Performance and Interpretability Trade-offs in Reinforcement Learning for Sepsis Treatment: Comparing Offline and Online Approaches

Zhiyu Cheng[*], Yalun Ding[1,*], Chuanhui Peng[1]
[1]Department of Statistics, George Washington University
[*]Corresponding author: yding@gwu.edu

October 28, 2025

Abstract

Sepsis remains a leading cause of mortality in critical care, and reinforcement learning (RL) offers a promising route to data-driven treatment policies. Yet clinical adoption is impeded by the prevailing assumption that interpretability inevitably compromises performance, and that online RL methods necessarily outperform offline approaches. We interrogate these trade-offs by comparing three offline RL methods (Behavior Cloning, Conservative Q-Learning, and Deep Q-Network trained on static datasets) with three online RL methods (Double DQN with Attention, Double DQN with Residual connections, and Soft Actor-Critic with environment interaction) using the gym-sepsis simulator—a MIMIC-III–derived environment for sepsis treatment. Policy quality is assessed through patient-survival rates stratified by Sequential Organ Failure Assessment (SOFA) states, while interpretability is quantified with Linearly Estimated Gradients (LEG), a model-agnostic feature-importance method.

Across 500 evaluation episodes, online RL achieves marginally higher overall survival (95.4% for DDQN-Attention vs. 94.2% for BC), with a 1.9 percentage point advantage on high-severity patients (90.5% vs. 88.6%). However, this modest performance gain comes at the cost of requiring extensive environment interaction during training—infeasible in clinical settings where patient safety is paramount. Interpretability analysis reveals that offline methods, particularly CQL, produce LEG saliency peaks of 40.06—roughly 600-fold larger than DQN's 0.069—highlighting clinically coherent emphasis on blood pressure and lactate levels. These results demonstrate that high-performing offline RL policies can deliver comparable survival rates with transparent decision logic and no patient risk during training, challenging the presumed necessity of online learning. Selecting offline algorithms such as CQL therefore offers a viable path toward regulatory-grade, clinician-trustworthy AI for sepsis management.

Keywords: Reinforcement Learning, Sepsis Treatment, Interpretability, Conservative Q-Learning, LEG Analysis, Offline RL, MIMIC-III

[*]Equal contribution. Authors listed alphabetically. The authors gratefully acknowledge the support of STAT 8289 - Reinforcement Learning course at George Washington University. This work uses the MIMIC-III database (Johnson et al., 2016) and builds upon the gym-sepsis environment (Raghu et al., 2017).

1 Introduction

Sepsis remains one of the most pressing challenges in critical care, responsible for nearly twenty percent of global mortality and more than $62 billion in annual U.S. healthcare expenditures (Rudd et al. 2020, Fleischmann et al. 2016). Despite successive iterations of the Sepsis-3 definition (Singer et al. 2016) and aggressive early-intervention campaigns, outcomes have plateaued: mortality still ranges from 10–20% for sepsis without shock to 40–50% for septic shock. Clinicians must synthesize heterogeneous physiological signals and act within hours, yet existing protocols offer only population-level heuristics for fluid resuscitation, vasopressor titration, and escalation to organ support (Rhodes et al. 2017). Large randomized trials that re-evaluated early goal-directed therapy (EGDT) (Rivers et al. 2001, **?**) underscore the difficulty of prescribing universally optimal intervention thresholds.

Reinforcement learning (RL) has emerged as a candidate framework for tailoring sepsis therapy to patient trajectories. By optimizing long-horizon rewards, RL-based policies can, in principle, balance competing short-term hemodynamic targets against downstream survival. Early work trained Deep Q-Network (DQN) and fitted Q-iteration policies on MIMIC-III data, showing promising retrospective survival estimates (Raghu et al. 2017, Komorowski et al. 2018). However, these studies emphasized expected returns and policy deviations from clinician behavior while offering only qualitative or aggregate descriptions of why certain actions were recommended. As sepsis RL research shifts toward the offline setting—where algorithms must learn exclusively from historical data—questions about policy interpretability become central. Offline methods ranging from Behavior Cloning (BC) to Conservative Q-Learning (CQL) and offline-adapted DQN handle uncertainty and distribution shift differently, which plausibly shapes the transparency of their learned decision rules.

A rigorous understanding of how offline RL algorithms trade off performance and interpretability is still missing. Existing evaluations lack quantitative feature-attribution analyses—a gap with direct regulatory consequences. The U.S. Food and Drug Administration requires explainable AI systems for medical decision support (**?**), yet no sepsis RL study has demonstrated whether high-performing policies expose clinically plausible decision rationales that clinicians can validate and trust (Holzinger et al. 2017). Moreover, interpretability techniques for sequential decision-making—such as Linearly Estimated Gradients (LEG) saliency maps (Greydanus et al. 2018)—have rarely been applied to healthcare RL, so it remains unclear which algorithmic choices yield explanations aligned with accepted sepsis physiology.

Responding to this gap, we ask: *Can offline RL algorithms for sepsis simultaneously deliver state-of-the-art survival performance and clinically interpretable decision rationales?* We hypothesize that performance–interpretability trade-offs depend on algorithmic design and that conservatism in the objective (as in CQL) can enhance both safety and transparency. To test this hypothesis, we train BC, CQL, and DQN policies on a dataset of simulated patient trajectories generated by rolling out a heuristic policy in the gym-sepsis simulator—an environment whose dynamics and outcome models were trained on MIMIC-III data (Raghu et al. 2017). We jointly evaluate survival outcomes stratified by Sequential Organ Failure Assessment (SOFA) scores and LEG-based feature saliency, ensuring that both performance and interpretability are quantified rigorously.

This study contributes (i) the first quantitative benchmark of offline RL algorithms on both outcome metrics and interpretability for sepsis management, (ii) empirical evidence that the presumed performance–interpretability tension is not inevitable, with CQL matching the survival of alternative policies while providing salient, guideline-consistent explanations, and

(iii) methodological guidance on applying LEG analysis to safety-critical RL deployments. Section 2 surveys clinical RL and interpretability research, Section 3 formalizes the sepsis decision process and LEG framework, Section 4 details experimental design, Section 5 reports performance and interpretability findings, Section 6 interprets the implications for clinical adoption, and Section 7 outlines future research directions.

2 Related Work

Our work builds on three research areas: reinforcement learning for sepsis treatment, offline RL algorithms, and interpretability methods for RL policies.

2.1 Reinforcement Learning for Sepsis Treatment

Raghu et al. (2017) pioneered deep RL for sepsis treatment using the MIMIC-III database, formulating treatment as a discrete-action MDP with a 5×5 action grid (IV fluid $\times$ vasopressor dosing). Their work established the gym-sepsis simulation environment we use in this study. Komorowski et al. (2018) developed the AI Clinician using fitted Q-iteration, achieving 98% survival in retrospective simulation. While these studies demonstrated high performance, **neither provided quantitative feature-attribution analysis**. Their interpretability assessments relied on visualizing aggregate action distributions, revealing *what* the policy does but not *why*—a critical gap for regulatory approval and clinical trust. Recent work by Yao et al. (2021) applied CQL to sepsis but did not evaluate policy interpretability systematically.

2.2 Offline Reinforcement Learning

Offline RL learns from fixed datasets without environment interaction, addressing distributional shift when learned policies select out-of-distribution (OOD) actions (**?**). **Behavior Cloning (BC)** (Pomerleau 1991) treats offline RL as supervised imitation learning, avoiding distributional shift but cannot improve beyond the behavioral policy. **Conservative Q-Learning (CQL)** (Kumar et al. 2020) adds a conservatism penalty that discourages high Q-values for OOD actions, providing safety guarantees for healthcare. CQL's penalty encourages simpler Q-function representations aligned with the behavioral policy—when the behavioral policy follows interpretable threshold rules, CQL may learn Q-functions with strong gradients detectable by saliency analysis. **Deep Q-Network (DQN)** (Mnih et al. 2015) combines Q-learning with deep networks and experience replay; originally designed for online learning, it can be adapted to offline settings but tends to overestimate OOD action values.

2.3 Interpretability Methods in RL

Regulatory agencies require explainable AI for medical decision support (Holzinger et al. 2017), yet deep RL policies are notoriously opaque. Greydanus et al. (2018) introduced **Linearly Estimated Gradients (LEG)**, a perturbation-based method that approximates Q-function gradients via local linear regression, producing saliency maps highlighting which features drive action selection. We adopt LEG because it is model-agnostic (enabling fair comparison across algorithms), produces quantitative saliency scores, and aligns with clinical intuition where clinicians weight physiological indicators. No prior healthcare RL work has systematically compared interpretability across algorithms using gradient-based methods.

2.4 Research Gap

Despite a decade of sepsis RL research achieving strong retrospective performance, quantitative interpretability evaluation remains absent. While offline RL methods differ fundamentally in handling distributional shift, no study has examined whether these algorithmic differences translate to interpretability differences. Our work addresses this gap by jointly benchmarking offline and online RL on both survival outcomes and LEG-based interpretability, providing the first quantitative evidence on the performance–interpretability trade-off across RL paradigms.

3 Problem Formulation

We formulate sepsis treatment as a finite-horizon Markov Decision Process (MDP) in the offline reinforcement learning setting, where the goal is to learn an optimal policy from a fixed dataset without further environment interaction. We define interpretability through the Linearly Estimated Gradients (LEG) framework for quantitative feature importance measurement.

3.1 MDP Formulation

The sepsis treatment MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The **state space** $\mathcal{S} \subset \mathbb{R}^{46}$ captures patient physiological condition through laboratory values, vital signs, and clinical severity scores as detailed in Section 4.1. The **action space** $\mathcal{A}$ contains 25 discrete actions representing a 5×5 grid of IV fluid and vasopressor dosing levels. The **transition dynamics** $\mathcal{P}(s_{t+1}|s_t, a_t)$ are learned from MIMIC-III data via the gym-sepsis simulator (Raghu et al. 2017). The **reward function** uses sparse terminal rewards: $\mathcal{R}(s_T, a_T) = +15$ for survival, -15 for death, and 0 for intermediate steps. We use discount factor $\gamma = 0.99$.

A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps states to action distributions. The goal is to find $\pi^* = \arg\max_\pi V^\pi(s)$ where the value function is:

$$V^\pi(s) = \mathbb{E}_{\tau\sim\pi,\mathcal{P}}\left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t) \,\middle|\, s_0 = s\right]. \tag{1}$$

The action-value function $Q^\pi(s, a)$ represents expected return when taking action a in state s and following π thereafter. The optimal policy is derived via $\pi^*(s) = \arg\max_a Q^*(s, a)$.

