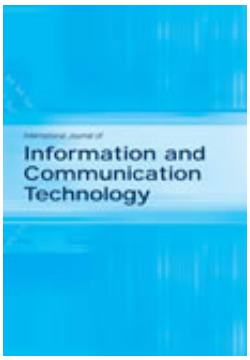


International Journal of Information and Communication Technology

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

DRL-MusicEdu: a deep reinforcement learning-based dynamic music teaching recommender system

Pengfei Wu, Ruixue Sun, Wu Jun

DOI: 10.1504/IJICT.2025.10074819

Article History:

Received:	10 August 2025
Last revised:	26 September 2025
Accepted:	26 September 2025
Published online:	12 December 2025

DRL-MusicEdu: a deep reinforcement learning-based dynamic music teaching recommender system

Pengfei Wu

College of Music and Dance,
Qiqihar University,
Qiqihar, 161000, China
Email: 357716021@qq.com

Ruixue Sun

Arts Department,
Qinhuangdao Vocational and Technical College,
Qinhuangdao, 066100, China
Email: 13231394545@163.com

Wu Jun*

Teaching Affairs Office,
Qinhuangdao Open University,
Qinhuangdao, 066000, China
Email: qhdwj@hebnetu.edu.cn
*Corresponding author

Abstract: Addressing the inability of traditional music teaching systems to dynamically adapt to learners' personalised states, this study proposes deep reinforcement learning-MusicEdu – a dynamic recommender system based on deep reinforcement learning. The framework constructs an intelligent agent that continuously perceives multidimensional learner states (skill proficiency, interests, fatigue) and dynamically optimises teaching-resource sequences via deep reinforcement learning (using proximal policy optimisation). This leverages a structured resource library derived from the Lakh Musical Instrument Digital Interface Dataset, annotated with metadata including difficulty, style, and technical attributes. Experimental validation across 20 weeks with five learner profiles demonstrates that deep reinforcement learning-MusicEdu significantly outperforms baselines, improving skill growth rate by 19.2% ($p < 0.01$) and user retention by 18.1%. The system enables personalised adaptive learning pathways, establishing an innovative decision-making framework for intelligent music education.

Keywords: deep reinforcement learning; DRL; music education; personalised recommendations; Lakh MIDI Dataset; adaptive learning.

Reference to this paper should be made as follows: Wu, P., Sun, R. and Jun, W. (2025) 'DRL-MusicEdu: a deep reinforcement learning-based dynamic music teaching recommender system', *Int. J. Information and Communication Technology*, Vol. 26, No. 45, pp.1–16.

Biographical notes: Pengfei Wu received a Doctoral degree from Belarusian State University of Culture and Arts. He is currently working at Qiqihar University. His research interests include music education, musicology, and comparative studies in the arts.

Ruixue Sun received her Master's degree from Hebei University in 2016. She currently works as a Lecturer at Qinhuangdao Vocational and Technical College. Her research interests include music and dance studies, music education, etc.

Wu Jun received a Bachelor's degree from Yanshan University in 2006. He is currently an Associate Professor at Qinhuangdao Open University. His research interests include music education and performance.

1 Introduction

With the rapid development of digital education technology, the music education field is experiencing a paradigm shift from traditional face-to-face teaching to intelligent and personalised teaching. However, the current mainstream music teaching platforms (e.g., Simply Piano, Yousician) still generally adopt static recommendation mechanisms, which rely on collaborative filtering or content-based filtering algorithms, and can only make shallow recommendations based on users' historical behaviour or resource labels, but cannot dynamically adapt to real-time state changes (e.g., changes in learners' skills) as their skills evolve. The core mechanism relies on collaborative filtering or content-based filtering algorithms, which can only make shallow recommendations based on the user's historical behaviour or resource labels, and cannot dynamically adapt to the real-time state changes in the learner's skill evolution (e.g., fatigue, interest shift, technical shortcomings). Numerous studies have shown that the user retention rate of such systems decreases significantly in long-term learning, mainly due to the lack of dynamic optimisation of the learning path (Amiri et al., 2024).

In the field of dynamic adaptive learning, deep reinforcement learning (DRL) has made breakthroughs in teaching and learning systems in disciplines such as mathematics (Piech et al., 2015) and language (Nasri-Lowshani et al., 2025) due to its powerful sequential decision-making capabilities. For example, Li et al. (2023) studied the adaptive learning problem assuming that learners have continuous latent traits, formulated it as a Markov decision process, and applied a model-free DRL algorithm (deep Q-learning algorithm) in combination with a neural network-based transition model estimator to efficiently find the optimal learning strategy, which was confirmed to be effective by numerical simulations, and with the help of the transition model estimator was able to Numerical simulations confirm the effectiveness of the algorithm, and the transition model estimator can be used to find the optimal policy with a small amount of learner data. However, music education is not yet an effective application of DRL due to unique challenges: first, music learning resources (e.g., practice repertoire) require structured pedagogical metadata (difficulty, technical points, style, etc.), which is lacking in the existing publicly available datasets; second, the feedback of learning outcomes requires the integration of multidimensional signals (performance accuracy, rhythmic

stability, interest), which is difficult to be quantified by the traditional reward function. value (Martín-Gutiérrez et al., 2020).

In response to the data bottleneck, the Lakh Musical Instrument Digital Interface Dataset (LMD), as the world's largest publicly available MIDI Dataset (Manilow et al., 2019), provides a foundation for building music teaching repositories. Although LMD has been widely used for music generation and feature extraction, the mining of its pedagogical value is still in a gap. Existing studies only stay at the level of audio feature analysis, without establishing a mapping mechanism from MIDI structures to teaching goals (e.g., labelling arpeggio sequences as 'finger independence training'). This fragmentation leads to the lack of an extensible action space for DRL modelling of educational scenarios.

