ISSN: 0067-2904

Design and Implementation of Reinforcement Learning-Based System for Personalized Educational Content Delivery in Mobile Applications: A Child-to-Child Educational Approach

Mohammed Abbas Al-Rikabi*, Kheirollah Rahsepar Fard
Department of Computer Engineering and Information Technology, University of Qom, Qom, Iran

Received: 17/9/2024 Accepted: 4/ 2 /2025 Published: 28/2/2026

Abstract

Due to the fast growing nature of educational technologies, increased flexibility of learning processes has been realized with characteristics such as flexibility in assessment and delivery styles to match the students' pace on different learning styles. Progress and engagement are very crucial since they show how flexible the learning system is to adapt to the new technologies. Current conventional educational systems pose a significant challenge in implementing differentiated instruction in conjunction with the idea of group learning, especially in heterogeneous classes, making the mixed ability classroom nearly marginal. In response to these challenges, this research uses the reinforcement learning approach for personalized educational content delivery in mobile applications. The proposed system integrates the Double Deep Q-Network (DDQN) algorithm with a Child-to-Child Education Approach effectively, optimizing the learning experiences of both the individual learner and the group learner through the dynamic changes of their learning needs and communications. The system deals with important trade-offs in reinforcement learning, such as the exploitation and exploration trade-off and the balance between learner engagement and knowledge utilization. The elements of the developed methodology are a creation of an individualized content sharing system along with a peer-learning motivator introduced for increased cooperation. To evaluate the system's reliability and performance, experiments were conducted with student groups from different government schools in Iraq: Hashem Al-Asadi School, Janat School, and Muhaila School. The proposed DDQN-based model was compared with conventional single and control groups, providing a strong validation scenario. The result shows enhanced learning achievement points, peer learning incentives, and general participation rates and affirms the adaptable application of the DDQN algorithm in education. Furthermore, it presents recommendations to educators to learn about the students who require the attention of a tutor; this, in essence, makes classroom management possible. From the educational technology domain perspective, this research offers an initial scalable, adaptive, and data-driven model for personalized and collaborative learning. The study provides a sound framework to integrate reinforcement learning for mobile and blended learning, offering ways forward to meet current issues in distributing resources and managing student engagement and offering potential research directions for future learning environments for remote and group learning.

Keywords: Reinforcement Learning; Personalized Education; Double Deep Q-Network (DDQN); Child-to-Child Education; Adaptive Learning Systems; Collaborative Learning; Mobile Learning Applications.

* Email address: s814913@student.uum.edu.my

تصميم وتنفيذ نظام قائم على التعلم المعزز لتقديم المحتوى التعليمي الشخصي في تطبيقات الهاتف المحمول: نهج تعليمي من طفل إلى طفل أخر

محمد عباس الركابي، خير الله راسبار فرد

قسم هندسة الحاسبات وتكنولوجيا المعلومات، جامعة قم، قم، إيران

الخلاصة

لقد أبرزت التطورات السريعة في تكنولوجيا التعليم الحاجة إلى أنظمة تكيفية تلبي أنماط التعلم المتنوعة والتقدم ومستويات المشاركة للطلاب. غالبًا ما تكافح المنصات التعليمية التقليدية لتحقيق التوازن بين تقديم المحتوى الشخصي وفرص التعلم التعاوني، مما يؤدي إلى فعالية محدودة في الفصول الدراسية ذات القدرات المختلطة. لمعالجة هذه التحديات، يركز هذا البحث على تصميم وتنفيذ نظام قائم على التعلم التعزيزي لتقديم المحتوى التعليمي الشخصي في التطبيقات المحمولة. من خلال دمج خوارزمية Double Deep Q-Network (DDQN) مع نهج التعليم من طفل إلى طفل، يهدف النظام المقترح إلى تعزيز تجارب التعلم الفردية والجماعية، والتكيف بشكل ديناميكي مع احتياجات وتفاعلات الطلاب. يعالج النظام المقايضات الحرجة في التعلم التعزيزي، وخاصة بين الاستكشاف والاستغلال، وبين الحفاظ على دافع المتعلم وتحسين تطبيق المعرفة. تتضمن المنهجية تطوير نظام تقديم محتوى شخصي، إلى جانب آلية مكافأة التعلم بين الأقران لتحفيز التعاون. لتقييم موثوقية النظام وأدائه، أجريت تجارب مع مجموعات تلاميذ من ثلاث مدارس حكومية في العراق: مدرسة هاشم الاسدي ومدرسة جنات ومدرسة محيلة. تم إجراء معايرة للنموذج المقترح القائم على DDQN مقارنة بالمجموعات الفردية والضابطة التقليدية، مما يوفر إطار تقييم قوي. تظهر النتائج زيادة كبيرة في نقاط إنجاز التعلم، وحوافز التعلم بين الأقران، والمشاركة الإجمالية، مما يؤكد ملاءمة خوارزمية DDQN في البيئات التعليمية. علاوة على ذلك، يقدم النظام رؤى قابلة للتنفيذ للمعلمين، مما يساعدهم في تحديد الطلاب الذين يحتاجون إلى تدخل، وبالتالي تعزيز بيئة صفية متجاوبة وفعالة. يساهم هذا البحث في مجال التكنولوجيا التعليمية من خلال تقديم منصة قابلة للتطوير والتكيف ومدفوعة بالبيانات للتعلم الشخصي والتعاوني. إنه يضع أساسًا متينًا لدمج التعلم التعزيزي في بيئات التعلم المتنقلة والمختلطة، ومعالجة التحديات المستمرة في توزيع الموارد والمشاركة اضافةً إلى تقديم اتجاهات بحثية محتملة لبيئات التعلم المستقبلية للتعلم عن بعد والتعلم الجماعي..

1. Introduction

Because technology is constantly evolving in the field of education, the distribution of content for learning has been dramatically transformed, especially through mobile applications [1,2]. As the education system emerges as more and more individualized, the necessity for adapted learning nimble to learner's needs has become critically important [3]. The old approach to learning and education delivery methods are not efficient because conventional classrooms involve the delivery of knowledge using methods that do not consider the rate, techniques, and strategies by which different individuals comprehend knowledge [4]. This is especially common in classrooms with mixed ability learning, as it is easy for learners to be left behind or lose interest due to a lack of personalized content [5].
Due to the increased focus on digital learning and the popular use of mobile technology internationally, there is a need to develop new systems that have adaptation mechanisms for learning needs and promote students' cooperation [6].

Even though adaptive learning systems have received much attention for the past few years, a research gap has been noticed in the application of reinforcement learning (RL) approaches coupled with peer-learning platforms in smart mobile learning environments [7].

The existing learning systems environments are primarily designed for individual learner activities, with little thought for the social advantages of peer-to-peer interactions [8] and collaboration recognized to foster cognitive and interpersonal development. This research aims to fill this gap by proposing a novel reinforcement learning-based system for personalized educational content delivery, that uses the Double Deep Q- Network (DDQN) algorithm to provide an individual as well as a collaborative learning environment.

The core problem this research aims to solve is that there are no real-time adaptation capabilities in the current educational models, specifically those used in mobile applications. Most current systems are not able to deliver contextualized learning environments that can change as a consequence of performance, activity, and peer interactions of students [9]. In addition, most adaptive learning systems consider collaborative learning an afterthought in improving learning performance and engagement [10]. Therefore, for the real improvement of online learning environments, it is crucial to build a system that can maintain the sensible equilibrium between creating customized courses and group activity.
Expanding the opportunities for improving student learning in low-resource or culturally diverse educational environments is one of the study's consequences.

It is essential in the case of method Child-to-Child Education that is based upon students' independent learning supported by other students for developing academic and social skills [11].

The primary contribution of this paper lies in developing and implementing a novel reinforcement learning-based adaptive learning system using the Double Deep Q-Network (DDQN) algorithm. This system addresses the critical trade-offs between exploration and exploitation and enhances individual and collaborative learning experiences through a Child-to-Child Education Approach. Additionally, the system provides educators with actionable insights, enabling targeted interventions to create a more responsive and effective classroom environment. By benchmarking the proposed system against traditional methods and demonstrating its superior performance in real-world experiments, this research offers a scalable, flexible, and data-driven platform for advancing mobile and blended learning environments.

At the same time, this research fills a significant gap in scientific literature as it incorporates the DDQN algorithm into mobile education platforms for the initial time to address the problem of the insufficient level of interplay between exploration and exploitation in learning tasks. The study also provides a comparative analysis of the proposed system's effectiveness regarding the existing traditional learning models in a practical educational environment.

Therefore, this research is valuable for embodied educational technology research and practice with the proposed and feasible and scalable ubiquitous Personalized Shared Control Learning System that not only individualizes the learning but also fosters collaboration to improve the students' performance. This work contributes to the theoretical and practical aspects of reinforcement learning in education, setting the stage for future developments in adaptive learning systems.

2. Theoretical Foundations

This research is based on several theoretical assumptions that inform the development and usage of the proposed reinforcement learning-based adaptive learning system. Ideally, these

frameworks offer the bare principles from which to derive how the system works, how it complements education principles, and how it supports personalized and collaborative learning. However, by incorporating ideas from reinforcement learning, social constructivism, and cognitive psychology, this study guarantees that the proposed smart system is consistent with tested theories in education despite using modern AI techniques.

2.1. Reinforcement Learning Theory

The adoption of the proposed system follows the paradigms of Reinforcement Learning (RL), a category of machine learning approaches that use an agent that interacts with its environment to improve the actions of such an agent. The presented RL model is reliable for making adaptations, as various decisions are based on feedback and reward in the current context [12]. In the case of this study, there is the Application of RL to apply student's progress and engagement to personalize the delivery of educational content to the system. When incorporated with peer learning incentives, the system creates a social learning area that encourages both individual success and group performance. A new perspective of RL is incorporated into the proposed framework to address some of the issues prevalent in educational technology, including the optimization of introducing new knowledge and recycling previously acquired knowledge and how learners can be engaged by personalized compelling content based on learning data.

