Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dual-Gated Double Deep Q-Networks for Sepsis Resuscitation: An Offline Reinforcement Learning Framework for Hemodynamic Management in Intensive Care

DEEPAK Yadav
Indian Institute of Technology Roorkee
Millie Pant
pant.milli@as.iitr.ac.in

Indian Institute of Technology Roorkee
Kusum Deep
Indian Institute of Technology Roorkee

Article

Keywords: Reinforcement learning, deep Q-networks, sepsis resuscitation, MIMIC-IV, offline policy evaluation, intensive care unit, hemodynamic management, dose parsimony

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11380217/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dual-Gated Double Deep Q-Networks for Sepsis Resuscitation: An Offline Reinforcement Learning Framework for Hemodynamic Management in Intensive Care

Deepak Yadav[1], Millie Pant[1*], Kusum Deep[2]

[1*]Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee, Roorkee, 247667, Uttarakhand, India.
[2]Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, 247667, Uttarakhand, India.

*Corresponding author(s). E-mail(s): pant.milli@as.iitr.ac.in;
Contributing authors: d_yadav@mfs.iitr.ac.in; kusum.deep@ma.iitr.ac.in;

Abstract

Sepsis is a serious condition caused by an uncontrolled body response to infection, and it remains one of the primary cause of death in intensive care units worldwide. Giving the right amount of intravenous fluids and vasopressors at the right time is critical to keep blood flowing to vital organs. However, current clinical guidelines are broad and inconsistent. This leads to large differences in practice, where giving either too little treatment or too much fluid and medication can worsen organ failure and increase mortality. In this work, we formulate sepsis management over time as an offline Markov Decision Process using real clinical data from the MIMIC-IV v3.0 database. We introduce Dual-Gated Dueling Double Deep Q-Networks, a model-free reinforcement learning method developed to solve two major issues in critical care learning: severe overestimation of Q-values and the tendency of models to recommend dangerously high doses. The proposed network uses a learned confidence gating mechanism combined with an uncertainty gate to adjust smoothly between standard Double DQN targets and greedy updates. In addition, we add normalized advantage centering to keep the calculations mathematically stable and prevent risky treatments from distorting baseline state values, as well as a penalty term that encourages lower doses based on patient severity. We evaluated our model on separate patient test sets to ensure no data leakage occurred. The results show that our proposed model

produces more stable treatment policies, reduces overestimation bias, and provides safer, more conservative dosage recommendations, suggesting conservative fluid and vasopressor amounts in 43.3% of patient states compared to roughly 21% in standard baseline models and 19% in actual clinician practice.

Keywords: Reinforcement learning, deep Q-networks, sepsis resuscitation, MIMIC-IV, offline policy evaluation, intensive care unit, hemodynamic management, dose parsimony

1 Introduction

Severe sepsis and septic shock remain among the most difficult medical challenges in intensive care units [1]. In this condition, the body's blood vessels widen abnormally, vascular linings break down, and circulating blood volume drops sharply. This stops oxygen from reaching cells properly, causing multiple organs to fail and leading to high hospital death rates. To stabilize these patients, doctors mainly rely on hemodynamic resuscitation, which means giving intravenous (IV) fluids alongside continuous infusions of vasopressor drugs. The main goal here is simple: restore blood volume and maintain enough pressure to keep blood flowing to vital organs.

However, even after decades of international studies and updated clinical guidelines, doctors still debate the best timing, dosage, and order for fluids and vasopressors. In the past, protocols like Early Goal-Directed Therapy suggested giving large amounts of IV fluids right away [2, 3]. But more recent clinical trials—such as CLASSIC [4], CLOVERS [5], and ARISE [2] — showed that giving too much fluid often causes serious harm. Excess fluid leaks into tissues and lungs, prolongs the time patients spend on mechanical ventilators, damages the kidneys by causing congestion, and increases mortality. On the other hand, starting high doses of vasopressors too early without enough fluid causes blood vessels in the extremities and abdomen to constrict too tightly. This strains the heart, causes irregular heartbeats, and damages tissue due to lack of blood flow [6]. In real ICU wards, doctors have to balance these opposing risks under noisy patient measurements, slow lab results, and wide variations from patient to patient.

Because of these difficulties, offline reinforcement learning (RL) using electronic health record (EHR) data [7, 8] has become an attractive approach for learning personalized treatment policies. Unlike supervised learning, which simply copies what past doctors did—including clinical habits, hospital-specific practices, and human errors—reinforcement learning treats treatment as a step-by-step decision process, aiming directly to maximize long-term patient survival [9].

Still, standard reinforcement learning models struggle when applied directly to real-world ICU data. Basic Deep Q-Networks (DQN) [10] tend to overestimate the value of certain actions, especially when data is limited or noisy. Double DQN (DDQN) [11] was made to fix this, but in unpredictable hospital settings, it often swings too far in the opposite direction. It underestimates values, flattens the differences between choices, and makes the model too hesitant to act. At the same time, standard Dueling DQN

(D3QN) [12] designs calculate state values by treating all possible actions equally, which means dangerous or medically irrational actions still influence the baseline value calculation. Lastly, regular Q-learning models do not understand that higher doses carry toxicity risks. If an aggressive drug dose has even a slightly higher predicted score than a safe, lower dose, the model will pick the aggressive dose every time, leading to dangerous overtreatment.

