Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Evaluasi Model Deep Reinforcement Learning Untuk Adaptasi Konten Pembelajaran Berdasarkan Performa Siswa

Fahad Abdul Aziz [a,*], *M. Yusril Helmi Setyawan* [b], *Cahyo Prianto* [c,]

[abc]*Program Studi Sarjana Terapan Informatika, Universitas Logistik dan Bisnis Internasional Jl.Sariasih No.54, Sarijadi, Kec, Sukasari, Bandung, 40151, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 14 Juli 2025
Revisi Akhir: 27 April 2026
Diterbitkan *Online*: 05 Mei 2026

KATA KUNCI

Deep Reinforcement Learning,

Pembelajaran Adaptif,

DQN,

DDQN

KORESPONDENSI

E-mail: fahadra96@gmail.com *

A B S T R A C T

Transformasi digital dalam pendidikan telah mempercepat integrasi (*Artificial Intelligence*/*AI*), khususnya pada sistem pembelajaran adaptif. Sistem konvensional sering kali gagal menyesuaikan materi dengan performa dan kecepatan belajar individu. Untuk mengatasi hal ini, penelitian ini mengimplementasikan *Deep Reinforcement Learning* (DRL) guna membangun model rekomendasi konten adaptif berdasarkan riwayat interaksi dan hasil belajar siswa. Dua model agen *Deep Q-Network* (DQN) dan *Double* DQN (DDQN) dikembangkan dan dievaluasi dalam lingkungan belajar simulatif menggunakan dataset EdNet-KT1, yang berisi data interaksi siswa dalam skala besar. Pelatihan dilakukan melalui formulasi *Markov Decision Process* (MDP), dengan vektor keadaan yang mencakup metadata soal, akurasi jawaban, dan waktu pengerjaan. Evaluasi model menggunakan tiga metrik utama: *reward* per episode, generalisasi terhadap pengguna baru (*unseen users*), dan akurasi prediksi. Hasil menunjukkan bahwa DDQN memiliki performa lebih unggul dibandingkan DQN dalam hal stabilitas, kemampuan generalisasi, dan akurasi. Rata-rata reward yang diperoleh DDQN melebihi 14 dalam sebagian besar skenario, dengan akurasi prediksi mencapai 78%, sedangkan DQN hanya mencapai 74%. Analisis kurva pembelajaran juga menunjukkan bahwa DDQN mengalami konvergensi lebih cepat dengan fluktuasi yang lebih rendah. Evaluasi model menggunakan tiga metrik utama: *reward* per episode, generalisasi terhadap pengguna baru (*unseen users*), dan akurasi prediksi. Hasil menunjukkan bahwa DDQN memiliki performa lebih unggul dibandingkan DQN dalam hal stabilitas, kemampuan generalisasi, dan akurasi. Rata-rata reward yang diperoleh DDQN melebihi 14 dalam sebagian besar skenario, dengan akurasi prediksi mencapai 78%, sedangkan DQN hanya mencapai 74%. Analisis kurva pembelajaran juga menunjukkan bahwa DDQN mengalami konvergensi lebih cepat dengan fluktuasi yang lebih rendah

1. PENDAHULUAN

Transformasi digital dalam pendidikan yang mendorong integrasi teknologi (*Artificial Intelligence*/*AI*), telah meningkatkan efektivitas dan personalisasi sistem pembelajaran. Salah satu tantangan utama dalam pembelajaran daring adalah keterbatasan sistem tradisional dalam mengakomodasi kebutuhan dan kemampuan individu siswa, yang sering kali mengalami varian gaya belajar dan pemahaman yang dinamis [1]. Pendekatan seragam dalam pembelajaran cenderung mengabaikan perbedaan ini, memerlukan penggunaan sistem pembelajaran adaptif untuk menyesuaikan konten dan jalur pembelajaran sesuai dengan performa siswa (Mirata et al., 2020). Sistem adaptif ini memungkinkan pembelajaran yang lebih personal dan potensi meningkatkan hasil belajar siswa secara signifikan [2].

Reinforcement Learning (RL), sebagai salah satu cabang dari kecerdasan buatan, menawarkan pendekatan berbasis *sequential decision making* yang dapat mengoptimalkan strategi pembelajaran melalui interaksi agen dengan lingkungan. Ketika dikombinasikan dengan *Deep Learning*, terbentuklah pendekatan *Deep Reinforcement Learning* (DRL) yang memiliki kemampuan untuk memproses representasi kompleks dari data pendidikan serta menghasilkan kebijakan adaptasi konten yang optimal [3].

https://doi.org/10.25077/TEKNOSI.v12i1.2026.105-112

DRL tidak hanya mampu belajar dari umpan balik eksplisit seperti skor kuis atau tingkat penyelesaian, tetapi juga dapat menggeneralisasi pola interaksi pengguna untuk menentukan konten selanjutnya yang paling sesuai guna meningkatkan capaian belajar siswa [4].

Berbagai studi sebelumnya telah mengeksplorasi penerapan RL dalam konteks pendidikan. Misalnya, [5] menerapkan DRL untuk sistem pembelajaran adaptif menggunakan dataset EdNet, menunjukkan efektivitas pendekatan ini dalam rekomendasi konten edukatif. [4] mengembangkan sistem *e-learning* adaptif yang merekomendasikan jalur pembelajaran personal menggunakan algoritma *Q-learning*. Selain itu, [6] menerapkan DRL untuk intervensi metakognitif, menunjukkan kemampuan sistem dalam beradaptasi secara dinamis terhadap perubahan kondisi belajar siswa. Namun, sebagian besar penelitian tersebut cenderung berfokus pada aspek prediksi performa atau rekomendasi umum, bukan pada analisis perbandingan varian algoritma DRL seperti *Deep Q-Network* (DQN) dan *Double Deep Q-Network* (DDQN) dalam konteks pendidikan berbasis data nyata.

