*energies*

MDPI

*Article*

# Monthly Power Outage Maintenance Scheduling for Power Grids Based on Interpretable Reinforcement Learning

**Wei Tang [1], Xun Mao [1], Kai Lv [1], Zhichen Cai [2] and Zhenhuan Ding [2,*]**

[1] State Grid Anhui Electric Power Co., Ltd., Hefei 230001, China
[2] Department of Electrical Engineering, Anhui University, Hefei 230601, China
* Correspondence: zding1@ahu.edu.cn

**Abstract**

This paper proposes an interpretable optimization method for power grid outage scheduling based on reinforcement learning. An outage scheduling optimization model is proposed, considering the convergence of power flow calculation, voltage violations, and operational economic behavior as objectives, while considering constraints such as simultaneous outage constraints, mutually exclusive constraints, and maintenance windows. Key features of the outage schedule are selected based on Shapley values to construct a Markov optimization model for outage scheduling. A deep reinforcement learning agent is established to optimize the outage schedule. The proposed method is applied to the IEEE-39 and IEEE-118 bus system for validation. Experimental results show that the proposed method outperforms existing algorithms in terms of voltage violation, total power losses, and computational time. The proposed method eliminates all voltage violations and reduces active power losses up to 5.7% and computation time by 6.8 h compared to conventional heuristic algorithms.

**Keywords:** power outage maintenance scheduling; reinforcement learning; interpretability; optimization

check for updates

Academic Editor: José Matas

Received: 12 September 2025
Revised: 3 October 2025
Accepted: 4 October 2025
Published: 16 October 2025

**Citation:** Tang, W.; Mao, X.; Lv, K.; Cai, Z.; Ding, Z. Monthly Power Outage Maintenance Scheduling for Power Grids Based on Interpretable Reinforcement Learning. *Energies* **2025**, *18*, 5454. https://doi.org/10.3390/en18205454

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

## 1. Introduction

Currently, the management of outage maintenance schedules is primarily handled by system operators. Various levels dispatch centers optimize the combination of outage requests submitted by maintenance units based on stability guidelines, dispatch procedures, event investigation protocols, and renewable energy penetration to formulate the grid outage maintenance schedule. The key to optimizing the grid outage maintenance schedule lies in exploring the near-optimal combination of tripped equipment to be maintained within a given period. Minimizing the impact of maintenance on grid safety and stability is crucial for ensuring the security, supply, and renewable penetration of the power system. However, optimization based on manual experience suffers from issues such as misjudgments, oversights, low scheduling efficiency, and difficulties in sharing and inheriting experience. The optimization of maintenance schedules is inherently a multi-objective combinatorial optimization problem. Presently, researchers are attempting to employ optimization algorithms for the scheduling of outage maintenance schedules: Literature [1] proposes a bi-level optimization model based on Mixed Integer Linear Programming (MILP) to address complex energy management issues in microgrids. Specifically considering preventive maintenance and demand response. Simulations were conducted on a single-node system. Literature [2] applies an improved Particle Swarm Optimization algorithm to the annual

optimization planning of distribution networks; however, the number of devices optimized for the entire year was less than the number of months, and the experimental results demonstrated limited practical engineering applicability. In [3], a restructured Independent System Operator used a cyclic algorithm to generate a Maximum-Reliability Maintenance Scheduling (MR-MS) plan and a Maximum-Profit Maintenance Scheduling (MP-MS) plan within the grid it manages. By continuously adjusting algorithm parameters, the MP-MS plan was made to approach the MR-MS plan in terms of reliability. Providing an effective framework for formulating maintenance schedules for generation equipment in restructured power systems. Literature [4] proposes a machine learning surrogate-based stochastic optimization method for medium-term outage scheduling in power systems. Considering multi-timescale decision-making and uncertainties. Optimizing the outage maintenance schedule through distributed computing. Ref. [5] proposes a coordinated maintenance framework for transmission components, using a Markov model and aperiodic strategies optimizing equipment maintenance timing and reliability. In [6] addresses maintenance issues in power systems during ice disasters by proposing a multi-stage conditional maintenance scheduling method, which considers changes in equipment state and optimizes maintenance strategies to improve system reliability. Literature [7] presents a medium-term preventive maintenance scheduling method based on decision trees and mixed integer linear programming, which optimizes the maintenance intervals and costs for thermal power units. Ref. [8] proposes an extended framework based on the Manufacturer Usage Description (MUD) standard. This framework enhances the expressiveness of MUD. It can automatically generate behavioral profiles. It extends security assessment results from the design phase to runtime. Thus, it achieves dynamic attack surface reduction. It also enforces mandatory execution of security policies. Ref. [9] introduces AI-driven solutions to combat IoT-specific threats and offers practical guidance for developing adaptive IDS in resource-limited environments. Literature [10] proposes an optimal allocation and sizing method for solar and wind power units based on the Gravitational Search Algorithm (GSA), which enhances system economic efficiency and renewable energy integration. Literature [11] proposes an adaptive optimal control strategy based on an integral reinforcement learning (IRL) algorithm for automatic generation control in islanded microgrids, which significantly enhances frequency regulation performance under various uncertainties. Literature [12] proposes a hierarchical game method for integrating grid-forming energy storage into multi-energy systems, employing a chaos-enhanced IMTBO algorithm that reduces voltage deviation by 23.77% and improves overall operational efficiency. Literature [13] proposes a computational model for medium-term transmission system scheduling, and optimizes long-term performance of maintenance policies, considering manpower constraints and operational uncertainties through Monte Carlo simulation and heuristic algorithms.

However, the aforementioned research on outage maintenance scheduling shares some common issues: (1) the excessively high level of detail in algorithmic models results in long processing duration; (2) outage maintenance scheduling constitutes a non-convex optimization problem, and the use of intelligent optimization methods easily leads to convergence at local optima.

As a major branch of machine learning, reinforcement learning has demonstrated exceptional performance in numerous optimization and decision-making scenarios in recent years [14,15]. Literature [16] employs a multi-agent coordinated deep reinforcement learning algorithm, designing a dual-network structure and a competitive reward mechanism to address the complex optimization challenges in distributed non-convex economic dispatch caused by grid expansion. The method enables real-time and stable optimization of power output under time-varying demands through collaborative learning of independent agents and joint reward updates, overcoming the convergence difficulties of traditional

methods in non-convex scenarios. Literature [17] investigates a model-free economic dispatch approach for multiple Virtual Power Plants (VPPs) that does not rely on accurate environmental models. It proposes an adversarial safety reinforcement learning method, which enhances the safety of actions and renders the model robust against discrepancies between the training environment and the testing environment. Literature [18] proposes an optimal scheduling method for shared energy storage systems (SESS) that accounts for distribution network (DN) operational risks. Specifically, it develops a multi-objective day-ahead scheduling model incorporating both explicit operational costs and reliability costs, integrates an island partitioning model with "maximum island partitioning—island optimal rectification" to accurately calculate reliability costs, and employs an improved genetic algorithm to solve the complex scheduling model—ultimately achieving efficient and reliable optimization of distribution networks under dynamic operating conditions. However, reinforcement learning strategies often operate as "black boxes" lacking transparent decision-making rationale. Their application in critical domains faces significant challenges, particularly concerning interpretability issues. Literature [19] proposes an interpretable anti-drone decision-making method based on deep reinforcement learning. By incorporating a double deep Q-network with dueling architecture and prioritized experience replay, the DRL agent was trained to intercept unauthorized drones. To enhance the transparency of the decision process, Shapley Additive Explanations (SHAP) were applied to identify key factors influencing the agent's actions. Additionally, experience replay was visualized to further illustrate the agent's learning behavior. Ref. [20] introduces an interpretable compressor design method combining deep reinforcement learning (DDPG) and decision tree distillation. The approach co-optimizes 25 design variables, achieving 84.65% isentropic efficiency and 10.75% surge margin in a six-stage axial compressor. SHAP and bio-inspired decision trees provide explicit design rules, enhancing both performance and interpretability for engineering applications.

