ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research on dynamic scheduling and service optimization of public service of national fitness based on reinforcement learning

Xiaoying Wang[1], *Xiaoliang Miao*[1], *Jingli Qu*[2,*] *and Shouyi Wang*[2]

[1] Department of Sports and Health, Handan Polytechnic College Handan 056001, Hebei, China

[2] Hebei University of Engineering, Handan 056001, Hebei, China

SUMMARY: *Under the background of digital transformation of national fitness public service, traditional manual scheduling and regular threshold scheduling are difficult to adapt to the fluctuation of fitness demand period, heterogeneous distribution of resources and changes in user feedback. For public sports venues, intelligent fitness equipment areas and community service areas, this paper constructs a dynamic scheduling and service optimization model based on PPO reinforcement learning. Multi-source data perception, time series demand prediction, resource load estimation and multi-objective reward function are used to realize the collaborative optimization of resource allocation, time period adjustment and user guidance. The model converts the reservation record, passenger flow monitoring, equipment status, external environment and user evaluation into computable state input, and improves the iterative stability of the strategy through feedback reward correction. The experimental results show that the average relative error of the model for the demand prediction of four types of fitness projects is 5.7%, and the accuracy of resource load level recognition reaches 92.7%. Compared with manual experience scheduling, PPO scheduling improves resource utilization to 89.3%, reduces average response time to 29.4 s, and user satisfaction reaches 91.8%. The research can provide technical support for the intelligent scheduling and fine governance of national fitness public services.*

KEYWORDS: *Reinforcement learning; Public service of national fitness; Dynamic scheduling; Service optimization*

1 Introduction

The public service of national fitness is an important support for improving the health level of residents, optimizing the allocation of urban public resources and promoting the balanced development of sports services [1]. In recent years, urban community sports venues, fitness paths, public gyms, smart gyms and online booking platforms have gradually increased, and the time, project and space selection of residents to participate in fitness activities show obvious dynamic changes [2]. Morning and evening peaks, holidays, events, weather changes and venue equipment status will all cause fluctuations in fitness demand, resulting in problems such as uneven utilization, response lag and service congestion of public fitness resources in different regions and different periods [3-5]. Traditional manual scheduling and rule-based scheduling methods mainly rely on fixed experience and static thresholds, which are difficult to perceive demand changes in time and form collaborative optimization among

*13503106552@163.com
https://doi.org/10.65102/is20261291 Published: 10 June, 2026

resource utilization, service fairness, user waiting time and satisfaction [6].

Under the background of digital public service construction, the national fitness service system has gradually accumulated multi-source data such as reservation records, admission passenger flow, equipment use time, user evaluation, venue opening status, traffic distance and weather conditions, which provides a data basis for intelligent scheduling [7]. Through the time series modeling and state representation of these data, the change law of residents' fitness demand can be more accurately described, and the calculation basis can be provided for venue allocation, service time adjustment, equipment maintenance reminder and pedestrian flow guidance. Reinforcement learning has the advantages of continuous trial and error oriented to dynamic environment, optimization strategy based on reward feedback and adaptation to complex constraint scenarios, which is suitable for dealing with problems with strong randomness of demand, multiple resource types and complex scheduling objectives in public services of national fitness [8-10]. In particular, PPO algorithm performs well in policy update stability, continuous decision optimization and complex state input processing, and can provide an effective technical path for the dynamic scheduling of public fitness resources [11].

Focusing on the problem of dynamic scheduling and service optimization of public services for national fitness, this paper constructs an intelligent scheduling model based on reinforcement learning. The model takes multi-source national fitness service data as input, extracts key features such as venue load, user demand, facility status, service period and feedback evaluation, and forms the operation state representation for scheduling decision. On this basis, the time series prediction method is introduced to identify the demand changes of different fitness projects and different service areas, and the scheduling action set is established by combining resource capacity, opening time, service radius and user waiting constraints. In order to make the model take into account both operational efficiency and public service attributes, this paper designs a multi-objective reward function integrating resource utilization, response time delay, service fairness and user satisfaction, and uses PPO strategy network to iteratively optimize the public fitness resource allocation scheme. The model continuously modifies the scheduling policy through the service feedback data, so that the system can maintain a relatively stable service response ability under the disturbance conditions such as peak demand, equipment anomaly and user cancellation.

The research work of this paper is mainly reflected in three aspects. Firstly, the spatiotemporal coupling of supply and demand and the collaborative modeling framework of heterogeneous resources are established for the national fitness public service scenario, and the public venues, fitness facilities, user behavior and service feedback are incorporated into the unified computing structure. Secondly, a dynamic scheduling model based on PPO reinforcement learning is constructed, and the multi-objective reward function is used to guide the policy network to comprehensively optimize the efficiency, fairness and satisfaction. Thirdly, the performance of the model in demand prediction, resource load identification, scheduling efficiency, disturbance stability and strategy convergence is verified by experimental simulation, which provides a computable and verifiable technical scheme for the intelligent governance of public services of national fitness.

2 Background and multidimensional problem modeling of dynamic scheduling of public service of national fitness

2.1 Spatiotemporal coupling constraint modeling of public fitness resources supply and demand

The core of public service scheduling for national fitness is to deal with the coupling relationship of "demand changes with time, resources are distributed with space, and service capacity is affected by the state of facilities". The demand of community residents for public fitness services is not evenly distributed, and the visits are often concentrated in the evening of workdays, the morning of weekends, and around holidays. However, there are differences in the capacity, opening hours, maintenance status and reachable distance of gyms, fitness paths, ball fields and smart fitness facilities in different areas. Therefore, the scheduling model needs to simultaneously depict user demand intensity, resource available capacity, regional service distance and waiting pressure, so that the reinforcement learning model can obtain a more complete state of the environment.

Suppose that the urban national fitness service area is divided into P areas, and the public fitness resources are divided into K types. In the time slice τ, the demand of area p for resource type k is $d_{pk\tau}$, the effective service capacity of this kind of resource is $a_{pk\tau}$, and the average distance of users to the resource point is l_{pk}. The queuing waiting pressure is $h_{pk\tau}$, then the spatiotemporal coupling pressure of supply and demand can be expressed as follows:

$$\Omega_\tau = \sum_{p=1}^{P} \sum_{k=1}^{K} \left(\alpha_1 \frac{\left| d_{pk\tau} - a_{pk\tau} \right|}{a_{pk\tau} + \varepsilon} + \alpha_2 \frac{l_{pk}}{L_{max}} + \alpha_3 \frac{h_{pk\tau}}{H_{max}} \right) \tag{1}$$

where, Ω_τ represents the coupling pressure of supply and demand of public fitness resources in time slice τ. α_1, α_2 and α_3 represent the weights of supply and demand deviation, space distance and waiting pressure, respectively. ε is a smoothing term that prevents the denominator from being zero. L_{max} denotes the maximum service distance; H_{max} denotes the maximum allowed waiting pressure. The formula can map the factors such as insufficient resources, too far away and queuing congestion into computable indicators, and provide state input for the subsequent PPO scheduling strategy.

In practical scheduling, resource allocation should also satisfy the constraints of capacity, open time and service fairness. Let $xp_{k\tau}$ denote the amount of resource service of the KTH type allocated to area p in time slice τ, $c_{pk\tau}$ denote the corresponding resource capacity, $\bar{w}_{p\tau}$ denote the average waiting time of the area, and $\eta_{p\tau}$ denote the service coverage level of the area. Then the constraint relation is as follows:

$$0 \le x_{pk\tau} \le c_{pk\tau}, \quad \bar{w}_{p\tau} \le W, \quad \eta_{p\tau} \ge \eta \tag{2}$$

where, W_{max} denotes the maximum acceptable waiting time and η_{min} denotes the minimum service coverage requirement. Through the above constraints, the model can not only avoid the overload of a single popular venue, but also prevent the edge area from being in a state of long-term low service level, so that the dynamic scheduling process takes into account both efficiency and public service fairness.

2.2 Analysis of collaborative characteristics of heterogeneous resources in national fitness service scheduling

The resource types faced by the public service scheduling of national fitness are obviously heterogeneous, and different resources are different in service ability, open cycle, spatial location, applicable population and operation state. Gymnasiums, ball fields, fitness trails, community fitness equipment and intelligent fitness stations belong to fixed space resources, and their service capacity is limited by the area of the venue, the number of equipment and the opening hours. Social sports instructors, volunteer service personnel and maintenance personnel are mobile resources, and their scheduling effect is affected by service area, professional ability and response time. Online reservation platform, passenger flow monitoring terminal and intelligent access control system belong to digital support resources, which can provide real-time status input for reinforcement learning model.

Suppose there are N types of heterogeneous service resources and V fitness service tasks in the time slice τ. $z_{nv\tau}$ represents whether resource n is assigned to task v, $\phi_{nv\tau}$ represents the matching degree between resource capacity and service tasks, $\rho_{n\tau}$ represents the real-time availability status of resources, and $\kappa_{nv\tau}$ represents the comprehensive adaptation value of resources for user reachability and service priority. $\delta_{nm\tau}$ represents the communication, waiting or transfer cost generated by resource n and resource m when they cooperate, then the heterogeneous resource cooperative utility can be expressed as follows:

$$\Psi_\tau = \sum_{v=1}^{V} \sum_{n=1}^{N} z_{nv\tau}(\beta_1 \phi_{nv\tau} + \beta_2 \rho_{n\tau} + \beta_3 \kappa_{nv\tau}) - \beta_4 \sum_{v=1}^{V} \sum_{n=1}^{N} \sum_{m \neq n}^{N} z_{nv\tau} z_{mv\tau} \delta_{nm\tau} \quad (3)$$

Where, $\Psi\tau$ represents the co-scheduling utility of heterogeneous resources within a time slice τ. $\beta1$, $\beta2$, $\beta3$, $\beta4$ represent the weights of capability matching, resource availability, service adaptation, and collaboration cost, respectively. The model can reflect the gain and loss of different types of resources when they jointly complete the fitness service task, and provide a quantitative basis for the subsequent reward design of reinforcement learning.