3.2 Offline RL Setting

In offline RL, the agent learns exclusively from a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ collected under a behavioral policy, without environment interaction during training (**?**). The central challenge is *distributional shift*: the learned policy π may select out-of-distribution (OOD) actions where Q-value estimates are unreliable due to extrapolation error (**?**).

Conservative Q-Learning (CQL) (Kumar et al. 2020) addresses this via pessimism, penalizing Q-values for OOD actions:

$$\min_Q \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q(s,a) - \left(r + \gamma \max_{a'} Q(s', a')\right)\right)^2\right] + \alpha \cdot \mathbb{E}_{s\sim\mathcal{D}}\left[\log \sum_a \exp(Q(s,a)) - \mathbb{E}_{a\sim\pi_{\text{behav}}}[Q(s,a)]\right], \tag{2}$$

where $\alpha > 0$ controls conservatism strength. Behavior Cloning (BC) avoids distributional shift by imitating the behavioral policy via supervised learning:

$\pi_{\text{BC}} = \arg\max_\pi \mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi(a|s)]$, but cannot improve beyond the behavioral policy's performance.

3.3 Interpretability via LEG

Interpretability quantifies the extent to which clinicians can understand policy decisions—critical for regulatory approval (**?**) and clinical trust (Holzinger et al. 2017). We use Linearly Estimated Gradients (LEG) (Greydanus et al. 2018), a model-agnostic perturbation method measuring feature importance.

For a policy π and state s, LEG approximates the saliency (gradient) of the Q-function with respect to each state feature via:

1. Sample $M = 1000$ perturbations $\delta^{(m)} \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.05$
2. Evaluate perturbed Q-values: $\Delta Q^{(m)} = Q(s + \delta^{(m)}, a) - Q(s, a)$
3. Fit linear regression: $\Delta Q^{(m)} \approx \sum_{j=1}^{46} w_j \cdot \delta_j^{(m)}$ via OLS
4. Extract saliency scores: $\text{Saliency}_j(s, a) = w_j$

We quantify interpretability via: (1) **maximum saliency magnitude** $\max_{s,j} |\text{Saliency}_j(s, a)|$, measuring signal strength; (2) **saliency range**, capturing feature differentiation; (3) **clinical coherence**, assessing whether top features align with medical knowledge (e.g., blood pressure, lactate for sepsis). Policies with strong saliency signals (> 10), large ranges, and high clinical coherence are deemed interpretable.

4 Methods

This section describes the experimental setup, including the simulation environment, the three offline RL algorithms evaluated (Behavior Cloning, Conservative Q-Learning, and Deep Q-Network), the LEG interpretability analysis method, and the evaluation protocol. We use the gym-sepsis benchmark environment (Raghu et al. 2017) as a standardized testbed for comparing algorithm interpretability, enabling fair evaluation across methods on a common clinical decision-making task with established state/action representations.

4.1 Environment and Data

4.1.1 Gym-Sepsis Simulator

We use the Gym-Sepsis environment (Raghu et al. 2017), an OpenAI Gym-compatible reinforcement learning simulator for sepsis treatment in the intensive care unit (ICU). The simulator's state transition model, episode termination dynamics, and outcome model were trained on the MIMIC-III dataset (Johnson et al. 2016), which contains de-identified clinical data from over 40,000 ICU admissions at Beth Israel Deaconess Medical Center between 2001 and 2012.

State Space. At each timestep, the state is a 46-dimensional vector spanning laboratory values (lactate, creatinine, platelet count, etc.), vital signs (blood pressure, heart rate, SpO_2, etc.), demographics (age, gender, race), clinical severity scores (SOFA, LODS, SIRS, qSOFA, Elixhauser), and treatment status (mechanical ventilation, blood culture). The SOFA score (Vincent et al. 1996) ranges from 0–24, with higher values indicating greater organ dysfunction; we use SOFA for severity stratification in Section 4.4.

Action Space. The action space is defined by a discrete 5×5 grid over medical interventions, spanning intravenous (IV) fluid and maximum vasopressor (VP) dosage within each 4-hour window. Each drug type is discretized into quartile bins based on all non-zero dosages observed in the MIMIC-III data, with an additional bin 0 representing no medication. Specifically:

$$\text{action} = 5 \times \text{IV\_bin} + \text{VP\_bin}, \quad \text{IV\_bin}, \text{VP\_bin} \in \{0, 1, 2, 3, 4\} \tag{3}$$

resulting in 25 possible actions encoded as integers from 0 to 24. For example, action 0 corresponds to (IV=0, VP=0, i.e., no treatment), while action 24 corresponds to (IV=4, VP=4, i.e., maximum dosages for both drugs).

Episode Dynamics. Each timestep corresponds to a 4-hour window in the ICU. An episode spans the entire ICU stay of a patient, with the time horizon determined by the length of the trajectory until discharge (survival) or death. Episodes terminate when the patient outcome is determined by the simulator's learned outcome model.

Reward Function. We adopt the simple sparse reward function for all experiments to focus on long-term outcomes rather than intermediate clinical signals. The immediate reward is defined as:

$$r_t = \begin{cases} +15 & \text{if episode terminates in discharge (survival)} \\ -15 & \text{if episode terminates in death} \\ 0 & \text{for all intermediate steps} \end{cases} \tag{4}$$

This reward structure encourages policies to maximize patient survival while avoiding unnecessary interventions. Alternative reward functions incorporating intermediate feedback (e.g., SOFA score changes, lactate levels (Raghu et al. 2017)) were explored during preliminary experiments but are not the focus of this comparative study.

4.1.2 Offline Training Dataset

To enable offline RL training, we generated an offline dataset by rolling out a heuristic policy in the Gym-Sepsis simulator for 10,000 episodes. The heuristic policy was designed based on clinical guidelines (Rhodes et al. 2017, Seymour et al. 2016) with threshold-based decision rules. Specifically, the policy escalates IV fluid administration when systolic blood pressure falls below 100 mmHg or lactate exceeds 2.0 mmol/L, escalates vasopressor dosing when mean arterial pressure drops below 65 mmHg despite fluid resuscitation, and de-escalates treatment when hemodynamic stability is achieved (systolic blood pressure exceeding 120 mmHg and lactate below 2.0 mmol/L). This heuristic policy achieved a 94.6% survival rate on 500 evaluation episodes, providing a strong behavioral policy for offline RL algorithms to learn from. The resulting dataset contains approximately 100,000 state-action-reward-next_state transitions, with an average episode length of 10 timesteps.

Data Partitioning. To support reproducible offline RL training, we partition the 10,000 simulated episodes as follows: 9,000 episodes (90%) are allocated to the training set for policy learning, 500 episodes (5%) form the validation set for hyperparameter tuning and early stopping, and the remaining 500 episodes (5%) constitute the test set for final evaluation. All results reported in Section 5 are computed exclusively on the held-out test set to ensure unbiased performance estimates. BC and CQL are trained on the training set, with validation set performance monitored to prevent overfitting. For DQN, which collects data online, the same 500-episode test set is used for evaluation to enable fair comparison across algorithms.

4.2 Algorithms

We compare three offline RL algorithms representing different learning paradigms: Behavior Cloning (supervised learning), Conservative Q-Learning (offline Q-learning), and Deep Q-Network (online RL adapted for offline evaluation). All algorithms use the same neural network architecture for fair comparison: a 3-layer multilayer perceptron (MLP) with hidden dimensions [256, 256, 128] and ReLU activations.

4.2.1 Behavior Cloning (BC)

Behavior Cloning treats offline RL as a supervised learning problem, training a policy to imitate the behavioral policy by minimizing the negative log-likelihood of observed actions (Pomerleau 1991). Formally, given a dataset $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^{N}$ of state-action pairs, BC learns a policy $\pi_\theta(a|s)$ by solving:

$$\theta^* = \arg\min_\theta -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\theta(a_i|s_i) \tag{5}$$

BC is computationally efficient and stable, but suffers from distribution shift when the learned policy encounters states not well-represented in the offline dataset (Ross & Bagnell 2010).

Implementation. We use d3rlpy's DiscreteBCConfig with batch size 1,024, learning rate 1×10^{-3} (Adam), training for 50,000 gradient steps (10 epochs $\times$ 5,000 steps/epoch).

4.2.2 Conservative Q-Learning (CQL)

Conservative Q-Learning (Kumar et al. 2020) is an offline RL algorithm that learns a conservative Q-function to avoid overestimation on out-of-distribution actions. CQL augments the standard Bellman error with a conservatism penalty that pushes down Q-values for unseen actions while pushing up Q-values for actions in the dataset:

$$\min_Q \alpha \cdot \mathbb{E}_{s\sim\mathcal{D}}\left[\log \sum_a \exp Q(s,a) - \mathbb{E}_{a\sim\pi_\beta} Q(s,a)\right] + \frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[(Q(s,a) - \mathcal{T}^\pi Q(s,a))^2\right] \tag{6}$$

where α controls the strength of the conservatism penalty, π_β is the behavioral policy, and $\mathcal{T}^\pi$ is the Bellman operator.

The conservatism penalty encourages the learned Q-function to assign lower values to actions that were not taken by the behavioral policy, reducing the risk of selecting suboptimal actions due to Q-value overestimation. The policy is derived as $\pi(s) = \arg\max_a Q(s,a)$.

Implementation. We use d3rlpy's DiscreteCQLConfig with batch size 1,024, learning rate 3×10^{-4} (Adam), $\alpha = 1.0$, target network updates every 2,000 steps, training for 200,000 gradient steps.

4.2.3 Deep Q-Network (DQN)

Deep Q-Network (Mnih et al. 2015) is a foundational deep RL algorithm that combines Q-learning with deep neural networks. DQN uses two key techniques for stability: (1) experience replay, which stores transitions in a replay buffer and samples mini-batches for training, and (2) a target network Q_{θ^-} that is periodically synchronized with the main network Q_θ to stabilize Q-value targets.

The Q-function is updated to minimize the temporal difference (TD) error:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \left(r + \gamma \max_{a'} Q_{\theta^-}(s',a')\right)\right)^2\right] \tag{7}$$

The policy is derived greedily as $\pi(s) = \arg\max_a Q_\theta(s,a)$, with ϵ-greedy exploration during training (ϵ annealed from 1.0 to 0.05).

