Applications of Deep Reinforcement Learning in Optimization of Offshore Wind Power Operation and Maintenance

Jinsheng Lai*
School of Civil and Transportation Engineering
South China University of Technology
Guangzhou, China
reallai@163.com

***Abstract*—Offshore wind power, as a key area to achieve energy transition, faces multiple challenges such as harsh environments, high costs and complex decision-making. Traditional operation and maintenance strategies based on fixed rules are difficult to optimize dynamically, and deep reinforcement learning provides a new way to solve such sequential decision-making problems. In this paper, an intelligent decision-making framework for offshore wind power operation and maintenance based on Deep Deterministic Policy Gradient (DDPG) algorithm is systematically constructed. The framework models the operation and maintenance problem as a Markov decision process, designs a state space integrating equipment health, environmental conditions, resource status and market information, and an action space combining discrete maintenance instructions and continuous scheduling decisions, and guides the agent to learn through a reward function integrating generation revenue, operation and maintenance costs and fault penalties. Empirical analysis shows that the DDPG model proposed in this paper outperforms the benchmark algorithms such as Deep Q Network (DQN) and Proximal Policy Optimization (PPO) in four core indicators: total net revenue (TNR, 4127.5 thousand yuan), number of failures (NF, as low as 4), Mean Time Between Failures (MTBF, 152.3 days) and Mean Time to Repair (MTTR, 28.5 hours). Robustness tests further show that the strategy can maintain stable performance under stress scenarios such as wind speed fluctuation increasing by 20% and wind turbine failure rate increasing by 15%. The research shows that deep reinforcement learning, especially DDPG algorithm, can effectively realize the adaptive optimization of offshore wind power operation and maintenance under multi-objective constraints, and provide feasible intelligent solutions for improving the economy and reliability of operation and maintenance.**

Keywords*—Offshore Wind Power, Operation and Maintenance, Deep Reinforcement Learning, DDPG

I. Introduction

Offshore wind power is an important form of renewable energy, and its large-scale development and efficient operation and maintenance have become the key path to achieve the goal of "double carbon". However, offshore wind farms are usually located in deep sea areas with harsh environments and face multiple challenges such as poor accessibility, high maintenance costs and complex failure risks. Traditional operation and maintenance strategies are mostly based on fixed periods or empirical rules, which are difficult to adapt to complex coupling factors such as dynamic degradation of wind turbine state, random and variable weather window, wake effect and uneven fatigue load distribution [1]. Therefore, exploring intelligent, adaptive and cost-effective operation and maintenance optimization methods has urgent practical significance for improving the reliability, availability and return on investment of offshore wind farms throughout their life cycle.

In recent years, deep reinforcement learning, as an artificial intelligence method that does not require accurate environmental models and can autonomously learn optimal decision sequences through interaction with complex environments, has shown significant advantages in solving high-dimensional continuous control and sequential decision problems. At its core, it is to maximize long-term cumulative rewards through trial-and-error mechanism, which is exactly in line with the nature of decision-making in offshore wind power operation and maintenance, which requires long-term cost-benefit trade-off under uncertain environment. Because of its effectiveness in dealing with continuous action space, the deep deterministic policy gradient algorithm is used in pitch control, wave compensation trestle adaptive control and sensor network energy efficiency optimization.

This paper aims to systematically discuss the overall application framework and modeling method of Policy Gradient (DDPG) algorithm in offshore wind power operation and maintenance optimization, in order to provide theoretical reference and practical ideas for promoting offshore wind power operation and maintenance to develop in a more intelligent, more economical and more reliable direction. The core novelty of this framework is the construction of a hybrid action space integrating discrete maintenance instructions and continuous scheduling decisions, breaking through the inherent limitation of existing DRL-based wind power O&M studies that only adopt a single type of action space. The designed state space comprehensively couples four types of dynamic information: equipment health, marine environment, O&M resources and electricity market, with more complete state dimensions than similar studies, and is more suitable for the complex and coupled operation scenarios of offshore wind power. The reward function innovatively unifies the modeling of power generation revenue, O&M costs, fault penalties and environmental safety constraints to achieve multi-objective collaborative optimization, which is different from the reward design idea of traditional studies focusing on single scenario objectives.

