Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 5 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i05.60003

PAPER

Offline Reinforcement Learning for Sepsis Management: A Dueling Double Deep Q-Network Approach

Endah Purwanti(✉), Fatima Hasya Puspa Kasih, Franky Chandra Satria Arisgraha

Biomedical Engineering, Universitas Airlangga, Surabaya, Indonesia

endah-p-1@fst.unair.ac.id

ABSTRACT

Sepsis requires rapid and individualized management of intravenous (IV) fluids and vasopressors, yet treatment strategies vary widely in clinical practice. This study develops an offline reinforcement learning (RL) framework based on a dueling double deep Q-network (DDQN) to model dosing policies for adult sepsis patients in the intensive care unit (ICU). Following preprocessing of a large ICU cohort under Sepsis-3 criteria, patient states were represented using 48 clinical variables, and actions were defined as 25 discrete IV fluid–vasopressor combinations. Reward estimation incorporated changes in SOFA score and lactate, with terminal rewards reflecting survival outcomes. The agent was trained using KL-regularized offline RL and evaluated using weighted importance sampling (WIS), fitted Q-evaluation (FQE), and weighted doubly robust (WDR) estimators. The selected model (β_KL = 0.0005) achieved higher estimated returns than the historical clinician policy across all OPE metrics on the held-out test set (WIS 9.18 vs. 8.60; FQE 18.42 vs. 17.76; WDR 17.80 vs. 17.50). Policy distribution analysis indicated differences in treatment allocation patterns across dosing bins. Permutation feature importance (PFI) identified systolic blood pressure, arterial pH, sodium, and INR among the most influential variables. These findings support the feasibility of offline RL for modeling treatment policies in sepsis management and motivate further validation in prospective or simulation-based settings.

KEYWORDS

sepsis, intensive care unit (ICU), offline reinforcement learning, dueling double deep Q-network (DDQN), clinical decision support systems, electronic health records (EHR)

1 INTRODUCTION

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection and remains a major global health burden [1]. Worldwide, it accounts for approximately 48–49 million cases and 11 million deaths annually, representing nearly 20% of all deaths [2], [3]. Despite advances in antimicrobial therapy, organ support, and standardized care bundles, sepsis continues to be a leading cause of ICU admission and mortality [4], [5].

Purwanti, E., Kasih, F. H. P., Arisgraha, F. C. S. (2026). Offline Reinforcement Learning for Sepsis Management: A Dueling Double Deep Q-Network Approach. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(5), pp. 156–170. https://doi.org/10.3991/ijoe.v22i05.60003

Article submitted 2025-12-16. Revision uploaded 2026-02-27. Final acceptance 2026-02-27.

The Sepsis-3 consensus emphasizes organ dysfunction, operationalized by an increase in the sequential organ failure assessment (SOFA) score of ≥2 points to identify patients at high risk of mortality [1]. Although early recognition, timely antibiotics, and appropriate volume resuscitation are essential, rigid protocolized approaches may not outperform individualized high-quality care [4–6]. Current guidelines recommend target-driven resuscitation and vasopressor titration while allowing clinician judgment in fluid and vasopressor management [4]. However, heterogeneous physiology and competing risks of under- or over-resuscitation complicate treatment decisions, motivating the development of data-driven strategies for individualized dosing in sepsis management.

Reinforcement learning (RL) provides a framework for learning sequential decision policies that maximize long-term reward from observed trajectories [7], [8]. Offline RL (batch RL) enables policy learning from historical clinical data while addressing safety and ethical constraints in medicine [7], [12]. Prior studies have explored RL for sepsis management, including the "Artificial Intelligence Clinician," which applied off-policy evaluation to infer fluid and vasopressor policies from retrospective ICU data [9], and subsequent deep RL approaches using continuous state-space models [10], [11].

However, much of the existing work has emphasized feasibility and simulated performance, with limited attention to conservative offline learning strategies, multi-estimator off-policy evaluation, and structured interpretability within clinically realistic discrete treatment spaces.

Recent advances in offline RL have introduced conservative and behavior-regularized approaches to mitigate distributional shift and improve safety in healthcare settings [18], [19]. Our KL-regularized formulation aligns with these developments while focusing on interpretable discrete treatment modeling.

Deep Q-network (DQN) enables learning in high-dimensional domains but is prone to overestimation bias [13], [15]. This limitation is addressed by double DQN (DDQN), which decouples action selection and evaluation, and further improved by the Dueling architecture that separates state-value and advantage streams to enhance stability and value estimation [13], [14]. Combining these approaches yields the dueling double DQN (dueling DDQN), well suited for high-dimensional sepsis states with a discrete yet rich fluid–vasopressor action space [10], [11].

At the same time, trust in RL-driven clinical decision support requires transparent evaluation and interpretability. Recent work on *off-policy evaluation* (OPE) provides tools for estimating policy value using observational data [19], [24–26]. Model-agnostic permutation feature importance (PFI), originally introduced for random forests and later generalized, offers a principled way to quantify how each predictor influences model performance, thereby improving clinical interpretability [16–17], [20–21].

From an online engineering perspective, modern ICU environments operate within interconnected electronic health record (EHR) systems and real-time digital infrastructures. Recent studies highlight the growing integration of machine learning into digital health ecosystems, including neural-network–based cardiac risk detection and AI-enabled mobile diagnostic platforms [27], [28]. Reinforcement learning–based treatment frameworks must therefore remain compatible with networked clinical decision-support systems (CDSS). Offline reinforcement learning provides a safety-aware mechanism for learning from historical data while enabling potential integration into online ICU platforms under clinician supervision. This positioning situates the present work at the intersection of biomedical and online engineering domains.