In the field of power systems, the literature [21] proposes Deep-SHAP, a backpropagation-based SHAP explainer tailored for deep reinforcement learning models in power system emergency control. It combines Shapley additive explanations with the efficient DeepLIFT backpropagation strategy to attribute DRL decisions to individual input features. A softmax layer converts raw SHAP scores into probability distributions, enabling intuitive interpretation. Features are categorized by their temporal and spatial properties, and both global (model-wide) and local (per-instance) analyses are delivered through stacked bar charts, bee-swarm plots, and 3D visualizations, allowing operators to see exactly which voltage and load variables drive load-shedding actions. Literature [22] proposed a green learning (GL) model for regional power load forecasting, which is structured into three modules: unsupervised representation learning, supervised feature learning, and supervised decision learning. By combining mixed feature extraction, seasonal climate features, and the Quantile Autoregressive Forest (QARF) algorithm, it optimizes parameters with seasonal aggregation and avoids the backpropagation mechanism, achieving the dual goals of reduced energy consumption and improved prediction accuracy while reducing computational complexity.

The major challenges for the current research can be concluded: (1) the optimal outage scheduling problem requests to allocate the maintenance tasks for the whole month, which is a non-convex, high dimension optimization problem; (2) due to the complexity of the problem, the traditional heuristic algorithms are with low computational efficiency; (3) interpretability and the transparency of the data-driven methods need to be enhanced.

To address the aforementioned challenges, this paper proposes an interpretable reinforcement learning-based optimization method for monthly outage maintenance scheduling in power grids. Inspired by Literature [23], this study focuses on the key factor of

reward in the reinforcement learning process and investigates a reward-level interpretability approach. First, a single-agent reward function is designed to ensure the existence of an optimal reward point within the action space. Then, the Shapley value method is employed to select the main features and reasonably decompose rewards. By leveraging the interpretability of Shapley values, we have overcome challenges such as an excessively large state space and the detrimental impact of meaningless states on optimization performance and speed. This approach enhances optimization effectiveness and model transparency, enabling its application in sensitive and confidential domains while maintaining strong engineering applicability. Next, the received monthly outage maintenance requirements are modeled as a Markov decision process, defining the corresponding state space, action space, and objective function. Finally, interpretable reinforcement learning is applied to optimize the monthly outage maintenance schedule for the power grid.

The contributions of this work can be concluded as follows: (1) A reinforcement learning (RL)-based methodology is proposed for monthly outage scheduling in power systems. (2) The rationale behind state space formulation is clarified and enhanced using Shapley values, improving the interpretability of feature selection. (3) Visual validation is conducted on the IEEE 39-bus system, demonstrating the effectiveness and transparency of the proposed approach.

The rest of this paper is constructed as follows: Section 2 formulates the optimization model for outage maintenance scheduling, including the construction of the objective function, the definition of the optimization variable set, and the establishment of operational constraints. Section 3 presents a reward decomposition and state shaping method based on Shapley values to enhance the interpretability of the decision process. Section 4 describes the selection of the reinforcement learning algorithm and the design of the state and action spaces, along with a visualization of the training results. Section 5 concludes that the integration of Shapley values with reinforcement learning contributes to a reduction in voltage violations, decreased active power losses, improved power flow convergence, enhanced operational security, and increased scheduling efficiency.

## 2. Power Outage Maintenance Schedule

### *2.1. Power Outage Maintenance Schedule Optimization Model*

In the problem of power outage maintenance schedules for $M$ devices, the length of the power outage maintenance schedules time window for device $m$ is denoted as $T_m$. In the power outage maintenance schedules submitted by the operation and maintenance units, the starting date of the power outage maintenance of device $m$ is $t_m$, and the power outage time window of device $m$ can be expressed as:

$$E_m = \{t_m, t_m + T_m - 1\} \tag{1}$$

The monthly power outage maintenance schedules can be defined as set $T$, which consists of the power outage time windows of all devices.

$$T = \{E_m | m = 1, 2, \ldots M\} \tag{2}$$

where a natural month ($NM$) is defined as:

$$NM = \{1, 2, \ldots, D_{max}\} \tag{3}$$

where $D_{max}$ means the maximum number of days in a natural month.

In order to intuitively show the current monthly maintenance arrangement of the power grid, this study introduces visualization technology to represent the outage schedule

in the form of a Gantt Chart. The Gantt chart, where $G(i,m) = 1$, indicates that equipment m is in a state of outage maintenance on day $i$ of the natural month; $G(i,m) = 0$ indicates that the device is operating normally on day $i$. Additionally, the power flow state of the typical grid model corresponding to the $i$-th day of the month is calculated using the power flow solver, denoted as $pf_i$, and the daily power flow results for the natural month are aggregated:

$$PF = \{pf_i | i = 1, 2, \ldots D_{max}\} \tag{4}$$

$pf_i$ is a vector that contains electrical quantities related to lines, nodes, and generators within the current network topology

A 30-day natural-month power outage maintenance schedules Gantt chart is shown in Figure 1. According to the Gantt chart, the power outage time window of some devices can be obtained:

$$E_1 = \{26, 27, 28\},\ E_3 = \{1, 2, 3\},\ E_{M-2} = \{2, 3, 4\},\ E_{M-1} = \{28, 29, 30\}.$$

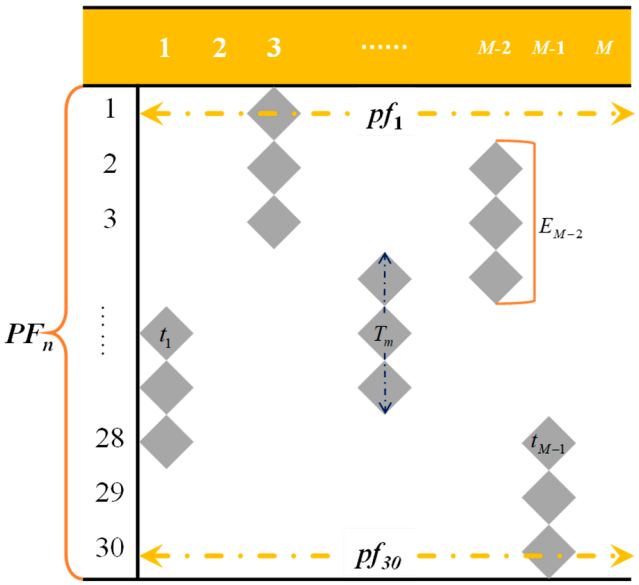



**Figure 1.** Gantt chart of power outage maintenance schedules.

### *2.2. Objective Function*

Considerations should be given to the secure and stable operation of the power grid, the assurance of power quality, and implementation of cost control when formulating power outage maintenance schedules. Accordingly, this study decomposes the objective function into three sub-functions, as defined by Equations (5)–(7).