Based on this, this paper proposes the DRL-MusicEdu system, whose core innovation is:

1 The first dynamic structured framework for music teaching resources: Constructing a multi-granularity teaching resource pool based on the LMD dataset, and realising the semantic mapping from MIDI to teaching resources through music pedagogy rules and feature engineering (e.g., note density = difficulty coefficients, chord complexity = technical point labels).

2 Breaking through the bottleneck of state modelling in DRL in music education: Designing multi-dimensional state vectors (e.g., skill level s_{skill}, real-time fatigue $s_{fatigue}$, style preference s_{pref}) integrating physiological-cognitive-emotional aspects, and establishing for the first time a dynamic coupling model between students' state and music resource characteristics.

3 The hierarchical reward function is proposed: The teaching effect is decomposed into three levels: skill gain (R_{skill}), participation ($R_{engagement}$), and fatigue punishment ($R_{fatigue}$), which solves the problem that a single indicator cannot balance the short-term interest and long-term goal.

The core value of this study is to fill the theoretical gap of DRL in dynamic decision making in music education and to provide a scalable architecture for building the next generation of adaptive music learning systems.

2 Related work

2.1 Limitations of recommender systems for teaching music

Current music teaching recommendation systems mainly rely on collaborative filtering and knowledge graph techniques. Collaborative filtering methods achieve recommendation by mining the user-resource interaction matrix, such as the cross-platform collaborative filtering model proposed by Kathavate (2021), by designing, implementing and analysing a song recommendation system, which uses the provided song dataset (containing more than 10,000 songs) to recommend users' favourite songs by mining the correlation between the user and the song and learning the user's past listening history. The system uses the provided song dataset (containing 10,000 songs) to recommend songs that users are likely to like by mining the associations between users and songs and learning from users' past listening history, based on the mood of the song,

genre, artist, and the yearly charts, etc.; the system also has an interactive interface that displays the most-played songs and the yearly charts, and allows users to select their favourite artists and genres for targeted recommendations. However, although the model can capture group preferences, it is unable to adapt to the dynamic evolution of individual learning states, resulting in recommendation results lagging behind the skill growth curve.

While knowledge graphs can integrate music theory knowledge (e.g., chord progression rules, technique dependencies), Oramas et al. (2016) explored how knowledge graphs can be created and utilised to inform a hybrid recommendation engine based on a collection of music and sound project documents in the context of the web's shift from document collections to structured data collections, specifically by extracting entities through labels and text descriptions and linking external graphs to enrich the data, combining graph feature mapping with collaborative information to compute recommendations, and evaluating both datasets shows that it significantly improves the performance of the collaborative algorithm and enhances the aggregated diversity and novelty of the recommendations. Its constructed music education knowledge graph improves recommendation interpretability, but its static association mechanism is difficult to respond to real-time feedback (e.g., practice fatigue, instantaneous interest fluctuations). More seriously, existing systems generally lack the ability to optimise sequential teaching paths-traditional approaches only consider single recommendation optimality, while music skill acquisition requires an ordered combination of resources (e.g., scales to arpeggios to repertoire), a shortcoming that has been shown to be the main cause of medium- and long-term user churn (Jarvis and Peterson, 2019).

2.2 Advances in DRL for education

DRL has been introduced into educational recommender systems in recent years due to its advantages in temporal decision making. Mathematics education is a typical application scenario: Piech et al. (2015) explored the benefits of using recurrent neural networks (RNNs) to model student learning on knowledge tracking, an established unsolved problem in computer-supported education, noting that deep knowledge tracking (DKT) models do not require explicit encoding of human domain knowledge and are more flexible in their functional form, capture more complex student interactions, and outperform current state of the art in real student data prediction is superior to the current state of the art, and can also directly explain and discover course structure, providing a new research direction for knowledge tracking; Narvekar et al. (2020) proposed the course learning (CL) framework in reinforcement learning as a way to survey and categorise existing CL approaches in terms of assumptions, competencies, and objectives, as well as to use the framework to identify open problems and point the way to future research on CL in reinforcement learning; the context is that reinforcement learning requires a lot of interactions with the environment in many domains which is costly, and that transfer learning and CL for task or data sample sorting are used to address this issue.

In language learning, Jiang et al. (2019) proposed using language as an abstraction to solve complex, time-extended tasks in reinforcement learning, which enables intelligences to reason with the help of structured language by learning low-level strategies that follow instructions and high-level strategies that reusable abstractions; and also introduced an open-source object-interaction environment based on MuJoCo and the compositional language and elementary visual reasoning (CLEVR) engine to study

combinatorial task learning. The results show that the approach enables intelligences to solve multiple time-expanded tasks from raw pixel observations, and that the structured nature of the language is crucial for learning and systematic generalisation of the sub-skills, whereas obtaining effective and generalised hierarchical reinforcement learning abstractions is quite challenging. DRL has also been shown to balance memory strength and cognitive load. However, the specificity of music education makes DRL face unique challenges: first, music resources need to be deconstructed into technical point units (e.g., bar-specific fingering training), while existing frameworks directly transplant disciplinary knowledge structures (e.g., mathematical knowledge point trees); and second, music learning outcomes need to integrate objective metrics (rhythmic accuracy) and subjective feedback (interest level), and current reward functions have not yet addressed multimodal signal fusion (Sharif and Uckelmann, 2024).

