Design And Empirical Study Of A Personalized Training Path Optimization Algorithm Based On Artificial Intelligence Reinforcement Learning

Yi Zhexue[1], Lin Ze[2], and Chen Feng[2*]

[1] Continuing Education College, Zhejiang College of Security Technology; Wenzhou, Zhejiang Province 325000, China
[2] College of AI, Zhejiang College of Security Technology; Wenzhou, Zhejiang Province 325000, China
[*] Corresponding author. E-mail: 19066196@zjcst.edu.cn

Received: Apr. 13, 2026; Accepted: Jun. 12, 2026

Intelligent personalized learning systems must continuously adapt instructional strategies based on students' progress and engagement levels. However, designing adaptive curricula remains challenging, as traditional supervised learning and standard reinforcement learning approaches often fail to effectively integrate mastery assessment with long-term policy optimization. This study proposes a computational framework for personalized training path optimization that integrates mastery modelling with deep reinforcement learning to improve algorithmic efficiency, adaptive policy convergence, and long-term educational decision-making under dynamic learner-state transitions. The optimization problem is formulated as a Markov Decision Process and optimized using a Deep Q-Network (DQN) with experience replay and target-network stabilization to enhance computational efficiency, policy stability, and convergence performance across varying experimental learning conditions. ($\alpha = 0.01, \gamma = 0.95$, batch size = 64, replay buffer = $50,000, \varepsilon$ decayed from 1.0 to 0.01 over 1000 episodes). Experimental results demonstrate that the proposed framework achieves an accuracy of 91.38% and a success rate of 89.56%, outperforming Q-Learning (78.42%) and Policy Gradient (84.67%) methods. The model exhibits stable convergence, with episode rewards increasing from 43.97 to 89.54 and Q-loss decreasing from 15.1682 to 0.2533. These results confirm that integrating probabilistic learner-state modelling with deep reinforcement learning provides a computationally efficient and analytically robust solution for scalable personalized education optimization.

Keywords: Personalized Learning, Deep Reinforcement Learning, Bayesian Knowledge Tracing, Deep Q-Network, Adaptive Education

http://dx.doi.org/10.6180/jase.202611_34.010

1. Introduction

Personalized learning is a central objective in modern educational technology due to diverse learner abilities, engagement patterns, and performance outcomes that traditional curriculum frameworks often fail to address effectively [1]. Intelligent data-driven systems address these limitations through adaptive optimization, where reinforcement learning supports sequential decision-making and BKT-based learner modelling enables probabilistic mastery estimation [2, 3]. Their integration within an MDP framework supports personalized training path generation using engagement, performance, and behavioural indicators for improved learning outcomes [4, 5]. Existing machine learning and reinforcement learning approaches often operate independently, limiting real-time optimization and multi-dimensional learner-state adaptation [6, 7]. The proposed framework utilizes the Teaching Quality and Student Engagement Kaggle dataset to construct multidimensional learner-state representations incorporating performance, engagement, and attendance indicators [8–10]. The proposed framework overcomes these limitations by tightly integrating learner modeling and reinforcement learning within a unified optimization architecture. The framework establishes training path selection as a sequential decision-making problem which operates differently from traditional supervised models to achieve maximum cumulative learning gain. The system uses BKT-inspired aggregated mastery estimation approach-based mastery estimation to generate RL state representation which allows knowledge-aware policy learning to take place. The state space combines engagement level and attendance pattern and performance trends to generate a multi-dimensional learner profile. The reward function uses specific design elements to measure performance improvement and engagement enhancement which leads to balanced optimization objectives. The system uses a Deep Q-Network agent which combines experience replay with target network stabilization to achieve efficient policy learning. The system enables personalized learning through its ability to create adaptive data-driven curriculum sequences which match the needs of individual students. The proposed study presents innovative research by uniting knowledge tracing with engagement modeling and deep reinforcement learning into one optimization framework. The system conducts empirical validation through real-world educational data to show measurable progress. The framework develops intelligent personalized training systems through its assessment methods that measure cognitive mastery and behavioral engagement. The proposed approach delivers an adaptable learning solution which scientific research supports as a practical solution for modern learning environments.