II. DDPG Algorithm

A. Overview of Offshore Wind Power Operation and Maintenance

Offshore wind power operation and maintenance refers to a series of planning, execution and monitoring activities carried out to ensure the safe, stable and efficient operation of wind turbines during the whole life cycle of wind farms. Compared with onshore wind power, the core challenges facing offshore operations stem from their unique operating environment. Severe oceanographic weather conditions such as high winds, waves and fog result in short and uncertain maintenance windows and significantly reduced accessibility for personnel and vessels [2]. In addition, the distance of the sea area makes the transportation cost high and the emergency response time of sudden failure significantly prolonged. Operation and maintenance activities include not only corrective maintenance and regular preventive maintenance of the mechanical, electrical and control systems of the wind turbine itself, but also predictive maintenance based on conditions, and inspection and maintenance of supporting facilities such as offshore substations and submarine cables. A complete operation and maintenance system also needs to consider the resources of spare parts inventory management, port logistics support, special vessel scheduling and technical personnel allocation. Traditional operation and maintenance strategies mostly rely on fixed time periods or manufacturers' basic recommendations. This static model is difficult to dynamically respond to differences in actual degradation rates of wind turbine components, and cannot effectively integrate real-time environmental risks such as wind speed and wave height with fluctuations in electricity market prices. Therefore, the goal of operation and maintenance optimization is to balance maintenance cost, fault loss and power generation income through scientific decision-making under the constraints of ensuring system reliability and power generation, and finally minimize the total cost of operation and maintenance in the whole life cycle or maximize the overall income. It is essentially a complex sequential decision problem under multiple uncertainties.

B. DDPG Algorithm

DDPG algorithm is a mainstream algorithm in the field of deep reinforcement learning, designed to deal with decision problems in continuous action spaces. It cleverly combines the idea of deep Q networks with a deterministic policy gradient framework, belonging to the actor-critic architecture. The core of the algorithm consists in training two deep neural networks simultaneously: the actor network and the critic network. The actor network acts as a decision-maker, directly observing the state of the environment and outputting a deterministic, continuous action value [3]. The critic network acts as an evaluator, receiving the current state and the actions given by the actor network and evaluating the expected long-term cumulative reward value of the state-action pair. During the training process, the learning direction of the actor network is to adjust its parameters along the direction of improving the value evaluation given by the critic network, so that the output actions can obtain higher long-term returns. In order to improve the stability of learning process, DDPG algorithm draws lessons from target networks and experience replay mechanism. The algorithm replicates target networks with the same structure for the actor network and the critic network respectively. Its parameters are updated slowly and are used to calculate stable target values, thereby alleviating fluctuations in training. Experience replay stores the transfer data generated by the interaction between agent and environment, and randomly samples them during training, breaking the temporal correlations among data. This design enables the agent to learn directly from the original high-dimensional state inputs and to learn a deterministic policy that approximates the optimal in an end-to-end manner in scenarios requiring continuous control outputs, such as wind turbine maintenance decisions and continuous power regulation [4].

The core mathematical formulations of the DDPG algorithm are given as follows.

The policy gradient for updating the actor network is:

$$\nabla_{\theta^\mu} J \approx \mathbb{E}_{s_t \sim \rho^\beta}\left[\nabla_{\theta^\mu}\mu(s_t|\theta^\mu)\nabla_a Q(s_t, a|\theta^Q)|_{a=\mu(s_t|\theta^\mu)}\right] \quad (1)$$

where θ^μ is the parameter of the actor network, θ^Q is the parameter of the critic network, $Q(\cdot)$ is the action-value function, $\mu(\cdot)$ is the deterministic policy, and ρ^β is the state distribution.

The target Q-value for critic network training is:

$$y_i = r_i + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1}\middle|\theta^{\mu'}\right)\middle|\theta^{Q'}\right) \quad (2)$$

where r_i is the immediate reward, γ is the discount factor, and Q', μ' are the target critic and target actor networks.

The soft update rule for target network parameters is:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta' \quad (3)$$

where τ is the soft update coefficient ($\tau \ll 1$), θ is the online network parameter, and θ' is the target network parameter.

The loss function of the critic network is the mean squared TD error:

$$L(\theta^Q) = \mathbb{E}_{(s,a,r,s')\sim R}[(y_i - Q(s,a|\theta^Q))^2] \quad (4)$$

where R is the experience replay buffer, and (s, a, r, s') is the state transition tuple.

III. Modeling of Offshore Wind Power Operation and Maintenance Optimization Based on DDPG

A. Model Framework Design

To apply deep reinforcement learning to offshore wind power operation and maintenance optimization decision-making, this paper constructs an intelligent decision-making framework based on deep deterministic policy gradient algorithm, and its core architecture is shown in Figure 1.