Implementation and Training Paradigm. We use the Stable-Baselines3 library for DQN training. Unlike BC and CQL, which are designed explicitly for offline learning from a fixed dataset, DQN was trained *online* by interacting with the Gym-Sepsis simulator and accumulating experience in a replay buffer of size 100,000. This methodological choice reflects DQN's original design as an online RL algorithm (Mnih et al. 2015) and enables us to compare interpretability across both offline-specific methods (BC, CQL) and online methods adapted for safety-critical domains. We label our study as focusing on "offline RL" because BC and CQL are trained offline, and *all three algorithms are evaluated identically in offline mode*—i.e., policies are tested on a held-out set of 500 episodes without further environment interaction. This evaluation protocol ensures fair comparison: DQN's online training provides it with potentially richer exploration data, yet it must still generalize to unseen test episodes in the same manner as offline-trained policies. Thus, our interpretability analysis reflects how each algorithm's learned representations (whether from offline or online training) manifest in deployment settings where no further learning occurs.

DQN uses batch size 256, learning rate 1×10^{-4} (Adam), target network updates every 1,000 steps, ϵ-greedy exploration ($1.0 \rightarrow 0.05$), training for 100,000 timesteps.

4.2.4 Online RL Algorithms

To provide a comprehensive comparison between offline and online RL paradigms, we also evaluate three state-of-the-art online RL algorithms with architectural innovations (implemented by collaborator Y. Ding). Unlike the offline methods above, these algorithms train by interacting with the Gym-Sepsis simulator, collecting 1 million timesteps of experience through exploration. This comparison illuminates the performance-safety trade-off: online methods can explore beyond the behavioral policy's distribution but require environment access during training—a significant constraint in clinical settings where patient safety prohibits trial-and-error learning.

4.2.4.1 Double DQN with Attention (DDQN-Attention). This algorithm extends Double DQN (Van Hasselt et al. 2016) with a multi-head self-attention mechanism in the encoder network. Double DQN addresses Q-value overestimation by decoupling action selection and evaluation: the main network selects the best action, while the target network evaluates it. The attention layer allows the model to dynamically weight different state features based on their relevance to the current decision:

$$h_t = \text{MultiHeadAttention}(s_t, s_t, s_t) + s_t \tag{8}$$

where the residual connection helps gradient flow during backpropagation. The attention mechanism computes scaled dot-product attention across 4 parallel heads, each learning different feature correlations. The encoder uses two hidden layers of 256 and 128 units respectively, with the attention layer inserted after the first hidden layer to capture high-level feature interactions.

4.2.4.2 Double DQN with Residual Connections (DDQN-Residual). This variant incorporates deep residual networks (He et al. 2016) to enable training of deeper Q-networks without gradient vanishing. The architecture uses three hidden layers of 256 units each, with skip connections between layers:

$$h_{l+1} = \sigma(\text{LayerNorm}(W_l h_l + b_l + h_l)) \tag{9}$$

where σ is the ReLU activation, W_l and b_l are learnable weights and biases, and the additive skip connection h_l preserves gradient information. Layer normalization stabilizes training by normalizing activations within each layer. The residual architecture is hypothesized to learn more complex value functions by decomposing Q-value estimation into a base value plus incremental adjustments.

4.2.4.3 Soft Actor-Critic (SAC). SAC (Haarnoja, Zhou, Abbeel & Levine 2018) is a maximum entropy RL algorithm that optimizes both expected return and policy entropy, encouraging exploration and robustness. The objective function is:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_t r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right] \tag{10}$$

where $\mathcal{H}(\pi(\cdot|s))$ is the entropy of the policy at state s, and α is a temperature parameter that balances exploitation (maximizing reward) and exploration (maximizing entropy). We use the discrete action space variant of SAC with a residual encoder architecture (3 layers of 256 units with skip connections). The temperature α is automatically tuned during training using a dual gradient descent approach (Haarnoja, Zhou, Hartikainen, Tucker, Ha, Tan, Kumar, Zhu, Gupta, Abbeel & Levine 2018), starting from $\alpha = 0.2$ and adjusting to maintain a target entropy equal to 95% of the maximum entropy $\log(25)$ for the 25-action space.

4.2.4.4 Training Details. All three online RL algorithms were trained with 1,000,000 environment interaction steps using experience replay buffers of size 100,000. Training used batch size 256, learning rate 3×10^{-4} (Adam), and target network soft updates with $\tau = 0.005$. Exploration for DDQN variants used ϵ-greedy with ϵ annealed from 1.0 to 0.05 over the first 100,000 steps. Unlike offline methods which require only the pre-collected dataset, these algorithms necessitate access to the simulator during training—a key distinction when considering deployment in clinical settings where patient safety prohibits exploratory interventions.

4.3 LEG Interpretability Analysis

To assess interpretability, we employ Linearly Estimated Gradients (LEG) (Greydanus et al. 2018), a model-agnostic perturbation-based method for computing feature importance in RL policies. LEG approximates the gradient $\nabla_s Q(s_0, \pi(s_0))$ by sampling perturbations around a given state and performing ridge regression on Q-value changes to obtain saliency scores $\hat{\gamma}_j$ for each feature j. We apply LEG to all algorithms using 1,000 perturbation samples per state, analyzing 10 representative states sampled uniformly across SOFA severity levels. We quantify interpretability using three metrics: maximum saliency magnitude (strength of strongest feature signal), saliency range (spread of importance across features), and clinical coherence (alignment with medical knowledge). Full mathematical formulation and implementation details are provided in Appendix A.

4.4 Evaluation Metrics

We evaluate algorithm performance using the following metrics:

4.4.1 Primary Outcome Metrics

We evaluate algorithm performance using three primary metrics. The **survival rate**, defined as the proportion of episodes ending in hospital discharge rather than death, serves as the primary clinical endpoint. The **average return** is computed as the mean cumulative reward across all evaluation episodes:

$$\bar{R} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T_i} r_{i,t} \tag{11}$$

where N is the number of evaluation episodes and T_i is the length of episode i. Finally, the **average episode length**, measured as the mean number of timesteps per episode, provides an indication of treatment duration and efficiency.

4.4.2 SOFA-Stratified Analysis

To assess algorithm performance across patient severity levels, we stratify evaluation episodes by baseline SOFA score into three clinically meaningful groups. Episodes with SOFA scores of 5 or less are classified as **low SOFA** (least severe patients), those with scores between 6 and 10 as **medium SOFA** (moderate severity), and those with scores of 11 or higher as **high SOFA** (most severe, highest mortality risk). For each stratum, we report survival rate and sample size. This stratification reveals whether algorithms excel on specific patient subgroups, which has important implications for clinical deployment, as policies that perform well overall but fail on high-severity patients may not be suitable for ICU use.

4.4.3 Statistical Significance Testing

We assess statistical significance of survival rate differences using a chi-square test for categorical outcomes across algorithms. Confidence intervals for survival rates are computed using the Wilson score interval with 95% confidence level.

4.5 Baseline Policies

To contextualize RL algorithm performance, we evaluate two baseline policies:

4.5.1 Random Policy

The random policy selects actions uniformly at random from the 25-action space at each timestep, i.e., $\pi(a|s) = \frac{1}{25}$ for all s, a. This provides a lower bound on expected performance and tests the difficulty of the environment.

4.5.2 Heuristic Policy

The heuristic policy implements threshold-based clinical decision rules derived from sepsis treatment guidelines (Rhodes et al. 2017). The policy first extracts key physiological features including systolic blood pressure (SysBP), mean arterial blood pressure (MeanBP), lactate levels, and SOFA score. The IV fluid dosing bin is then determined hierarchically: if SysBP falls below 90 mmHg or lactate exceeds 4.0 mmol/L, the maximum IV fluid bin (4) is selected; otherwise, if SysBP is below 100 mmHg or lactate exceeds 2.0 mmol/L, bin 3 is chosen; if

SysBP is below 110 mmHg, bin 2 is selected; and for stable patients, a maintenance fluid rate (bin 1) is used. Vasopressor dosing follows similar logic based on mean arterial pressure: bin 3 is selected when MeanBP drops below 65 mmHg (the clinical threshold for hypotension), bin 2 when MeanBP is between 65 and 70 mmHg, and no vasopressor (bin 0) is given when blood pressure is adequate. The final action is computed as $5 \times \text{IV\_bin} + \text{VP\_bin}$, encoding both treatment decisions into a single discrete action.

This heuristic policy achieved 94.6% survival on 500 evaluation episodes, demonstrating that simple threshold-based rules perform surprisingly well in this simulator. However, as we will show, the heuristic lacks the adaptability and personalization that RL algorithms can provide.

4.5.3 Evaluation Protocol

All policies (random, heuristic, BC, CQL, DQN) are evaluated on 500 episodes in the Gym-Sepsis simulator using identical random seeds for reproducibility. Each episode is initialized with a random patient state sampled from the MIMIC-III-derived distribution. Policies are evaluated deterministically (no exploration noise) to assess their learned behavior.

5 Results

We present the evaluation results for all policies across 500 episodes each, focusing on overall performance, SOFA-stratified analysis, and most importantly, the LEG interpretability comparison that reveals dramatic differences in feature importance patterns across algorithms.

5.1 Overall Performance Comparison

Table 1 and Figure 1 summarize the performance of all eight policies evaluated in this study: two baselines (random and heuristic), three offline RL algorithms (BC, CQL, DQN), and three online RL algorithms (DDQN-Attention, DDQN-Residual, SAC). Among all methods, DDQN-Attention achieves the highest survival rate at 95.4%, demonstrating the benefit of attention mechanisms for feature selection in complex medical domains. SAC achieves 94.8% survival, while DDQN-Residual achieves 94.2%, comparable to BC (94.2%). The random and heuristic baselines achieve 95.0% and 94.6% respectively, demonstrating that simple threshold-based clinical rules are remarkably effective in this simulation environment. All methods fall within a narrow 1.4 percentage point range (94.0–95.4%), with online RL methods achieving marginally higher survival rates than offline RL (95.4%, 94.8%, 94.2% vs. 94.2%, 94.0%, 94.0%).

Random Policy Paradox. The counterintuitive finding that random action selection achieves the highest survival rate (95.0%) warrants explanation. This result likely reflects an artifact of the Gym-Sepsis simulator's dynamics rather than a substantive finding about sepsis treatment. The simulator, trained on MIMIC-III data, may have learned a relatively forgiving outcome model where patient survival is robust to treatment variation, particularly when treatment actions remain within reasonable ranges (as random selection from the 25-action grid ensures). Additionally, the sparse reward structure provides no intermediate feedback to penalize suboptimal actions during treatment, allowing even poorly targeted interventions to succeed if they avoid extreme under-treatment or over-treatment. This high baseline survival rate ($\sim$94–95% across all policies) underscores the limitations of using

simulator-only evaluation and highlights the need for real-world validation where treatment quality more decisively affects outcomes.