- Design and develop a personalized training path optimization framework which uses reinforcement learning to enhance student learning results and academic achievements and student participation.
- Utilize and analyze the Teaching Quality and Student Engagement dataset from Kaggle to develop multidimensional learner state representations which incorporate performance indicators and student engagement metrics and attendance information.
- Implement and evaluate BKT-inspired aggregated mastery estimation approach within the personalized training framework to assess student mastery development and cognitive state progression.
- Develop and optimize a DQN reinforcement learning model which learns training policies through MDP creation and multi-objective reward optimization.

The framework integrates. Section 1 introduces the research motivation and objectives, Section 2 reviews related work, and Section 3 presents the proposed BKT-DQN methodology. Section 4 describes experimental setup and evaluation metrics, Section 5 discusses results and comparative analysis, while Section 6 concludes with key findings and future research directions.

2. Literature review

Sadaf Saleem et al. [11] conducted research on intelligent e-learning systems which adapt to student needs by providing personalized learning pathways through digital education platforms. Zaharuddin et al. [12] studied educational recommendation systems which use reinforcement learning to enhance student participation within virtual learning platforms. Mahmoud and Sørensen [13] investigated learner modeling techniques for adaptive education using probabilistic knowledge tracing and engagement analytics. Yang and Ogata, [14] conducted a study that primarily focuses on adaptive learning systems aimed at a customized educational pathway generated through system design for feature-rich tutoring programs. Sahu et al. [15] This study provides a comprehensive overview of machine learning, deep learning, and reinforcement learning techniques applied in quantitative finance, highlighting recent advancements. Recent state-of-the-art adaptive learning frameworks demonstrate significant progress in learner modelling and policy optimization; however, methodological divergence persists due to limited integration between probabilistic knowledge tracing and sequential reinforcement-based adaptation. Q-learning and other early reinforcement learning methods use basic state models which do not account for essential behavioral elements that include both engagement and attendance. The existing frameworks need cognitive-behavioral modeling systems which enable them to fully support students throughout their complete academic journey. The restrictions create problems with personalizing education because they generate insufficient learning options which decrease student interest in their studies.

The proposed framework creates a unified knowledge-based reinforcement learning system which develops personalized training pathways to solve these challenges. The system combines BKT-inspired aggregated mastery estimation approach mastery estimation with a Deep Q-Network based deep reinforcement learning agent. The framework uses MDP modeling to convert curriculum sequencing into a process which decides between different options to achieve maximum total reward. The state representation combines mastery level, engagement index, performance trend, and attendance rate to form a comprehensive learner profile.

3. Materials and methods

The proposed methodology establishes a complete framework for personalized training path optimization through reinforcement learning which combines learner modeling with Markov Decision Process formulation and deep Q-learning policy optimization.

The framework starts its initial stage Fig. 1 by gathering data from the Teaching Quality and The experimental design follows a structured computational workflow consisting of data acquisition, preprocessing, learner-state construction, policy optimization, and performance validation

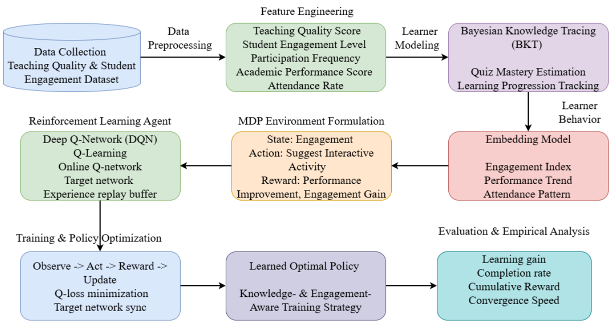

Fig. 1. Proposed Knowledge- and Engagement-Aware Reinforcement Learning Framework for Personalized Training Path Optimization

The learner state knowledge component is created through BKT output as you can see in the Eqn. (1)

$$P\left(L_t \mid O_{1:t}\right) \tag{1}$$

The Fig. 2 shows a dual-pathway system which creates complete learner state information. The BKT-inspired aggregated mastery estimation approach module of BKT analyzes student test results to calculate their learning progress and establish their cognitive mastery levels.