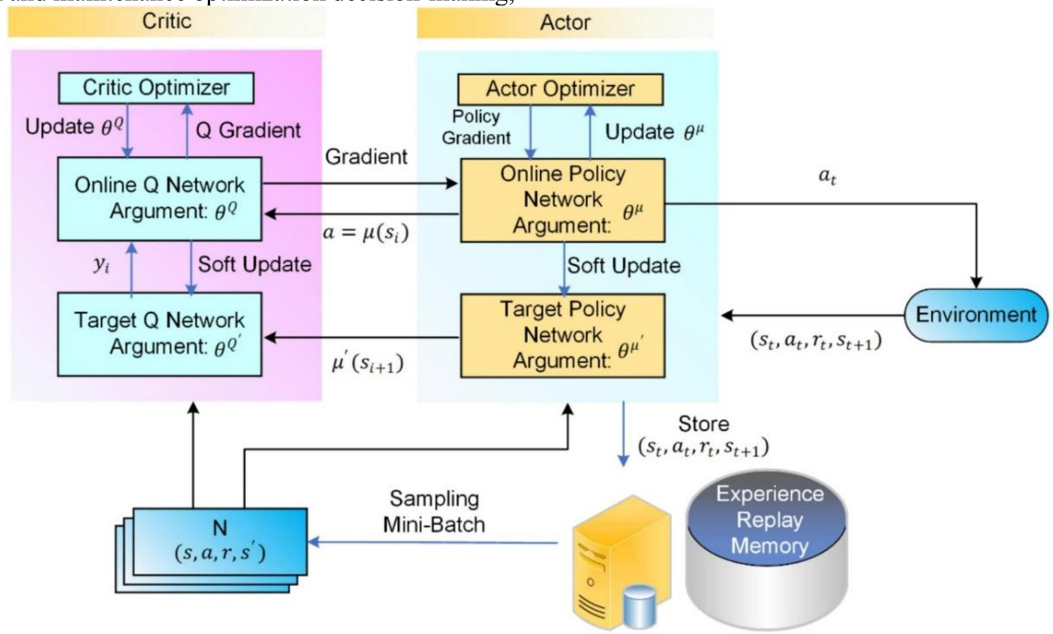

Fig. 1 DDPG Model for Optimization of Offshore Wind Power Operation and Maintenance

The framework abstracts the offshore wind power operation environment into a standard Markov decision process, and learns the optimal policy through continuous interaction between agents and the environment. The frame body consists of actor-critic modules. The actor module is responsible for the strategy function, and its online strategy network directly outputs the determined operation and maintenance actions according to the current wind farm system state (such as unit health, weather forecast, wave condition, power market price, etc.), such as the timing arrangement of specific maintenance tasks and resource allocation scheme. The critic module is responsible for the value function, and its online Q network is responsible for evaluating the long-term expected value of the actions output by the actor module in the current state and providing gradient directions for policy updates.

To ensure the stability of training, target networks and experience replay mechanism are introduced into the framework. Actors and critics have separate online and target networks, and target network parameters track the online network slowly through soft updates to provide reliable value targets. Experience replay memory stores the state transition sequences generated by the agent's exploration environment, from which small batches of samples are randomly sampled to break the temporal correlations among data during training [5]. Specifically, the agent performs actions at each interaction step, observes rewards and new states from the environment, and stores that experience. The data sampled from the memory pool is then used to update both the critic network and the actor network: the critic network improves the accuracy of the value assessment by minimizing the time difference error; the actor network adjusts the policy parameters in the direction of improving the long-term cumulative reward according to the policy gradient provided by the critic network. Through this closed-loop "interaction-storage-learning" mechanism, the framework enables agents to adaptively learn to make sequential decisions with optimal cost and benefit in complex and uncertain operation environments.