In this study, we develop an offline RL framework for adult sepsis management based on a dueling DDQN agent that recommends IV fluid and vasopressor dosing strategies in the ICU. Patient states are represented using 48 clinical features, and

actions correspond to discretized fluid–vasopressor combinations [29]. The reward incorporates changes in organ failure and lactate, with terminal components reflecting hospital survival [10], [11]. The agent is trained using KL-regularized offline RL and evaluated with multiple off-policy estimators (WIS, FQE, and WDR) [24–26]. PFI is used to assess physiological drivers of the learned policy. Specifically, this study makes three contributions: (i) integrating KL-regularized dueling DDQN for conservative offline policy learning; (ii) employing multiple off-policy evaluation estimators for robust value assessment; and (iii) enhancing interpretability through permutation feature importance analysis. Our overarching goal is to evaluate whether the proposed agent can learn clinically interpretable strategies associated with improved estimated returns relative to historical treatment.

2 METHODS

2.1 Study design and data source

This retrospective study utilized the MIMIC-III v1.4 database, which contains detailed clinical data from over 50,000 ICU admissions at a tertiary academic hospital in the United States [22]. Data were accessed via Google BigQuery and processed using Python. Adult ICU stays (≥18 years) with complete outcome documentation were included.

2.2 Sepsis cohort definition and episode construction

Sepsis episodes were identified according to Sepsis-3 criteria, defined as suspected infection with a ≥2-point increase in SOFA score. Suspected infection was operationalized by the temporal proximity between intravenous antibiotic administration and blood culture sampling. The earliest qualifying timestamp was defined as sepsis onset.

Exclusion criteria included ICU length of stay <24 hours, age <18 years, missing outcome labels, and early treatment withdrawal. Clinical observations were resampled into fixed 4-hour windows spanning –24 to +56 hours relative to sepsis onset, yielding up to 20 decision points per patient [9], [23]. Each patient trajectory was represented as sequential state–action–reward transitions suitable for reinforcement learning. The overall workflow is illustrated in Figure 1.

Fig. 1. Overview of the study workflow

2.3 State representation and preprocessing

Each 4-hour window was encoded as a 48-dimensional state vector including demographic variables, vital signs, laboratory measurements, ventilation parameters, and fluid balance indicators (refer to Table 1). Missing values were imputed using k-nearest neighbors. Variables approximating Gaussian distributions were standardized using z-scores, whereas skewed variables were log-transformed

prior to scaling. Binary variables were mean-centered. These preprocessing steps improved numerical stability and feature comparability across heterogeneous clinical measurements [30].

Table 1. 48 clinical indicators as state representation

Category	Count	Features
Demographics	5	Age, Gender, Weight, Elixhauser Score, ICU readmission
Vital Signs	11	SOFA, SIRS, GCS, Heart Rate (HR), Systolic Blood Pressure (SysBP), Diastolic Blood Pressure (DiaBP), Mean Arterial Pressure (MeanMP), Shock Index, Oxygen Saturation (SpO_2), Body Temperature, Respiratory Rate (RR)
Laboratory Results	28	Potassium, Sodium, Chloride, Glucose, Creatinine, Magnesium, Calcium, SGOT, Ionized Calcium, HCT (Hematocrit), SGPT, Bilirubin, Albumin, Hemoglobin, White Blood Cell Count (WBC Count), PTT, PT, BUN, Platelet Count, INR, pH, PaO_2, $PaCO_2$, Base Excess, Lactate, PaO_2/FiO_2 Ratio, HCO_3
Ventilation	2	Mechanical Ventilation (MechVent), Fraction of Inspired Oxygen (FiO_2)
Fluid Status	2	Urine Output, Cumulative Fluid Balance (CFB)

2.4 Action space specification

Two therapeutic interventions—intravenous fluid administration and vasopressor dose—were extracted for each four-hour interval. Both variables were discretized into quartiles (bins 0–4), and their Cartesian product generated 25 discrete treatment actions. This discretization preserves clinically observed dosing patterns while maintaining interpretability. The resulting action grid is shown in Figure 2, with detailed bin definitions provided in Tables 2–3.

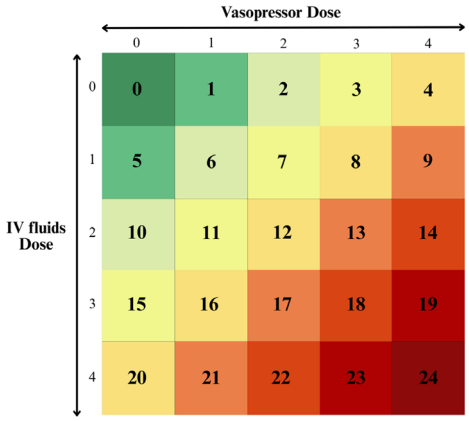

Fig. 2. Discrete action space formed by the Cartesian product of binned intravenous fluid volumes and vasopressor doses (quartiles 0–4), resulting in 25 treatment combinations

2.5 Reward function

The reward function was designed to capture short-term physiological improvement and long-term survival outcomes. Intermediate rewards were defined based on temporal changes in SOFA score and lactate levels:

$$\Delta SOFA = SOFA_{t+1} - SOFA_t,\ \Delta Lactate = Lactate_{t+1} - Lactate_t$$

The stepwise was defined as:

$$r_t(s_t, a_t) = C_0 1(\Delta SOFA > 0) + C_1 \Delta SOFA + C_2 \tanh(\Delta Lactate) \quad (1)$$

where C_0, C_1, C_2 are scaling coefficients. A terminal reward of +15 was assigned for survival and –15 for in-hospital mortality. The agent optimized the discounted cumulative return:

$$G_t = \sum_{k=0}^{T} \gamma^k r_{t+k}$$

2.6 Dueling double deep Q-network agent

We implemented a dueling double deep Q-network (DDQN) agent to approximate the action–value function $Q(s, a)$ over the discrete treatment action space [10], [11]. The network architecture consisted of two fully connected hidden layers with 384 and 256 units, respectively, each followed by rectified linear unit (ReLU) activation and batch normalization. The dueling architecture decomposes the Q-function into state-value and advantage components:

$$Q(s,a;\theta) = V(s;\theta) + \left(A(s,a;\theta) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a';\theta) \right) \quad (2)$$

where $V(s)$ denotes the estimated state-value function, $A(s, a)$ the advantage function, and $|\mathcal{A}|$ the number of discrete actions.

