Journal of Information Systems and Informatics

Vol. 7, No. 3, September 2025 e-ISSN: **2656-4882** p-ISSN: **2656-5935**

DOI: 10.51519/journalisi.v7i3.1270 Published By DRPM-UBD

Actor-Critic Reinforcement Learning for Personalized STEM Learning Path Optimization

Muhammad Hatta[1], Lena Magdalena[2], Dwi Pasha Anggara Putra[3], Yohanes Michael Fouk Runa[4], Ananda Irfansyah[5], Fernando Valentino[6]

[1,2,3,4,5]Information System Departement, Catur Insan Cendekia University, Cirebon, Indonesia
Email: [1]muhammad.hatta@cic.ac.id, [2]lena.magdalena@cic.ac.id, [3]dwi.putra.si.22@cic.ac.id, [4]yohanes.runa.si.23@cic.ac.id, [5]ananda.irfansyah.si.22@cic.ac.id, [6]fernando.valentino.si.22@cic.ac.id

Abstract

This study addresses the critical need for adaptive learning in non-formal education settings, particularly Community Learning Centres (PKBM) in Indonesia, where student heterogeneity and limited resources challenge conventional teaching methods. We developed a personalized learning path optimization model using Actor-Critic Reinforcement Learning (RL) to enhance STEM competency development. The novel framework integrates cognitive, affective, and personality features to dynamically adjust material difficulty based on real-time analysis of student cognitive states (quiz performance, completion rate) and affective conditions (emotional level), moving beyond static predictive approaches. Experimental results on a synthetic dataset demonstrate that the Actor-Critic agent achieves statistically significant higher rewards (-2.92 vs -3.01, $p<0.05$) and greater output stability compared to a random baseline. Although the absolute reward difference is modest, it reflects more consistent adaptive policy performance, despite limited effect size (Cohen's d=0.0317). Feature importance analysis confirms that quiz_score and emotion_level are the dominant factors influencing adaptive recommendations, while personality traits show negligible impact. The framework offers a viable pathway for scalable, personalized learning in resource-constrained environments. Future work should validate the model with real-world student data and refine reward functions to strengthen practical impact.

Keywords: Adaptive Learning, Reinforcement Learning, STEM Competency, Personalized Learning Paths, Non-Formal Education.

1. INTRODUCTION

Adaptive learning has become increasingly vital in modern education, significantly improving learning outcomes particularly in low and middle-income countries by tailoring instruction to individual student needs [1, 2]. This approach effectively addresses student diversity and resource limitations through real-time, personalised support [3, 4]. In Indonesia, however, adaptive learning implementation faces significant challenges in non-formal education settings such as Community Learning Centres (PKBMs), which serve learners outside the formal education system. PKBMs typically employ generalized, non-adaptive teaching methods and have yet to effectively integrate educational technologies, despite the growing importance of STEM skills in the digital era [5].

Machine learning techniques including collaborative filtering, neural networks, and supervised learning have been utilized to develop adaptive learning systems that recommend content based on learning styles or historical performance [6]. However, these approaches remain predominantly static and predictive, relying on historical data without adapting to real-time changes in student behavior [6],[7],[8]. In contrast, Reinforcement Learning (RL) particularly the actor-critic method enables dynamic personalization by continuously balancing exploration (actor) and evaluation (critic), allowing for real-time adaptation of learning paths based on ongoing student interactions [9], [10], [11], [12].

Despite the demonstrated potential of RL in adaptive learning, actor-critic-based RL has emerged as a pivotal foundation for the development of adaptive learning systems that respond to students' needs and performance [10], [11], [13]. However, its application in non-formal educational contexts, particularly in Indonesian PKBMs for STEM competency development, remains limited. Current adaptive learning implementations in Indonesian community learning centres predominantly follow linear, generalised approaches that fail to accommodate student heterogeneity and teacher limitations [17]. Existing studies largely rely on historical data, resulting in inflexible learning pathways [14], [15]. Despite evidence that RL can enhance engagement and outcomes through dynamic personalisation, there is a significant research gap in developing RL-based models specifically designed for the unique constraints and student diversity characteristic of PKBMs [12], [14], [15], [16].

This study aims to develop and evaluate an Actor-Critic RL model for personalized learning path recommendation in PKBMs. Specifically, we seek to: (1) design a dynamic learning environment that integrates cognitive, affective, and personality features; (2) compare the performance of the Actor-Critic agent against a random baseline; and (3) identify which student factors most influence adaptive rewards. The research addresses a critical gap in non-formal STEM education by offering a scalable, real-time personalization framework suitable for resource-constrained settings.

2. METHODS

3.1. Dataset and Features

The dataset under consideration in this study is comprised of multi-session learning interactions of community learning centres students, with the objective of capturing the dynamics of learning. The features employed are categorised into three distinct groupings: firstly, cognitive features, which include metrics such as time_spent, quiz_score, material_completed, and num_attempts; Second, personality, which consists of the five dimensions of the Big Five Personality: openness, conscientiousness, extroversion, agreeableness, and neuroticism; Third, affective, represented by emotion_level as an indicator of students' emotional state when interacting with learning materials, can be seen in Table 1. The label employed is "reward," which evaluates the suitability of the material in relation to the conditions and characteristics of the students. Consequently, this dataset facilitates the training of RL models to assess the efficacy of adaptive learning pathways.

Table 1. Features and Descriptions

Feature	Category	Description
time_spent	Cognitive	Time spent by students studying the material (in minutes).
quiz_score	Cognitive	Quiz scores obtained by students (0–100).
material_completed	Cognitive	Material completion status (0 = not yet, 1 = completed).
num_attempts	Cognitive	Number of attempts made by students to complete assignments/quizzes.
openness	Personality	Level of openness of students to new experiences (0–1).
conscientiousness	Personality	Level of discipline and responsibility of students (0–1).
extraversion	Personality	Level of social involvement of students (0–1).
agreeableness	Personality	Level of friendliness and cooperation of students (0–1)
neuroticism	Personality	Level of students tendency towards stress or anxiety (0–1).
emotion_level	Affective	Students' emotional state while learning, e.g. motivation or frustration (0–1).
reward (target)	Outcome	A score representing the success of the learning path based on interactions.

Figure 1. Histogram of feature distributions in the synthetic learner dataset. The figure illustrates the distribution of cognitive and performance-related variables (time_spent, quiz_score, material_completed, num_attempts), personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism), and affective state (emotion_level), showing variation among learners across multiple sessions.

Based on the descriptive statistical data presented in Table 2, there is diversity in the learning and psychological characteristics of students in the dataset. The average learning time (time_spent) is around 32 minutes with a relatively high standard deviation (≈16 minutes), reflecting variations in learning styles, from students who complete tasks quickly to those who need more time.