B. State and Action Space Design

The state space should describe the dynamic situation of wind farm system at decision time comprehensively. Specifically, the state vector mainly includes the following dimensions. The health status of a wind turbine unit is usually quantified by degradation indicators or predicted remaining useful life of key components (blades, gearbox, generator, pitch system). Environmental conditions, including real-time and forecast wind speeds, wave heights, tidal cycles and visibility, directly determine the time window and safety risks of maintenance operations [6]. The status of operation and maintenance resources covers the location and type of vessels available for maintenance, the number and skills of technicians, and the inventory level of spare parts in the port. Grid and market conditions, such as current feed-in tariffs, system load demand or dispatch orders, also constitute important inputs [7]. Time information (such as calendar date, accumulated operating hours) and historical maintenance records are also used as supplementary state variables. Key state indicators adopt standardized quantification methods: unit health is characterized by the normalized value of remaining useful life (RUL) of key components, environmental conditions are quantified by significant wave height and 1-hour wind speed forecast value, and O&M resources take the number of available vessels and spare parts inventory adequacy rate as core quantification indicators. All quantified indicators are min-max normalized to the [0,1] interval, and temporal sliding mean features are extracted to ensure unified state input dimensions and stable convergence of model training.

Action space defines the operations that an agent can perform in a given state. At its core, each turbine in the wind farm is assigned a specific maintenance order, which is a discrete action that usually includes "no operation (continue operation)", "planned preventive maintenance" and "corrective failure maintenance". More elaborate designs expand Planned Preventive Maintenance into a continuous action that specifies specific points in time (e.g., hours in the future) when maintenance is recommended. At the same time, the action space also needs to contain the dimension of resource allocation, which can express the intensity of resource investment through continuous variables or represent the resource allocation scheme through discrete selection.

This multidimensional, hybrid design of state and action spaces is designed to enable models to receive sufficient environmental information and output detailed, executable operations scheduling and resource allocation instructions.

C. *Reward Function Construction*

Reward function is the core mechanism to guide the learning direction of agent, and its design must accurately reflect the multiple objectives of economy and reliability pursued by offshore wind power operation and maintenance optimization. The reward signal is generated after each decision step and is essentially the sum of the immediate benefits and costs of the actions taken within that step. Positive incentives mainly come from generation revenue. This is usually related to the availability status of the wind turbine during the period, the actual wind resource conditions and the feed-in tariff.

The output power of the wind turbine is calculated as:

$$P(v) = \begin{cases} 0, v < v_{in} \text{ or } v > v_{out} \\ P_r \cdot \frac{v^3 - v_{in}^3}{v_r^3 - v_{in}^3}, v_{in} \leq v \leq v_r \\ P_r, v_r < v \leq v_{out} \end{cases} \quad (5)$$

where $P(v)$ is the real-time power, v is the wind speed, v_{in}/v_{out} is the cut-in/cut-out wind speed, v_r is the rated wind speed, and P_r is the rated power.

The comprehensive reward function for offshore wind O&M is defined as:

$$R_t = \eta \cdot P(v_t) \cdot p_t - C_{om,t} - \lambda \cdot F_t \quad (6)$$

where R_t is the total reward at step t, η is the unit conversion coefficient, p_t is the electricity price, $C_{om,t}$ is the O&M cost, λ is the fault penalty coefficient, and F_t is the fault indicator (1 for fault, 0 for normal).

The availability status is determined by whether the wind turbine is operating and has no faults [8]. The power generation can be estimated from the wind power curve. The final benefit is the product of the power generation and the tariff. Negative rewards arise from costs and penalties. The most important cost is the cost of operation and maintenance, including the fixed rental cost of sending vessels, fuel consumption, personnel man-hour cost and replacement of spare parts. These costs are closely related to the type of action, the resources required, and the operating conditions. Another key negative reward is the penalty for failure of the wind turbine, which is usually set to a large negative value far higher than the cost of preventive maintenance, to strongly incentivize the agent to take preventive measures before the failure occurs and thus learn the value of preventive maintenance. The weight coefficients η and λ in the reward function are calibrated through multiple rounds of pre-experiments combined with measured data of O&M costs, power generation revenue and fault losses in the offshore wind power industry, ensuring that the multi-objective weight ratio conforms to engineering practice and the optimization direction is reasonable. Other leading terms can also be introduced into the reward function. For example, risk penalties may be imposed for forcing work in a severe weather window to encourage safe decision-making, or small positive rewards may be given for successfully completing preventive maintenance before a failure occurs to encourage proactive behavior. The reward function converts complex long-term economic optimization objectives into real-time feedback signals that can be perceived at each step of the agent, driving it to learn an optimal strategy to balance short-term investment and long-term return under uncertainty.