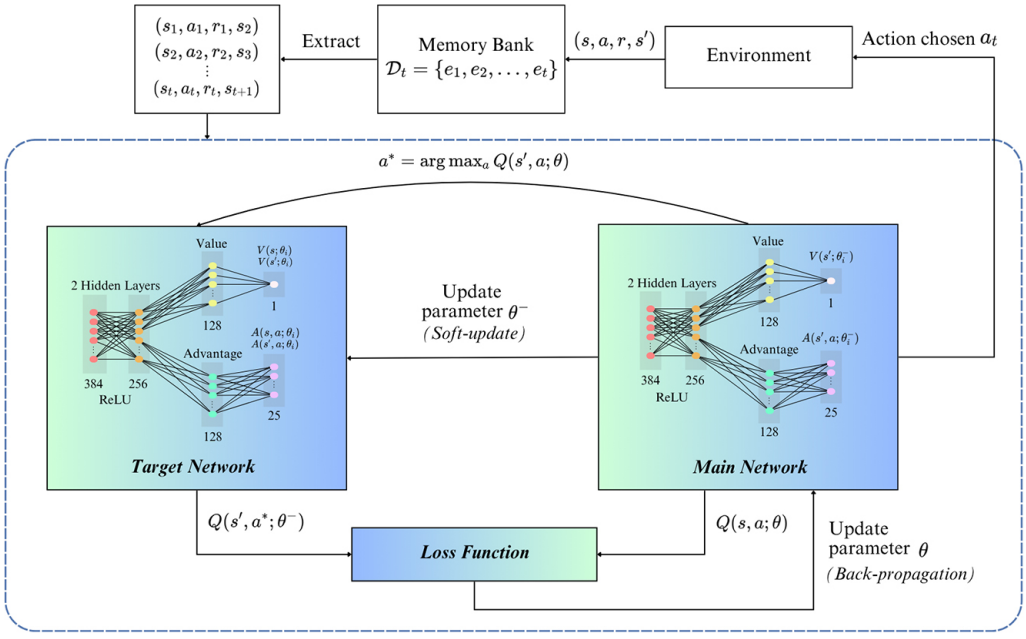

Fig. 3. Dueling Double DQN training pipeline illustrating replay-buffer sampling, value–advantage decomposition, temporal-difference optimization (Equation 3), and soft parameter updates from the main network to the target network ($\tau = 0.005$)

Figure 3 summarizes the training pipeline, including replay-buffer sampling, value–advantage decomposition, double DQN target computation, and soft parameter updates between the main and target networks.

In double DQN, the greedy next action is selected using the main network and evaluated using the target network:

$$a* = \arg\max_{a'} Q(s', a'; \theta)$$

The temporal-difference (TD) loss is then defined as:

$$L(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(r + \gamma Q(s', a*; \theta^{-}) - Q(s,a;\theta)\right)^2\right] \quad (3)$$

where $\gamma \in (0, 1]$ is the discount factor, θ and θ^{-} denote the parameters of the main and target networks, respectively.

A separate main network was updated via backpropagation, while the target network parameters were updated using a soft-update mechanism with a rate $\tau = 0.005$ [31].

To mitigate extrapolation error in the offline setting, KL-regularized Q-learning was applied:

$$L_{total}(\theta) = L(\theta) + \beta_{KL}\, \mathbb{E}_{s\sim\mathcal{D}}[KL(\pi(\cdot \mid s) \,\|\, \pi_b(\cdot \mid s))] \quad (4)$$

where $L(\theta)$ denotes the standard temporal-difference loss; β_{KL} is the KL-regularization coefficient controlling the degree of policy conservatism; $\pi(\cdot \mid s)$ represents the learned policy parameterized by θ; $\pi_b(\cdot \mid s)$ denotes the behavior policy inferred from historical clinician data; D is the empirical state distribution from the offline dataset; and $KL(\cdot \,\|\, \cdot)$ denotes the Kullback–Leibler divergence between the learned and behavior policy distributions.

Training used the Adam optimizer with a learning rate 1×10^{-4}, discount factor $\gamma = 0.97$, batch size 128, and 100 epochs. Episodes were split at the patient level (80% training, 10% validation, 10% testing) to avoid overlap.

2.7 Off-policy evaluation

Policy value was estimated using three established off-policy evaluation (OPE) methods: Weighted Importance Sampling (WIS), Fitted Q-Evaluation (FQE), and Weighted Doubly Robust (WDR). These methods estimate the expected return of a target policy π under data generated by a behavior policy π_b, based on samples $(s,a,r,s') \sim D$. In general, OPE aims to estimate:

$$V^{\pi} = \mathbb{E}_{\tau\sim\pi}[G_{\tau}]$$

where G_{τ} denotes the discounted return of trajectory τ. WIS, FQE, and WDR provide complementary bias–variance trade-offs for estimating V^{π} in the offline setting.

2.8 Feature importance and policy analysis

Permutation feature importance was used to quantify the contribution of each state variable to the learned policy. For each feature x_j, importance was defined as the reduction in the OPE-estimated policy value after random permutation:

$$PFI(x_j) = V^{\pi} - V^{\pi}_{perm(x_j)}$$

where $V^{\pi}_{\mathrm{perm}(x_j)}$ denotes the estimated policy value after permuting feature x_j. The 15 most influential features are presented in Figure 8.

Policy behavior was further analyzed by comparing clinician and RL action distributions across all 25 discrete bins (Figure 7).

3 RESULTS AND DISCUSSIONS

3.1 Pre-processing Outcomes and Cohort Formation

The pre-processing workflow converted raw ICU records into structured trajectories for offline reinforcement learning. Of 26,299 patients with suspected sepsis, 24,948 had data within the ±80-hour window surrounding sepsis onset. After applying exclusion criteria, the final cohort comprised 18,830 adults contributing 364,563 four-hour windows. Window counts varied substantially across patients, reflecting heterogeneous ICU trajectories and resulting in non-uniform episode lengths. Window duration and episode structure were standardized to ensure consistent trajectory representation.