Penelitian ini bertujuan untuk mengembangkan dan mengevaluasi model DRL untuk adaptasi konten pembelajaran berdasarkan performa siswa. Model yang diusulkan mengimplementasikan DQN dan DDQN untuk membentuk kebijakan rekomendasi konten yang mempertimbangkan histori interaksi pengguna, akurasi jawaban, serta fitur-fitur pembelajaran lainnya. Evaluasi dilakukan menggunakan *dataset* EdNet-KT1, yang merupakan representasi autentik dari interaksi pengguna dalam lingkungan pembelajaran daring berskala besar [7]. Penelitian ini menitikberatkan pada kemampuan model dalam melakukan adaptasi konten secara *real-time* berdasarkan performa siswa. Dengan pendekatan tersebut, diharapkan penelitian ini memberikan kontribusi terhadap pengembangan sistem pembelajaran cerdas yang lebih personal, adaptif, dan efisien, serta memperkaya literatur mengenai penerapan DRL dalam domain teknologi pendidikan.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan metode berbasis DRL untuk mengembangkan dan mengevaluasi model DQN dan DDQN dalam adaptasi konten pembelajaran. Proses eksperimen dilakukan dalam lingkungan simulatif yang dibangun secara artifisial, dengan tiga tahapan utama: (1) Persiapan Data & Pengaturan Lingkungan, (2) Pelatihan & Optimisasi Agen DRL, dan (3) Evaluasi. Pada tahap pertama, data interaksi siswa diproses dan elemen *Markov Decision Process* (MDP) diformulasikan, sehingga agen DRL dapat menginterpretasi *state*, *action*, dan *reward* dengan efisien [8]. Tahapan kedua melibatkan pelatihan agen DQN dan DDQN menggunakan arsitektur jaringan saraf, di mana strategi eksplorasi *epsilon-greedy* diaplikasikan untuk mengoptimalkan pelatihan [9] [10]. Pada tahap evaluasi, metrik seperti *reward* per episode dan akurasi untuk pengguna yang belum pernah terlihat digunakan untuk menilai efektivitas rekomendasi konten yang dihasilkan oleh model [11].

Gambar 1. Flowchart Metodelogi penelitian

2.1. Data Preparation & Environment Setup

2.1.1. Dataset dan Sumber data

Dataset yang digunakan dalam penelitian ini adalah EdNet-KT1, yang merupakan bagian dari proyek *open-source* EdNet yang dirilis oleh Riiid Labs melalui repositori GitHub: https://github.com/riiid/ednet. EdNet merupakan salah satu dataset pembelajaran daring terbesar yang pernah dirilis, mencakup lebih dari 131 juta interaksi pengguna, 945.000 siswa, dan lebih dari 10.000 soal dari *platform* pembelajaran komersial [7] dan mendukung penelitian di bidang teknologi pendidikan, *knowledge tracing*, serta pembelajaran adaptif berbasis kecerdasan buatan. Dataset ini terbagi dalam beberapa subset, termasuk KT1, KT2, KT3, KT4, serta file *questions.csv* dan *lectures.csv* [10] Penelitian ini memfokuskan pada subset KT1 yang berisi *log* interaksi siswa dengan sistem pembelajaran, dan *questions.csv* yang berisi metadata mengenai karakteristik soal [11] Dari total data, eksperimen ini menggunakan 1000 pengguna pertama dari KT1 untuk efisiensi komputasi dan mempertahankan keragaman perilaku pengguna yang representatif dalam pelatihan model DQN dan DDQN.

2.1.2. Data Preprocessing

Sebelum pelatihan model DRL, data dari KT1.csv dan questions.csv digabung dan diproses untuk memastikan validitas dan kesiapan numerik. Tahapan praproses meliputi:

1. *Cleaning* untuk menghapus baris dengan nilai kosong pada *user_answer* dan menyaring *elapsed_time* dalam rentang 1–300 detik. Kolom duplikat hasil join juga dihapus.
2. *Encoding* untuk kolom *question_id* dan *tags* diubah menjadi bentuk numerik (*_enc) menggunakan *LabelEncoder.*

https://doi.org/10.25077/TEKNOSI.v12i1.2026.105-112

3. Normalisasi *elapsed_time* dinormalisasi ke [0, 1] dengan *MinMaxScaler*. Kolom *correct* dikonversi ke format biner.
4. Seleksi Fitur hanya lima atribut digunakan: *user_id*, *question_id_enc*, *tags_enc*, *elapsed_time_norm*, dan *correct.*

Dataset akhir disimpan sebagai rl_ready.csv dan digunakan dalam lingkungan KTEnv. Tabel 1 menyajikan contoh dari data hasil prapemrosesan yang digunakan sebagai input dalam lingkungan pelatihan agen.

