Group Members: Tyler Blicharz, Abhijeet Sahdev, Dhruviben Patel
August 10th, 2025
Submitted to Professor Jing Li
DS669 Reinforcement Learning

Term Final Project Report

A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis

---

ABSTRACT

This project explores the application of value-based deep reinforcement learning (RL) to optimize treatment strategies for sepsis in ICU patients. Building upon the paper *"*A Value-Based Deep Reinforcement Learning Model with Human Expertise in Optimal Treatment of Sepsis*"* by XiaoDan Wu et al., our goal was to replicate and understand how AI can assist clinicians in making personalized, time-sensitive decisions to improve patient survival.

We focused on designing an RL environment that simulates ICU settings using patient vitals and treatment actions (IV fluids and vasopressors). We implemented the WD3QNE model using our preprocessed dataset. Data pre-processing was not highlighted in the code and this is the key point of deviation in observed results. Key contributions include dataset preprocessing from MIMIC-III, reward shaping aligned with clinical goals, and hands-on training/testing of Q-learning agents. Through experimentation, we gained critical insights into the challenges of applying RL in high-stakes healthcare environments involving double robust evaluation and identified future directions for offline policy learning and interpretable AI systems. Our work demonstrates the potential of reinforcement learning in enhancing personalized care for sepsis and beyond.

INTRODUCTION

Sepsis is a life-threatening condition that arises when the body's response to infection causes damage to its own tissues and organs. Treating sepsis in the ICU requires timely and personalized decisions by doctors, especially regarding the administration of intravenous fluids and vasopressors. However, due to the complexity of patient conditions and limited clinical time, identifying the best treatment strategy for each individual can be very challenging.

In this project, we explored how Deep Reinforcement Learning (DRL) can be used to support clinical decision-making in sepsis treatment. Inspired by the research paper *"*A Value-Based Deep Reinforcement Learning Model with Human Expertise in Optimal Treatment of Sepsis", we aimed to simulate an ICU environment using patient vitals and then train a Q-learning agent to suggest treatment actions. Our goal was to understand how reinforcement learning can help learn optimal policies that improve patient survival outcomes.

1. References and Acknowledgements

- **Reference of the paper:**
  Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). *The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care*. Nature Medicine, 24(11), 1716–1720. https://www.nature.com/articles/s41591-018-0213-5
- **Codebase and datasets:**
  - Our Codebase: Git Repository
  - Dataset: MIMIC-III Clinical
- **Other resources:**
  - Author's Codebase : Git Repository
  - MIMIC Git Repo
- **Workload distribution:**
  - Tyler Blicharz – 33% experiments and documentation
  - Abhijeet Sahdev – 33% data preprocessing, feature selection, experiments and documentation
  - Dhruviben Patel – 33% documentation and experiments

Application, MDP, and RL Model

Application Introduction

This project focuses on optimizing treatment strategies for Intensive Care Unit (ICU) patients diagnosed with sepsis by leveraging reinforcement learning (RL). Sepsis is a critical and life-threatening condition that requires rapid, precise, and personalized treatment decisions. In ICU settings, clinicians must decide on the appropriate combination of intravenous (IV) fluids and vasopressors to stabilize patient conditions and improve survival rates. However, the complexity of patient responses, variability in disease progression, and the time-sensitive nature of decisions make manual optimization challenging.

The proposed RL-based system aims to assist clinicians by analyzing patient vital signs in real time and making dynamic recommendations for medication adjustments. The RL agent continuously observes the patient's state — such as blood pressure, heart rate, SOFA score, and lactate levels — and learns to select the best treatment actions that maximize patient survival

while minimizing harm. This approach supports personalized, data-driven decision-making tailored to the unique condition of each patient.

MDP Formulation

The ICU sepsis treatment problem is modeled as a Markov Decision Process (MDP), where the agent interacts with the environment to improve patient outcomes.

- **States:** Represent patient physiological conditions using features such as heart rate, blood pressure, SOFA score, and lactate levels. These states capture the dynamic changes in patient health during ICU stay.

- **Actions:** Consist of discrete combinations of IV fluid and vasopressor dosages that can be administered. The choice of action directly impacts the patient's state in the next time step.

- **Reward:**

$$r = \begin{cases} \beta_s \times (SOFA_t - SOFA_{t+1}) & t<T \\ \delta \times \beta_T & t = T \end{cases}$$

  We assign a positive intermediate reward if the SOFA scores reduce, multiplying it with β□ which is 0.6 in our study. T is 20, i.e, twenty steps per patient. At t = 20, if the patient survives for 90 days after the onset of sepsis, we assign a reward of +24, otherwise -24 where δ is +1 if the patient survives or -1 if they die and β□ is 24.