To avoid resource overload and task omission, heterogeneous resource collaboration also needs to satisfy basic allocation constraints:

$$\sum_{n=1}^{N} z_{nv\tau} \geq 1, \quad \sum_{v=1}^{V} z_{nv\tau} g_{v\tau} \leq b_{n\tau} \quad (4)$$

where, $g_{v\tau}$ represents the service load generated by task v in time slice τ, and $b_{n\tau}$ represents the loadable service capacity of resource n. Through this constraint, the model can ensure that each type of fitness service demand at least obtains resource response, while avoiding excessive occupation of popular venues, key equipment or service personnel. It can be seen that the collaborative characteristics of heterogeneous resources not only determine the boundary of scheduling actions, but also directly affect the optimization direction of the reinforcement learning model among efficiency, fairness and stability.

3 Reinforcement learning based public service dynamic scheduling and service optimization model construction for national fitness

3.1 Multi-source national fitness service data perception and running state representation

In order to realize effective decision-making of public service dynamic scheduling of national fitness, the premise is to be able to characterize the running state of the service system continuously, accurately and computably. Different from general single-scene scheduling, national fitness service is affected by multiple factors such as reservation behavior, museum passenger flow, equipment operation, environmental changes and user feedback at the same time. The data sources are scattered, the sampling frequency is different, and the dimension difference is obvious. Therefore, this paper constructs a multi-source data perception and state representation method at the input of the model, and maps the reservation platform logs, gate entry and exit records, instrument sensing data, museum opening information, weather and holiday information, and user evaluation information into the scheduling environment uniformly, which provides structured state input for the subsequent PPO policy network.

Let the total number of service units be U and the number of data source channels be C. At time t, the original observation collected by the u-th service unit from the c-th data source is denoted as oc,t(u). In order to eliminate the dimension differences of different indicators and consider the different credibility of each data source, the normalized processing method with confidence weight is adopted to obtain the standardized observation quantity:

$$\hat{o}_{c,t}^{(u)} = \chi_c \frac{o_{c,t}^{(u)} - \underline{o}_c}{\overline{o}_c - \underline{o}_c + \iota_0} \tag{5}$$

where, $\hat{o}_{c,t}^{(u)}$ is the normalized observation value, χ_c represents the credibility weight of the CTH type of data source, and ι_0 is the smoothing constant. After this process, heterogeneous variables such as the number of appointments, passenger density, equipment use time and evaluation scores can be mapped into a unified numerical space, which is convenient for subsequent fusion calculation.

Considering the obvious time continuity of public fitness service scheduling, the current state of the system is not only affected by the current observation, but also related to the operation trajectory of the previous several periods. To this end, this paper uses the combination of current observation and sliding window historical observation to construct service unit operation description:

$$r_t^{(u)} = \sum_{c=1}^{C} \omega_c \, \hat{o}_{c,t}^{(u)} + \sum_{q=1}^{Q} \lambda_q \, \bar{o}_{t-q}^{(u)} \tag{6}$$

where, $r_t^{(u)}$ represents the running status score of the u-th service unit at time t, ω_c is the fusion weight of different data channels at the current time, Q is the length of the history window, λq is the time attenuation coefficient of the QTH historical delay term. $\bar{o}_{t-q}^{(u)}$ denotes the average level of various standardized observations of the u -th service unit at time t−q. This representation can simultaneously retain the immediate load characteristics and short-term trend characteristics in the state-aware stage, thereby enhancing the recognition

ability of the model for peak congestion, local idle and abnormal fluctuations.

After the service unit level state extraction is completed, it is further integrated with the appointment demand vector, the device health vector and the environment context vector to form a global state representation for reinforcement learning decision making:

$$s_t = \tanh(W_r r_t + W_b b_t + W_m m_t + W_f f_t + b_s) \quad (7)$$

where, s_t is the global environment state vector at time t, r_t represents the vector composed of the operation state of each service unit, b_t represents the feature vector of reservation demand, m_t represents the health state vector of equipment and venue facilities, f_t represents the context feature vector of weather, holidays and time periods, W_r, W_b, W_m and W_f are the corresponding mapping matrix, and b_s is the bias term. Through nonlinear fusion, the model can describe the coupling relationship of "demand change, resource load, facility condition and external environment" in a unified state space.Figure 1 shows the framework of reinforcement learning dynamic scheduling model for public service of national fitness.

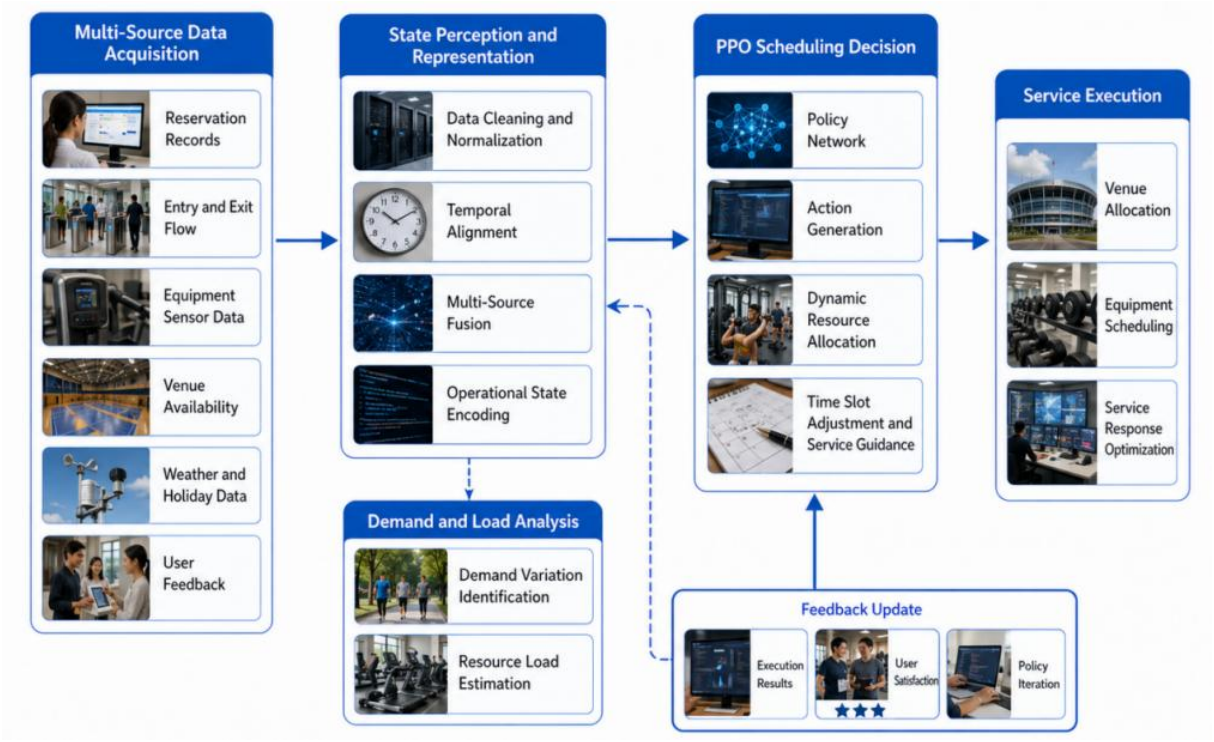

Figure 1: Framework diagram of reinforcement learning dynamic scheduling model for public service of national fitness

As shown in Figure 1, the reinforcement learning dynamic scheduling model framework for public service of national fitness constructed in this paper consists of data collection layer, state perception layer, demand identification layer, PPO scheduling decision layer and feedback update layer. The data acquisition layer is responsible for accessing multi-source information such as reservation, passenger flow, equipment, environment and evaluation. The state-aware layer completes normalization, fusion and running state coding. The requirement identification layer further forms the key representations required by scheduling. PPO decision layer output site allocation, service time adjustment and resource guidance strategy. The feedback update layer modifies the subsequent policies according to the execution results. This framework provides a stable data basis and state representation basis for demand change

recognition, reward function construction and dynamic scheduling optimization.

3.2 Fitness demand change identification and resource load estimation based on time series prediction

The public service demand of national fitness has obvious time fluctuations and regional differences, and it is difficult to find the potential congestion risk in time by scheduling only based on the current reservation volume or instant passenger flow. In order to improve the perception ability of reinforcement learning model for future service pressure, this paper introduces a time series prediction module before scheduling decision, which identifies the demand changes of different service areas, venue types and fitness programs in advance, and estimates the resource load level synchronously. In order to avoid mixing with the previous formula notation, this section adopts a new notation system. Let σ denote the fitness service combination unit,ℓ denote the prediction period, $\zeta_{\sigma,\ell}$ denote the historical service demand intensity, and $\mathrm{y}_{\sigma,\ell}$ denote the external influence characteristics such as weather, holidays, sports events and regional travel heat. In this paper, the historical demand and external features in a continuous time window are jointly encoded to obtain the time series prediction input vector:

$$\mathbf{v}_{\sigma,\ell} = \mathrm{Concat}\left(\zeta_{\sigma,\ell-\Delta+1}, \zeta_{\sigma,\ell-\Delta+2}, \cdots, \zeta_{\sigma,\ell}, \mathbf{y}_{\sigma,\ell}\right) \tag{8}$$

where, $\mathrm{v}_{\sigma,\ell}$ denotes the timing input vector of the σ th service composition unit at time period ℓ, Δ denotes the length of the historical observation window. The input structure can simultaneously retain short-term passenger flow variation, reservation accumulation trend and external disturbance factors, so that the model does not rely on single point observation, but identifies the direction of demand change from continuous operation trajectory.