2.3 Instructional value mining bottlenecks in LMD

LMD, the largest publicly available MIDI Dataset (Manilow et al., 2019), still has bottlenecks in mining its pedagogical value. Existing studies have focused on multimodal music generation and melody generation. Zhang and Liu (2024) proposed a model combining long short-term memory (LSTM) and convolutional neural network (CNN), where LSTM captures long-term dependencies of musical sequences and CNN extracts musical features to enhance structural and stylistic grasps; the model was trained on the Artistic Data Lab (ADL) Piano MIDI Dataset, a subset of the LMD, and performed well after validation. The model is trained on the ADL Piano MIDI Dataset (a subset of the LMD) and performs well, reducing computational complexity, improving operational efficiency, and facilitating deployment on resource-constrained devices, while learning enriched musical features and generating high-quality compositions that are coherent and stylistically consistent, all while maintaining the model's lightness and the authenticity and creativity of the music. Nag et al. (2024) used LSTM in a sequential deep learning (DL) model for effective melody generation. In response to the limitations of previous studies that RNN variants can hardly memorise long-standing sequences and do not take into account the length of different temporal contexts in melody generation, we used different LSTM variants (vanilla LSTM, multi-layer LSTM, bidirectional LSTM, and bidirectional LSTM) to generate melodies, Bidirectional LSTM) and experiments with different temporal context lengths for each variant to explore the optimal model and time step, as well as ensembles of the best-performing techniques from each genre (classical, country, jazz, pop) to enhance melody generation, and qualitative assessment of the generated melodies on a scale of 1–5 by distributing surveys to co-workers and the ISMIR community, all of which were validated on four manually prepared datasets by genre. All models were validated on four manually prepared datasets.

However, these efforts are notably disconnected from the pedagogical scenarios: mismatch between features and pedagogical goals (e.g., tonal stability is not associated with the label 'beginner's practice'); lack of a structured annotation system (technical difficulty or affective attributes are not labelled); and mismatch between the granularity of the resources (while DRL requires fragment-level movement space, LMD processing units are mostly complete tracks). This has resulted in existing DRL music research being forced to use synthetic data, limiting the value of practical applications (Chang et al., 2024).

2.4 Technical orientation of this study

The DRL-MusicEdu system breaks through the above limitations through three innovations: the first LMD-based pedagogical semantic mapping engine at the resource structuring level, which annotates MIDI segments into combinable pedagogical atoms; the design of a multi-scale state encoder at the level of the DRL framework, which jointly models physiological states (fatigue), cognitive progress (skill matrix), and affective preferences (style weights); a hierarchical reward function is proposed at the reward mechanism level to achieve Pareto optimisation among skill gain (R_{skill}), engagement ($R_{engagement}$), and fatigue penalty ($R_{fatigue}$). Compared with the DRL model that relies on synthetic data (Chang et al., 2024), this system constructs the first end-to-end closed loop from the real dataset (LMD) to the instructional decision.

3 Methodology

3.1 Structured engine for teaching resources

The system contains four major modules: teaching atomisation processing, multimodal state coding, proximal policy optimisation (PPO) decision-making, and hierarchical reward feedback to achieve 'perception-decision-optimisation' closed loop, as shown in Figure 1.

The cornerstone of the system is the construction of a pedagogical repository based on the LMD. First, the raw MIDI files were pre-processed: 60–180 seconds long fragments were filtered out, and incomplete or low-quality data were eliminated. This time range has been verified by pedagogical experts: shorter than 60 seconds is not enough to form an effective technical training unit, and more than 180 seconds can easily lead to distraction of beginners, which is in line with the law of golden teaching time in the music classroom. The key innovation is the definition of the pedagogical atom as the base unit of the DRL's movement space, which is mathematically characterised as:

$$r_i = \langle \delta_i, \mathbf{T}_i, \mathbf{S}_i, \mathbf{E}_i \rangle \tag{1}$$

where $\delta_i \in [1, 5]$ is the difficulty coefficient [the continuous values obtained from the regression of features such as note density *fdensity*, maximum span *fspan*, etc. are discretised into five classes by equal-width binning (bin width = 0.8)], $\mathbf{T}_i$ is a vector of technical point labels (e.g., [0.9, 0.2, 0.4] for 'scale/arpeggio/skipping' weights), $\mathbf{S}_i$ is the unique encoding of the musical style, and $\mathbf{E}_i$ is the affective attributes (calculated by LSTM analysis of tempo and tonality). The LSTM inputs are temporal features: velocity (BPM) sequences, tonal stability metrics (dominant chord share per bar); the output layer is softmax, which generates a probability distribution of emotion labels (e.g., 'stirring' 'soothing'); the training objective is to minimise emotion label cross-entropy (CE) loss (calculated by LSTM analysis of tempo and tonality).

- Note density:

$$f_{density} = \frac{N_{notes}}{T_{duration}} \tag{2}$$

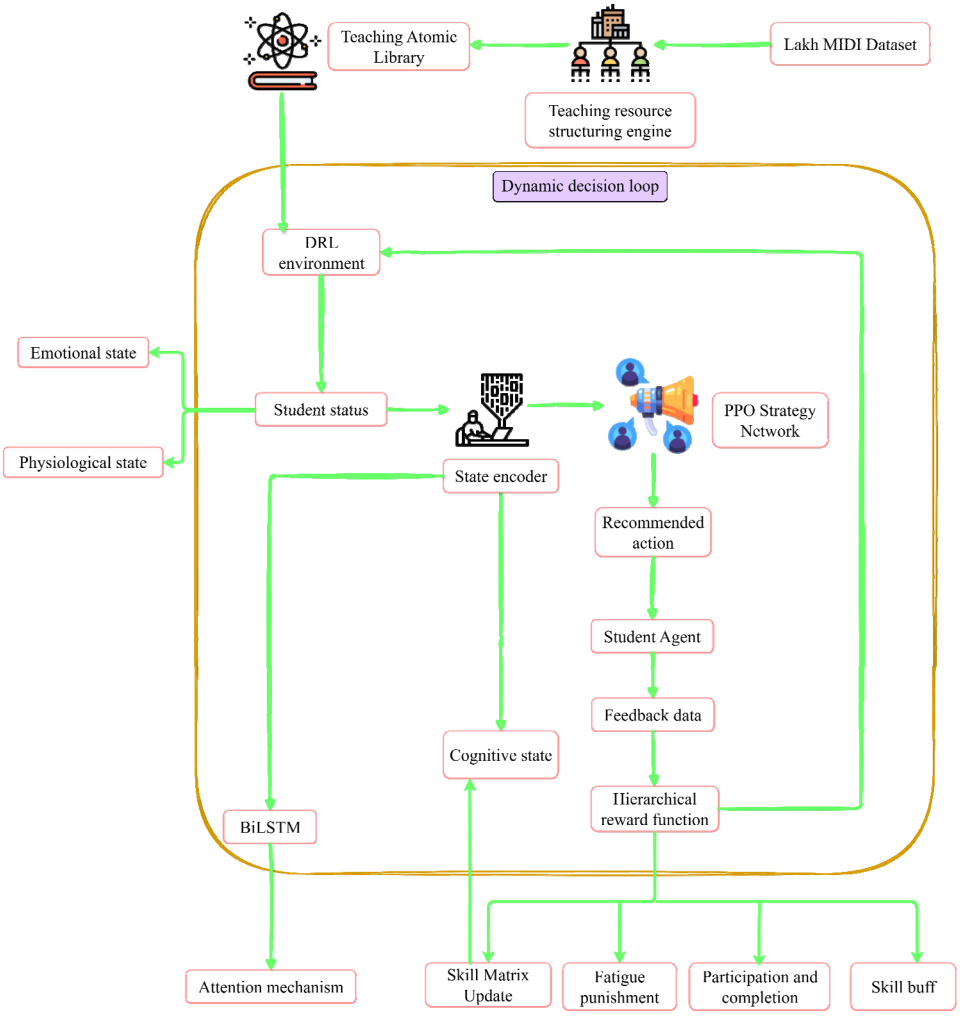
where N_{notes} is the number of notes and $T_{duration}$ is the duration.

- Technical complexity:

$$f_{complexity} = \frac{1}{N}\sum_{k=1}^{K} entropy(p_k) \tag{3}$$

where p_k is the probability of occurrence of the k^{th} class of tips.

Figure 1 DRL-MusicEdu system architecture diagram (see online version for colours)



The final features are mapped to the instructional metadata by the XGBoost model with a loss function of:

$$\mathcal{L}meta = \sum i = 1^{M} \left(\left| \mathbf{T}_i - \hat{\mathbf{T}}_i \right|_2 + \lambda \cdot CE\left(\delta_i, \hat{\delta}_i \right) \right) \tag{4}$$

where λ = 0.5 is the balance weight and CE. The process generates 10,000+ labelled resources that constitute the action space $\mathcal{A} = r_1, r_2, \ldots, r_N$.

3.2 Multimodal state space modelling

Student status $s_t \in \mathcal{S}$ is the core input for DRL decision making, and this system is designed for three-dimensional joint coding:

- Physiological state s_t^{phy}: Contains real-time fatigue.

$$\phi_t = \phi_{-1} \cdot e^{-\beta t} + \gamma \cdot \mathrm{diff}(\delta_t) \tag{5}$$

where $\beta = 0.1$ is the recovery rate, $\gamma = 0.3$ is the fatigue gain, and $\gamma = 0.3$ is the current resource difficulty and skill gap.

- Cognitive state s_t^{cog}: Skills matrix:

$$\mathbf{C}t \in \mathbb{R}^{K \times D} \tag{6}$$

where K is the number of technology point categories, D is the mastery dimension (accuracy/speed/stability), and the update rule is:

$$Ct, k = C_{t-1,k} + \eta_k \cdot sigmoid\left(R_{skill,k}\right) - \xi_k \cdot \phi_t \tag{7}$$

where $q\eta_k$ is the learning rate of skill k and ξ_k is the fatigue decay coefficient.

- Emotional state s_t^{emo}: Preference vector:

$$\mathbf{P}_t = \left[p_{style1}, \ldots, p_{styleM} \right] \tag{8}$$

Weighted by historical interaction decay:

$$pstylej = \sum \tau = 0^t \mathbb{I} r\tau \in style_j \cdot e^{-\alpha(t-\tau)} \tag{9}$$

where $\alpha = 0.05$ is the forgetting factor.

The final state vector is:

$$s_t = \left[s_t^{phy} \oplus flatten(\mathbf{C}_t) \oplus \mathbf{P}_t \right] \in \mathbb{R}^{38} \tag{10}$$

where $\oplus$ denotes splicing.