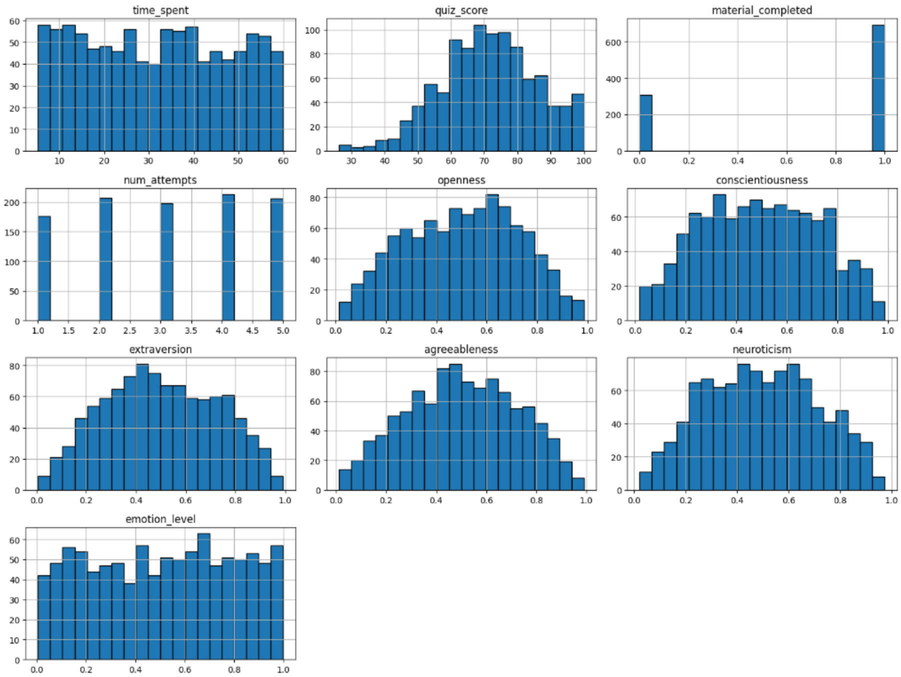


Figure 1. Distribution of features in the dataset)

In terms of cognitive achievement, the average quiz score (quiz_score) was 71 with a standard deviation of 15, indicating heterogeneity in academic ability among participants. Most students (69%) completed the material provided, although there were still some who did not. Meanwhile, the average number of attempts (num_attempts) was 3 times with a variation of 1–5 times, showing that some students needed more iterations to achieve success.

In terms of personality, the Big Five Traits profile shows a stable mean value of around 0.49–0.50—close to the midpoint of the scale (0–1). The fairly wide distribution (standard deviation ≈0.22–0.23) confirms the diversity of personality traits in the sample, a crucial factor in the adaptive approach.

On the affective side, the average level of student emotion (emotion_level) was 0.51 with significant variation (standard deviation ≈0.28). The range of values from a minimum close to 0.00 to a maximum close to 1.00 confirms a broad spectrum of emotions, from frustration to enthusiasm, which need to be responded to differently in an adaptive learning system.

Table 2. Descriptive statistics of features in the synthetic learner dataset

Feature	Mean	Std. Dev	Minimum	Maximum
time_spent	31.964110	16.067555	5.254761	59.984472
quiz_score	71.353469	14.546905	26.179743	100.000000
material_completed	0.692000	0.461898	0.000000	1.000000
num_attempts	3.066000	1.394844	1.000000	5.000000
openness	0.504103	0.223974	0.012692	0.985998
conscientiousness	0.495263	0.230097	0.017333	0.986921
extraversion	0.502862	0.226195	0.006451	0.991470
agreeableness	0.504233	0.220437	0.010810	0.990429
neuroticism	0.495056	0.218873	0.021306	0.973673
emotion_level	0.512548	0.288622	0.004337	0.999268

3.2. Environment Design (Adaptive Learner Env)

The objective of this study was to design a simulation environment (AdaptiveLearnerEnv) for adaptive learning systems, with the overall workflow illustrated in Figure 2. The environment state comprises ten features representing student characteristics across cognitive, affective, and personality dimensions. The RL agent selects actions corresponding to material difficulty levels (easy, medium, hard), with the reward function calculated based on action suitability relative to student preferences and cognitive achievements.
The reward function is defined as:

$$R = \alpha \cdot 1(action = preferred) + \beta \cdot material\_completed + \delta \cdot \frac{quiz\_score}{100} - \gamma \cdot penalty \quad (1)$$

where adjustment weights α, β, δ, and γ were determined through grid search optimization to balance the competing objectives of cognitive achievement, material completion, emotional well-being, and learning efficiency. Specifically, the weights were tuned to maximize cumulative reward during preliminary experiments, resulting in values of α=0.4, β=0.3, δ=0.2, and γ=0.1. This multi-objective formulation ensures the model not only pursues quiz scores but also balances material completion, emotional state, and student effort, thereby generating truly adaptive and personalized learning paths.

Figure 2. Adaptive Learning System Workflow: (a) Student interacts with learning materials, (b) System extracts cognitive, affective, and personality features, (c) Actor-Critic agent processes state information, (d) Agent selects difficulty level action, (e) Environment returns reward and updated state.