Voltage violation can damage the power grid and user equipment. Therefore, preventing voltage violation is the core task for maintaining the basic power quality of the power grid. The objective function of the voltage violation is defined as:

$$\theta(v) = \sum_{i=1}^{D_{max}} \left(V_{day} \leq 0.95\right) \cdot 1 + \sum_{i=1}^{D_{max}} \left(V_{day} \geq 1.05\right) \cdot 1 \tag{5}$$

where vector $V_{day}$ represents the per-unit value of the voltage of each bus bar on specific days. 1 denotes a unit vector.

Considering the economic operation of the power grid, the active power loss of the power grid is considered as one of the optimization objectives. The objective function of the active power loss is defined as:

$$\sigma(P) = \sum_{i=1}^{D_{max}} P_{loss} \tag{6}$$

where $P_{loss}$ represents the active power loss in the transmission line, which is calculated using the power flow solver.

In addition, in the scheduling process of the power outage maintenance schedules, there is a risk of non-convergence in the power flow calculation, which affects the feasibility and safety of the maintenance plan. Therefore, the objective function of power flow convergence is defined as:

$$\xi(flag) = \sum_{i=1}^{D_{max}} flag_{success} \tag{7}$$

where $flag_{success} = 0$ indicates that the power flow calculation of the day is not convergent, and $flag_{success} = 1$ indicates that the power flow calculation of the day is convergent.

Finally, based on the above sub-objective functions, the overall optimization objective function is defined as:

$$F = min[\alpha \cdot \theta(V) + \beta \cdot \sigma(P) + \gamma \cdot \xi(flag)] \tag{8}$$

where $\alpha$, $\beta$, and $\gamma$ are coefficients of the opimizaition problem.

### *2.3. Optimization Constraints for Power Outage Maintenance Scheduling*

#### 2.3.1. Mutual Exclusion Constraint

When multiple devices are scheduled for maintenance simultaneously, mutual exclusion issues may arise, leading to unnecessary power supply interruptions in some buses in the power grid. To avoid such problems, the agent should consider the mutually exclusive relationship of the equipment when formulating the maintenance plan, and arrange the maintenance time of the mutually exclusive equipment through peak shifting to ensure the continuous and stable operation of the power system as much as possible. For any two mutually exclusive devices $m$ and $n$ ($m, n \in \{1, 2, \ldots, M\}, n \neq m$), in the system, the mutually exclusive constraints in the monthly power outage maintenance schedules can be expressed as:

$$E_m \cap E_n = \varnothing \tag{9}$$

where $t_m$ denotes the start date of the equipment maintenance and $T_m$ denotes the length of the maintenance time window of the device.

#### 2.3.2. Co-Termination Constraint

When the power outage maintenance of one device affects the normal operation of other devices, the two devices should be maintained simultaneously to minimize the number of outages and ensure the quality of the power supply. For equipment $m$ and $n$ ($m, n \in \{1, 2, \ldots, M\}, n \neq m$) that need to be overhauled simultaneously, the same stop constraint condition can be expressed as follows:

$$E_m \cap E_n = \begin{cases} E_m, length(E_m) > length(E_n) \\ E_n, length(E_m) \leq length(E_n) \end{cases} \tag{10}$$

where $E_m$ and $E_n$ are the maintenance time windows of devices $m$ and $n$, respectively, defined according to Equation (1).

#### 2.3.3. Power Outage Time Window Constraints

From the perspective of human resource allocation and maintenance efficiency, the outage maintenance time window of a single equipment should remain continuous. In the

monthly outage schedule, for the device $m$ ($m, n \in \{1, 2, \ldots, M\}, n \neq m$) the outage time window constraint can be expressed as:

$$length(T_m) \in N \tag{11}$$

## 3. Reinforcement Learning Method Based on Interpretability

This paper proposes a power outage maintenance schedule optimization method based on interpretable reinforcement learning. By formalizing the power outage maintenance schedule optimization problem as a Markov decision process (MDP), we construct the state space, action space, and reward function, and optimize the scheduling strategy based on DQN.

### *3.1. Key Feature Interpretability*

Since reinforcement learning strategies typically exist as "black boxes," researchers face significant challenges when applied in critical fields, particularly in terms of interpretability. To address this issue, Reference [23] proposes an interpretability analysis method based on the Shapley value. This method treats each state feature in a Markov decision process as a participant in a cooperative game. By calculating the Shapley value of each feature during the decision optimization process, it quantifies the contribution of each state feature to the final optimized result, thereby providing an intuitive and quantitative explanation. This mechanism helps enhance users' understanding and trust in the decision-making process of the agent, thereby improving the explainability and acceptability of reinforcement learning models.

In a cooperative game with $N$ participants, the Shapley value can be used to calculate the average marginal benefit of participant $i$, expressed as:

$$\Phi_i(v) = \frac{1}{N!} \sum_{\pi \in \Pi(N)} \{\tau(P_i{}^{\pi} \cup \{i\}) - \tau(P_i{}^{\pi})\} \tag{12}$$

where $i = \{1, 2, \ldots, N\}$ denotes the cooperating parties; $\Pi(N)$ denotes all permutations and combinations of $N$ participants, containing $n!$ elements; $v$ represents the cooperative profit function, also known as the characteristic function, whose domain is all subsets of $\Pi(N)$. It satisfies $v(\varnothing) = 0$ and superadditivity, for any $S, T \subseteq N$ such that $S \cap T = \varnothing$, we have $v(S \cup T) \geq v(S) + v(T)$.

The interpretability discussed in this paper pertains to the reward level. The Shapley value calculates the marginal contribution of a specific participant to the reward function, representing the importance of a particular plan or feature within the power outage maintenance schedules. Furthermore, the reward-level interpretability explored in this paper encompasses two aspects: first, reward decomposition based on the Sharpley value method; second, state shaping, which utilizes the Sharpley value to extract key states for optimizing power outage maintenance schedules, thereby enhancing the optimization efficiency of reinforcement learning agents.

### *3.2. State Space*

Markov decision processes (MDPs) exhibit a memoryless property, meaning the system's next-time-step state is determined solely by the current state and the action executed.

The $N$ minus $k$ criterion framework in power systems [24] proposes that a system comprising $N$ units can tolerate failures in up to $k$ units, while the failure of $k + 1$ units may cause grid collapse. Based on interpretable results and the $N$ minus $k$ criterion this paper defines the first-dimensional state $S_1$ as follows, using the daily set of equipment undergoing maintenance in the current month's Gantt chart as the state:

$$s_1 = \max_{day \in [1, D_{max}]} \sum_{m=1}^{M} G(i, m) \tag{13}$$

where $G(i, m)$ denotes the element in row $i$ and column $m$ of the Gantt chart, representing the maintenance status of equipment $m$ on day $i$.

Additionally, since planned power outages affect power flow, the source-load status of the power grid is also included in the state space.

$$s_2 = \{\boldsymbol{P_{gen}}, \boldsymbol{P_{load}}\} \tag{14}$$

where $\boldsymbol{P_{gen}}, \boldsymbol{P_{load}}$ denote the power generation and load vectors of the power grid.

