

[11] A. Nawaz, J. Wu, and C. Long, "Distributed optimal energy scheduling for grid connected multi-microgrids with architecturized load characteristics," *Energy Reports*, vol. 8, no. 2, pp. 11259–11270, 2022.

[12] Y. Du and F. Li, "A hierarchical real-time balancing market considering multi-microgrids with distributed sustainable resources," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 72–83, 2020.

[13] Z. Wang, B. Chen, J. Wang, M. M. Begovic, and C. Chen, "Coordinated energy management of networked microgrids in distribution systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 45–53, 2015.

[14] X. Zheng, Q. Li, C. Bai, Y. Nie, and C. Huang, "Energy trading management based on stackelberg game theory to increase independence of microgrids," *Energy Reports*, vol. 139, no. 3, pp. 771–779, 2022, doi: 10.1016/j.egyr.2022.10.168.

[15] R. Zhang, X. Li, L. Fu, T. Jiang, G. Li, and H. Chen, "Network-aware energy management for microgrids in distribution market: A leader-followers approach," *Applied Energy*, vol. 332,no. 2, 2023.

[16] Ö. Erol and Ü. B. Filik, "A stackelberg game approach for energy sharing management of a microgrid providing flexibility to entities," *Applied Energy*, vol. 316, no.2, 2022.

[17] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1066–1076, 2020.

[18] B. Zhang, W. Hu, X. Xu, T. Li, Z. Zhang, and Z. Chen, "Physical-model-free intelligent energy management for a grid-connected hybrid wind-microturbine-PV-EV energy system via deep reinforcement learning approach," *Renewable Energy*, vol. 200, no. 7, pp. 433–448, 2022.

[19] Y. Ji, J. Wang, J. Xu, and D. Li, "Data-driven online energy scheduling of a microgrid based on deep reinforcement learning," *Energies*, vol. 14, no. 8, 2021.

[20] Y. Ji, J. Wang, J. Xu, X. Fang, and H. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, 2019.

[21] L. Fan, J. Zhang, Y. He, Y. Liu, T. Hu, and H. Zhang, "Optimal scheduling of microgrid based on deep deterministic policy gradient and transfer learning," *Energies*, vol. 14, no. 3, 2021.

[22] Y. Liu, Y. Wang, Y. Li, H. B. Gooi, and H. Xin, "Multi-agent based optimal scheduling and trading for multi-microgrids integrated with urban transportation networks," *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 2197–2210, 2021.

[23] B. Zhang, W. Hu, A. M. Ghias, X. Xu, and Z. Chen, "Multi-agent deep reinforcement learning based distributed control architecture for interconnected multi-energy microgrid energy management and optimization," *Energy Conversion and Management*, vol. 277, no.4, 2023.

[24] J. Li, T. Zhou, H. Keke, H. Yu, H. Du, S. Liu, and H. Cui, "Distributed quantum multiagent deep meta reinforcement learning for area autonomy energy management of a multiarea microgrid," *Applied Energy*, vol. 343,no. 5, 2023.

[25] X. Fang, Q. Zhao, J. Wang, Y. Han, and Y. Li, "Multi-agent deep reinforcement learning for distributed energy management and strategy optimization of microgrid market," *Sustainable Cities and Society*, vol. 74, no.4, 2021.

[26] C. Guo, X. Wang, Y. Zheng, and F. Zhang, "Optimal energy management of multi-microgrids connected to distribution system based on deep reinforcement learning," *International Journal of Electrical Power & Energy Systems*, vol. 131 ,no.6 2021.

[27] X. Liu, S. Li, and J. Zhu, "Optimal coordination for multiple network-constrained vpps via multi-agent deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3016–3031, 2023.