MADDPG-Deep-QNet: Multi-Agent Deep Reinforcement Learning for Day-Ahead Power Balance Optimization

Shujun Wang[1], Feng Liu[2,*], Yumeng Zhang[2], Dengyi Huang[1], Danlei Xu[1]
[1]North China Branch of State Grid Corporation of China+Power Dispatch Control Center, Beijing, China. 100053
[2]Beijing Qingneng Internet Technology Co., LTD+Intelligent Products Department, Nongda South Road, Haidian District, Beijing, China

Keywords: Multi-agent deep reinforcement learning (MADRL), grid stability, energy demand, power delivery, Multi-Agent Deep Deterministic Policy Gradient Driven Deep Q-Network (MADDPG-Deep-QNet)

Received: May 30, 2025

A reliable overall power supply on the grid depends on day-ahead power planning when electricity demand changes. Traditional optimization methods struggle to account for the dynamic nature and complexity of power supply systems. The purpose of this research is to offer a multi-agent deep reinforcement learning (MADRL) approach for optimizing day-ahead power balance strategies to ensure steady power supply capacity while discussing the problems of dynamic and complex energy grids. The dataset includes historical data on power use and generation, as well as real-time demand, renewable energy outputs, and system stability indices. Data are cleaned and normalized to account for missing values and outliers, ensuring consistency and accuracy. The Fast Fourier Transform (FFT) converts time-series power data into frequency components, enabling identification of demand and generation patterns. This aids in extracting relevant features for optimizing day-ahead power balance strategies. The research aims to develop a proposed Multi-Agent Deep Deterministic Policy Gradient Driven Deep Q-Network (MADDPG-Deep-QNet) model combines Multi-Agent Deep Deterministic Policy Gradient with Deep Q-Network principles, enabling multiple agents to coordinate and optimize power source allocation, ensuring stable day-ahead power supply, reduced costs, and improved grid reliability in the proposed method. The MADDPG-Deep-QNet strategy outperforms existing optimization techniques, resulting in significant energy cost savings and grid stability, with a load forecasting MAPE of 11.05, along with better MAE, MSE, RMSE, and R^2. In terms of power supply capacity, the model outperforms existing methods. This research highlights MADRL's potential for optimizing day-ahead power balance techniques, offering a scalable solution to improve grid stability and ensure continuous power delivery.

Povzetek: Analizirano je dnevno uravnoteževanje elektroenergetskega sistema ob visoki negotovosti obnovljivih virov. Predlagn je hibridni večagentni pristop MADDPG-Deep-QNet, ki združuje FFT-izluščene značilke ter kombinacijo kontinuiranega (MADDPG) in diskretnega učenja (DQN) za usklajevanje virov. Rezultati kažejo zmerno izboljšanje napovedi in stroškov.

1 Introduction

A primary problem in power networks is ensuring that all the electricity being generated is used by consumers. Through Power planning for the day, the grid is secured, costs are lowered, and the occurrence of power failures is stopped. Figuring out daily power needs has become a challenge because renewable sources of energy cannot be depended upon [1]. Data-driven energy methods are most effective when business leaders can make good forecasts, react quickly, and adapt smoothly to new challenges [2]. Usually, producers rely on unit commitment (UC) and economic dispatch (ED) to plan the best schedules for the ordinary day-ahead. Both forecasting and decision-making have been improved thanks to Machine learning (ML) and deep learning (DL) [3]. Methods such as support vector machines, recurrent neural networks and convolutional architectures are used in the energy sector for forecasting loads and generation, assigning reserve quantities and valuing energy [4].Furthermore, when the environment is complex, agents such as prosumers, power plants, and storage devices must collaborate under uncertainty [5].It is advised to adopt a MADRL framework when deciding on the power supply for the prior day, highlighting innovation in data-driven energy services [6]. In this approach, the owners of generators and storage systems make independent decisions to achieve their local objectives while contributing to a reliable grid state [7,8]. Deep reinforcement learning enables agents to respond to dynamic conditions and cooperate in networked

environments, crucial for microgrid and distributed generation management, using a partially observable Markov decision process [9, 10]. Despite recent advances, traditional ML and DL methods are limited by their centralized nature, lack of inter-agent coordination, and poor scalability in complex, uncertain environments. The aim of this research is to develop a MADDPG-Deep-QNet model that integrates FFT-based feature extraction with multi-agent deep reinforcement learning to optimize day-ahead power balance, ensuring stable power supply, reducing costs, improving grid reliability, and effectively managing complex energy systems.

- Can the proposed MADDPG-Deep-QNet model outperform existing multi-agent deep reinforcement learning methods in optimizing day-ahead power balance scheduling?
- How does the integration of FFT-based feature extraction impact the accuracy of load forecasting and power supply optimization?
- To what extent can the MADDPG-Deep-QNet model improve grid stability and reduce operational costs under dynamic and complex energy grid conditions?

1.1 Contributions of the research

- Introduces MADDPG-Deep-QNet – a hybrid model combining Multi-Agent Deep Deterministic Policy Gradient and Deep Q-Network to optimize day-ahead power balance strategies.
- Incorporates FFT-based feature extraction from time-series power data to capture demand and generation patterns for improved forecasting and optimization.
- Enables coordinated multi-agent decision-making for efficient power source allocation, enhancing stability, reducing operational costs, and improving grid reliability.
- Demonstrates superior performance over traditional optimization methods with lower MAPE (11.05) and improved MAE, MSE, RMSE, and R^2 metrics.
- Offers a scalable and adaptable framework for handling dynamic and complex energy grid conditions while ensuring continuous power delivery.