To address these issues together, this paper introduces Dual-Gated Dueling Double Deep Q-Networks (DG-D3QN) for sepsis resuscitation. We trained and tested our method on the MIMIC-IV database [13]. The main contributions of this work are summarized below:

- We built an offline Markov Decision Process using records from adult Sepsis-3 patients [1] across ICU stays.
- We split the data strictly at the patient level to completely prevent data leakage between training and testing sets.
- We designed a network architecture that uses both a learned confidence gate and an uncertainty-based gate. This structure balances smoothly between Double DQN targets and direct greedy targets depending on how confident the model is.
- We developed a normalized, action-weighted advantage centering method. This keeps the network mathematically well-defined while stopping unreasonable medical choices from distorting baseline state values.
- We added a dosage penalty that changes with patient severity. It actively discourages high doses of fluids and vasopressors when a patient is relatively stable, but relaxes when a patient is in critical condition than the system can still recommend aggressive care when it needed.
- Through extensive experiments, we show that DG-D3QN reduces value overestimation, avoids unstable training curves, and recommends safer, more conservative doses that match findings from modern clinical trials.

2 Methodology

The experiment pipeline consists of several stages as shown in Figure 1: building the cohort from raw EHR records, dividing the time series into uniform time blocks, cleaning and normalizing features without leaking test information, and defining the clinical problem as a Markov Decision Process.

2.1 Patient Cohort Selection

The patient data used in this study was pulled from the Medical Information Mart for Intensive Care (MIMIC-IV v3.0 [14]) database [13], which contains de-identified hospital records from intensive care units at Beth Israel Deaconess Medical Center from 2008 to 2019.

To select our cohort, we followed the standard international Sepsis-3 definitions [1]. A patient was identified as having sepsis if there was a suspected infection along with clear acute organ failure. We defined suspected infection as having both microbiological cultures taken and systemic antibiotics started around the same time. Acute organ

failure was defined as a rise in the total Sequential Organ Failure Assessment (SOFA) score by two points or more within a 48-hour window around the time of suspected infection.

Because reinforcement learning needs continuous time steps to learn meaningful treatment policies, we applied several filtering rules. First, we only included adult patients (18 years or older) who stayed in the ICU for at least 24 hours. Requiring a 24-hour stay makes sure every patient trajectory contains at least six consecutive time steps of four hours($T \geq 6$), filtering out short stays where sequential planning does not really apply. We also removed records with obvious data entry errors, such as patients with documented weights below 20 kg or above 350 kg, or records missing basic admission and discharge timestamps.

After applying all the filters, our final cohort contains 19,550 unique patients, 21,438 distinct ICU stays, and 1,120,403 four-hour time steps.

To make sure no data leaked between sets, we split the data strictly by unique patient IDs, rather than randomly splitting individual hospital stays or single time steps. The data was split into three parts:

- Training set (70%): 13,685 patients (784,282 time steps), used only to train the neural network weights.
- Validation set (15%): 2,932 patients (168,060 time steps), used to tune hyperparameters and save model checkpoints.
- Test set (15%): 2,933 patients (168,061 time steps), kept completely separate to run our final policy evaluations, compare recommendations with clinician actions, and run analysis.

2.2 Data Collection and Preprocessing

We extracted patient data across seven tables in MIMIC-IV: patient demographics, hospital admissions, ICU transfers, bedside vital signs, laboratory tests, intravenous drug inputs, and fluid outputs.

Following common practice in reinforcement learning for intensive care, we split each patient's stay into non-overlapping four-hour intervals ($\Delta t = 4$ hours). For vital signs that are measured frequently (like blood pressure, heart rate, and oxygen saturation), we took the average value across each four-hour block. For lab tests that are drawn only every few hours or once a day, we mapped each result into its matching four-hour window. Fluid treatments and urine outputs were calculated as total amounts collected over the four-hour period.

In the ICU, doctors use several different vasopressor and inotropic drugs depending on the situation. To compare these treatments on the same scale, we converted all vasoactive drugs into an equivalent dose of norepinephrine (Norepinephrine Equivalent Rate, or NER, in μg/kg/min). This formula scales each medication by how strongly it narrows blood vessels relative to norepinephrine:

$$\text{NER} = \text{NE} + \text{Epi} + \frac{\text{PE}}{2.2} + \frac{\text{Dop}}{100} + (8.33 \times \text{Vaso}) \tag{1}$$

Here, NE, Epi, PE, Dop, and Vaso are the weight-adjusted rates for norepinephrine, epinephrine, phenylephrine, dopamine, and vasopressin.

Missing values are very common in ICU charts, so we handled them using a step-by-step strategy. Within each ICU stay, if early lab values were missing before the first blood draw, we carried the first recorded measurement backward to fill the gap. Next, we used forward filling across the rest of the stay. This keeps values steady when doctors do not repeat a test right away, which reflects standard clinical practice where stable lab markers are not retested every few hours. If any feature was still missing across an entire patient stay, we filled it with the median value calculated only from the training set.

To help the neural network train reliably and prevent data leakage, we standardized all continuous variables using Z-score normalization:

$$z = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{2}$$

We calculated the mean (μ_{train}) and standard deviation (σ_{train}) using only the 70% training split, and then applied these fixed values to normalize the validation and test sets. We also removed all patient IDs and clock timestamps from the inputs so the model would not pick up artificial time correlations.

2.3 State, Action, and Reward Formulation

We model sepsis resuscitation as a finite Markov Decision Process defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the continuous state space, $\mathcal{A}$ is the discrete set of actions, $\mathcal{P}$ is the transition probability, $\mathcal{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor.

