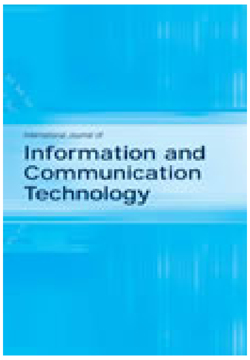
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

A reinforcement learning - enabled system for personalised sports training plan generation

Linli Zhou, Kaili Zhou

DOI: 10.1504/IJICT.2025.10075262

Article History:

Received:	18 October 2025
Last revised:	14 November 2025
Accepted:	15 November 2025
Published online:	12 January 2026

A reinforcement learning – enabled system for personalised sports training plan generation

Linli Zhou*

School of Physical Education,
Sichuan Technology and Business University,
Meishan, 620000, China
Email: s578604595@163.com
*Corresponding author

Kaili Zhou

Chengdu Qingling Qi'an Technology Co., Ltd.,
Sichuan Technology and Business University,
Chengdu, 610051, China
Email: kuangfabin@163.com

Abstract: Personalised sports training has emerged as a critical component in optimising athletic performance and minimising injury risks. Nevertheless, conventional approaches predominantly depend on coaches' subjective expertise, which often falls short in delivering dynamically precise adaptations. In response, this study introduces a reinforcement learning-based framework for generating individualised training regimens. By formulating the training process as a Markov decision process, the system enables an intelligent agent to interact with a simulated training environment, producing optimised training actions derived from real-time user status information. Evaluations conducted on the public FitRec dataset indicate that, relative to conventional baseline techniques, the proposed system yields an average improvement of 15% in predicted performance indicators, while concurrently lowering the incidence of training overload by 30%. These findings highlight the potential of the proposed framework as an effective new paradigm for automated and individualised sports science training.

Keywords: reinforcement learning; RL; personalised sports training; Markov decision process; MDP; reward function; FitRec dataset.

Reference to this paper should be made as follows: Zhou, L. and Zhou, K. (2025) 'A reinforcement learning – enabled system for personalised sports training plan generation', *Int. J. Information and Communication Technology*, Vol. 26, No. 50, pp.72–87.

Biographical notes: Linli Zhou is a Lecturer in the School of Physical Education at Sichuan Technology and Business University. She obtained a Master's degree from Chengdu Institute of Physical Education in 2020. Her research interests include sports industry economy, sports education and sports industry management.

Kaili Zhou is a Technology and Safety Expert in the Chengdu Qingling Qi'an Technology Co., Ltd. He received a Bachelor's degree from Sichuan University in 2012. His interests include education and training, artificial intelligence and smart security.

1 Introduction

As national fitness initiatives gain momentum and the pursuit of healthy lifestyles becomes increasingly prevalent, the demand for scientific and individualised sports training has moved to the forefront of exercise science. Conventional training programs, often designed with a one-size-fits-all philosophy, fail to accommodate the considerable variations among individuals in terms of physiological traits, recovery ability, and performance goals (Akenhead and Nassis, 2016). Such limitations not only hinder the optimal development of athletic potential but also elevate the risk of injury caused by inappropriate training loads (Halson, 2014). Common injuries stemming from inappropriate loads include musculoskeletal overuse conditions like muscle strains, stress fractures, and tendinopathies. If mismanaged, excessive training can also lead to non-functional overreaching and overtraining syndrome, significantly hindering athletic development. Consequently, there is a compelling need to develop intelligent systems capable of generating adaptive training plans that align closely with users' evolving physiological states. Although coaches' expertise currently serves as the primary means of personalisation, this approach is inherently subjective, difficult to scale, and inadequate for processing high-volume, multi-dimensional physiological data in real-time – posing a major obstacle to the broad implementation of tailored training methodologies.

The rapid evolution of wearable sensor technology in recent years has enabled continuous, large-scale acquisition of physiological and kinematic data – such as heart rate, blood oxygen saturation, acceleration, and global positioning system (GPS) trajectories – from athletes and fitness participants (Cossich et al., 2023). This capability provides a robust technical foundation for data-driven training design. In this context, researchers have increasingly turned to computational intelligence to advance sports science. Early attempts included expert systems that operated on predefined rule sets to generate recommendations, yet their rigidity limited applicability in complex, dynamically changing training environments. Subsequent adoption of traditional machine learning techniques, including support vector machines and random forests, allowed for predictions of training outcomes or injury risks (Claudino et al., 2019). However, these models primarily performed static, single-timepoint predictions, failing to conceptualise training as a continuous and adaptive decision-making sequence. As a result, they were unable to resolve the essential sequential decision problem: 'What is the optimal action to take in the current state to maximise long-term training benefits? ' – which lies at the heart of true personalisation. These inherent shortcomings motivate the pursuit of more expressive and dynamic modelling paradigms.