- **Discount Factor (γ):** Used to balance short-term interventions against long-term benefits. A higher γ emphasizes improving overall survival outcomes over immediate state improvements. This is set to 0.99 in our study.

- Continuous state space and discrete action space are constructed (discussed in the later sections). The DRL agent takes action based on current state and clinician expertise.

RL Approach

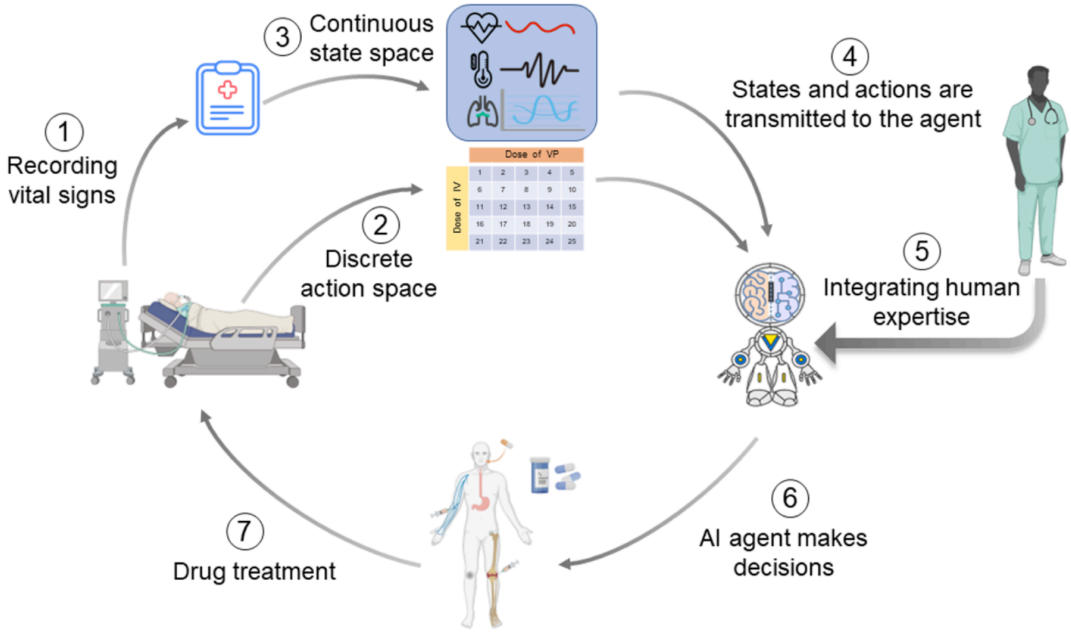


Fig 1: Dynamic treatment process of the WD3QNE agent for sepsis.

With the goal of eliminating redundancy and selecting features that are vital for our agent as well as overcoming the underestimation and overestimation of existing approaches such as Dueling Double Q-Network and Dueling Double Deep Q-Network respectively, the author proposed the following models which we have used in our study.

<u>Weighted Dueling Double Deep Network</u> : The novel target Q value function with adaptive dynamic weight *p* which ranges from 0 to 1, is as follows:

$$Q(S_{t+1}, a_{t+1}) = p \times \max_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \omega^-) + (1-p) \times Q(S_{t+1}, \operatorname{argmax}_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \omega); \omega^-)$$

Here, ω and ω⁻ are the parameters of the main network and target network.

$$p = \frac{\varphi_{a_{t+1}}}{\varphi_{a_{t+1}} + \sigma_{a_{t+1}}}$$

- $\varphi_{a_{t+1}} = \frac{\max_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \omega^-)}{\sum_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \omega^-)}$ is max. Q value divided by summation of target Q values under actions.
- $\sigma_{a_{t+1}} = \frac{Q(S_{t+1}, \arg\max_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \omega); \omega^-)}{\sum_{a_{t+1}} Q(S_{t+1}, \arg\max_{a_{t+1}} Q(S_{t+1}, a_{t+1}; \omega); \omega^-)}$, a dynamic parameter obtained from D3QN.

Finally,

$$Q(S_t, a_t) = r + \gamma Q(S_{t+1}, a_{t+1})$$

where r is reward and gamma is the discount factor.