The reinforcement learning state space is composed of $s_1$ and $s_2$, forming $s = \{s_1, s_2\}$.

### *3.3. Action Space*

For equipment $m$ with a maintenance time window length of $T_m$, if maintenance commences at date $T_m$, the action space $A_m$ for equipment $m$ can be defined as:

$$A_m = [-(D_{max} - T_m), \ldots, -1,\ 0,\ 1, \ldots, D_{max} - T_m] \tag{15}$$

Negative numbers indicate the number of days the maintenance window is advanced, while positive numbers indicate the number of days it is delayed.

The overall action space encompassing all device action spaces is defined as:

$$A = \{A_1, A_2, \ldots, A_M\} \tag{16}$$

### *3.4. Reward Function*

The reward function determines the reward an agent receives for selecting a specific action when transitioning from state $s_t$ at time t to a new state $s_{t+1}$. Using Equation (8) as the mapping rule, the state evaluation function can be expressed as follows to assess the state change from $s_t$ to $s_{t+1}$ before and after taking an action:

$$f(s_t) = \begin{cases} -\theta(V) \times 10 - \sigma(P); & \xi(flag) = D_{max} \\ -10,000; & \xi(flag) < D_{max} \end{cases} \tag{17}$$

In the optimization process, the number of voltage limit violations in the 39-bus system over one month is on the order of hundreds. The amount of power loss remains largely within the range of 1000 to 2000 MW. Therefore, we set a coefficient 10 in (17). Given that active network losses are typically measured in megawatts (MW) per hour and occur at the thousand-level magnitude, while the number of daily voltage exceedances in a given month is around a hundred, we set the weight for voltage exceedances to 10.

Based on the state evaluation function, the reward $r_t$ can be expressed as the difference between $f(s_{t+1})$ and $f(s_t)$. Furthermore, to minimize labor costs in maintenance scheduling and enhance the action efficiency and reliability of the scheduling strategy, it is necessary to reduce the number of actions required to achieve optimal results.

$$r_t = f(s_{t+1}) - f(s_t) - \sum_{m=1}^{M} |\Delta t_m| \tag{18}$$

### *3.5. Optimization Process Based on Deep Reinforcement Learning*

To achieve intelligent optimization of power outage maintenance schedules, this paper employs a reinforcement learning approach based on Deep Q-Learning ($DQN$). First, using the method described in Section 2.1, the raw power outage maintenance

schedules submitted by operational units are converted into an initial Gantt chart $G_0$, and the corresponding power flow state $PF_0$ is calculated. Based on this, the outage maintenance plan is iteratively optimized through reinforcement learning. Action selection within the action space employs an $\varepsilon$-greedy strategy. During the action selection phase, the action index in $DQN$ must be mapped to the corresponding action space and ultimately to the corresponding Gantt chart, as illustrated below.

Define $size(\cdot)$ as the number of elements in the current set. The number of equipment units undergoing maintenance in the current month, $k$, can be expressed as:

$$k = size(T) \tag{19}$$

The action index of a single device can be expressed as:

$$num_{am} = size(A_m) = 2(D_m - T_m) + 1 \tag{20}$$

In each iteration, the current state $s_t$ is computed according to Equation (2), and device $m$ is selected. Then, an action $a_t$ is chosen from the action space $A_m$ of device $m$. Consequently, the maintenance time window for device $m$ changes from $E_m$ to $E'_m$, where $E'_m$ can be expressed as:

$$E'_m = [t_m + a, t_m + T_m + a_t - 1] \tag{21}$$

At this point, the Gantt chart $G_t$ is iterated to $G_{t+1}$ based on $E'_m$, and the new state $s_{t+1}$ is computed using Equation (12).

Additionally, the reward value $r_t$ is computed using Equations (16) and (17). Finally, the current state value $Q(s_t, a_t)$ is calculated based on the Bellman equation, and the value function is updated, completing a full iteration of the value function.

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\left[r_t + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1})\right] \tag{22}$$

The implementation process of deep reinforcement learning is briefly described as follows: Construct a deep neural network ($DQN$) based on the state space and action space to approximate the $Q$-function. The number of neurons in the output layer is 270, the network is designed with 3 hidden layers, and each layer has 256 neurons. learning rate is 0.001. The size of the replay buffer is 2000. At each time step, an action is selected using an $\varepsilon$-greedy policy based on the current state. The selected action is applied to the environment, observing the new state, reward, and termination status. The current experience tuple $(s_t, a_t, r_t, s_{t+1})$ is stored in the experience replay buffer. Randomly sample a batch of experience tuples from the experience replay buffer; compute the target $Q$-value for each tuple; update the neural network parameters via backpropagation using the mean squared error loss function to make predicted $Q$-values closer to target $Q$-values; Periodically update the parameters of the target network to align with those of the main network, stabilizing the training process; Continuously repeat the steps of sampling, storing, sampling-based updates, and target network updates until the agent converges within the environment.

## 4. Case Study and Result Analysis

In order to validate the effectiveness and applicability of the proposed reinforcement learning-based interpretable optimization method, this paper conducts simulation tests on the IEEE 39-node system. By comparing the reinforcement learning optimization algorithm put forward herein with the conventional Genetic Algorithm (GA) approach [25], an assessment is made of the differences in their optimization performance with regard to the outage maintenance scheduling problem.

### *4.1. Simulation System Setup and Agent Training*

This study utilizes the IEEE 39-bus system [21] as the environmental platform. The system consists of 46 transmission lines and 10 generators, for a total of 56 devices ($M = 56$). For the purpose of maintenance scheduling optimization simulation, the natural month length is set to 30 days ($D_{MAX} = 30$). During the simulation, five units are selected at random for maintenance. Their numbers in the system topology are designated as 2, 10, 28, 31, and 44, respectively. The system topology is illustrated in Figure 2.