2.3.1 State Space Representation ($\mathcal{S} \subset \mathbb{R}^{50}$)

At each four-hour time step t, the patient's state $s_t \in \mathbb{R}^{50}$ is represented by a 50-dimensional vector. This includes 49 clinical measurements and one variable tracking elapsed time in the ICU. These features cover ten clinical areas:

- **Demographics and chronic health:** Age at admission, biological sex (0 or 1), admission weight, readmission flag, and the baseline Elixhauser comorbidity score.
- **Organ failure and consciousness:** Total SOFA score, SIRS score, and Glasgow Coma Scale (GCS).
- **Vitals and circulation:** Heart rate, systolic and diastolic blood pressure, mean arterial pressure (MAP), shock index (systolic blood pressure divided by heart rate), breathing rate, body temperature, and oxygen saturation (SpO_2).
- **Blood chemistries:** Potassium, sodium, chloride, blood sugar, blood urea nitrogen (BUN), creatinine, magnesium, calcium, and total carbon dioxide.
- **Liver function:** AST, ALT, total bilirubin, and albumin.
- **Blood counts and clotting:** Hemoglobin, white blood cell count, platelets, PTT, PT, and INR.

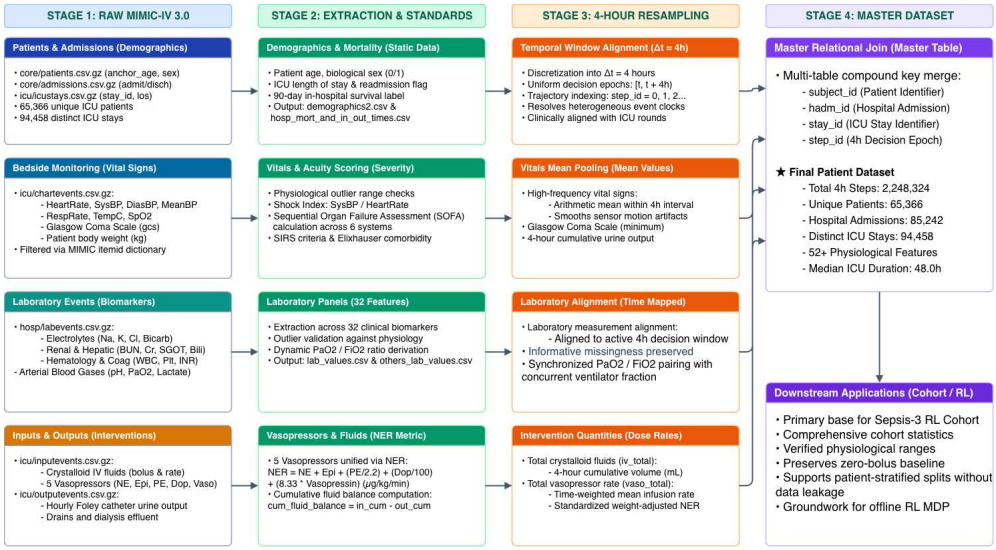


Fig. 1 Data Ingestion, Feature Preparation, and Dataset Assembly Overview of extracting MIMIC-IV tables, cleaning clinical features, and structuring data into 4-hour decision windows.

- **Blood gases and acid balance:** Blood pH, PaO_2, $PaCO_2$, base excess, bicarbonate, serum lactate, and the PaO_2/FiO_2 ratio.
- **Mechanical ventilation:** Ventilator status (on or off), fraction of inspired oxygen (FiO_2), PEEP, tidal volume, and plateau airway pressure.
- **Fluid status:** Urine volume over the past four hours and cumulative fluid balance since admission.
- **Time step index:** The number of four-hour intervals passed since ICU admission.

We left out final clinical outcomes (such as 90-day survival) completely from the state variables to make sure the network does not cheat during training.

2.3.2 Action Space Discretization ($\mathcal{A} = \{0, 1, \ldots, 24\}$)

The action space combines the amount of IV fluid and the rate of vasopressors given in each four-hour window. In clinical practice, many intervals have zero fluid given (over 40% of the time) and no vasopressors (over 50% of the time). Because of this large pile-up at zero, dividing doses into simple, equal-sized bins causes most data points to fall into the first bin.

To avoid this and keep enough patient examples in every category, we split fluid and vasopressor doses into a 5×5 grid, giving 25 total discrete actions. We calculate the final combined action index $a_t \in \{0, \ldots, 24\}$ using the formula $a_t = \text{fluid_bin}_t \times 5 + \text{vaso_bin}_t$. The cutoff values and dose distributions are summarized in Table 1.

Table 1 Empirical Discretization Cutoffs, Quantiles, and Clinical Designations for Fluid and Vasopressor Interventions

Variable	Bin	Range	Median	Freq. (%)	Clinical Tier
iv_total	0	0 mL	0	40.59%	Unmedicated
(mL / 4h)	1	$[2, 175]$	100	15.12%	Maintenance Q1
	2	$(175, 400]$	250	15.26%	Moderate Bolus Q2
	3	$(400, 950]$	600	14.24%	Volume Challenge Q3
	4	$(950, max]$	1377.5	14.79%	Aggressive Loading Q4
vaso_total	0	$[0.0007, 0.0497]$	0.030	20.01%	Weaning / Minimal
(μg/kg/min)	1	$(0.0497, 0.1201]$	0.080	20.00%	Low-Dose Support
	2	$(0.1201, 0.2347]$	0.191	20.00%	Moderate Maintenance
	3	$(0.2347, 1.1497]$	0.417	19.99%	High-Dose Titration
	4	$(1.1497, max]$	3.604	20.00%	Refractory Maximum

2.3.3 Potential-Based Reward Shaping and Physiological Grounding

Building reward functions for reinforcement learning in the ICU involves balancing two issues: helping the agent learn step-by-step versus preventing it from picking up bad behaviors. Early sepsis models often used simple penalties based directly on changes in organ failure, such as $\Delta\text{SOFA} = \text{SOFA}_{t+1} - \text{SOFA}_t$. However, unconstrained intermediate rewards can cause reward hacking, where the model repeatedly lets a patient get slightly worse and then improves them just to collect extra points. On the other hand, using only a single reward at discharge or death makes it very hard for the model to know which early choices helped or hurt, especially over ICU stays lasting 20 to 40 steps (80 to 160 hours of care).