Amid this exploration, reinforcement learning (RL) has emerged as a promising machine learning framework specialised in sequential decision-making. At its core, RL operates through an agent that progressively learns an optimal policy by interacting with an environment and maximising cumulative rewards via trial and error. This learning mechanism bears a natural resemblance to the decision-making process of a coach, who continuously adapts training content (actions) according to an athlete's real-time physiological feedback (state) in pursuit of long-term performance gains (reward). Notably, RL has already achieved significant success in healthcare applications – such as personalised dosage optimisation and chronic disease management – demonstrating its strength in handling long-horizon, uncertain sequential decisions that require a careful balance between benefits and risks (Mulani et al., 2019). Nevertheless, the direct application of RL to sports training prescription entails several domain-specific challenges: how to construct a state space that accurately captures the user's physiological status and accumulated fatigue; how to design reward functions that promote performance gains while mitigating overtraining risks; and how to overcome issues of data sparsity and privacy in real-world deployment. Practical implementation faces barriers such as the requirement for high-quality, continuous physiological data and non-trivial computational resources. Furthermore, integration into existing coaching infrastructures presents a significant adoption hurdle.

This study is motivated by the need to overcome the challenges outlined above (Ghosh et al., 2023). We posit that an effective personalised training system must account for the longitudinal, dynamic, and safety-critical nature of athletic development. To that end, this work investigates both methodological advances and practical applications of RL in this context. Central to our approach is the construction of a Markov decision process (MDP) framework, built around a multidimensional state representation that combines real-time physiological indicators with historical training loads, as well as a composite reward function designed to balance immediate training stimuli against long-term adaptation, and performance gains against injury prevention. By integrating systematic environment modelling with judicious algorithm selection, our goal is to equip the learning agent with the capability to emulate the decision quality of an expert coach – producing personalised training plans that are scientifically grounded, safe, and adaptive.

2 Related work

2.1 Research status of personalised sports training systems

The development of personalised training plans has long been a central pursuit in exercise science and sports engineering. Early computer-aided training systems were predominantly built on rule-based expert systems, which formalised the empirical knowledge of expert coaches into deterministic logic – for instance, 'if heart rate exceeds threshold X, reduce intensity' (Bonidia et al., 2018). Early systems, such as the research prototype TRAINER and commercial technologies like Firstbeat, formalised expert rules – for instance, using heart rate thresholds to recommend intensity adjustments – laying the groundwork for automated training prescription. While these systems introduced a degree of automation, their rigidity made them ill-suited to handle complex or unforeseen scenarios beyond predefined rules, and they incurred substantial costs in knowledge engineering and maintenance. With the rise of machine learning, researchers turned to static predictive models to support personalisation (Rein and Memmert, 2016). Regression techniques, for example, were used to forecast performance under varying training loads, while classification models helped estimate injury risk (Lames and McGarry, 2007). Yet, these methods largely focused on isolated, single-step predictions, treating each training session as an independent event rather than part of a continuous, adaptive process. By failing to model training as a sequential decision-making problem with long-term dependencies, such systems cannot dynamically adapt future sessions in response to an athlete's evolving state – a fundamental limitation in pursuing truly optimised, long-term training planning. In current practice, coaches often compensate for this limitation through periodic reassessments and manual plan adjustments. However, this approach is inherently subjective, difficult to scale, and lacks the dynamic responsiveness of an automated system.

2.2 Applications of reinforcement learning in health and management

RL has attracted considerable interest and achieved notable successes in healthcare and personalised management, owing to its strengths in solving sequential decision-making problems – offering valuable references for potential applications in sports training. In clinical settings, RL has been utilised to create tailored treatment regimens, such as optimising chemotherapy dosages to balance therapeutic effects against side effects, or devising adaptive insulin administration strategies for diabetic patients. In clinical settings, RL has shown promise in areas like personalised insulin dosing for diabetes management and optimised chemotherapy scheduling in oncology. These successes demonstrate its capability for adaptive, long-term decision-making based on individual patient states. In the realm of health behaviour management, RL has also been applied to generate individualised nutrition plans or psychological intervention strategies to promote sustained healthy habits (Gottesman et al., 2019). These applications share a fundamental characteristic: the need to make sequential decisions over extended periods, maximising cumulative benefits based on the user's dynamically changing state – such as physiological indicators and behavioural feedback (Yu et al., 2021). These achievements underscore RL's adaptability and efficacy in addressing personalised, temporal decision tasks. That said, sports training presents distinct challenges compared to clinical contexts: performance improvements – as reward signals – are often more delayed and harder to quantify (Bartlett et al., 2015). Moreover, the state representation must integrate not only physiological parameters but also complex motor performance metrics, all while ensuring high levels of training safety in real-time (Buchheit and Simpson, 2017). As a result, direct transfer of medical RL models is infeasible; domain-specific innovations are required (Clifton and Laber, 2020).

2.3 Applicable to reinforcement learning algorithms for training generation

The selection of RL algorithms plays a decisive role in determining the performance and reliability of the resulting system. In training plan generation tasks, where the action space – encompassing variables such as intensity and duration – is inherently continuous, classical algorithms such as Q-learning and deep Q-network (DQN) become unsuitable due to their limitation to discrete action domains. While deep deterministic policy gradient (DDPG) and its variants can handle continuous control, they often exhibit sensitivity to hyperparameter settings and are prone to training instability. In contrast, proximal policy optimisation (PPO) introduces a clipped objective function that constrains policy update steps, thereby significantly improving training stability and sample efficiency without compromising performance (Wang et al., 2020). Training stability is paramount in this domain to prevent the generation of erratic and potentially harmful training plans. PPO's constrained policy updates ensure consistent and safe recommendations, directly supporting athlete well-being. This attribute makes PPO particularly suitable for sports training applications, where data availability is often limited – such as in athlete-specific cases – and operational safety is paramount. As a result, PPO and its derivatives are frequently adopted in studies demanding robust performance under complex and constrained conditions. Ultimately, algorithm selection must carefully balance problem characteristics, data scale, and stringent requirements for safety and stability. Our methodological section will subsequently elaborate on the specific algorithmic choices and enhancements made in this work.