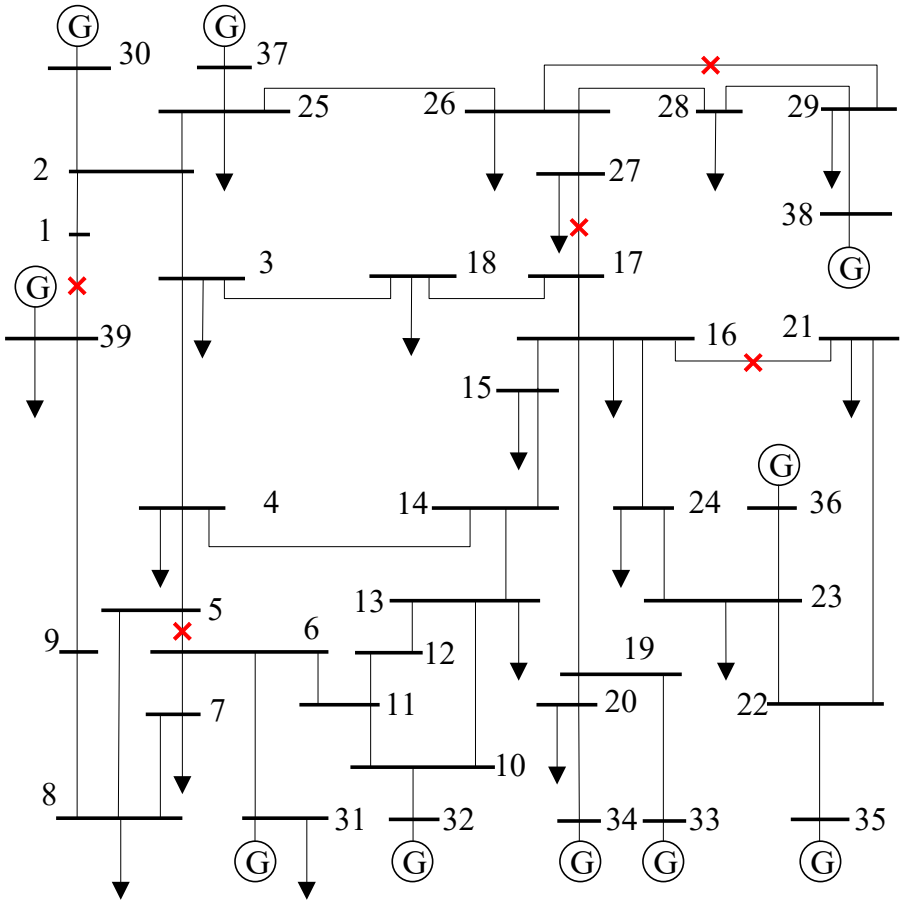



**Figure 2.** Maintenance schedule topology of the IEEE 39-bus system.

In order to standardize the maintenance conditions, the time window length for all equipment awaiting maintenance is set to $T_m = 3$ days, and the initial power outage schedules are randomly assigned. The hyperparameter settings for optimization are as follows: 1200 training iterations; an initial step size of 80 for each iteration; and an exploration rate of 0.9.

Figure 3 presents the variation trend of reward values across individual training episodes during the reinforcement learning training process. The horizontal axis represents the training episode number, and the vertical axis denotes the reward value associated with all actions performed in each episode. As observed in Figure 3, the reward value increases consistently as training progresses and eventually stabilizes. This pattern indicates that the model gradually converges throughout the learning process, thereby achieving favorable training performance.

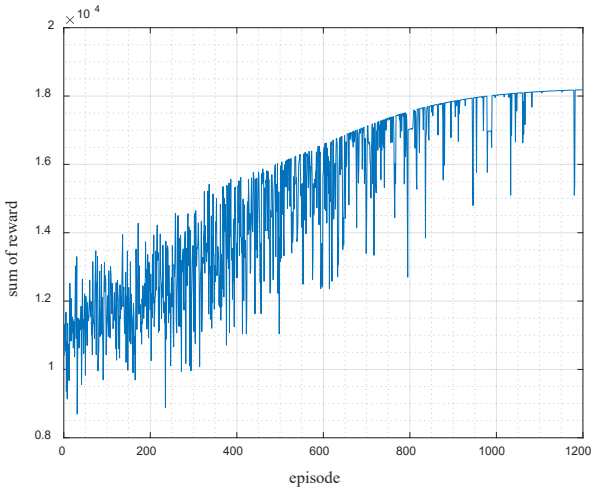



**Figure 3.** Sum of reward in the IEEE 39-bus system.

*4.2. Interpretable Quantification Based on Shapley Value*

In order to enhance the interpretability of decisions made by reinforcement learning algorithms, the present study employs the Shapley value to quantitatively analyze the marginal effects of different equipment combinations in outage scheduling.

The interpretability of key features in power outage maintenance schedules is illustrated by the following example. The selected set of maintenance equipment is represented by the following numbers: $\{2, 10, 28, 31, 44\}$. These numbers represent the participants in a cooperative game. According to the principles of cooperative game theory, the set of all possible cooperative combinations, denoted by $\Pi(N)$, can be expressed as follows:

$$\Pi(N) = \{\{2\}, \{10\}, \ldots, \{2, 10\}, \{2, 28\} \ldots, \{2, 10, 28, 34, 44\}, var(T), K\} \quad (23)$$

where var$(T)$ denotes the variance of each element in the power outage maintenance schedules $T$; $K$ represents the maximum number of outage facilities at any given moment within the outage plan.

The Shapley value calculation uses the composite objective function defined in Equation (8) as the cooperative benefit function $v(\cdot)$. Using historical data from the existing grid topology, we evaluate the marginal contribution of each equipment combination to grid operational performance.

As illustrated in Figure 4, the Shapley values for each feature are presented. The horizontal axis signifies the corresponding influencing factor, with each factor encompassing all historical samples. The vertical axis of this figure indicates the Shapley value for a given factor, while the color axis denotes the magnitude of the feature value for each historical sample. In this context, purple signifies smaller values and red indicates larger values. The greater the dispersion of data points across different colors, the more significant the feature's impact on optimization results, warranting particular attention during the optimization process. Figure 3 reveals that feature $K$ exhibits the most pronounced impact, with a distinct red-purple boundary. The red data points, which correspond to larger $K$ values, exert a negative influence on the objective function and reduce grid stability. In contrast, purple data points correspond to smaller $K$ values, which have been shown to enhance grid stability. However, the variance of the power outage maintenance schedules (var$(T)$) exhibits a weaker influence. The data points, which are represented by various colors, are intermingled along this feature axis, thereby obscuring the clear delineation of correlation boundaries. This finding suggests that the variance of the power outage

maintenance schedules ($var(T)$) offers a limited degree of insight into grid performance. The other operational condition features demonstrated an absence of distinct color boundaries, suggesting their negligible contribution to the objective function. The problem of optimizing the outage schedule can be conceptualized as a non-convex, high-dimensional NP-hard problem. Without extracting key features, the problem becomes significantly more difficult to solve. The present study employs the Shapley value method, a technique that has been demonstrated to enhance computational efficiency and improve interpretability of reinforcement learning.

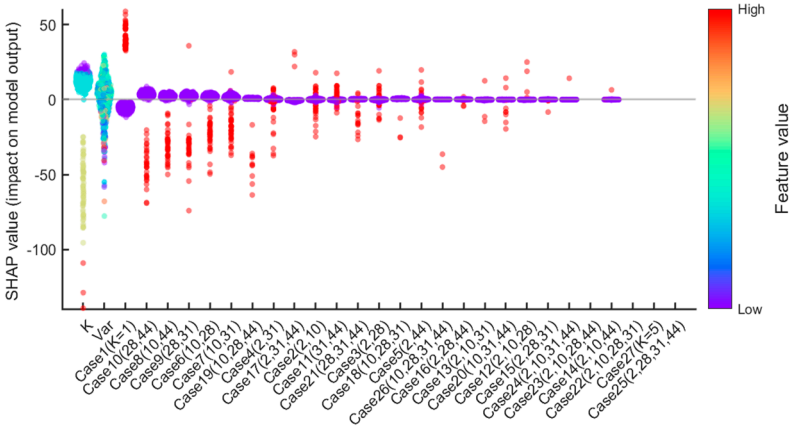



**Figure 4.** Shapley value characteristic analysis in the IEEE 39—bus system.

### *4.3. Analysis of Optimization Results*

This study employs the reinforcement learning algorithm proposed in this paper to optimize the power outage maintenance schedules for the IEEE 39-bus system. Herein, the action space for a single equipment can be formulated as follows:

$$A_m = \{-27, \ldots 0, \ldots, 27\}, m \in \{2,\ 10,\ 28,\ 31,\ 44\} \tag{24}$$

The operational space of all equipment can be represented as:

$$A = \{\, A_2, A_{10}, A_{28}, A_{31}, A_{44}\} \tag{25}$$

Figure 5 illustrates the differences in the equipment maintenance schedule before and after optimization; Figure 6 employs the Shapley value to demonstrate the interpretability of power outage maintenance schedules optimization. Figures 7 and 8, in turn, present the number of violation occurrences, active power losses, and the number of power flow non-convergence events.