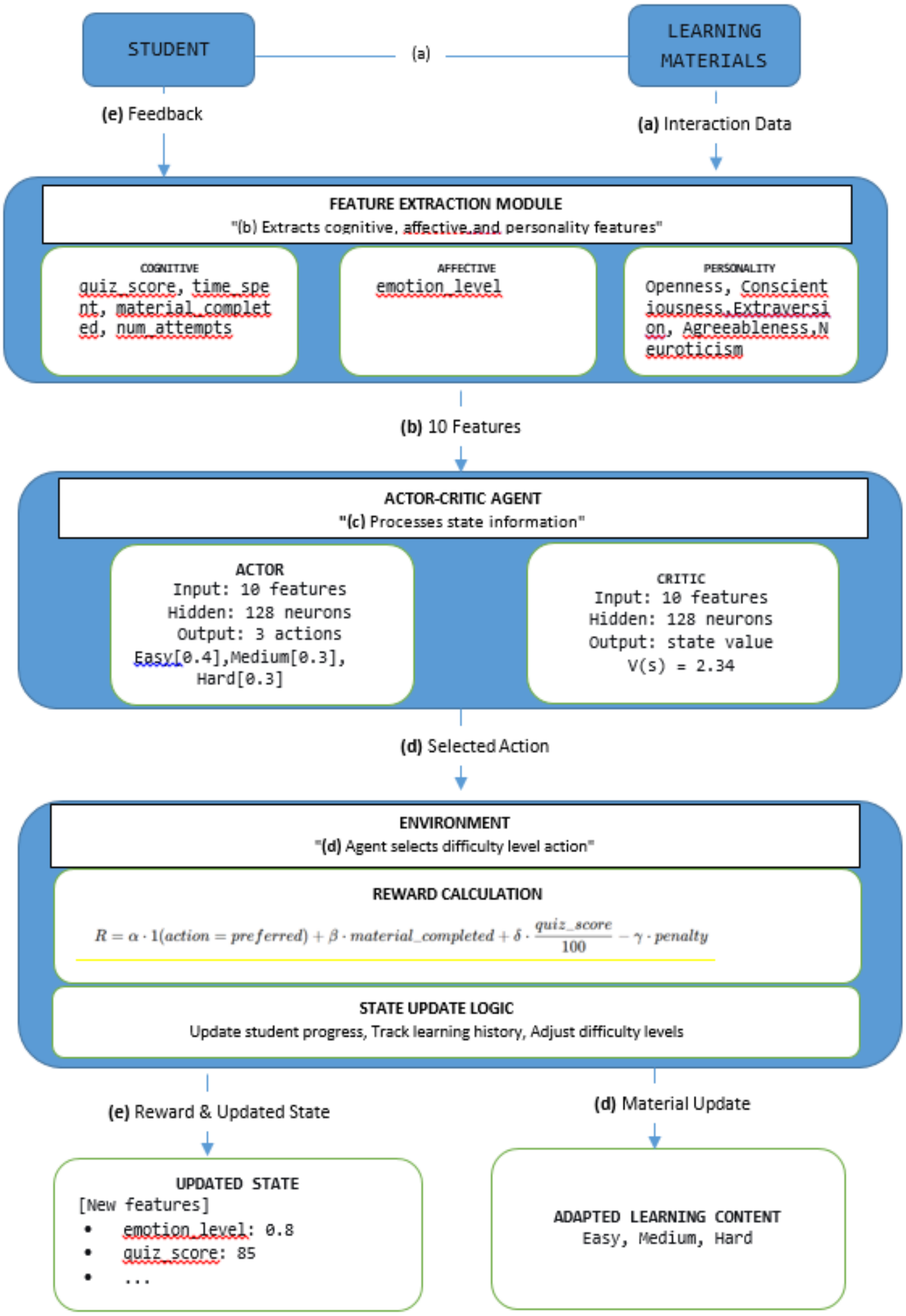

STUDENT
(a)
LEARNING MATERIALS
(e) Feedback
(a) Interaction Data
FEATURE EXTRACTION MODULE
"(b) Extracts cognitive, affective and personality features"
COGNITIVE
quiz_score, time_spent, material_completed, num_attempts
AFFECTIVE
emotion_level
PERSONALITY
Openness, Conscientiousness,Extraversion, Agreeableness,Neuroticism
(b) 10 Features
ACTOR-CRITIC AGENT
"(c) Processes state information"
ACTOR
Input: 10 features
Hidden: 128 neurons
Output: 3 actions
Easy[0.4],Medium[0.3], Hard[0.3]
CRITIC
Input: 10 features
Hidden: 128 neurons
Output: state value
V(s) = 2.34
(d) Selected Action
ENVIRONMENT
"(d) Agent selects difficulty level action"
REWARD CALCULATION
$R = \alpha \cdot 1(action = preferred) + \beta \cdot material\_completed + \delta \cdot \frac{quiz\_score}{100} - \gamma \cdot penalty$
STATE UPDATE LOGIC
Update student progress, Track learning history, Adjust difficulty levels
(e) Reward & Updated State
(d) Material Update
UPDATED STATE
[New features]
• emotion_level: 0.8
• quiz_score: 85
• ...
ADAPTED LEARNING CONTENT
Easy, Medium, Hard

Figure 2. Adaptive Learning System Workflow

3.3. Actor-Critic Model Architecture

The model architecture employed in this study is Actor-Critic, which is constructed as a Multi-Layer Perceptron (MLP) using the Python framework, PyTorch. The model comprises two core components: an actor that generates a probability distribution for action selection, and a critic that estimates state quality. It is evident that both systems share the same initial layer, the purpose of which is to extract features from states that have 10 dimensions. The separation of tasks enables the model to concurrently learn the optimal policy and evaluate the quality of each observed state.

The actor component is implemented through an MLP with a single hidden layer of 128 neurons, activated using the ReLU function, followed by an output layer that generates a probability distribution over three possible actions (easy, medium, hard). Meanwhile, the critic employs a similar MLP architecture, but its output layer consists of a single neuron that represents the estimated state value V(s). Therefore, the actor focuses on the policy π(a | s) that maps states to actions, while the critic learns the value function V(s)=E[Rt|st=s] that evaluates the expected return of a state.

The model was trained with the Adam optimizer, with a learning rate of 0.001, to ensure stable convergence. During the training process, the actor was updated using the log-likelihood-based policy gradient of the actions taken, while the critic was updated by minimizing the value loss between the estimate and the target return. The combined objective function can be expressed as shown in Equation 2.

$$L(\theta) = -\mathbb{E}[\log \pi_\theta(a_t|s_t) \cdot A_t] + \frac{1}{2}(R_t - V_\phi(s_t))^2 - \beta H(\pi_\theta) \qquad (2)$$

The function A_t is defined as the advantage function, R_t is defined as the discounted return, and H(πθ) is defined as the entropy of the policy distribution to maintain exploration. Integrating the actor and critic into a unified architecture facilitates more stable learning in comparison to pure policy gradient methods. This integration enables the model to generate more personalized and dynamic adaptive learning paths, as illustrated in Figure 3.

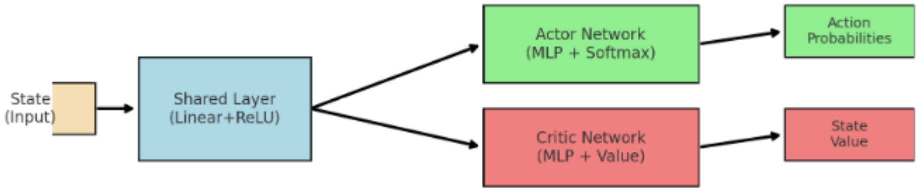

State (Input)
Shared Layer (Linear+ReLU)
Actor Network (MLP + Softmax)
Critic Network (MLP + Value)
Action Probabilities
State Value

Figure 3. Actor-Critic Model Architecture

3.4. Training Setup

The training process was conducted in the form of episode simulations, where each episode represented a complete series of interactions between the agent and the environment. The initial number of episodes was set at 50 as the basic configuration, then expanded to 500 episodes to test the stability and consistency of the model. Each episode comprised a series of steps, initiated by the resetting of the environment. Subsequently, the agent selected actions based on the probability distribution of the actor policy. The calculation of the reward was then executed in accordance with the predetermined reward function.

A systematic comparison of two agents was conducted to assess the effectiveness of the method. Initially, the Actor-Critic Agent is trained using policy gradient and value function to update the neural network weights. The actor generates action probabilities $\pi(a|s;\theta)$, while the critic estimates the state value $V(s;\theta_v)$. The formula for the combined loss is as shown in Equation 3.

$$L(\theta, \theta_v) = -\log \pi(a|s;\theta) \cdot A(s,a) + \frac{1}{2}(R - V(s;\theta_v))^2 - \beta H(\pi(s;\theta)), \quad (3)$$

In this model, $A_{(s,a)}$ is defined as the advantage, R is defined as the cumulative reward, and H is defined as the entropy of the action distribution to maintain exploration. The optimizer employed in this study is Adam.

For the purpose of comparison, Random Agent is utilized as a baseline. This agent's actions are selected at random, devoid of any learning mechanism, resulting in a learning path that manifests as non-adaptive performance. A comparison between Actor-Critic and Random Agent algorithms enables the analysis of the extent to which Actor-Critic-based RL can produce a more adaptive learning trajectory than random strategies. During the training phase, the rewards received from both agents are systematically documented for each individual episode. These rewards are then represented graphically in the form of average reward curves. By utilizing this approach, the efficacy of the model can be evaluated through empirical means. Pseudocode training can be written as show in Code 1.

```
Initialize  Actor-Critic  network  with  parameters  θ  (actor)  and  θv
(critic)
Set optimizer = Adam(θ, θv), learning_rate = 0.001
For episode = 1 to N_episodes do:
    state ← reset(environment)
    total_reward ← 0
    For step = 1 to episode_length do:
        # === Actors choose actions based on probability distributions.
===
        action, log_prob ← sample_from_actor(state)
        # === The environment executes actions and gives rewards. ===
        next_state, reward, done ← env.step(action)
        total_reward += reward
        # === Critic calculates state value ===
        value ← V(state; θv)
        next_value ← V(next_state; θv)
        # === Calculate the advantage ===
        advantage ← reward + γ * next_value - value
        # === Calculate combined loss ===
        policy_loss ← -log_prob * advantage
        value_loss  ← (reward + γ * next_value - value)^2
        entropy_loss ← -β * entropy(action_probs)
        total_loss ← policy_loss + 0.5 * value_loss + entropy_loss
        # === Update parameters ===
        optimizer.zero_grad()
        total_loss.backward()
        optimizer.step()
        state ← next_state
        If done then break
    Record total_reward for episode
```

Code 1. Training Pseudocode

3.5. Evaluation Metrics

The evaluation of the model was conducted through a comparative analysis of the Actor-Critic Agent with the Random Agent as a baseline, employing a range of quantitative metrics and supplementary analyses. Table 3 presents a comprehensive overview of the evaluation metrics employed in this study.

