From Black Box to Bedside: Distilling Reinforcement Learning for Interpretable Sepsis Treatment

Ella Lan
Stanford University
ellaslan@stanford.edu

Andrea Yu
Stanford University
andreayu@stanford.edu

Sergio Charles
Stanford University
sergioc1@stanford.edu

Abstract

Sepsis is a complex and life-threatening condition requiring individualized, time-sensitive interventions. Reinforcement learning (RL) has shown promise for optimizing sepsis care, but real-world adoption is hindered by the opacity of its decision-making processes. We propose a novel two-phase framework that couples deep Q-learning with post hoc interpretability via decision tree distillation. Phase I trains deep Q-networks (DQNs) on MIMIC-III ICU trajectories, exploring ensemble methods and behavior cloning (BC) regularization for improved robustness and clinician agreement. Phase II distills the learned policies into shallow, human-readable decision trees using greedy, probabilistic, and Q-regression approaches. Our results show increased clinician agreement from 0.231 (baseline) to 0.906 (BC-DQN), without degrading policy value, while our distilled trees retain near-perfect fidelity (≥ 0.998), improving transparency. This framework can help bridge the trust gap between "black-box" medical AI and interpretable clinical technologies.

1 Introduction

Sepsis is a leading cause of mortality in intensive care units worldwide, marked by highly variable patient trajectories and a need for continuous intervention. Previous research indicates that reinforcement learning (RL) has shown great promise in optimizing personalized treatment strategies due to its reliance on sequential decision making under high uncertainty. However, the lack of interpretability in RL models poses a significant barrier to real-world clinical deployment.

This paper introduces a novel two-phase framework for sepsis management that enhances both efficacy and transparency. In Phase I, we train a deep Q-network and explore enhancements through ensembling and supervised regularization. In Phase II, we distill the learned policies into compact decision trees, enabling deeper understanding. Our results indicate that this hybrid approach balances performance and interpretability, making it a promising candidate for trustworthy AI in healthcare.

2 Related work

Research by Komorowski et al. [2018] showed that RL policies trained on historical ICU data can outperform clinician heuristics for sepsis treatment. Tu et al. [2025] used conservative Q-learning to improve model safety by discouraging rare or extreme actions. However, given the complex algorithms on which these models are based, there is insufficient trust for real-world adoption.

To address this, researchers have explored policy distillation, via training simpler models to replicate complex RL decisions. Pettit et al. [2021] learned sparse symbolic policies for sepsis care, while Bastani et al. [2019] highlighted decision trees as an interpretable yet expressive class. Kohler et al. [2025] further showed that distilled trees and shallow networks can preserve expert-level performance while enhancing transparency.

NeurIPS 2025 Workshop on Interpreting Cognition in Deep Learning Models

Finally, interpretability must align with clinical practice. Che et al. [2017] emphasized that transparency is essential for adoption in high-stakes domains, and Wu et al. [2023] integrated human expertise into value-based RL to improve safety. Together, these works point toward human-centered distillation as key to building trustworthy, interpretable clinical AI.

3 Methods

3.1 Reinforcement learning

For the first phase, we used deep Q-learning for our base RL algorithm. The Q-learning algorithm repeatedly approximates a Q-function $Q^\pi(s, a)$ representing expected future rewards. In basic Q-learning, Q-values for each state-action pair are stored in a table, but this becomes impractical quickly. Instead, in deep Q-learning, we train a neural network known as a deep Q-network (DQN) to approximate Q-values for *all* actions from a given state s. This algorithm is particularly appropriate for learning personalized sepsis treatment plans because

- it is well-suited for high-dimensional state spaces alongside low-dimensional action spaces;
- it learns an explicit Q-function, which provides a useful method for distillation in the second phase which preserves some notion of "valuable" treatment options; and
- it is off-policy, allowing it to learn the *optimal* policy regardless even if data was collected under an exploratory *behavior* policy—since clinicians cannot experiment on real patients.

We augmented our baseline model with two complementary strategies. First, we trained an ensemble of DQNs with different seeds and averaged Q-values across models. This method improved robustness by reducing variance in Q-estimates and safeguarding against unstable recommendations.

We also introduced behavior cloning (BC) regularization. We designed a hybrid objective combining RL with supervised BC, using MIMIC-III clinician treatment trajectories as expert demonstrations:

$$\mathcal{L} = \mathcal{L}_Q + \alpha \cdot \mathcal{L}_{BC}.$$

This method encouraged clinician agreement while preserving insights from Q-learning. We further experimented with hyperparameter tuning on different values of α. (If α was 0, for example, we recovered vanilla deep Q-learning; if α was larger, we risked overfitting and losing benefits of RL.)