Figure 5 provides a visual comparison of the equipment maintenance schedule before and after optimization. Figure 5a illustrates the initial maintenance plan. The horizontal axis represents equipment IDs, and the vertical axis denotes dates within calendar months. As can be seen in the figure, all equipment outages in the original plan were concentrated at the beginning of the month, indicating significant resource congestion and operational risk. Figure 5b shows the scheduling results after optimizing the reinforcement learning algorithm. Maintenance periods for each piece of equipment are now dispersed throughout the month, which significantly improves resource allocation and system safety. From the perspective of state variables, the pre-optimization system state was $K = 5$, meaning the maximum number of equipment units undergoing maintenance on a single day was five. Post-optimization, the state decreased to $K = 1$, leading to a more balanced distribution of system load. This validates the significant optimization achieved in state scheduling performance.

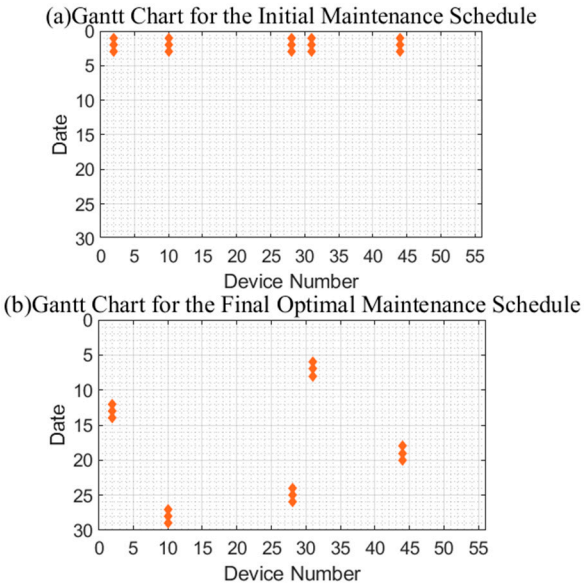



**Figure 5.** Comparison of Gantt charts of the IEEE 39-bus system before and after optimization.

To demonstrate the interpretability of the power outage maintenance schedules optimization results, this paper treats each entry in the power outage maintenance schedules as a participant in a cooperative game. It randomly records the adjustment outcomes of 1000 iterations, which means, randomly adjusting each power outage plan, obtaining reward/penalty values based on the reward function, normalizing them, and then calculating the marginal effect of each entry's contribution to the adjustment. The results are shown in Figure 6. It can be observed that the marginal effect of contribution value for adjustment item 2 is the highest. Combined with the optimization results in Figure 5, the second item exhibits the largest adjustment magnitude for equipment, explaining the optimization outcomes.

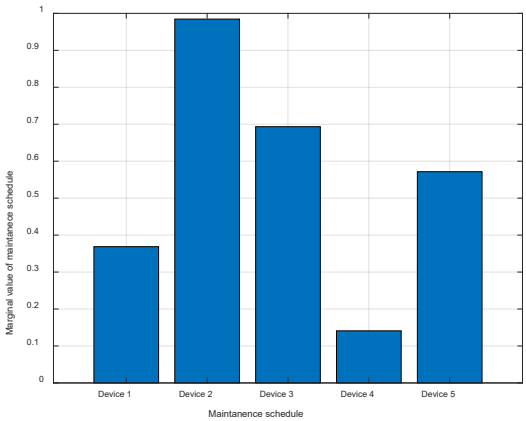



**Figure 6.** Quantification of interpretability for outage scheduling.

Figure 7 shows a comparison of active power losses across system branches before and after optimization. The horizontal axis represents the date, the vertical axis denotes branch numbers, and the depth dimension (*z*-axis) shows the active power loss value for each branch on a given date. Figure 7a is plotted based on the initial maintenance plan, showing that the active network losses of each branch across different dates are relatively scattered and generally at a high level. Figure 7b, by contrast, is based on the optimized scheme; the active power losses of all branches are significantly reduced and more uniformly distributed, which reflects the effectiveness of the proposed method in

lowering the network's operational energy consumption. In the case study, the average daily reduction in active power losses reaches 50 MWh per month.