Table 3. Summary of evaluation metrics used in the study

Metric	Description	Purpose
Average Total Reward	Average cumulative reward per episode ($R_{total}=\sum^{T}_{t=1}$ rt)	Measuring the main performance of the model in generating adaptive learning paths.
Reward Distribution	Visualisation of rewards (line plot, box plot, violin plot) per episode.	Observing stability, variation, and consistency of performance between episodes.
Independent t-test	Average reward of Actor-Critic vs Random Agent.	Statistically significant performance ($p<0.05p < 0.05$).
Cohen's d	Measure of the effect of the difference in average reward between agents.	Providing practical interpretations of the magnitude of performance differences.
Correlation Heatmap	Pearson correlation between student features and reward.	Identifying the relationship between student characteristics and adaptive reward achievement.
Feature Importance (Permutation)	Measurement of the relative contribution of each feature to reward prediction.	Determining the dominant characteristics that influence adaptive learning pathways.

The employment of diverse evaluation metrics facilitates a comprehensive analysis of model performance, encompassing both quantitative and interpretative dimensions. The mean reward and its distribution are indicative of the model's effectiveness and consistency. Moreover, statistical tests (t-test, Cohen's d) ensure that the differences with the baseline are academically and practically significant. The implementation of correlation analysis and feature importance (permutation and SHAP) serves to enhance the interpretation by elucidating the primary factors that influence the adaptive learning trajectory.

3. RESULTS AND DISCUSSION

3.1. Experimental Performance

1) Training Performance

The Actor-Critic model training process demonstrates a pattern of cumulative reward increase as the number of episodes increases. In the initial phase (episodes 0–50), the reward fluctuates, but it begins to stabilize after more than 400 episodes.

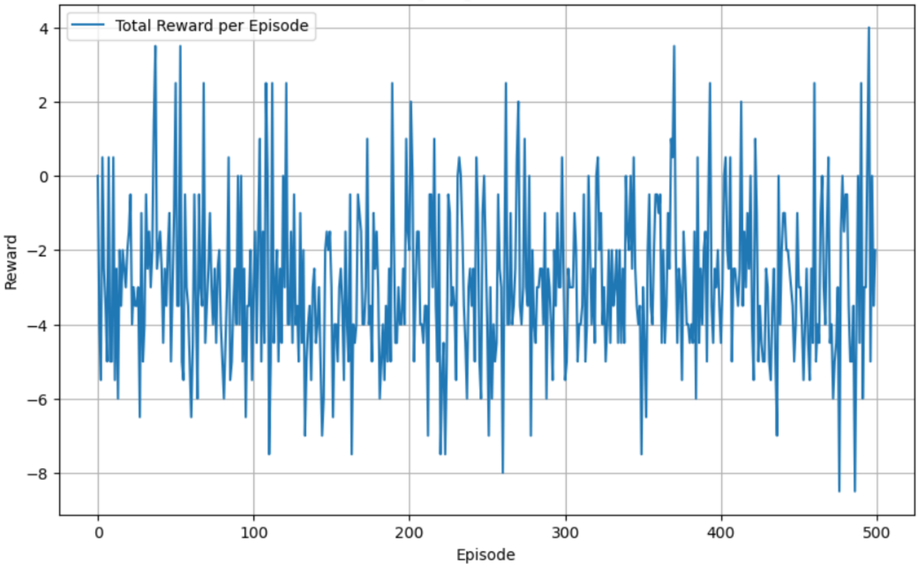


Figure 4. Traning Progress Actor-Critic.

As illustrated in Figure 4, the reward trend in the Actor-Critic training process from 500 episodes demonstrates a consistent pattern. In general, the reward fluctuates within the range of -9 to +4, indicating that the learning process is unstable. In the range of episodes 451–500, the reward pattern remains predominantly characterized by negative values, ranging from -8.5 to +4.0.

However, there are notable exceptions, including episodes 461 (+2.5), 491 (+2.5), and 496 (+4.0), which exhibit positive rewards. This pattern suggests that the model is beginning to identify an adaptive learning trajectory, although this trajectory is not yet consistent, as shown in Figure 5.

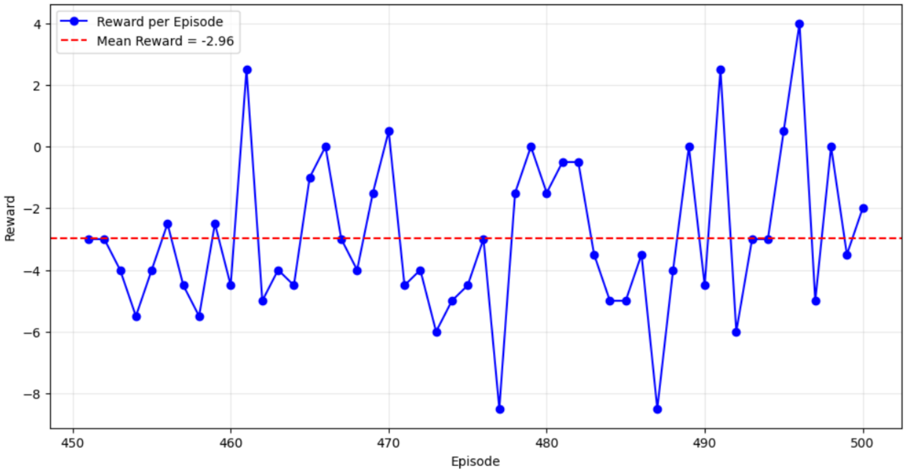


Figure 5. Traning Progress Episode 451-500

2) Evaluating Models in the Environment

The evaluation of the model was conducted through the implementation of a test of the Actor-Critic agent within the AdaptiveLearnerEnv environment, utilizing a dataset. The evaluation process was executed for a total of 100 episodes, with the objective of obtaining an overview of the model's performance stability. In each episode, the agent initiated from the initial condition (reset) and interacted repeatedly with the environment until the episode was completed. The actions taken at each step were determined by the training policy, without a gradient update process, thereby reflecting the model's generalisation ability.

As illustrated in Figure 6 above, the results of the Actor-Critic model evaluation over 100 episodes are presented. The reward per episode exhibits significant variability, ranging from approximately –7 to +3, with an average value of –2.67, as indicated by the red dotted line. This pattern suggests that, while the model demonstrates the capacity to generate multiple episodes that yield positive rewards, the learned policy remains unstable. Consequently, the model tends to exhibit suboptimal decisions in terms of selecting the level of difficulty for the material.