3 Methodology

3.1 Problem formulation: a Markov decision process framework

To transform the problem of generating personalised training plans into one solvable by RL, we first formalise it as a MDP (Sutton and Barto, 1998). An MDP can be represented by a quintuple: $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$. This formulation mirrors a coach's intuitive process: evaluating the athlete's current condition (state), selecting a training regimen (action), and aiming for long-term performance peaks (maximising cumulative reward), thereby structuring this adaptive decision-making loop. Here, $\mathcal{S}$ represents the state space, encompassing all information that can describe the user's state; $\mathcal{A}$ represents the action space, i.e., all possible training actions the system can recommend; $\mathcal{P}(s' \mid s, a)$ is the state transition probability, indicating the probability of transitioning from state s to state s' after executing action a in state s; $R(s, a)$ is the reward function, quantifying the quality of performing action a in state s; $\gamma \in (0, 1)$ is the discount factor, balancing the importance of immediate rewards versus future long-term rewards.

Within this framework, the agent (our system) observes the current state of the environment $s_t \in \mathcal{S}$ at each discrete time step t (typically representing a training day or cycle) (Mnih et al., 2015). Based on this state, the agent takes an action $a_t \in \mathcal{A}$ (i.e., generates a training plan). The environment (i.e., the user model or simulator) transitions to a new state $s_{t+1} \sim \mathcal{P}\left(s_{t+1} \mid s_t, a_t\right)$ according to this action and provides the agent with a scalar reward feedback $r_i = R(s_i, a_i)$. The agent's objective is to learn a policy $\pi(a|s)$ that maximises the expected cumulative discounted reward obtained starting from the initial state, i.e., the return $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$.

$$(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma) \tag{1}$$

where $\mathcal{S}$ represents the state space, encompassing all information that can describe the user's state; $\mathcal{A}$ represents the action space, i.e., all possible training actions that the system can recommend; $\mathcal{P}(s' \mid s, a)$ denotes the state transition probability, representing the likelihood of transitioning from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ after executing action $a \in \mathcal{A}; R(s, a)$ is the reward function, quantifying the immediate reward obtained by executing action a in state s; $\gamma \in (0, 1)$ is the discount factor, used to balance the importance of immediate rewards versus future long-term rewards.

3.2 State space design

State s_t must comprehensively and effectively characterise the user's physical condition, fatigue level, and training history at time t (Bourdon et al., 2017). Our state space design incorporates two major categories of features: instantaneous features and historical window features.

Let F_t denote the set of all instantaneous physiological and subjective features observable at time step t, with dimension m. Simultaneously, we consider a historical window of length L, containing historical features and action information from time steps $t–L$ to $t–1$. The state vector s_t is formed by concatenating these features:

$$s_t = (F_t, H_t) \tag{2}$$

where $F_t = \left(f_t^1, f_t^2, \ldots, f_t^m\right)$ contains the current instantaneous features, such as: f_t^1 is resting heart rate (RHR); f_t^2 is heart rate variability (HRV); f_t^3 is previous day's Rate of Perceived Exhaustion (RPE); f_t^4 is sleep quality score after the previous training session; H is represents historical features, computed by calculating statistics (such as mean and standard deviation) from data over the past L days, for example: H_t^1 is average training load over the past 7 days; H_t^2 is variability of training load over the past 7 days (standard deviation); H_t^3 is acute-to-chronic workload ratio (ACWR) over the past 3 days relative to the past 28 days, used to quantify injury risk; H_t^4 is total training duration over the past week.

This design ensures that the state not only reflects the user's immediate response but also encodes their recent training trends and accumulated fatigue, providing a robust basis for the agent to make informed decisions (Scott et al., 2016). Instantaneous features provide a snapshot of the user's immediate physiological state and readiness. In contrast, historical features encapsulate recent training trends and accumulated fatigue, together offering a holistic view crucial for robust decision-making.

3.3 Action space design

Action a_t represents the personalised training plan generated for the user at time step t. We define it as a continuous, multidimensional action vector to enable fine-grained control.

$$a_t = \left(a_t^{type}, a_t^{\text{int}\,ensity}, a_t^{duration}\right) \tag{3}$$

where $a_t^{type}, \in [0,1]$: A continuous value representing the training type bias. 0 denotes pure aerobic endurance training, 1 denotes pure strength training, and intermediate values represent mixed training. $a_t^{intensity}, \in [0,1]$: A continuous value representing the relative intensity of the training session. 0 corresponds to extremely low intensity (e.g., active recovery), while 1 corresponds to extremely high intensity (e.g., interval sprints). This value can be mapped to a specific percentage range of the user's maximum heart rate (HRmax) or maximum oxygen uptake (VO$_2$max). $a_t^{duration} \in [0,1]$: A continuous value representing the relative duration of the training session. 0 corresponds to the shortest training duration (e.g., 20 minutes), while 1 corresponds to the longest training duration (e.g., 120 minutes). This value will be linearly mapped to a predefined duration range.