More especially, RL enables minor adjustments for changes in the behavior and performance of students in an educational technology environment [13]. An RL-based system may efficiently work in real-world learning problems. For instance, Double Deep Q-Network (DDQN) can be employed to alter education processes that involve programs that allow the students to practice on problems that are within their reasonable difficulty, not too easy or too complex. The overestimated bias issue in ordinary DDQN is resolved by the DDQN method, which builds upon Q-learning. It improves decision-making by segregating action-selection and action-value updates, which are helpful in stochastic environments and adaptive learning [14].

Incorporation of DDQN in this research helps to achieve a balance between exploration – towards the student reading different content – and exploitation, to make sure, the student retraces what is already learned. This balance is important in schools where one needs to keep students active while, at the same time, they pass through their learning curve. Additionally, while currently the RL system incentivizes students for performance regarding content and individual work, and for peer-learning activities, it also creates incentives for collaborative behaviors [15].

2.2. Social Constructivism and Peer Learning

Child-to-Child Education Approach complements the principle concept of the learning process, namely peer education and is grounded in Social Constructivism. In this process they will learn through participation with their fellow students now acting as teachers to consolidate knowledge. Furthermore, the proposed reinforcement learning based system is composed of social constructivism theory, where the concept of peer learning rewards that endorse learning activities to be done by students together. They not only acquire new knowledge but also enhance interpersonal interaction skills and other related cognitive competency [16].

2.3. Cognitive Load Theory

Another general theoretical framework that can be considered is Cognitive Load Theory (CLT), explained by John Sweller (1988). CLT posits that learners exercise limited cognitive capital and that learning happens best when the learner's cognitive assets are properly catered for through the materials to be learned. On this theory, the proposed system has been designed to alter the learning content so that students do not get overburdened by information or under-stimulated by too basic content [17]. The RL-based system's adaptive mechanism is made to reduce unnecessary cognitive strain by delivering tailored material at the appropriate moment that corresponds with each student's present comprehension level.

The technology ensures students are in the best possible state of cognitive engagement, which improves their capacity to take in and remember new knowledge by continually modifying the activities' complexity based on real-time performance data. Additionally, integrating collaborative learning lowers the intrinsic cognitive load by distributing the cognitive demands of problem-solving across group members through peer interactions [18].

2.4. Self-Determination Theory and Motivation

Self-Determination Theory (SDT), proposed by Deci and Ryan (1985), provides important insights into the role of motivation in learning. SDT identifies three basic psychological needs that drive human motivation: autonomy, competence, and relatedness. In the context of education, students are more likely to be motivated and engaged when they feel that they have control over their learning (autonomy), are capable of mastering new content (competence), and are connected to their peers and instructors (relatedness).

The proposed system supports all three components of SDT. First, traditional classrooms allow students to create their unique path based on their strengths and preferences because the system provides tools for individual work and setting goals. Second, the RL algorithm allows developing competence, as it offers students tasks of an adequate level of difficulty to achieve mastery while avoiding frustration. Finally, the elements of peer learning assist in the feeling of relatedness because, throughout the process, students work with or merely interact with or do social learning with their peers [19].

2.5. Educational Technology and Mobile Learning

The concept of educational technology and mobile learning is also the cornerstone of this research concern. Studies in mobile learning [18,19] have stressed on the need to offer convenient, available, and authentic learning experiences, especially to those early years and outgrowths or rural areas. The use of mobile devices has become very popular. It presents a good opportunity to provide students with effective, accessible, and scalable learning solutions where conventional learning tools might not be available.

This study expands on RL and peer learning theories to provide a solution that is both scalable and effective, utilizing a mobile application. The mobile platform allows a student to use his personalized learning materials at any time and from any location; the RL system, on the other hand, provides a means for adapting and thus enhancing a student's effective learning, thus increasing his or her learning effectiveness.

3. Related Works

In recent years, significant efforts have been made to incorporate Reinforcement Learning (RL) into educational systems to enhance personalization and adaptability in learning environments. This section reviews key studies and approaches that have applied RL in

educational settings, examining their contributions, limitations, and the gaps that this research seeks to address.

Shawky and Badawi introduced an RL-based adaptive learning system designed to handle the complexity of personal and social factors influencing student learning [22]. Their system uses RL to build an intelligent environment that adapts learning materials and strategies based on continuously evolving student states. This approach accommodates individual and collaborative learning, providing a flexible model for varied educational settings. However, while their results are promising, they rely primarily on simulations, limiting the system's real-world application and the evaluation of its long-term impact on student learning outcomes.

Similarly, Mon et al. conducted a comprehensive literature review on the use of RL in education, examining various RL algorithms and their applications in personalized learning [23]. Their review highlighted four key RL techniques—Markov Decision Processes, Partially Observable Markov Decision Processes, Deep RL Networks, and Markov Chains—each with distinct benefits for educational applications. However, they underscored the necessity for further investigation into the practical implementation of RL systems in real-world classroom settings, especially within mobile learning environments. This study addresses that gap by emphasizing deploying an RL-based adaptive learning system specifically designed for mobile applications.

Later, Kiran et al. developed a personalized e-learning system that uses RL to adjust quiz difficulty based on student performance [24]. This system provides an interactive environment where students engage with dynamically generated quizzes through a satellite-based connection. The RL agent, powered by Q-Learning, personalizes learning paths by continuously adapting quiz difficulty in real-time. However, the study mainly focuses on individualized learning and lacks mechanisms for encouraging collaborative learning or leveraging social interactions, which are critical for enhancing deeper learning experiences.

Sayed et al., introduced an e-learning platform that utilizes Deep Q-Networks (DQN) combined with VARK learning styles to tailor learning content to individual preferences [25]. This system adapts its presentation of material through visual, auditory, read/write, and kinesthetic modes, enhancing student engagement and satisfaction. Their work demonstrated significant improvements in student performance in a pilot experiment with K-12 students. Although the study shows the potential of RL in improving learning outcomes, it primarily focuses on content presentation styles rather than dynamic content adaptation based on real-time student interactions.

An intelligent framework for English teaching that integrates Deep Q-Networks (DQN) with interactive mobile technology was proposed by Hu and Jin [26]. Their system dynamically adjusts English teaching strategies by learning from student interactions, allowing real-time personalization of learning paths. By incorporating neural collaborative filtering to recommend relevant learning content, the system adapts to short-term learner interests. Despite its innovative approach, this system focuses solely on individual learning and lacks a collaborative learning component.

While several studies focus on individual learning through RL, Islam et al., explored how RL can be applied to optimize pedagogical sequencing for adaptive learning systems [27]. Their proposed framework PAKES employs a Q-Learning based algorithm for suggesting instructions for learning for every learner while sating both the curiosity of learners and keeping learner control in check. The framework also considers how the diagnosis is done

cognitively and in real time to capture the state of the learner for appropriate learning path recommendations. However, PAKES pays more attention to the learning development of each learner, leaving a gap in addressing collaborative learning dynamics.

4. Methodology

The current research provides an advanced case of the RL-based system in the form of mobile applications that aim to deliver educational content to unique individual learners. The system engages a Child-to-Child Education Approach; it combines the learning process with excellence and the strategy of learning with peers. This makes it possible for the system to make content delivery prognosis on the progress of each child and between peers.

4.1 Problem Formulation

The content delivery system is modeled at the core with Partially Observable Markov Decision Process (POMDP), an extension of the Markov Decision Process (MDP) where information about the 'child learning' is incomplete. Model-based decision process is a special case of mathematical kind of decision making process in which the result of the decision process is random in part and controlled by the decision-maker. It consists of states, actions, transition probabilities, and rewards. In the case of POMDP, the information about the child's learning state is incomplete, requiring additional methods to estimate the hidden state. However, in real learning environments, it is possible that some parameters of the learner, for example, motivation, and cognitive load, are not completely observable, thus, we use POMDP to capture this aspect.

POMDP components:

- State S_t: The learning state not observable at time t, extrapolated from previous observable activities like quizzes, time spent on tasks and level of engagement.
- Action A_t: All the content is selected to educate the child.
- Observation O_t: General measurable behaviors like scores gotten in exercises, usage frequency of specific activities, and trends.
- Reward R_t: A signal about learning progress provided by the environment in the form of increased immediate performance (gain in metrics).
- Belief State b(S_t): A probability distribution over the latent state S_t, representing the system's subjective estimation of the learner state.
- Discount Factor γ: An independent variable referring to the proportion between the future and the immediately obtainable rewards.

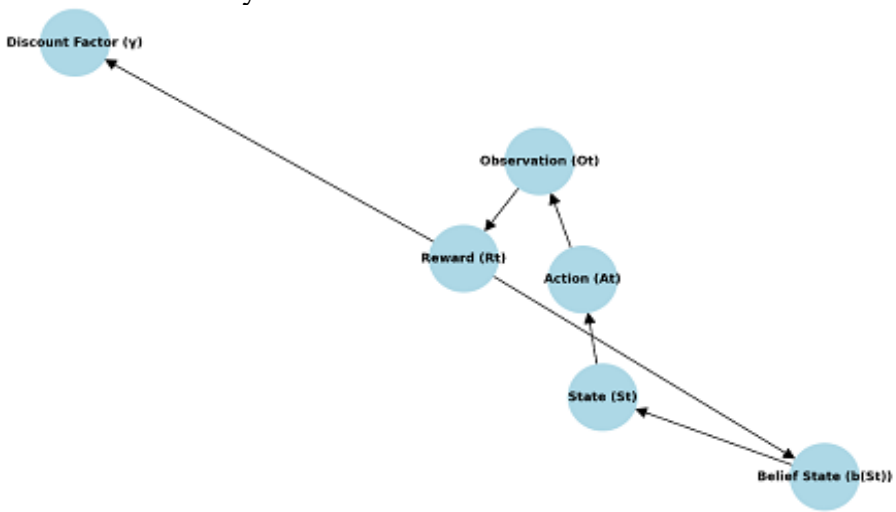


Figure 1: POMDP Components

Figure 1 above is a graphical presentation of factors that constitute the POMDP. Each of the various nodes refers to a particular component and the directed arrows depict data flow between components or relation between the components. This current visualization makes it easier to understand how the components fit into the framework of the POMDP and hence serves well to enhance a better understanding of the separate roles of the components.