3.2 Clinical Profile of the Final Sepsis Cohort

Figure 4 summarizes the baseline characteristics of the final cohort. The population consisted primarily of older adults (mean age 65.16 ± 16.27 years) with slight male predominance (55.7%). Hospital survival was 73.6%.

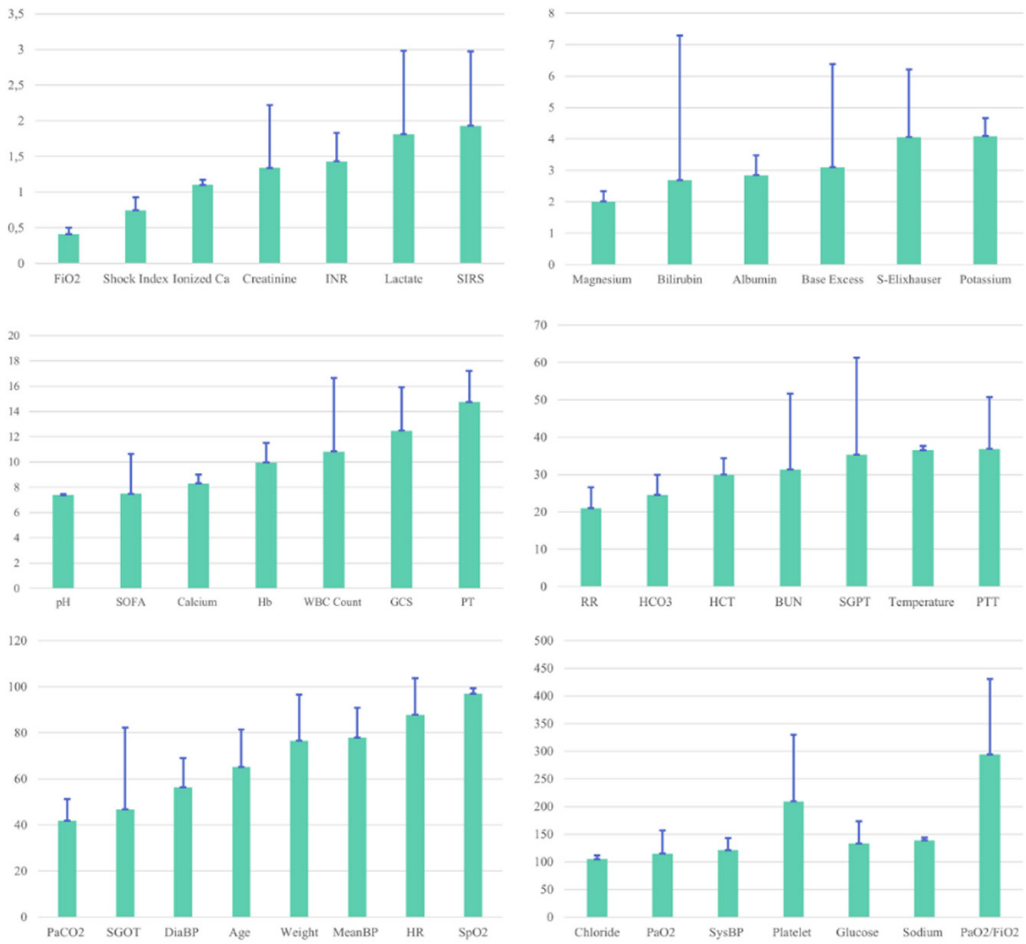

Fig. 4. Baseline distribution of key clinical variables in the final sepsis cohort (mean ± variability indicators)

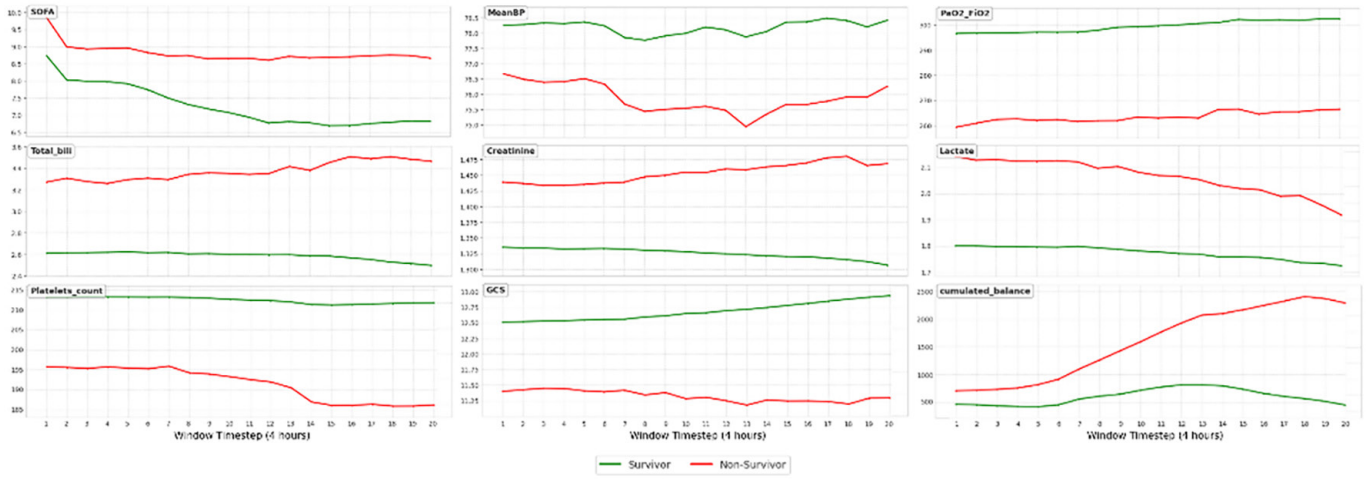

Fig. 5. Temporal trends of selected state variables in survivors (green) and non-survivors (red) across the early trajectory window

Illness severity at sepsis onset varied substantially, with a mean SOFA score of 7.49 (SD 3.15), spanning moderate to advanced organ dysfunction. Laboratory variables exhibited heterogeneous dispersion patterns. Some markers (e.g., electrolytes and renal parameters) showed relatively tight distributions, whereas others—including bilirubin (CV 171%), SGOT, SGPT, platelet count, and PaO_2/FiO_2—displayed greater variability. Figure 5 illustrates differences in selected state variables between survivors and non-survivors across the early trajectory window.