Tabel 1. Dataset Hasil *Processing*

user_ id	question_id_ enc	tags_e nc	elapsed_time _norm	correct
u1	6666	1685	0.123746	0
u1	6340	1682	0.076923	1
u1	5988	4	0.224080	1
u1	6465	1694	0.137124	0

2.1.3. Definisi Komponen *Reinforcerment Learning*

Masalah adaptasi konten dalam penelitian ini diformulasikan sebagai permasalahan MDP, yang digunakan untuk memodelkan alur pembelajaran interaktif antara agen dan siswa [12]. Komponen-komponen utama dari MDP dalam eksperimen ini dijelaskan sebagai berikut:

1. *State* (s)
   Setiap kondisi siswa pada satu waktu direpresentasikan oleh sebuah *state vector* berdimensi 4, yang mencakup informasi sebagai berikut:
   a. *question_id_enc* untuk indeks numerik dari soal yang diberikan terakhir.
   b. *tags_enc* kategori/topik dari soal tersebut.
   c. *elapsed_time_norm* waktu pengerjaan soal sebelumnya, yang telah dinormalisasi ke rentang [0, 1].
   d. *correct*: hasil jawaban siswa pada interaksi sebelumnya (0 atau 1).

   State diwakili dalam bentuk vektor *real-valued*:

$$s_t = \left[\frac{question\_id\_enc}{N_q}, \frac{tags\_enc}{N_t}, elapsed\_time\_norm, correct\right] \quad (1)$$

   di mana N_q dan N_t adalah jumlah total soal dan tag unik, yang digunakan untuk normalisasi indeks soal dan tag.
2. *Action* (a)
   Aksi yang dilakukan agen berupa pemilihan satu soal untuk direkomendasikan dari daftar soal yang tersedia bagi siswa tertentu. Setiap *action* a_t adalah nilai *integer* yang merepresentasikan ID soal (*question_id_enc*) yang valid untuk siswa aktif.
3. *Reward* (r)
   Reward diberikan setelah siswa menyelesaikan soal yang direkomendasikan. Dalam eksperimen ini, *reward* tidak hanya mempertimbangkan kebenaran jawaban, tetapi juga waktu pengerjaan sebagai indikator efisiensi. Fungsi *reward* dirancang sebagai:

$$r_t = correct_t + 0.5 \times (1 - elapsed\_time\_norm_t) \quad (2)$$

   Dimana *correct* hasil jawaban (1 jika benar, 0 jika salah), $elapsed\_time\_norm_t$ waktu pengerjaan soal pada t, telah dinormalisasi.
4. *Transition*
   Setelah agen merekomendasikan suatu soal dan *reward* diterima, sistem berpindah ke *state* berikutnya:

$$s_{t+1} = f(s_t, a_t, r_t) \quad (3)$$

   Transisi ini bersifat deterministik berdasarkan data interaksi historis yang tersedia. Setiap episode interaksi dibatasi hingga maksimum 10 langkah atau sampai data siswa habis.

2.1.4. Environment Setup

Lingkungan pembelajaran dikembangkan menggunakan *library OpenAI Gym* dalam kelas KTEnv, yang mensimulasikan interaksi agen dengan siswa berdasarkan data historis. Setiap episode dimulai dengan pemilihan acak satu siswa, lalu state awal dibentuk dari empat elemen: *question_id_enc, tags_enc, elapsed_time_norm*, dan *correct*. Agen kemudian memilih soal valid sebagai aksi, dan menerima *reward* berdasarkan kebenaran jawaban serta kecepatan pengerjaan. Episode berakhir saat langkah mencapai batas atau data siswa habis. Lingkungan ini dirancang untuk mendukung pelatihan DQN dan DDQN secara berulang dengan konteks belajar yang realistis, sejalan dengan pendekatan modular dalam pengembangan *environment reinforcement learning* berbasis *OpenAI Gym* seperti dijelaskan oleh [13].

2.2. DRL *Agent Training*

Penelitian ini mengimplementasikan dua pendekatan DRL sebagai agen pembelajar, yaitu DQN dan DDQN, untuk melakukan adaptasi konten pembelajaran berdasarkan performa siswa. Keduanya dirancang untuk mempelajari kebijakan rekomendasi soal dengan memaksimalkan *reward* yang diperoleh melalui interaksi dengan lingkungan pembelajaran simulatif (KTEnv) [4].

2.2.1. DQN

DQN merupakan algoritma DRL yang menggunakan satu jaringan saraf untuk mengaproksimasi fungsi nilai Q. Tujuannya adalah mempelajari nilai dari setiap aksi a'yang dapat diambil pada suatu *state*, agar agen dapat memilih aksi optimal yang memaksimalkan *reward* kumulatif jangka Panjang.
Fungsi nilai target dalam DQN dihitung dengan persamaan:

$$Q_{\text{target}} = r_t + \gamma \cdot \max_{a'} Q\,(s_{t+1}, a') \quad (4)$$

Formulasi ini merupakan pondasi dari DQN dan banyak digunakan dalam berbagai eksperimen DRL berbasis Q-learning [14].
Keterangan:

r_t = *reward* pada langkah ke$-t$
γ = faktor diskonto untuk *reward* masa depan
s_{t+1} = *state* setelah aksi dilakukan
a' = semua aksi yang mungkin dilakukan dis_{t+1}
Q_{target} = estimasi nilai Q yang digunakan untuk pembaruan

2.2.2. DQN

DDQN adalah pengembangan dari DQN yang mengatasi *overestimation* bias dalam prediksi nilai Q. Pendekatan ini menggunakan dua jaringan saraf terpisah, pertama *Online network* (*policy network*) digunakan untuk memilih aksi terbaik, dan target *network* digunakan untuk mengevaluasi nilai Q dari aksi tersebut.