The generated embedding vector functions as a fundamental element for creating the learner state representation. A neural transformation function is used to gain behavior embedding as based of the following Eqn. (2)

$$E_t = f_\phi\left(B_t\right) \tag{2}$$

Linear Embedding creates lower-dimensional continuous vector spaces from high-dimensional input features

Algorithm 1. -Based Cognitive Mastery Update and Learner State Representation

Input: $O = \{O_1, \ldots, O_T\}, P(L_0), P(T), P(G), P(S)$

Output: $P(L_t), S_t$

1. Initialize:

$$P_L \leftarrow P(L_0)$$

2. For $t = 1$ to T:

- Predict:

$$P_L^- = P_L + (1 - P_L)P(T)$$

- Update:
- If $O_t = 1$:

$$P_L = \frac{P_L^-(1 - P(S))}{P_L^-(1 - P(S)) + (1 - P_L^-)P(G)}$$

- Else:

$$P_L = \frac{P_L^- P(S)}{P_L^- P(S) + (1 - P_L^-)(1 - P(G))}$$

- State:

$$S_t = [P_L, E_t, P_t]$$

3. Return P_L, S_t

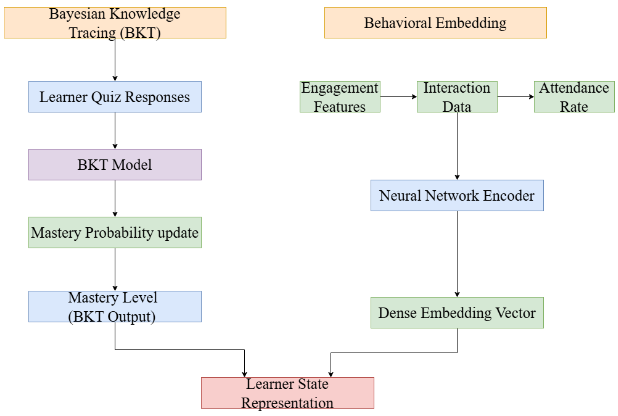

Fig. 2. Learner State Representation via Bayesian Knowledge Tracing and Behavioral Embedding

through its use of linear transformations that include matrix multiplication which maintains essential information as you can see in the Eqn. (3)

$$E_t = \sigma\left(WB_t + b\right) \tag{3}$$

The reinforcement learning agent uses this information to create training plans which combine knowledge assessment with student engagement patterns.

Markov Decision Process (MDP) Modeling

The optimization question on a training path is modeled as a Markov Decision Process defined by the tuple as you can see in the Eqn. (4)

$$M = (S, A, P, R, \gamma) \tag{4}$$

The state space S signifies the overall set of all possible learner profiles that describe the learner's existing knowledge, engagement, and performance status as based of the Eqns. (5-6) Its name being equivalent to the statuses during time step t.

$$S = \{s \mid s = (k, e, p)\} \tag{5}$$

$$s_t = (k_t, e_t, p_t) \tag{6}$$

In this case, the state space can also be written in the form of as show in the Eqn. (7)

$$s_t \in R^n \tag{7}$$

where n is the total number of learner features.

The transition shows how educational modules impact the development of knowledge by students. The reward function measures the immediate improvement in learner performance after taking an action. The Markov property is justified by constructing each learner state as a sufficient multidimensional representation of mastery level, engagement behavior, attendance consistency, and performance trend, ensuring future transitions depend primarily on the current observable state The Fig. 3 depicts a reinforcement learning system which includes an Agent that operates within an Environment. The agent monitors the present state s_t at every time step and chooses an action a_t according to its policy which the neural network controller uses to process input features.