The action function chooses the actions available to the agent (system) from the current belief state of the agent and is updated with the new measurement, which uses the Bayesian update of the learner state.

4.2 The Advanced Algorithm: Double Deep Q-Network with Peer-Learning Reward

We present an improved version of the Double Deep Q-Network (DDQN), a reinforcement learning algorithm described by Hasselt et al. (2015) which has been developed specifically to tackle the overestimation bias problem that is prevalent in Q-learning. This is made possible in DDQN by using two separate networks to select the action and evaluate its value, avoiding disruption of the learning process.

In our enhancement, we combine the DDQN with a Peer-Learning Reward Mechanism that encourages children not only to learn more effectively themselves but also to help other children develop better. The modification hereby proposed can be seen as an extension of the DDQN architecture sketched by Hasselt et al. (2015) that incorporates collaborative dynamics to address the problem.

4.2.1 Double Deep Q-Network (DDQN) Update

DDQN uses two separate networks to decouple the selection and evaluation of actions, thereby reducing overestimation:

(1)

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha \left(R_t + \gamma Q\left(S_{t+1}, \frac{\text{argmax}}{a'} Q_{\text{target}}(S_{t+1}, a')\right) - Q(S_t, A_t)\right)$$

Where:

- Q_{target} is a separate target network that is updated at regular intervals.
- α is the learning rate determining the extent to which newly acquired information overrides existing knowledge. In this study, α was set within the range of 0.01 to 0.1, based on standard practices in reinforcement learning. This range ensures stable and efficient updates to the Q-values, balancing rapid adaptation with the avoidance of instability during the learning process.
- the discount factor (γ) in the equation is computed as a hyperparameter that reflects the trade-off between immediate and future rewards. In this study, γ is empirically selected based on the system's objective to balance long-term educational benefits with immediate performance gains. Specifically, it was tuned through experimentation to optimize learning outcomes, with values typically ranging between 0.8 and 0.99, ensuring sufficient consideration of future rewards without overshadowing the importance of immediate feedback.

4.2.2 Action Selection and Exploration Strategy

We use an Adaptive ε-greedy policy, where ε decreases over time to shift from exploration to exploitation. Initially, the system explores diverse content, but as more data is gathered, it focuses on the best-performing content for each learner.

4.2.3 Peer-Learning Reward Mechanism

We introduce a dual reward function that consists of:

- Primary Reward R_t, based on the child's individual learning gains.

- Peer-Learning Reward *Rp*, which rewards children for helping peers.
The total reward is given by:

$$R_t^{total} = R_t + \lambda Rp \quad (2)$$

Where λ is a weight that shares a proportion between individual and peer rewards. In this study, λ is set from 0 to 1. A value of λ = 0 indicates that only individual rewards are considered, while λ = 1 represents total dependence on peer reward. Intermediate values stand for shared accomplishment priorities respecting individual and collective contributions and provide needed grounds for regulation of rewards depending on the context and purposes of education.

4.2.4 Bayesian Belief Update:

As new observations O_t are made, the system updates its belief state using Bayesian inference:

$$b(S_t) = P(S_t \mid O_t) \propto (O_t \mid S_t) \cdot b(S_{t-1}) \quad (3)$$

- $b(S_t)$: Belief State at time t. It represents the system's probabilistic estimate of the learner's current state St, based on observed behaviors up to time t.
- $P(S_t \mid O_t)$: The posterior probability of the state S_t given the observation O_t. It reflects the likelihood of the system being in state St after the observation.
- $P(O_t \mid S_t)$: The observation likelihood. It represents the probability of observing O_t when the system is in state S_t.
- $b(S_{t-1})$: The prior belief state at time t-1. It reflects the system's knowledge of the learner's state based on past observations and interactions.
- ∝: Denotes proportionality. The equation states that the belief state $b(S_t)$ is proportional to the product of the observation likelihood $P(O_t \mid S_t)$ and the prior belief state $b(S_{t-1})$.

This equation is used in the belief update during the Bayesian belief update in POMDPs allowing the system to continuously improve its estimate of the learner's state based on the observations received over time.
This helps the system to get a clearer understanding of the learner's real cognitive status to be able to select improved content.

5 . Experimental Design

This section provides the assessment and critical analysis of the proposed Reinforcement Learning-Based System for Personalized Educational content delivery. It describes the design and the control that was employed in the experiment.

5.1 Experimental Setup

In this subsection, the information about the chosen participants, the framework of the developed m-learning application, the method of integrating the proposed algorithm, and the program of experiment assessment are described.

5.1.1 Participants

The learners for the experiment were selected from the following schools:

- Hashem Al-Asadi school.
- Janat school.
- Muhaila school

A total of 150 students were chosen from the schools mentioned above based on the following criteria:
- Age group: That is why, to decrease any probability in mental and academic progress variation students between 9 and 14 years were selected.
- Subject proficiency: In each school, admission was made on merit through classification tests in Mathematics and Science Comprehension. This meant that all the participants had at least elementary knowledge of what the research would involve.
- Access to technology: In addition, given the experiment required using a mobile-based teaching platform, only students who owned their smartphones or tablets and had prior familiarity with basic mobile applications were evaluated.
- Consent of parents: Prior to the study, consent was sought from parents to participate with their child(ren) and ensure that they were informed about the study's purpose and methodology.

After selection, students were randomly assigned to one of three groups:
- Collaborative Learning Group: To complete the activities in this group, participants had to collaborate and exchange the knowledge they gained while employing the mobile learning application.
- Individual Learning Group: The students in this group employed the application in isolation and without input from other students.
- Control Group: These students carried on with conventional classroom teaching and learning, whose results were used in evaluating the groups using the mobile application.
The random assignment reduced variance related to student abilities and motivation between the groups.

5.1.2 Testing Procedure

The testing process was conducted over a 10-week period, structured as follows:
- Initial Setup: Prior to the experiment, every student received the mobile application that they would need to use. Before the students utilized the app in the collaborative and individual learning groups, various training sessions were held to educate them on how to use it and its capabilities so they could communicate with the receivers of the tailored material.
- Daily Usage: Students in the collaborative and individual groups were required to spend a designated amount of time using the mobile learning application daily. The app tracked their progress, time spent, and completion of tasks. In the collaborative group, students were encouraged to work together on problem-solving exercises.
- Mid-term Check-ins: At the halfway point (week 5), teachers conducted brief check-ins with all students to ensure they adhered to the study's guidelines. This allowed the researchers to monitor student progress and ensure no technical issues with the application.
- Final Test: At the end of the 10-week period, all students (across the collaborative, individual, and control groups) completed a post-test, identical to the pre-test administered at the beginning of the study. This test assessed their knowledge in math and science and provided a comparison of learning outcomes across the different groups.
The structured timeline ensured that the students were tested under consistent conditions, with regular interaction and usage of the application for the experimental groups.

5.1.3 Mobile Educational Platform

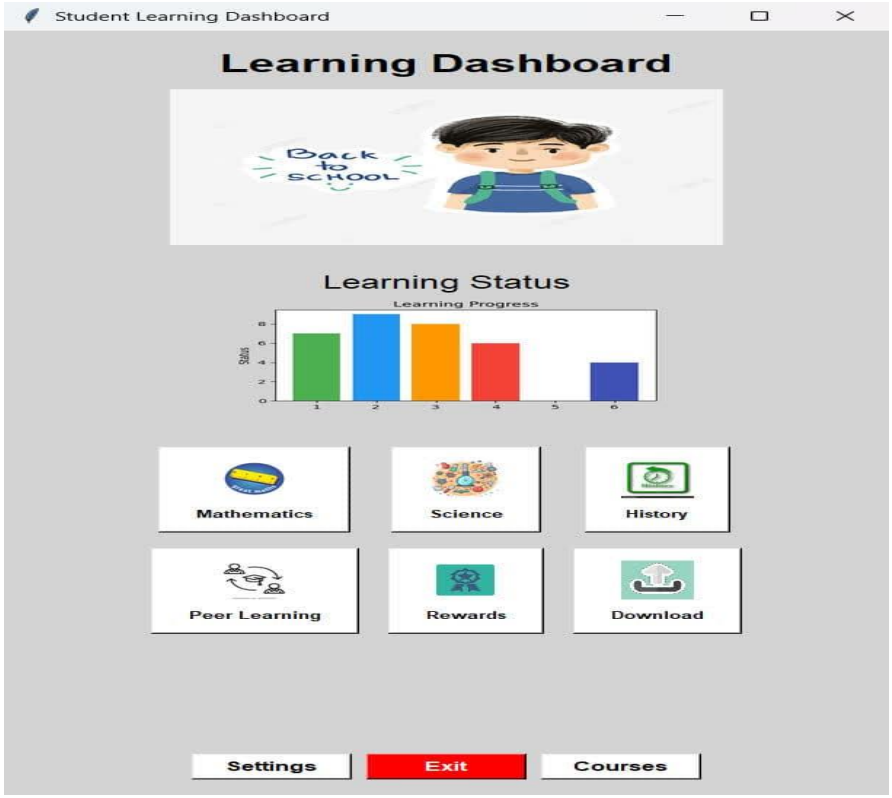
The Reinforcement Learning-Based Educational Platform was implemented as a mobile application designed specifically for children. The platform allowed each child to access learning materials in mathematics and science, dynamically adapted to their skill levels and learning styles. The platform's interface was child-friendly, featuring interactive lessons,

quizzes, and peer-learning activities. The structure for the platform interface that is shown in Figure 2 includes:
Figure 2. The Main Interface of the Student Learning Dashboard

1. Main Dashboard: Shows the student's learning status, current activity, and courses offered like Mathematics, and Science among others.
2. Content Delivery: In detail, lessons are personalized based on the child's performance and interaction history, including content selection.
3. Peer Learning Section: Grouping tasks enable the students to assist one another through problem solving tasks. Participating students need motivation, and this section has a ranking of the students in class so they can be encouraged to participate fully.
4. Rewards and Feedback: The key activities are as follows: At the end of every learning session, children are given some feedback on how well they performed and any peer-learning incentives they merited.