IV. Empirical Analysis

A. *Experimental Design*

The data set used in the experiment comes from the combination of simulated and public datasets of an offshore wind farm group. Core data include: Time series of historical operation status of 20 × 5 MW offshore wind turbines within three years, which integrates component degradation simulation data based on physical model and fault records in public dataset to reflect performance attenuation and random faults of key components such as gearbox and generator; high-precision meteorological and marine data of corresponding time period, including wind speed, wind direction, effective wave height and wave period, with time resolution of 1 hour; Operation and maintenance log records the specific type of historical maintenance task, time consumption, spare parts consumed, type and sailing time of dispatched vessels, as well as the daily electricity price data of regional electricity market. The experimental hardware environment is configured as a workstation equipped with Intel Core i9-13900K processor, 128GB memory and NVIDIA RTX 4090 graphics card. The software environment is Ubuntu 20.04 operating system, using Python 3.9 as programming language, and relying on PyTorch 1.13 deep learning framework to implement DDPG algorithm and its training process.

Key hyperparameter settings for the model are shown in Table 1. These parameters are determined through multiple rounds of pre-experimental adjustments to balance the algorithm's exploration and exploitation capabilities and ensure the stability of the training process.

TABLE I. Experimental Parameter Settings

Name of parameter	Parameter values
Actor network learning rate	0.0001
Critics network learning rate	0.001
Discount factor (γ)	0.99

Soft Update Coefficient (τ) of Target Networks	0.005
Experience replay cache capacity	1,000,000
Training batch size	128
Random noise parameters (exploration)	Initial 0.1, linear decay

This paper selected four core evaluation indicators. First, Total Net Revenue (TNR), which is the net value of total generation revenue minus total operation and maintenance cost during the evaluation period, is the primary comprehensive indicator to measure economy. The second is the Number of Failures (NF), which records the total number of unplanned downtime failures that occur during the evaluation cycle and directly reflects the reliability of the system. The third is Mean Time Between Failures (MTBF), which refers to the average continuous operation time of a single fan between adjacent failures. It is a key indicator to measure the preventive effect of operation and maintenance strategies. The higher the value, the more effective the preventive maintenance strategy is. Mean Time to Repair (MTTR), defined as the average time from failure or planned maintenance instruction to completion of maintenance grid connection. This indicator comprehensively reflects the efficiency of maintenance resource scheduling and the ability to utilize weather windows, and is an important indicator for evaluating the ability to respond to special challenges in offshore operation and maintenance.

B. *Experimental Results and Analysis*

To evaluate the performance of operation and maintenance optimization model based on DDPG, this paper compares it with deep Q network (DQN) and proximal policy optimization (PPO). Under the same simulation environment and evaluation cycle, the comprehensive performance comparison of each algorithm is shown in Table 2.

TABLE II. PERFORMANCE COMPARISON RESULTS OF DIFFERENT ALGORITHMS

Indicator	DDPG	DQN	PPO
TNR (thousand yuan)	4127.5	3568.2	3894.1
NF	4	11	7
MTBF (days)	152.3	89.7	125.6
MTTR (hours)	28.5	41.2	33.8

As shown in Table 2, the DDPG model proposed in this paper significantly outperforms the comparison benchmark on all core indicators. In terms of economy, DDPG strategy achieves a total net return (TNR) of 4127.5 thousand yuan, which is 15.7% and 6.0% higher than DQN and PPO strategies respectively. This is mainly due to its better cost-benefit trade-off ability. In terms of reliability, the number of failures (NF) under DDPG strategy is only 4 times, far lower than 11 times of DQN and 7 times of PPO, and the corresponding mean time between failures (MTBF) reaches 152.3 days, which is 69.8% higher than DQN (89.7 days) and 21.2% higher than PPO (125.6 days). This indicates that DDPG learned strategies have more proactive and precise preventive maintenance capabilities. In terms of operation and maintenance efficiency, the MTTR under DDPG strategy is 28.5 hours, which is 30.8% shorter than DQN and 15.7% shorter than PPO. It proves that DDPG strategy has more advantages in scheduling resources and utilizing limited operation windows. DQN is the weakest because of its limited discrete action space, and PPO is stable, but its long-term strategy optimization ability is slightly inferior to DDPG in this paper.

To test the adaptability and stability of DDPG strategy, this paper further designed robustness test to observe the change of core indicators by changing key environmental parameters. The results are shown in Table 3.