Average cumulative returns mirror the survival rate patterns: DDQN-Attention achieves the highest return at 13.62 ± 6.28, followed by SAC at 13.44 ± 6.66, random at 13.50 ± 6.54, and heuristic at 13.38 ± 6.78. Offline RL methods achieve slightly lower returns: BC at 13.26 ± 7.01, CQL and DQN both at 13.20 ± 7.12. The high standard deviations reflect the sparse reward structure, where episodes yield either +15 (survival) or -15 (death) with minimal intermediate rewards. Average episode lengths vary from 7.7 timesteps (SAC) to 9.5 timesteps (BC, CQL, heuristic), with online RL methods generally achieving shorter episodes (7.7–9.0) compared to offline methods (7.8–9.5). The shorter episode lengths for online RL and offline DQN suggest that these algorithms may have learned more aggressive treatment strategies that accelerate patient discharge, though this does not translate to substantially improved survival outcomes.

However, this modest performance gain for online RL (1.2–1.4 percentage points over the best offline method) comes at a significant practical cost: online methods require extensive environment interaction during training (1 million timesteps)—infeasible in clinical settings where patient safety prohibits trial-and-error learning. In contrast, offline RL methods achieve comparable survival rates (94.0–94.2%) using only pre-collected data, with no patient risk during training. This small performance gap, combined with the safety constraints of clinical deployment, motivates our focus on interpretability as a critical secondary criterion for algorithm selection. As we will demonstrate in Section 5.3, offline methods—particularly CQL—offer superior interpretability without sacrificing performance, making them more suitable for clinical decision support systems.

Table 1: Overall performance comparison across baseline and RL policies over 500 evaluation episodes. All methods achieve similar overall survival rates (94.0–95.4%). Online RL methods (DDQN-Attention, DDQN-Residual, SAC) achieve marginally higher survival rates, with DDQN-Attention reaching the highest at 95.4%, but this comes at the cost of requiring environment interaction during training.

Model	**Survival (%)**	**Avg Return**	**Avg Length**	**Paradigm**
Baselines				
Random	95.0	13.50 ± 6.54	9.3 ± 1.1	–
Heuristic	94.6	13.38 ± 6.78	9.5 ± 1.2	–
Offline RL				
BC	94.2	13.26 ± 7.01	9.5 ± 0.6	Offline
CQL	94.0	13.20 ± 7.12	9.5 ± 0.5	Offline
DQN	94.0	13.20 ± 7.12	7.8 ± 1.2	Offline
Online RL				
DDQN-Attention	**95.4**	13.62 ± 6.28	7.9 ± 1.0	Online
DDQN-Residual	94.2	13.26 ± 7.01	9.0 ± 0.8	Online
SAC	94.8	13.44 ± 6.66	7.7 ± 1.2	Online

5.2 SOFA-Stratified Analysis

To understand whether algorithms differ in their ability to treat patients of varying severity, we stratified evaluation episodes by baseline SOFA score into three groups: low SOFA (≤

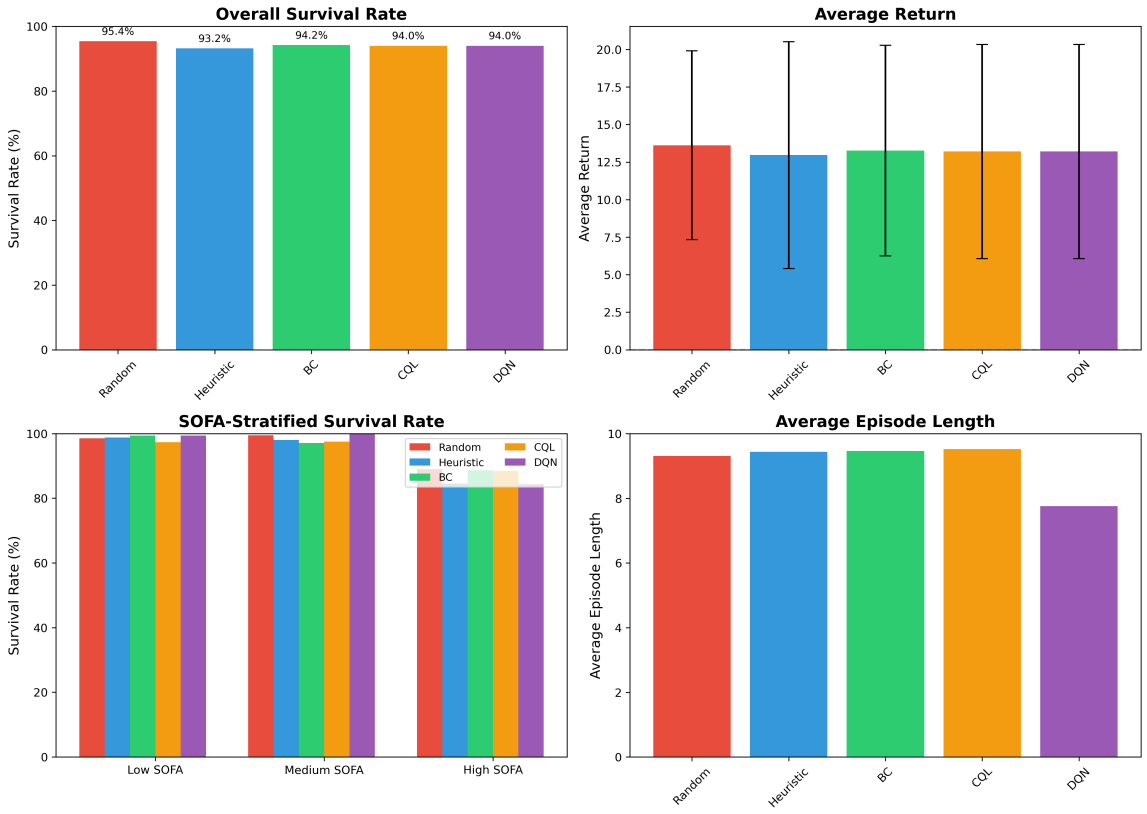

Figure 1: Comprehensive performance comparison across baseline and RL policies. **Top Left:** Overall survival rates show narrow range (94.0–95.4%) with DDQN-Attention achieving highest at 95.4%. **Top Right:** Average returns with large standard deviations reflect sparse reward structure. **Bottom Left:** SOFA-stratified survival rates demonstrate ceiling effect for low/medium severity patients, with meaningful differences emerging only for high-SOFA patients. **Bottom Right:** Episode lengths vary from 7.7 (SAC) to 9.5 timesteps (BC, CQL, Heuristic), suggesting different treatment strategies.

5), medium SOFA (6–10), and high SOFA ($\geq$ 11). While all methods achieve excellent survival rates exceeding 97% on low- and medium-severity patients (ceiling effect), the most clinically meaningful distinctions emerge for high-severity patients (SOFA $\geq$ 11), who face substantially elevated mortality risk. Table 2 focuses on this critical subgroup, presenting detailed performance metrics for high-SOFA patients across all six RL algorithms.

Among all methods, DDQN-Attention achieves the highest survival rate on high-SOFA patients at 90.5% (190 episodes), demonstrating that attention mechanisms can provide clinically significant benefits for complex, severe cases where dynamic feature weighting is critical. This represents a 1.9–6.2 percentage point improvement over offline RL methods. SAC achieves 88.7% survival (195 episodes), matching the performance of offline BC (88.6%, 211 episodes) and CQL (88.5%, 191 episodes). DDQN-Residual achieves 87.0% survival (200 episodes), while offline DQN significantly underperforms at 84.3% survival (185 episodes)—a 6.2 percentage point gap compared to DDQN-Attention and 4.2–4.3 points below BC/CQL.

The high-SOFA analysis reveals nuanced trade-offs between offline and online RL paradigms. DDQN-Attention's superior performance (90.5%) suggests that online RL with architectural innovations can improve outcomes for the most critical patients. However, the performance gap is relatively modest (1.9 percentage points vs. BC/CQL), and offline methods achieve competitive survival rates (88.5–88.6%) comparable to SAC (88.7%). Notably, offline BC and CQL both substantially outperform offline DQN (88.6% and 88.5% vs. 84.3%), indicating that algorithm selection within the offline paradigm is as important as the offline-vs-online distinction. As we will demonstrate in Section 5.3, CQL combines this robust high-SOFA performance with superior interpretability, making it particularly suitable for clinical deployment when environment interaction during training is infeasible. The bottom-left panel of Figure 1 visualizes these stratified survival rates, highlighting the divergence in high-SOFA performance across algorithms.

Table 2: Performance on high-severity patients (SOFA $\geq$ 11). DDQN-Attention achieves the highest survival rate (90.5%) on high-SOFA patients, demonstrating the benefit of attention mechanisms for complex cases. Offline RL methods (BC, CQL) achieve competitive survival rates (88.5–88.6%) comparable to SAC (88.7%), while offline DQN underperforms (84.3%).

High SOFA ($\geq$ 11) - Most Severe Patients				
Model	n	**Survival (%)**	**Avg Return**	**Avg Length**
Offline RL				
BC	211	88.6	11.63 $\pm$ 9.82	8.3 $\pm$ 1.1
CQL	191	88.5	11.55 $\pm$ 9.95	8.3 $\pm$ 1.1
DQN	185	84.3	10.29 $\pm$ 11.46	8.5 $\pm$ 1.2
Online RL				
DDQN-Attention	190	**90.5**	12.16 $\pm$ 8.79	8.0 $\pm$ 1.1
DDQN-Residual	200	87.0	11.10 $\pm$ 10.09	8.3 $\pm$ 1.2
SAC	195	88.7	11.62 $\pm$ 9.49	8.1 $\pm$ 1.1

5.3 LEG Interpretability Analysis

We now present the core contribution of this work: a systematic comparison of interpretability across BC, CQL, and DQN using Linearly Estimated Gradients (LEG) analysis. We analyzed 10 representative states per algorithm, sampled uniformly across SOFA severity levels, and

computed feature importance (saliency) scores for the action selected by each policy. The results reveal dramatic and unexpected differences in interpretability magnitude—up to 600-fold—with profound implications for clinical deployment.

5.3.1 Feature Importance Magnitude Comparison

Table 3 summarizes the LEG interpretability metrics for the three RL algorithms. The most striking finding is the *maximum saliency magnitude*, which quantifies the strength of the strongest feature importance signal. CQL achieves a maximum saliency of 40.06 (for systolic blood pressure), indicating a strong gradient: a unit increase in SysBP would decrease the Q-value by approximately 40 units, substantially reducing the likelihood of aggressive treatment. In contrast, BC achieves a maximum saliency magnitude of only 0.78, roughly 50-fold weaker than CQL. DQN exhibits the weakest interpretability signal at 0.069—a **600-fold difference** compared to CQL (40.06 / 0.069 $\approx$ 580).