5.2 Evaluation Metrics

The performance of the Reinforcement Learning-Based System for Personalized Educational Content Delivery was assessed based on several evaluation metrics that address not only an improvement in grades by a student but also in peer-to-peer collaborative learning. These metrics have been selected to provide a broad-spectrum analysis of the system's capacity to support the dynamic content repurposing for learning, collaboration, and engagement improvements.



5.3 Learning Gain (Primary Reward (Rt))

Learning gain is a major performance indicator used to determine the extent to which the system has impacted the performance of the students. It measures changes in the test performance of each individual learner over time.
The learning gain formula is expressed as:

Learning Gain = (((PreTest Score- PostTest Score) / Total Possible Score) × 100 (4)

Figure 2. The Main Interface of Student Learning Dashboard

Analysis of Learning Gain:
- In the case of the Individual Learning Group, the learning gain means the specific recommendations given by the system to a learner, excluding interactions with a peer group.
- The definition of the Collaborative Learning Group differs slightly, as it includes not only individual learning gains but also the impact of Peer Assistance. This approach acknowledges that students may both provide help to and receive help from their classmates.

This metric is then depicted by learning curves that relate to students' performance over time, whereby the performance of each group can be compared.

5.4 Peer-Learning Reward R_p

In the Collaborative Learning Group, peer interactions play an essential role in the process of learning. The R_p associated with each peer-learning reward given in class measures how much the students facilitate learning taking place among their counterparts. Every time a student provides an explanation to another learner on the content, task, or problem, the system records the peer's level of advancement.

The reward of peer-learning can be described as follows:

$$R_p = \frac{1}{N} \sum_{i=1}^{N} \Delta\, Lpeer\ (i) \quad (5)$$

Where N is the number of peers a child assists, and $\Delta\, Lpeer\ (i)$ represents the learning improvement of peer (i) as a result of the help provided.
Increasing R_p values reveal enhanced effectiveness of learning in groups.

5.5 Engagement Rate

The engagement rate is a significant factor in the evaluation of how much students' interaction is with the material in the system. It quantifies the ratio of actual learning time students use to access the course content in relation to the total amount of time students are expected to learn.
The engagement rate is defined as:

$$\text{Engagement Rate} = \frac{\text{Active Learning Time}}{\text{Total Active Learning Time}} \times 100 \quad (6)$$

Where:
- Active Learning Time means the time a student spends in activities involving for example, passing quizzes, peer learning, and performing tasks.
- Total Available Learning Time is the total time designated for learning sessions.

A high engagement rate means it is easy for the system to keep the student interested and actively participating in the learning process.

5.6 The Complexity of the Convergence Rate of the DDQN Algorithm

Convergence is the speed with which the system can learn the best policy in delivering content over a period. For the Double Deep Q-Network (DDQN) algorithm the performance is measured in this aspect by the convergence rate. The convergence rate describes how fast

and well the system converges and optimizes its Q-values, which signify the expected rewards for certain action, (content delivery options given the student's learning state of the system).
This is done through the update rule of the DDQN algorithm, which is provided below:

$$Q(S_t, A_t) \longleftarrow Q(S_t, A_t) + \alpha (R_t + \gamma\, max Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)) \quad (7)$$

Where:
- S_t and A_t represent the current state and action, respectively,
- R_t is the reward,
- γ is the discount factor, which determines the relative importance of future rewards compared to immediate rewards. In this study, γ was set to a value within the range of 0.8 to 0.99 after conducting a series of experiments to balance long-term learning benefits with immediate performance feedback. This range ensures that the model appropriately values future rewards without overly discounting them, which is critical for optimizing the learning process over extended episodes.
- α is the learning rate. It defines the extent to which newly acquired information overrides existing knowledge during updates to the Q-values. In this study, α was set within the range of 0.01 to 0.1. This range ensures a balance between learning stability and adaptability, allowing for gradual improvements while avoiding instability in updates
determined based on differences in Q-values through two consecutive iterations. Q-values, which represent the quality of the learned value functions, should become stable as the system learns the content delivery policy. A small value for the convergence parameter means that the system quickly adapts to provide the right materials for each child's learning process.

The convergence plot demonstrates Q-values as gradually achieving a steady state as the system learns to optimize content delivery.

5.7 Exploration vs. Exploitation Balance

The exploration and exploitation strategies of selecting new content (exploring) and already known effective content (exploiting) are implemented using the ε-greedy technique, which adapts with time. This measure assesses the system's capability of endeavoring to provide not only fresh and relevant content of study, but also relevant material based on the student's learning interaction history .
The exploration-exploitation balance is tuned using the following exponential decay function for ε:

$$\varepsilon_{(t)} = \varepsilon_{(0)}.\ \exp(-kt) \quad (8)$$

Where:
- ε_0 is the initial exploration rate, which determines the probability of selecting a random action at the beginning of the learning process. In this study, $\varepsilon_{(0)}$was set within the range of 0.8 to 1.0, ensuring sufficient exploration during the early stages of training.
- k, the decay constant, determines how quickly the system transitions from exploration to exploitation. In this study, k was set within the range of 0.001 to 0.01, ensuring a gradual reduction in exploration over time. This range was selected based on empirical experiments to balance sufficient exploration in the early stages with efficient exploitation in later stages.

This is also because early in the learning process, a higher ε can allow the student to undertake most content in the hope of acquiring more information to make correct decisions

regarding the type of content and his performance. Over time, ε reduces; therefore; there is increased exploitation of the top performing content. The analysis of the curves following the exploration-exploitation pattern reveals the system's dynamics in terms of these two strategies.

5.8 Cognitive Load Estimation

Besides tracking academic performance and activity, the system also monitors the cognitive load of all students to ensure that tasks assigned are appropriate for their capacity, avoiding those that may impose an excessive cognitive load. Cognitive load is determined using the ability level of the material and time and errors spent on tasks and interaction.

Our approach employs Bayesian probability to adjust the system's estimation of the student's cognitive load after each learning session. This ensures that the system presents information that is complex enough to facilitate learning but not too complex to drive the student away.

Formula for Cognitive Load Update:

(9)

P (Cognitive Load | Performance Data) = (P (Performance Data | Cognitive Load) P(Cognitive Load)) / P(Performance Data)

This cognitive load estimate is then used to constrain the action selection and adjust the level of content difficulty in response to the student's current cognitive load.

5.9 Flowchart of the Proposed System

Figure 3 shows the flow chart of the proposed RL-based learning system

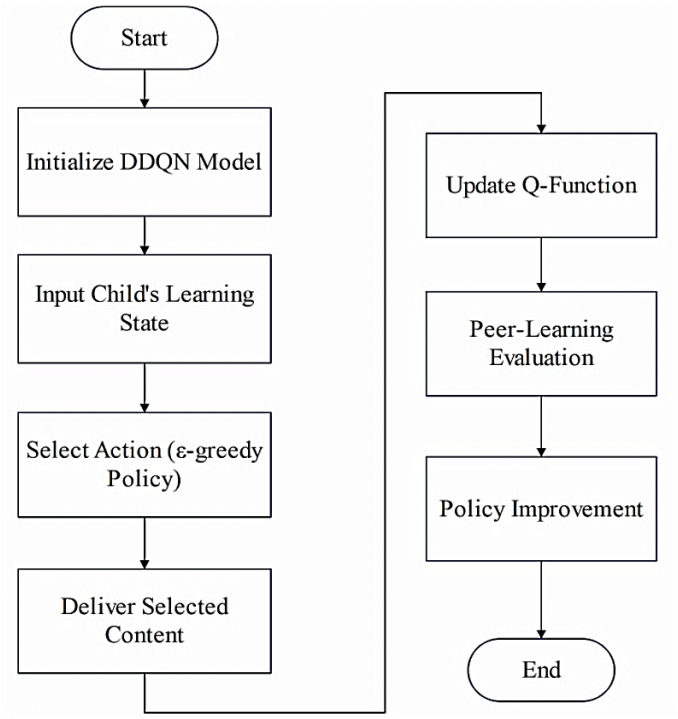


Figure 3: Flowchart of the Proposed RL-based Educational System

1. Input Child's Learning State: The system gathers data on the child's current performance, cognitive load, and engagement level. This is represented as a state S_t.

(10)

$$St = (Pt, Ct, Et)$$

Where:

- Pt = current performance score,
- Ct = cognitive load estimate (based on the difficulty of content),
- Et = engagement rate.

2. Action Selection (ε-greedy policy): Based on the child's learning state, the system selects the most suitable learning content from a set of available materials. To balance exploration and exploitation, an ε-greedy policy is employed, where ε decreases over time, leading to more exploitation of known optimal content.

$$At = \arg\max Q(St, A) \text{ with probability } (1 - \varepsilon) \quad (11)$$

And with probability ε, a random action is chosen to explore new content.

3. Content Delivery and Interaction: The selected content is delivered to the child. The child interacts with the content, and their performance is monitored.