Fungsi nilai target dalam DDQN didefinisikan sebagai:

$$Q_{\text{target}} = r_t + \gamma \cdot Q_{\text{target}}\left(s_{t+1}, \arg\max_{a'} Q_{\text{policy}}\,(s_{t+1}, a')\right) \quad (5)$$

Formulasi ini merupakan pondasi dari DDQN dan banyak digunakan dalam berbagai eksperimen DRL berbasis *Q-learning* [14].

Keterangan:

r_t = *reward* pada langkah ke$-t$
γ = faktor diskonto untuk *reward* masa depan
s_{t+1} = *state* setelah aksi dilakukan
a' = semua aksi yang mungkin dilakukan dis_{t+1}
Q_{target} = estimasi nilai Q yang digunakan untuk pembaruan
Q_{policy} = nilai Q dari *policy network*

2.2.3. Arsitektur Model dan Optimisasi Hiperparameter

Kedua agen menggunakan jaringan *feedforward* tiga lapis dengan dua *hidden layer* (128 neuron) sebagai *approximator Q-value*. Input berupa vektor *state* (dimensi 4), *output* berupa vektor nilai Q untuk seluruh aksi valid. Pelatihan dilakukan selama 100 episode dengan strategi *epsilon-greedy* untuk eksplorasi dan *experience replay buffer* untuk stabilisasi pembelajaran.

Tabel 2. Hiperparameter Pelatihan Agen DRL

Parameter	Nilai	Keterangan
Jumlah episode	100	Durasi pelatihan per agen
Batch size	64	Ukuran mini-batch saat training
Learning rate (α)	0.001	Kecepatan pembelajaran
Gamma (γ)	0.99	Diskonto *reward* masa depan
Epsilon awal	1.0	Probabilitas eksplorasi awal
Epsilon minimal	0.05	Batas eksplorasi minimum
Epsilon decay rate	0.995	Penurunan epsilon per episode
Replay buffer size	10.000	Jumlah maksimum pengalaman yang disimpan
Target update freq	Setiap 10 ep	Sinkronisasi target network (DDQN & DQN)

Dengan strategi ini, model dapat mengadopsi mekanisme pembelajaran bertahap: dimulai dari eksplorasi penuh (ε tinggi), kemudian beralih ke eksploitasi berdasarkan pengalaman (*epsilon decay*), sembari menjaga stabilitas melalui *replay buffer* dan pembaruan target *network.*

2.2.4. Evaluasi

Evaluasi dilakukan untuk menilai efektivitas, generalisasi, dan stabilitas model DQN dan DDQN dalam adaptasi konten pembelajaran. Tiga metrik utama digunakan:

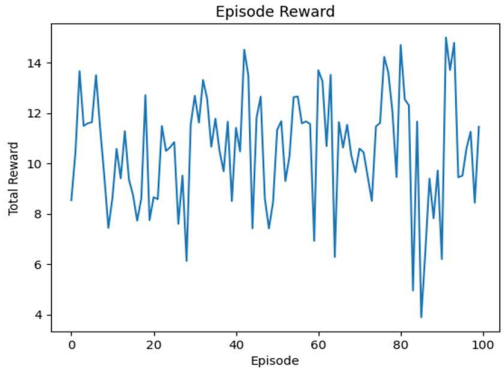
1. *Reward* per Episode Mengukur rata-rata *reward* selama pelatihan sebagai indikator kualitas rekomendasi konten oleh agen.
2. *Generalization on Unseen Users* Menguji kemampuan model beradaptasi pada 200 siswa baru yang tidak dilibatkan saat pelatihan.
3. *Learning Curve* Menganalisis dinamika pelatihan (*reward*, *loss*, akurasi) untuk menilai stabilitas dan konvergensi model.

3. HASIL

Bab ini menyajikan hasil evaluasi dari implementasi dua model *Deep Reinforcement Learning*, yaitu DQN dan DDQN, dalam konteks adaptasi konten pembelajaran berbasis performa siswa. Empat aspek utama yang dievaluasi meliputi *reward* per episode, generalisasi terhadap pengguna baru, pola konvergensi selama pelatihan (*learning curve*), dan akurasi prediksi.

3.1. Evaluasi *Reward* per Episode

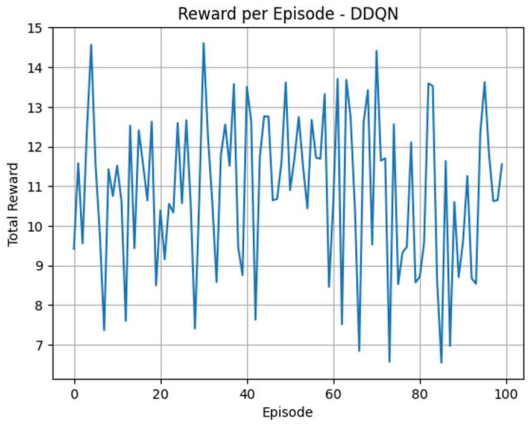
Evaluasi awal dilakukan dengan menganalisis total *reward* kumulatif yang diperoleh oleh agen pada setiap episode pelatihan. *Reward* digunakan sebagai indikator utama untuk mengukur efektivitas pembelajaran model dalam menyusun kebijakan rekomendasi soal yang adaptif terhadap performa siswa [11] [8]. Nilai reward yang lebih tinggi menunjukkan bahwa agen berhasil memilih soal yang sesuai dengan kemampuan siswa, sehingga mendorong peningkatan respons yang benar.