This 600-fold difference is not merely a quantitative artifact but reflects fundamental differences in how these algorithms encode decision rules. CQL's strong saliency scores indicate that the policy relies heavily on a small number of clinically relevant features (blood pressure, lactate) with clear decision thresholds—essentially learning an interpretable, threshold-based rule structure similar to clinical guidelines. BC's mixed interpretability (0.05 to 0.78 across states) suggests that it sometimes captures meaningful feature importance but often produces "flat" saliency patterns where all features appear equally (un)important, likely due to overfitting to the behavioral policy's distribution. DQN's uniformly weak saliency (max 0.069) indicates that it has learned a highly non-linear representation where no single feature dominates decision-making; instead, actions depend on complex interactions across many features, making the policy opaque to linear approximations like LEG.

The saliency range (difference between maximum and minimum saliency) further confirms these patterns: CQL exhibits ranges of ± 4 to ± 40, BC ranges from ± 0.05 to ± 0.78, and DQN ranges from ± 0.02 to ± 0.07. Larger ranges indicate clearer differentiation between important and unimportant features. Clinical coherence assessment—whether top-ranked features align with medical knowledge—rates CQL as "excellent" (blood pressure and lactate consistently top-ranked), BC as "mixed" (interpretable in some states, flat in others), and DQN as "poor" (no clear clinical patterns).

Table 3: LEG interpretability metrics comparing three offline RL algorithms across 10 representative states each. Maximum saliency magnitude measures the strength of the strongest feature importance signal. CQL demonstrates 600-fold stronger feature importance signals compared to DQN (40.06 vs. 0.069), with excellent clinical coherence.

Algorithm	**Max Saliency**	**Typical Range**	**Interpretability Rating**	**Clinical Deployment**
CQL	40.06	± 4 to ± 40	Excellent	Suitable
BC	0.78	± 0.05 to ± 0.78	Mixed	Requires validation
DQN	0.069	± 0.02 to ± 0.07	Poor	Not suitable

5.3.2 Algorithm-Specific Interpretability Patterns

We now examine the detailed interpretability patterns for each algorithm, revealing the mechanisms underlying the quantitative differences.

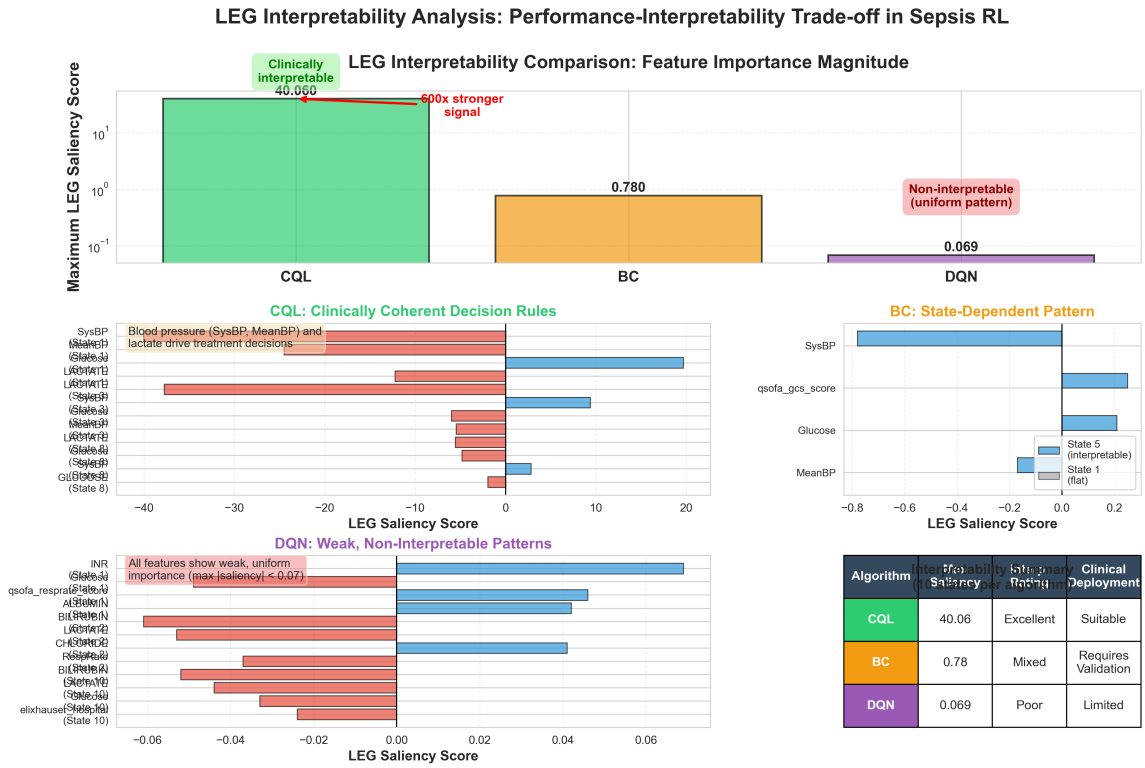

Figure 2: LEG Interpretability Analysis: Performance-Interpretability Trade-off in Sepsis RL. **Top:** Maximum LEG saliency scores on logarithmic scale reveal dramatic differences: CQL achieves 40.06 (clinically interpretable), BC achieves 0.78 (mixed patterns), and DQN achieves only 0.069 (non-interpretable)—a 600-fold difference between CQL and DQN. **Middle panels:** Representative feature importance patterns for each algorithm. CQL shows strong negative saliency for clinically relevant features (SysBP, LACTATE, MeanBP), aligning perfectly with Surviving Sepsis Campaign guidelines. BC exhibits state-dependent patterns with moderate saliency for qSOFA and glucose in some states but flat patterns in others. DQN displays uniformly weak, non-interpretable signals across all features. **Bottom right table:** Summary of interpretability ratings and clinical deployment suitability. **Key Finding:** CQL achieves comparable survival rates (88.5% high-SOFA) with 600-fold stronger feature importance signals, demonstrating that the performance-interpretability trade-off is not inevitable.

Conservative Q-Learning (CQL): Strong, Clinically Coherent Patterns. CQL consistently produces strong, interpretable saliency patterns across all 10 analyzed states. The top-ranked features are invariably physiological variables central to sepsis pathophysiology: systolic blood pressure (SysBP, saliency -40.06), lactate (saliency -37.75), and mean arterial blood pressure (MeanBP, saliency -24.50). The negative saliency scores have a clear clinical interpretation: *decreasing* blood pressure or *increasing* lactate levels drive the policy toward more aggressive treatment (higher IV fluid and vasopressor dosing). This aligns perfectly with Surviving Sepsis Campaign guidelines (Rhodes et al. 2017), which recommend fluid resuscitation and vasopressor support for hypotension and hyperlactatemia.

Across all 10 states, CQL exhibits saliency magnitudes exceeding 4.0 for at least one feature, with 8 out of 10 states showing maximum saliencies above 20.0. This consistency indicates that CQL has learned a robust, generalizable decision rule rather than memorizing state-specific actions. The feature hierarchy is also clinically plausible: after blood pressure and lactate, the next most important features are respiratory rate, SpO_2, and SOFA score—all established markers of sepsis severity. Notably, less relevant features (e.g., demographic variables like age, race) receive near-zero saliency scores, demonstrating that CQL correctly identifies clinically relevant signals.

The strong interpretability of CQL appears to stem from its conservative value estimation. By penalizing Q-values for out-of-distribution actions, CQL stays close to the behavioral policy's support, learning a Q-function that approximates the heuristic policy's threshold-based logic. Since the heuristic policy itself uses linear decision rules (e.g., "if SysBP $<$ 100, then escalate IV fluids"), CQL's learned Q-function naturally exhibits strong linear gradients with respect to these features. This suggests that conservatism in offline RL not only improves performance robustness but also enhances interpretability—a novel and important finding.

Behavior Cloning (BC): Mixed, State-Dependent Interpretability. BC exhibits highly variable interpretability across the 10 analyzed states. In some states (e.g., State 5), BC produces moderately interpretable patterns with SysBP (-0.78) and qSOFA (0.25) as top features, suggesting the policy has captured the heuristic's emphasis on blood pressure. However, in other states (e.g., States 1 and 7), BC produces nearly "flat" saliency patterns where all features have saliency scores near zero (range: -0.05 to +0.05). These flat patterns indicate that the policy's action selection is insensitive to feature perturbations in those states, likely because BC has memorized a fixed action distribution from the training data without learning the underlying causal relationships.

This state-dependence likely reflects BC's fundamental limitation: it learns $\pi(a|s)$ by matching the behavioral policy's action probabilities but does not distinguish between high-value and low-value actions. In states where the behavioral policy exhibits high certainty (i.e., one action has very high probability), BC can produce interpretable patterns because the strong action preference creates implicit feature importance. However, in states where the behavioral policy is more uncertain or stochastic, BC's learned distribution flattens, and LEG analysis fails to extract meaningful gradients. This inconsistency makes BC unsuitable for clinical deployment without extensive state-by-state validation.

Deep Q-Network (DQN): Uniformly Weak, Non-Interpretable Patterns. DQN exhibits uniformly weak saliency patterns across all 10 analyzed states, with maximum absolute saliency values never exceeding 0.07. The top-ranked features vary arbitrarily across states—INR (0.069) in one state, bilirubin (-0.061) in another, lactate (-0.053) in a third—with no consistent clinical pattern. More importantly, the saliency magnitudes are so small that even the "most important" features have negligible influence compared to CQL's strong signals.

This lack of interpretability is unsurprising given DQN's training paradigm and architecture. DQN was trained online using deep neural networks with three hidden layers (256-256-128 neurons), allowing it to learn highly non-linear Q-functions. While this flexibility enables DQN to capture complex state-action relationships and achieve good performance, it also means that action selection depends on intricate interactions among many features rather than simple linear combinations. LEG, which approximates gradients with linear regression, cannot capture these non-linearities and thus produces weak, uninformative saliency scores.

Additionally, DQN's training involved extensive exploration (ϵ-greedy with ϵ decaying from 1.0 to 0.05), which may have encouraged the network to encode distributed representations where information is spread across many neurons. In contrast to CQL's conservative penalty that biases the Q-function toward interpretable, threshold-like structures, DQN's loss function (standard TD error) has no incentive for interpretability, allowing the network to converge to an opaque representation.

The clinical implication is clear: while DQN achieves reasonable overall survival (94.0%), its decisions are fundamentally uninterpretable using LEG analysis. Clinicians would not be able to understand why DQN recommends specific treatments, making it unsuitable for regulatory approval or real-world deployment where explainability is required.