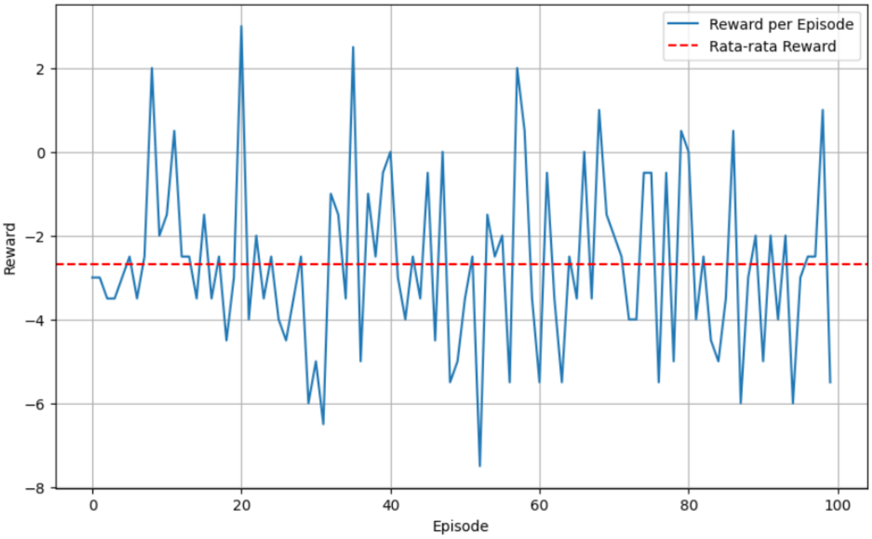


Figure 6. Evaluation of the Actor-Critic Model.

3) Learning Path Analysis

An analysis of the action distribution of the Actor-Critic model reveals a predominant tendency to recommend material at the easy level. This phenomenon is illustrated in Figure 7, which shows that all model recommendations are concentrated in the Easy category, with no variation observed at the Medium or Hard levels. This condition suggests that the model exhibits a propensity to adopt a conservative strategy in determining the learning path, thereby frequently directing students to fundamental material.

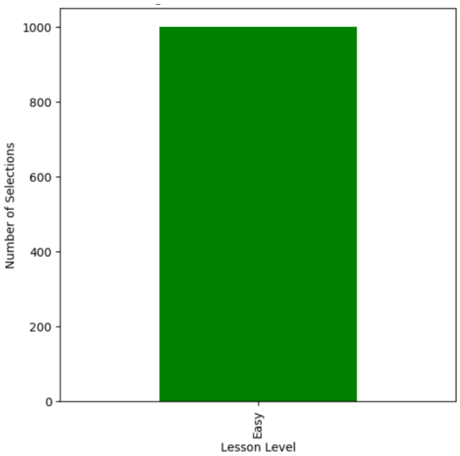


Figure 7. Evaluation of the Actor-Critic Model.

4) Comparison with Random Agent

The present study evaluates the performance of the model by comparing the Actor-Critic agent to the Random Agent as a baseline. The Actor-Critic model is trained using student state representations, while the Random Agent chooses actions randomly. An evaluation of the results across a total of 1,000 episodes reveals that the Actor-Critic method attains an average reward of -2.92, which is marginally higher than the -3.01 average achieved by the Random Agent. Despite the negligible numerical disparity, the Actor-Critic reward distribution manifests a heightened degree of stability.

As illustrated in Figure 8, the Actor-Critic algorithm generates a reward pattern that is more concentrated around its mean, while the Random Agent demonstrates higher variability. This finding suggests that Actor-Critic has developed a more targeted policy. However, its inability to consistently exceed baseline performance indicates that this architecture still requires further refinement, such as in its reward function or exploration strategy.

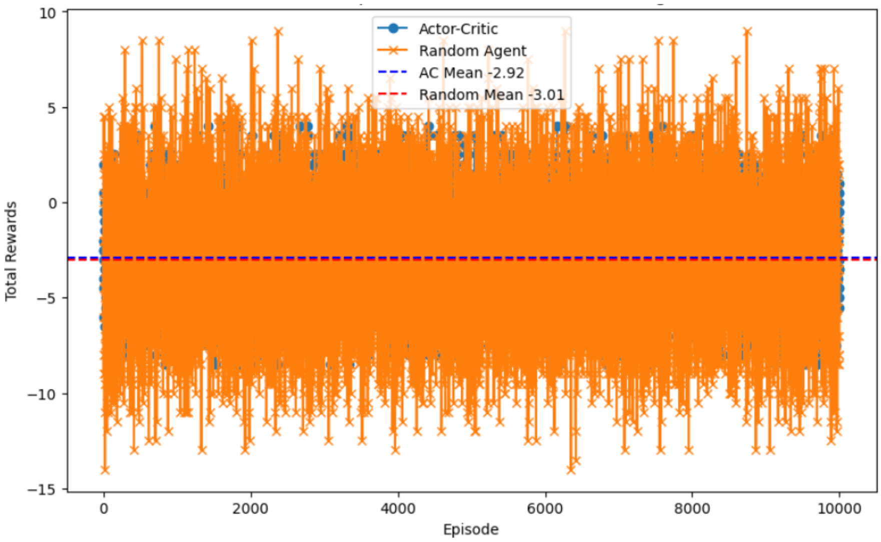


Figure 8. Reward Comparison: Actor-Critic vs. Random Agent.

5) Independent t-test

The results of the independent t-test with Welch's t-test revealed a T-statistic value of 2.2426 and a p-value of 0.024935, indicating a statistically significant difference in performance between the Actor-Critic Agent and the Random Agent at a significance level of $p < 0.05$. The findings demonstrate that Actor-Critic yields a higher mean reward in comparison to the random baseline, thereby substantiating its role as a more efficacious and reactive approach in deriving adaptive learning pathways based on student data dynamics, as illustrated in Table 3.

Tabel 3. Independent t-test Statistical Test Results

Test	T-statistic	p-value	Decision
Actor-Critic vs Random Agent	2.2426	0.024935	Significant (p < 0.05): Actor-Critic is better

6) Effect Size: Cohen's d

A meticulous examination of the effect size, as measured by Cohen's d, yielded a value of 0.0317, thereby classifying the effect as negligible. While a statistically significant discrepancy in performance was observed between the Actor-Critic Agent and the Random Agent, the magnitude of the effect was found to be virtually negligible. This finding suggests that an enhancement in the model's performance was identified; however, its effect remains negligible within the context of its implementation at community learning centres. Consequently, further development is necessary, such as refining the reward function or incorporating contextual features to augment its practical impact.

Table 4. Statistical Test and Effect Size Results

Analysis	Value	Interpretation
T-statistic (Welch's t)	2.2426	Positive values indicate that Actor-Critic tends to perform better than Random Agent
P-value	0.0249	Significant (p < 0.05), Actor-Critic differs significantly from Random Agent.
Cohen's d	0.0317	Negligible effect → significant difference but with very small effect strength.

7) Visualisasi Distribusi Reward

As illustrated in Figure 9, the results of the visual data analysis, employing box plots and violin plots, indicate that the distribution of Actor-Critic rewards demonstrates a heightened concentration around the median, accompanied by diminished variation. Conversely, Random Agent exhibits a broader distribution, marked by a heightened prevalence of negative extreme values. These findings suggest that, while it does not invariably yield higher rewards, the Actor-Critic method is capable of ensuring consistency in learning path recommendations. This stability is more conducive to the sustainability of the student learning process in the context of community learning centers than fluctuating results.