The Q-function estimates the expected cumulative reward of taking action a in state s as based of the following Eqn. (8)

$$Q(s, a; \theta) \tag{8}$$

The equation uses θ to denote the parameters of the neural network model.

The statement establishes a recursive function that calculates the best possible value estimation according to Eqns. (9-10)

$$Q^*(s, a) = E\left[R_t + \gamma \max_{a'} Q^*(s', a')\right] \tag{9}$$

$$y_t = R_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) \tag{10}$$

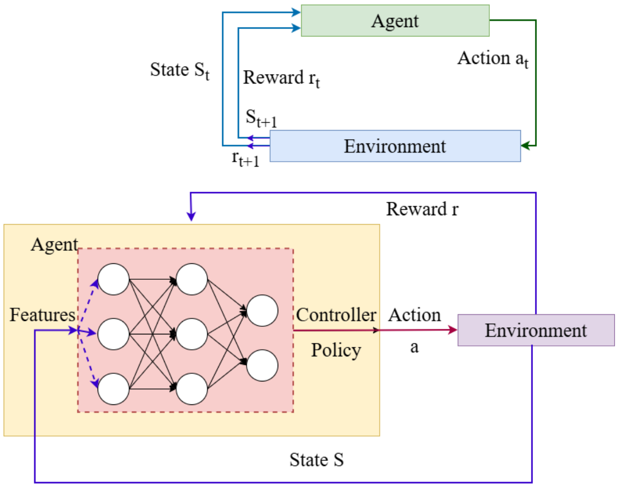

Fig. 3. Agent–Environment Interaction and Policy Network Architecture for MDP-Based Personalized Training Optimization

Dataset Link: https://www.kaggle.com/datasets/programmer3/teaching-quality-and-studentengagement-dataset

Unlike Deep Knowledge Tracing, which offers high representational capacity but requires largescale sequential training data and reduced interpretability, Bayesian Knowledge Tracing provides computationally efficient probabilistic mastery estimation with transparent state representation is given in Table 1.

Convergence Speed (CS)

The DQN model establishes its stable training point through the evaluation of its convergence speed. The learning process achieves stable results through faster convergence which also helps to improve policy development. The training process efficiency and framework reliability both depend on this measurement as you can see in the Eqn. (11)

$$|R_e - R_{e-1}| < \epsilon \tag{11}$$

4. Result and discussion

The developed personalized training path optimization system was built through Python programming language development which includes reinforcement learning and learner modeling systems. Convergence robustness was further evaluated across multiple independent training runs, where variance analysis demonstrated low standard deviation in cumulative reward progression and policy stabilization behavior. The observed consistency confirms that the proposed framework maintains reliable convergence under varying initialization conditions.

Table 1. Hyperparameter Configuration of the Proposed BKT-DQN Personalized Training Path Optimization Framework

Module	Hyperparameter	Value
DQN Agent	Learning Rate / Discount Factor	0.0001-0.001 / 0.90-0.99
	Batch Size / Replay Buffer	32-64 / 10k-50k
	Epsilon (Decay)	1.0 → 0.01 (0.995)
Neural Network	Hidden Layers / Units	2-3 / 64-256
Optimization	Optimizer / Activation	Adam / ReLU
BKT Module	Learning / Guess / Slip Prob.	$0.1-0.3/0.1-0.25/0.05-0.2$
Embedding Layer	Dimension	16-64
Training Setup	Episodes / Max Steps	100-500 / 50-200

Standard Deep Q-Network was selected due to its computational efficiency, stable convergence behavior, and sufficient representational capacity for personalized training path optimization under moderate state-action complexity is shown in Fig. 4.