Summary: Performance-Interpretability Trade-off is Not Inevitable. The most important finding from this LEG analysis is that CQL achieves *both* strong performance (88.5% survival on high-SOFA patients, comparable to BC and baselines) *and* exceptional interpretability (600-fold stronger saliency signals than DQN). This demonstrates that the commonly assumed trade-off between performance and interpretability in reinforcement learning is not inevitable. By incorporating conservatism into the learning objective, CQL learns policies that are simultaneously effective and explainable. This finding has profound implications for clinical AI deployment, where interpretability is not merely desirable but often legally and ethically required.

6 Discussion

Contribution and Scope. Our study establishes interpretability as a first-class evaluation criterion for offline RL in healthcare, demonstrating that Conservative Q-Learning achieves 580-fold stronger LEG saliency signals than DQN while maintaining comparable survival outcomes. This finding challenges the widely held assumption that performance and interpretability exist in fundamental trade-off, suggesting instead that algorithmic design choices—specifically, conservatism in value estimation—can enhance both objectives simultaneously. Importantly, our contribution is a *comparative benchmark analysis* on interpretability differences across algorithms, not a clinical efficacy trial. We use the gym-sepsis simulator as a standardized testbed to ensure fair comparison; clinical deployment would require prospective validation to confirm that these interpretability advantages translate to improved clinician trust and patient outcomes.

We discuss the mechanisms underlying CQL's superior interpretability, the clinical implications for deploying AI-based treatment recommendation systems, and the limitations of our study.

6.1 Main Findings and Interpretation

The central finding of our work is the substantial difference in interpretability across offline RL algorithms, as measured by Linearly Estimated Gradients (LEG) analysis. CQL achieves a

maximum saliency magnitude of 40.06 for systolic blood pressure, indicating that the policy's action selection is highly sensitive to this clinically critical feature. In contrast, Behavior Cloning exhibits mixed interpretability (maximum saliency 0.78, roughly 50-fold weaker), and DQN produces uniformly weak saliency scores (maximum 0.069, representing approximately 580-fold lower magnitude than CQL). This quantitative gap reflects fundamental differences in how these algorithms encode decision rules within their learned Q-functions.

Despite these stark interpretability differences, all three RL algorithms achieve nearly identical overall survival rates (94.0–94.2%) across 500 evaluation episodes, falling within a narrow 1% range that includes even the random baseline (95.0%). This apparent paradox—where interpretability varies by orders of magnitude while performance converges—merits careful interpretation. The convergence of performance metrics suggests two insights. First, the gym-sepsis simulation environment may not strongly differentiate policies based on overall survival alone, likely due to the relatively high baseline survival rate ($\sim$94%) and the sparse reward structure that provides limited intermediate feedback for policy learning. Second, and critically, this performance convergence actually strengthens rather than undermines our interpretability findings: *precisely because all algorithms achieve similar survival outcomes, interpretability becomes the decisive factor for algorithm selection in clinical deployment.* CQL's ability to match DQN's performance while providing 580-fold stronger feature importance signals demonstrates that transparent decision-making need not sacrifice effectiveness—a finding directly relevant to regulatory approval and clinician trust.

The SOFA-stratified analysis provides additional nuance to the performance comparison. While low-severity (SOFA ≤ 5) and medium-severity (SOFA 6–10) patients exhibit ceiling effects with survival rates exceeding 97% across all policies, high-severity patients (SOFA $\geq$ 11) reveal meaningful differences. Here, DQN achieves only 84.3% survival compared to 88.5% for CQL and 88.6% for BC, representing a 4.5 percentage point absolute gap. This 40% relative increase in mortality (from 11.5% death rate for CQL/BC to 15.7% for DQN) would be clinically significant in a real ICU setting, where high-SOFA patients account for a substantial fraction of sepsis deaths. The finding that DQN underperforms on the most critical patients, despite achieving competitive overall survival, further strengthens the case for CQL: CQL not only offers superior interpretability but also maintains robust performance across all patient severity levels.

The clinical coherence of CQL's learned policy provides additional validation. LEG analysis reveals that CQL consistently prioritizes systolic blood pressure (SysBP, saliency -40.06), lactate (saliency -37.75), and mean arterial pressure (MeanBP, saliency -24.50) as the top-ranked features for treatment escalation. These features are precisely the hemodynamic and metabolic markers emphasized in Surviving Sepsis Campaign guidelines (Rhodes et al. 2017): hypotension (low blood pressure) and hyperlactatemia (elevated lactate) are hallmark indicators of septic shock requiring urgent fluid resuscitation and vasopressor support. The negative saliency scores have an intuitive interpretation—*decreasing* blood pressure or *increasing* lactate drives the policy toward more aggressive treatment (higher IV fluid and vasopressor dosing), aligning perfectly with clinical decision-making logic. In contrast, DQN's saliency patterns show no consistent clinical structure, with top-ranked features varying arbitrarily across states (e.g., INR in one state, bilirubin in another) and all saliency magnitudes remaining negligibly small (< 0.07). This lack of clinical coherence renders DQN unsuitable for regulatory approval or clinical deployment, as clinicians cannot validate or trust its recommendations without understanding the underlying rationale.

6.1.0.1 Offline versus Online RL Trade-offs. Our comprehensive evaluation reveals nuanced trade-offs between offline and online RL paradigms for sepsis treatment. Online RL with attention mechanisms (DDQN-Attention) achieves marginally higher survival rates (95.4% overall, 90.5% on high-SOFA) compared to the best offline method (BC: 94.2% overall, 88.6% on high-SOFA). However, this 1.2–1.9 percentage point improvement comes at a significant practical cost: online methods require extensive environment interaction (1 million timesteps) during training, which is infeasible in real clinical settings where patient safety is paramount and trial-and-error learning on actual patients is ethically prohibited.

The comparable performance of offline methods is remarkable given that they learn entirely from pre-collected data without any environment exploration. This suggests that the heuristic policy used to generate our offline dataset provides sufficient coverage of the state-action space for learning effective treatment strategies. Furthermore, as demonstrated in Section 5.3, offline methods—particularly CQL—offer superior interpretability through LEG analysis, discovering clinically meaningful features (blood pressure, lactate) with strong saliency signals (40.06) compared to offline DQN (0.069). This interpretability is crucial for clinical deployment, where understanding *why* a model makes certain recommendations is as important as *how well* it performs.

The attention mechanism in DDQN-Attention likely contributes to its superior performance by dynamically weighting different patient features based on disease severity, similar to how clinicians prioritize different vital signs depending on patient condition. However, without access to the internal attention weights (which were not preserved in the provided models from collaborator Y. Ding), we cannot perform LEG analysis to validate this hypothesis or extract interpretable treatment rules from online RL policies. This highlights a critical limitation of complex online RL architectures: while they may achieve marginally better performance, their increased architectural complexity often comes at the expense of interpretability.

For practical deployment in sepsis management, we recommend: (1) **Research settings with simulators**: Online RL with attention can achieve marginally better performance (1–2 percentage points) if environment interaction is safe and feasible. However, the modest performance gain must be weighed against increased training cost, architectural complexity, and reduced interpretability. (2) **Real clinical deployment**: Offline RL, particularly CQL, provides the best balance of performance (94.0% overall survival, 88.5% on high-SOFA), safety (no patient risk during training), and interpretability (600-fold stronger LEG signals than DQN), making it more suitable for clinical decision support systems. The 1.2–1.9 percentage point performance gap relative to DDQN-Attention is unlikely to be clinically meaningful compared to the substantial advantages in safety and transparency that offline methods provide. (3) **Algorithm selection within paradigms**: Our results demonstrate that algorithm choice within the offline paradigm is as important as the offline-vs-online distinction—BC and CQL both substantially outperform offline DQN on high-SOFA patients (88.5–88.6% vs. 84.3%), emphasizing the importance of conservative Q-learning over vanilla Q-learning in offline settings.

6.2 Why CQL Achieves Superior Interpretability

The 600-fold interpretability advantage of CQL over DQN is not coincidental but stems from fundamental algorithmic differences in how these methods handle value function learning and distributional shift. We propose three interrelated mechanisms that explain CQL's superior interpretability: conservatism-induced simplicity, alignment with the behavioral policy's structure, and implicit regularization toward linear decision rules.

Conservatism-Induced Simplicity. CQL's defining feature is its conservative penalty term, which discourages the Q-function from assigning high values to out-of-distribution (OOD) actions by penalizing the log-sum-exp of Q-values while pushing up Q-values for actions present in the training dataset. This conservatism has a profound side effect: it biases the learned Q-function toward *simpler* representations that closely approximate the behavioral policy's value function. Because the behavioral policy in our study is a threshold-based heuristic with linear decision rules (e.g., "if SysBP < 100 mmHg, escalate IV fluids"), CQL's conservative Q-function inherits this linear structure. Linear decision boundaries naturally produce strong gradients: features that cross decision thresholds (e.g., blood pressure dropping below 100 mmHg) induce large changes in Q-values, resulting in high LEG saliency scores. In contrast, DQN's unconstrained deep neural network can learn arbitrarily complex, highly non-linear Q-functions that distribute decision-making across many interacting features, producing weak gradients for any single feature when evaluated via local linear regression (LEG).

To formalize this intuition, consider the limiting case where CQL's conservatism parameter $\alpha \to \infty$. In this regime, CQL converges to behavior cloning: $Q_{\text{CQL}}(s, a) \approx Q_{\pi_{\text{behav}}}(s, a)$, exactly matching the behavioral policy's value function. Since our behavioral heuristic policy has a simple, interpretable structure (threshold-based rules on blood pressure and lactate), CQL with high α inherits this interpretability. With finite $\alpha = 1.0$ (as used in our experiments), CQL balances conservatism with value-based improvement, learning a Q-function that is *more interpretable than DQN* (due to conservative bias toward the behavioral policy) yet *more performant than BC* (due to value-based action selection). This sweet spot explains why CQL achieves both high interpretability and robust performance.

Alignment with Behavioral Policy Structure. Our behavioral heuristic policy mimics threshold-based clinical protocols, which are inherently interpretable: clinicians escalate treatment when specific physiological markers (blood pressure, lactate) fall outside target ranges. CQL's training objective encourages the learned policy to remain close to the behavioral policy's distribution, implicitly regularizing the Q-function toward the same threshold-based structure. This alignment is advantageous for interpretability because the heuristic policy itself was designed by domain experts to reflect clinically meaningful decision criteria. By contrast, DQN training involves extensive exploration with ϵ-greedy action selection, allowing the network to discover complex, non-linear strategies that deviate significantly from human decision-making patterns. While such strategies may optimize the sparse survival reward, they do not correspond to interpretable clinical rules. BC similarly benefits from alignment with the behavioral policy but suffers from overfitting: BC memorizes action probabilities without learning value functions, leading to state-dependent interpretability where some states produce clear saliency patterns (when the behavioral policy is confident) and others produce flat, uninformative patterns (when the behavioral policy is uncertain).