Gambar 2. DQN Episode *Reward*

Gambar 2 menunjukkan pola reward yang dihasilkan oleh agen DQN selama 100 episode pelatihan. Teramati bahwa *reward* mengalami fluktuasi cukup tajam, terutama pada fase awal hingga pertengahan pelatihan. Pola ini mencerminkan tingginya intensitas eksplorasi akibat penerapan strategi *epsilon-greedy* yang masih dominan di awal pelatihan. Selain itu, penggunaan satu jaringan *Q-network* sebagai estimator utama pada DQN membuat model rentan terhadap *overestimation* bias dan ketidakstabilan nilai Q [14] [15].

https://doi.org/10.25077/TEKNOSI.v12i1.2026.105-112



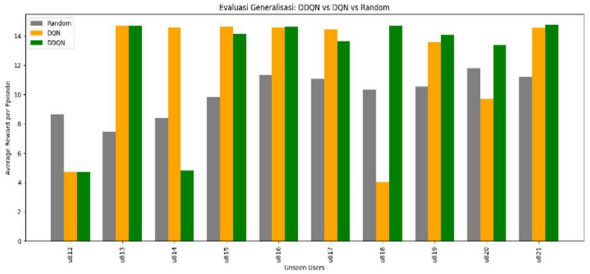
Gambar 3. DQN Episode *Reward*

Gambar 3 memperlihatkan tren *reward* yang diperoleh oleh model DDQN sepanjang pelatihan. Berbeda dengan DQN, model DDQN menunjukkan peningkatan reward yang lebih konsisten dan stabil. Hal ini mencerminkan kemampuan arsitektur DDQN dalam mengatasi bias estimasi melalui pemisahan antara jaringan policy dan target, yang secara teoritis telah terbukti mampu mengurangi ketidaktepatan estimasi nilai Q dan meningkatkan stabilitas proses pembelajaran [16] [17].

Kedua model menunjukkan kapabilitas dalam mempelajari kebijakan rekomendasi yang efektif, dengan rata-rata reward yang diperoleh melebihi 10 poin per episode. Hasil ini sejalan dengan studi terdahulu yang menyatakan bahwa model berbasis *Deep Reinforcement Learning* mampu beradaptasi terhadap interaksi siswa dan menghasilkan rekomendasi yang relevan secara personal [7]

3.2. Generalisasi Pada Pengguna Baru

Untuk mengukur kemampuan model dalam melakukan generalisasi terhadap data yang tidak tersedia selama proses pelatihan, dilakukan evaluasi terhadap sepuluh pengguna acak yang tidak pernah dilibatkan sebelumnya (*unseen users*). Evaluasi ini penting untuk menilai robustitas model dalam menghasilkan rekomendasi yang relevan di luar konteks pelatihan awal. *Metrik* yang digunakan adalah rata-rata reward per episode untuk masing-masing pengguna. Dalam pengujian ini, tiga pendekatan dibandingkan, yaitu *random policy* sebagai *baseline*, DQN, dan DDQN [8].

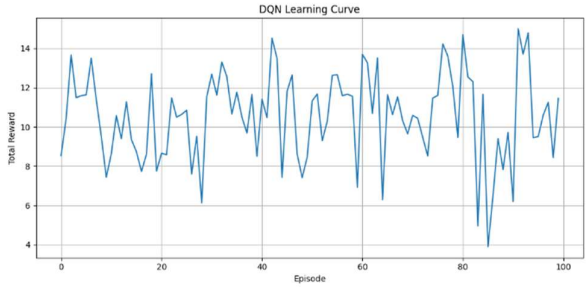


Gambar 4. Evaluasi Generalisasi DQN, DDQN, *Random*

Hasil eksperimen menunjukkan bahwa model DDQN secara konsisten mencatatkan *reward* tertinggi pada mayoritas pengguna uji, yaitu 7 dari 10 peserta. Hal ini menunjukkan bahwa DDQN memiliki kemampuan adaptasi yang lebih baik dalam mengakomodasi pola interaksi pengguna baru. Sementara itu, DQN menunjukkan performa yang relatif kompetitif namun disertai fluktuasi antar pengguna yang lebih tinggi. Random policy memberikan *reward* terendah secara konsisten, yang menegaskan pentingnya pembelajaran berbasis pengalaman historis dalam menghasilkan rekomendasi yang kontekstual [4]

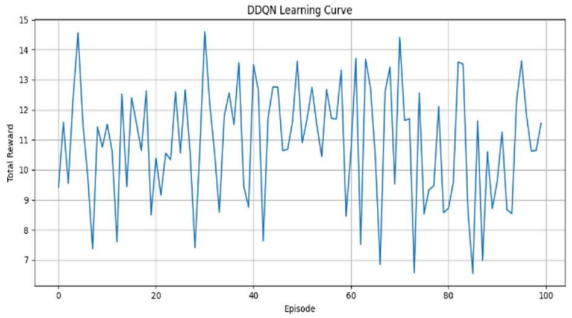
3.3. Analisis *Learning Curva*

Untuk mengevaluasi dinamika pembelajaran dari kedua model, dilakukan analisis terhadap kurva pembelajaran (*learning curve*) yang merepresentasikan akumulasi *reward* per episode selama proses pelatihan berlangsung. Kurva ini memberikan gambaran sejauh mana agen berhasil memperbaiki kebijakan rekomendasi soal berdasarkan umpan balik dari lingkungan simulasi. Gambar 5 dan Gambar 6 masing-masing menunjukkan kurva *reward* model DQN dan DDQN selama 100 episode pelatihan.