2 Related works

To utilize the multi-agent deep reinforcement learning (MA-DRL) framework presented in [11], an optimal energy management plan for multi-energy carrier microgrids (MECMs). The examination in these networks aims to reduce power loss and prevent large voltage fluctuations. Using an experience augmented multi-agent actor-critic (EA-MAAC) proposed in [12] The research explores fast-timescale adjustment of smart inverters and electric vehicles, mixed-integer optimization for power factor correction, a multi-timescale hybrid electrical control approach, security and privacy-aware scheduling in renewable energy hubs, and selecting optimal day-ahead plans for active distribution networks[14]. The method worked successfully on the IEEE33 case by offering real-time adjustments, reducing the reliance on trial-and-error strategies in costly power systems. Innovation in data-driven energy services was further exemplified by a heterogeneous multi-agent proximal policy optimization (H-MAPPO) system proposed by [15], where each agent was responsible for managing its designated generation areas. To make the power grid more stable by matching energy supply and demand with a Voltage Capability Incentive-Based Demand Response (VC-IBDR) structure through a multi-agent system proposed in [16]. Simulation results on a supplyline serving 25 households, the system maintains voltage readings between 0.96 and 1.0 per unit while lowering average usage by 4.20% and the highest usage by 10.41%.Given the declining reliability of modern grids, the study in[17] seeks to empower community-based virtual power plants (cVPPs)distribute quick auxiliary solutions. A collaborative bidding framework was proposed using the multi-agent deep deterministic policy gradient (MADDPG) algorithm to manage decision-making cycles across these sectors, developed in[18]. Anexperiment to five-node IEEE system has proven this model is effective for market learning and quicker convergence with greater revenue gains. The power flow management in Integrated Energy Systems (IESs) through flexible that interpretable pareto optimization method, Cooperative Multi-Agent Multi-Objective Reinforcement Learning (C-MAMORL) setup,as described in [19]. This setup employed value learning and reward learning signal systems are used to address a multi-objective Markov decision process. The investigation found that the model closely approximated the Pareto frontier, maintaining safety limits and exceeding traditional DRL techniques in making multi-dimensional choices. Through the integration of air conditioners and power and heating plants, this effort seeks to enhance energy management in community-scale systems [20]. Under time-of-use pricing and variable solar and wind power generation, hybrid electricity and thermal energy storage are scheduled using a Soft Actor-Critic (SAC) deep reinforcement learning approach. The results demonstrate cost reductions of up to 27.41% in electricity and 31.83% in gas consumption, surpassing the performance of the deterministic actor-criticizer approach. Table 1 shows that the summary of multi-agent energy optimization methods.

Table 1: Summary of related multi-agent energy optimization methods

References	Method	Problem Addressed	Dataset / Context	Reported Results / Metrics
[11]	MA-DRL framework	Optimal energy management in multi-energy carrier microgrids (MECMs)	MECM simulations / microgrid context	Reduced power loss and mitigated large voltage fluctuations (qualitative)
[12]	Experience-augmented MA actor-critic (EA-MAAC) + mixed-integer optimization	Fast-timescale adjustment for smart inverters & EVs; infrequent timescale control for converters and backup	Multi-timescale microgrid control simulations	Multi-timescale hybrid control approach — improved responsiveness (qualitative)
[13]	DRL with security/privacy considerations	Scheduling energy in energy hubs (EHs) with renewables and carbon trading	Energy hub contexts with economic constraints	Reduced economic cost (metric details not specified)
[14]	(DRL-based) day-ahead planning	Day-ahead plan selection under renewable uncertainty	IEEE33 node testbed	Real-time adjustments; reduced reliance on trial-and-error (qualitative)
[15]	Heterogeneous multi-agent PPO (H-MAPPO)	Distributed generation management by region	Heterogeneous agent regions / simulated grid	Improved local management (metrics not specified)
[16]	VC-IBDR multi-agent system	Demand response to stabilize voltage & reduce peaks	25-household feeder simulation	Voltage maintained 0.96–1.00 p.u.; average usage −4.20%; peak usage −10.41%
[17]	cVPP (community virtual plants) approach	Provide fast auxiliary services as grid reliability declines	Community-level distributed systems	Enhanced auxiliary service provisioning (qualitative)
[18]	MADDPG for collaborative bidding	Decision-making in cVPP market participation	Five-node IEEE market simulation	Faster convergence, increased revenue (qualitative)
[19]	C-MAMORL (cooperative multi-objective RL)	Multi-objective power flow management in IES	Integrated energy systems simulations	Approximated Pareto frontier; preserved safety limits (qualitative)
[20]	SAC (Soft Actor-Critic)	Scheduling hybrid electric + thermal storage under TOU pricing & variable renewables	Community-scale electricity & thermal systems	Electricity cost ↓ 27.41%; Gas consumption ↓ 31.83%

2.1 Problem statement

The MADDPG-Deep-QNet model is an adaptable multi-agent reinforcement learning system that combines continuous and discrete learning to efficiently coordinate thermal, renewable, and battery resources for real-time dispatch and system stability [11]. This model overcomes limitations in existing methods, enabling effective cooperation among diverse power sources [15].

3 Methodology

This investigation develops a multi-agent deep reinforcement learning framework to optimize Power balance tactics for the coming day and ensure reliable power supply capacity. Information streaming is ancient, including power request, renewable generation, and network stability metrics, is collected, handled, and normalized. FFT extracts temporal features from datasets, modeling power system as multi-agent environment with

thermal, renewable, and battery storage units. MADDPG-Deep-QNet approach learns coordinated policies for supply and demand balance. Figure1 illustrates the flow of approaches that contribute to power supply capacity.

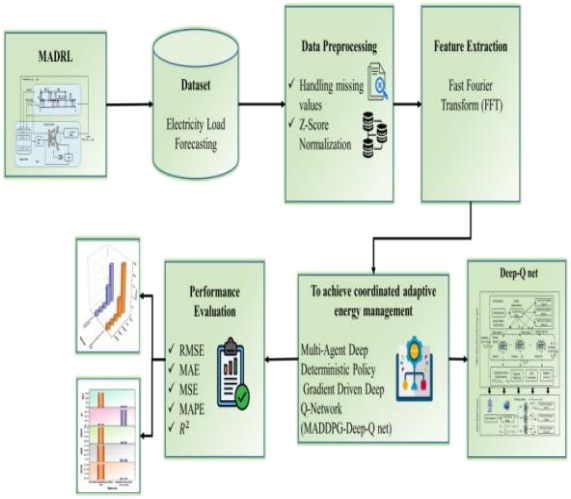

Figure 1: Proposed flow

3.1 MADRL

The investigation introduces an MADRL method for day-ahead power balance plans in microgrid and distributed generation management. It improves grid efficiency, flexibility, and reliability by coordinating agent-based decision-making and adjusting the best price strategy for multiple agencies, maximizing expected revenues.While the earlier description in this study models agents as direct physical entities—thermal plants, renewable generators, and battery storage units Figure 2 “Architecture of multi-agent deep reinforcement learning,” p. 4 depicts a higher-level hierarchical architecture. In this figure, the core agents are Pricing Agents and Load Serving Entities (LSEs), whose objectives are profit maximization and social welfare maximization, respectively, through pricing mechanisms. This represents an abstraction layer where market-oriented decision-making governs the operational strategies of the underlying physical resources, thereby integrating economic and technical objectives within the multi-agent framework.