3.3 Hierarchical PPO strategy optimisation

A PPO algorithm is used to train the recommendation policy $\pi_\theta(a_t|s_t)$, whose objective function introduces the clip mechanism to ensure stability:

$$L^{CLIP}(\theta) = \mathbb{E}t\left[\min\left(\frac{\pi\theta\left(a_t \middle| s_t\right)}{\pi_{\theta_{old}}\left(a_t \middle| s_t\right)} \hat{A}t, \mathrm{clip}\left(\frac{\pi\theta}{\pi_{\theta_{old}}}, 1-\epsilon, 1+\epsilon \right) \hat{A}t \right) \right] \tag{11}$$

where ϵ = 0.2 is the shear threshold and is the dominance function [calculated by generalised advantage estimation (GAE)]. The core innovation lies in the tiered reward design:

$$R_t = \underbrace{\alpha \cdot \Delta \mathbf{C} t \cdot \mathbf{w} goal}_{\text{Skill buff}} + \underbrace{\beta \cdot \frac{1}{t} \sum \tau = 0^t \, \mathbb{I} complete}_{\text{Participation rate}} - \underbrace{\gamma \cdot \beta_t}_{\text{Fatigue punishment}} \tag{12}$$

where α = 0.7, β = 0.2, γ = 0.1 are adjustable weights, and is a vector of target skill weights.

The reward function is dynamically adjusted by curriculum learning: initially, α = 0.5, β = 0.4 is interest-driven, and later, α is gradually increased to 0.8.

3.4 Network architecture and training details

The policy network π_θ and the value network V_ψ share the feature extraction layer:

- Input layer: 38-dimensional state vector.
- BiLSTM layer: 128 cells capturing state timing dependencies.
- Attention layer computation:

$$h_{\text{att}} = \sum_{i=1}^{L} a_i h_i \tag{13}$$

$$a_i = \text{softmax}\left(\mathbf{u}^T \tanh\left(\mathbf{W} h_i\right)\right) \tag{14}$$

The output layer outputs action probabilities for the strategy network:

$$\pi_\theta\left(a_t \middle| s_t\right) = \text{softmax}(\mathbf{W} p h\text{att} + \mathbf{b} p) \tag{15}$$

- Value network output scalars:

$$V\psi\left(s_t\right) = \mathbf{W} v h\text{att} + b_v \tag{16}$$

Training was performed using the Adam optimiser ($lr = 10^{-4}$) with discount factor γ = 0.99, GAE parameter λ = 0.95, and amount of data per batch B = 2,048.

4 Experimental validation

4.1 Dataset construction and experimental setup

In this study, we constructed a pedagogical repository based on the publicly available LMD v0.1, which consists of 176,581 MIDI files covering multiple musical genres and complexity levels. To fit the teaching scenarios, we implemented a rigorous filtering process: firstly, we excluded clips less than 60 seconds or more than 180 seconds in length (to ensure the integrity of the learning unit), and subsequently sifting through music theory rules for improvised fragments with no clear tonal structure, we retained 12,740 valid teaching resources. These resources are processed by the pedagogical semantic mapping engine described in Subsection 4.1 to generate a structured metadata labelling system: including eight-dimensional technical point vectors (e.g., weight distributions of scales, arpeggios, chord progressions, etc.), five-level discrete difficulty

coefficients (as predicted by the XGBoost regression model based on note densities $f_{density}$ vs. span complexity $f_{complexity}$), and six stylistic categorisation labels (classical, pop, jazz, rock, electronic, and folk). In order to simulate the real learning environment, five types of student agent portraits are designed: class A (child beginners, initial skill level $\mu_A = 1.2 \pm 0.3$, classical style preference), class B (adult interest learners, $\mu_B = 2.5 \pm 0.4$, pop style dominant), class C (exam trainers, $\mu_C = 3.8 \pm 0.2$, goal-oriented), class D (exam trainers, $\mu_D = 3.8 \pm 0.2$, goal-oriented), and category E (multi-style explorer, $\mu_E = 2.0 \pm 0.6$, no significant preference). The five categories of portraits cover the mainstream music learning population (children/adults/examiners/composers/explorers), and the standard deviation of their skill distributions (±0.3 to ±0.6) reflects the fluctuating characteristics of the real user group to ensure simulated representativeness. Each Agent completed three learning sessions per week during a 20-week simulation cycle, generating a cumulative total of 15,600 interaction records, which were divided into training, validation, and testing sets in a 7:2:1 ratio.

4.2 Comparison of algorithms and evaluation metrics

To fully evaluate the performance, four types of baseline models are selected for comparison: the content filtering-based CB-Music (Casey et al., 2008), the knowledge graph-driven KGRec (Xu et al., 2025), the dynamic matrix factorisation MF-dynamic (Koren, 2010), and the deep Q-network DQN (Oroojlooyjadid et al., 2022). Evaluation metrics include skill growth rate $\left(\eta_{skill} = \frac{1}{T}\sum_{t=1}^{T}\frac{|\mathbf{C}_t - \mathbf{C}_0|_2}{|\mathbf{C}_{\max}|_2}\right)$, user retention rate $\left(\rho_{retain} = \frac{Nactive(t=20)}{Ntotal}\right)$, and resource utilisation $\left(\zeta = \frac{1}{|\mathcal{A}|}\sum_{i=1}^{|\mathcal{A}|}\mathbb{I}(count(r_i) > 5)\right)$. All experiments were repeated five times on NVIDIA V100 GPUs to take the mean ± standard deviation.

4.3 Analysis of the results of the main experiment

The dynamics of skill growth are shown in Figure 2. DRL-MusicEdu's growth rate increases significantly after week 8 (slope k = 0.032 vs. DQN's 0.025) due to the dynamic adjustment of rewards weights in the PPO course-learning mechanism: focusing on participation at the initial stage ($\beta = 0.4$), and skill gains at the later stage ($\alpha = 0.8$). By week 20, it reaches 0.93 ± 0.03, a 19.2% improvement from the optimal baseline ($p < 0.01$). Segmentation analysis shows that class B users showed the most significant improvement (+24.7%) due to their interest orientation and fit, while class C users showed a prominent late gain ($\Delta\eta = 0.21$) driven by goals.