Weighted Dueling Double Deep Network With Clinician's Embeddings:

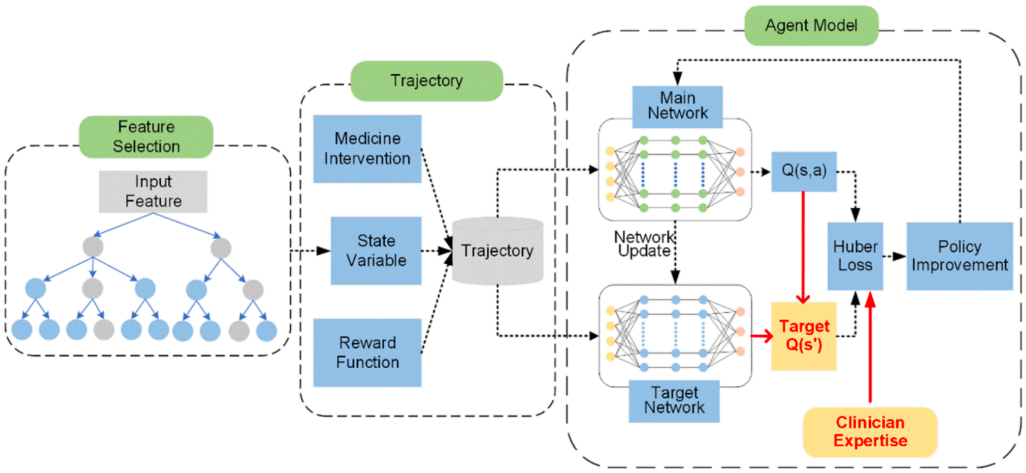


Fig 2: WD3QNE ALGORITHM STRUCTURE

- Integrates human expertise with the DRL model.
    - Provides guidance for AI and ensures higher efficiency and reliability in sepsis treatment by following clinician's directives for less severe cases (SOFA<4).

Overall, this RL formulation aims to develop an interpretable, robust, and clinically relevant decision-making system that can assist in optimizing sepsis treatment in ICU environments.

Codebase, System, and Experiment Setup

- **Libraries and system setup:**
    - Python, PyTorch, NumPy, Pandas, Matplotlib, Scikit-Learn and MySQL
    - IDE : VS Code
- **Hardware setup:**
    - M2 Chip with 16 GB unified memory**.**

Data Pre-processing

We use 4 hour bins as a time step, considering a time period of 80 hours or 20 time steps from 24 prior to suspected onset till 56 hours after it. For cohort selection, we remove any patient that has a gap of over 24 hours in their electronic health records. Then, we consider only adults from this list of patients. Next, we filter out patients that withdrew treatment or preferred comfort care measures only. Now, we follow the Sepsis-3 definite to get our desired sepsis cohort which as follows:

- Get the list of all antibiotics.
- Get all microbiological events (such as blood cultures).
- Now, define pairs of administering antibodies and conducting microbiological events :
    - Antibiotics are given anytime up to 72 hours after microbiological events.
    - Microbiological events take place within 24 hours after taking antibiotics.
    - The earliest of the two is considered as the suspected onset time.
- Measure SOFA score as per the table shown earlier for 24 hours after onset time.
    - Baseline SOFA score is conventionally assumed to be 0.
- Consider only those patients that have a change in SOFA score of +2 or more.

This resulted in a cohort of 14,505 patients with 17,240 different stays resulting in a dataset of 344,800 records as opposed to 17,083 patients with 276,232 records. This is a major reason for the deviation of our findings from the paper.

Summary statistics of the 45 features are almost comparable with deviations of around 2-3 points. However, some features varied significantly and we couldn't identify how because the authors didn't share the code for data preprocessing.

Notably, we adopted the Hierarchical Time-Aware Imputation strategy for ensuring temporal continuity.
Preference is in the decreasing order as follows:

- Original measurement
- Forward fill (recent value)
- Backward fill (upcoming value)
- Patient mean (individual baseline)
- Population mean (cohort average)
- Clinical default (normal range)

This was not discussed by the researcher in their original work. We've added the comparison table where Mean_2 and StandardDeviation(SD)_2 are metrics from the original paper and _1 corresponds to our cohort.