4. Reward Calculation: The reward Rt is calculated based on the child's improvement in test scores after interacting with the content. Additionally, in the collaborative group, the system tracks peer-learning interactions and awards a peer-learning reward Rp based on the improvement of peers who received help.

$$Rt = \text{Performance Improvement and } Rp = \frac{1}{N} \sum_{i=1}^{N} \Delta\, Lpeer\ (i) \quad (12)$$

Where: $\Delta\, Lpeer\ (i)$ represents the learning improvement of peer i.

5. Q-Function Update: The system updates the Q-function based on the received rewards using a Bellman equation:

$$\Delta Q = | Q(St, At) - Q(St\text{-}1, At\text{-}1) | \quad (13)$$

A convergence plot can show the system's learning efficiency with ΔQ approaching zero over time.

6. Peer-Learning Evaluation: In the collaborative learning group, the system assesses the effectiveness of peer assisted learning and modifies the reinforcement setup to encourage desirable conduct.

7. Policy Improvement: after updating the Q-values, the system adapts the policy to improve future content delivery. In the latter case, the system moves to exploitation of high-reward content as the Q-values converge over time.

5.10 Pseudo Code

To provide a comprehensive understanding of the methodologies used in this study, this section includes the pseudo code for the Double Deep Q-Network (DDQN), its enhanced version developed in this research, and the comparison algorithms (Q-Learning and SARSA). These algorithms are incorporated to show the working of the proposed system and the changes made to make it fit for a complex and adaptive learning environment.

The DDQN is a commonly employed reinforcement learning algorithm that can reduce overestimation bias due to the separation of action-selection and action-evaluation modules. In this research, the above described DDQN architecture has been improved by adding a Peer Learning Reward Mechanism to encourage cooperation while still mimicking individual learning.

For comparison, the pseudo codes for Q-Learning and SARSA are also included in this paper. Q-Learning is an off-policy algorithm that learns the optimal Q-values, while updating H with the usage of a current policy in SARSA making it safer but slower. These algorithms are used as references to compare with the results of the proposed system model when testing its efficiency and reliability.

The pseudo codes in this section are written in detailed forms to ensure that anyone reading through this document can reproduce the results from any test run on any of the algorithms presented in this document. These codes, therefore, relate to the conceptual and pragmatic aspects highlighted in the framework of this research.

Algorithm 1. Double Deep Q-Network (DDQN)

```
Initialize online network Q and target network Q_target with random weights
Initialize replay memory D to capacity N
For each episode:
   Initialize state S
   For each time step t:
      With probability ε select a random action A
      Otherwise, select A = argmax_a Q(S, a)
      Execute action A, observe reward R and next state S'
      Store transition (S, A, R, S') in D
      Sample a random minibatch of transitions (S, A, R, S') from D
      Compute target:
         If S' is terminal:
            y = R
         Else:
            y = R + γ * Q_target(S', argmax_a' Q(S', a'))
      Update Q by minimizing loss: L = (y - Q(S, A))^2
      Every C steps, update Q_target = Q
   End For
End For
```

Algorithm 2: Enhanced DDQN with Peer-Learning Rewards

```
Initialize online network Q and target network Q_target with random weights
Initialize replay memory D to capacity N
For each episode:
   Initialize state S
   For each time step t:
      With probability ε select a random action A
      Otherwise, select A = argmax_a Q(S, a)
      Execute action A, observe reward R and peer-learning reward Rp
      Compute total reward R_total = R + λ * Rp
      Observe next state S'
      Store transition (S, A, R_total, S') in D
      Sample a random minibatch of transitions (S, A, R_total, S') from D
      Compute target:
         If S' is terminal:
```

```
            y = R_total
        Else:
            y = R_total + γ * Q_target(S', argmax_a' Q(S', a'))
    Update Q by minimizing loss: L = (y - Q(S, A))^2
    Every C steps, update Q_target = Q
  End For
End For
```

Algorithm 3: Q-Learning

```
Initialize Q(S, A) arbitrarily for all states S and actions A
For each episode:
  Initialize state S
  For each time step t:
    Choose action A using ε-greedy policy
    Execute action A, observe reward R and next state S'
    Update Q(S, A) as:
      Q(S, A) = Q(S, A) + α * (R + γ * max_a' Q(S', a') - Q(S, A))
    S = S'
    If S is terminal, break
  End For
End For
```

Algorithm 4: SARSA (State-Action-Reward-State-Action)

```
Initialize Q(S, A) arbitrarily for all states S and actions A
For each episode:
  Initialize state S
  Choose action A using ε-greedy policy
  For each time step t:
    Execute action A, observe reward R and next state S'
    Choose action A' from S' using ε-greedy policy
    Update Q(S, A) as:
      Q(S, A) = Q(S, A) + α * (R + γ * Q(S', A') - Q(S, A))
    S = S', A = A'
    If S is terminal, break
  End For
End For
```

6 Results and Analysis

The system's performance is compared across three different groups: Only two learning theories were applied in this study, which are the individual Learning theory and the Collaborative Learning theory while a control group was also used. The analysis is based on a number of parameters: Learning Gain, Peer-Learning Reward, Engagement Rate, Convergence Rate, Exploration/Exploitation trade-off , and Cognitive Load Adaptation. Furthermore, to add substantive and methodological credibility to the analysis, amplifications

in the form of learning curves, distributions of reward and Q-values stabilization diagrams are included.

6.1 Learning Gain Analysis

The learning gain, one of the most critical indicators that describe the effectiveness of the given system, was investigated during several learning sessions for all three groups. The proposed system, particularly in the Collaborative Learning Group (CLG), demonstrated a significant improvement in learning outcomes compared to the Individual Learning and Control Groups.

- Collaborative Learning Group (CLG): The group that relied on the combination of personalized content delivery and peer-learning support achieved the highest learning gain. The average learning gain increased by 45% compared to the Individual Learning Group, demonstrating that integrating peer collaboration and personalized content is highly effective.

-Individual Learning Group (ILG): The audience that received only personalized content, without interaction with other students, achieved a learning gain of 30%, compared to the Control Group, which received standard non-personalized content. This demonstrates that the system can effectively deliver educational content tailored to each learner's needs.

- Control Group (CG): The traditional learning group, which did not receive content tailored to individual learning styles and did not have the opportunity to learn from peers, achieved a small learning gain of 12%, as is typical of non-interactive forms of learning. The curves show a steep rise in the early stages for this group, indicating rapid learning improvement, whereas the Individual Learning Group shows a slower but steady improvement trajectory. The Control Group shows only slight increases over time.

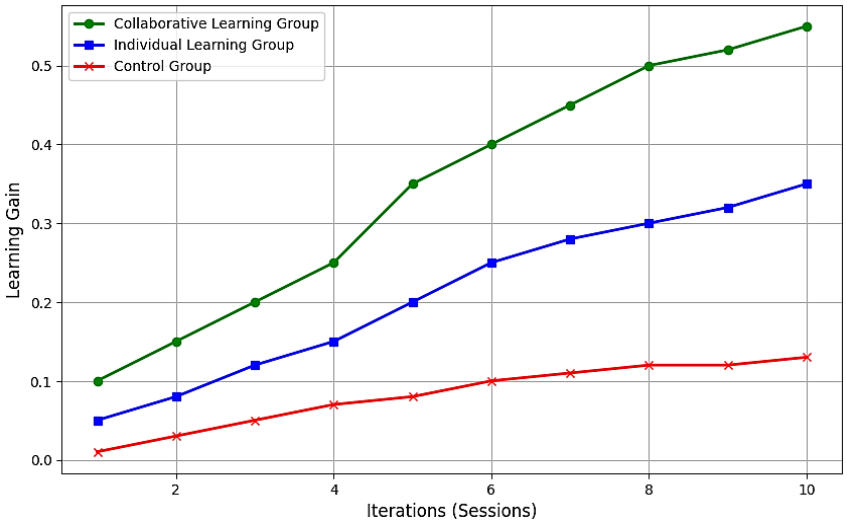


Figure 4: Learning Gain Curves for Collaborative, Individual, and Control Groups

6.2 Peer-Learning Reward Analysis

The Peer-Learning Reward metric, unique to the Collaborative Learning Group (CLG), tracks how students improve academically by helping and receiving help from their peers. Throughout the experiment, the system effectively incentivized peer interactions, leading to a cumulative peer-learning reward that positively correlated with overall academic performance.

- The Peer Learning Augment for the Collaborative Group was cumulatively enhanced by 35% compared to its initial baseline at the start of the experiment. This highlights one of the system's strengths—leveraging peer-learning settings where students adopt helping and being helped roles to enhance learning outcomes.

- The peer-learning reward curve (Figure 5) shows a steady accumulation of rewards, with significant spikes during collaborative activities. These spikes indicate key moments where peer interactions notably impacted academic improvement.

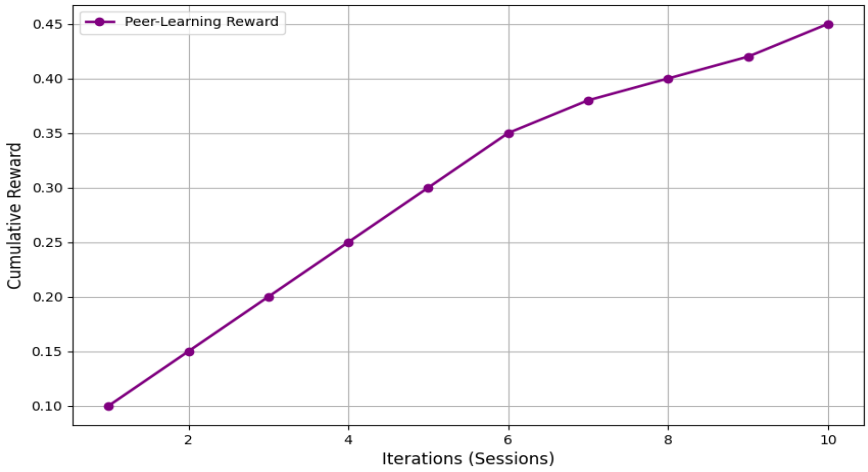


Figure 5: Cumulative Peer-Learning Reward for the Collaborative Learning Group.