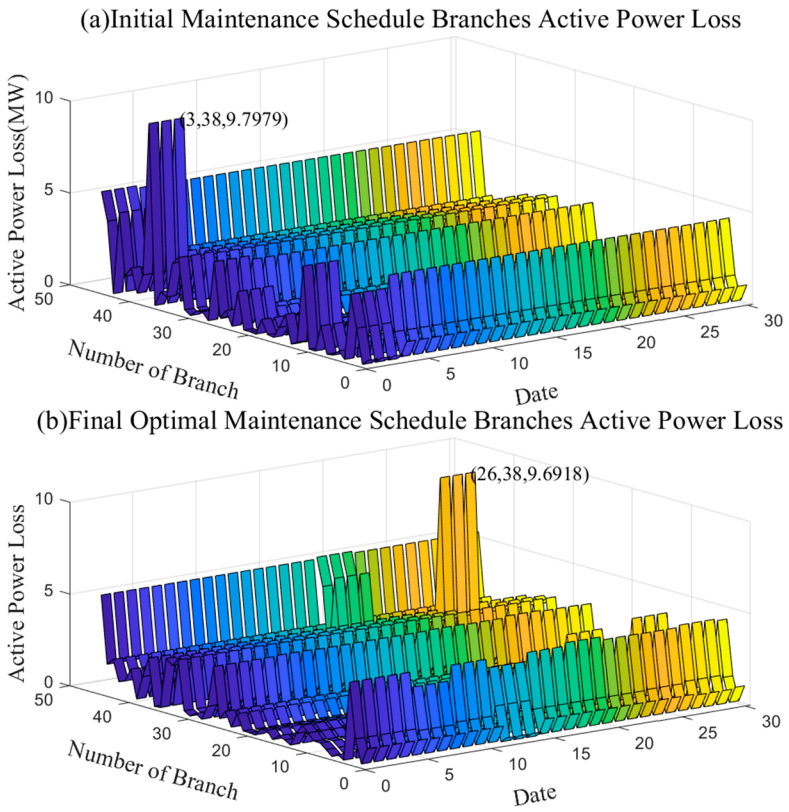



**Figure 7.** Active power loss on each branch in the IEEE 39-bus system before and after optimization.

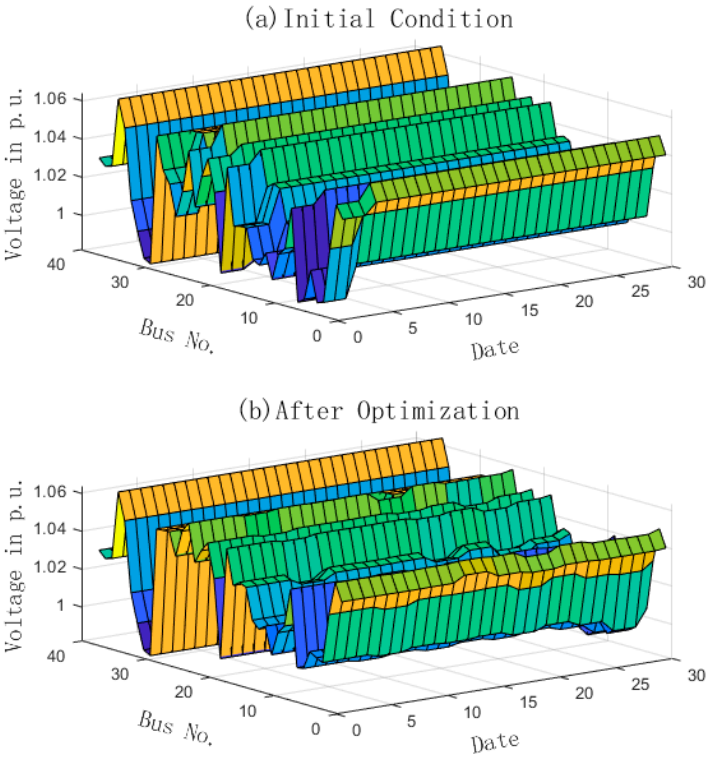



**Figure 8.** Voltage per-unit value on each bus in the IEEE 39-bus system before and after optimization.

Figure 8 compares the voltage magnitudes at various nodes before and after optimization. Figure 8a illustrates the voltage scaling value distribution under the original plan,

revealing significant voltage violation issues in some nodes. Figure 8b corresponds to the optimized plan. In this plan, voltages are maintained within a more stable range. This results in an improvement in system voltage quality.

To further validate the comprehensive performance of the algorithm proposed in this paper, this study selects the Genetic Algorithm (GA) optimization algorithm [20] as a benchmark. Additionally, the reinforcement learning algorithm results presented in the table correspond to scenarios where the Shapley value is not considered—specifically, the maximum number of outage devices at a single time (denoted as $K$) in the outage plan is not incorporated into the state space. In contrast, the interpretable reinforcement learning algorithm accounts for the interpretable insights derived from the Shapley value. This study conducts comparative analyses on metrics, including monthly voltage violation occurrences, total active power losses, and computational time within the optimization objectives. Table 1 presents a comparison of the results of the five algorithms. Power outage maintenance schedules optimization can be considered an NP-hard problem. By introducing the Shapley value, this research enhances the interpretability and computational efficiency.

**Table 1.** Compared to the GA method.

| **Algorithm Performance** | **Method** | | | | |
|---|---|---|---|---|---|
| | **Genetic Algorithm [19]** | **Particle Swarm Algorithm** | **Simulated Annealing** | **Reinforcement Learning Algorithms** | **Interpretable Reinforcement Learning Algorithms** |
| Monthly voltage violation (*Times*) | 6 | 8 | 10 | 0 | 0 |
| Monthly active power Losses (GWh) | 974.2 | 976 | 976.8 | 975 | 919 |
| Computational time (*Hours*) | 14 | 8.7 | 7 | 10 | 0.2 |

The experimental results indicate that the explainable reinforcement learning method proposed in this paper not only outperforms traditional algorithms in terms of optimization quality but also demonstrates higher efficiency in terms of optimization time. It can complete maintenance scheduling optimization tasks in complex power grid environments more rapidly, thereby exhibiting strong potential for engineering applications.

According to Table 1, the last column represents the method proposed in this article, and compared to the first three columns, each row represents an improvement in different aspects. The host processor used in this paper is an Intel i9 processor with an 8-core and 16-thread architecture. The memory used is 512 GB, and the GPU is GeForce RT 3070 Ti. The computational time depends on the hyperparameter tuning. Above mentioned algorithms have been tuned based on experience, but parameters can be changed in different scenarios. By integrating interpretability, the key feature of the problem is quantified and involved in the solving process, which accelerates the processing duration.

### *4.4. Scalability Study in IEEE 118 Bus System*

To evaluate the performance of the proposed method, we implement out algorithm into IEEE 118 bus system. The accumulated reward in each episode is given in Figure 9 and the optimization results are demonstrated in Table 2. The computational time for IEEE 118 bus system is 1.2 h.

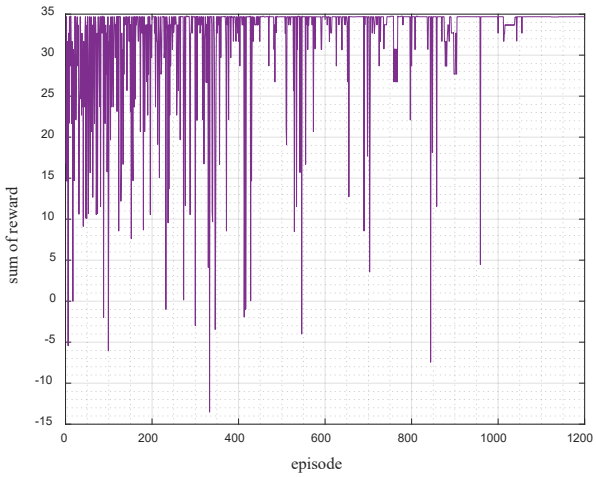



**Figure 9.** Sum of reward in the IEEE 118—bus system.

**Table 2.** Optimal Outage Maintenance Scheduling.

| Device Index | Initial Maintained Date | Optimized Outage Date |
|---|---|---|
| 24 | [1, 2, 3] | [7, 8, 9] |
| 118 | [1, 2, 3, 4] | [8, 9, 10, 11] |
| 152 | [1, 2, 3] | [14, 15, 16] |
| 169 | [1, 2, 3] | [24, 25, 26] |
| 170 | [1, 2, 3, 4, 5] | [2, 3, 4, 5, 6] |

We select five lines to maintain in the month, whose index are 24, 118, 152, 169, 170. The initial start date is set to the first day for all selected lines. The outage duration is randomly from three days to five days. After optimization, outage schedule for device 24 is from 7th to 9th of the month, device 118 closes from 8th to 11th, device 152 maintains from 14th to 16th, device 169 trips from 24th to 26th, and the outage schedule of device 170 is from 2nd to 4th of the day. The reward results show that the agent converges to average maximum reward after 1000 episodes. The optimal maintenance outage scheduling improves the system performance in voltage violation, and losses. The number of the voltage

## 5. Conclusions

This paper proposes an interpretable reinforcement-learning optimization method for power system outage maintenance scheduling. The approach simultaneously targets security, stability, power quality, and economy, while explicitly handling constraints during the solution process. Shapley values are employed to identify the most influential features of outage maintenance scheduling, after which a Markov decision process is formulated and solved by a reinforcement-learning agent. Case studies show that the method reduces voltage violations, lowers total active power losses, improves power flow convergence and system security.

Future work will incorporate user-side load forecasting into power outage maintenance planning considerations. Grid status indicators such as line load factors, bus bar voltage levels, and rotating reserve capacity will be factored in, while graph neural networks will be employed to model the grid topology. A weighted multi-objective reward function framework will be adopted, with explicit justification provided for the weighting scheme. Methods such as post-event experience replay or curriculum learning will be employed. These approaches will be integrated with power outage scheduling optimization plans.