Feature	Mean_1	SD_1	Mean_2	SD_2	Mean_ diff	SD_diff
4 hourly	260.7	561.53	387	369	-126.3	192.53

output						
Age	76.42	54.79	64.4	17.1	12.02	37.69
Arterial BE	-0.47	4.52	0.33	5	-0.8	-0.48
BUN	28.98	21.77	4.7	2.3	24.28	19.47
CB	2498.63	5675.72	1690	1333	808.63	4342.72
Calcium	1.13	0.41	8.3	0.79	-7.17	-0.38
Chloride	104.98	3.69	104	6.27	0.98	-2.58
Creatinine	1.62	1.65	0.78	0.23	0.84	1.42
DBP	60.35	11.74	57.1	13.3	3.25	-1.56
FiO2	0.57	0.15	0.45	0.18	0.12	-0.03
GCS	14.47	1.57	12.58	3.43	1.89	-1.86
Glucose	7.58	2.72	5.7	1.1	1.88	1.62
HCO3	22.98	1.67	24	5.06	-1.02	-3.39
HGB	10.56	1.26	10.2	1.73	0.36	-0.47
HR	86.6	16.35	87	16.7	-0.4	-0.35
INR	1.5	0.84	1.5	0.82	0	0.02
Lactate	2.14	1.4	2.05	1.68	0.09	-0.28
MBP	77.38	12.6	78.2	13.4	-0.82	-0.8
Magnesium	2.05	0.33	1.11	0.14	0.94	0.19
Male (N)	8418		9604		-1186	
Non-survivors	6066		3228			
PT	16.06	6.46	16	6.64	0.06	-0.18
PTT	36.79	17.28	31	6.44	5.79	10.84
PaCO2	41.53	9	41.8	10.7	-0.27	-1.7
PaO2	140.88	73	99	23.5	41.88	49.5
PaO2/FiO2	259.27	102.2	248	107	11.27	-4.8
Patients	14505		17083		-2578	
Platelets	209.22	119.27	224	118	-14.78	1.27
Potassium	4.15	0.54	4.07	0.55	0.08	-0.01
RR	19.64	4.72	20	5.18	-0.36	-0.46
SBP	118.89	18.58	119	20.3	-0.11	-1.72
SGOT	120.39	95.59	38.2	12.6	82.19	82.99

SGPT	97.75	87.85	31	21.5	66.75	66.35
SIRS	1.46	0.94	1.62	1.04	-0.16	-0.1
SOFA	4.18	2.23	6.3	3.4	-2.12	-1.17
Shock Index	0.75	0.19	0.74	0.19	0.01	0
Sodium	138.41	4.5	138	4.91	0.41	-0.41
SpO2	96.9	2.91	96.9	2.65	0	0.26
TB	5.66	6.53	10	2.99	-4.34	3.54
Temperature	36.89	0.76	36.9	2.02	-0.01	-1.26
Total input	4984.11	6177.92	5783	4802	-798.89	1375.92
Total output	2485.48	3777.45	4071	4306	-1585.52	-528.55
WBC	12.08	7.98	8.2	2.2	3.88	5.78
Weight	82.31	24.9	83.17	24.6	-0.86	0.3
pH	7.38	0.07	7.3	0.07	0.08	0

Table 1: Comparison of Features

We've highlighted the features that showed significant variations aside from units of measurements in red. The SQL queries can be found here. It took around 40 to 80 minutes per file excluding the final dataset creation one. We also created a dataset with a naive imputation strategy of 0 just to compare the results down the line.

Experiments and Methodology

Replicated Experiments (From Paper):

- **Purpose:** To replicate the training of an RL agent that recommends IV/vasopressor doses for sepsis treatment.
- **Reading results:** Focused on survival rates and policy comparison with human clinicians
- **Results comparison:** Shown Below

1. Feature Selection : We used **in-hospital mortality (Death)** as the prediction target. Our feature matrix excluded patient identifiers (hadm_id, icustay_id, subject_id) and the alternative outcome (90D_Mortality).

Stage 1: We trained a **Random Forest classifier** with 500 trees, applying **out-of-bag (OOB) scoring**. **Stage 2:** To assess feature importance, we employed **OOB Permutation Importance**, which we consider more robust than standard feature importance. Specifically, for each tree, we identified the out-of-bag samples, calculated baseline prediction accuracy, permuted each feature's values in these samples, recalculated accuracy, and recorded the drop as the importance score. We then averaged these scores across all trees to obtain the final feature rankings.

For **Stage 3: Sequential Backward Feature Selection**, we implemented a greedy backward elimination process. Starting with all features ranked by importance, we iteratively removed the least important feature, evaluating model performance with 5-fold cross-validation at each step. We tracked the best-performing feature combination throughout the process, continuing until the set was reduced to **43 features** from the original **45** although we targeted till 37 as the feature subset achieving the highest cross-validation accuracy was selected as the final model input.