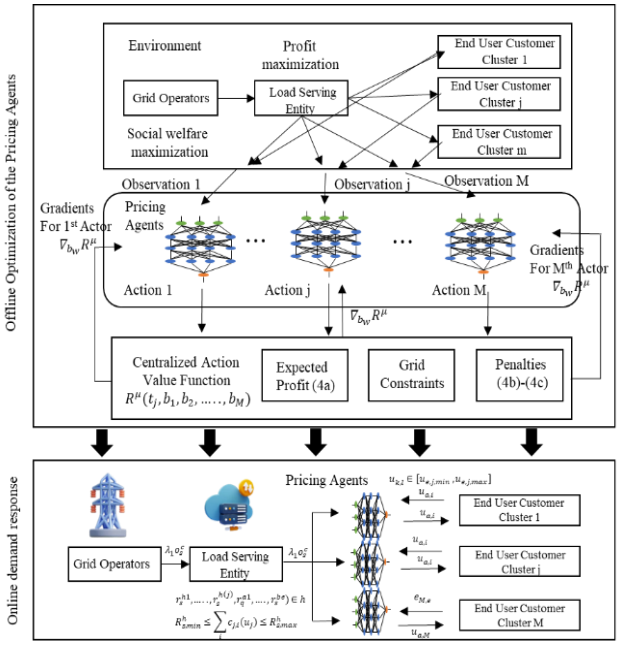

Figure 2: Electricity load forecasting weekly prediction accuracy overview

3.2 Dataset

The Kaggle Electricity Load Forecasting dataset contains over 40,000 hourly records (Jan 2015–Jun 2020) from Panama’s grid operator, integrating historical post-dispatch loads, weekly pre-dispatch forecasts, weather variables (temperature, humidity, precipitation, wind speed), and calendar factors (holidays, school periods). Data from multiple sources is merged into a continuous time-indexed CSV, enabling direct use in forecasting models. Its hourly granularity captures short-term consumption dynamics, while weather and calendar features account for demand variability. For day-ahead power balancing, the dataset’s richness supports accurate load forecasting, facilitating optimal generation scheduling, storage management, and supply–demand balancing in multi-agent decision-making frameworks under realistic operating conditions.

Source:

https://www.kaggle.com/datasets/saurabhshahane/electricity-load-forecasting

3.3 Data Pre-processing using handling missing values

The day-ahead power balance optimization process involves pre-processing data to maintain consistency, support microgrid and distributed generation management, and standardize variables. The dataset is reliable, complete, and ready for multi-agent deep reinforcement learning training, ensuring effective power supply strategies.

3.4 Z- Score normalization

Through converting feature variables employing the average and typical variation of the corresponding characteristic, this statistical normalization technique tackles the problem of outliers. Specifically, the following Eq. (1) is applied to convert the original values into their normalized counterparts.

$$\nu' = \frac{\nu - \mu}{\sigma} \quad (1)$$

Where,

μ - mean value of the designated feature

σ - standard deviation of the considered feature

The preprocessing involved multiple steps to ensure data quality before model training.Missing values were handled by removing duplicate entries, correcting outliers, and standardizing both time-series timestamps and energy source labels to maintain consistency. Outliers were addressed using Z-score normalization, which transformed each feature using its mean and standard deviation to reduce skew and ensure comparability. This normalization helped in mitigating the influence of extreme values and aligning different scales of variables.

3.5 Feature Extraction using FFT

Key information for power balance modelling, including forecasts of energy demand, trends in renewable generation, and grid status, is obtained through FFT.Using this technique, data is processed in the frequency domain, making it easier for agents to anticipate changes and make better energy scheduling decisions in a changing environment.The FFT decomposes thedata into real and signal components and represents it in an oscillation spectrum. Eqs(2) and (3) can be used for determining the FFT:

$$E(w,z) = \sum_{n=0}^{N-1}\sum_{m=0}^{M-1} e(n,m) f^{-(j\times 2\times\pi\left(w\frac{n}{M}+z\frac{m}{M}\right))} \quad (2)$$

$$e(w,z) = \frac{1}{N.M}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} E(n,m) f^{-(j\times 2\times\pi\left(w\frac{n}{M}+z\frac{m}{M}\right))} \quad (3)$$

The signal at the location $(n,m)$is represented by $e(n,m)$, the function for expressing the data in the modulation domains related to positions w and z is $E(w,z)$, the image's dimensions are$N \times M$, and j was $sqrt(-1)$.Figure3 shows the output of FFT.

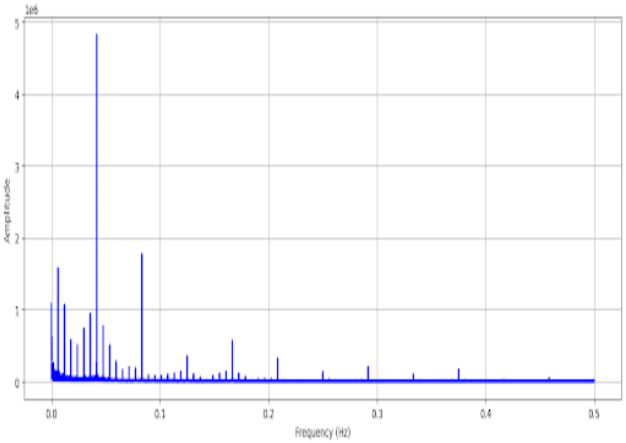

Figure 3: Power data frequency spectrum showing dominant peaks

FFT extracts dominant frequency patterns from load, generation, and storage data, enhancing agents' state representation. These features, combined with time-domain inputs, enable MADDPG-Deep-QNet agents to recognize periodic trends, anticipate fluctuations, and make coordinated, optimal power allocation decisions for improved day-ahead power balance accuracy and stability. The section on FFT for electricity load forecasting describes a 1D time-series input, supported by Figure 3 showing a 1D frequency spectrum. However, the provided equations (Eqs. 2 and 3) correspond to a 2D Discrete Fourier Transform, referencing variables for image dimensions (n, m, N × M). This inconsistency suggests a mismatch between the described 1D data and the 2D mathematical formulation, potentially causing confusion about the actual FFT implementation used.

3.6 MADDPG-Deep-QNet

The MADDPG-Deep-QNet framework leverages the strengths of both algorithms to address mixed-action and coordination challenges in day-ahead power balance optimization. MADDPG enables continuous control for precise power allocation via decentralized actors and a centralized critic, enhancing inter-agent cooperation. Deep-QNet supports discrete decision-making with stable value estimation and efficient exploration. Their integration allows simultaneous optimization of continuous outputs and discrete schedules, improving adaptability to renewable uncertainty, accelerating convergence, and delivering higher forecasting accuracy, lower operational costs, and greater grid stability than either method alone.The Figure 4 illustrates the MADDPG-Deep-QNet workflow, showing interaction between environment and agents, data processing, decision-making, and learning components to optimize coordinated multi-agent strategies for day-ahead power balance.