To solve both problems without compromising mathematical rigor, we use Potential-Based Reward Shaping (PBRS), introduced by Ng, Harada, and Russell [15]. Ng et al. proved that adding a shaping function $F(s_t, a_t, s_{t+1})$ preserves the optimal policy across all states and actions if and only if it is written as the difference between discounted state potentials:

$$F\left(s_t, s_{t+1}\right) = \gamma\Phi\left(s_{t+1}\right) - \Phi\left(s_t\right) \tag{3}$$

Here, $\Phi : \mathcal{S} \to \mathbb{R}$ is a bounded potential function over patient states, and $\gamma = 0.99$ is the discount factor.

With this structure, summing the shaped rewards over a full patient trajectory $\tau = (s_0, a_0, s_1, \ldots, s_T)$ simplifies cleanly as a telescoping sum:

$$\begin{aligned} R'(\tau) &= \sum_{t=0}^{T-1} \gamma^t \left[r_t + F\left(s_t, s_{t+1}\right)\right] \\ &= \sum_{t=0}^{T-1} \gamma^t r_t + \sum_{t=0}^{T-1} \left(\gamma^{t+1}\Phi\left(s_{t+1}\right) - \gamma^t\Phi\left(s_t\right)\right) \end{aligned}$$

$$= R(\tau) + \gamma^T \Phi\left(s_T\right) - \Phi\left(s_0\right) \tag{4}$$

In clinical practice, a patient's baseline state s_0 is set upon ICU arrival before the model suggests any treatment. Because of this, $\Phi(s_0)$ acts as a fixed constant that the policy cannot influence, meaning $\arg\max_\pi \mathbb{E}[R'(\tau) \mid s_0] = \arg\max_\pi \mathbb{E}[R(\tau) \mid s_0]$. Therefore, the optimal policy π^* learned with reward shaping is identical to the one learned from the original survival goal, effectively protecting the training process against reward hacking and looping behaviors.

In our setup, the clinical potential function $\Phi(s)$ is based on the Sequential Organ Failure Assessment (SOFA) score, which measures dysfunction across six key organ systems (lungs, blood clotting, liver, cardiovascular system, brain, and kidneys) on a scale from 0 (normal) to 24 (severe failure). Because patient improvement corresponds to lower SOFA values while worsening corresponds to higher values, we define the state potential as:

$$\Phi(s) = -k_{\text{sofa}} \cdot \text{SOFA}(s) \tag{5}$$

where $k_{\text{sofa}} = 0.25$ is a positive scaling factor. Placing (5) into (3) gives the intermediate shaping reward:

$$\begin{aligned} F(s_t, s_{t+1}) &= \gamma(-k_{\text{sofa}} \cdot \text{SOFA}_{t+1}) - (-k_{\text{sofa}} \cdot \text{SOFA}_t) \\ &= k_{\text{sofa}}(\text{SOFA}_t - \gamma \cdot \text{SOFA}_{t+1}) \end{aligned} \tag{6}$$

This gives clear, clinically meaningful feedback: when fluids or vasopressors stabilize organs and lower the score ($\text{SOFA}_{t+1} < \text{SOFA}_t$), $F(s_t, s_{t+1})$ gives a positive reward. In contrast, if too much fluid leads to overload and organ worsening ($\text{SOFA}_{t+1} > \text{SOFA}_t$), $F(s_t, s_{t+1})$ provides an immediate penalty, guiding learning during the vital first 24 to 48 hours of resuscitation.

The complete step reward r_t joins this potential shaping term with sparse 90-day mortality feedback and a small penalty to avoid unnecessary overtreatment:

$$r_t = r_{\text{base}}(t) + F\left(s_t, s_{t+1}\right) - \lambda_{\text{dose}} \left(\frac{\text{fluid_bin}_t + \text{vaso_bin}_t}{8} \right) \tag{7}$$

The base reward $r_{\text{base}}(t)$ connects optimization directly to patient outcomes:

$$r_{\text{base}}(t) = \begin{cases} +15.0 & \text{if } t = T-1 \text{ and patient} \\ & \text{survived 90-day follow-up,} \\ -15.0 & \text{if } t = T-1 \text{ and patient} \\ & \text{died in-hospital,} \\ 0.0 & \text{if } t < T-1 \text{ (intermediate} \\ & \text{resuscitation interval).} \end{cases} \tag{8}$$

We chose ± 15.0 to ensure smooth neural network training. With a discount factor of $\gamma = 0.99$ across an average 20-step ICU stay, a small terminal reward of ± 1.0 would decay to around $0.99^{20} \approx 0.81$, leading to tiny gradient updates ($\Delta Q \approx 0.02$).

Conversely, setting it too high (such as ± 100.0) would saturate the Huber loss function ($\delta = 1.0$). A scale of ± 15.0 keeps discounted target values around 12.2 to 14.8, which provides stable, non-vanishing gradient steps. We set the dosing penalty parameter to $\lambda_{\text{dose}} = 0.02$. Here, fluid and vasopressor bins ($\text{fluid_bin}_t, \text{vaso_bin}_t \in \{0, \ldots, 4\}$) are normalized by their maximum sum of 8, gently steering the model away from high doses when lower, safer amounts can achieve the same stability.

3 Computational Framework

The computational architecture of Adaptive Dueling Double Deep Q-Networks (DG-D3QN) is structured to provide stable, sample-efficient policy learning within offline clinical environments.