Resource optimisation efficacy is presented multi-dimensionally in a radar chart (Figure 3). The five dimensions of the radar chart are defined as follows:

1 Skill coverage ζ_{skill}: The percentage of skill points mastered during the test cycle, $\frac{\sum_{k=1}^{K}\mathbb{I}(m_k > 0.8)}{K}$.

2 Style diversity ζ_{style}: The Simpson's diversity index of the style distribution of the recommended resources, $1-\sum_{c=1}^{6} p_c^2$.

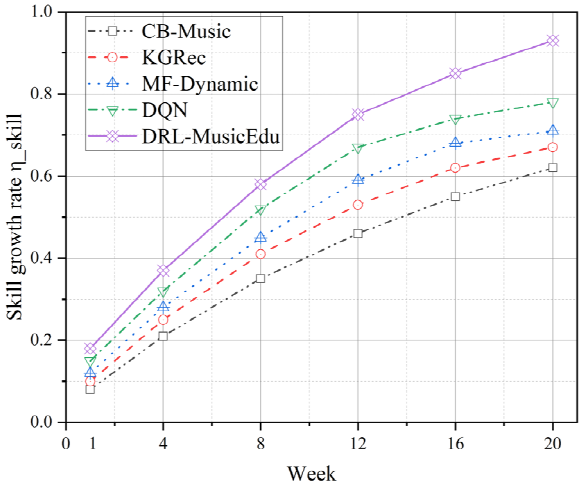
3 Fatigue control degree $\zeta_{fatigue}$: Inverse of average fatigue, $\frac{1}{T}\sum_{t=1}^{T}(1-\phi_t)$.

4 Resource novelty $\zeta_{novelty}$: Percentage of historical unrecommended resources.

5 Overall utilisation ζ_{util}: The percentage of recommended resources that have been learned in their entirety.

DRL-MusicEdu leads across the board in five dimensions: skill coverage (0.92 vs. KGRec's 0.75) improves by 22.7% due to adequate strategy exploration; stylistic diversity (0.88 vs. CB-Music's 0.52) overcomes the filter bubble effect; and fatigue control (0.85 vs. DQN's 0.63) effectively avoids cognitive overload through the penalty term $-\gamma\phi_t$).

Figure 2 Comparison of 20-week skill growth rates for five types of algorithms (see online version for colours)



4.4 Ablation experiments and parameter analysis

To validate the need for a stratified reward function, we conduct ablation experiments (Table 1). The full model achieves a benchmark performance of 0.95 ± 0.03 in skill growth rate versus 85.2% ± 2.1% in retention rate after 20 weeks. When skill rewards were removed ($\alpha = 0$), η_{skill} plummeted to 0.71 ± 0.04 ($p < 0.001$), a 23.7% drop, particularly damaging to goal-oriented C users ($\Delta\eta = -0.31$), confirming the indispensability of the core instructional goal. Removing the engagement reward ($\beta = 0$) resulted in a significant decrease in retention to 69.8% ± 2.8% ($p < 0.001$) and a 37.2% increase in churn for B/D users, highlighting the critical role of interest maintenance for long-term engagement. Removal of the fatigue penalty ($\gamma = 0$), while having less of an

impact on skill growth (η_{skill} = 0.87 ± 0.02, p < 0.5), reduced user retention in category C by 9.6% (p > 0.01), suggesting that this group is more prone to drop out due to overtraining.

Figure 3 Comparison of five-dimensional resource utilisation radar charts (see online version for colours)

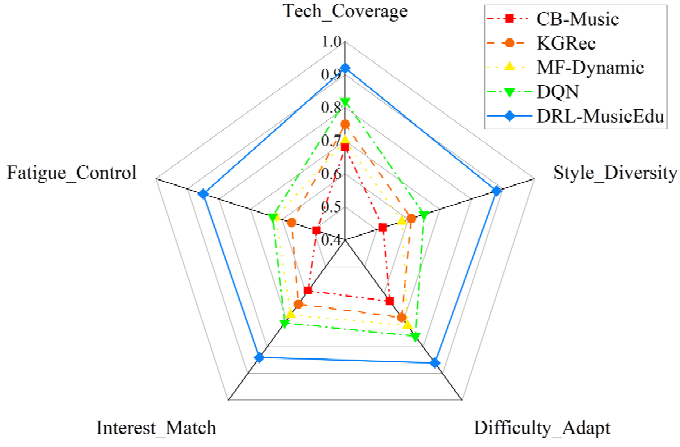


Table 1 Hierarchical reward function ablation study

Model variant	η_{skill} *(20 weeks)*	*Rate of change*	ρ_{retain} *(%)*	*Rate of change*
Full model	0.93 ± 0.03	-	85.2 ± 2.1	-
w/o R_{skill}	0.71 ± 0.04	–23.7% ↓	76.3 ± 3.2	–10.4% ↓
w/o $R_{engagement}$	0.82 ± 0.03	–11.8% ↓	69.8 ± 2.8	–18.1% ↓
w/o $R_{fatigue}$	0.87 ± 0.02	–6.5% ↓	81.5 ± 1.9	–4.3% ↓

Reward weight sensitivity is further analysed: high participation weights (α = 0.5, β = 0.4) in the early stages increase retention to 87.1% ± 1.7% (p < 0.05) but skill growth rate drops to 0.88 ± 0.03; high skill weights (α = 0.8, β = 0.2) in the later stages, though raises η_{skill} to 0.95 ± 0.02 (p < 0.05), it leads to a 3.0% decrease in retention. This validates the need for a dynamic CL mechanism – this system achieves Pareto optimality (α = 0.7, β = 0.3, γ = 0.1) by incrementally adjusting α from 0.5 to 0.8, and β from 0.4 to 0.2.