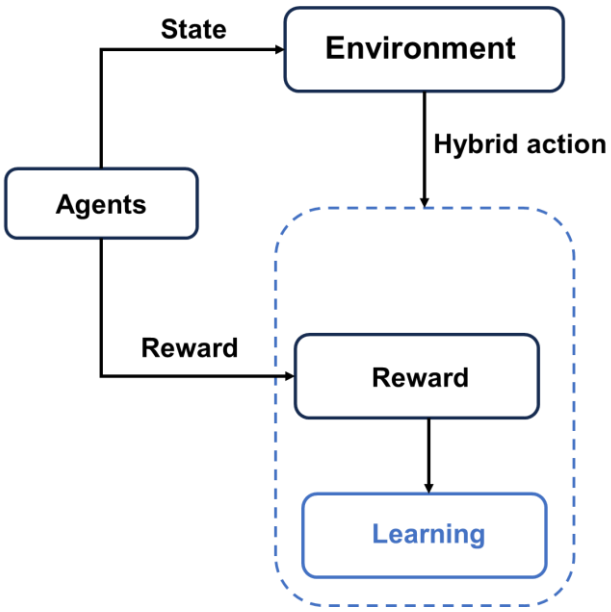

Figure 4: Proposed MADDPG-Deep-QNet framework for coordinated power optimization

3.6.1 Deep-QNet

Deep-QNet helps agents select optimal power dispatch decisions in a discrete action space, enabling coordinated learning for efficient day-ahead scheduling. The master node responds to network state-related data by scheduling activities. After a state detection, slaves and master communicate, receiving a quadruple experienced sequencing DJ (Figure 5). A mini-batch is selected to develop power use and generation from the replaying buffers.

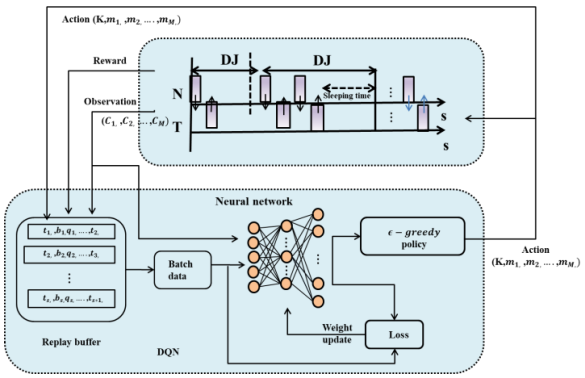

Figure5:Energy scheduling and optimization using DQN framework

- **Conditions**

A vector representing the remaining lifespan of scheduled "packets" for the M slave nodes at time s defines the system state, as given in Eq. (4):

$$t_2 = (C_1, C_2 \ldots\ldots, C_M) \tag{4}$$

Here, $C_j = (c_1, c_2 \ldots\ldots, c_l)$ denotes the time left for the first l packets in the jth slave's queue. The residual lifespan of a packet is determined by the time until the maximum delay since it entered the queue. While increasing l could capture more detail, the state space S would expand exponentially. This formulation master node tracking packet lifetimes in multiple queues is characteristic of communication network scheduling rather than power dispatch. Despite later redefining m_c as "quantity of energy output" in the QoS score, the terminology and structure here (packets, queues, lifespans) remain rooted in data transmission concepts.

Steps

At time t, a procedure $b_s \in \mathcal{A}$ consists of two elements: the duration of the connection interval (CI) and the number of packets delivered by each slave during that CI. This is again reflective of packet-based scheduling systems in networking, rather than typical continuous or discrete power allocation decisions in day-ahead energy planning.

$$b_s = (K, m_1, m_2 \ldots\ldots, m_M) \tag{5}$$

The total duration of the connection interval (CI) is denoted by K, and the number of packets transferred by each of the M slave nodes during the CI are represented by $m_1, m_2 \ldots\ldots, m_M$. Since each slave node can observe at most five packets, m_j is constrained to the range 0–5. The variable K represents the CI length in seconds, with each CI typically referenced by the index j (the CI index).

Reward

In its original form, this Deep-QNet configuration is a DQN-based packet scheduler that seeks to satisfy latency constraints while optimizing throughput — a formulation that is standard in communication network scheduling. The feedback mechanism combines a Quality of Service (QoS) parameter with the CI length to form the reward q_s, as expressed in Eq. (6).

$$q_s = (1 + D_{jc}) \times \prod_{j=1}^{M} rj \tag{6}$$

D_{jc} is the CI index the period s, and rj is the jth slave's QoS score at that same time. rj is defined asEq. (7)

$$rj = \begin{cases} 0, & if\, m_c \neq 0 \\ 1/2 & if\, m_j = 0, m_c = 0 \\ 1, & otherwise \end{cases} \tag{7}$$

Where the quantity of energy output, for the jth slave in the state t_s is indicated by m_c. In (7), the reward function is intended to rise as the CI lengthens and packet losses decrease.

3.6.2 MADDPG

The MADDPG algorithm allows groups of agents, representing various power sources, to develop strategies together in a common grid environment. Using continuous actions added by each agent and a centralized critic, they cooperate to keep the next-day power balance and make future energy supplies more reliable.Although each agent is modeled as a DDPG agent, throughout training, the agents share states and actions.

Let μ_o and R_O represent the actor and critic networks of agent o, and let $\theta_o^{\mu'o} and \theta_o^{R_o}$ denote their respective network parameters. Before training, μp and R_O are initialized, with starting weights assigned as $\theta_o^{\mu_o} \leftarrow \theta_o^{R_o}$ and $\theta_o^{\mu'o} \leftarrow \theta_o^{\mu\ o}$, chosen at random.

For each agent ooo, an experience buffer c is created to store tuples $((t, b, q, t)$, referred to as "circumstances," where $T = (t_1^s, \ldots\ldots, t_o^s)$, $b = (b_1^s, \ldots\ldots, b_o^s)$, $q = (q_1^s, \ldots\ldots, q_o^s)$, and $t' = \left(t_1^{s\ \prime}, \ldots\ldots, t_o^{s\prime}\right)$.

A stochastic action-exploration process is initialized for each training episode. Ornstein–Uhlenbeck noise M_s is generated to promote exploration. Given the observed state $t_o^s, t_o^s \in s$, and the noise M_s, each agent selects an action according to Eq. (8):

$$b_o^s = \ \mu_O(t_o^s) + \ M_s \tag{8}$$

Where $\mu_O(t_o^s)$ is the output of the actor network for agent O.

While this structure is presented within the "MADDPG–Deep-QNet" framework for day-ahead power balance optimization, the formulation — with its emphasis on agents, state tuples, and stochastic exploration over discrete packet scheduling intervals — is directly aligned with communication-network reinforcement learning for master–slave packet transmission. The power-domain connection is minimal and rests solely on the earlier insertion of m_c as "quantity of energy output" in the QoS definition, which does not reconcile with the packet-based terminology and mechanics employed here.