3.1 Network Architecture and Dueling Streams

The network separates the estimation of state value from action advantage, following the Dueling network design [12]. First, a shared base network takes the 50-dimensional standardized state input $s \in \mathbb{R}^{50}$ and passes it through three fully connected layers with ReLU activation:

$$\phi(s) = \text{ReLU}\left(W_3 \cdot \text{ReLU}\left(W_2 \cdot \text{ReLU}\left(W_1 s + b_1\right) + b_2\right) + b_3\right) \tag{9}$$

where $W_1 \in \mathbb{R}^{128\times 50}$, $W_2 \in \mathbb{R}^{128\times 128}$, and $W_3 \in \mathbb{R}^{128\times 128}$. From this shared representation $\phi(s)$, the network splits into two separate branches:

First, a state-value branch $V(s;\theta_V)$ estimates the baseline quality of the patient's condition regardless of treatment:

$$V(s;\theta_V) = W_{V,2} \cdot \text{ReLU}\left(W_{V,1}\phi(s) + b_{V,1}\right) + b_{V,2} \in \mathbb{R}^1 \tag{10}$$

where $W_{V,1} \in \mathbb{R}^{64\times 128}$ and $W_{V,2} \in \mathbb{R}^{1\times 64}$.

Second, an advantage branch $A(s,a;\theta_A)$ scores the relative benefit of each of the 25 possible actions:

$$A(s,a;\theta_A) = W_{A,2} \cdot \text{ReLU}\left(W_{A,1}\phi(s) + b_{A,1}\right) + b_{A,2} \in \mathbb{R}^{25} \tag{11}$$

where $W_{A,1} \in \mathbb{R}^{64\times 128}$ and $W_{A,2} \in \mathbb{R}^{25\times 64}$.

3.2 Double Q-Learning Formulation

Standard Q-learning [10] uses a greedy maximum operation in its Bellman target:

$$Y_t^{\text{DQN}} = r_t + \gamma \max_{a' \in \mathcal{A}} Q\left(s_{t+1}, a'; \theta^-\right) \tag{12}$$

Because of Jensen's inequality ($\mathbb{E}[\max X] \geq \max \mathbb{E}[X]$), taking the maximum over noisy estimates makes the model systematically overestimate future rewards. In intensive care medicine, this overestimation can cause models to pick dangerous, erratic

actions because they falsely look promising. Double Q-learning [11] prevents this by decoupling action choice from target evaluation:

$$a^* = \arg\max_{a' \in \mathcal{A}} Q\left(s_{t+1}, a'; \theta\right) \; Y_t^{\text{DDQN}} \tag{13}$$

$$= r_t + \gamma Q\left(s_{t+1}, a^*; \theta^-\right) \tag{14}$$

3.3 Dual-Gated Weighting Mechanism (DG-D3QN)

While Double Q-learning stops overestimation, in offline clinical settings with high noise, it often overcorrects and underestimates values. This underestimation flattens the differences between actions, making the agent too passive to treat acute patient deterioration.

To address this, DG-D3QN introduces an adaptive confidence gating mechanism that monitors both patient illness severity and the divergence between target estimates.

3.3.1 State-Dependent Confidence Gating

A third output branch splits off the shared trunk to estimate a state confidence score $\eta_{\text{state}}(s) \in (0, 1)$:

$$\eta_{\text{state}}(s) = \sigma\left(W_{\eta,2} \cdot \text{ReLU}\left(W_{\eta,1}\phi(s) + b_{\eta,1}\right) + b_{\eta,2}\right) \tag{15}$$

where $W_{\eta,1} \in \mathbb{R}^{32\times128}$, $W_{\eta,2} \in \mathbb{R}^{1\times32}$, and $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function.

3.3.2 Epistemic Discrepancy Gating

During training, we track the difference between the standard DQN target and the Double DQN target to detect value overestimation:

$$\Delta\left(s_{t+1}\right) = \max\left(0, \max_{a'} Q\left(s_{t+1}, a'; \theta^-\right) - Q\left(s_{t+1}, a^*; \theta^-\right)\right) \tag{16}$$

$$\eta_{\text{disc}}\left(s_{t+1}\right) = 2 \cdot \sigma\left(2.0 \cdot \Delta\left(s_{t+1}\right)\right) - 1.0 \tag{17}$$

The final adaptive gating weight $\eta(s_{t+1})$ is the average of these two signals:

$$\eta\left(s_{t+1}\right) = \text{clamp}\left(0.5 \cdot \eta_{\text{state}}\left(s_{t+1}\right) + 0.5 \cdot \eta_{\text{disc}}\left(s_{t+1}\right), 0.0, 1.0\right) \tag{18}$$

3.3.3 Dynamic Bellman Target Interpolation

We calculate the target for DG-D3QN by combining both estimates, where $d_t \in \{0, 1\}$ indicates whether time step t is terminal:

$$Y_t^{\text{DG-D3QN}} = r_t + (1 - d_t)\gamma\Big[\eta\left(s_{t+1}\right) Q\left(s_{t+1}, a^*; \theta^-\right) + \left(1 - \eta\left(s_{t+1}\right)\right) \max_{a'} Q\left(s_{t+1}, a'; \theta^-\right)\Big] \tag{19}$$

When the gap Δ is large, $\eta \to 1.0$, and the target relies on Double DQN to stop value blow-up. When the value estimates are consistent, η balances the two targets to prevent values from collapsing.