3.3 State representation and physiologic separation

Each four-hour window was encoded as a 48-dimensional state vector. After imputation and normalization, the distributions of selected features were compared between survivors and non-survivors (see Figure 5). Several variables demonstrated noticeable separation, including SOFA score, mean blood pressure, bilirubin, creatinine, lactate, platelet count, and GCS. These differences indicate that clinically relevant physiological variation is reflected within the constructed state space.

3.4 Action space design

The clinical interventions—IV fluid administration and vasopressor dosing—were discretized into quartiles, resulting in a 25-action grid (refer to Tables 2–3). The discretization reflects observed dosing ranges in the cohort and defines the discrete treatment space used for policy learning.

Table 2. Quartile discretization of IV fluids and vasopressor

Bin	IV fluids (mL/4 hours)	Vasopressor (μg/4 hours)
0	0	0
1	(0, 1112.73]	(0, 0.08]
2	(1112.73, 2956.74]	(0.08, 0.20]
3	(2956.74, 6009.85]	(0.20, 0.45]
4	> 6009.85	> 0.45

Table 3. Dose range interpretation per action label

a_n	Interpretation
0	No Vasopressor and No IV Fluids Administered
1	Given 0 – 0.08 μg Vasopressor without IV Fluids Administration
2	Given 0.08 – 0.20 μg Vasopressor without IV Fluids Administration
3	Given 0.20 – 0.45 μg Vasopressor without IV Fluids Administration
4	Given > 0.45 μg Vasopressor without IV Fluids Administration
5	Given 0 – 1112.73 mL IV fluids without Vasopressor Administration
6	Given 0 – 0.08 μg Vasopressor and 0 – 1112.73 mL IV fluids
7	Given 0.08 – 0.20 μg Vasopressor and 0 – 1112.73 mL IV fluids
8	Given 0.20 – 0.45 μg Vasopressor dan 0 – 1112.73 mL IV fluids
9	Given > 0,45 μg Vasopressor dan 0 – 1112.73 mL IV fluids
10	Given 1112.73 – 2956.74 mL IV fluids without Vasopressor Administration
11	Given 0 – 0.08 μg Vasopressor and 1112.73 – 2956.74 mL IV fluids
12	Given 0.08 – 0.20 μg Vasopressor and 1112.73 – 2956.74 mL IV fluids
13	Given 0.20 – 0.45 μg Vasopressor and 1112.73 – 2956.74 mL IV fluids
14	Given > 0.45 μg Vasopressor and 1112.73 – 2956.74 mL IV fluids
15	Given 2956.74 – 6009.85 mL IV fluids without Vasopressor Administration
16	Given 0 – 0.08 μg Vasopressors and 2956.74 – 6009.85 mL IV fluids
17	Given 0.08 – 0.20 μg Vasopressor and 2956.74 – 6009.85 mL IV fluids
18	Given 0.20 – 0.45 μg Vasopressors and 2956.74 – 6009.85 mL IV fluids
19	Given > 0.45 μg Vasopressors and 2956.74 – 6009.85 mL IV fluids
20	Given > 6009.85 mL IV fluids without Vasopressor Administration
21	Given 0 – 0.08 μg Vasopressor and > 6009.85 mL IV fluids
22	Given 0.08 – 0.20 μg Vasopressor and > 6009.85 mL IV fluids
23	Given 0.20 – 0.45 μg Vasopressor and > 6009.85 mL IV fluids
24	Given > 0.45 μg Vasopressor and > 6009.85 mL IV fluids