Gambar 5. DQN *Learning Curve*

Gambar 5 menunjukkan bahwa model DQN mengalami fluktuasi *reward* yang cukup tajam selama pelatihan. Ketidakstabilan ini disebabkan oleh kombinasi antara eksplorasi agresif pada tahap awal pelatihan dan keterbatasan arsitektur *single Q-network* dalam menangani pembaruan nilai Q secara akurat. Estimator tunggal seperti yang digunakan dalam DQN cenderung menghasilkan varians tinggi dan kurang mampu menyaring *noise* dari distribusi reward [15].



Gambar 6. DDQN *Learning Curve*

Sebaliknya, pada Gambar 6, model DDQN menunjukkan pola reward yang jauh lebih terstruktur dan stabil. *Reward* meningkat secara bertahap dan cenderung terkonsolidasi lebih cepat. Fenomena ini menunjukkan bahwa penggunaan dua jaringan terpisah *policy network* dan target *network* dalam arsitektur DDQN berhasil mengurangi *overestimation* bias dan meningkatkan stabilitas proses pembelajaran [17]. Dengan demikian, DDQN tidak hanya lebih efisien dalam mencapai konvergensi, tetapi juga lebih andal dalam mempertahankan kinerja yang stabil selama pelatihan.

3.4. Evaluasi Akurasi Prediksi

Sebagai metrik pelengkap untuk menilai efektivitas model dalam menghasilkan rekomendasi soal yang sesuai dengan tingkat kemampuan siswa, dilakukan evaluasi terhadap akurasi prediksi pada data pengguna yang tidak terlibat dalam proses pelatihan (*unseen users*). Pengukuran akurasi ini penting untuk memastikan bahwa agen tidak hanya unggul dalam memperoleh reward selama pelatihan, tetapi juga dapat memprediksi interaksi pengguna dengan ketepatan yang tinggi.

Tabel 3. Hasil Akurasi Prediksi Model

Model	Akurasi
DQN	74.00%
DDQN	78.00%

Berdasarkan hasil pada Tabel 3, model DDQN menunjukkan akurasi prediksi sebesar 78%, unggul 4% dibandingkan model DQN yang mencapai 74%. Meskipun selisih ini terlihat relatif kecil secara numerik, peningkatan tersebut memiliki signifikansi praktis dalam konteks sistem pembelajaran adaptif, di mana akurasi prediksi berbanding lurus dengan relevansi dan efektivitas rekomendasi konten. Ketepatan dalam pemilihan soal yang sesuai dapat meningkatkan keterlibatan siswa serta mempercepat pencapaian hasil belajar yang optimal [18].

4. PEMBAHASAN

Hasil evaluasi menunjukkan bahwa model DDQN secara konsisten memberikan performa yang lebih stabil dan efektif dibandingkan DQN. Hal ini terlihat dari pola *reward* per episode yang diperoleh selama pelatihan, di mana DDQN menunjukkan tren peningkatan *reward* yang lebih terstruktur. Temuan ini selaras dengan penelitian [17] dan [16] yang menyatakan bahwa arsitektur double estimator dalam DDQN mampu mengurangi bias estimasi nilai Q (*overestimation* bias) yang sering terjadi pada model berbasis *single estimator* seperti DQN. Sementara itu, fluktuasi *reward* yang tinggi pada DQN mengindikasikan lemahnya konsistensi pembaruan nilai Q, sebagaimana dilaporkan oleh [15].

Evaluasi terhadap pengguna baru (*unseen users*) memberikan bukti tambahan mengenai superioritas DDQN dalam aspek generalisasi. Model ini berhasil mempertahankan reward tinggi pada sebagian besar pengguna yang tidak dilibatkan dalam pelatihan, mengindikasikan kemampuannya dalam mengenali dan mengadaptasi pola interaksi yang baru. Studi oleh [13] memperkuat temuan ini dengan menyatakan bahwa pemisahan antara *policy network* dan target *network* dalam DDQN berkontribusi signifikan terhadap kestabilan nilai estimasi Q dalam konteks data yang belum dikenali.

Meskipun DQN menunjukkan performa yang kompetitif, hasil *reward* yang diperoleh cenderung lebih fluktuatif ketika diuji terhadap pengguna baru. Hal ini menandakan bahwa model tersebut belum sepenuhnya robust terhadap variasi karakteristik pengguna di luar domain pelatihan. [19] menyatakan bahwa pendekatan single estimator lebih rentan terhadap *overfitting* dan bias estimasi, sehingga kurang optimal dalam situasi yang menuntut fleksibilitas model.

Sementara itu, hasil dari pendekatan *random policy* secara konsisten menunjukkan *reward* terendah. Ketidakterlibatan elemen pembelajaran dalam strategi acak ini menyebabkan rendahnya relevansi antara soal yang direkomendasikan dan kebutuhan aktual pengguna. Temuan ini juga sejalan dengan [17], yang menunjukkan bahwa sistem berbasis DRL secara signifikan lebih unggul dibanding pendekatan non-adaptif dalam lingkungan yang kompleks dan dinamis.