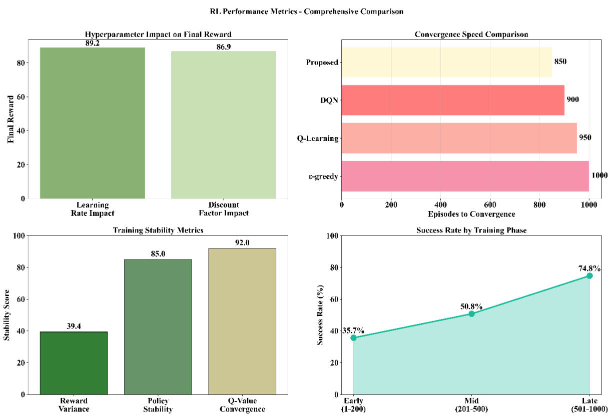

Fig. 4. Comprehensive Reinforcement Learning Performance Metrics

Table 2 learning rate values, 0.1 turns out to gradually make minitone updates to the Q-values so as to avert extreme undulations during policy learning. The discount factor of 0.95 serves as the best option because it allows assessment of both long-term rewards and current learning progress. The system uses its parameters to create balanced convergence which also supports personalized training optimization during extended learning sessions.

Table 2. Optimal Reinforcement Learning Hyperparameters for Training Path Optimization

Optimal Learning Rate	0.1
Optimal Discount Factor	0.95

The suggested DQN model demonstrates its performance under various learning rates (α) The results demonstrate that moderate learning rates ($\alpha = 0.01$ and $\alpha = 0.3$) produce higher final rewards of 84.14 and 89.19 compared to the lower learning rates of 0.01 and 0.05. The learning rate of $\alpha = 0.01$ delivers consistent results with minimal variation yet it produces a restricted final reward outcome is given in Fig. 5.

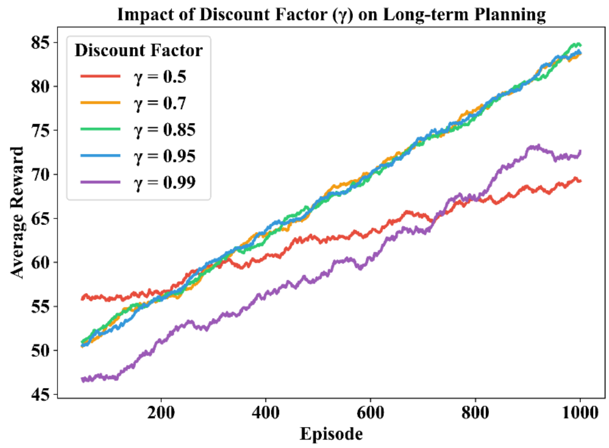

Fig. 5. Impact of Discount Factor (γ) on Long-Term Reward Optimization in the Proposed DQN Model

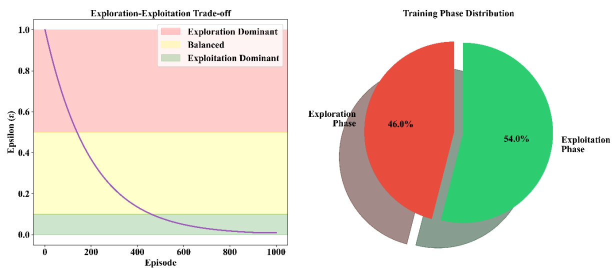

Fig. 6. Exploration–Exploitation Trade-off in Reinforcement Learning

The Fig. 6 demonstrates how reinforcement learning maintains its training process through the two competing methods of exploration and exploitation. The left side contains a line graph that displays how epsilon decreases throughout the episodes, because high epsilon values at first enable players to test all available move possibilities.

The results show that the proposed DQN-based framework outperforms traditional Q-Learning and Policy Gradient methods, achieving 91.38% accuracy and 89.56% success rate, indicating superior decision-making capability is shown in Table 3.

Experimental robustness was validated through repeated training runs, with performance stability confirmed by low variance across cumulative reward progression and convergence consistency. Mean performance values with standard deviation demonstrate reliable policy optimization under varying initialization conditions.