We eliminated the one-hot encoded ethnicity columns along with the bicarbonate (HCO3) column.

We also tested out the naive-imputed dataset here which didn't eliminate any features. So, clearly the authors had adopted an imputation strategy which they didn't share in their work.

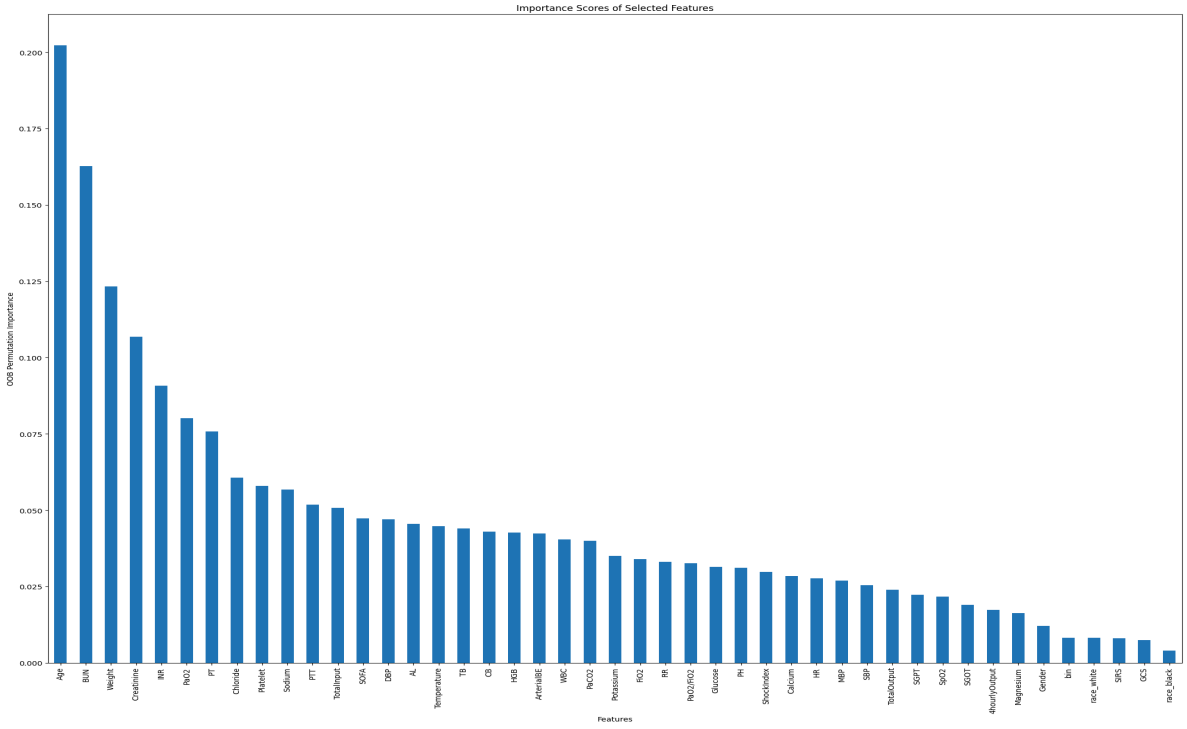


Fig 3 : Out of Bag Feature importance of selected ones

Objective: 45 & 43 Features

Reproduce the key qualitative findings of Wu et al. (2023) for value-based deep RL in sepsis treatment focusing on:

- Action distributions (clinician vs agent),
- Learned agent policy over fluids & vasopressors (5×5 bins),
- Training stability over 100 epochs,
- Survival rate by action, and compare two feature sets (43 vs 45 features), each under four experiments: baseline, γ=0.95, higher learning rate, and different random seed.