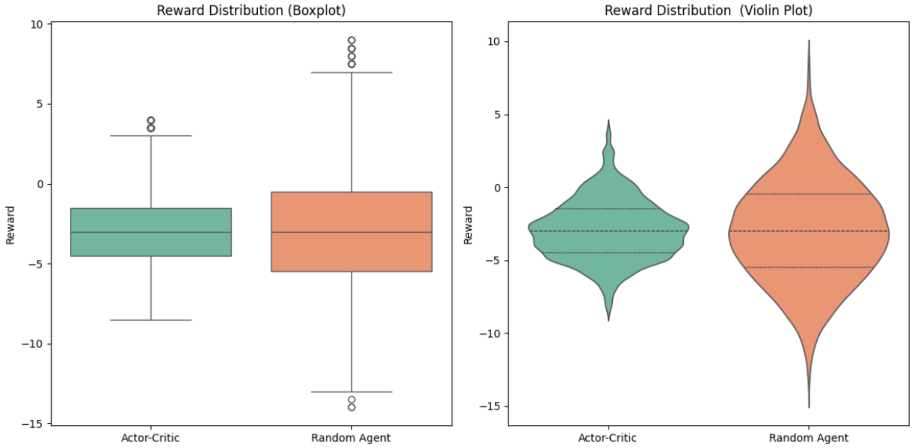


Figure 9. Visualisasi Distribusi Reward

As illustrated in Table 5, the median reward of the Actor-Critic is marginally lower than that of the Random Agent; however, the interquartile range (IQR) is more constrained, suggesting enhanced stability in the outcomes. In contrast, the Random Agent exhibits a more pronounced distribution of rewards, with a minimum value of -15 and a maximum of 9, resulting in increased variability compared to the Actor-Critic.

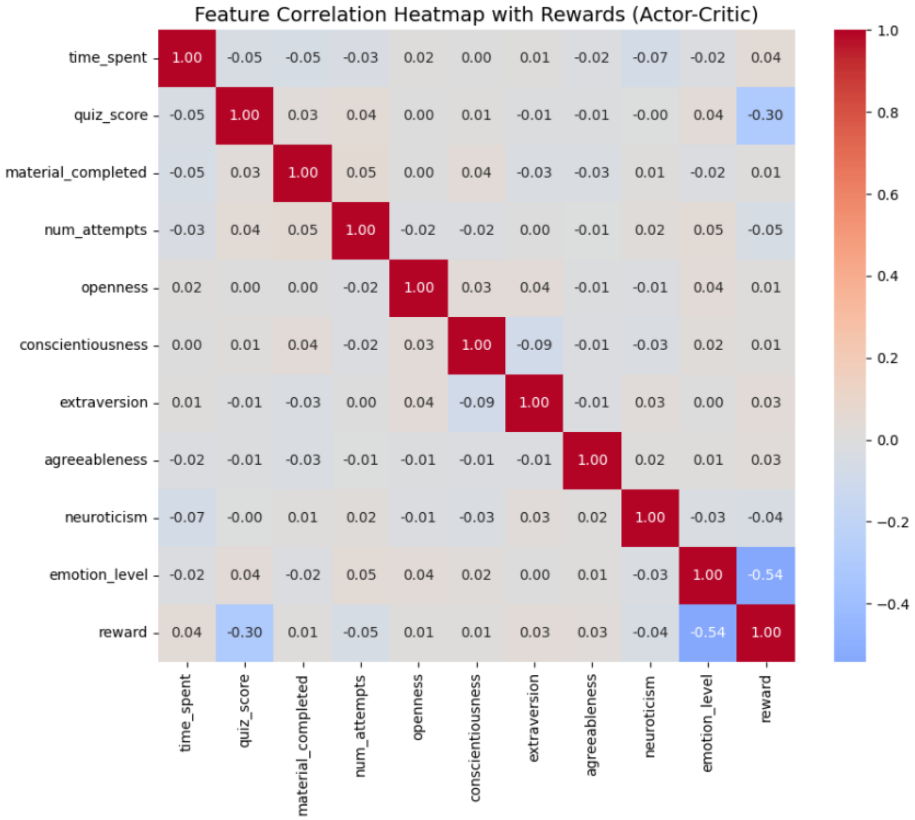
Table 5. Summary of Reward Distribution Statistics

Model	Median	IQR (Q3–Q1)	Minimum	Maximum
Actor-Critic	-3.00	3.00	-9.00	4.00
Random Agent	-2.50	6.00	-15.00	9.00

8) Correlation Heatmap

Figure 10 presents the results of Pearson's correlation analysis, which utilizes a heat map as a graphical representation, to ascertain the contribution of ten input features to the reward. The findings of the study indicate that the majority of the features, particularly the personality dimensions, exhibit negligible correlations. However, two features stand out with negative correlations: quiz_score (r = -0.30) and emotion_level (r = -0.54). This pattern suggests that a discrepancy between the cognitive complexity of the material and the emotional state and performance of students can diminish the reward value.

These findings suggest that affective variables and quiz performance exert a more substantial influence on rewards than personality factors. The implication is that the model must be equipped with the ability to respond more adaptively to students' emotional states and academic performance. The analysis results also open up opportunities to improve reward formation by giving greater weight to affective factors, so that the recommendations generated are more appropriate and support student learning well-being.



	time_spent	quiz_score	material_completed	num_attempts	openness	conscientiousness	extraversion	agreeableness	neuroticism	emotion_level	reward
time_spent	1.00	-0.05	-0.05	-0.03	0.02	0.00	0.01	-0.02	-0.07	-0.02	0.04
quiz_score	-0.05	1.00	0.03	0.04	0.00	0.01	-0.01	-0.01	-0.00	0.04	-0.30
material_completed	-0.05	0.03	1.00	0.05	0.00	0.04	-0.03	-0.03	0.01	-0.02	0.01
num_attempts	-0.03	0.04	0.05	1.00	-0.02	-0.02	0.00	-0.01	0.02	0.05	-0.05
openness	0.02	0.00	0.00	-0.02	1.00	0.03	0.04	-0.01	-0.01	0.04	0.01
conscientiousness	0.00	0.01	0.04	-0.02	0.03	1.00	-0.09	-0.01	-0.03	0.02	0.01
extraversion	0.01	-0.01	-0.03	0.00	0.04	-0.09	1.00	-0.01	0.03	0.00	0.03
agreeableness	-0.02	-0.01	-0.03	-0.01	-0.01	-0.01	-0.01	1.00	0.02	0.01	0.03
neuroticism	-0.07	-0.00	0.01	0.02	-0.01	-0.03	0.03	0.02	1.00	-0.03	-0.04
emotion_level	-0.02	0.04	-0.02	0.05	0.04	0.02	0.00	0.01	-0.03	1.00	-0.54
reward	0.04	-0.30	0.01	-0.05	0.01	0.01	0.03	0.03	-0.04	-0.54	1.00

Figure 10. Feature Correlation Heatmap with Rewards (Actor-Critic)

The findings of the present study indicate a robust negative correlation between emotional state and rewards, suggesting that a decline in emotional state contributes to a decline in students' adaptive achievement. Furthermore, the observed negative correlation between quiz scores and learning path recommendations suggests a potential causal relationship, whereby lower quiz scores are associated with less optimal learning path recommendations. In contrast, the duration of time spent exhibited a positive correlation, albeit a very weak one, suggesting that the amount of time devoted to studying does not have a significant direct impact on rewards. This phenomenon is evident in Table 6.