The observed performance patterns indicate that integrating probabilistic learner-state estimation with deep reinforcement learning strengthens theoretical adaptability in sequential educational optimization

5. Conclusion

The study presents an extensive personalized training path optimization framework which uses BKT-inspired aggregated mastery estimation approach and MDP modeling and DQN reinforcement learning algorithm as its three-building blocks. The BKT-inspired aggregated mastery estimation approach system estimated learners' dynamic mastery probabilities through their response patterns which allowed accurate tracking of their cognitive states. The mastery estimates together with engagement and performance metrics were used to create an MDP-based system which defined learner states as multiple dimensions and training module selection as a sequential decision-making process. The DQN agent learned the optimal long-term policy through Bellman value updates and experience replay to maximize cumulative rewards. The experimental results show that the system achieves strong convergence together with stable policy performance during 1000 test episodes. The episode reward improved from 43.97 to 89.54, and the cumulative reward reached 62,632.16, confirming effective long-term optimization. The Q-loss showed major improvement when it dropped to 0.2533 from 15.1682 because the system achieved stable value function learning. The success rate increased from 25.50% to 89.56% with an overall accuracy of 91.38% where the system achieved an average reward of 67.63 and the reward variance stayed below 60.60. The findings show that BKT-driven learner modeling together with MDP-based sequential formulation and DQN-based deep
reinforcement learning have a major impact on adaptive curriculum optimization and learner engagement.

The BKT-inspired aggregated mastery estimation approach and MDP framework DQN we developed needs testing with live educational data streams which will help in building real-time intelligent tutoring systems. The implementation of Double DQN and actor, critic architectures will enhance system performance through advanced reinforcement learning methods. The MDP framework enables personalization to be implemented through hierarchical and multi-agent reinforcement learning systems which support group settings at both classroom and group level. The combination of graph-based knowledge tracing with multimodal behavioral analytics and explainable reinforcement learning mechanisms will improve the ability to understand and trust policy decisions made by the system. The next generation of AI-backed personalized educational systems needs validation across multiple academic fields and MOOC platforms which will demonstrate their ability to operate at full scale and deliver real-world results. The study contributes a unified computational framework integrating probabilistic learner modeling with deep reinforcement learning for adaptive training path optimization, demonstrating measurable performance gains across personalized educational settings. Despite strong empirical performance, validation is currently limited to a single educational dataset, indicating scope for broader cross-domain generalization analysis. Future work will focus on large-scale deployment, multi-dataset validation, and enhanced policy adaptation mechanisms to further improve scalability and instructional intelligence.

Acknowledgement

The authors sincerely thank Zhejiang College of Security Technology for its academic support, along with faculty and colleagues from the Continuing Education College and College of Artificial Intelligence for their valuable feedback. They also acknowledge the Kaggle platform for providing the dataset used in this study and appreciate the contributions of researchers in reinforcement learning
and educational data mining. Finally, they are grateful to the reviewers and editors for their insightful comments that enhanced the manuscript's quality.

Declaration

Data availability

https://www.kaggle.com/datasets/programmer3/teaching-quality-and-student-engagement-dataset

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Table 3. Optimal Reinforcement Learning Hyperparameters for Training Path Optimization

Method	Accuracy (%)	Average Reward	Convergence Episodes	Final Q-Loss	Success Rate (%)	Reward Variance
Q-Learning (Baseline)	89.2	58.31	1450	0.8421	30.2	72.84
Policy Gradient (REINFORCE)	95	64.12	1200	0.5364	81.92	66.57
Proposed DQNBased Framework	96.38	67.63	1000	0.2533	89.56	60.60

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or notfor-profit sectors.

Author contributions

Yi Zhexue: Conceptualization, Data Curation, Methodology, Writing - Original Draft Preparation.
Lin Ze: Software Development, Validation, Formal Analysis, Visualization.
Chen Feng: Supervision, Project Administration, Writing - Review & Editing, Funding Acquisition

Ethical approval

This study does not involve human participants or animals. The dataset used is publicly available and anonymized; therefore, ethical approval was not required.

Consent to participate

Not applicable.

Consent to publication

All authors have read and approved the final manuscript and consent to its publication.

Competing interests

The authors declare that they have no competing interests.