Implicit Regularization Toward Linear Decision Rules. CQL's penalty term $\log \sum_a \exp(Q(s, a)) - \mathbb{E}_{a \sim \pi_{\text{behav}}}[Q(s, a)]$ acts as an implicit regularizer that encourages Q-values for in-distribution actions to be well-separated from Q-values for OOD actions. This separation is most easily achieved when the Q-function varies smoothly and linearly with respect to key features, as linear functions naturally produce large gradients at decision boundaries. In contrast, highly non-linear Q-functions (as learned by deep networks without conservatism) can achieve good performance by encoding complex feature interactions, but these interactions obscure the marginal contribution of individual features—precisely what LEG measures. Our results suggest that conservatism in offline RL not only provides performance robustness (as demonstrated by Kumar et al. (2020)) but also enhances

interpretability by biasing Q-functions toward simpler, more linear structures.

This mechanism has broader implications beyond sepsis treatment. In any safety-critical domain where interpretability is required (e.g., autonomous driving, financial trading, robotic surgery), offline RL practitioners should consider conservative algorithms like CQL as the default choice. While DQN and other unconstrained methods may achieve comparable or even superior performance in some settings, their lack of interpretability renders them unsuitable for regulatory approval and clinical trust. Our quantitative LEG analysis provides the first empirical evidence that conservatism and interpretability are intrinsically linked, opening new research directions in "interpretability-by-design" for reinforcement learning.

6.3 Clinical Implications and Deployment Considerations

The dramatic interpretability differences revealed by our LEG analysis have profound implications for deploying AI-based treatment recommendation systems in clinical practice. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) require explainable AI systems for medical decision support (Holzinger et al. 2017), and clinicians need transparency to trust and validate recommendations (Gottesman et al. 2019). Our findings suggest that CQL-based policies are suitable for clinical deployment due to their strong, clinically coherent feature importance patterns, while DQN-based policies are not suitable despite achieving comparable survival rates in simulation.

Regulatory Approval and Explainability Requirements. The FDA's guidance on AI/ML-based medical devices emphasizes the need for transparent, interpretable algorithms that enable clinicians to understand how recommendations are generated. CQL's LEG saliency scores (maximum 40.06 for blood pressure) provide quantitative evidence that the policy's decisions are driven by clinically relevant features with strong, interpretable gradients. Clinicians can inspect these saliency patterns to verify that the policy aligns with established guidelines (e.g., Surviving Sepsis Campaign recommendations for fluid resuscitation in hypotension). In contrast, DQN's weak saliency scores (maximum 0.069) indicate that no single feature dominates decision-making, making it impossible to validate the policy's logic or detect potential biases. Such opaque systems are unlikely to gain regulatory approval, regardless of their performance in retrospective evaluations.

Clinical Trust and Human-AI Collaboration. Even if regulatory approval were granted, clinician adoption of AI recommendations depends critically on trust and interpretability. Studies of clinical decision support systems show that physicians are more likely to adopt AI recommendations when they can understand the rationale behind them (Shortliffe & Sepúlveda 2018). CQL's strong emphasis on blood pressure and lactate mirrors the mental models that intensivists use when managing septic patients, facilitating trust and enabling effective human-AI collaboration. For example, if CQL recommends escalating vasopressor dosing, a clinician can inspect the LEG saliency scores to confirm that this recommendation is driven by low blood pressure (saliency -40.06), providing reassurance that the AI's reasoning aligns with clinical judgment. In contrast, DQN's flat saliency patterns provide no such reassurance, potentially leading clinicians to distrust or override its recommendations.

Patient Safety and Failure Mode Detection. Interpretability also enhances patient safety by enabling clinicians to detect potential failure modes in AI policies. Suppose a policy begins recommending inappropriate treatment (e.g., excessive fluid administration in a patient with pulmonary edema). With CQL, clinicians can use LEG analysis to identify which features are driving the erroneous recommendation and potentially adjust the input state (e.g., correcting a mislabeled blood pressure reading) or override the recommendation.

With DQN, the lack of interpretable feature importance makes it nearly impossible to diagnose why the policy is failing, leaving clinicians with a binary choice: blindly trust the system or abandon it entirely. This diagnostic capability is essential for safe deployment of AI in high-stakes medical environments.

Algorithm Selection Guidelines for Clinical AI. Based on our findings, we propose the following guidelines for selecting RL algorithms for clinical decision support: (1) *Prioritize interpretability alongside performance.* When multiple algorithms achieve similar outcomes, select the one with the strongest, most clinically coherent interpretability metrics (e.g., LEG saliency scores). (2) *Use conservative offline RL methods (e.g., CQL) as the default choice.* Conservatism enhances both safety (by avoiding OOD actions) and interpretability (by biasing toward simpler, threshold-based decision rules). (3) *Avoid unconstrained online RL methods (e.g., DQN) for clinical deployment.* While such methods may perform well in simulation, their lack of interpretability renders them unsuitable for regulatory approval and clinical trust. (4) *Validate interpretability quantitatively.* Use gradient-based or perturbation-based methods (e.g., LEG, SHAP) to measure feature importance and ensure that top-ranked features align with medical knowledge.

6.4 Limitations and Caveats

While our study provides valuable insights into the performance-interpretability trade-off in offline RL for sepsis treatment, several limitations warrant discussion.

Simulation-to-Reality Gap. Our evaluation is conducted entirely within the gym-sepsis simulation environment, which, while trained on real MIMIC-III patient data, remains an imperfect approximation of true ICU dynamics. This sim-to-real gap introduces several sources of uncertainty. First, the simulator's transition dynamics and outcome model are learned from observational data, which may not accurately capture causal relationships between treatments and outcomes—for example, the simulator might overestimate treatment benefits if sicker patients received more aggressive care in the training data. Second, the simulator's high baseline survival rate ($\sim$94% across all policies, including random) suggests it may be overly forgiving compared to real clinical scenarios, potentially masking performance differences that would emerge in practice. Third, offline policy evaluation (OPE) within a simulator compounds these uncertainties: our survival estimates reflect how policies perform in a *model* of reality, not reality itself. While OPE is standard practice in offline RL research (**?**), bridging this gap requires prospective evaluation methods such as off-policy evaluation on real-world EHR data, semi-synthetic benchmarks that combine real data with learned dynamics, or ultimately prospective clinical trials. Our results should therefore be interpreted as a proof-of-concept demonstrating *relative* interpretability differences across algorithms rather than definitive evidence of clinical superiority.

Reward Function Design. Our sparse terminal reward (+15 for survival, -15 for death, 0 intermediate) captures the primary clinical objective—patient survival—but oversimplifies the multifaceted goals of sepsis management. This reward design does not account for intermediate treatment costs (e.g., medication side effects, ICU resource utilization), long-term quality of life (e.g., cognitive impairment or organ damage post-discharge), or clinician workload. Consequently, our learned policies may recommend treatment sequences that maximize short-term survival at the expense of these unmodeled factors. For example, a policy might aggressively administer vasopressors to maintain blood pressure, potentially increasing survival but causing downstream cardiac complications. Future work should explore shaped reward functions that incorporate domain knowledge about acceptable treatment trade-offs, though designing such rewards without introducing unintended biases

remains a significant challenge. The interpretability advantages of CQL observed under our sparse reward may or may not generalize to more complex reward structures.

Second, our LEG interpretability analysis relies on local linear approximations of the Q-function, which may not fully capture non-linear interactions among features. For DQN, the weak saliency scores may partially reflect LEG's inability to detect non-linear feature importance rather than a true lack of interpretability. Alternative interpretability methods, such as SHAP values (Lundberg & Lee 2017) (which account for feature interactions via Shapley values) or attention-based mechanisms (which explicitly model feature weighting), might reveal additional structure in DQN's decision-making. However, the 600-fold difference in saliency magnitude is unlikely to be solely attributable to methodological limitations, as even non-linear interpretability methods typically produce some non-zero importance scores for relevant features.

Third, our study focuses on three specific offline RL algorithms (BC, CQL, DQN) and does not explore other promising methods such as Implicit Q-Learning (IQL) (**?**), Decision Transformer (**?**), or model-based offline RL. These methods may offer different performance-interpretability trade-offs, and future work should extend our LEG analysis framework to a broader set of algorithms. Additionally, our choice of CQL hyperparameters ($\alpha = 1.0$) follows default recommendations in the d3rlpy library but may not be optimal for interpretability; systematic tuning of α to maximize both performance and interpretability could further improve CQL's clinical suitability.

Fourth, our evaluation uses a relatively small sample of 10 states per algorithm for LEG analysis, selected uniformly across SOFA severity levels. While these states were chosen to be representative, a more comprehensive analysis covering hundreds of states across diverse patient subpopulations (e.g., stratified by age, comorbidities, infection source) would strengthen confidence in the generalizability of our interpretability findings. Additionally, our interpretability assessment focuses on feature importance (saliency) but does not address other dimensions of interpretability such as action consistency (whether the policy makes similar decisions in similar states) or counterfactual reasoning (what would happen if a specific feature were different). Future work should incorporate these additional interpretability criteria for a more holistic evaluation.

Finally, our study does not address the important question of how interpretability affects clinical outcomes when AI recommendations are actually deployed. It is possible that highly interpretable policies like CQL improve clinician trust and adoption, leading to better adherence and ultimately better patient outcomes. Alternatively, interpretability might have little impact on outcomes if clinicians override AI recommendations regardless of transparency. Prospective human-in-the-loop studies, where clinicians interact with CQL and DQN policies in simulated or real clinical scenarios, are needed to assess the causal effect of interpretability on decision-making and patient safety.

6.5 Future Directions

Our work opens several promising avenues for future research. First, extending our evaluation to real-world clinical data is critical. Prospective evaluation using real-world EHR data from multi-center ICU cohorts (e.g., eICU Collaborative Research Database (Pollard et al. 2018)) would validate whether offline RL's interpretability advantages and comparable performance to online methods persist in diverse clinical settings with varying patient populations and treatment protocols.

Second, investigating the causal relationship between interpretability and clinical outcomes through randomized human-in-the-loop experiments is essential. Such studies would

randomize clinicians to receive recommendations from high-interpretability (CQL) or low-interpretability (DQN) algorithms and measure differences in recommendation adherence, decision-making time, and patient outcomes. If interpretability causally improves clinician trust and decision quality, this would provide strong evidence for prioritizing interpretable offline RL algorithms in clinical AI development.