Continuous action spaces offer finer-grained and more flexible control capabilities than discrete actions (such as ‘low, medium, high’), enabling agents to generate an infinite variety of possible training plan combinations. While continuous action spaces provide finer control and greater flexibility, they can introduce higher computational complexity and demand more interaction data for the policy to converge effectively compared to discrete action representations.

3.4 Reward function design

The reward function *R*(*s*, *a*) is key to guiding the agent in learning correct behaviours (Ng et al., 1999). Our goal is to simultaneously enhance athletic performance and ensure training safety, so the reward function is designed as a weighted sum of multiple sub-rewards.

$$R(s_t, a_t) = w_{perf} \cdot R_{perf}(s_t, a_t) + w_{safe} \cdot R_{safe}(s_t, a_t) + w_{pref} \cdot R_{pref}(s_t, a_t) \tag{4}$$

- Performance reward: this reward incentivises actions that effectively enhance users’ physical fitness (Moesch et al., 2018). Since performance improvements are long-term and delayed, we employ a proxy metric based on state changes. For example, we can utilise the predicted outputs from validated fitness models (such as the Banister model) as the foundation for rewards.

$$R_{perf}(s_t, a_t) = \Delta Fitness\ predicted - \Delta Fatigue\ predicted \tag{5}$$

- Safety reward: this reward is applied to penalise actions that may lead to overtraining or excessive injury risk (Gabbett, 2016). It functions as a penalty.

$$R_{safe}(s_t, a_t) = -\mathbb{I}\left(ACWR > \theta_{high}\right) \cdot \alpha - \mathbb{I}\left(RPE > \theta_{rpe}\right) \cdot \beta \tag{6}$$

- Preference reward: this reward makes the plan more aligned with the user’s personal preferences, such as encouraging the selection of training types historically preferred by the user to enhance plan adherence (Deci and Ryan, 2000).

$$R_{pref}(s_t, a_t) = similarity\left(a_t^{type}, user_p ref_{type}\right) \tag{7}$$

3.5 Algorithm: proximal policy optimisation

We employ the PPO algorithm to learn the optimal policy $\pi_\theta(a|s)$. PPO is a policy gradient algorithm widely favoured for its training stability, sample efficiency, and ease of parameter tuning. Its core principle involves constraining the magnitude of change between the new and old policies at each update to prevent disruptive policy shifts.

PPO updates the policy parameters θ by maximising a surrogate objective function:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, clip\left(r_t(\theta), 1-\in, 1+\in\right)\hat{A}_t\right)\right] \tag{8}$$

where $r_t(\theta) = \dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}$ is the probability ratio, representing the change in the probability of selecting an action under the new policy relative to the old policy. $\hat{A}_t$ is the estimated value of the advantage function at time step *t*, measuring the relative strength of action a_t compared to the average. We employ generalised advantage estimation (GAE) to compute $\hat{A}_t$ (Li and He, 2023).

$$\hat{A}t^{GAE(\gamma,\lambda)} = \sum l = 0^{\infty} (\gamma\lambda)^l \delta_{t+l}\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t) \tag{9}$$

where $V_\phi(s)$ is a state value function network parameterised by ϕ, which estimates the expected return obtainable starting from state *s*. The network is updated by minimising the mean squared error between its output and the target return:

$$L^{VF}(\phi) = \hat{\mathbb{E}}t\left[\left(V\phi(s_t) - V_t^{t\arg}\right)^2\right] \tag{10}$$

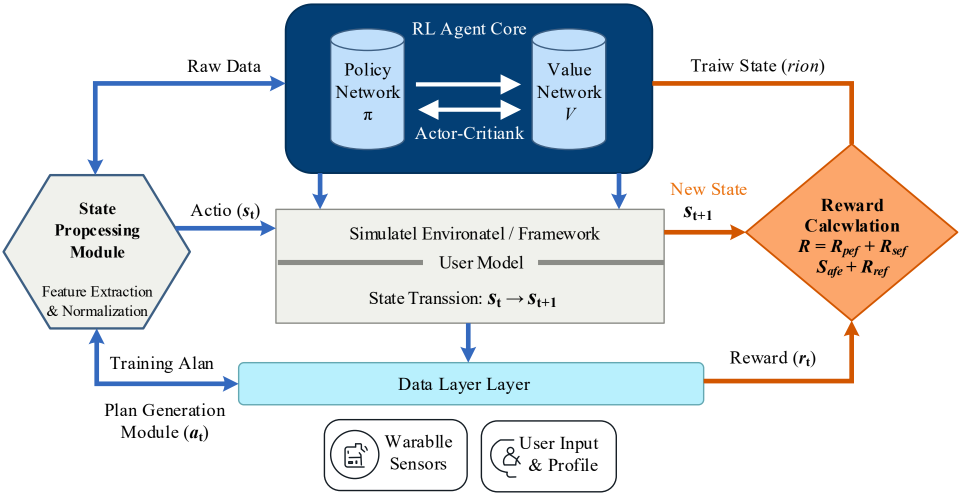
where is the target value of the value function.