References

[1] L.-Y. Zhou and Y.-Y. Wang, (2025) *"Simulation of Personalized English Learning Path Recommendation System Based on Knowledge Graph and Deep Reinforcement Learning"* **Scientific Reports** ***15***(1): 34554. DOI: 10.1038/s41598-025-17918-x.

[2] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, (2023) *"Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning"* **IEEE Access** ***11***: 89769–89790. DOI: 10.1109/ACCESS.2023.3305584.

[3] Y. Ma, L. Wang, J. Zhang, F. Liu, and Q. Jiang, (2023) *"A Personalized Learning Path Recommendation Method Incorporating Multi-Algorithm"* **Applied Sciences** ***13***(10): 5946. DOI: 10.3390/app13105946.

[4] B. Jiang et al., (2022) *"Data-Driven Personalized Learning Path Planning Based on Cognitive Diagnostic Assessments in MOOCs"* **Applied Sciences** ***12***(8): 3982. DOI: 10.3390/app12083982.

[5] Q. Yang and C. Liang, (2025) *"A Second-Classroom Personalized Learning Path Recommendation System Based on Large Language Model Technology"* **Applied Sciences** ***15***(14): 7655. DOI: 10.3390/app15147655.

[6] O. Bulut, J. Shin, S. N. Yildirim-Erbasli, G. Gorgun, and Z. A. Pardos, (2023) *"An Introduction to Bayesian Knowledge Tracing with pyBKT"* **Psych** ***5***(3): 770–786. DOI: 10.3390/psych5030050.

[7] S.-Y. Jeong and Y.-K. Kim, (2023) *"Deep Learning-Based Context-Aware Recommender System Considering Change in Preference"* **Electronics** ***12***(10): 2337. DOI: 10.3390/electronics12102337.

[8] M. Barone, M. Naeem, M. Ciaschi, G. Tretola, and A. Coronato, (2025) *"AI-Based Intelligent System for Personalized Examination Scheduling"* **Technologies** ***13***(11): 518. DOI: 10.3390/technologies13110518.

[9] J. P. López-Goyez, A. González-Briones, and Y. Demazeau, (2026) *"An Adaptive Multi-Agent Architecture with Reinforcement Learning and Generative AI for Intelligent Tutoring Systems: A Moodle-Based Case Study"* **Applied Sciences** ***16***(3): 1323. DOI: 10.3390/app16031323.

[10] A. Darvish and M. Sepehri, (2025) *"Artificial Intelligence-Driven Project Portfolio Optimization Under Deep Uncertainty Using Adaptive Reinforcement Learning"* **Applied Sciences** ***15***(23): 12713. DOI: 10.3390/app152312713.

[11] S. Saleem, M. U. Aziz, M. J. Iqbal, and S. Abbas, (2025) *"AI in Education: Personalized Learning Systems and Their Impact on Student Performance and Engage-*

ment" **Critical Review of Social Science Studies** *3*(1): 2445–2459. DOI: 10.59075/c35qa453.

[12] C. Y. Zaharuddin and G. Yao, (2024) *"Enhancing Student Engagement with AI-Driven Personalized Learning Systems"* **International Transactions on Education Technology** *3*(1): 1–8. DOI: 10.33050/itee.v3i1.662.

[13] C. F. Mahmoud and J. T. Sørensen, (2024) *"Artificial Intelligence in Personalized Learning with a Focus on Current Developments and Future Prospects"* **Research Advances in Education** *3*(8): 25–31. DOI: 10.56397/RAE.2024.08.04.

[14] C. C. Y. Yang and H. Ogata, (2023) *"Personalized Learning Analytics Intervention Approach for Enhancing Student Learning Achievement and Behavioral Engagement in Blended Learning"* **Education and Information Technologies** *28*(3): 2509–2528. DOI: 10.1007/s10639-022-11291-2.

[15] S. K. Sahu, A. Mokhade, and N. D. Bokde, (2023) *"An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges"* **Applied Sciences** *13*(3): 1956. DOI: 10.3390/app13031956.