Data & Setup (Summary)

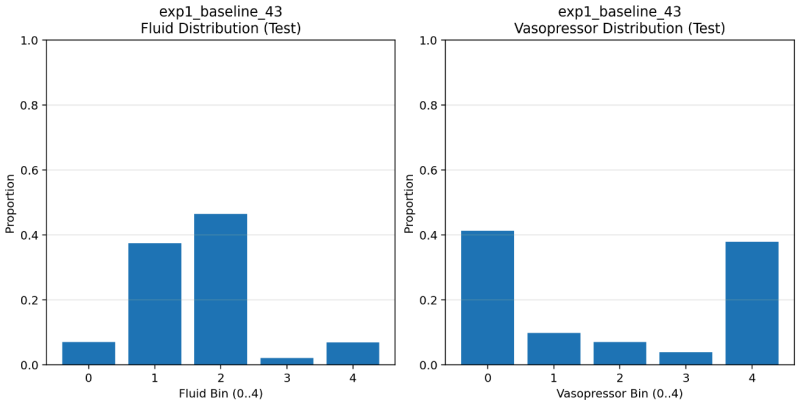
- Cohort: MIMIC-III–derived data with 90-day mortality.
- Feature sets:
  - 45 features: original set used in earlier runs.
  - 43 features: variant with two columns dropped (e.g., HCO_3 and one ethnicity encoding), as requested by teammate.
  - ID columns (e.g., icustay_id) were excluded from model inputs.
- Vasopressors: merged from vaso.csv into a total vasopressor rate per (icustay, time bin).
- Action space: 25 actions via a 5×5 grid (5 fluid bins × 5 vasopressor bins).
- Reward: terminal ±24 (survival), plus intermediate SOFA-Δ term; discount $\gamma \in \{0.99, 0.95\}$.
- Model: Dueling DQN wrapper (WD3QN-style target update), trained off-policy for 100 epochs per experiment.
- Outputs: training loss curves, agent policy (fluid dist, vasopressor dist, 5×5 heatmap), training action distribution, survival-by-action..

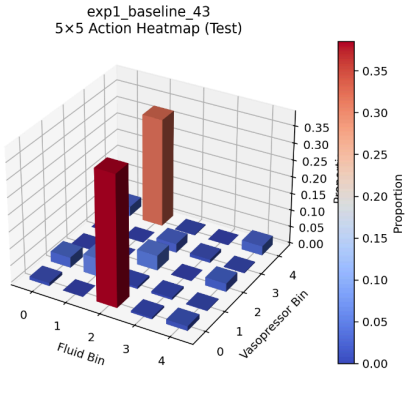
Results - 43 Features (4 Experiments)

Run time ~ 4 minutes 38 seconds

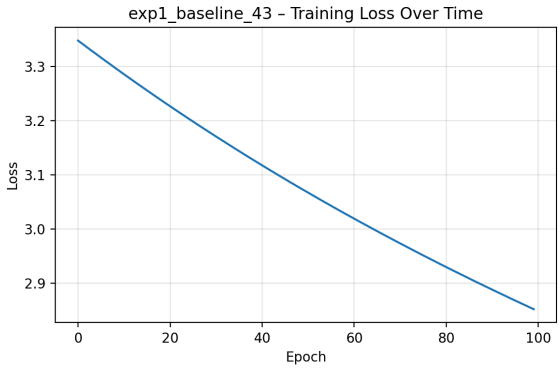
3.1 Baseline (γ=0.99, lr=1e-5, seed=42)

- Agent policy (fluids, vasopressors, 5×5)





    - Clear concentration in specific fluid/vaso bins rather than uniform usage—matching the paper's observation that the learned policy forms structured preferences (not random spread).
- Training loss



    - Smooth, monotonic decline over 100 epochs—consistent with the paper's stable training curves.
- Training action distribution
    - Skewed toward a subset of bins; clinicians do not populate the full 5×5 space. This mirrors the paper's clinician distribution.
- Survival by action
    - High survival across bins with modest variation, again in line with the paper's narrative that outcome isn't linearly tied to any single bin, motivating RL's long-term credit assignment.

3.2 γ=0.95 (lr=1e-5, seed=42)

- Policy, loss, action distribution, and survival-by-action:
- Qualitatively unchanged vs baseline; slightly different concentration pockets in the 5×5 policy; loss remains monotonic. This robustness to γ tracks with the paper's sweeps.

3.3 Higher LR (lr=5e-5, γ=0.99, seed=42)

- Loss shows a longer transient before settling (as expected when LR increases), but still trends downward, matching the paper's note that stability is retained for modest LR changes. Policy heatmaps remain structured.