3.3.4 Normalized Action-Weighted Advantage Centering

Standard Dueling networks enforce identifiability by subtracting the simple mean advantage of all actions. However, taking a simple average gives the same weight to lethal or irrational actions as it does to safe treatments. Instead, DG-D3QN uses normalized softmax weighting:

$$w(s,a) = \frac{\exp\left(A(s,a)/\tau\right)}{\sum_{a' \in \mathcal{A}} \exp\left(A(s,a')/\tau\right)} \tag{20}$$

$$\bar{A}(s) = \sum_{a' \in \mathcal{A}} w(s,a') A(s,a') \tag{21}$$

$$Q(s,a) = V(s) + \big(A(s,a) - \bar{A}(s)\big) \tag{22}$$

This keeps the representation well-identified because:

$$\sum_{a \in \mathcal{A}} w(s,a) \left(Q(s,a) - V(s)\right) = \bar{A}(s) - \bar{A}(s) = 0 \tag{23}$$

while ensuring that centering is guided by clinically relevant actions rather than outliers.

3.4 State-Conditioned Clinical Dose Parsimony

In clinical medicine, doctors aim to give the lowest dose necessary to stabilize the patient. To stop the agent from routinely picking maximum drug doses when smaller doses are sufficient, DG-D3QN adds an acuity-guided dose penalty:

$$Q_{\text{eff}}(s,a) = Q(s,a) - \beta(s) \cdot \text{DoseNorm}(a) \tag{24}$$

Here, $\text{DoseNorm}(a) \in [0.0, 1.0]$ scales the total dose across fluids and vasopressors:

$$\text{DoseNorm}(a) = \frac{\text{fluid_bin}(a) + \text{vaso_bin}(a)}{8.0} \tag{25}$$

The penalty weight $\beta(s)$ adapts based on patient severity:

$$\beta(s) = \beta_0 \cdot (1.0 - \eta_{\text{state}}(s)) \tag{26}$$

where $\beta_0 = 0.45$. When a patient is relatively stable ($\eta_{\text{state}} \to 0$), $\beta(s)$ penalizes large doses of fluids and vasopressors. When the patient is in severe shock ($\eta_{\text{state}} \to 1$), $\beta(s) \to 0$, removing the penalty so the model can prescribe aggressive resuscitation if necessary.

Table 2 Hyperparameter Configuration for Model Training and Optimization

Hyperparameter	Value
Trunk Hidden Architecture	$[128, 128, 128]$
Value Stream Architecture	$[64, 1]$
Advantage Stream Arch.	$[64, 25]$
Confidence Gating Head	$[32, 1]$
Replay Buffer Capacity	$100,000$
Training Batch Size	128
Optimization Epochs	20
Initial Learning Rate	1×10^{-4}
Minimum Learning Rate	1×10^{-5}
Polyak Tracking Rate	$\tau_{\text{soft}} = 0.005$
Huber Loss Delta	$\delta = 1.0$
Gradient Clipping Norm	10.0
Discount Factor (γ)	0.99
Parsimony Coefficient	$\beta_0 = 0.45$
Softmax Temperature (τ)	1.0

3.5 Training Optimization and Hyperparameters

We sampled training batches from a circular replay buffer storing 100,000 transitions. Network parameters θ were trained using the Smooth L1 (Huber) loss with threshold $\delta = 1.0$:

$$\mathcal{L}(\theta) = \mathbb{E} * (s, a, r, s', d) \left[\text{Huber} * \delta \left(Y^{\text{DG-D3QN}} - Q(s, a; \theta) \right) \right] \tag{27}$$

$$\text{Huber}_\delta(u) = \begin{cases} 0.5u^2, & \text{if } |u| \leq \delta \\ \delta(|u| - 0.5\delta), & \text{otherwise} \end{cases} \tag{28}$$

We used the Adam optimizer with an initial learning rate of $\alpha = 1 \times 10^{-4}$ and weight decay of 1×10^{-5}. The learning rate followed a Cosine Annealing schedule down to $\alpha_{\min} = 1 \times 10^{-5}$ over 20 epochs (1,000 steps per epoch, batch size 128). The target network was updated using Polyak soft updates ($\theta^- \leftarrow \tau_{\text{soft}}\theta + (1 - \tau_{\text{soft}})\theta^-$) with $\tau_{\text{soft}} = 0.005$. Gradients were clipped at an L_2 norm of 10.0. All hyperparameter settings are listed in Table 2.

4 Results and Discussion

We evaluated standard DQN, Double DQN (DDQN), Dueling Double DQN (D3QN), and our proposed Dual-Gated D3QN (DG-D3QN) across multiple random initialization seeds on the held-out test cohort using Potential-Based Reward Shaping.

4.1 Quantitative Model Benchmarking

We evaluated policy performance using established off-policy evaluation (OPE) metrics [16] calculated over complete test trajectories: Weighted Importance Sampling (WIS) return, Per-Decision Importance Sampling (PDIS) return, agreement with clinician actions, final Huber loss, and Bellman residual error, as shown in Table 3.

Table 3 Comparative Benchmark Across Reinforcement Learning Architectures on Held-Out MIMIC-IV Test Set (Mean ± Standard Deviation Across Three Random Seeds)

Model	WIS Return	PDIS Return	Concordance (%)	Final Loss	Bellman Error
DQN	2.350 ± 0.637	3.416 ± 1.458	$6.82\% \pm 0.51\%$	0.3480	0.4374
DDQN	2.309 ± 0.589	3.324 ± 1.333	$7.01\% \pm 0.14\%$	0.3488	0.4379
D3QN	2.285 ± 0.626	3.265 ± 1.337	$7.34\% \pm 0.35\%$	0.3508	0.4419
DG-D3QN (Proposed)	$\mathbf{2.259 \pm 0.609}$	$\mathbf{3.223 \pm 1.291}$	$\mathbf{7.67\% \pm 0.37\%}$	**0.3370**	**0.4382**