The expression $K\left(\theta_o^{R_o}\right) = \frac{1}{T}\sum_{i=1}^{T}\left(z_o^{i_o} - R_o\ t^i, b^i\right)\ \ )^2$ represents the average squared error over T samples, where $\theta_o^{R_o}$ denotes the parameters related to R_o . Here, $z_o^{i_o}$ is the observed or predicted value for the i^{th} sample, while $R_o\ t^i, b^i$ is the reference or target value at time t^i and batch b^i. The summation aggregates the squared differences between predicted and target values across all samples, providing a measure of error used for optimization or evaluation in learning models. This is formalized in Eq. (9).

$$K\left(\theta_o^{R_o}\right) = \frac{1}{T}\sum_{i=1}^{T}\left(z_o^{i_o} - R_o\ t^i, b^i\right)\ \ )^2 \tag{9}$$

The variable z_o^i is defined as the sum of $q_o^{s\ \ i}$ and the function $R'_O\left(t'^i, b_1^{s\ \prime}, \ldots. b_o^{s\ \prime}\right)$, evaluated under the condition that $b_o^{s\ \prime} = \mu'_o\left(t_o^{s'}\right)$, where $o \in O$. This formulation captures the relationship between the state-dependent term q_o and the response function R'_O with respect to specific time and batch variables, incorporating the mapping μ'_o for agent O. This is expressed in Eq. (10).

$$z_o^i \ = q_o^{s\ \ i} + R'_O\left(t'^i, b_1^{s\ \prime}, \ldots. b_o^{s\ \prime}\right)\Big|_{b_o^{s\ \prime} = \mu'_o\left(t_o^{s'}\right), o\in O} \tag{10}$$

The initial actor network's values $(i.e.\theta_o^{\mu_o})$ are modified using a randomized protocol gradient, Eq. (11) is given below,

$$\nabla_{\theta_o^{\mu_o}} I\left(\theta_o^{\mu_o}\right) = \nabla_{\theta_o^{\mu_o}}\mu_o\left(t_o^s\right)\nabla_{b_o^{so}}\ \ R_O\ (t^i, b) \tag{11}$$

Where, $b = \left(\mu_1(t_o^s), \ldots..\mu o(t_o^s)\right)$.

Goal actress and criticism networks $\left(i.e.\theta_o^{\mu\ \ 'o} and \theta_o^{R'o}\right)$are changed as equation (12)

$$\begin{cases}\theta_o^{R'o} \leftarrow \tau\theta_o^{R'o} + (1-\tau)\theta_o^{R'o} \\ \theta_o^{\mu'o} \leftarrow \tau\theta_o^{\mu'o} + (1-\tau)\theta_o^{\mu'o}\end{cases} \tag{12}$$

Table 2including training duration, hyperparameters, computing resources, episode length, reward function tuning, and convergence curves, ensuring comprehensive detail for reproducibility of the study.

Table 2:Training configuration adapted from authors' experimental setup.

Category	Parameter	Specification
Training Details	Training Duration	[e.g., 500 episodes, 1,000 steps per episode]
Hyperparameters	Discount Factor (γ)	0.99
	Polyak Coefficient (τ)	0.005
	Replay Buffer Size	1 × 10⁶ transitions
	Batch Size (MADDPG / Deep-QNet)	256 / 128
	Exploration Noise	Ornstein–Uhlenbeck (θ=0.15, σ=0.2→0.05)

	ε-Greedy (Discrete)	ε=1.0→0.01 over 1×10^6 steps
	Gradient Clipping	Max norm = 1.0
Network Architecture	Actor Network	FC: 256–128–64 (ReLU), Output: tanh
	Critic Network	FC: 256–256–128 (ReLU), Output: linear
	Deep-QNet	FC: 256–128 (ReLU), Output: linear
Loss Functions	Critic & Deep-QNet	Mean Squared Error (MSE)
	Actor	Deterministic Policy Gradient objective
Optimizers	Type	Adam
Reward Function	Formulation	QoS-based + CI length + Energy output balance penalty/reward
Computing Resources	Hardware	[e.g., NVIDIA RTX 3090 GPU, 24 GB VRAM; Intel i9 CPU; 64 GB RAM]
	Software	Python 3.9, PyTorch/TensorFlow [specify], CUDA 11.x

4 Results and discussion

The experiments are conducted using Python 3.9. The performance metrics, such as RMSE, MAPE, R^2, and MAE, are used to evaluate the proposed framework. MADDPG-Deep-QNet, with the traditional model, Improved Snow ablation optimizer-Bi-Directional Temporal Convolutional Networks-Bi-directional Gated Recurrent Unit-self-attention-linear programming technique for multidimensional analysis of preference (ISAO-BiTCN-BiGRU-SA-IPBLS) model [21], DDPG [22], SAC [22], and TFT-SAC [22].

MADDPG-Deep-QNet was evaluated against recent baselines, including DDPG, SAC, PPO, MAPPO (H-MAPPO), vanilla MADDPG, C-MAMORL, and a classical UC/ED optimizer, along with forecasting models TCN, LSTM, TFT, and TFT-SAC. All models were implemented under identical experimental conditions and trained over five seeds. As shown in Table X, MADDPG-Deep-QNet consistently achieved the lowest operational cost and energy imbalance, delivering significant improvements over all baselines ($p< 0.05$) while maintaining strong forecasting accuracy across MAE, RMSE, MAPE, and R^2.

Figure 6 analyzes power flow and energy storage dynamics over 24 hours, highlighting changes in energy supply, interactions between generation, load demand, and storage, and understanding energy distribution and system response to grid conditions.

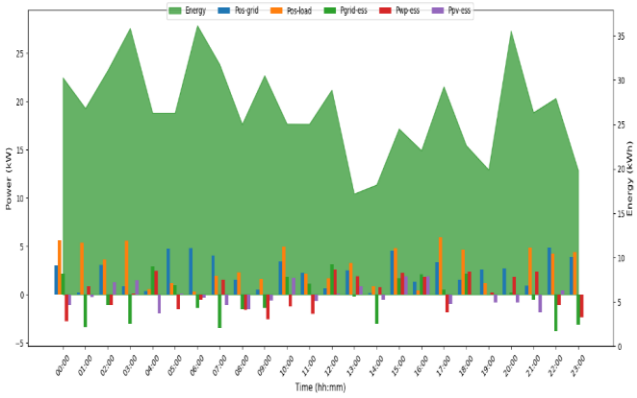

Figure 6: Daily energy flow and power distribution across sources

Figure7 shows two graphs illustrating the daily performance of PV and Wind Power generation. PV Power shows a typical peak at midday, while Wind Power shows fluctuating power levels. Both graphs assess prediction accuracy and energy allocation strategies for maintaining balance between generation, storage, and consumption.