In the stage of demand forecasting, this paper uses a time-series neural network to map the input vector non-linearly and output the estimated value of demand intensity in the next period. The calculation process is expressed as follows:

$$\hat{\zeta}_{\sigma,\ell+1} = \mathrm{Softplus}\left(\mathrm{A}_1 \mathrm{v}_{\sigma,\ell} + \mathrm{a}_1\right) \tag{9}$$

where,$\hat{\zeta}_{\sigma,\ell+1}$ represents the predicted fitness demand intensity in the next period, A_1 is the demand prediction mapping matrix, a_1 is the bias vector, and Softplus(·) is used to ensure that the prediction result is non-negative. The prediction results can be used to determine whether resources such as basketball courts, badminton courts, fitness equipment areas, and running trails will be intensively occupied in advance, and provide prior information for reinforcement learning scheduling action generation.

The demand change recognition not only focuses on the size of the forecast value, but also needs to determine whether the demand increases rapidly or decreases abnormally. To this end, this paper constructs a demand fluctuation identification index:

$$\Gamma_{\sigma,\ell+1} = \frac{\left|\hat{\zeta}_{\sigma,\ell+1} - \zeta_{\sigma,\ell}\right|}{\zeta_{\sigma,\ell} + \varpi_0} \tag{10}$$

where, $\Gamma_{\sigma,\ell+1}$ represents the magnitude of demand change in the next period, and ϖ_0 is a smoothing constant. When the value is high, it indicates that the current service unit may enter the state of sudden increase in demand, centralized reservation or passenger flow decline, and the system needs to adjust the opening time in advance, guide the diversion of users or

increase the support of service personnel. This metric can convert demand changes into computable signals and avoid the scheduling model to respond passively after the demand is already crowded.

In resource load estimation, only demand prediction is not enough. It is also necessary to combine service capacity, facility availability status and average service time to determine the degree of resource pressure. Let θ_{σ} denote the average service occupancy time of the σ th service composition unit, $\mu_{\sigma,\ell+1}$ denote the unit service capacity that can be provided in the next period, and $\upsilon_{\sigma,\ell+1}$ denote the facility health and open availability coefficients, then the resource load estimate can be expressed as follows:

$$\Lambda_{\sigma,\ell+1} = \frac{\hat{\zeta}_{\sigma,\ell+1}\theta_{\sigma}}{\mu_{\sigma,\ell+1}\nu_{\sigma,\ell+1} + \varpi_0} \tag{11}$$

where, $\Lambda_{\sigma,\ell+1}$ represents the resource load level in the next time period. The higher the value, the closer the predicted demand is to saturation with respect to the available service capacity, and the more the model needs to take scheduling actions such as time-sharing reservation, cross-venue guidance, instrument rotation or service staff recruitment. Through the above time series prediction and load estimation, the system can input the "future demand intensity, change amplitude and resource pressure degree" into the reinforcement learning decision-making process, so that the PPO policy network has stronger forward-looking and stability when generating scheduling schemes.

3.3 Reinforcement learning decision objective modeling for service efficiency fairness and satisfaction collaboration

Different from the single efficiency-oriented resource allocation task, the public service scheduling of national fitness needs to simultaneously cover the service response speed, resource utilization balance, regional fairness and user experience feedback. If the reinforcement learning model only pursues the museum utilization, it may lead to the hot area being continuously full loaded and the edge area being idle. If the waiting time is only reduced, it may cause excessive diversion and frequent switching of resources, and increase the cost of service management. Therefore, this paper transformed the public service scheduling goal of national fitness into a multi-objective reinforcement learning reward modeling problem, so that the PPO policy network could comprehensively judge "whether the resources are fully used, whether users can obtain services in time, whether different regions maintain basic fairness, and whether the feedback results continue to improve" in the scheduling process.

Table 1 presents the reinforcement learning scheduling decision objective and reward function composition. The reward design focuses on four elements: service efficiency, fairness constraint, satisfaction feedback and scheduling penalty, so that the model can not only learn efficient scheduling strategies, but also avoid excessive concentration of public service resources in local high-demand areas.

Table 1: Reinforcement learning scheduling decision objective and reward function constitute table

Reward Objective	Calculation Basis	Optimization Direction	Scheduling Function
Service Efficiency Reward	Service completion ratio, average waiting time, resource utilization level	Improve service response speed and facility utilization efficiency	Guide the model to preferentially relieve congestion under high-demand scenarios
Service Fairness Reward	Service coverage level of each area, differences in service opportunities	Narrow the service accessibility gap among different areas	Prevent resources from being concentrated in popular venues or core areas for a long time
User Satisfaction Reward	User rating score, reuse intention, complaint ratio	Improve user experience and service acceptance	Enable scheduling results to form a closed loop with real resident feedback
Resource Overload Penalty	Venue full-load status, excessive equipment use time, staff service pressure	Reduce the risk of local resource overuse	Avoid single-point congestion and facility operation abnormalities
Scheduling Switching Penalty	Cross-region guidance frequency, service time adjustment range	Reduce unnecessary frequent adjustments	Maintain the stability and executability of scheduling strategies

In the service efficiency reward modeling, let the decision round be ν, the service completion level be $\mathcal{F}_{\nu}^{done}$, the standardized waiting pressure be $\mathcal{Q}_{\nu}^{\mathrm{wait}}$, and the resource effective use level be $\mathcal{G}_{\nu}^{\mathrm{use}}$, then the efficiency reward is expressed as follows:

$$\mathcal{R}_{\nu}^{eff} = \gamma_1 \mathcal{F}_{\nu}^{done} + \gamma_2 \left(1 - \mathcal{Q}_{\nu}^{wait}\right) + \gamma_3 \mathcal{G}_{\nu}^{use} \tag{12}$$

where, $\mathcal{R}_{\nu}^{eff}$ represents the service efficiency reward, and γ_1, γ_2, and γ_3 are the corresponding weight coefficients. The reward item can promote the model to improve the service completion ratio while reducing the waiting pressure of users, and improve the effective use of public fitness venues, equipment and personnel resources.

In the fairness reward modeling, suppose that there are B service grids, the service coverage level of the BTH service grid under decision round ν is $\mathcal{L}_{b,\nu}^{cov}$, and the average coverage level of all service grids is $\overline{\mathcal{L}}_{\nu}^{cov}$, then the fair reward is expressed as follows:

$$\mathcal{R}_{\nu}^{fair} = 1 - \sqrt{\frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} \left(\mathcal{L}_{b,\nu}^{cov} - \overline{\mathcal{L}}_{\nu}^{cov}\right)^2} \tag{13}$$

where, $\mathcal{R}_{\nu}^{fair}$ denotes the service fair reward. The smaller the difference in coverage level is, the higher the fair reward is, and the more inclined the PPO policy network is to generate a scheduling scheme that considers both the core area and the edge area, so as to reduce the regional deviation in the allocation of public fitness resources.

The user satisfaction reward is used to reflect the actual service feeling of the scheduling result. Assuming that the user evaluation score is Ivscore, the reuse intention is Ivreturn, and the complaint intensity is Ivcomplaint, the satisfaction reward is as follows.

$$\mathcal{R}_v^{sat} = \gamma_4 I_v^{score} + \gamma_5 I_v^{return} + \gamma_6\left(1 - I_v^{complaint}\right) \tag{14}$$

where, $\mathcal{R}_v^{sat}$ represents the user satisfaction reward, and γ_4, γ_5, and γ_6 are the satisfaction related weights. This term can incorporate user evaluation, reappointment tendency and complaint feedback into the reinforcement learning training process, so that the model not only focuses on the numerical efficiency of scheduling results, but also reflects the actual acceptance of residents to public fitness services.

Considering efficiency, fairness and satisfaction objectives, and adding resource overload and scheduling switch penalty, the total reward function is obtained as follows:

$$\mathcal{R}_v = \gamma_7 \mathcal{R}_v^{eff} + \gamma_8 \mathcal{R}_v^{fair} + \gamma_9 \mathcal{R}_v^{sat} - \gamma_{10} \mathcal{C}_v^{over} - \gamma_{11} \mathcal{C}_v^{shift} \tag{15}$$

where, R_v represents the integrated reward under decision round v, $\mathcal{C}_v^{\text{over}}$ represents the resource overload penalty, $\mathcal{C}_v^{shift}$ represents the frequent schedule switching penalty, and γ_7 to γ_{11} are the integrated reward weights. Through the decision objective modeling, the PPO policy network can form scheduling preferences for public service scenarios during the training process, so that the dynamic scheduling scheme can maintain a good synergy among resource efficiency, regional fairness and user experience.