3.5 Reward dynamics and clinical interpretability

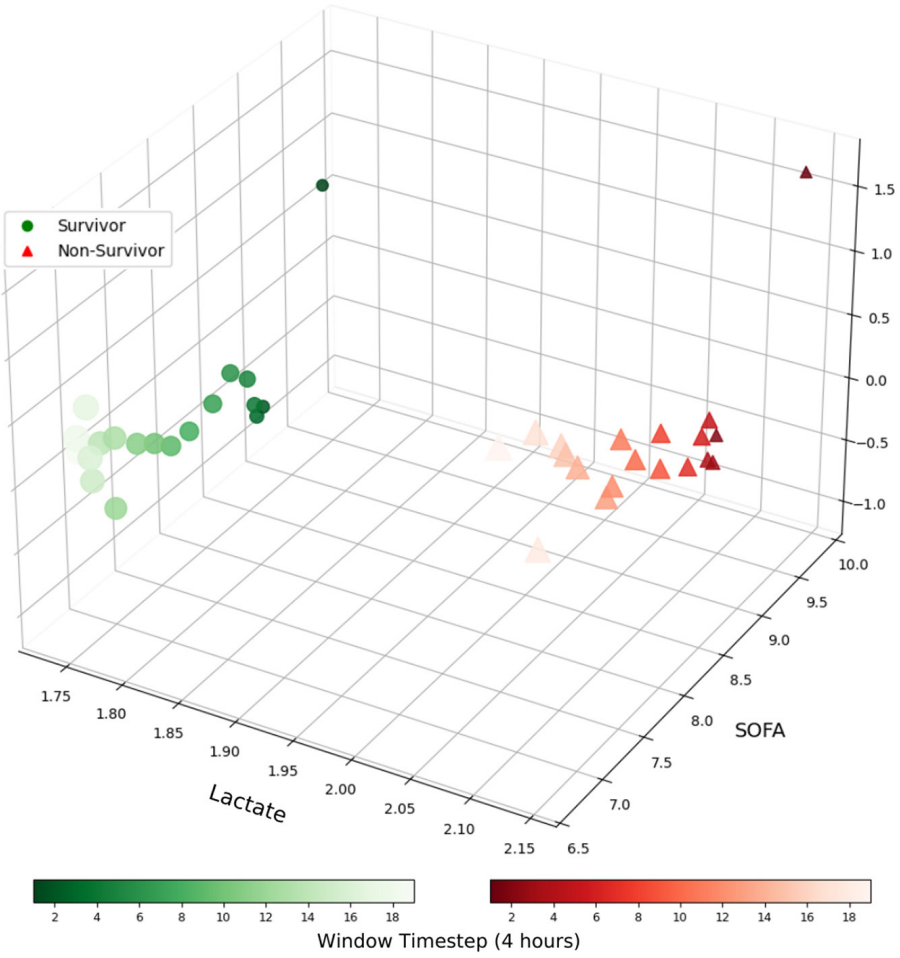

Fig. 6. Reward trajectories across time steps for survivors and non-survivors

Reward trajectories differed between survivors and non-survivors (see Figure 6). Survivors exhibited progressively increasing intermediate rewards, corresponding to improvements in SOFA and lactate over time, whereas non-survivors showed declining reward trends. The resulting net reward separation—+0.949 in survivors versus –0.685 in non-survivors—demonstrates distinct trajectory patterns between outcome groups. These differences indicate that the reward formulation captures clinically meaningful physiological variation across trajectories.

3.6 Training dynamics and model selection

The five KL-regularized dueling DDQN variants displayed distinct learning patterns during training, as shown in Figure 7, which illustrates the evolution of loss and Q-values across epochs. These trajectories reflect the influence of the KL regularization parameter in balancing exploration and adherence to clinician-like behavior.

At the highest regularization level ($\beta\_KL = 0.005$), learning dynamics were strongly constrained, with elevated loss values and limited Q-value growth. Conversely, minimal regularization ($\beta\_KL = 0.00025$) produced higher Q-values but increased variability in later epochs.

Among the tested configurations, β_KL = 0.0005 demonstrated comparatively stable learning dynamics. Loss decreased smoothly and converged to a lower final value, while Q-values increased steadily and stabilized over training. This configuration was therefore selected for subsequent evaluation.

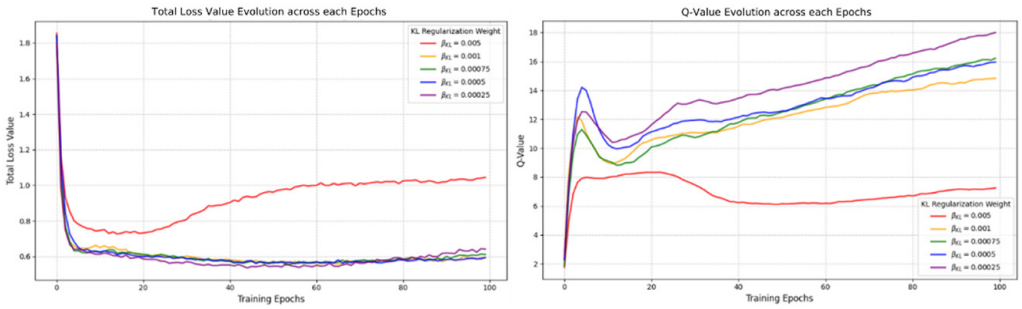

Fig. 7. Evolution of total loss and Q-values across training epochs for different KL-regularization levels

3.7 Policy behavior: Clinicians vs. RL agents

As shown in Figure 8, the RL agents adopted treatment patterns that differed from historical clinician behavior.

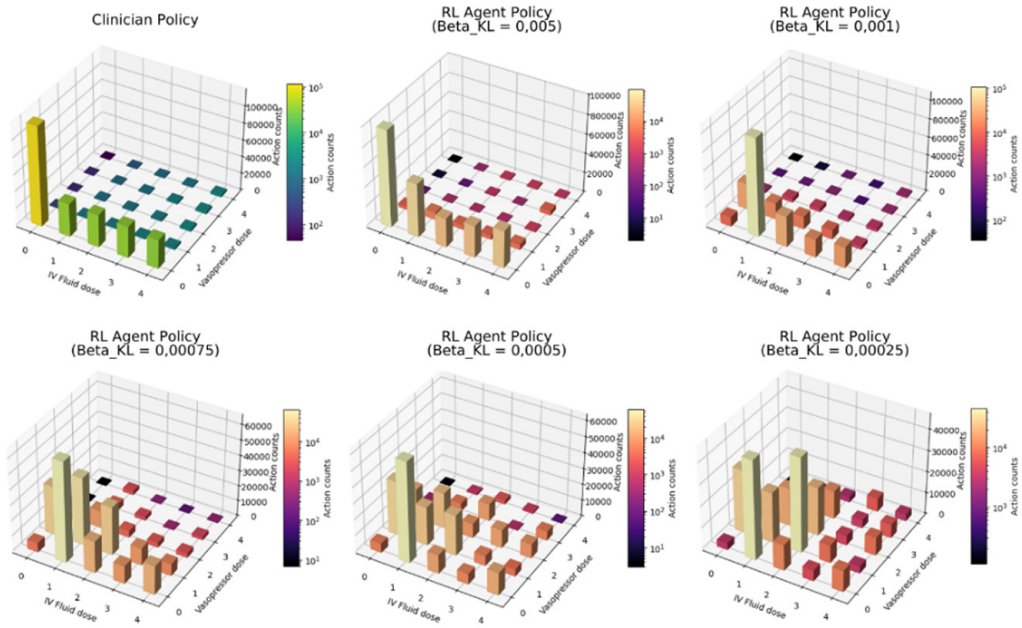

Fig. 8. Action distribution comparison between clinician policy and RL agent variants across discrete treatment bins

Clinicians tended to favor lower-intensity intervention bins, consistent with conservative management in the presence of hemodynamic instability and fluid balance concerns. In contrast, RL agents—particularly those trained with lower KL penalties—allocated actions across a broader range of dosing bins. The agent trained with β_KL = 0.0005 demonstrated a relatively balanced action distribution, with decisions concentrated in moderate dosing ranges. These distributional differences reflect distinct treatment allocation patterns between clinician and RL-derived policies.