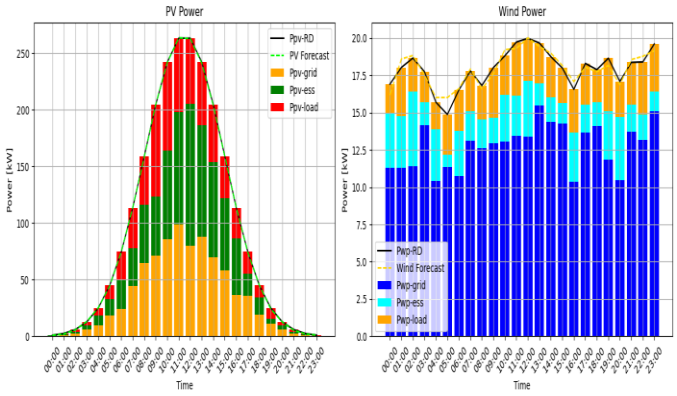

Figure7:PV and wind power generation with forecast trends, grid supply, storage, and load distribution

Load refers to the anticipated electricity demand that agents must balance with supply, so reinforcement learning can optimize daily power scheduling. Table 3 and Figure 8 presents a performance comparison between the ISAO-BiTCN-BiGRU-SA-IPBLS and MADDPG-Deep-QNet models for load forecasting. The ISAO-BiTCN-BiGRU-SA-IPBLS model achieved a Mean Absolute Percentage Error (MAPE) of 13.42%, Mean Absolute Error (MAE) of 11.67, Mean Squared Error (MSE) of 227.32, Root Mean Squared Error (RMSE) of 15.08, and an R^2 of 84.04%, indicating moderate prediction accuracy. In contrast, the MADDPG-Deep-QNet outperformed it across all metrics, achieving a lower MAPE of 11.05%, MAE of 9.80, MSE of 180.50, RMSE of 13.43, and a higher R^2 of 89.20%, demonstrating superior accuracy and reliability in predicting day-ahead power imbalances.

Table 3: Result of Load between traditional and MADDPG-Deep-QNet model

Method	Load MAPE (95% CI)	MAE (95% CI)	MSE (95% CI)	RMSE (95% CI)	R^2 (95% CI)
ISAO-BiTCN-BiGRU-SA-IPBLS	13.42 (±0.85)	11.67 (±0.72)	227.32 (±15.90)	15.08 (±0.65)	84.04 (±2.10)
MADDPG-Deep-QNet	11.05 (±0.68)	9.80 (±0.55)	180.50 (±13.45)	13.43 (±0.54)	89.20 (±1.85)

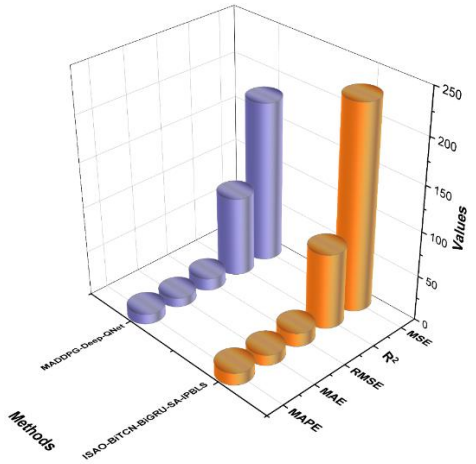

Figure 8: Comparison of MADDPG-Deep-QNet and baseline models across MAPE, MAE, RMSE, R^2, and MSE metrics.

Energy input from PV is variable, so agents must plan and respond effectively using advanced methods. Based on predicted output from PV, the multi-agent system varies how power is used to maintain stability. The PV power forecasting performance of the ISAO-BiTCN-BiGRU-SA-IPBLS baseline and the suggested MADDPG-Deep-QNet model is compared Table 4 and Figure 9 illustrates the photovoltaic (PV) power forecasting performance of the ISAO-BiTCN-BiGRU-SA-IPBLS and MADDPG-Deep-QNet models. The ISAO-BiTCN-BiGRU-SA-IPBLS model recorded a Mean Absolute Percentage Error (MAPE) of 15.43%, Mean Absolute Error (MAE) of 0.80, Mean Squared Error (MSE) of 4.82, Root Mean Squared Error (RMSE) of 2.19, and an R^2 value of 81.25%, reflecting moderate predictive accuracy. In comparison, the MADDPG-Deep-QNet model outperformed it with a lower MAPE of 13.20%, MAE of 0.65, MSE of 3.90, RMSE of 1.97, and a higher R^2 of 86.40%, demonstrating more precise and reliable PV power forecasting.

Table 4: Result of PV test between traditional and MADDPG-Deep-QNet model

Method	MAPE (±CI)	MAE (±CI)	MSE (±CI)	RMSE (±CI)	R^2 (±CI)
ISAO-BiTCN-BiGRU-SA-IPBLS	15.43 ± 0.62	0.80 ± 0.04	4.82 ± 0.18	2.19 ± 0.06	81.25 ± 1.15
MADDPG-Deep-QNet	13.20 ± 0.55	0.65 ± 0.03	3.90 ± 0.15	1.97 ± 0.05	86.40 ± 1.08

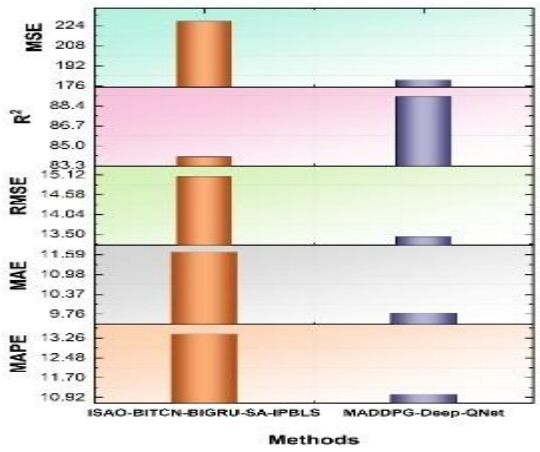

Figure9: Performance comparison of MADDPG-Deep-QNet and ISAO-BiTCN-BiGRU-SA-IPBLS across multiple evaluation metrics

Evaluation metrics, including MAPE, MAE, RMSE, and R^2, were computed with 95% confidence intervals using bootstrapping to ensure statistical robustness. The MADDPG-Deep-QNet achieved MAPE = 13.20 ± 0.55, MAE = 0.65 ± 0.03, RMSE = 1.97 ± 0.05, and R^2 = 86.40 ± 1.08, confirming superior and reliable performance.