3.4 Dynamic scheduling method of public fitness resources based on PPO policy network

After the multi-source state representation, demand change identification and decision goal modeling are completed, the key of public fitness resource scheduling is to map the environment state into an executable scheduling policy. Considering the characteristics of rapid demand fluctuation, many constraints, scheduling feedback delay and high stability requirements of strategy update in the public service scenario of national fitness, this paper uses the PPO strategy optimization method to construct a dynamic scheduling model of public fitness resources. By limiting the update range between the old and new policies, PPO can improve the efficiency of policy learning while ensuring the stability of training, and is suitable for dealing with continuous iterative decision-making tasks such as venue allocation, equipment scheduling, time period adjustment and service guidance.

Let the environment embedding state under the ω scheduling round be n_ω, and the set of candidate scheduling actions be P, where each action corresponds to a resource allocation and service organization scheme. The policy network outputs the selection probability of each candidate action according to the current environment state, which is expressed as follows:

$$\pi_\Theta(\psi_\omega \mid n_\omega) = \frac{\exp(g_\Theta(n_\omega, \psi_\omega))}{\sum_{\psi' \in \mathcal{P}} \exp\left(g_\Theta(n_\omega, \psi')\right)} \tag{16}$$

where, $\pi_\Theta(\cdot)$ denotes the policy network with parameter Θ, ψ_ω denotes the scheduling action selected at round ω, and $g_\Theta(\cdot)$ denotes the policy scoring function. Through this probability distribution, the model can complete the adaptive selection among the actions such as "diversion of popular venues, rotation of equipment, fine adjustment of service hours, and cross-region guidance", so as to improve the adaptability of scheduling decisions to complex scenes.

In order to prevent the training shock caused by too large policy update, this paper uses the PPO pruning objective function to optimize the policy network. Let Θ^- be the parameters of the old policy before updating, and ϱ_ω be the probability ratio of the old and new policy.

Then the PPO optimization objective can be written as follows:

$$\mathcal{J}_{\mathrm{PPO}}(\Theta) = \mathbb{E}_{\omega}[\min(\varrho_{\omega}\mathcal{A}_{\omega}, \mathrm{clip}(\varrho_{\omega}, 1-\epsilon_{c}, 1+\epsilon_{c})\mathcal{A}_{\omega})] \tag{17}$$

Type, J_{PPO} (Θ) PPO strategy objective function, $\varrho_{\omega} = \pi_{\Theta}(\psi_{\omega} \mid n_{\omega})/\pi_{\Theta^{-}}(\psi_{\omega} \mid n_{\omega})$ for the old and new strategy estimate probability than $\mathcal{A}_{\omega}$ said advantage, ϵ_{c} for cutting threshold. The objective function can control the policy change within a reasonable range, so that the model can maintain good update stability in the face of peak bookings, temporary equipment failures or sudden changes in museum load.

Advantage estimation is used to measure the degree of the current action relative to the benchmark value. In this paper, the advantage term is constructed by combining multi-step returns with value estimation:

$$\mathcal{A}_{\omega} = \left(\sum_{s=0}^{S-1} \vartheta^{s} \mathcal{U}_{\omega+s}\right) + \vartheta^{S} V_{\Phi}(\mathbf{n}_{\omega+S}) - V_{\Phi}(\mathbf{n}_{\omega}) \tag{18}$$

where $U_{\omega+s}$ represents the immediate reward obtained from round ω+s, ϑ is the discount factor, S is the payoff step, and $V_{\Phi}(\cdot)$ represents the value network with parameter Φ. Through this estimation method, the model not only considers the direct impact of the current scheduling action on resource efficiency and waiting time, but also pays attention to its long-term effect on the subsequent fairness and satisfaction, so that the policy learning is more in line with the goal of continuous optimization of public services. Figure 2 shows the dynamic scheduling process of public fitness resources based on PPO policy network.

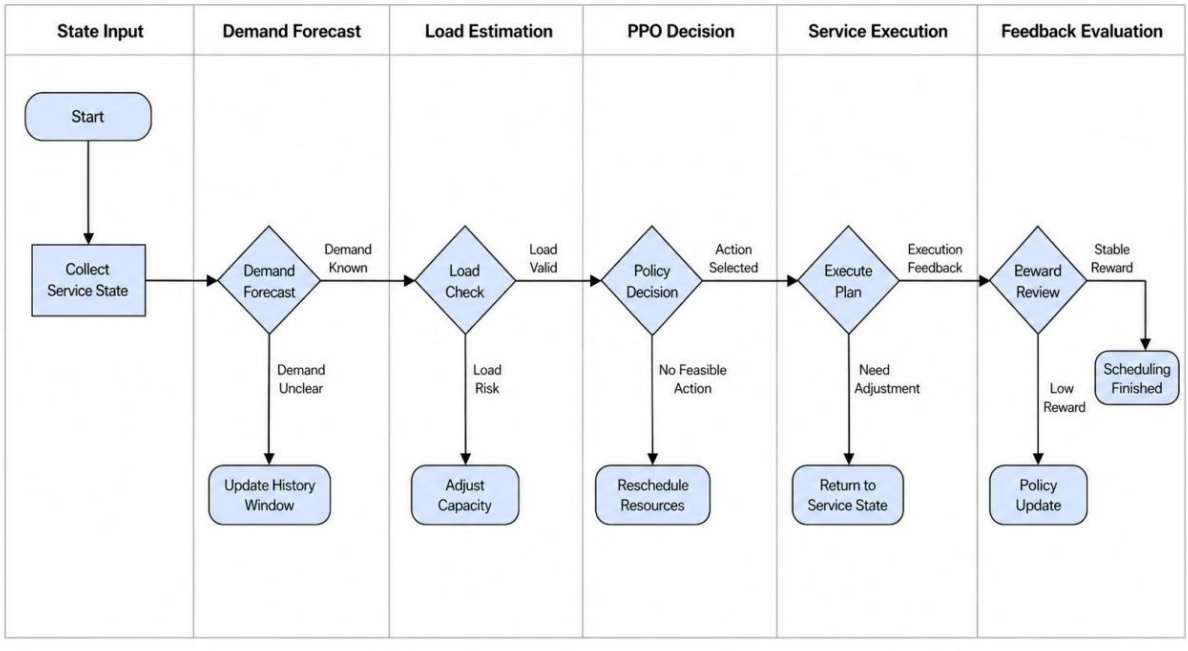

Figure 2: PFlow chart of public fitness resource dynamic scheduling based on PPO policy network

As shown in Figure 2, the dynamic scheduling process of public fitness resources based on PPO policy network includes five links: state input, policy generation, action execution, environment feedback and parameter update. Firstly, the system receives the operation state vector, demand prediction results and resource load estimation results constructed by the

above, and inputs them into the policy network and value network. Then, the policy network outputs the probability of candidate scheduling actions, and the system selects actions according to probability sampling or maximum probability to form specific venue allocation, service guidance and time period adjustment schemes. After the action is executed, the environment returns information such as service completion, waiting change, area coverage and user feedback, and generates an immediate reward according to the reward function in Section III. The policy network uses the sampling trajectory for batch training, and the value network synchronously updates the state value estimate until the model gradually converges in multiple rounds of interaction.

The advantage of this method is that it continuously learns the strategy pattern of "when to distribute, how to distribute, and how to balance fairness and satisfaction" through continuous interaction. For the public service of national fitness, this dynamic environment-oriented policy optimization method is more suitable for dealing with complex scheduling problems with multiple scenarios, multiple constraints and multiple objectives, and also lays a method foundation for subsequent service feedback-driven policy iterative optimization.

3.5 Iterative optimization of reinforcement learning scheduling policy with user feedback

The public fitness resource scheduling strategy will be continuously affected by user experience feedback in actual operation. Scheduling only relying on reservation volume, passenger flow density and facility load can reflect the operating pressure of the service system, but it is difficult to fully reflect the subjective feelings of residents on venue allocation, waiting time, equipment availability, service guidance and sports experience. Therefore, this paper introduces the user feedback closed loop after the execution of PPO scheduling strategy, and converts the evaluation score, reuse intention, complaint record, crowding perception and waiting feeling into policy iteration signals, so that the reinforcement learning model can continuously revise the scheduling tendency according to the real service results.

Suppose that in the feedback iteration ξ, the system collects feedback indicators such as user evaluation, waiting feeling, reuse intention, complaint intensity and crowding perception, and normalize them into a unified feedback feature vector:

$$\mathbf{F}_{\xi}^{fb} = \mathrm{Norm}\left(\overline{E}_{\xi}^{score}, \overline{E}_{\xi}^{wait}, \overline{E}_{\xi}^{reuse}, \overline{E}_{\xi}^{complaint}, \overline{E}_{\xi}^{crowd}\right) \tag{19}$$

where, $\mathrm{F}_{\xi}^{\mathrm{fb}}$ represents the user feedback feature vector in round ξ, $\overline{\mathrm{E}}_{\xi}^{\mathrm{score}}$ represents the average evaluation score, $\overline{E}_{\xi}^{wait}$ represents the user's perceived waiting intensity, $\overline{\mathrm{E}}_{\xi}^{\mathrm{reuse}}$ represents the willingness to use again,$\overline{\mathrm{E}}_{\xi}^{\mathrm{complaint}}$ represents the complaint intensity, and $\overline{\mathrm{E}}_{\xi}^{\mathrm{crowd}}$ represents the user's perceived crowding degree. Through this processing, scattered user feedback can be transformed into numerical signals that can be received by reinforcement learning models.