Ultimately, the overall objective function of the PPO is the sum of the clipped surrogate objective, the value function loss term, and the policy entropy reward term (Williams, 1992):

$$L^{Total}(\theta, j) = \hat{\mathbb{E}}_t\left[L^{CLIP}(\theta) - c_1 L^{VF}(j) + c_2 S\pi_\theta\right] \tag{11}$$

where $V_t^{t\arg} = \tilde{A}t + V\phi_{old}(s_t)$ denotes the entropy of policy π_e in state s_t, serving to incentivise exploration; c_1 and c_2 are coefficients.

Figure 1 The architecture of the proposed reinforcement learning-based personalised training plan generation system (see online version for colours)



3.6 System architecture and training

Our system architecture is shown in Figure 1, with the training process conducted within a simulated environment. This environment is constructed based on historical data from public datasets, where the state transition function $\mathcal{P}$ is implemented by a pre-trained probabilistic model (such as a recurrent neural network or Gaussian process). This model predicts the next state s_{t+1} and immediate reward r_t based on the current state s_t and action a_t. The agent (PPO algorithm) interacts with this simulated environment, collecting experience traces (s_t, a_t, r_t, s_{t+1}). It continuously updates the policy network π_e and value network V_ϕ using this data until the policy converges.

4 Experimental verification

To comprehensively evaluate the performance of this system (hereafter referred to as PPO-TP), we designed a series of experiments aimed at addressing the following key questions:

1. Does PPO-TP demonstrate superior advantages over existing methods in enhancing athletic performance and ensuring safety?
2. Are all components of the designed reward function both necessary and effective?
3. Can the system generate genuinely personalised training plans?

4.1 Experimental setup

4.1.1 Dataset and pre-processing

This experiment utilises the publicly available FitRec dataset (Liu et al., 2023). This dataset contains activity watch records from over a thousand users, covering multiple exercise modes such as running and cycling. Data dimensions include heart rate, speed, elevation, GPS tracks, and user-reported metadata such as height, weight, and RHR. We performed rigorous data pre-processing: First, we selected active user samples with complete records and continuous activity exceeding 90 days. Second, we filled missing values using linear interpolation from preceding and subsequent time steps. Finally, we calculated the training load for each training day based on the training impulse algorithm and exercise duration, further deriving key features such as the ACWR (Banister and Calvert, 1980). After pre-processing, we obtained a total of 18,000 valid training days from 200 users. We allocated 80% of the user data for training, 10% for validation, and the remaining 10% for final testing.

4.1.2 Baseline methods

For fair comparison, we selected the following representative baseline methods:

1. Rule-based (ACWR): a baseline based on widely accepted rules in the field of exercise science (Hulin et al., 2016). It recommends training load levels based on the user’s real-time ACWR value range (< 0.8: decrease; 0.8–1.3: maintain; > 1.3:

increase). This rule is widely used for load management in professional athletes and serves as a robust empirical baseline.

2 XGBoost: a powerful gradient-boosted tree model. We employ it as a static prediction model, inputting the current state to predict the 'optimal' training action for the next training day. This baseline is used to contrast the performance differences between sequential decision methods and static prediction approaches.

3 DDPG: a classic DDPG algorithm, representing a leading approach for continuous action space RL problems. We train it using the exact same state space and reward function as PPO-TP.

4 A2C: an actor-critic synchronous policy gradient algorithm, serving as another advanced policy gradient method for comparison with PPO.

4.1.3 Evaluation metrics

Since genuine long-term physiological feedback cannot be obtained in the simulation environment, we employ the following domain-validated proxy metrics for assessment:

1 Performance metric (higher is better): evaluate the cumulative value of predicted fitness gains (predicted fitness gain) at the end of the entire testing cycle using a pre-trained predictive model (e.g., a model predicting VO_2max changes) on a retained test set.

2 Safety metric (lower is better): percentage of days during the entire testing cycle where the user's state shows ACWR > 1.5 (high-risk threshold).

3 Personalisation metric (higher is better): calculate the cosine similarity between the generated plan sequence and the user's historical preference sequence (e.g., training type distribution).

4.1.4 Implementation details

Both the strategy network and the value network of PPO-TP are two-layer fully connected neural networks (256–128 neurons) using the rectified linear unit (ReLU) activation function. The discount factor $\gamma = 0.99$, the GAE parameter $\lambda = 0.95$, and the tailoring range $\in = 0.2$. The reward function weights are set to $w_{perf} = 1.0$, $w_{safe} = 0.6$, and $w_{pref} = 0.3$. All RL baselines are trained with the same network structure and reward function.

4.2 Results and analyses

4.2.1 Comparative experimental analysis

The average performance of all methods on the test set is shown in Table 1.