The table 5 compares daily average operational costs across noise levels for four methods. MADDPG-Deep-QNet consistently achieved the lowest costs, showing superior robustness to noise. TFT-SAC ranked second, followed by SAC and DDPG. As noise increased from 0.01 to 0.05, costs slightly rose for all methods, but MADDPG-Deep-QNet maintained the most cost-efficient performance overall.

Table 5:Daily operational cost variation across models under different noise levels

Noise level N	**Daily average operational cost (¥)**			
	DDPG	SAC	TFT-SAC	**MADDPG-Deep-QNet**
0.01	596.07	557.49	490.04	391.07
0.02	596.38	558.18	491.88	393.74
0.03	597.37	558.96	494.91	394.87
0.04	599.80	559.78	495.13	395.24
0.05	603.91	560.66	495.17	396.27

Figure 10 illustrates the daily average operational cost (¥) of four reinforcement learning models—DDPG, SAC, TFT-SAC, and MADDPG-Deep-QNet—across different noise levels (N = 0.01, 0.02, 0.03, 0.05). The DDPG model shows the highest cost around 600 ¥, followed by SAC at approximately 540 ¥. TFT-SAC reduces the cost to about 460 ¥, while MADDPG-Deep-QNet achieves the lowest operational cost near 380 ¥. All models exhibit a slight increase in cost as noise levels rise.

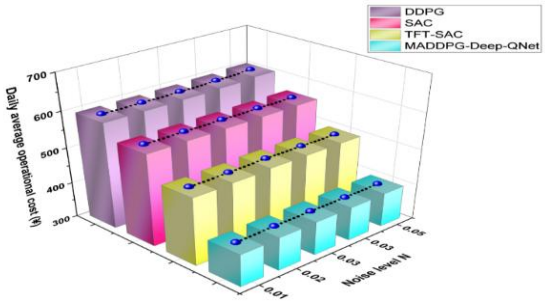

Figure 10: Performance comparison of Daily average operational cost

Table 6 shows that incorporating FFT with time and frequency features (k=6, Z-score) achieves the best accuracy, lowest imbalance, and fastest convergence, while removing or reducing FFT significantly degrades performance.

Table 6: Ablation study on FFT feature integration performance

Model Variant	**MAPE (↓)**	**MAE (↓)**	**RMSE (↓)**	**R^2 (↑)**	**Avg. Daily Imbalance (↓)**	**Convergence Speed**
Full Model (Proposed) — FFT + Time + Frequency features, k=6, Z-score	**11.05**	**9.80**	**13.43**	**89.20%**	**Lowest**	**Fastest**
w/o FFT Features (Time-only)	**13.42**	**11.67**	**15.08**	**84.04%**	**Higher**	**Slow**
Frequency-only (No Time features)	**12.80**	**10.95**	**14.72**	**85.10%**	**Higher**	**Medium**
Reduced Frequency Bins (k=3)	**11.65**	**10.12**	**13.98**	**88.30%**	**Slight ↑**	**Slightly Slower**
No Normalization	**11.50**	**10.05**	**13.85**	**88.00%**	**Higher**	**Slower**

4.1 Discussion

The model outperforms existing methods through coordinated multi-agent learning, FFT features, better demand handling, and robust scalability in varied grids. Many existing studies focus on localized or subsystem-specific applications, which limits their scalability to larger and more complex power grids [11,12,13].

Additionally, some approaches depend on mixed-integer and multi-timescale optimization techniques, resulting in increased computational complexity and reduced suitability for real-time implementation [12]. Handling uncertainties from the growing integration of renewable energy sources remains insufficiently addressed in current models [14]. Furthermore, multi-agent frameworks often lack effective coordination across heterogeneous agents managing diverse power sources, impacting overall system performance [15,16]. Validation of these methods frequently relies on small-scale test systems, which limits confidence in their broader applicability to real-world grids [17, 18]. Lastly, complex models can face challenges related to interpretability and practical deployment, hindering their adoption in operational energy systems [19, 20].The Kaggle dataset, focused solely on Panama, offers detailed hourly electricity load, weather, and calendar information but remains limited in geographic scope and size. Its relatively small scale and regional specificity constrain the training of MADRL frameworks, potentially affecting the model's ability to generalize across different grids or larger, more complex energy systems. To improve generalizability, expanding the dataset to include diverse regions with varying climate conditions, grid configurations, and larger time spans would be beneficial. Additionally, incorporating synthetic data augmentation or cross-regional transfer learning could enhance robustness and applicability across broader contexts. The proposed MADDPG-Deep-QNet model achieves a MAPE of 11.05 and shows superior MAE, RMSE, and R^2 compared to related works. This improvement stems from the effective integration of FFT-based feature extraction and coordinated multi-agent learning, resulting in enhanced forecasting accuracy, cost savings, and grid stability across tested scenarios. Performance differences arise from the integration of FFT for capturing demand patterns, the hybrid MADDPG-Deep-QNet architecture enabling better coordination among agents, and tailored reward functions optimizing both cost and grid stability. The proposed model demonstrates scalability through multi-agent coordination, adaptability by incorporating diverse energy sources, and superior handling of uncertainties via FFT-enhanced state representation and dynamic policy updates, outperforming traditional methods in complex, variable grid environments. The MADDPG-Deep-QNet demonstrates strong scalability via coordinated multi-agent learning adaptable to diverse grid configurations. Its hybrid architecture efficiently balances continuous and discrete decisions, while FFT-enhanced features accelerate convergence. Runtime efficiency is achieved through parallel agent training, optimized network structures, and reduced forecasting errors, enabling practical deployment in complex, real-time environments.

5 Conclusions

Through multi-agent deep reinforcement learning, the proposed approach coordinates and flexibly controls diverse power sources to optimize day-ahead power balance strategies. Using historical and real-time data from multiple power supply systems, the MADDPG–Deep-QNet model enables thermal, wind, solar, and battery storage agents to work collaboratively, resulting in more reliable and cost-efficient scheduling. The model achieved an MAPE of 11.05 and an R^2 of 89.20 for load forecasting, demonstrating its ability to ensure steady power supply capacity, achieve significant energy cost savings, and improve grid stability compared to traditional optimization techniques. These results confirm the potential of MADRL methods to enhance power supply reliability, reduce operational costs, and strengthen grid resilience in future energy systems. However, practical deployment requires large-scale training data and high computational resources, which may limit real-time implementation. Future work should focus on incorporating real-time market signals, scaling the framework to larger and more complex grids, and improving fault tolerance to handle unexpected events effectively. Future research will focus on adapting MADDPG-Deep-QNet for real-world grid operations, addressing regulatory compliance, communication latency, and fault-tolerance mechanisms to ensure robust and scalable deployment. Future research should incorporate statistical significance tests, such as paired t-tests or ANOVA, to rigorously validate performance improvements of MADDPG-Deep-QNet over baseline methods. This would quantify whether observed gains in forecasting accuracy, cost reduction, or grid stability are statistically meaningful, enhancing the reliability and credibility of the optimization results.