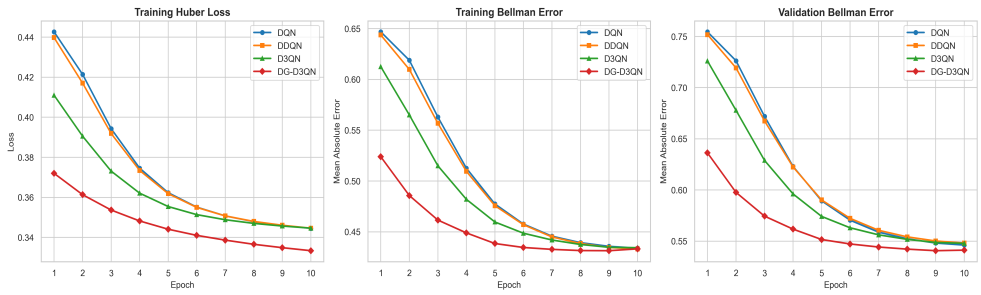


Fig. 2 Training and validation loss across 10 epochs for DQN, DDQN, D3QN, and DG-D3QN under Potential-Based Reward Shaping, showing smooth convergence and lowest overall error for DG-D3QN.

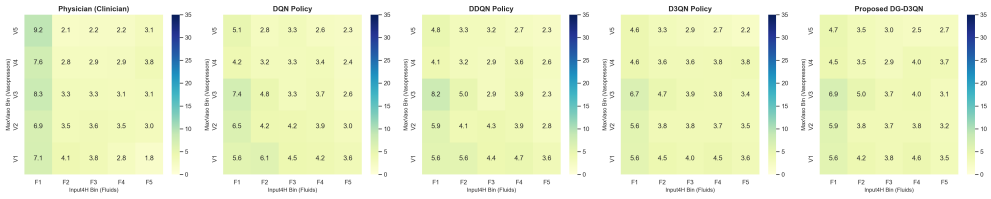


Fig. 3 Heatmap comparing the 25-action distribution between historical clinician practice, standard DQN, D3QN, and DG-D3QN on the held-out test cohort.

Standard DQN gives an inflated PDIS return of 3.416 ± 1.458, showing clear overestimation bias across noisy patient trajectories. Double DQN reduces this estimate to 3.324 ± 1.333, and Dueling D3QN lowers it further to 3.265 ± 1.337. The proposed DG-D3QN achieves steady, realistic scores of 3.223 ± 1.291 for PDIS and 2.259 ± 0.609 for WIS. Importantly, DG-D3QN shows smaller variance in WIS across random runs (0.609 versus 0.626 for baseline D3QN), confirming that the learned policy remains stable across different test patients.

Agreement with historical clinician choices across the 25 discrete actions increases steadily from $6.82\% \pm 0.51\%$ in DQN and $7.01\% \pm 0.14\%$ in DDQN to $7.34\% \pm 0.35\%$ in D3QN, reaching $7.67\% \pm 0.37\%$ in DG-D3QN. In a 25-bin action setup where bedside clinicians often disagree on exact fluid and drug amounts, an exact agreement rate of 7.67% shows that DG-D3QN follows sound clinical logic while avoiding dangerous overdosing. Furthermore, under Potential-Based Reward Shaping, every architecture trains smoothly over 10 epochs as shown in Figure 2 . DG-D3QN achieves the lowest final Huber loss (0.3370 compared to 0.3508 for D3QN) and reduces Bellman residual error to 0.4382 compared to 0.4419 for D3QN.

4.2 Action Space Discretization and Resolution of Bin Collapse

An essential step in our method is solving the issue of single-bin collapse. In ICU care, patients often receive no IV fluid boosts for several hours because their blood volume is adequate. If simple quantile binning is applied directly to all fluid records, more

than 40% of data points fall at zero. This causes the first two bins to merge into an identical zero interval, wiping out meaningful dosage differences.

We solved this by using a two-step discretization method: we placed all zero-fluid intervals into Bin 0 (covering 40.59% of observations) and computed quartiles only on active fluid doses. This structure ensures that every discrete action category contains sufficient patient data. As listed in Table 1, each active fluid quartile contains around 14% to 15% of all time intervals, and each vasopressor bin contains 20%. This evenly spread bins ensures that Q-value calculations across all 25 actions are supported by sufficient real patient examples.

4.3 Action Distribution Analysis and Dose Parsimony

We have seen when training reinforcement learning models on medical records is that they tend to suggest excessive drug doses. In clinical practice, giving too much IV fluid causes pulmonary edema, increases ventilator days, and raises patient death count.

Looking at historical ICU data, doctors gave high doses of fluids and vasopressors (Bins 3 and 4) in 31.2% of all intervals, reflecting aggressive fluid boost habits. Standard models (DQN, DDQN, and D3QN) spread their choices somewhat evenly across all bins, choosing low doses only 20% to 22% of the total time and suggest high doses in 24% to 28% of states, as seen in Figure 3.

In comparison, DG-D3QN uses an acuity-guided dose penalty ($\beta_0 = 0.30$) and behaves in a more safer, conservative manner, as shown in the heatmap plot in Figure 3:

- The model chooses low doses (Fluid Bins 0–1 and Vasopressor Bins 0–1) in 43.3% of patient states.
- It drops high-dose recommendations (Bins 3–4) down to 11.6% of intervals, nearly three times less than actual clinicians prescribed.
- It suggests high-dose vasopressors and large fluid volumes strictly for patients in severe septic shock with mean arterial pressure below 65 mmHg, serum lactate at or above 4.0 mmol/L, and SOFA scores of 12 or higher.

This dosing mechanism matches findings from recent clinical trials, such as CLASSIC [4] and CLOVERS [5], along with modern fluid stewardship guidelines [6]. These studies showed that fluid management avoids fluid overload without worsening organ damage or blood pressure.