Analysis of Table 1 reveals that our proposed PPO-TP method significantly outperforms all baseline methods in both physical fitness gain prediction and personalised similarity metrics. Simultaneously, it matches the best rule-based method in safety metrics while substantially outperforming other data-driven approaches. Specifically: rule-based methods exhibit the highest safety but the lowest performance gains, indicating that conservative rules mitigate risks while simultaneously limiting training effectiveness maximisation; XGBoost static models achieve notable performance gains but demonstrate the poorest safety. This stems from their lack of long-term planning capabilities, making them prone to recommending plans that appear effective in the short term but accumulate excessive risks over extended periods; The performance of DDPG and A2C confirms the effectiveness of RL methods for this task, though both are less stable than PPO (DDPG's high standard deviation reflects its training instability). PPO achieves more robust performance and higher safety through its clipping mechanism.

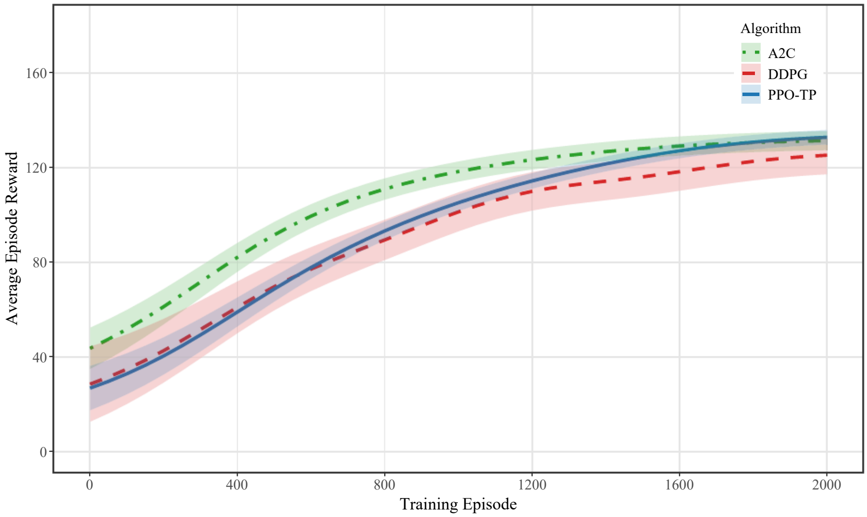
Table 1 Comparison of average performance across methods on the test set (mean ± standard deviation)

Methodologies	*Predicted fitness gains* (*Arbitrary unit*)	*Proportion of days with ACWR > 1.5* (%)	*Personalised similarity*
Rule-based (ACWR)	100.0 ± 10.5	5.2 ± 2.1	0.65 ± 0.12
XGBoost	118.3 ± 11.2	15.7 ± 3.8	0.71 ± 0.10
DDPG	125.6 ± 19.8	8.5 ± 3.5	0.68 ± 0.14
A2C	130.1 ± 12.6	7.1 ± 2.9	0.73 ± 0.11
PPO-TP (Ours)	138.5 ± 9.7	4.8 ± 1.9	0.79 ± 0.08

4.2.2 *Training stability and convergence analysis*

We plotted the average reward per round over training for PPO-TP alongside DDPG and A2C, as shown in Figure 2.

Figure 2 Comparison of average episode reward during training (see online version for colours)



Notes: The solid lines represent the mean over 5 random seeds, and the shaded regions represent the standard deviation

It can be observed that the reward curve of PPO-TP exhibits the smoothest and most stable upward trend throughout the training process, ultimately converging to the highest value. In contrast, the DDPG curve fluctuates dramatically and experiences performance collapse in the later stages, confirming its well-known instability issues. Although A2C converges rapidly, its final performance falls below that of PPO-TP.

4.2.3 Personalised generation case study

To visually demonstrate personalised outcomes, we selected three distinct user profiles from the test dataset:

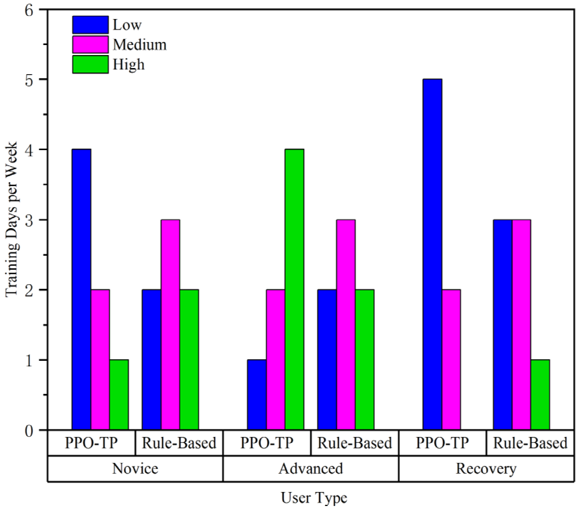
a Beginner user: low historical load, prefers low-intensity aerobic exercise

b Advanced runner: high historical load, prefers high-intensity interval training

c Recovery-phase user: recent ACWR elevated, requiring load reduction.

Figure 3 presents grouped bar charts illustrating the distribution of training intensity across weekly plans generated by PPO-TP and the rule-based baseline for three representative user types.