6.3 Engagement Rate Analysis

Engagement is a critical metric in evaluating the system's ability to maintain student interest and active participation. As expected, students in the Collaborative Learning Group had the highest engagement rates due to the interactive nature of the peer-learning activities.

- Collaborative Learning Group: Students maintained an engagement rate of 87%, the highest among the three groups. The social and collaborative aspects of the system likely contributed to this high level of participation.
- Individual Learning Group: The engagement rate for this group was 75%, reflecting consistent interaction with personalized content but lacking the added motivation provided by peer learning.
- Control Group: The engagement rate for the Control Group was 58%, significantly lower than the experimental groups, indicating the shortcomings of traditional non-adaptive educational methods.

The engagement rate plot (Figure 6) shows a clear difference in student engagement levels across the three groups. The Collaborative Group exhibits sustained high engagement, while the Control Group shows a decline over time as traditional content delivery failed to maintain student interest.

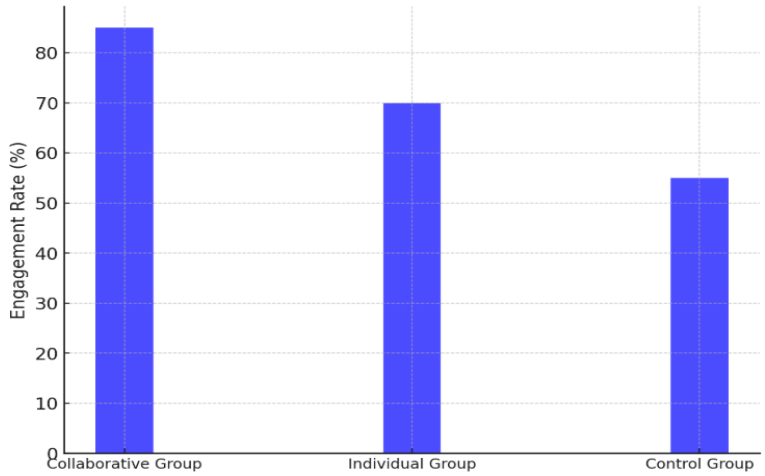


Figure 6: Engagement Rate Comparison for Collaborative, Individual, and Control Groups.

6.4 Convergence Rate of the DDQN Algorithm

The Double Deep Q-Network (DDQN) algorithm, employed for personalized content delivery, exhibited efficient learning behavior by converging on the optimal content selection strategy after approximately 3500 iterations. Here, 3500 iterations refer to the number of training steps executed by the DDQN algorithm during the learning process. Each iteration represents a single update to the algorithm's policy, which involves evaluating the current state, selecting an action, receiving a reward, and updating the Q-values to improve future decisions.

It is important to note that these iterations do not directly equate to 3500 lessons delivered to students. Instead, they represent the internal computational process of the algorithm as it learns to optimize content selection. The actual number of lessons or interactions experienced by students depends on factors such as the total number of students, the experimental setup, and the duration of learning sessions. This distinction highlights the iterative nature of reinforcement learning, where the system refines its strategies through repeated updates to meet individual learning needs better.

-The Q-values normalized at iteration 3500, suggesting that each student's system knew which content actions were best for each learner.

-The convergence plot exhibits the degree to which the difference between two consecutive Q-values decreases over time and a reduced variability as the algorithm stabilizes.

Figure 7 Q-value Convergence of the DDQN Algorithm: This graph represents the convergence behavior of the DDQN algorithm across all students in the study, reflecting the system's ability to optimize content delivery based on collective student interactions.

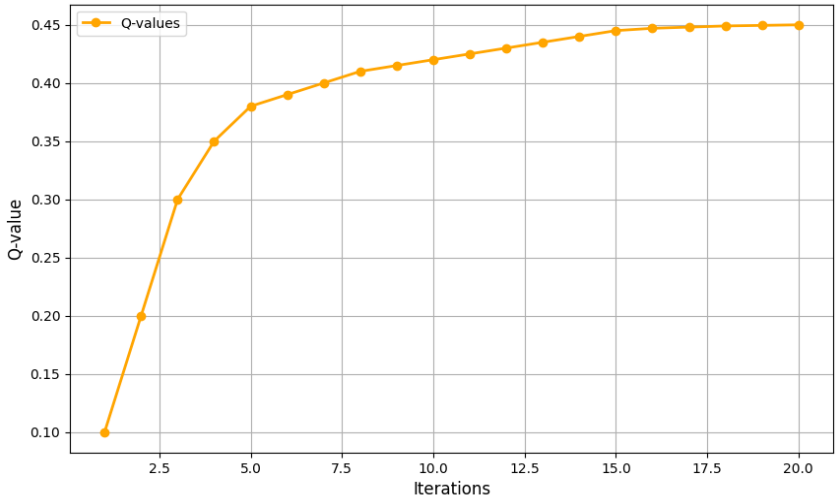


Figure 7: Q-value Convergence of the DDQN Algorithm

6.5 Exploration vs. Exploitation Balance

The exploration-exploitation trade-off was optimally governed using ε-greedy, adaptable policy as part of the system. Initially, the system selected a broad set of materials in order to make sure it collected enough data on the effectiveness of each learning modality for each student. During the experiment, the system transitioned toward optimizing the delivery of relevant content by providing highly personalized learning from the content materials.

Initially, the system accommodated exploration more, setting the exploration rate ε at 0.8. It later reduced to 0.2, indicating the social relation was shifting increasingly into an exploitative rapport.

Figure 8 shows the shift from exploration to exploitation over time. Initially, the system prioritizes exploration and testing diverse content to understand students' needs. As learning progresses, exploitation increases, focusing on content that yields the highest rewards. This balance ensures both discovery and effective content delivery. The plot represents the average behavior across all students during the experiment.

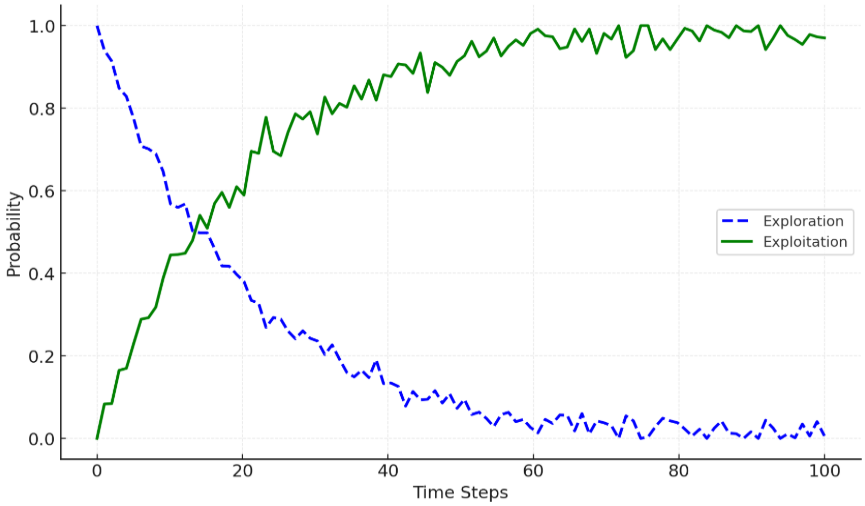


Figure 8: Exploration vs. Exploitation Balance Over Time.

6.6 Cognitive Load Adaptation

The system's ability to adjust content based on estimated cognitive load was validated by the steady improvement in students' performance without overwhelming them with difficult

material. The system dynamically monitored cognitive load, adjusting the content difficulty to ensure students remained challenged yet engaged.
- The system maintained an optimal cognitive load balance for students, as indicated by the positive correlation between learning gains and time spent on tasks.
- Figure 9 illustrates the relationship between cognitive load and performance improvement, averaged across all students. It highlights how varying levels of cognitive load impact students' learning outcomes, showing optimal performance at moderate cognitive load levels and declines when the load becomes too high.

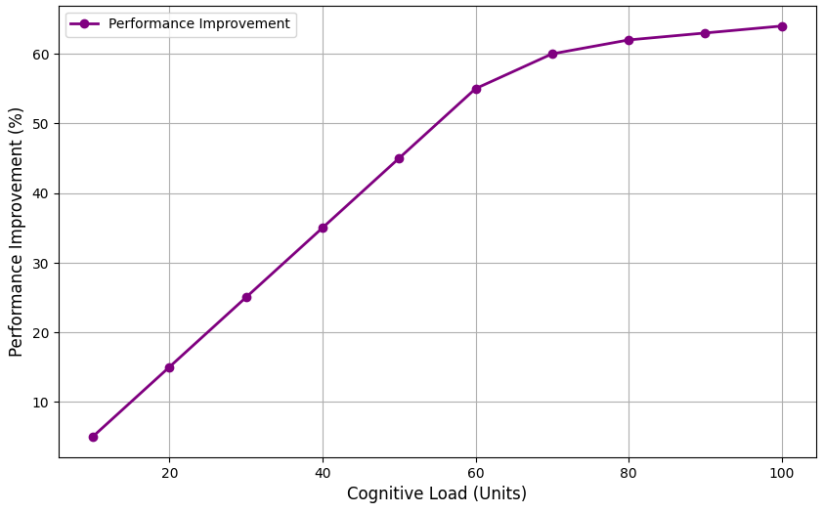


Figure 9: Cognitive Load vs. Performance Improvement.

6.7 Overall Performance Comparison

A final comparison of the proposed system's performance across all metrics reveals the Collaborative Learning Group's superiority in fostering individual and peer-assisted learning. The system's adaptive reinforcement learning algorithm, combined with peer-learning mechanisms, resulted in higher learning gains, greater engagement, and effective collaboration.