Considering that user feedback may have problems such as insufficient samples, extreme evaluation and repeated submission, this paper further constructs feedback credibility coefficient to control the influence strength of feedback signal on policy update:

$$\mathcal{T}_{\xi}^{fb} = \frac{\mathcal{M}_{\xi}^{valid}}{\mathcal{M}_{\xi}^{total} + \epsilon_f}\left(1 - \mathcal{N}_{\xi}^{noise}\right) \tag{20}$$

where, $\mathcal{T}_{\xi}^{\mathrm{fb}}$ represents feedback credibility, $\mathcal{M}_{\xi}^{valid}$ represents the number of effective feed-back, $\mathcal{M}_{\xi}^{total}$ represents the total number of feed-back, $\mathcal{N}_{\xi}^{noise}$ represents the proportion of abnormal feedback, and ϵ_f is the smoothing term. When the proportion of effective feedback is high and the abnormal feedback is few, the feedback credibility is improved, and the model will fully absorb user experience information. When the feedback samples are weak or noisy, the system reduces its perturbation to the policy update.

In the reward correction link, feedback credibility and user experience indicators are introduced into the comprehensive reward, so that the scheduling strategy can be directionally calibrated according to the service results:

$$\tilde{\mathcal{R}}_{\xi} = \mathcal{R}_{\xi} + \mathcal{T}_{\xi}^{fb}\left(\varsigma_1 F_{\xi}^{score} + \varsigma_2 F_{\xi}^{reuse} - \varsigma_3 F_{\xi}^{wait} - \varsigma_4 F_{\xi}^{complaint}\right) \tag{21}$$

where, $\tilde{\mathcal{R}}_{\xi}$ represents the modified reward after integrating user feedback, $\mathcal{R}_{\xi}$ represents the original scheduling reward, $\mathrm{F}_{\xi}^{\mathrm{score}}$, $\mathrm{F}_{\xi}^{\mathrm{reuse}}$, $\mathrm{F}_{\xi}^{\mathrm{wait}}$ and $\mathrm{F}_{\xi}^{\mathrm{complaint}}$ represent the feedback components corresponding to evaluation score, reuse intention, waiting pressure and complaint intensity, respectively, and ς_1 to ς_4 are the feedback correction weights. In the policy update stage, the feedback correction reward does not directly replace the original training process of PPO, but acts as a supplementary signal for policy gradient update. To avoid drastic parameter changes caused by feedback fluctuations, a gradient clipping constraint is introduced:

$$\Theta_{\xi+1} = \Theta_{\xi} + \mathfrak{h}_{\xi} clip\left(\nabla_{\Theta}\tilde{\mathcal{R}}_{\xi}, -\mathcal{G}_{lim}, \mathcal{G}_{lim}\right) \tag{22}$$

where Θ_{ξ} represents the policy network parameters in round ξ, h_{ξ} represents the update step size under feedback guidance, $\nabla_{\Theta}\tilde{\mathcal{R}}_{\xi}$ represents the gradient of the correction reward to the policy parameters, and $\mathcal{G}_{\mathrm{lim}}$ represents the gradient clipping boundary. Through this update method, the model can maintain training stability while absorbing user feedback, and avoid frequent hopping of scheduling strategy caused by short-term evaluation fluctuations.

After the above iterative optimization, the public fitness service system can form a closed-loop process of "scheduling execution-user feedback-reward correction-strategy update". This process enables the reinforcement learning model to not only make scheduling choices according to resource states, but also continuously adjust service allocation preferences according to residents' real experience, thereby improving the acceptability and continuous optimization ability of dynamic scheduling schemes.

4 Experimental simulation and result analysis

4.1 Experimental environment Configuration and National Fitness Service data set Construction

In order to verify the effectiveness of the reinforcement learning based dynamic scheduling model for public service of national fitness, this paper constructs a simulation experiment environment for community sports venues, public fitness paths, intelligent fitness equipment areas and ball games venues. The experimental platform uses Python 3.10 and PyTorch 2.1 to build the reinforcement learning training framework, the demand prediction module is implemented by a time series neural network, and the PPO policy network is composed of a two-layer fully connected network and a value evaluation network. The hardware environment is configured with Intel Core i7 processor, 32 GB memory, NVIDIA RTX 3060

GPU, and Windows 11 operating system. According to the operation characteristics of urban community national fitness service, the simulation environment sets up service areas, venue capacity, opening hours, reservation rules, equipment status and user feedback channels, and simulates the dynamic service demand under working days, weekends, holidays and peak hours.

The data set mainly consists of reservation platform logs, museum visitor flow records, equipment sensing status, user feedback evaluation and external environment information. In the data preprocessing stage, repeated appointments, abnormal admission, missing equipment status and extreme evaluation are cleaned and aligned according to time slices. The time granularity is set to 30 min, and each sample includes fields such as service area, fitness project, reservation number, actual attendance number, resource capacity, average waiting time, equipment availability and satisfaction evaluation. The dataset was divided into training set, validation set and test set at 7:2:1, which were used for model training, parameter tuning and final performance evaluation. The composition of the national fitness service dataset is shown in Table 2.

Table 2: National fitness service data set composition table

Data Category	Collected Content	Sample Size	Preprocessing Method	Technical Function
Service Area Data	12 community service areas with different service radii and population densities	12 areas	Area encoding and service radius normalization	Represent the spatial distribution of public fitness resources
Venue Resource Data	Gymnasiums, ball-game courts, fitness trails, and smart fitness equipment areas	48 resource points	Capacity calibration and opening-hour encoding	Construct resource capacity and availability status
Reservation Record Data	User reservation time, project type, cancellation records, and attendance status	86,420 records	Deduplication, abnormal reservation removal, and time-slice alignment	Support fitness demand prediction
Passenger Flow Monitoring Data	Number of entries, number of exits, peak passenger flow, and stay duration	124,800 records	Missing value imputation and sliding-average smoothing	Identify real-time service load
Equipment Status Data	Equipment occupancy rate, fault status, maintenance records, and operating duration	32,560 records	Status discretization and outlier correction	Determine facility availability
User Feedback Data	Satisfaction scores, waiting perception, complaint records, and reuse intention	18,730 records	Text labeling and score normalization	Revise reinforcement learning reward signals
External Environment Data	Weather, holidays, sports events, and temperature changes	1,920 records	Category encoding and continuous feature standardization	Assist in explaining demand fluctuations
Training Sample Split	Training set, validation set, and test set	60,494 / 17,284 / 8,642 records	Chronological split	Ensure temporal consistency in model validation

Table 2 shows that the data set covers the information of demand side, resource side, environment side and feedback side, and can completely describe the operation state of the public service of national fitness. After unified coding and time slice alignment, the multi-source data can be directly input into the demand prediction module and PPO scheduling model, which provides an experimental basis for subsequent verification of resource utilization, service response efficiency and strategy stability.

4.2 Fitness demand prediction and resource load identification experiments

The fitness demand prediction experiment is mainly used to verify the ability of the time series prediction module to identify the demand changes of different fitness services. In this paper, four typical items such as basketball court, badminton court, intelligent fitness equipment area and running trail are selected as test objects, and the real reservation volume is compared with the model prediction results. Figure 3 shows the demand prediction results of different fitness service projects.

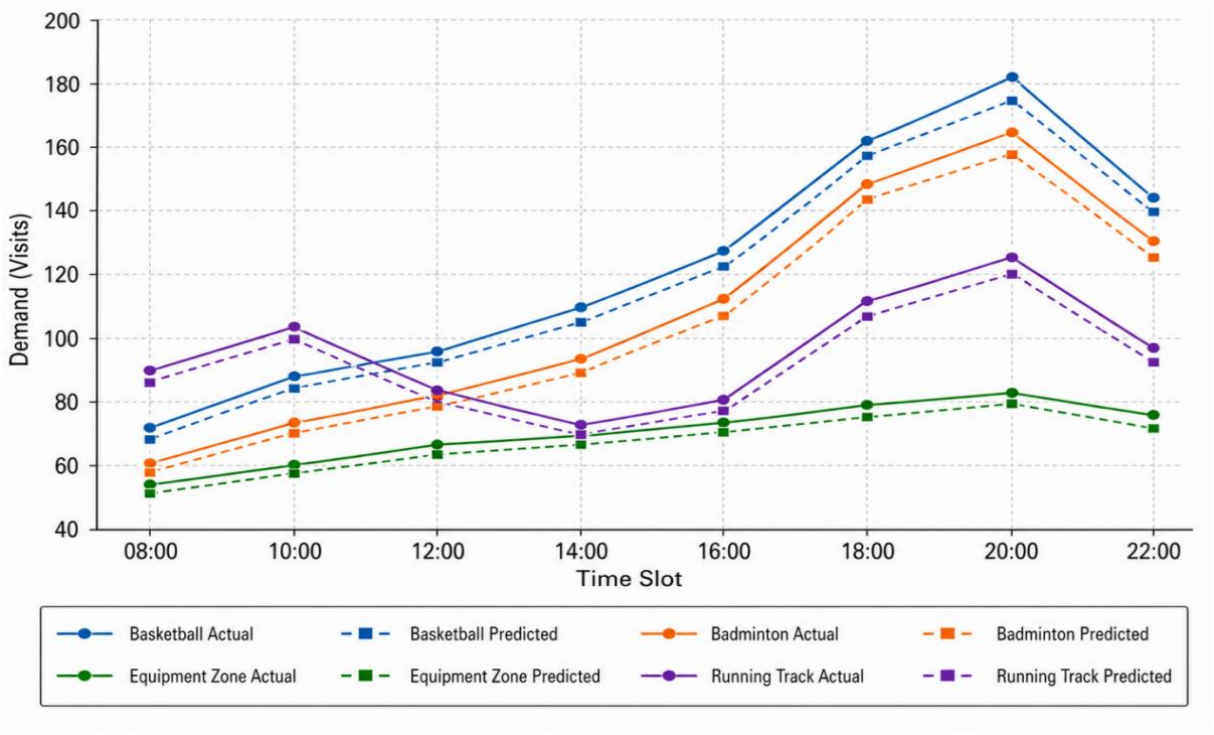

Figure 3: Comparison of demand prediction results of different fitness service projects