Figure 3 Weekly training intensity distribution for three user types (see online version for colours)



The results in Figure 3 clearly demonstrate the personalised adaptation of PPO-TP compared to the rule-based method. While the rule-based approach generates nearly identical plans across all user categories with only minor ACWR-based adjustments, PPO-TP produces distinctly tailored plans: for novice users, it emphasises low-intensity training (4 days per week); for advanced runners, it incorporates substantial high-intensity sessions (4 days per week); and for recovery-phase users, it significantly reduces overall training load by eliminating high-intensity sessions altogether. This demonstrates PPO-TP's exceptional personalisation capabilities and adherence to safety constraints.

- Ablation study: to validate the necessity of each component in the reward function, we conducted ablation experiments, with results shown in Table 2.

Table 2 Melting experiment results

Model variant	*Predicted fitness gains (Arbitrary unit)*	*Proportion of days with ACWR > 1.5 (%)*
PPO-TP (Full)	138.5 ± 9.7	4.8 ± 1.9
w/o R_{safe}	142.1 ± 11.3	18.2 ± 4.5
w/o R_{pref}	135.2 ± 10.1	5.1 ± 2.0
w/o R_{safe} and R_{pref}	145.0 ± 20.1	22.5 ± 5.8

Experiments demonstrate that removing the safety reward (w/o R_{safe}) yields slightly higher performance gains for the model, but its risk (proportion of days with ACWR > 1.5) surges dramatically, proving that R_{safe} is crucial for constraining model behaviour and ensuring safety. Removing the preference reward (w/o R_{pref}) resulted in performance and safety metrics that were largely comparable to the full model, but personalised similarity significantly decreased (from 0.79 to 0.68, not shown in the table), indicating that this component primarily enhances the alignment of the plan. Simultaneously removing both rewards caused the model to become extremely aggressive and unstable, validating the necessity of multi-objective reward function design.

4.3 Discussion and limitations

The experimental findings confirm the effectiveness and reliability of our proposed PPO-TP system in generating personalised training plans. This success stems from three key factors:

1 the accurate formulation of the training planning problem as a MDP

2 the design of a multi-objective reward function that effectively balances performance, safety, and user preference

3 the adoption of the stable and sample-efficient PPO algorithm.

Nevertheless, several limitations should be acknowledged. First, the experimental validation was conducted in a simulated environment, where system performance is inherently dependent on the accuracy of the state transition prediction model. Although we utilised a publicly available dataset with rigorous pre-processing, the generalisability of our conclusions requires further verification through real-world clinical trials with human participants. Second, the current system implementation primarily incorporates physiological indicators; future enhancements could integrate additional multimodal data streams – such as movement posture information captured by inertial sensors or computer vision techniques – to further refine training prescriptions and mitigate injury risks.

5 Conclusions

This study presents the design and implementation of a personalised training plan generation system (PPO-TP) based on the PPO algorithm, which formulates training prescription as a MDP. By constructing a carefully designed state representation, continuous action space, and a composite reward function that incorporates performance, safety, and user preference objectives, the system learns to produce long-term training strategies that dynamically adjust according to the athlete's evolving status. Evaluation on the public FitRec dataset shows that PPO-TP outperforms both rule-based methods and traditional machine learning models in terms of predicted performance gains, while simultaneously maintaining lower levels of injury risk and demonstrating stronger personalisation capability. Ablation studies further confirm the contribution of each reward component, supporting the rationality of the multi-objective optimisation strategy.

The main theoretical contribution of this work is a RL-based framework for data-driven personalised training generation. This framework not only illustrates the capability of deep RL in addressing complex sequential decision-making tasks, but its MDP formulation and reward design – such as using ACWR to model injury risk – also provide meaningful references for other domains involving adaptive personalised planning.

On the practical side, this study offers a viable technical pathway toward building intelligent training assistance systems. The proposed approach can serve as a decision support tool for professional coaches or as the core reasoning module in fitness applications, delivering scientifically-grounded, individualised training guidance that promotes performance improvement and reduces injury incidence.

Despite these promising results, this work has several limitations. The experimental evaluation was conducted primarily in simulation, and the actual efficacy of the system needs to be further verified in real-world settings through longitudinal user studies. Future efforts will focus on deployment in practical training environments and the incorporation of multimodal data sources – such as movement technique analysis from video – to improve the precision and applicability of the generated plans.

Declarations

All authors declare that they have no conflicts of interest.

References

Akenhead, R. and Nassis, G.P. (2016) 'Training load and player monitoring in high-level football: current practice and perceptions', *International Journal of Sports Physiology and Performance*, Vol. 11, No. 5, pp.587–593.

Banister, E.W. and Calvert, T.W. (1980) 'Planning for future performance: implications for long term training', *Canadian journal of applied sport sciences. Journal Canadien Des Sciences Appliquees Au Sport*, Vol. 5, No. 3, pp.170–176.

Bartlett, J.D., Hawley, J.A. and Morton, J.P. (2015) 'Carbohydrate availability and exercise training adaptation: too much of a good thing?', *European Journal of Sport Science*, Vol. 15, No. 1, pp.3–12.