**Author Contributions:** Conceptualization, W.T.; methodology, W.T. and Z.D.; validation, K.L. and Z.C.; formal analysis, W.T.; investigation, X.M.; writing—original draft preparation, Z.D.; writing—review and editing, Z.D.; funding acquisition, W.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Science and Technology Project of State Grid (Research and application of multi-agent interpretable reinforcement learning general technology for comprehensive optimization of power grid blackout planning). Grant number: 5108-202420053A-1-1-ZN.

**Data Availability Statement:** Data is contained within the article.

**Conflicts of Interest:** Authors Wei Tang, Xun Mao and Kai Lv were employed by the State Grid Anhui Electric Power Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Science and Technology Project of State Grid. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

## References

1. Mazidi, P.; Bobi, M.A.S. Strategic maintenance scheduling in an islanded microgrid with distributed energy resources. *Electr. Power Syst. Res.* **2017**, *148*, 171–182. [CrossRef]
2. Li, R.; Liu, Y.; Ouyang, G.; Long, H.; Liu, P. Research on Optimization of Planned Interruption Non-Effect for Distribution Network Planning Based on Improved Particle Swarm Optimization Algorithm. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *252*, 032071. [CrossRef]
3. Conejo, A.; Garcia-Bertrand, R.; Diaz-Salazar, M. Generation Maintenance Scheduling in Restructured Power Systems. *IEEE Trans. Power Syst.* **2005**, *20*, 984–992. [CrossRef]
4. Dalal, G.; Gilboa, E.; Mannor, S.; Wehenkel, L. Chance-Constrained Outage Scheduling Using a Machine Learning Proxy. *IEEE Trans. Power Syst.* **2019**, *34*, 2528–2540. [CrossRef]
5. Moslemi, N.; Kazemi, M.; Abedi, S.M.; Khatibzadeh-Azad, H.; Jafarian, M. Maintenance Scheduling of Transmission Systems Considering Coordinated Outages. *IEEE Syst. J.* **2017**, *12*, 3052–3062. [CrossRef]
6. Niu, T.; Huang, Q.; Fang, S.; Chen, G.; Liao, R.; Zhang, L.; Folly, K.A. Multistage Condition-Based Resilient Maintenance Scheduling for Power Systems During Ice Storms. *IEEE Trans. Ind. Appl.* **2024**, *60*, 2201–2213. [CrossRef]
7. Abiri-Jahromi, A.; Fotuhi-Firuzabad, M.; Parvania, M. Optimized Midterm Preventive Maintenance Outage Scheduling of Thermal Generating Units. *IEEE Trans. Power Syst.* **2012**, *27*, 1354–1365. [CrossRef]
8. García, S.N.M.; Sánchez-Cabrera, A.; Schiavone, E.; Skarmeta, A. Integrating the manufacturer usage description standard in the modelling of cyber–physical systems. *Comput. Stand. Interfaces* **2023**, *87*, 103777. [CrossRef]
9. Ahanger, T.A.; Ullah, I.; Algamdi, S.A.; Tariq, U. Machine learning-inspired intrusion detection system for IoT: Security issues and future challenges. *Comput. Electr. Eng.* **2025**, *123*, 110265. [CrossRef]
10. Mukherjee, A.; Sarker, J. Optimal placement and sizing of renewable energy source-based generations in transmission system. *I-Manag. J. Electr. Eng.* **2025**, *18*, 7.
11. Muduli, R.; Jena, D.; Moger, T. Automatic generation control of is-landed micro-grid using integral reinforcement learning-based adaptive optimal control strategy. *Electr. Eng.* **2025**, *107*, 3135–3150. [CrossRef]
12. Chen, H.; Yang, L.B.; Gao, J.; Ji, B.-X. Hierarchical game optimization of integrated energy systems with grid-forming energy storage: Improved mountaineering team optimization algorithm. *J. Ind. Prod. Eng.* **2025**, *42*, 235–254. [CrossRef]
13. Liu, Z.; Liu, Y.; He, J.; Huang, X.; Ding, Z. Double DQN-based Power System Transient Stability Emergency Control with Protection Coordinations. In Proceedings of the 2023 IEEE 6th International Electrical and Energy Conference (CIEEC), Hefei, China, 12–14 May 2023; pp. 4182–4187.
14. Ding, Z.; Huang, X.; Liu, Z.; Zhang, Z. A Two-Level Scheduling Algorithm for Battery Systems and Load Tap Changers Coordination in Distribution Networks. *IEEE Trans. Power Deliv.* **2022**, *37*, 3027–3037. [CrossRef]
15. Zhao, J.; Zhang, D.; Zhou, Y.; Chen, S.; Tang, J.; Yao, R. Interpretable object detection method for remote sensing image based on deep reinforcement learning. *Pattern Recognit. Artif. Intell.* **2021**, *34*, 777–786.
16. Chen, L.; Xu, C.; Guo, J.; Zhou, X. Adaptive Optimization Model and Calculation Method of Target Frame of Intelligent Distribution Network. In Proceedings of the 2025 5th International Conference on Mechanical, Electronics and Electrical and Automation Control (METMS), Chongqing, China, 9–11 May 2025; pp. 707–710.

17. Hamma, A.H.; Manjang, S.; Kitta, I. Implementation of Ant Colony Optimization and Particle Swarm Optimization for Power Plant Operation Optimization. In Proceedings of the 2025 4th International Conference on Electronics Representation and Algorithm (ICERA), Yogyakarta, Indonesia, 12–12 June 2025; pp. 12–17.
18. Chen, J.; Sun, B.; Zeng, Y.; Jing, R.; Dong, S.; Wang, J. An Optimal Scheduling Method of Shared Energy Storage System Considering Distribution Network Operation Risk. *Energies* **2023**, *16*, 2411. [CrossRef]
19. Çetin, E.; Barrado, C.; Salamí, E.; Pastor, E. Analyzing deep reinforcement learning model decisions with Shapley additive explanations for counter drone operations. *Appl. Intell.* **2024**, *54*, 12095–12111. [CrossRef]
20. Zhang, M.; Miao, Z.; Nan, X.; Ma, N.; Liu, R. Explainable Reinforcement Learning for the Initial Design Optimization of Compressors Inspired by the Black-Winged Kite. *Biomimetics* **2025**, *10*, 497. [CrossRef]
21. Zhang, K.; Zhang, J.; Xu, P.-D.; Gao, T.; Gao, D.W. Explainable AI in Deep Reinforcement Learning Models for Power System Emergency Control. *IEEE Trans. Comput. Soc. Syst.* **2022**, *9*, 419–427. [CrossRef]
22. Huang, H.H.; Huang, Y.H. A novel green learning artificial intelligence model for regional electrical load prediction. *Expert Syst. Appl.* **2024**, *256*, 124907. [CrossRef]
23. Wang, H.; Dang, R. Interpretability analysis and model update research of a transient stability assessment model based on Transformer. *Power Syst. Prot. Control* **2023**, *51*, 15–25.
24. Ghosh, S.; Isbeih, Y.J.; El Moursi, M.S. Assessment of Bus Inertia to Enhance Dynamic Flexibility of Hybrid Power Systems with Renewable Energy Integration. *IEEE Trans. Power Deliv.* **2023**, *38*, 2372–2386. [CrossRef]
25. Eyal, W. *Hands-On Genetic Algorithms with Python: Apply Genetic Algorithms to Solve Real-World AI and Machine Learning Problems*; Packt Publishing: Birmingham, UK, 2024.

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