Dengan demikian, hasil generalisasi ini menguatkan posisi DDQN sebagai model yang tidak hanya unggul dalam proses pelatihan internal, tetapi juga memiliki kapabilitas untuk melakukan transfer pembelajaran ke skenario baru secara efektif. Keunggulan ini menjadikan DDQN sebagai kandidat yang layak untuk diimplementasikan dalam sistem pembelajaran adaptif berbasis data nyata, seperti *Learning Management System* (LMS), di mana kemampuan generalisasi merupakan aspek krusial terhadap keberhasilan sistem [16]

Temuan dari analisis learning curve turut memperkuat perbedaan mendasar antara DQN dan DDQN dalam hal stabilitas pelatihan. Kurva *reward* DQN menunjukkan fluktuasi tinggi yang mencerminkan ketidakkonsistenan proses pembelajaran, terutama akibat penggunaan estimator tunggal yang tidak mampu menyaring *noise* secara efektif. Hal ini diperkuat oleh[15], yang menyebutkan bahwa DQN sering memerlukan teknik tambahan, seperti *multi-step bootstrapping* atau *prioritized experience replay*, untuk menstabilkan pelatihan.

Sebaliknya, performa DDQN yang lebih stabil konsisten dengan desain arsitektur dua jaringan yang memang dikembangkan untuk mengatasi kelemahan utama DQN. Penggunaan target *network* yang diperbarui secara terpisah dari *policy network* terbukti mampu menjaga nilai estimasi Q agar tetap konservatif dan terhindar dari pembaruan ekstrim yang berisiko. [16] dan [18] menyimpulkan bahwa pendekatan ini tidak hanya mempercepat proses konvergensi, tetapi juga meningkatkan akurasi jangka panjang secara signifikan.

Berdasarkan hasil evaluasi *learning curve*, dapat disimpulkan bahwa DDQN lebih unggul dalam aspek stabilitas dan efisiensi pelatihan. Karakteristik ini menjadikan DDQN sebagai model yang lebih tepat untuk implementasi di sistem pembelajaran adaptif skala besar, terutama yang menuntut performa konsisten dalam jangka panjang.

Akurasi prediksi juga menjadi aspek krusial dalam mengukur keberhasilan sistem rekomendasi edukatif. Pada pengujian ini, DDQN menunjukkan akurasi sebesar 78%, unggul 4% dibandingkan DQN yang mencatatkan akurasi sebesar 74%. Meskipun perbedaan tersebut terlihat kecil secara numerik, dalam konteks sistem pembelajaran berbasis personalisasi, selisih tersebut sangat signifikan. Ketepatan dalam memilih soal yang sesuai dengan kemampuan siswa sangat berpengaruh terhadap efektivitas proses belajar [18].

Akurasi prediksi yang lebih tinggi pada DDQN juga menunjukkan bahwa *reward* tinggi yang diperoleh selama pelatihan tidak hanya mencerminkan kecocokan terhadap data pelatihan, tetapi juga berdampak positif terhadap performa

https://doi.org/10.25077/TEKNOSI.v12i1.2026.105-112

prediktif terhadap pengguna baru. Hal ini memperkuat argumen bahwa DDQN memiliki generalisasi yang kuat secara struktural. Studi oleh [17] menegaskan bahwa peningkatan kecil dalam akurasi model dapat memberikan pengaruh besar terhadap pengalaman belajar, kepuasan pengguna, dan efisiensi sistem secara keseluruhan.

Oleh karena itu, dari berbagai aspek yang telah dievaluasi *reward*, generalisasi, kestabilan pelatihan, dan akurasi prediksi model DDQN secara konsisten menunjukkan performa yang superior dibandingkan DQN. Keunggulan ini membuktikan bahwa DDQN sangat layak untuk diimplementasikan sebagai algoritma dasar dalam sistem pembelajaran adaptif berbasis *Deep Reinforcement Learning.*

5. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi penerapan DRL dalam sistem pembelajaran adaptif berbasis performa siswa, dengan membandingkan dua arsitektur model, yaitu DQN dan DDQN. Hasil eksperimen menunjukkan bahwa model DDQN secara konsisten memberikan kinerja yang lebih unggul dibandingkan DQN maupun strategi *baseline random policy*. Keunggulan DDQN terlihat dari rata-rata *reward* per episode yang lebih tinggi dan stabil selama pelatihan, mengindikasikan efektivitas model dalam menyusun kebijakan rekomendasi soal yang relevan dengan kemampuan siswa.

Selain itu, DDQN menunjukkan performa yang lebih baik dalam aspek generalisasi, dengan menghasilkan *reward* tertinggi pada 7 dari 10 pengguna baru yang tidak terlibat dalam proses pelatihan. Kemampuan ini mencerminkan bahwa DDQN memiliki kapabilitas dalam mengenali dan menyesuaikan diri terhadap pola pembelajaran yang belum dikenali sebelumnya. Dari sisi kurva pembelajaran, DDQN memperlihatkan konvergensi yang lebih cepat dan fluktuasi *reward* yang lebih rendah, menunjukkan stabilitas proses pelatihan yang lebih andal dibandingkan DQN.

Dalam aspek akurasi prediksi, DDQN mencatatkan nilai sebesar 78%, mengungguli DQN yang hanya mencapai 74%. Walaupun selisih akurasi terlihat kecil secara numerik, peningkatan tersebut signifikan dalam konteks sistem pembelajaran adaptif yang menuntut ketepatan tinggi dalam pemilihan konten. Seluruh proses pelatihan dan evaluasi berhasil dilakukan dalam lingkungan simulasi KTEnv berbasis data EdNet-KT1, yang merepresentasikan kondisi nyata dalam interaksi pembelajaran daring. Berdasarkan temuan ini, DDQN direkomendasikan sebagai pendekatan yang lebih efektif dan layak diimplementasikan dalam pengembangan sistem pembelajaran adaptif berbasis DRL.

.