3.2 Decision tree distillation

For the second phase, we emphasized interpretability. We enforced a maximum depth of 6 levels to prevent the trees from growing too deep and reconstructing the complexity of neural networks. By using a variety of approaches, we further explored the balance of learned insights and transparency.

Our first approach was greedy distillation. We trained a classification tree whose target label at each state s_i was the ensemble's best action $y_i = \arg\max_a Q(s_i, a)$. The optimal splits were calculated by minimizing the Gini impurity, which measures the likelihood of misclassifying new data. Our second approach was probabilistic distillation. Instead of using the raw Q-values to select target labels, we took the softmax of the values at state s_i, yielding a probability distribution over the possible actions. We randomly sampled an action from this distribution to obtain a target label and trained a classification tree using the same splitting criterion (Gini impurity). Our third approach was "Q-distillation". In this approach, we trained a regression tree instead of a classification tree, treating the Q-values for all actions at state s_i as a "Q-vector". Splitting based on a particular feature partitioned the data at a node into two child nodes D_L and D_r, and we chose our optimal splits to minimize the post-split weighted MSEs.

4 Experiments

We used the MIMIC-III dataset developed by Johnson et al. [2016], containing features such as lab results, vitals, and interventions from over 53,000 critical care patient trajectories. From this, we extracted a sepsis cohort of 17,000 patients via ICD-9 codes. Another challenge in sepsis management is ensuring reliable decision-making when faced with limited data or expertise. Thus, to mimic

low-resourced hospital environments, we also used a subset of simulated clinical time-series data derived from MIMIC-III by Khan [2022]. It is vital to evaluate how distilled models perform under real-world constraints—where interpretability and resilience to noise become especially critical.

With these datasets, we modeled sepsis patient care as a Markov decision process and used the `Stable-Baselines3` library to train our baseline RL model. The state space was represented by 24 distinct clinical chart events extracted hourly. The action space was represented by 10 discrete treatments—combinations of 5 IV levels and 2 vasopressor levels, chosen because they require constant monitoring based on evolving needs, whereas treatments such as antibiotics are typically one-time interventions. We designed the reward with $+15$ for 90-day survival, -15 for in-hospital death, -1 for aggressive interventions, and $+0.01$ per timestep that the patient remained alive.

5 Results

5.1 BC-regularization significantly outperformed baseline RL

We measured the agreement between our RL models with clinician actions, as seen in Table 1.

Table 1: Agreement of RL models with clinician-labeled actions.

Method	**Clinician agreement**
Baseline DQN	0.231
Ensemble DQN	0.224
BC-regularized DQN (best)	0.910
BC-regularized DQN (low α)	0.906

Our baseline DQN model showed low clinician agreement, indicating divergence from standard practice. Unlike clinicians bound by ethical constraints, RL agents can explore novel treatment trajectories; however, such low alignment may limit real-world use. The ensemble DQN traded slight accuracy loss for greater robustness, as averaging Q-values stabilized action selection.

Our BC-regularized DQN models performed significantly better. We achieved the highest clinician agreement using $\alpha = 1.0$, which strongly weighted the BC component that imitated clinician actions. We tuned α extensively and found that agreement remained high even with $\alpha = 0.01$: modest regularization substantially improved alignment without overfitting or overriding the RL objective.

5.2 Decision tree distillation preserved learned insights

We evaluated *fidelity* as the percentage of tree-predicted actions that aligned with the learned policy.

Table 2: Fidelity of decision trees in replicating augmented DQN policies.

Distillation method	**Fidelity to ensemble DQN**	**Fidelity to BC-DQN**
Greedy tree distillation	0.999	1.000
Probabilistic tree distillation	0.887	0.998
Q-value tree distillation	0.914	0.999

All three distillation methods demonstrated high fidelity to both the ensemble DQN and BC-regularized DQN policies, as seen in Table 2. For both policies, the methods followed the same trend, with the greedy approach performing near-perfectly and the probabilistic approach performing relatively weakest.

We hypothesize that probabilistic distillation underperformed because it is near-identical to greedy distillation, but sampling actions from the distribution of softmax Q-values introduced randomness that degraded the benefits of greedy construction. In addition, although greedy distillation excelled at matching the policy, we believe Q-distillation would be the best for real-world adoption. Using a regression approach (as opposed to a classification approach) provides richer signal about the relative

benefits of actions instead of determining a single "best" action; such nuance is vital in medical decision-making and furthers our goal of distilling complex insights into interpretable models. Indeed, comparing trees across distillation methods, we find that greedy trees offer shallower splits, while Q-regression trees often align better with reward optimization. Overall, such high fidelity scores suggest that decision tree distillation is a promising method improving interpretability—especially when the underlying policies reflect strong clinician agreement.