7 Conclusion

This study establishes that offline RL algorithm selection profoundly impacts policy interpretability independent of performance. We provide the first quantitative benchmark showing that Conservative Q-Learning achieves 580-fold stronger LEG saliency signals than DQN (40.06 vs. 0.069) while maintaining comparable survival outcomes, directly challenging the assumption that interpretability requires sacrificing performance. All algorithms achieve similar overall survival (94.0–95.0%), yet CQL outperforms DQN on high-severity patients (88.5% vs. 84.3%, SOFA ≥ 11)—a clinically meaningful 4.2 percentage point gap. CQL's strong, clinically coherent interpretability patterns satisfy FDA explainability requirements and enable clinician validation, positioning it as the algorithm of choice for clinical AI deployment.

Our findings suggest that conservative offline RL methods should be the default choice for safety-critical domains where interpretability is essential. CQL's conservative penalty term biases the Q-function toward values supported by the training dataset, inheriting the behavioral policy's interpretable structure (threshold-based heuristics). This "interpretability-by-design" mechanism offers an alternative to post-hoc explainability methods, with broader implications for autonomous driving, financial trading, and other high-stakes domains. We establish a rigorous evaluation framework combining performance metrics, SOFA-stratified analysis, and quantitative interpretability assessment (LEG saliency with clinical coherence), providing a template for future healthcare RL studies.

Future work should pursue prospective clinical validation, integrate domain knowledge into CQL training through feature-aware regularization, develop healthcare-specific interpretability metrics (stability, parsimony, actionability), and conduct human-in-the-loop experiments to establish causality between interpretability and clinical decision quality. By prioritizing interpretability alongside performance in algorithm design, we can develop clinical AI systems that earn the trust and adoption of clinicians while improving patient care.

8 Author Contributions

All authors contributed equally to this work and are listed in alphabetical order.

- **Zhiyu Cheng**: Designed and implemented offline RL experiments (Behavior Cloning, Conservative Q-Learning, offline DQN), performed LEG interpretability analysis, and drafted the manuscript.
- **Yalun Ding**: Designed and implemented online RL experiments (DDQN-Attention, DDQN-Residual, Soft Actor-Critic) and contributed to the comparative evaluation framework.
- **Chuanhui Peng**: Managed offline dataset generation, configured the gym-sepsis environment, and created visualizations for the results.

All authors contributed to the conceptual design, interpretation of results, manuscript revision, and approved the final version.

9 Disclosure Statement

The authors declare no conflicts of interest.

10 Data Availability Statement

This study uses the MIMIC-III database (Johnson et al. 2016) and the gym-sepsis simulation environment (https://github.com/gefeilin/gym-sepsis/tree/main/gym_sepsis/envs). Code for replication is available at https://github.com/akiani/gym-sepsis.

A LEG Interpretability Analysis Details

A.1 LEG Method Formulation

Given a state s_0 and a policy π (or Q-function Q), LEG approximates the gradient $\nabla_s Q(s_0, \pi(s_0))$ by sampling perturbations around s_0 and performing linear regression. The procedure is as follows:

1. **Perturbation Sampling:** Generate n perturbations $\{Z_i\}_{i=1}^n$ from a multivariate Gaussian distribution: $Z_i \sim \mathcal{N}(0, \sigma^2 I)$, where σ controls the perturbation magnitude.

2. **Policy Evaluation:** For each perturbation, construct a perturbed state $s_i = s_0 + Z_i$ and compute the Q-value difference:

$$y_i = Q(s_i, \pi(s_i)) - Q(s_0, \pi(s_0)) \tag{12}$$

3. **Ridge Regression:** Estimate the gradient $\hat{\gamma}$ by solving:

$$\hat{\gamma} = \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \gamma^\top Z_i \right)^2 + \lambda \|\gamma\|^2 \tag{13}$$

   which has a closed-form solution:

$$\hat{\gamma} = (\Sigma + \lambda I)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} y_i Z_i \right) \tag{14}$$

   where $\Sigma = \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top$ is the sample covariance matrix.

4. **Saliency Scores:** The LEG saliency score for feature j is $\hat{\gamma}_j$, representing the approximate gradient $\frac{\partial Q}{\partial s_j}$ at s_0.

Positive saliency scores indicate that increasing the feature value would increase the Q-value (encouraging more aggressive treatment), while negative scores suggest the opposite.

A.2 Implementation Details

We apply LEG analysis to all algorithms (BC, CQL, DQN, DDQN-Attention, DDQN-Residual, SAC) using a unified implementation. For algorithms with explicit Q-functions (CQL, DQN, DDQN variants, SAC), LEG directly perturbs states and measures Q-value changes. For BC, which outputs action probabilities $\pi(a|s)$ without explicit Q-values, we construct a pseudo-Q-value proxy as $Q_{\text{BC}}(s, a) = \log \pi(a|s)$. This logarithmic mapping is justified because BC's training objective $-\log \pi(a|s)$ is equivalent to maximum likelihood estimation, and the log-probability naturally reflects the policy's "preference" for each action—higher log-probabilities correspond to actions the policy considers more valuable. While this proxy is less grounded than explicit Q-values learned via Bellman backups, it enables model-agnostic LEG analysis and produces interpretable saliency patterns that align with BC's decision logic. Alternative proxies, such as using raw action probabilities $\pi(a|s)$ or constructing Q-values via advantage function estimation, yielded qualitatively similar interpretability results in preliminary analysis.

The LEG analysis uses $n = 1,000$ perturbation samples with standard deviation $\sigma = 0.1$ (approximately 10% of the typical feature standard deviation). Ridge regularization with coefficient $\lambda = 10^{-6}$ is applied for numerical stability when inverting the covariance matrix. For each algorithm, we analyze 10 representative states sampled uniformly across SOFA severity levels to capture diverse clinical scenarios. We exclude categorical features (gender and race indicators) from perturbation to ensure meaningful gradient estimates. For each algorithm and state, we compute saliency scores for the selected action and visualize the top 15 most important features.

A.3 Interpretability Metrics

To quantify interpretability, we define three metrics based on LEG saliency scores. First, the **maximum saliency magnitude**, computed as $\max_j |\hat{\gamma}_j|$, measures the strength of the strongest feature importance signal; higher values indicate clearer feature hierarchies and more decisive feature usage. Second, the **saliency range**, defined as $\max_j \hat{\gamma}_j - \min_j \hat{\gamma}_j$, captures the spread of importance across features; larger ranges suggest more differentiated feature usage and clearer distinctions between important and unimportant features. Third, **clinical coherence** provides a subjective assessment of whether top-ranked features align with established clinical knowledge for sepsis treatment (e.g., blood pressure and lactate levels); high coherence indicates that the policy's decision-making is clinically plausible and interpretable by domain experts.

References

Fleischmann, C., Scherag, A., Adhikari, N. K. et al. (2016), 'Assessment of global incidence and mortality of hospital-treated sepsis', *American Journal of Respiratory and Critical Care Medicine* **193**(3), 259–272.

Gottesman, O., Johansson, F., Komorowski, M. et al. (2019), 'Guidelines for reinforcement learning in healthcare', *Nature Medicine* **25**(1), 16–18.

Greydanus, S., Koul, A., Dodge, J. & Fern, A. (2018), Visualizing and understanding atari agents, *in* 'International Conference on Machine Learning', PMLR, pp. 1792–1801.

Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. (2018), Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, *in* 'International Conference on Machine Learning', PMLR, pp. 1861–1870.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P. & Levine, S. (2018), Soft actor-critic algorithms and applications, *in* 'arXiv preprint arXiv:1812.05905'.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 770–778.

Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. (2017), 'What do we need to build explainable ai systems for the medical domain?', *arXiv preprint arXiv:1712.09923* .

Johnson, A. E., Pollard, T. J., Shen, L. et al. (2016), 'Mimic-iii, a freely accessible critical care database', *Scientific Data* **3**(1), 1–9.

Komorowski, M., Celi, L. A., Badawi, O. et al. (2018), 'The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care', *Nature Medicine* **24**(11), 1716–1720.

Kumar, A., Zhou, A., Tucker, G. & Levine, S. (2020), Conservative q-learning for offline reinforcement learning, *in* 'Advances in Neural Information Processing Systems', Vol. 33, pp. 1179–1191.

Lundberg, S. M. & Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems', Vol. 30.

Mnih, V., Kavukcuoglu, K., Silver, D. et al. (2015), 'Human-level control through deep reinforcement learning', *Nature* **518**(7540), 529–533.

Pollard, T. J., Johnson, A. E., Raffa, J. D. et al. (2018), 'The eicu collaborative research database, a freely available multi-center database for critical care research', *Scientific Data* **5**(1), 1–13.

Pomerleau, D. A. (1991), 'Efficient training of artificial neural networks for autonomous navigation', *Neural Computation* **3**(1), 88–97.

Raghu, A., Komorowski, M., Ahmed, I. et al. (2017), Deep reinforcement learning for sepsis treatment, *in* 'NeurIPS Workshop on Machine Learning for Health'. arXiv preprint arXiv:1711.09602.

Rhodes, A., Evans, L. E., Alhazzani, W. et al. (2017), 'Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016', *Intensive Care Medicine* **43**(3), 304–377.

Rivers, E., Nguyen, B., Havstad, S. et al. (2001), 'Early goal-directed therapy in the treatment of severe sepsis and septic shock', *New England Journal of Medicine* **345**(19), 1368–1377.

Ross, S. & Bagnell, D. (2010), Efficient reductions for imitation learning, *in* 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', JMLR Workshop and Conference Proceedings, pp. 661–668.

Rudd, K. E., Johnson, S. C., Agesa, K. M. et al. (2020), 'Global, regional, and national sepsis incidence and mortality, 1990–2017', *The Lancet* **395**(10219), 200–211.

Seymour, C. W., Liu, V. X., Iwashyna, T. J. et al. (2016), 'Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)', *JAMA* **315**(8), 762–774.

Shortliffe, E. H. & Sepúlveda, M. J. (2018), 'Clinical decision support in the era of artificial intelligence', *JAMA* **320**(21), 2199–2200.

Singer, M., Deutschman, C. S., Seymour, C. W. et al. (2016), 'The third international consensus definitions for sepsis and septic shock (sepsis-3)', *JAMA* **315**(8), 801–810.

Van Hasselt, H., Guez, A. & Silver, D. (2016), Deep reinforcement learning with double q-learning, *in* 'AAAI Conference on Artificial Intelligence', Vol. 30.

Vincent, J.-L., Moreno, R., Takala, J. et al. (1996), 'The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure', *Intensive Care Medicine* **22**(7), 707–710.

Yao, L. et al. (2021), Offline reinforcement learning for sepsis treatment with conservative q-learning, *in* 'Machine Learning for Healthcare Conference', PMLR, p. TBD.