DAFTAR PUSTAKA

[1] V. Mirata, F. Hirt, P. Bergamin, and C. van der Westhuizen, "Challenges and contexts in establishing adaptive learning in higher education: findings from a Delphi study," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, Dec. 2020, doi: 10.1186/s41239-020-00209-y.

[2] I. Katsaris and N. Vidakis, "Adaptive e-learning systems through learning styles: A review of the literature," *Adv Mobile Learn Educ Res*, vol. 2021, no. 2, pp. 124–145, 2021, doi: 10.25082/AMLER.2021.02.007.

[3] Xiao Li, Hanchen Xu, Jinming Zhang, and Hua-hua Chang, "Psychometrika Submission," 2020.

[4] R. Mustapha, G. Soukaina, Q. Mohammed, and A. Es-Sâadia, "Towards an Adaptive e-Learning System Based on Deep Learner Profile, Machine Learning Approach, and Reinforcement Learning." [Online]. Available: www.ijacsa.thesai.org

[5] Z. Li, F. Hu, S. Qi, R. Hu, Y. Zhou, and Y. Bai, "Deformation characteristics of the shear zone and movement of block stones in soil-rock mixtures based on large-sized shear test," *Applied Sciences (Switzerland)*, vol. 10, no. 18, Sep. 2020, doi: 10.3390/APP10186475.

[6] M. Abdelshiheed, J. W. Hostetter, T. Barnes, and M. Chi, "Leveraging Deep Reinforcement Learning for Metacognitive Interventions across Intelligent Tutoring Systems," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.09821

[7] Y. Choi *et al.*, "EdNet: A Large-Scale Hierarchical Dataset in Education," Jul. 2020, [Online]. Available: http://arxiv.org/abs/1912.03072

[8] P. Tam, S. Math, A. Lee, and S. Kim, "Multi-agent deep q-networks for efficient edge federated learning communications in software-defined iot," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3319–3335, 2022, doi: 10.32604/cmc.2022.023215.

[9] Y. Li, T. Guo, Q. Li, and X. Liu, "Optimized Feature Extraction for Sample Efficient Deep Reinforcement Learning," *Electronics (Switzerland)*, vol. 12, no. 16, Aug. 2023, doi: 10.3390/electronics12163508.

[10] B. S. Ciftler, M. Abdallah, A. Alwarafy, and M. Hamdi, "DQN-Based Multi-User Power Allocation for Hybrid RF/VLC Networks," in *IEEE International Conference on Communications*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021. doi: 10.1109/ICC42927.2021.9500564.

[11] F. Rasheed, K. L. A. Yau, R. M. Noor, C. Wu, and Y. C. Low, "Deep Reinforcement Learning for Traffic Signal Control: A Review," *IEEE Access*, vol. 8, pp. 208016–208044, 2020, doi: 10.1109/ACCESS.2020.3034141.

[12] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction Second edition, in progress."

[13] A. Schuderer, S. Bromuri, and M. van Eekelen, "Sim-Env: Decoupling OpenAI Gym Environments from Simulation Models," Feb. 2021, doi: 10.1007/978-3-030-85739-4_39.

[14] S. Kumar, "Balancing a CartPole System with Reinforcement Learning -- A Tutorial," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.04938

[15] A. Ly, R. Dazeley, P. Vamplew, F. Cruz, and S. Aryal, "Elastic Step DQN: A novel multi-step algorithm to alleviate overestimation in Deep QNetworks," Oct. 2022, [Online]. Available: http://arxiv.org/abs/2210.03325

[16] Z. Ren, G. Zhu, H. Hu, B. Han, J. Chen, and C. Zhang, "On the Estimation Bias in Double Q-Learning," Jan. 2022, [Online]. Available: http://arxiv.org/abs/2109.14419

[17] P. Liu, C. Yao, C. Li, S. Zhang, and X. Li, "A Caching-Enabled Permissioned Blockchain Scheme for Industrial Internet of Things Based on Deep Reinforcement Learning," *Wirel Commun Mob Comput*, vol. 2023, 2023, doi: 10.1155/2023/2852085.

[18] Z. Zhou, C. Allen, K. Asadi, and G. Konidaris, "Characterizing the Action-Generalization Gap in Deep Q-Learning," May 2022, [Online]. Available: http://arxiv.org/abs/2205.05588

[19] X. Zhang, "Application and Optimization of Reinforcement Learning Based on Deep Q-Network (DQN) in Complex Environments," 2025.

BIODATA PENULIS

Fahad Abdul Aziz
Merupakan mahasiswa Program Studi Sarjana Terapan Informatika di Universitas Logistik dan Bisnis Internasional. Saat ini sedang menyelesaikan tugas akhir yang berfokus pada penerapan model DRL dalam sistem pembelajaran adaptif.

M. Yusril Helmi Setyawan
Merupakan dosen tetap pada Program Studi Sarjana Terapan Informatika di Universitas Logistik dan Bisnis Internasional. Menyelesaikan pendidikan S1 di bidang Teknik Informatika di STMIK Tasikmalaya dan S2 pada bidang Sistem Informasi di STMIK LIKMI Bandung.

Cahyo Prianto
Merupakan dosen pada Program Studi Sarjana Terapan Informatika di Universitas Logistik dan Bisnis Internasional. Meraih gelar Sarjana Pendidikan Fisika dari Universitas Pendidikan Indonesia (UPI) dan Magister Teknik Elektro dari Institut Teknologi Bandung (ITB). Saat ini sedang menempuh studi doktoral (S3) di bidang Teknik Informatika di UPI.