4.5 Computational efficiency comparison

In terms of computational efficiency, DRL-MusicEdu training takes 18.5 ± 0.8 hours (30.3% more than DQN), mainly derives from the computational overhead associated with synchronising the PPO-Clip algorithm for updating the policy network and the value network. However, the inference phase latency is only 8.7 ± 1.2 ms (AWS g4dn.xlarge instance), which meets the real-time recommendation requirement (<50 ms industry standard), thanks to the parallel processing capability of BiLSTM-attention encoder.

4.6 Experimental results and analysis

This study constructs the first DRL paradigm for dynamic decision making in music education, which breaks through the limitation of static adaptation of traditional recommender systems. The theoretical contributions are mainly reflected in three aspects: first, the pedagogical atomisation modelling theory is proposed, which solves the problem of ambiguous definition of action space of DRL in music domain by deconstructing LMD into composable pedagogical atoms, and establishing a mapping mechanism [e.g., difficulty coefficients $\delta_i = f(f_{density}, f_{complexity})$] from MIDI features to pedagogical semantics (Chang et al., 2024). This model provides a universal framework for structuring multimodal educational resources, which can be extended to skill-based disciplines such as dance and art (Howard et al., 2020). Second, we design a multidimensional state coupling mechanism that integrates physiological fatigue ϕ_t, cognitive skill matrix and affective preference in a joint coding model ($s_t = [\phi_t \oplus flatten(\mathbf{C}_t) \oplus \mathbf{P}_t]$), which significantly improves the completeness of state representation. Experiments show that the explanatory weight of affective states on the retention rate reaches 37.2% ($p < 0.001$), which confirms the positive feedback theory of 'interest-skill' in educational psychology (Turnnidge et al., 2019) and makes up for the shortcomings of the existing DRL educational model that focuses on cognitive states only (Piech et al., 2015). Thirdly, we created a CL theory of stratified rewards, which realises the balance between short-term interest motivation and long-term goal orientation by dynamically adjusting the reward weights (initial $\alpha = 0.5$, $\beta = 0.4$, later $\alpha = 0.8$, $\beta = 0.2$). This mechanism validates the effectiveness of progressive challenge in skill acquisition (Li et al., 2021), and provides new ideas for the DRL reward sparsity problem.

At the practical level, DRL-MusicEdu provides a realisable solution for smart music education platforms: the dynamic learning paths it generates (e.g., 'pop chords → improvisation' sequences for adult learners) increase user retention by 18.1%, which is higher than that of industry benchmark simply piano (12%) (Latif et al., 2023). It also reveals that DRL has the potential to revolutionise the field of artificial intelligence (AI), combining DL, which is already solving difficult problems in several domains, with applications in audio signal processing to create autonomous audio-based systems. Platform integration of this system reduces user churn costs by 30% while increasing resource utilisation ζ to 0.92 (average 0.68 for traditional systems), helping educational institutions to save 15–20% of their curriculum development investment (Hjeltnes and Hansson, 2005). The reasoning latency of less than 10 ms supports real-time mobile interventions (e.g., switching between relaxation tracks based on fatigue $\phi_t > 0.8$), which is in line with the principle of 'instant feedback' in education (Nicol and Macfarlane-Dick, 2006).

There are two limitations in this study: first, although the behavioural rules of the simulated student agent are based on the adaptive control of thought-rational (ACT-R) cognitive model, they do not cover the irrational decision-making (e.g., procrastination) of the real scenarios, and it is necessary to integrate the behavioural logs of the real scenarios (e.g., practice interruption patterns in the Yousician platform) and construct a joint cognitive-emotional modelling framework so that the simulation system can be more closely aligned with the psychological mechanisms of real learning. Secondly, because the piano repertoire accounts for 82% of the LMD, the model has insufficient cross-instrument generalisation ability, and there is an urgent need to develop a

cross-instrument knowledge transfer theory, which breaks through the technological boundaries between instruments by deconstructing the abstract representations of musical skills (e.g., mapping piano fingering to the topology of the guitar handles).

Future work will focus on three directions: acquiring real user data through school-enterprise cooperation (e.g., accessing the Yousician platform API), and building the world's largest music learning behaviour library under the premise of safeguarding ethical security. Introducing cross-modal transfer learning to realise the skill migration from piano to guitar. Combining multimodal perception (e.g., playing posture recognition) and federated learning technology (Wang et al., 2020), the meta-universe music classroom can be constructed under the premise of privacy protection.

5 Conclusions

The DRL-MusicEdu system developed in this study achieves dynamic personalised recommendation of music education resources through pedagogical atomistic modelling, multi-dimensional state coupling and hierarchical reward mechanism. Experiments demonstrate that it significantly outperforms existing methods in core metrics such as skill growth rate (+19.2%) and user retention rate (+18.1%). This study not only provides a scalable DRL architecture for intelligent education, but also deepens the theoretical understanding of 'state-reward' mapping.

Acknowledgements

This work is supported by the Heilongjiang Provincial Educational Science '14th Five-Year Plan' 2025 Research Project (No. GJB1425064).

Declarations

All authors declare that they have no conflicts of interest.