Tabel 6. Top 3 Feature Correlations with Reward (Actor-Critic)

Feature	Pearson Correlation (r)	Interpretation
Emotion Level	-0.54	Strong negative correlation; low emotions decrease reward.
Quiz Score	-0.30	Moderate negative correlation; high quiz scores do not always correlate with reward.
Num Attempts	-0.05	Very weak correlation; repeated student efforts have little effect on reward.

9) Feature Importance (Permutation)

As illustrated in Table 7, the results of the permutation importance analysis with logistic regression reveal the relative contribution of each feature to positive reward prediction. The results of the study indicate that quiz_score is the most influential feature (importance value 0.0328), consistent with the reward function design that emphasizes cognitive performance. The variable "emotion_level" (0.0093) occupies the second position, thereby underscoring the significance of affective conditions. In a similar vein, "num_attempts" (0.0033) demonstrates a modest yet noteworthy contribution.

Conversely, features such as openness, neuroticism, and time_spent have negative or near-zero importance values, indicating their very limited influence. These findings clearly confirm that in this model, cognitive and affective factors play a more dominant role than personality factors in determining the success of adaptive learning.

Tabel 7. The Importance of Features (Permutations) on Rewards

Feature	Importance Mean	Importance Std
quiz_score	0.0328	0.0068
emotion_level	0.0093	0.0053
num_attempts	0.0033	0.0042
extraversion	0.0028	0.0025
conscientiousness	0.0025	0.0025
agreeableness	0.0000	0.0000
material_completed	0.0000	0.0000
time_spent	-0.0005	0.0037
neuroticism	-0.0015	0.0032
openness	-0.0020	0.0033

3.2. Discussion

This study aimed to develop and evaluate an Actor-Critic Reinforcement Learning (RL) model for personalized learning path recommendations in the context of Community Learning Centres (PKBMs) in Indonesia. The results indicate that, while the Actor-Critic model demonstrates a higher level of consistency in learning path recommendations compared to a random baseline, several challenges persist in terms of achieving stable, optimal performance.

The training performance of the Actor-Critic model showed that the cumulative reward steadily increased over time, with fluctuations in the early stages of training. This behavior is typical of RL models during the exploration phase, where the agent adjusts its policies. As seen in Figure 4, the model's reward fluctuated initially, stabilizing after 400 episodes, but did not exhibit a consistent upward trajectory. In the evaluation phase, despite a significant amount of variability, the model achieved an average reward of -2.67 across 100 test episodes. This suggests that while the model showed some capacity for generating adaptive learning trajectories, the learned policy was still far from optimal. The significant variability in reward values, ranging from approximately -7 to +3, reflects the instability of the learning path recommendations. One key finding from the model evaluation was that the Actor-Critic agent demonstrated a consistent policy, but it often opted for easier material levels, as seen in Figure 7. This could indicate that the model adopts a conservative strategy in selecting the material difficulty, potentially in an effort to maximize learning success, though this might result in suboptimal learning experiences for students who could benefit from more challenging content.

A direct comparison between the Actor-Critic model and the Random Agent baseline showed that the Actor-Critic agent performed slightly better, achieving a marginally higher average reward of -2.92 compared to -3.01 for the Random Agent. However, the reward distribution for the Actor-Critic agent was notably more stable, with rewards concentrated around the median. In contrast, the Random Agent exhibited a broader distribution, characterized by a higher incidence of negative extreme values. This suggests that while the Actor-Critic model was not dramatically more effective in terms of cumulative rewards, it was more consistent, which is an important trait for adaptive learning systems. The improved stability of the Actor-Critic model is a positive outcome, particularly in educational contexts where consistent engagement is crucial for student success.

The statistical analysis, including the independent t-test, revealed a significant difference between the Actor-Critic and Random Agent ($p < 0.05$), reinforcing the notion that the Actor-Critic agent is a more effective method for generating adaptive learning paths. However, the effect size, measured by Cohen's d, was found to be negligible, indicating that while the Actor-Critic agent outperformed the baseline, the practical significance of the difference is limited. This suggests that additional improvements to the model, such as refining the reward function or enhancing exploration strategies, are necessary for achieving a more substantial improvement in performance.

The correlation heatmap and feature importance analysis revealed interesting insights into the factors influencing the rewards. Notably, emotional state (emotion_level) exhibited a strong negative correlation with rewards, with a Pearson correlation of -0.54, indicating that poor emotional states tend to decrease the effectiveness of the learning path recommendations. This suggests that emotional factors play a crucial role in the adaptive learning process and should be given greater weight in future models to improve student engagement and well-being. The model should be better equipped to identify and respond to fluctuations in students' emotional states, which are highly dynamic and can significantly affect learning outcomes.

In terms of cognitive factors, quiz score demonstrated a moderate negative correlation ($r = -0.30$) with reward, suggesting that performance on quizzes is not always directly aligned with reward outcomes. This finding implies that the model might benefit from incorporating a more nuanced understanding of how quiz performance interacts with other factors like emotional state and time spent on tasks. Interestingly, personality traits, including openness, conscientiousness, extraversion, agreeableness, and neuroticism, had negligible correlations with rewards. This highlights the limited impact of these factors in shaping adaptive learning paths within the context of this study. These findings are in line with previous research suggesting that, in certain adaptive learning systems, cognitive and affective variables may be more predictive of learning outcomes than personality traits.

The results of this study have significant implications for the application of adaptive learning technologies in non-formal educational settings like PKBMs. Despite the limited success of the Actor-Critic model in consistently outperforming the baseline agent, the findings indicate the potential for reinforcement learning-based systems to provide personalized learning experiences, especially when student emotional and cognitive states are considered. The ability to dynamically adjust learning materials based on real-time feedback could be particularly beneficial in resource-constrained environments where teachers are often unable to provide individualized attention to every student. However, several challenges remain. The current model's tendency to favor easy material suggests that it is not yet fully optimized to provide an appropriately challenging learning path for all students. This highlights the need for further refinements in the reward function, particularly in balancing cognitive achievement with emotional well-being. Additionally, the inclusion of more contextual features—such as real-time engagement metrics, learning preferences, or external factors affecting students' emotional states—could further enhance the model's performance.

While the study demonstrates the potential of reinforcement learning for adaptive learning in non-formal education, there are several avenues for future research. First, future models should explore more sophisticated reward functions that better balance cognitive, affective, and behavioral dimensions. Second, incorporating more diverse datasets, including data from various learning environments, could improve the generalizability of the model. Furthermore, exploring other RL algorithms, such as Proximal Policy Optimization (PPO) or Deep Q-Networks (DQN), may offer more stable and efficient learning pathways. The study's limitations include the relatively small dataset and the fact that the model was tested in a simulation environment rather than in real-world PKBM settings. A more comprehensive evaluation involving actual students in real PKBMs would provide more robust insights into the practical applicability of the model.