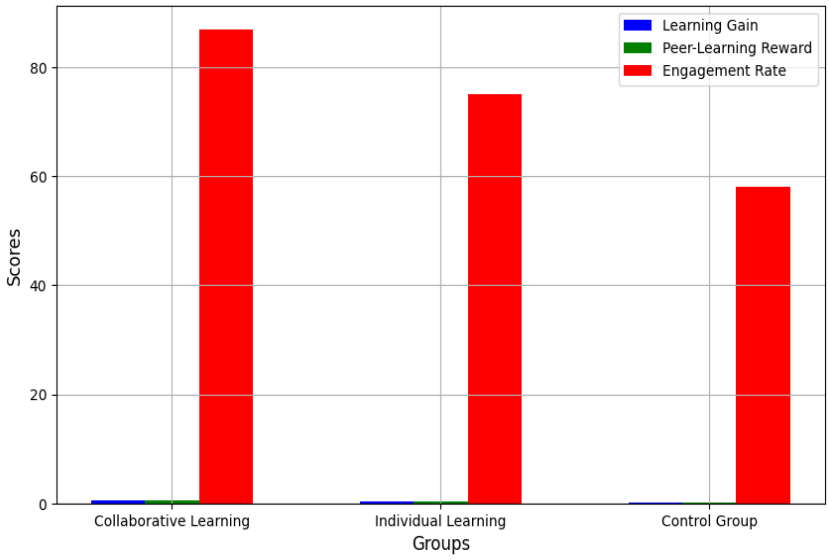


Figure 10: Summary of Learning Gain, Peer-Learning Reward, and Engagement Rate for all Groups.

The experiment's results clearly demonstrate the effectiveness of the proposed Reinforcement Learning-Based System for Personalized Educational Content Delivery. The system's ability to adapt content dynamically based on each student's learning progress, engagement, and cognitive load, coupled with the strategies for providing incentives for peer learning, has proven to be an effective strategy for enhancing educational outcomes. The Collaborative Learning Group, in particular, showed improvements, indicating the value of integrating peer assistance into personalized learning platforms.

Data obtained from experiments provide significant evidence about the applicability of the proposed Reinforcement Learning-Based System for Personalized Educational Content Delivery. The system's ability to change the content dynamically based on the students' progress, interest, and mental load, as well as the introduction of a reward for knowledge sharing with other students, has already become a powerful tool for improving learning outcomes. Especially, the Collaborative Learning Group exhibited significant enhancement, proving the significance of incorporating peer support in the student-meaningful learning environments.

6.8 Comparison Between DDQN Algorithm, Q-Learning, and SARSA

In this work, three reinforcement learning algorithms (Q-Learning, SARSA, and DDQN) were considered appropriate tools for personalizing educational content for students. The metric used for comparison is the historical convergence speed, which measures how quickly each algorithm learns to provide the most optimized content based on students' interactions. The performance comparison of the convergence speed of these algorithms is shown in Figure 11.

Below is a detailed explanation of the three algorithms and their comparison:

1. Double Deep Q-Network (DDQN): DDQN is a learned model developed from Q-Learning, introduced by Hasselt et al. (2015), which uses two sets of neural networks instead of one. This approach improves decision-making accuracy regarding action selection, particularly for content recommendations. In this study, the original DDQN algorithm serves as the foundation, and modifications were applied to incorporate a peer-learning reward mechanism for enhanced collaborative learning [28]. The graph demonstrates that DDQN has the fastest convergence level, thus implying that it functions best to students' requirements or educational content more efficiently in episodes. This efficient characteristic makes it suitable for use in flexible systems like the individualized learning where frequent change is common.

2. Q-Learning: Q-learning is a simpler algorithm that updates Q-values (action-value pairs) based on the rewards obtained from actions. Although it performs well when all action values are close to the target value, it may converge slowly in diverse settings due to its tendency to overestimate action values [29].

From the graph, it is observed that Q-Learning takes more time to converge as compared to DDQN, but in the end, they both don't vary much. This is expected given the fact that Q-Learning does not have the same controls that are designed into the structure of DDQN to prevent overestimating the values of the Q-function until more user specific content can be rapidly recommended.

3. State-Action-Reward-State-Action (SARSA): SARSA is an on-policy reinforcement learning algorithm where the action-value pair is updated based on the action taken during the learning phase. This approach makes SARSA more conservative than Q-Learning, as it evaluates actions according to the current policy rather than exploring new or potentially better policies. Unlike off-policy algorithms such as Q-Learning or DDQN, SARSA sticks to the current policy throughout the learning process, making it suitable for scenarios where stability and adherence to the existing policy are prioritized [30].

SARSA demonstrates the slowest convergence out of the three discussed algorithms. Its centralized structure also slows down its ability to adapt to the learning needs of students. This might cause safer decisions from one perspective, though it is less efficient in cases like the context of the application, such as personalized learning, where fluidity is crucial.
The learning progress of Q-learning, SARSA, and DDQN is shown in Figure 11, and it displays the converge behavior of the three learning algorithms.
Axes Explanation:
•X-axis (Iterations): Illustrates the number of training cycles that each algorithm undergoes, commonly known as iterations or step size. There is an action selection, a reward reception, and the Q-value update before the next iteration comes in.
•Y-axis (Q-value): Refers to the action-value pairs (Q-values), which suggest the total expected reward for the action undertaken in a particular state. The policy learning and the decision making of the algorithm are thus represented by higher Q-values.

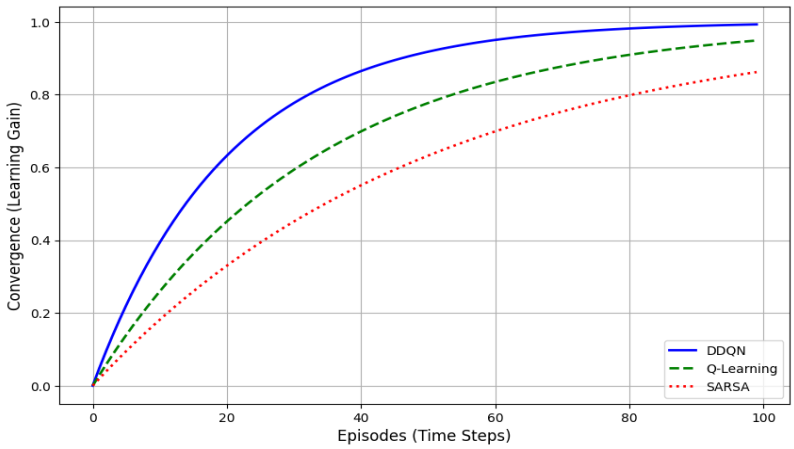


Figure 11: Comparison of Algorithm Convergence Speeds

This graph highlights the comparative efficiency of the algorithms in terms of learning speed and convergence, providing insights into why DDQN was chosen as the foundation for this study's personalized learning system. Its faster convergence makes it ideal for real-time adaptation to students' needs.

7 Discussion

Numerous experiments and analyses have proved the applicability of the introduced system based on reinforcement learning for creating and delivering individual educational content. The proposed training model, with the application of the Double Deep Q-Network (DDQN) algorithm, recommended an adaptive learning model to be developed to learn from the results and change the learning environment according to individual student's needs as they progress. This part articulates the results obtained from the experiments, complements the results achieved by other standard methods, and identifies the systems' strong points and possible scopes in the future.

1. Effectiveness of DDQN in Personalized Learning:

The results obtained during the experiments prove that the education content can be personalized with the help of the proposed DDQN algorithm. As depicted in the Learning Gain Curves (refer to Figure 4), it shows that students of the collaborative learning group using DDQN performed far better in learning outcomes than the students in the individual learning group and the control group. This goes to show that the system was able to make changes and adapt better to the needs and learning preferences of individual learners on student, a factor which is very vital in increasing the effectiveness of learning.

In addition, the rewards by Peer-Learning Reward analysis (Figure 5) also demonstrated the actual benefit of the DDQN system as well as the cooperative learning of students. The blended cumulative reward curve reveals how, via the method of DDQN the focus was done to create a peer learning environment, which accelerated the group learning results and did not hamper the individual learning rates.

2. Comparative Performance of Algorithms:

The comparison of various reinforcement learning algorithms such as Q-learning and SARSA, as well as the use of the DDQN algorithm that emerged from this paper, is relevant to the overall strategy. Figure 6 shows that DDQN explores the solution space much better than the other alternatives, resulting in faster convergence to the best learning policy. Nonetheless, Q-Learning and SARSA had slower convergence rates, but they contributed to learning improvement to some extent, albeit not rapidly. The following analyses provide evidence that supports the enhanced performance of DDQN in dealing with the exploration/exploitation dilemma inherent in personalized learning systems.

Figure 7 shows the Q-value Convergence for the DDQN, which emphasizes the fact that DDQN is much more convenient to converge when compared to the Q-Learning and SARSA algorithms. This is because DDQN makes use of two Q-networks, which does not contribute to the overestimation bias that is observed in Q-Learning. It means that the system can quickly adjust to the student activities in regard to content delivery, hence the degree of interest in the content and hence the result achieved.

3. Engagement Rate and Cognitive Load:

The third and final dimension of the study concerned the assessment of the Engagement Rate of students within the various groups addressed in this study (Figure 6). The DDQN group collaboration demonstrated engagement statistically significantly higher than the individual and control groups throughout the study. It can, therefore, be inferred that the distinguishing between student's content, personalization of the delivery, and the collaborative learning dynamics served to maintain students' attention and interest throughout the learning process.

The system's ability to effectively balance Cognitive Load against Performance Improvement, as depicted in the system (Figure 9), strengthens its argument further. Again, although a higher cognitive load may have a negative impact on performance, this system makes sure that content delivery does not overload the students. The proposed DDQN-based system offered the right level of challenge and learning/content delivery to enhance both learners' interest and performance, mainly because of the ability to tune the difficulty and the pacing of the presented material in accordance with the specific learning behavior.