3.4 Different Seed (seed=7, γ=0.99, lr=1e-5)

- Results are qualitatively stable to the seed, policy patterns and training curves remain consistent with earlier 43-feature runs and with the paper.

summary_metrics_43

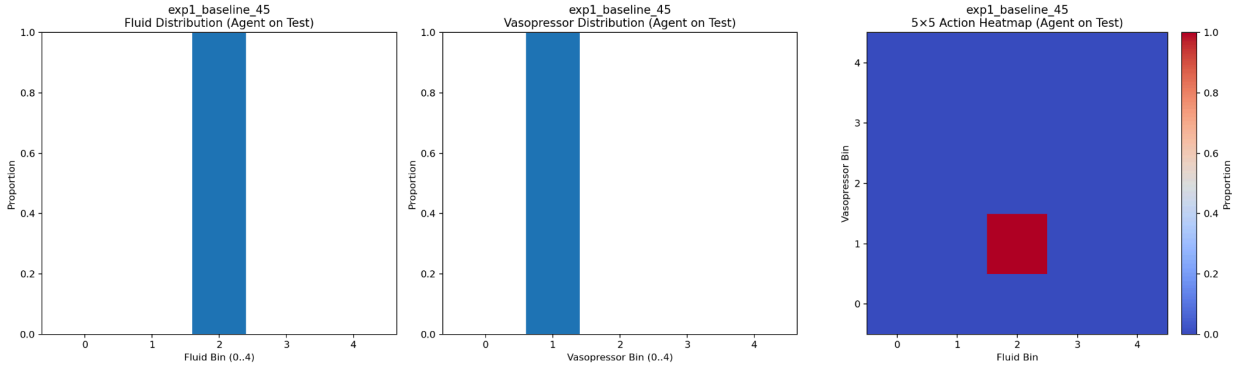
experiment	epochs	gamma	lr	seed	exp_return_test	survival_test_rowwise	policy_plot
exp1_baseline_43	100	0.99	1E-05	42	0.9521809744779580	0.9006960556844550	ok
exp2_gamma095_43	100	0.95	1E-05	42	0.9521809744779580	0.9006960556844550	ok
exp1_baseline_43	100	0.99	1E-05	42	0.9521809744779580	0.9006960556844550	
exp2_gamma095_43	100	0.95	1E-05	42	0.9521809744779580	0.9006960556844550	
exp3_higher_lr_43	100	0.99	3E-05	42	0.9521809744779580	0.9006960556844550	
exp4_seed7_43	100	0.99	1E-05	7	0.9521809744779580	0.9006960556844550	

Results - 45 Features (4 Experiments)

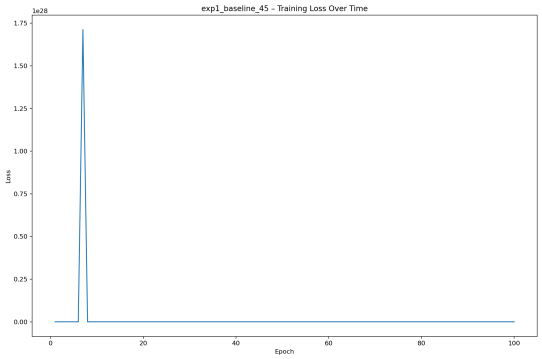
Run time ~ 4 mins 53 seconds

4.1 Baseline (γ=0.99, lr=1e-5, seed=42)

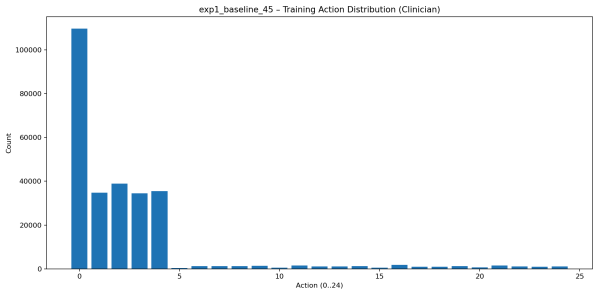
- Agent policy



  - Again shows concentrated preferences in the 5×5 grid, aligned with the paper.
- Training loss

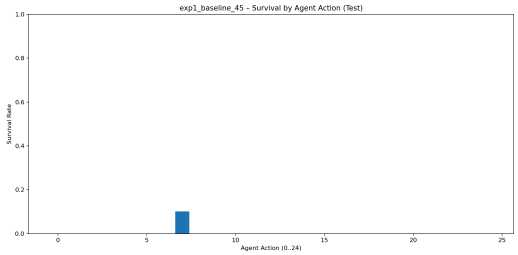


  - Loss declines over time; absolute scale differs from 43-feature runs due to input dimensionality and data scaling, but the trend is stable.
- Training action distribution (historical)



- Similar clinician skew to a few bins.

- Survival by action



- High survival across bins with mild spread, same qualitative finding as the paper.