Figure 3 shows that the model has a good fitting effect on the demand change trend of the four types of fitness service projects. The basketball court and badminton court showed obvious peak demand in the evening, the real peak demand was 182 and 164, respectively, and the predicted value was 176 and 158, respectively, the deviation was controlled within 4%. The demand fluctuation of intelligent fitness equipment area is relatively smooth, and the average prediction error is 5.3%. The running trail is significantly affected by weather and time of day, but the model can still identify two use peaks in the morning and evening. Overall, the average absolute error of the four types of projects is 6.8 people, and the average relative error is 5.7%, which indicates that the demand forecasting module can provide a more stable prior input for subsequent dynamic scheduling.

After completing the demand prediction, this paper further identifies the resource load

levels of different service regions, and divides the service operation status into four categories: low load, medium load, high load and overload risk, which are used to determine whether each region needs to divert in advance, increase service supply or adjust opening hours. The identification results of resource load levels in different service areas are shown in Figure 4.

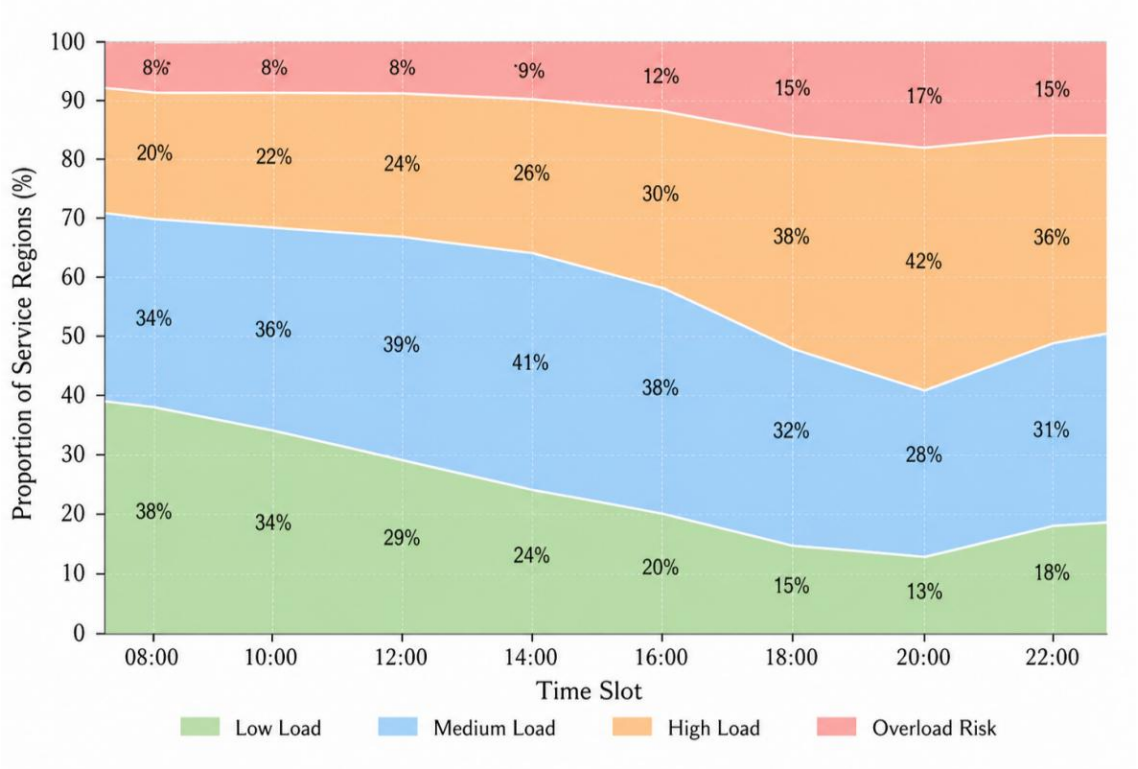

Figure 4: Stack area diagram of resource load level identification in different service areas

Figure 4 shows that the service area load level shows obvious changes with time. In the morning, low load and medium load accounted for a high proportion, totaling 72.4%. After the evening peak, the proportion of high load rose to 38.6%, and the proportion of overload risk reached 14.8%, mainly concentrated in the ball field and the central community gymnasium. The overall recognition accuracy of the model for four types of load levels is 92.7%, of which the recognition accuracy of high load is 91.3%, and the recognition accuracy of overload risk is 89.6%. The results show that the resource load identification module can accurately capture the changes of service pressure in different regions, which provides a reliable basis for PPO scheduling model for venue diversion, equipment allocation and user guidance.

4.3 Comparative analysis of resource utilization and service response efficiency under different scheduling schemes

In order to further evaluate the comprehensive performance of the proposed model at the level of resource scheduling, this paper selects four scheduling schemes, including manual experience scheduling, rule threshold scheduling, DQN scheduling and PPO scheduling, and makes comparative analysis from three dimensions of resource utilization, average service response time and user satisfaction, where bubble size represents user satisfaction. Figure 5 shows the trade-off between resource utilization and service response time under different scheduling schemes.

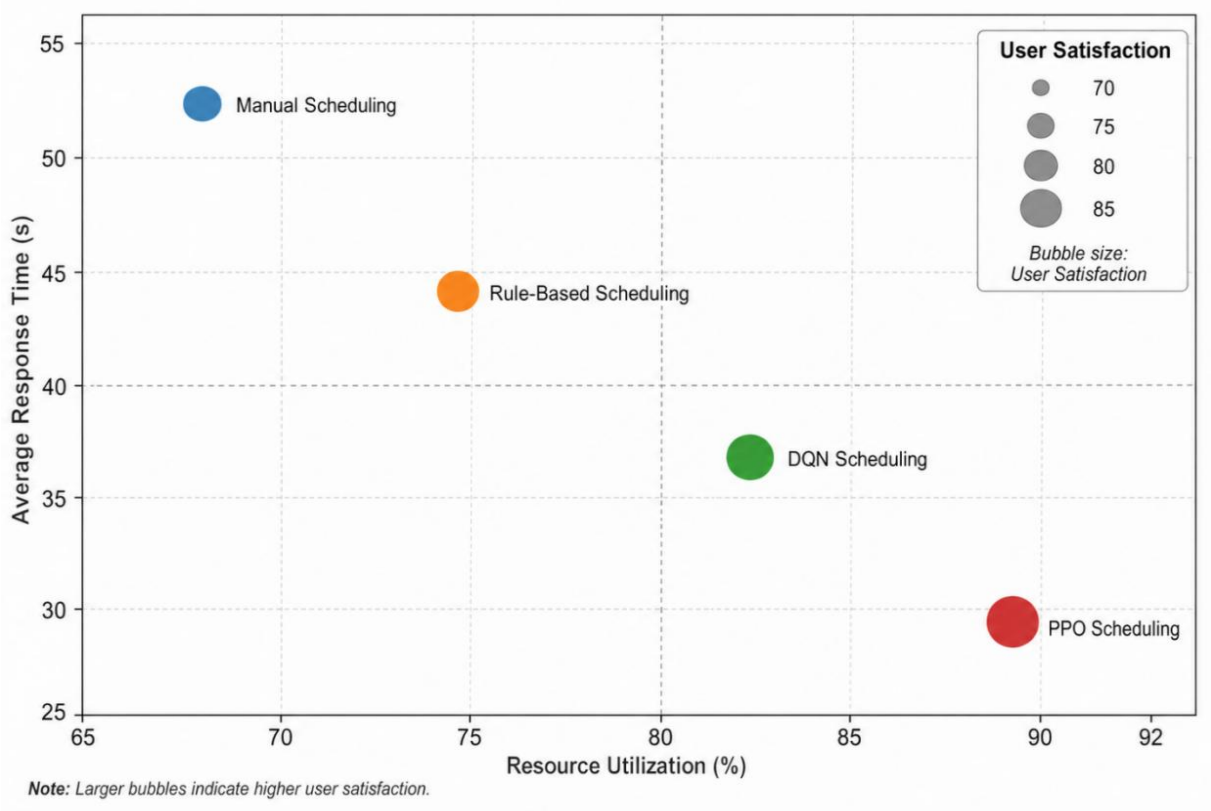

Figure 5: Scatter bubble plots of resource utilization and service response time tradeoffs under different scheduling schemes

Figure 5 shows that each scheduling scheme presents obvious differences between efficiency and response speed. The resource utilization rate of manual experience scheduling is only 68.4%, the average response time is 52.6 s, and the bubble size is also the smallest, indicating that the user satisfaction is relatively low. Although the regular threshold scheduling improves the resource utilization to 74.9% and reduces the response time to 44.1 s, its adaptability to peak demand is still limited. DQN scheduling further improves the resource utilization to 82.7%, reduces the average response time to 36.8 s, and the overall performance has been significantly improved. PPO scheduling is located in the "high utilizationlow response time" advantage region in the figure, resource utilization reaches 89.3%, average response time drops to 29.4 s, and user satisfaction reaches 91.8%. The results show that the proposed PPO scheduling model achieves a better balance between resource allocation efficiency and service response speed.