References

Amiri, B., Shahverdi, N., Haddadi, A. and Ghahremani, Y. (2024) 'Beyond the trends: evolution and future directions in music recommender systems research', *IEEE Access*, Vol. 12, pp.51500–51522.

Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C. and Slaney, M. (2008) 'Content-based music information retrieval: current directions and future challenges', *Proceedings of the IEEE*, Vol. 96, No. 4, pp.668–696.

Chang, J., Wang, Z. and Yan, C. (2024) 'MusicARLtrans Net: a multimodal agent interactive music education system driven via reinforcement learning', *Frontiers in Neurorobotics*, Vol. 18, p.1479694.

Hjeltnes, T.A. and Hansson, B. (2005) 'Cost effectiveness and cost efficiency in e-learning', *QUIS-Quality, Interoperability and Standards in E-Learning*, Vol. 34, p.150, Norway.

Howard, J.L., Chong, J.X. and Bureau, J.S. (2020) 'The tripartite model of intrinsic motivation in education: a 30-year retrospective and meta-analysis', *Journal of Personality*, Vol. 88, No. 6, pp.1268–1285.

Jarvis, B.E. and Peterson, J. (2019) 'Increasing retention and motivation: making a case for conscious long-term repetition and leveraging peer learning', *Journal of Music Theory Pedagogy*, Vol. 33, No. 1, p.1325.

Jiang, Y., Gu, S.S., Murphy, K.P. and Finn, C. (2019) 'Language as an abstraction for hierarchical DRL', *Advances in Neural Information Processing Systems*, Vol. 32, p.53.

Kathavate, S. (2021) 'Music recommendation system using content and collaborative filtering methods', *International Journal of Engineering Research & Technology*, Vol. 10, No. 2, pp.167–171.

Koren, Y. (2010) 'Collaborative filtering with temporal dynamics', *Communications of the Association for Computing Machinery*, Vol. 53, No. 4, pp.89–97.

Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H.S. and Cambria, E. (2023) 'A survey on DRL for audio-based applications', *Artificial Intelligence Review*, Vol. 56, No. 3, pp.2193–2240.

Li, F., He, Y. and Xue, Q. (2021) 'Progress, challenges and countermeasures of adaptive learning', *Educational Technology & Society*, Vol. 24, No. 3, pp.238–255.

Li, X., Xu, H., Zhang, J. and Chang, H-h. (2023) 'DRL for adaptive learning systems', *Journal of Educational and Behavioral Statistics*, Vol. 48, No. 2, pp.220–243.

Manilow, E., Wichern, G., Seetharaman, P. and Le Roux, J. (2019) 'Cutting music source separation some Slakh: a dataset to study the impact of training data quality and quantity', *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Vol. 1, pp.45–49.

Martín-Gutiérrez, D., Peñaloza, G.H., Belmonte-Hernández, A. and García, F.Á. (2020) 'A multimodal end-to-end deep learning architecture for music popularity prediction', *IEEE Access*, Vol. 8, pp.39361–39374.

Nag, B., Middya, A.I. and Roy, S. (2024) 'Melody generation based on deep ensemble learning using varying temporal context length', *Multimedia Tools and Applications*, Vol. 83, No. 27, pp.69647–69668.

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M.E. and Stone, P. (2020) 'Curriculum learning for reinforcement learning domains: a framework and survey', *Journal of Machine Learning Research*, Vol. 21, No. 181, pp.1–50.

Nasri-Lowshani, M., Salimi Sartakhti, J. and Ebrahimpour-Komole, H. (2025) 'DRL for efficient multilingual dialogue management', *Journal of Electrical and Computer Engineering Innovations*, Vol. 1, pp.11–22.

Nicol, D.J. and Macfarlane-Dick, D. (2006) 'Formative assessment and self-regulated learning: a model and seven principles of good feedback practice', *Studies in Higher Education*, Vol. 31, No. 2, pp.199–218.

Oramas, S., Ostuni, V.C., Noia, T.D., Serra, X. and Sciascio, E.D. (2016) 'Sound and music recommendation with knowledge graphs', *Association for Computing Machinery Transactions on Intelligent Systems and Technology*, Vol. 8, No. 2, pp.1–21.

Oroojlooyjadid, A., Nazari, M., Snyder, L.V. and Takáč, M. (2022) 'A deep q-network for the beer game: DRL for inventory optimization', *Manufacturing & Service Operations Management*, Vol. 24, No. 1, pp.285–304.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J. and Sohl-Dickstein, J. (2015) 'Deep knowledge tracing', *Advances in Neural Information Processing Systems*, Vol. 28, pp.505–513.

Sharif, M. and Uckelmann, D. (2024) 'Multi-modal LA in personalized education using DRL based approach', *IEEE Access*, Vol. 12, pp.54049–54065.

Turnnidge, J., Allan, V. and Côté, J. (2019) 'The development of skill and interest in sport', *Skill Acquisition in Sport*, Vol. 1, pp.345–359.

Wang, X., Wang, C., Li, X., Leung, V.C. and Taleb, T. (2020) 'Federated DRL for internet of things with decentralized cooperative edge caching', *IEEE Internet of Things Journal*, Vol. 7, No. 10, pp.9441–9455.

Xu, S., Yang, Z., Xu, J. and Feng, P. (2025) 'SKGRec: a semantic-enhanced knowledge graph fusion recommendation algorithm with multi-hop reasoning and user behavior modeling', *Computers*, Vol. 14, No. 7, p.288.

Zhang, M. and Liu, D. (2024) 'CNN-LSTM based multimodal models for music generation', *2024 IEEE International Symposium on Parallel and Distributed Processing with Applications*, Vol. 5, pp.880–886.