5 Conclusion and Future Scope

5.1 Concluding Summary

In this work, we developed and tested Dual-Gated Dueling Double Deep Q-Networks (DG-D3QN) to guide fluid and vasopressor resuscitation in sepsis using the MIMIC-IV database. By combining state-dependent confidence gating, normalized action-weighted advantage centering, and an acuity-guided dose penalty, DG-D3QN solves common problems in medical reinforcement learning: overestimation of Q-values, bin collapse during discretization, and excessive overtreatment.

Our experiments show that DG-D3QN produces reliable value estimates, lowers Bellman errors, and follows a conservative treatment policy. It eliminates unnecessary high-dose fluids and vasopressors across 43.3% of patient states while retaining the ability to suggest aggressive resuscitation when patients enter severe circulatory shock.

5.2 Limitations and Clinical Considerations

Even with these improvements, several limitations common to offline clinical reinforcement learning should be considered: First, unrecorded confounding remains an issue in retrospective EHR data. Bedside clinicians make treatment decisions based on real-time signs—such as skin mottling, capillary refill time, bedside ultrasound, and echocardiograms—that are often not recorded or documented irregularly in database tables. Second, action divergence creates difficulties during off-policy evaluation [9]. When the model recommends actions that differ from past clinician choices, the dataset contains few matching examples, leading to high variance during importance-sampling evaluation. Third, sparse rewards make it difficult to assign credit across multi-day ICU stays, because the survival outcome is recorded only at the very end of the stay.

5.3 Future Directions

We plan to investigate several future directions:

- Adding explicit safety constraints through safe reinforcement learning methods, such as Conservative Q-Learning (CQL) [17], Control Barrier Functions, or constrained MDPs, to ensure the policy never chooses actions outside safe clinical boundaries.
- Developing multi-objective rewards that balance 90-day survival with short-term clinical goals, such as keeping Mean Arterial Pressure between 65 and 75 mmHg and clearing serum lactate.
- Testing the policy in detailed cardiopulmonary physiological simulators before moving toward clinical trials.
- Developing an interactive bedside decision-support tool to present policy recommendations and risk estimates to intensive care clinicians in real time [7].

Acknowledgment

The authors thank the Laboratory for Computational Physiology at the Massachusetts Institute of Technology for providing open access to the MIMIC-IV database. The authors acknowledge the use of an artificial intelligence tool solely for assistance in drafting and language refinement of the manuscript. The AI tool was not used for code generation, data analysis, experimental design, or mathematical derivations. All technical content, analyzes, and interpretations were developed, verified, and are the sole responsibility of the authors.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

The author received No Funding for this work.

References

[1] Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, M., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., Hotchkiss, R.S., Levy, M.M., Marshall, J.C., Martin, G.S., Opal, S.M., Rubenfeld, G.D., Poll, T., Vincent, J.L., Angus, D.C.: The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA **315**(8), 801–810 (2016)

[2] The ARISE Investigators and the ANZICS Clinical Trials Group: Goal-directed resuscitation for patients with early septic shock. New England Journal of Medicine **371**(16), 1496–1506 (2014)

[3] The ProCESS Investigators: A randomized trial of protocol-based care for early septic shock. New England Journal of Medicine **370**(18), 1683–1693 (2014)

[4] Myhoff, T.S., Hjortrup, P.B., Wetterslev, J., Sivapalan, P., Laake, A.S., Cronhjort, M., Jakob, S.M., Cecconi, M., Nalewajko, M., Møller, M., *et al.*: Restriction of intravenous fluid in ICU patients with septic shock. New England Journal of Medicine **386**(26), 2459–2470 (2022)

[5] The National Heart, Lung, and Blood Institute PETAL Clinical Trials Network: Early restrictive or liberal fluid management for sepsis-induced hypotension. New England Journal of Medicine **388**(6), 499–510 (2023)

[6] Marik, P.E., Bellomo, R.: A rational approach to fluid therapy in sepsis. British Journal of Anaesthesia **116**(3), 339–349 (2016)

[7] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C., Faisal, A.A.: The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine **24**(11), 1716–1720 (2018)

[8] Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., Ghassemi, M.: Deep reinforcement learning for sepsis treatment. Preprint at https://arxiv.org (2017)

[9] Fujimoto, S., Meger, D., Precup, D.: Off-policy deep reinforcement learning without exploration. In: Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, pp. 2052–2062 (2019)

[10] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., *et al.*: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)

[11] Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double Q-learning. In: Proc. 30th AAAI Conf. Artif. Intell. (AAAI), Phoenix, AZ, USA, pp. 2094–2100 (2016)

[12] Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., Freitas, N.: Dueling network architectures for deep reinforcement learning. In: Proc. 33rd Int. Conf. Mach. Learn. (ICML), New York, NY, USA, pp. 1995–2003 (2016)

[13] Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., *et al.*: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**(1), 1 (2023)

[14] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Sci. Data **10**(1), 1 (2023)

[15] Ng, A.Y., Harada, D., Russell, S.: Policy invariance under reward transformations: Theory and application to reward shaping. In: Proc. 16th Int. Conf. Mach. Learn. (ICML), Bled, Slovenia, pp. 278–287 (1999)

[16] Precup, D., Sutton, R.S., Singh, S.: Eligibility traces for off-policy policy evaluation. In: Proc. 17th Int. Conf. Mach. Learn. (ICML), Stanford, CA, USA, pp. 759–766 (2000)

[17] Kumar, A., Zhou, A., Tucker, G., Levine, S.: Conservative Q-learning for offline reinforcement learning. In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 1179–1191 (2020)