4.4 Stability verification of dynamic scheduling under peak demand and sudden disturbance scenarios

In order to verify the stability of the proposed model under complex operating conditions, this paper further sets up a joint scenario of peak demand and sudden disturbance, including the surge of reservations at night, the temporary closure of local venues, fitness equipment failure, and centralized cancellation of reservations by users, and compares the PPO scheduling scheme with the regular threshold scheduling and DQN scheduling. The investigated indexes include the dynamic changes of resource utilization, service completion rate and average response time to evaluate the adaptive scheduling ability of the model in a disturbed environment.Figure 6 shows the scheduling performance variation under peak demand and

burst disturbance.

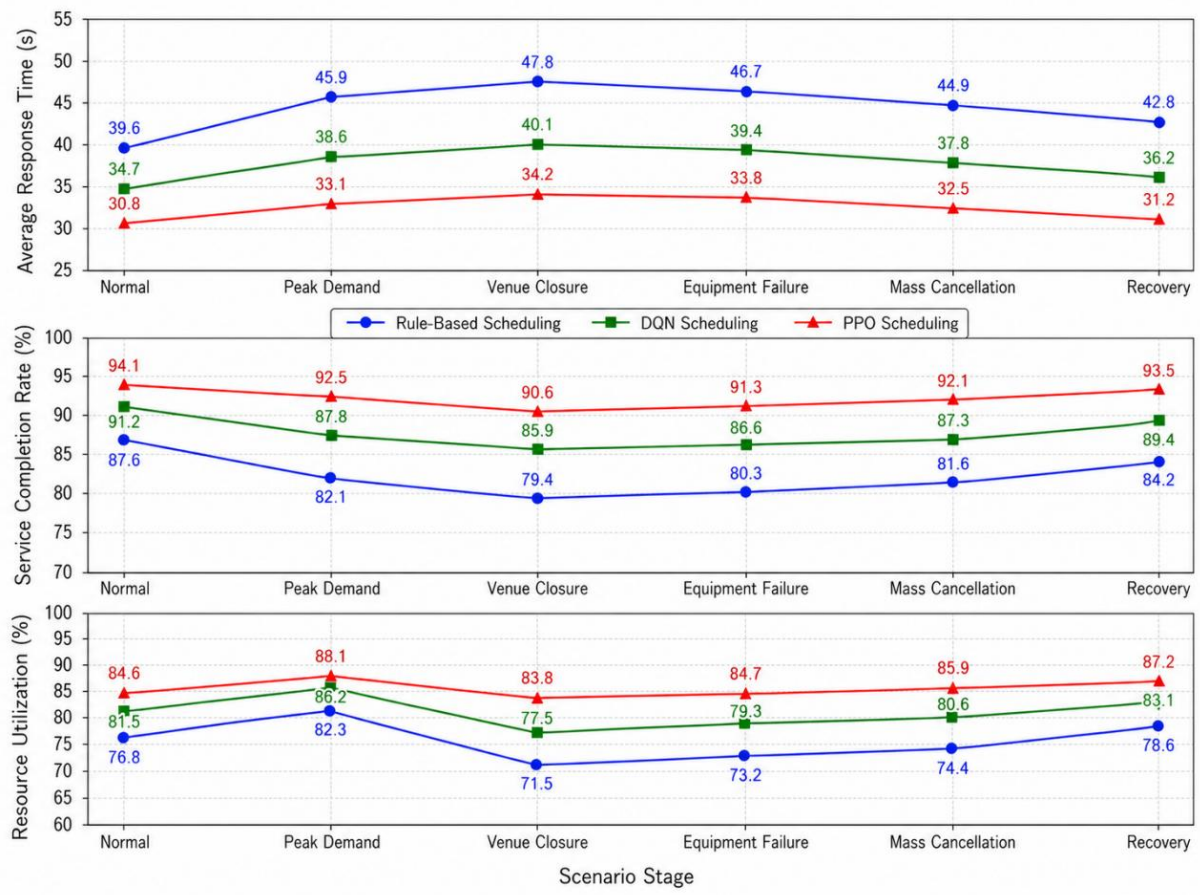

Figure 6: Curve of scheduling performance variation under peak demand and burst disturbance

Figure 6 shows that the performance of each scheduling scheme fluctuated to some degree after the sudden disturbance superimposed on the peak demand, but the PPO scheduling scheme was more stable overall. Compared with the regular threshold scheduling, the response time of PPO scheme was controlled at 34.2 s in the stage of museum closure disturbance, while the regular threshold scheduling reached 47.8 s. The corresponding service completion rates were 90.6% and 79.4%, respectively. Compared with DQN scheduling, the fluctuation range of resource utilization of PPO scheme is significantly reduced, and the fluctuation range is controlled within 4.3 percentage points in the disturbance stage, indicating that it can still maintain good dynamic coordination ability under the conditions of venue closure and demand sudden increase. The results show that the proposed method has better robustness and scheduling stability.

4.5 Convergence analysis of PPO model and service optimization effect

In order to further verify the training stability and effectiveness of the proposed model, this paper analyzes the convergence process of PPO strategy network and the ablation experimental results. The former is used to observe whether the reward growth and the policy loss decline are stable, and the latter is used to test the contribution of key modules such as demand prediction, resource load estimation, fairness reward and user feedback update to the

service optimization effect. Figure 7 shows the convergence of PPO training reward and policy loss.

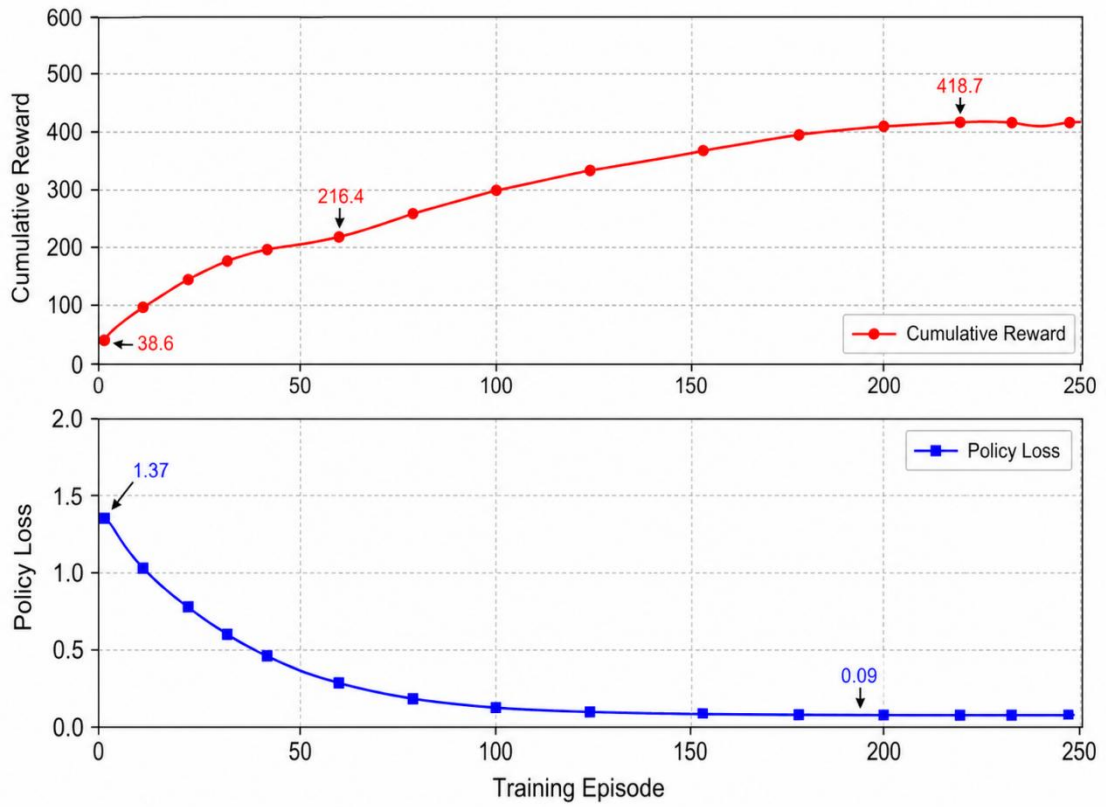

Figure 7: Convergence curve of PPO training reward and policy loss

Figure 7 shows that with the increase of training rounds, the cumulative reward of PPO model shows a continuous upward trend, rapidly increasing from 38.6 to 216.4 in the first 60 rounds, indicating that the policy network has initially learned to adjust resource allocation according to demand fluctuations. The cumulative reward continued to grow after the 120th round and stabilized around 418.7 around the 220th round, and the fluctuation range was controlled within ±8.5. At the same time, the strategy loss gradually decreases from 1.37 in the initial stage to 0.09, and is basically stable after the 200th round, indicating that the model update process is relatively stable, and there is no obvious oscillation or divergence. On the whole, the PPO model can complete effective convergence within a limited number of training rounds, and form a public fitness resource scheduling strategy that takes into account both efficiency and stability.

In order to investigate the specific influence of each component module on the service optimization effect, this paper further conducts ablation experiments, and the results are shown in Table 3.