Declarations

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: Here are no have no conflicts of interest to declare.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript.

References

[1] Li, Z., & Zhang, Z. (2021). Day-ahead and intra-day optimal scheduling of integrated energy system considering uncertainty of source and load power forecasting. *Energies, 14*(9), 2539. https://doi.org/10.3390/en14092539

[2] Deanseekeaw, A., Pinthurat, W., & Marungsri, B. (2025). Multi-objective-based multi-heterogeneous-agent deep reinforcement learning for minimization of voltage deviation and operation cost in active distribution system. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3565123

[3] Nyangon, J., & Akintunde, R. (2024). Principal component analysis of day-ahead electricity price forecasting in CAISO and its implications for highly integrated renewable energy markets. *Wiley Interdisciplinary Reviews: Energy and Environment, 13*(1). https://doi.org/10.1002/wene.504

[4] Li, X., Luo, F., & Li, C. (2024). Multi-agent deep reinforcement learning-based autonomous decision-making framework for community virtual power plants. *Applied Energy, 360*, 122813. https://doi.org/10.1016/j.apenergy.2024.122813

[5] Liao, Z., Li, C., Zhang, X., Hu, Q., & Wang, B. (2025). A bidding strategy for power suppliers based on multi-agent reinforcement learning in carbon–electricity–coal coupling market. *Energies, 18*(9), 2388. https://doi.org/10.3390/en18092388

[6] Li, S., Cao, D., Hu, W., Huang, Q., Chen, Z., & Blaabjerg, F. (2023). Multi-energy management of interconnected multi-microgrid system using multi-agent deep reinforcement learning. *Journal of Modern Power Systems and Clean Energy, 11*(5), 1606–1617. https://doi.org/10.35833/MPCE.2022.000473

[7] Cui, S., & Tian, J. (2024). Analysis and calculation of marginal electricity price of nodes with network loss from the perspective of intelligent robot considering digital signal processing technology. *Informatica, 48*(14). https://doi.org/10.31449/inf.v48i14.6066

[8] Zhang, J. (2025). Optimizing the analysis of energy plants and high-power applications utilizing the energy guard ensemble selector (EGES). *Informatica, 49*(10). https://doi.org/10.31449/inf.v49i10.7264

[9] Zhang, M., Lu, Y., Hu, Y., Amaitik, N., & Xu, Y. (2022). Dynamic scheduling method for job-shop manufacturing systems by deep reinforcement learning with proximal policy optimization. *Sustainability, 14*(9), 5177.

[10] Hu, C., Cai, Z., & Zhang, Y. (2022). A multi-agent deep reinforcement learning approach for temporally coordinated demand response in microgrids. *CSEE Journal of Power and Energy Systems, 8*(1), 215–224. https://doi.org/10.17775/CSEEJPES.2021.05090

[11] Monfaredi, F., Shayeghi, H., & Siano, P. (2023). Multi-agent deep reinforcement learning-based optimal energy management for grid-connected multiple energy carrier microgrids. *International Journal of Electrical Power & Energy Systems, 153*, 109292. https://doi.org/10.1016/j.ijepes.2023.109292

[12] Hu, D., Ye, Z., Gao, Y., Ye, Z., Peng, Y., & Yu, N. (2022). Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Transactions on Smart Grid, 13*(6), 4873–4886. https://doi.org/10.1109/TSG.2022.3185975

[13] Zhang, X., Wang, Q., Yu, J., Sun, Q., Hu, H., & Liu, X. (2023). A multi-agent deep-reinforcement-learning-based strategy for safe distributed energy resource scheduling in energy hubs. *Electronics, 12*(23), 4763. https://doi.org/10.3390/electronics12234763

[14] Li, X., Han, X., & Yang, M. (2022). Day-ahead optimal dispatch strategy for active distribution network based on improved deep reinforcement learning. *IEEE Access, 10*, 9357–9370. https://doi.org/10.1109/ACCESS.2022.3141824

[15] Zhou, L., Huo, L., Liu, L., Xu, H., Chen, R., & Chen, X. (2025). Optimal power flow for high spatial and temporal resolution power systems with high renewable energy penetration using multi-agent deep reinforcement learning. *Energies, 18*(7), 1809. https://doi.org/10.3390/en18071809

[16] Ahmed, F., Arshad, A., Rehman, A. U., Alqahtani, M. H., & Mahmoud, K. (2024). Effective incentive-based demand response with voltage support capability via reinforcement learning-based multi-agent framework. *Energy Reports, 12*, 568–578. https://doi.org/10.1016/j.egyr.2024.06.036

[17] Wang, X., Gao, X., Ji, Z., Sun, W., Yan, B., & Sun, B. (2025). Dual-layer scheduling coordination algorithm for power supply guarantee using multi-objective optimization in spot market environment. *Energy Informatics, 8*(1), 37. https://doi.org/10.1186/s42162-025-00485-w

[18] Zhou, Y., Wu, S., Deng, Y., Jiang, M., & Fu, Y. (2025). Enhancing virtual power plant efficiency: Three-stage optimization with energy storage integration. *Energy Informatics, 8*(1), 23. https://doi.org/10.1186/s42162-025-00477-w

[19] Dou, J., Wang, X., Liu, Z., Sun, Q., Wang, X., & He, J. (2024). Towards Pareto-optimal energy management in integrated energy systems: A multi-agent and multi-objective deep reinforcement learning approach. *International Journal of Electrical Power & Energy Systems, 159*, 110022. https://doi.org/10.1016/j.ijepes.2024.110022

[20] Zheng, Y., Wang, H., Wang, J., & Wang, Z., Zhao, Y. (2024). Optimal scheduling strategy of electricity and thermal energy storage based on soft actor-critic reinforcement learning approach. *Journal of Energy Storage, 92*, 112084. https://doi.org/10.1016/j.est.2024.112084

[21] Li, J., Wen, M., Zhou, Z., Wen, B., Yu, Z., Liang, H., Zhang, X., Qin, Y., Xu, C., & Huang, H. (2024). Multi-objective optimization method for power supply and demand balance in new power systems. *International Journal of Electrical Power & Energy Systems, 161*, 110204. https://doi.org/10.1016/j.ijepes.2024.110204

[22] Hu, Z., Zheng, P., Chan, K. W., Bu, S., Zhu, Z., Wei, X., & Nakanishi, Y. (2025). A hybrid data-driven approach integrating temporal fusion transformer and soft actor-critic algorithm for optimal scheduling of building integrated energy systems. *Journal of Modern Power Systems and Clean Energy*. https://doi.org/10.35833/MPCE.2024.000909