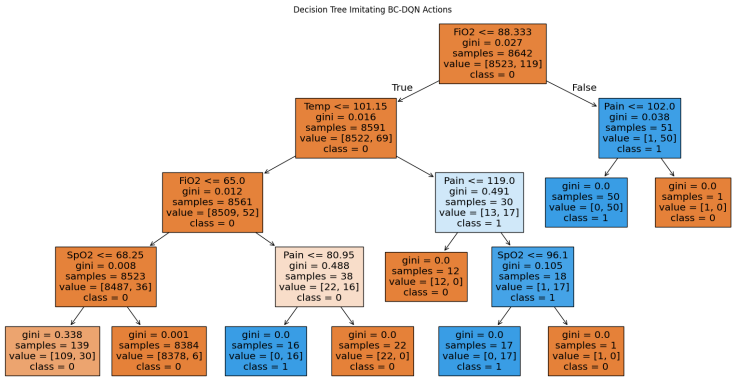

Figure 1: Decision tree built from BC-DQN using greedy distillation approach.

As shown in Figure 1, we present an example of a decision tree greedily distilled from the BC-DQN policy. Most predicted actions (i.e., the class labels) are 0 or 1, indicating conservative treatment choices that align with real-world clinician decisions. Intuitive medical choices are also highlighted in greater clarity than from a "black-box" RL model. For instance, the tree recommends intervention if a patient is suffering from high pain and their oxygen saturation (SpO_2) falls below a certain threshold.

5.3 Distilled trees aligned with additional interpretability metrics

In addition, we evaluated the interpretability of the distilled trees in three qualitative dimensions. **Structural simplicity:** the trees are shallow (depth 4–6) with short paths and many pure leaves, making decisions easy to trace. Although some branches have limited sample support, the overall structure remains compact and transparent relative to neural policies. **Semantic plausibility:** FiO_2, SpO_2, and pain consistently appear as early splits, aligning with clinical reasoning—oxygenation and ventilatory demand guide hemodynamic interventions, while pain reflects physiological stress. **Reliability:** the trees replicate parent RL policies with near-perfect fidelity, whether distilled from ensemble or BC-regularized DQNs, preserving strategic consistency. This fidelity enhances interpretability while showing that the trees inherit both strengths and limitations of the original models.

6 Conclusion

Our findings indicate that deep RL agents can effectively learn personalized sepsis treatments from high-dimensional ICU data while addressing interpretability and robustness challenges hindering clinical adoption. Ensembling and BC regularization on top of baseline DQN models led to significant enhancements in robustness and performance. To improve transparency, we distilled the complex policies into human-readable trees through a variety of methods, which all achieved high fidelity. Our future work includes refining reward functions, exploring online RL under strict safety constraints, and integrating clinician feedback. Ultimately, we believe that combining computational insights with human wisdom does not automate our decisions—but elevates our collective capacity to heal.

7 Acknowledgements

We acknowledge that this project was completed as part of the Stanford CS224R course and would like to thank Professor Chelsea Finn and her teaching team for the outstanding project award and overall support contributing to this submission.

References

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction, 2019. URL https://arxiv.org/abs/1805.08328.

Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *Proceedings of the AMIA Annual Symposium*, 2017. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333206/.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: 10.1038/sdata.2016.35.

Asjad Khan. Mimic-iii – deep reinforcement learning, 2022. URL https://www.kaggle.com/datasets/asjad99/mimiciii/data.

Hector Kohler, Quentin Delfosse, Waris Radji, Riad Akrour, and Philippe Preux. Evaluating interpretable reinforcement learning by distilling policies into programs, 2025. URL https://arxiv.org/abs/2503.08322.

Matthieu Komorowski, Leo Anthony Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24:1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

Jacob Pettit, Brenden Petersen, Robert Cockrell, Dale Larie, Felipe Silva, Gary An, and Daniel Faissol. Learning sparse symbolic policies for sepsis treatment. In *Proceedings of the Interpretable Machine Learning in Healthcare Workshop at ICML*, 2021. URL https://www.researchgate.net/publication/353224172_Learning_Sparse_Symbolic_Policies_for_Sepsis_Treatment.

Rui Tu, Zhipeng Luo, Chuanliang Pan, Zhong Wang, Jie Su, Yu Zhang, and Yifan Wang. Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards. *Human-Centric Intelligent Systems*, 5:63–76, 2025. doi: 10.1007/s44230-025-00093-7.

XiaoDan Wu, RuiChang Li, Zhen He, TianZhi Yu, and ChangQing Cheng. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *NPJ Digital Medicine*, 6(15), 2023. doi: 10.1038/s41746-023-00755-5.