Bonidia, R.P., Rodrigues, L.A., Avila-Santos, A.P., Sanches, D.S. and Brancher, J.D. (2018) 'Computational intelligence in sports: a systematic literature review', *Advances in Human-Computer Interaction*, Vol. 2018, No. 1, p.3426178.

Bourdon, P.C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M.C., Gabbett, T.J., Coutts, A.J., Burgess, D.J. and Gregson, W. (2017) 'Monitoring athlete training loads: consensus statement', *International Journal of Sports Physiology and Performance*, Vol. 12, No. s2, pp.161–170.

Buchheit, M. and Simpson, B.M. (2017) 'Player-tracking technology: half-full or half-empty glass?', *International Journal of Sports Physiology and Performance*, Vol. 12, No. s2, pp.S2-35–S2-41.

Claudino, J.G., Capanema, D.d.O., de Souza, T.V., Serrão, J.C., Machado Pereira, A.C. and Nassis, G.P. (2019) 'Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review', *Sports Medicine-Open*, Vol. 5, No. 1, p.28.

Clifton, J. and Laber, E. (2020) 'Q-learning: theory and applications', *Annual Review of Statistics and Its Application*, Vol. 7, No. 1, pp.279–301.

Cossich, V.R., Carlgren, D., Holash, R.J. and Katz, L. (2023) 'Technological breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality, and modern data visualization in performance analysis', *Applied Sciences*, Vol. 13, No. 23, p.12965.

Deci, E.L. and Ryan, R.M. (2000) 'The' what' and' why' of goal pursuits: human needs and the self-determination of behavior', *Psychological Inquiry*, Vol. 11, No. 4, pp.227–268.

Gabbett, T.J. (2016) 'The training – injury prevention paradox: should athletes be training smarter and harder?', *British Journal of Sports Medicine*, Vol. 50, No. 5, pp.273–280.

Ghosh, I., Ramasamy Ramamurthy, S., Chakma, A. and Roy, N. (2023) 'Sports analytics review: artificial intelligence applications, emerging technologies, and algorithmic perspective', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 13, No. 5, p.e1496.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F. and Celi, L.A. (2019) 'Guidelines for reinforcement learning in healthcare', *Nature Medicine*, Vol. 25, No. 1, pp.16–18.

Halson, S.L. (2014) 'Monitoring training load to understand fatigue in athletes', *Sports Medicine*, Vol. 44, No. 2, pp.139–147.

Hulin, B.T., Gabbett, T.J., Lawson, D.W., Caputi, P. and Sampson, J.A. (2016) 'The acute: chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players', *British Journal of Sports Medicine*, Vol. 50, No. 4, pp.231–236.

Lames, M. and McGarry, T. (2007) 'On the search for reliable performance indicators in game sports', *International Journal of Performance Analysis in Sport*, Vol. 7, No. 1, pp.62–79.

Li, H. and He, H. (2023) 'Multiagent trust region policy optimization', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 9, pp.12873–12887.

Liu, X., Gao, B., Suleiman, B., You, H., Ma, Z., Liu, Y. and Anaissi, A. (2023) 'Privacy-preserving personalized fitness recommender system P3FitRec: a multi-level deep learning approach', *ACM Transactions on Knowledge Discovery from Data*, Vol. 17, No. 6, pp.1–24.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K. and Ostrovski, G. (2015) 'Human-level control through deep reinforcement learning', *Nature*, Vol. 518, No. 7540, pp.529–533.

Moesch, K., Kenttä, G., Kleinert, J., Quignon-Fleuret, C., Cecil, S. and Bertollo, M. (2018) 'FEPSAC position statement: mental health disorders in elite athletes and models of service provision', *Psychology of Sport and Exercise*, Vol. 38, pp.61–71.

Mulani, J., Heda, S., Tumdi, K., Patel, J., Chhinkaniwala, H. and Patel, J. (2019) 'Deep reinforcement learning based personalized health recommendations', *Deep Learning Techniques for Biomedical and Health Informatics*, Vol. 85, pp.231–255.

Ng, A.Y., Harada, D. and Russell, S. (1999) 'Policy invariance under reward transformations: Theory and application to reward shaping', Icml, Vol. 35, No. 3, pp. 278-287.

Rein, R. and Memmert, D. (2016) 'Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science', *SpringerPlus*, Vol. 5, No. 1, p.1410.

Scott, M.T., Scott, T.J. and Kelly, V.G. (2016) 'The validity and reliability of global positioning systems in team sport: a brief review', *The Journal of Strength and Conditioning Research*, Vol. 30, No. 5, pp.1470–1490.

Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, MIT Press Cambridge, Vol. 28, No. 3, pp.297–300.

Wang, Y., He, H. and Tan, X. (2020) 'Truly proximal policy optimization', *Uncertainty in Artificial Intelligence*, Vol. 21, No. 125, pp.113–122.

Williams, R.J. (1992) 'Simple statistical gradient-following algorithms for connectionist reinforcement learning', *Machine Learning*, Vol. 8, No. 3, pp.229–256.

Yu, C., Liu, J., Nemati, S. and Yin, G. (2021) 'Reinforcement learning in healthcare: a survey', *ACM Computing Surveys*, Vol. 55, No. 1, pp.1–36.