Table 3: Comparison table of ablation experiments and service optimization effects

Model Configuration	Service Completion Rate / %	Average Response Time / s	Resource Utilization Rate / %	User Satisfaction / %	Scheduling Stability / %
Without Demand Prediction Module	87.8	35.6	84.2	86.5	87.9
Without Resource Load Estimation Module	88.6	34.9	85.1	87.4	88.6
Without Fairness Reward Term	89.4	33.8	86.7	85.9	89.1
Without User Feedback Update Module	90.1	32.7	87.5	88.2	89.8
Complete Model	93.5	29.4	89.3	91.8	92.6

Table 3 shows that the full model outperforms the ablation model in all metrics. After removing the demand prediction module, the service completion rate decreases to 87.8%, and the average response time increases to 35.6 s, which indicates that forward-looking demand identification plays an important role in scheduling decisions. After removing the fairness reward item, the resource utilization rate remained at 86.7%, but the user satisfaction decreased to 85.9%, indicating that the fairness target could not be ignored in the public service scenario. After removing the user feedback update module, the satisfaction and stability decreased by 3.6 percentage points and 2.8 percentage points, respectively. In contrast, the full model improves the service completion rate to 93.5%, and compresses the average response time to 29.4 s, indicating that the constructed PPO scheduling model has good comprehensive performance in convergence and service optimization effect.

5 Conclusion

Focusing on the problems of demand fluctuation, uneven distribution of resources and insufficient response efficiency in the public service of national fitness, this paper proposes a dynamic scheduling and service optimization method based on PPO reinforcement learning. The model constructs the running state representation through reservation records, passenger flow monitoring, equipment status, environmental information and user feedback, and combines time series prediction and resource load identification results to generate scheduling actions. In the reward design, the service completion rate, resource utilization rate, response time, fairness and satisfaction are comprehensively considered, so that the policy network can be continuously optimized under multi-objective constraints. Experiments show that the service completion rate of the proposed model reaches 93.5%, the resource utilization rate reaches 89.3%, and the average response time is compressed to 29.4 s. Under the peak demand and sudden disturbance, the response time of PPO scheme in the typical disturbance stage remains at 34.2 s, and the service completion rate reaches 90.6%. The training results show that the cumulative reward stabilizes around 418.7, and the policy loss decreases to 0.09. In the future, real city-level venue data and edge computing deployment mechanisms can be further introduced to improve the promotion value of the model in complex public sports

service scenarios.

Funding

This work was supported by Key Annal Project of Handan Municipal Social Science Planning, 2025. (Grant No. 2025591)

References

[1] Mazyavkina N, Sviridov S, Ivanov S, Burnaev E. Reinforcement learning for combinatorial optimization: A survey[J]. Computers & Operations Research, 2021, 134: 105400. DOI: 10.1016/j.cor.2021.105400.

[2] OroojlooyJadid A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning[J]. Applied Intelligence, 2023, 53: 13677-13722. DOI: 10.1007/s10489-022-04105-y.

[3] Wong A, Bäck T, Kononova A V, Plaat A. Deep multiagent reinforcement learning: Challenges and directions[J]. Artificial Intelligence Review, 2023, 56(6): 5023-5056. DOI: 10.1007/s10462-022-10299-x.

[4] Orr J, Dutta A. Multi-agent deep reinforcement learning for multi-robot applications: A survey[J]. Sensors, 2023, 23(7): 3625. DOI: 10.3390/s23073625.

[5] Zuccotto M, Castellini A, La Torre D, Mola L, Farinelli A. Reinforcement learning applications in environmental sustainability: A review[J]. Artificial Intelligence Review, 2024, 57: 88. DOI: 10.1007/s10462-024-10706-5.

[6] Park J, Chun J, Kim S H, Kim Y, Park J. Learning to schedule job-shop problems: Representation and policy learning using graph neural network and reinforcement learning[J]. International Journal of Production Research, 2021, 59(11): 3360-3377. DOI: 10.1080/00207543.2020.1870013.

[7] Lee Y H, Lee S. Deep reinforcement learning based scheduling within production plan in semiconductor fabrication[J]. Expert Systems with Applications, 2022, 191: 116222. DOI: 10.1016/j.eswa.2021.116222.

[8] Jayanetti A, Halgamuge S, Buyya R. Multi-agent deep reinforcement learning framework for renewable energy-aware workflow scheduling on distributed cloud data centers[J]. IEEE Transactions on Parallel and Distributed Systems, 2024, 35(4): 604-615. DOI: 10.1109/TPDS.2024.3360448.

[9] Hoang D T, Nguyen D N, Pham Q V. Multi-agent deep reinforcement learning for online resource allocation in wireless edge computing[J]. IEEE/ACM Transactions on Networking, 2023, 31(6): 2761-2776. DOI: 10.1109/TNET.2023.3263538.

[10] Hortelano D, Gayán J, García-García A, et al. A survey on deep reinforcement learning for resource allocation in communication networks[J]. Journal of Network and Computer Applications, 2023, 216: 103669. DOI: 10.1016/j.jnca.2023.103669.

[11] Hribar J, Marinescu A, Chiumento A, DaSilva L A. Energy-aware deep reinforcement learning scheduling for sensors correlated in time and space[J]. IEEE Internet of Things Journal, 2022, 9(9): 6732-6744. DOI: 10.1109/JIOT.2021.3114102.

[12] Yarahmadi H, Shiri M E, Challenger M, Navidi H, Sharifi A. Multi-agent credit assignment and bankruptcy game for improving resource allocation in smart cities[J]. Sensors, 2023, 23(4): 1804. DOI: 10.3390/s23041804.

[13] Haydari A, Yilmaz Y. Deep reinforcement learning for intelligent transportation systems: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(1): 11-32. DOI: 10.1109/TITS.2020.3008612.

[14] Kiran B R, Sobh I, Talpaert V, Mannion P, Al Sallab A A, Yogamani S, Pérez P. Deep reinforcement learning for autonomous driving: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 4909-4926. DOI: 10.1109/TITS. 2021.3054625.

[15] Aradi S. Survey of deep reinforcement learning for motion planning of autonomous vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(2): 740-759. DOI: 10.1109/TITS.2020.3024655.

[16] Alipio M, Bureš M. Deep reinforcement learning perspectives on improving reliable transmissions in IoT networks: Problem formulation, parameter choices, challenges, and future directions[J]. Internet of Things, 2023, 23: 100846. DOI: 10.1016/j.iot.2023. 100846.

[17] Bouktif S, Fiaz A, Ouni A, Serhani M A. Deep reinforcement learning for traffic signal control with consistent state and reward design approach[J]. Knowledge-Based Systems, 2023, 267: 110440. DOI: 10.1016/j.knosys.2023.110440.

[18] Ta-Dinh Q, Pham T S, Hà M H, Rousseau L M. A reinforcement learning approach for the online dynamic home health care scheduling problem[J]. Health Care Management Science, 2024, 27(4): 650-664. DOI: 10.1007/s10729-024-09692-5.

[19] Silva-Aravena F, Morales J, Jayabalan M, Sáez P. Optimizing MRI scheduling in high-complexity hospitals: A digital twin and reinforcement learning approach[J]. Bioengineering, 2025, 12(6): 626. DOI: 10.3390/bioengineering12060626.

[20] Abualrous R, Zouzou H, Zgheib R, Hasan A, Hijazi B, Kermani A. Fairness-aware intelligent reinforcement: An AI-powered hospital scheduling framework[J]. Information, 2025, 16(12): 1039. DOI: 10.3390/info16121039.

[21] Barbier A, Evrard B, Dermit-Richard N. Predictive modelling of sports facility use: A model of aquatic centre attendance[J]. Sustainability, 2023, 15(5): 4142. DOI: 10.3390/su15054142.

[22] Kuvaja-Köllner V, Kankaanpää E, Laine J, et al. Municipal resources to promote adult physical activity: A multilevel follow-up study[J]. BMC Public Health, 2022, 22: 1213. DOI: 10.1186/s12889-022-13617-8.

[23] Høyer-Kruse J, Schmidt E B, Faber A, Pedersen M R L. The interplay between social

environment and opportunities for physical activity within the built environment: A scoping review[J]. BMC Public Health, 2024, 24: 2361. DOI: 10.1186/s12889-024-19733-x.

[24] Hewamalage H, Bergmeir C, Bandara K. Recurrent neural networks for time series forecasting: Current status and future directions[J]. International Journal of Forecasting, 2021, 37(1): 388-427. DOI: 10.1016/j.ijforecast.2020.06.008.

[25] Lara-Benítez P, Carranza-García M, Riquelme J C. An experimental review on deep learning architectures for time series forecasting[J]. International Journal of Neural Systems, 2021, 31(3): 2130001. DOI: 10.1142/S0129065721300011.

[26] Masini R P, Medeiros M C, Mendes E F. Machine learning advances for time series forecasting[J]. Journal of Economic Surveys, 2023, 37(1): 76-111. DOI: 10.1111/joes.12429.

[27] Perera A T D, Kamalaruban P. Applications of reinforcement learning in energy systems[J]. Renewable and Sustainable Energy Reviews, 2021, 137: 110618. DOI: 10.1016/j.rser.2020.110618.

[28] Jendoubi I, Bouffard F. Multi-agent hierarchical reinforcement learning for energy management[J]. Applied Energy, 2023, 332: 120500. DOI: 10.1016/j.apenergy.2022.120500.