4. CONCLUSION

This study has successfully demonstrated that the Actor-Critic Reinforcement Learning model significantly enhances the quality of adaptive learning pathways compared to a random approach, as evidenced by higher achieved rewards and a more stable distribution of outcomes. Our analysis revealed that cognitive metrics, particularly quiz scores, and affective states, such as emotional levels, serve as the most influential factors in the adaptation process, whereas personality traits exhibited a limited impact. The primary contribution of this work is the development of a reinforcement learning-based adaptive learning framework specifically tailored for Community Learning Centres (PKBMs), with a focused aim on improving STEM competencies. For practical implementation in PKBMs, we recommend a phased adoption strategy beginning with STEM subjects, leveraging existing digital infrastructure to minimize costs, and providing teacher training focused on interpreting system recommendations and offering emotional support. Pilot programs targeting diverse learner groups are advised to initial validate the approach. Looking forward, future research should prioritize validation with real-world student data, explore multi-agent RL and curriculum learning for enriched educational dynamics, and refine reward functions with more nuanced affective and subject-specific parameters. This framework underscores the potential for scalable, personalized STEM education in resource-constrained environments, offering a viable pathway to bridge educational gaps and foster inclusive competency development in non-formal settings.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the Ministry of Research, Technology and Higher Education of the Republic of Indonesia for the Lecturer Research Grant program with the Junior Lecturer Research scheme (Skema Penelitian Dosen Pemula) with Contract number: 8020/LL4/PG/2025 who has provided full support and guidance so that this paper can be realized. Special appreciation is also extended to Catur Insan Cendekia University for the academic and administrative assistance during the research process. Lastly, the authors are grateful to all students and colleagues who contributed to the implementation of this study.

REFERENCES

[1] E. Aymane, D. Aziz, H. Abdelfatteh, and A. Abdelhak, "Enabling sustainable learning: a machine learning approach for an eco-friendly multi-factor adaptive E-learning system," *Procedia Comput. Sci.*, vol. 236, pp. 533–540, 2024.

[2] L. Major, G. A. Francis, and M. Tsapali, "The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis," *Br. J. Educ. Technol.*, vol. 52, no. 5, pp. 1935–1964, Sept. 2021, doi: 10.1111/bjet.13116.

[3] M. Taşkın, "Artificial Intelligence in Personalized Education: Enhancing Learning Outcomes Through Adaptive Technologies and Data-Driven Insights," *Hum. Comput. Interact.*, 2025, doi: 10.62802/ygye0506.

[4] F. Naseer, M. N. Khan, M. Tahir, A. Addas, and S. M. H. Aejaz, "Integrating deep learning techniques for personalized learning pathways in higher education," *Heliyon*, vol. 10, 2024, doi: 10.1016/j.heliyon.2024.e32628.

[5] C.-J. Ku, K.-Y. Lin, H. Kwon, and T. R. Kelley, "A Six-Stage Instructional Design Model for Collaborative Implementation of Integrated STEM Education," *J. Technol. Educ.*, vol. 36, no. 2, pp. 25–60, 2025.

[6] S. G. Essa, T. Çelik, and N. Human-Hendricks, "Personalized Adaptive Learning Technologies Based on Machine Learning Techniques to Identify Learning Styles: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 48392–48409, 2023, doi: 10.1109/ACCESS.2023.3276439.

[7] A. A. F. Osman, "A New Approach to Machine Learning Algorithms in Adaptive E-Learning Systems," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 15s, pp. 419–432, Mar. 2025, doi: 10.52783/jisem.v10i15s.2480.

[8] Z. Ersozlu, S. Taheri, and I. Koch, "A review of machine learning methods used for educational data," *Educ. Inf. Technol.*, vol. 29, no. 16, pp. 22125–22145, 2024.

[9] J. Xie, "The Role of Reinforcement Learning in Enhancing Education: Applications in Psychological Education and Intelligent Tutoring Systems," *Highlights Sci. Eng. Technol.*, 2025, doi: 10.54097/rkxbvk42.

[10] C. W. Fernandes, T. Miari, S. Rafatirad, and H. Sayadi, "Unleashing the Potential of Reinforcement Learning for Enhanced Personalized Education," in *2023 IEEE Frontiers in Education Conference (FIE)*, Oct. 2023, pp. 1–5, doi: 10.1109/FIE58773.2023.10342902.

[11] Y. Li, "Research on the Optimization Path of Multi-scene Teaching Strategies for English in Higher Vocational Railway Industry Based on Reinforcement Learning," *J. Comb. Math. Comb. Comput.*, 2025, doi: 10.61091/jcmcc127b-007.

[12] B. Fahad Mon, A. Wasfi, M. Hayajneh, A. Slim, and N. Abu Ali, "Reinforcement Learning in Education: A Literature Review," *Informatics*, vol. 10, no. 3, p. 74, Sept. 2023, doi: 10.3390/informatics10030074.

[13] A. Singla, A. N. Rafferty, G. Radanovic, and N. T. Heffernan, "Reinforcement Learning for Education: Opportunities and Challenges," July 15, 2021, arXiv: arXiv:2107.08828, doi: 10.48550/arXiv.2107.08828.

[14] X. Li, H. Xu, J. Zhang, and H. H. Chang, "Deep reinforcement learning for adaptive learning systems," *J. Educ. Behav. Stat.*, vol. 48, no. 2, pp. 220–243, 2023.

[15] S. W. Sayed, A. M. Noeman, A. Abdellatif, M. Abdelrazek, M. G. Badawy, A. Hamed, and S. El-Tantawy, "AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging E-learning platform," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3303–3333, 2023.

[16] R. Han, K. Chen, and C. Tan, "Curiosity-Driven Recommendation Strategy for Adaptive Learning via Deep Reinforcement Learning," Oct. 12, 2019, arXiv: arXiv:1910.12577, doi: 10.48550/arXiv.1910.12577.

[17] H. A. El-Sabagh, "Adaptive e-learning environment based on learning styles and its impact on development students' engagement," *Int. J. Educ. Technol. High. Educ.*, vol. 18, no. 1, p. 53, Oct. 2021, doi: 10.1186/s41239-021-00289-4.

[18] W. Strielkowski, V. Grebennikova, A. Lisovskiy, G. Rakhimova, and T. Vasileva, "AI-driven adaptive learning for sustainable educational transformation," *Sustainable Development*, vol. 33, no. 2, pp. 1921–1947, 2025.

[19] F. Stasolla, A. Zullo, R. Maniglio, A. Passaro, M. Di Gioia, E. Curcio, and E. Martini, "Deep Learning and Reinforcement Learning for Assessing and Enhancing Academic Performance in University Students: A Scoping Review," *AI*, vol. 6, no. 2, p. 40, 2025.

[20] H. E. Sari, B. Tumanggor, and D. Efron, "Improving educational outcomes through adaptive learning systems using AI," *Int. Trans. Artif. Intell.*, vol. 3, no. 1, pp. 21–31, 2024.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

[21] C. Song, S. Y. Shin, and K. S. Shin, "Implementing the dynamic feedback-driven learning optimization framework: a machine learning approach to personalize educational pathways," *Appl. Sci.*, vol. 14, no. 2, p. 916, 2024.

[22] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, "Smart E-learning framework for personalized adaptive learning and sequential path recommendations using reinforcement learning," *IEEE Access*, vol. 11, pp. 89769–89790, 2023.

[23] R. Mustapha, S. Soukaina, Q. Mohammed, and A. Es-Sâadia, "Towards an adaptive e-learning system based on deep learner profile, machine learning approach, and reinforcement learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, 2023.

[24] F. Rahioui, M. El Ghzaoui, M. A. T. Jouti, M. O. Jamil, and H. Qjidaa, "Machine Learning with Reinforcement for Optimal and Adaptive Learning," in *Int. Conf. Digital Technol. Appl.*, Jan. 2023, pp. 142–149. Cham: Springer Nature Switzerland.

[25] F. Zini, F. Le Piane, and M. Gaspari, "Adaptive cognitive training with reinforcement learning," *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 1, pp. 1–29, 2022.