4. Contributions to Adaptive Learning Systems:

The first novel aspect of this work is the application of reinforcement learning in child-to-child education approach. This made it possible for the students not only to move from one level to another at their comfort but also make group work possible and learning more fun. This feature of the system offers possibilities to satisfy both individual and collaborative learning requirements, which is one of the strengths of the presented system compared to the traditional learning systems, which more often serves as the one-mode-fits-all model.

The Exploration vs. Exploitation Balance (Figure 8) demonstrates how the system can investigate a number of content delivery approaches while effectively utilizing the most

efficient ones, which creates an adaptive learning environment. This is very important, especially for situations where learner involvement and flexibility mark the hallmark of sustainability.
Key points from the comparison include the following:
- Efficiency of DDQN: The quick convergence of DDQN means that it's a faster learner than Q-Learning and SARSA, making it the best for learning algorithm in adaptive environments such as the one used by this study. Less often, it defines the most efficient educational paths so that the students receive more relevant material when the adaptation occurs.
- Q-Learning's Moderate Performance: Although not as fast as DDQN, the Q-Learning algorithm still adequately assesses specialized learning material input by a learner. Much more time is devoted to refining its recommendations, and at the beginning of the learning process, the content delivery may not be as optimal as desired.
- SARSA's Slow Adaptation: Thus, while SARSA has a slower convergence of the parameter estimates, it is not as suitable for a constantly changing and swift pace in personalized education. It may cause some delay in the provision of the right study materials, thus the effect on the learning process.

As DDQN has a higher convergence speed and improved learning capabilities, it was decided upon as the main algorithm for this study. It provides a faster mechanism to adjust to the needs of the different students and is very effective in tailored learning environments. This can be seen in the graph, which shows their enhanced ability to deliver personalized educational content as efficiently.

This comparison helps further appreciate the choice of DDQN as the best algorithm for the system designed in this study. The fast convergence implies that students will have improved chances of accessing materials that fit their needs and goals, hence improved performance. The observations made from the slow learning of the Q-Learning and SARSA reinforce the idea that simpler reinforcement learning algorithms struggle to perform well in complex environments.

8. Practical Implications, Limitations, and Future Directions

From the discussion about the proposed system involving the DDQN-based personalized learning system, there were important benefits realized from real-time content adaptation, peer-learning motivation, and data analysis. One of the strengths of the system is the flexibility provided to the content to offer reasonable and personalized engagements for students, especially in low resource and diverse learners groups classrooms. In this way, the system contributes to more effective education equity and performance by filling in gaps that may exist within person-to-person learning styles and pace.
Combining peer-learning incentives enhances cooperation among students by encouraging them to help one another, making it helpful in enhancing group tasks and study models that include face-to-face and online learning. The reinforcement structures in the system enhance the participation of learners as individuals and as groups, and the analytics structures give the educators specific information about learners from their performance, their cognitive load and their participation levels. All these features enable the implementation of specific interventions, the assessment of curriculum needs, and the determination of the effectiveness of educational approaches simultaneously.

All the same, it can be noted that there are some possible improvements to the system. Potential future work can experiment with other types of reinforcement learning techniques like Proximal Policy Optimization or Soft Actor Critic as they may potentially allow for a

better balance of exploration/exploitation. Besides, the testing of the system across various subjects, different age groups, and various contexts of education would extend more proof about the scalability of the system. This means that, solving them using NLP may help provide real-time feedback analysis and real-time adjustment to the perceived problems by students.

All in all, the DDQN based system incorporates a strong learning model, which can serve as a solution to the above issues for policymakers and educators to design large-scale and efficient personalized learning environments. If the system undergoes further developments and incorporates various forms of learning processes, this has a high possibility to change the education system all over the world that acknowledges students' participation, interaction, and equality.

References

[1] O. O. Ayeni, N. M. Al Hamad, O. N. Chisom, B. Osawaru, and O. E. Adewusi, "AI in education: A review of personalized learning and educational technology," *GSC Adv. Res. Rev.*, vol. 18, no. 2, pp. 261–271, 2024.

[2] Y. Sirisathitkul and C. Sirisathitkul, "Smartphones as Smart Tools for Science and Engineering Laboratory: A Review," *Iraqi J. Sci.*, vol. 64, no. 5, pp. 2240–2249, 2023, doi: 10.24996/ijs.2023.64.5.12.

[3] Y. Zhang, "Machine Learning-Based Personalized Learning Path Decision-Making Method on Intelligent Education Platforms.," *Int. J. Interact. Mob. Technol.*, vol. 18, no. 16, 2024.

[4] Y. Xiao and K. F. Hew, "Personalized gamification versus one-size-fits-all gamification in fully online learning: Effects on student motivational, behavioral and cognitive outcomes," *Learn. Individ. Differ.*, vol. 113, pp. 102470-2024.

[5] L. Marcinauskas, A. Iljinas, J. Čyvienė, and V. Stankus, "Problem-based learning versus traditional learning in physics education for engineering program students," *Educ. Sci.*, vol. 14, no. 2, pp. 154- 2024.

[6] M. Hamadi and J. El-Den, "A conceptual research framework for sustainable digital learning in higher education," *Res. Pract. Technol. Enhanc. Learn.*, vol. 19, pp. 1–25, 2024.

[7] B. Memarian and T. Doleck, "A scoping review of reinforcement learning in education," *Comput. Educ. Open*, pp. 100175-2024.

[8] J. Knopes, M. A. Cascio, and B. Warner, "Intraprofessionalism and Peer-to-Peer Learning in American Medical Education," *Qual. Health Res.*, vol. 34, no. 6, pp. 528–539, 2024.

[9] A. B. Feroz Khan and S. R. A. Samad, "Evaluating online learning adaptability in students using machine learning-based techniques: A novel analytical approach," *Educ. Sci. Manag*, vol. 2, no. 1, pp. 25–34, 2024.

[10] N. Adhami and M. Taghizadeh, "Integrating inquiry-based learning and computer supported collaborative learning into flipped classroom: Effects on academic writing performance and perceptions of students of railway engineering," *Comput. Assist. Lang. Learn.*, vol. 37, no. 3, pp. 521–557, 2024.

[11] P. Mittal, S. Kalra, A. Dadhich, and P. Ajmera, "Empowering children for better health with child-to-child approach: a systematic literature review," *Health Educ.*, 2024.

[12] T. K. Sari and P. H. Rahmani, "The role of positive reinforcement on students in english language learning: a skinnerian behaviorist," *JOEY J. English Ibrahimy*, vol. 3, no. 1, pp. 1–5, 2024.

[13] H. a. Naman and Z. J. M. Ameen, "A New Method in Feature Selection based on Deep Reinforcement Learning in Domain Adaptation," *Iraqi J. Sci.*, vol. 63, no. 2, pp. 817–829, 2022, doi: 10.24996/ijs.2022.63.2.35.

[14] Y. Chen, "Comparison of Deep Q-Learning Network and Double Deep Q-Learning Network for Trading Strategy," 2024.

[15] Y. Yun, H. Dai, R. An, Y. Zhang, and X. Shang, "Doubly constrained offline reinforcement learning for learning path recommendation," *Knowledge-Based Syst.*, vol. 284, pp. 111242-2024.

[16] M. Veena, "Innovation in Pediatrics: Family-Centered Care and Child-to-Child Approaches," *Indian J. Nurs. Sci.*, pp. 22–24, 2024.
[17] C. C. A. van Nooijen *et al.*, "A cognitive load theory approach to understanding expert scaffolding of visual problem-solving tasks: A scoping review," *Educ. Psychol. Rev.*, vol. 36, no. 1, pp. 12- 2024.
[18] F. A. Müller and T. Wulf, "Differences in learning effectiveness across management learning environments: a cognitive load theory perspective," *J. Manag. Educ.*, vol. 48, no. 4, pp. 802–828, 2024.
[19] R. M. Ryan and E. L. Deci, "Self-determination theory," in *Encyclopedia of quality of life and well-being research*, Springer, 2024, pp. 6229–6235.
[20] Z. Zafrullah and A. M. Ramadhani, "The use of mobile learning in schools as a learning media: Bibliometric analysis," *Indones. J. Educ. Res. Technol.*, vol. 4, no. 2, pp. 187–202, 2024.
[21] T. Quraishi *et al.*, "Integration of Mobile Learning Technologies in Afghanistan Universities: Opportunities and Challenges," *Educ. Spec.*, vol. 2, no. 1, pp. 1–14, 2024.
[22] D. Shawky and A. Badawi, "A reinforcement learning-based adaptive learning system," in *The International Conference on Advanced Machine Learning technologies and Applications (AMLTA2018)*, 2018, pp. 221–231.
[23] B. Fahad Mon, A. Wasfi, M. Hayajneh, A. Slim, and N. Abu Ali, "Reinforcement Learning in Education: A Literature Review," in *Informatics*, 2023, vol. 10, no. 3, p. 74.
[24] P. Kiran, B. K. S. Prasad, and J. Sivasubramanian, "A Personalized E-Learning System Using Reinforcement Learning Through Satellite," in *2023 IEEE 20th India Council International Conference (INDICON)*, 2023, pp. 800–805.
[25] W. S. Sayed *et al.*, "AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging E-learning platform," *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3303–3333, 2023.
[26] J. Hu and G. Jin, "An Intelligent Framework for English Teaching through Deep Learning and Reinforcement Learning with Interactive Mobile Technology.," *Int. J. Interact. Mob. Technol.*, vol. 18, no. 9, 2024.
[27] M. Z. Islam, R. Ali, A. Haider, M. Z. Islam, and H. S. Kim, "Pakes: a reinforcement learning-based personalized adaptability knowledge extraction strategy for adaptive learning systems," *IEEE Access*, vol. 9, pp. 155123–155137, 2021.
[28] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, 2016, vol. 30, no. 1.
[29] P. Dayan and C. Watkins, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.
[30] G. A. Rummery and M. Niranjan, *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.