3.8 Off-policy evaluation on the test set

As shown in Table 4, the selected RL agent obtained higher estimated returns than the historical clinician policy across all off-policy evaluation (OPE) metrics. The agent's WIS score (9.1763) exceeded that of clinicians (8.5987). Similar differences were observed in FQE (18.4194 vs. 17.7555) and WDR (17.8040 vs. 17.5044). These results indicate that, under the adopted OPE estimators, the learned policy is associated with higher estimated returns relative to the behavior policy. Consistency across multiple OPE methods supports the robustness of this comparative finding within the evaluation framework.

Table 4. Comparison of estimated expected returns between clinician behavior policy and selected RL agent under three OPE estimators

Policy (π)	WIS	FQE	WDR
Clinician	8.5987	17.7555	17.5044
RL Agent	9.1763	18.4194	17.8040

3.9 Feature importance

Permutation feature importance analysis (see Figure 9) identified systolic blood pressure as the most influential variable in the agent's decision process.

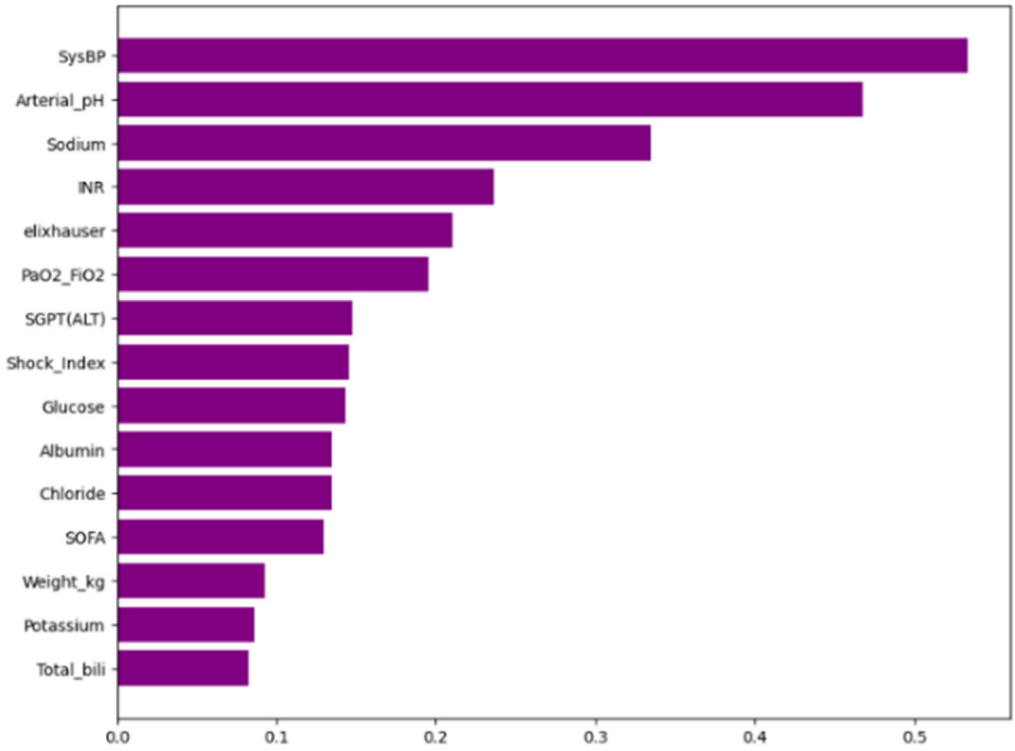

Fig. 9. Permutation feature importance (PFI) scores of the 15 most influential state variables in the trained RL policy

Arterial pH, sodium, INR, and elixhauser score also demonstrated relatively high importance scores. These features correspond to hemodynamic, metabolic,

and coagulation-related variables commonly monitored in sepsis management. The observed importance ranking indicates that the learned policy is sensitive to physiologic markers frequently associated with organ dysfunction. Overall, the feature importance distribution reflects the multi-dimensional structure of the state representation used during training.

4 CONCLUSION

This study demonstrates that an offline reinforcement learning framework based on a dueling DDQN can learn sepsis treatment policies that differ from historical clinician behavior and achieve higher estimated returns under multiple off-policy evaluation metrics. By leveraging physiologically structured state representations and reward signals tied to organ dysfunction trajectories, the model captures treatment allocation patterns across discrete dosing ranges. Feature importance analysis further indicates reliance on clinically relevant physiologic variables. While limited by retrospective data and action discretization, the findings support the potential of offline reinforcement learning for modeling treatment strategies in critical care. Future work should focus on multi-center validation, continuous dosing strategies, and prospective or simulation-based evaluation to assess safety and clinical applicability.

5 REFERENCES

[1] M. Singer *et al.*, "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *JAMA,* vol. 315, no. 8, pp. 801–810, 2016. https://doi.org/10.1001/jama.2016.0287

[2] L. La Via *et al.*, "The global burden of sepsis and septic shock," *Epidemiologia*, vol. 5, no. 3, pp. 456–478, 2024. https://doi.org/10.3390/epidemiologia5030032

[3] World Health Organization, "Global report on the epidemiology and burden of sepsis: Current evidence, identifying gaps and future directions," 2020. https://www.who.int/publications/i/item/9789240010789

[4] L. Evans *et al.*, "Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021," *Intensive Care Medicine*, vol. 47, no. 11, pp. 1181–1247, 2021. https://doi.org/10.1007/s00134-021-06506-y

[5] P. R. Mouncey *et al.*, "Trial of early, goal-directed resuscitation for septic shock," *New England Journal of Medicine*, vol. 372, no. 14, pp. 1301–1311, 2015. https://doi.org/10.1056/NEJMoa1500896

[6] P. E. Marik, L. Byrne, and F. van Haren, "Fluid resuscitation in sepsis: The great 30 mL per kg hoax," *Journal of Thoracic Disease*, vol. 12, no. Suppl 1, pp. S37–S47, 2020. https://doi.org/10.21037/jtd.2019.12.84

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction,* 2nd ed. Cambridge, MA: The MIT Press, 2018.