TABLE III. ROBUSTNESS EXPERIMENTAL RESULTS

Test scenarios	TNR rate of change	MTBF rate of change	MTTR rate of change
Business as usual	0%	0%	0%
20% increase in wind speed volatility	-3.2%	-1.8%	+2.1%
Fan failure rate increased by 15%	-5.7%	-8.3%	+4.5%
10% increase in maintenance costs	-4.1%	-0.5%	+1.2%

As shown in Table 3, DDPG strategy exhibits good robustness under various stress test scenarios. When wind speed fluctuation increases by 20%, TNR decreases by 3.2% and MTBF decreases by 1.8% by adjusting maintenance schedule dynamically. When the basic failure rate of wind turbine increases by 15%, the strategy responds by moderately increasing preventive maintenance investment, resulting in a decrease of TNR by 5.7%, but effectively restraining the deterioration of reliability indicators, MTBF decreases by 8.3%, which is better than the increase in failure rate. When the cost of maintenance activities increased by 10%, the strategy optimized the resource mix and task packaging logic, so MTBF was almost unaffected (only decreased by 0.5%) and TNR decreased (4.1%) less than the cost increase. The increase of MTTR is controlled within 5% in all disturbance scenarios; MTTR rises slightly under wind speed fluctuation and increased maintenance costs, and although it increases moderately with higher fan failure rate, it remains within a reasonable range, confirming that the strategy can stably maintain O&M response and repair efficiency. These results show that the strategy can adjust its decision adaptively according to the changing environment, and maintain the balance of effectiveness and economy under variable conditions.

V. CONCLUSIONS

Experimental results show that the DDPG model constructed in this paper outperforms traditional deep reinforcement learning baseline algorithms significantly in comprehensive performance. Specifically, it achieves the highest total net benefit while minimizing unplanned failures and achieving the longest mean time between failures and shortest mean time to repair. This proves that the model not only has excellent long-term economic optimization ability, but also performs well in improving system reliability and operation efficiency. Its core advantage lies in its ability to accurately grasp the timing and intensity of preventive maintenance through end-to-end learning, and achieve the best balance between failure risk and direct operation and maintenance cost. Robustness analysis further confirms that the strategy exhibits good adaptability in the face of environmental parameter fluctuations, and can mitigate the negative effects brought by external changes by dynamically adjusting decisions, which proves its potential application in real complex environments.

The main contribution of this paper lies in systematically applying deep reinforcement learning to the full chain decision optimization of offshore wind power operation and maintenance, and providing specific modeling paradigm and verification. The research results show that deep reinforcement learning can effectively overcome the limitations of traditional methods and provide a data-driven and autonomous decision support path for maximizing the value of offshore wind farms throughout their life cycle. Future work could further explore multi-agent frameworks to coordinate collaborative operation and maintenance of large-scale wind farms, integrate transfer learning to accelerate strategy training for new sites, and combine digital twinning technology to achieve a higher fidelity training environment and online strategy adjustment.

References

[1] Yang C, Jia J, He K, et al. Comprehensive analysis and evaluation of the operation and maintenance of offshore wind power systems: a survey[J]. Energies, 2023, 16(14): 5562.

[2] Akdağ O. The operation and applicability to hydrogen fuel technology of green hydrogen production by water electrolysis using offshore wind power[J]. Journal of Cleaner Production, 2023, 425: 138863.

[3] Desalegn B, Gebeyehu D, Tamrat B, et al. Onshore versus offshore wind power trends and recent study practices in modeling of wind turbines' life-cycle impact assessments[J]. Cleaner Engineering and Technology, 2023, 17: 100691.

[4] Yu H, Yang X, Chen H, et al. Energy storage capacity planning method for improving offshore wind power consumption[J]. Sustainability, 2022, 14(21): 14589.

[5] Yang B, Cai A, Lin W. Analysis of early fault vibration detection and analysis of offshore wind power transmission based on deep neural network[J]. Connection Science, 2022, 34(1): 1005-1017.

[6] Kim S H, Shin Y J. Optimize the operating range for improving the cycle life of battery energy storage systems under uncertainty by managing the depth of discharge[J]. Journal of Energy Storage, 2023, 73: 109144.

[7] Ye Z, Cai Z, Yang H, et al. Joint optimization of maintenance and quality inspection for manufacturing networks based on deep reinforcement learning[J]. Reliability engineering & system safety, 2023, 236: 109290.

[8] Dolati M, Sayyaf N. Towards Reliable Deep Reinforcement Learning for Industrial Applications: A DDPG-based Algorithm with Improved Performance[J]. AUT Journal of Mechanical Engineering, 2026, 10(1): 61-74.