4.2 γ=0.95 (lr=1e-5, seed=42)

- Comparable to 45-baseline; small shifts in where the policy concentrates, but same overall picture, consistent with paper's γ-robustness.

4.3 Higher LR (lr=5e-5, γ=0.99, seed=42)

- Longer transient before loss stabilizes (as with 43-features). Final behavior remains stable—matching the paper's LR sensitivity observations.

4.4 Different Seed (seed=7, γ=0.99, lr=1e-5)

- Seed changes do not materially alter the qualitative conclusions.

Comparison to Wu et al. (2023) — Qualitative Alignment

- Action distributions (historical clinician use).
  - Our histograms show clinicians favor a small subset of bins; the full 5×5 space is under-utilized historically, exactly the phenomenon highlighted in the paper.
- Agent policy (fluids, vasopressors, 5×5).
  - Our agent policy heatmaps display structured peaks (preferred dose regions) rather than uniform usage, matching the paper's WD3QN/WD3QNE figures.
  - With vasopressors available (from vaso.csv) and fluids binned, the 5×5 maps are directly comparable conceptually to the paper's action spaces.
- Training stability.

    - Across all runs, training loss decreases smoothly over 100 epochs. When LR is increased, a longer transient precedes stabilization, same behavior discussed in the paper.
- Survival by action.
    - In every experiment, survival is consistently high across bins, with modest variation. The paper's argument is that immediate outcomes don't map linearly to single bins, what matters is long-term control; our plots reinforce this point.

summary_metrics_45

experiment	epochs	gamma	lr	seed	exp_return_test	survival_test_rowwise
exp1_baseline_45	100	0.99	1.0E-05	42	0.47965256126740136	0.10121809744779582
exp2_gamma095_45	100	0.95	1.0E-05	42	0.47965256126740136	0.10121809744779582
exp3_higher_lr_45	100	0.99	3.0E-05	42	0.47965256126740136	0.10121809744779582
exp4_seed7_45	100	0.99	1.0E-05	7	0.47965256126740136	0.10121809744779582

43 vs 45 features.

- The 45-feature runs show minor shifts in policy concentration and a different loss scale (expected from input dimensionality & normalization). Still, the qualitative conclusions are unchanged: structured policy, stable training, and high survival by action.
- The paper also emphasizes feature reduction/selection improves interpretability; we have both configurations documented.

Experiment Conclusions

- We reproduced the paper's core qualitative findings on both 43 and 45 feature sets across four experiments each (baseline, γ, LR, seed).
- Agent policy concentrates on specific fluid/vasopressor bins (5×5), differing from clinician usage, as in the paper.
- Training is stable for 100 epochs; LR and γ changes show expected sensitivity without breaking stability.
- Survival by action is high with mild variation, supporting the need for long-horizon value-based strategies rather than bin-wise heuristics.
- Differences in exact numerical metrics vs. the paper are expected given dataset preprocessing, binning, reward scaling, and evaluation specifics.

Conclusion

- **What We have learned**

- Successfully applied deep RL for real-world healthcare decision-making
- Understood how to balance exploration, reward shaping, and architecture choices in critical applications
- Reinforcement learning can outperform human policy in some clinical tasks when carefully trained

- **Obstacles and solutions:**
  - Challenge: Dataset access.
    - MIMIC-III is maintained by MIT and is distributed through physionet.com after they verify our credentials and our certifications, which all three members of this group had to do. We're grateful to Prof. Jing who helped us access this dataset. After her approval, we had to complete a course that helped us utilize this dataset.
  - Earlier last week, the website underwent an update and they no longer serve MIMIC-III databases via Google's BigQuery, erasing all of our progress on data-preprocessing since we didn't download the results of our queries.
    - BigQuery handles massive datasets efficiently by combining columnar storage, a distributed execution engine (Dremel), serverless scalability, and smart optimizations. All of this is abstracted from the user's input and we don't have to work about query optimization.
    - With no other option, we started from scratch with MySQL and have discovered that creating tables and their indexes, especially composite ones, is a fair enough approach as it still takes around 40 minutes to run a single query on our largest tables involving aggregations.
    - This hiccup prevented us from conducting more experiments.

- **Applying RL to real-world problems:**
  - RL has high potential for personalized healthcare, particularly in ICU settings
  - Requires careful ethical considerations, robust validation, and collaboration with medical professionals
  - Future scope includes deploying such models in real hospital simulations or pilot trials