[8] L. Roggeveen *et al.*, "Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis," *Artificial Intelligence in Medicine*, vol. 111, p. 102003, 2021. https://doi.org/10.1016/j.artmed.2020.102003

[9] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The Artificial Intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, no. 11, pp. 1716–1720, 2018. https://doi.org/10.1038/s41591-018-0213-5

[10] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, "Continuous state-space models for optimal sepsis treatment: A deep reinforcement learning approach," in *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 2017, pp. 147–163.

[11] T. Zhang *et al.*, "Optimizing sepsis treatment strategies via a reinforcement learning model," *Biomedical Engineering Letters*, vol. 14, no. 2, pp. 279–289, 2024. https://doi.org/10.1007/s13534-023-00343-2

[12] S. Lange, T. Gabel, and M. Riedmiller, "Batch reinforcement learning," in *Reinforcement Learning*, M. Wiering and M. van Otterlo, Eds., Springer, 2012, pp. 45–73. https://doi.org/10.1007/978-3-642-27645-3_2

[13] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016. https://doi.org/10.1609/aaai.v30i1.10295

[14] Z. Wang *et al.*, "Dueling network architectures for deep reinforcement learning," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 1995–2003.

[15] S. Tang and J. Wiens, "Model selection for offline reinforcement learning: Practical considerations for healthcare settings," in *Proceedings of Machine Learning Research*, vol. 149, 2021, pp. 2–35.

[16] F. K. Ewald *et al.*, "A guide to feature importance methods for scientific inference," in *Explainable Artificial Intelligence*, L. Longo, S. Lapuschkin, and C. Seifert, Eds., Springer, 2024, pp. 440–464. https://doi.org/10.1007/978-3-031-63797-1_22

[17] S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng, "Reinforcement learning for clinical decision support in critical care: Comprehensive review," *Journal of Medical Internet Research*, vol. 22, no. 7, p. e18477, 2020. https://doi.org/10.2196/18477

[18] C. Gao *et al.*, "Behavior-regularized diffusion policy optimization for offline reinforcement learning," in *Proceedings of the 42nd International Conference on Machine Learning*, 2025, pp. 18630–18657.

[19] R. Tu *et al.*, "Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards," *Human-Centric Intelligent Systems*, vol. 5, pp. 63–76, 2025. https://doi.org/10.1007/s44230-025-00093-7

[20] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[21] C. Molnar *et al.*, "Relating the partial dependence plot and permutation feature importance to the data generating process," in *xAI 2023*, L. Longo Ed., Springer, 2023, pp. 456–479. https://doi.org/10.1007/978-3-031-44064-9_24

[22] J. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016. https://doi.org/10.1038/sdata.2016.35

[23] B. M. Emr, A. M. Alcamo, J. A. Carcillo, R. K. Aneja, and K. P. Mollen, "Pediatric sepsis update: How are children different?" *Surgical Infections*, vol. 19, no. 2, pp. 176–183, 2018. https://doi.org/10.1089/sur.2017.316

[24] A. R. Mahmood, H. van Hasselt, and R. S. Sutton, "Weighted importance sampling for off-policy learning with linear function approximation," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2014, pp. 3014–3022.

[25] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 3703–3712.

[26] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 2139–2148.

[27] C. B. Chandrakala, S. Pooja, C. Pujari, S. Ketavarapu, S. Awatramani, and S. Gohil, "Health-lens: A health diagnosis companion," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 12, pp. 68–102, 2025. https://doi.org/10.3991/ijim.v19i12.51525

[28] F. Alebeisat, A. M. A. Awwad, A. Qatawneh, and S. Al-Suhemat, "A real-time heart attack detection and warning system for drivers using neural network," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 20, pp. 183–204, 2025. https://doi.org/10.3991/ijim.v19i20.55789

[29] Y. Kotani, A. Di Gioia, G. Landoni, A. Belletti, and A. K. Khanna, "An updated 'norepinephrine equivalent' score in intensive care as a marker of shock severity," *Critical Care*, vol. 27, no. 1, p. 29, 2023. https://doi.org/10.1186/s13054-023-04322-y

[30] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, no. Part B, p. 105524, 2020. https://doi.org/10.1016/j.asoc.2019.105524

[31] T. Kobayashi and W. E. L. Ilboudo, "t-soft update of target network for deep reinforcement learning," *Neural Networks*, vol. 136, pp. 63–71, 2021. https://doi.org/10.1016/j.neunet.2020.12.023

6 AUTHORS

Endah Purwanti received B.Sc. degree and the M.T. degree and is currently pursuing the Ph.D. degree in Electrical Engineering. She is a Lecturer in Biomedical Engineering at Universitas Airlangga, Surabaya, Indonesia. Her research focuses on artificial intelligence for healthcare, including medical image analysis, deep learning for segmentation and classification, and reinforcement learning for clinical decision support (E-mail: endah-p-1@fst.unair.ac.id).

Fatima Hasya Puspa Kasih is a student researcher in Biomedical Engineering at Universitas Airlangga with research interests in artificial intelligence and healthcare analytics. Her work focuses on machine learning and reinforcement learning approaches for medical decision-support applications (E-mail: fatimahasyap20@alumni.unair.ac.id).

Franky Chandra Satria Arisgraha is a Lecturer in Biomedical Engineering at Universitas Airlangga, Surabaya, Indonesia. He received the bachelor's and master's degrees in engineering and is currently pursuing the doctoral degree in science. His research interests include biomedical instrumentation, sensor-based medical technologies, embedded systems, and applied healthcare engineering solutions (E-mail: franky-c-s-a@fst.unair.ac.id).