UNIVERSITAS AIRLANGGA | FAKULTAS SAINS & TEKNOLOGI
SI TEKNIK BIOMEDIS

Monday, November 3rd, 2025

THESIS SEMINAR

OPTIMIZATION OF SEPSIS TREATMENT USING *OFFLINE REINFORCEMENT LEARNING*–BASED *DUELING DOUBLE DEEP* Q-*NETWORK*

Fatima Hasya P. K. | NIM 082111733030

Supervisor I
Endah Purwanti S.Si., M.T.

Supervisor II
Franky Chandra Satria A, S.T., M.T.

Examiner I
Prof. Dr. Khusnul Ain, S.T., M.Si.

Examiner II
Inten Firdhausi Wardhani, S.T., M.Sc.

SMART UNIVERSITY | TOP #287 QS WORLD UNIVERSITY RANKINGS | ZONA INTEGRITAS menuju wilayah bebas dari korupsi

Background

Sumber: Government of Western Australia, 2024, *"Mum's Mission to Highlight Early Warning Signs After "Surviving Deadly Sepsis Infection"*

Sumber: WHO, 2024, *"Sepsis"*

Sumber: 410Medical, 2020, *"Case Write-Up: 8-Year-Old Male with Sepsis and Pneumonia Resuscitated with LifeFlow"*

Sumber: The Independent, 2023, *" 'I Thought It Was a Bad Cold' – I Ended in Intensive Care with Sepsis"*

Sepsis

- **Sepsis Prevalence**

- **48,9 million** cases/year
- Mortality up to 11 million (**≈ 20% of global deaths**)

- **Prevalence: 30,29%**
- **Mortality** rate: **22,5 - 52%**

- Accounts for **40,1%** of all **ICU patients**
  - **25,2% sepsis** (without shock)
  - **15,0%** present **with shock**
- Mortality in **septic shock** is **significantly higher** (68,3% vs 37,7%)

- **Progressivity: Sepsis → Septic Shock**

Septic shock: sepsis + tissue perfusion abnormalities and persistent hypotension despite adequate fluid resuscitation

30,5% sepsis ↓ septic shock

62,9% were already in septic shock upon ICU admission

8,9% deteriorated during treatment

Bottomline :

- **Septic shock** is a progressive consequence that often **becomes unavoidable** when sepsis is **not optimally managed**
- **Early and appropriate management** of sepsis is essential to reduce mortality and improve patient prognosis

Treatment Protocols

***One-Hour Sepsis Bundle* in *"Surviving Sepsis Campaign* (SSC)"**

Intravenous fluids (IV *fluids*)
30 mL/kg BW crystalloid

Vasopressor
0,01 - 0,03 µg/kg BW/minute norepinephrine

MAP < 65 mmHg

Limitations :

Too general

- SSC alone ≠ significant impact
- Heavily depends on clinical judgment and clinicians' experience

Risks of inaccurate dosing

- IV fluids, too little → hypoperfusion; too much → organ edema
- Vasopressor, too little → hypotension; too much → organ ischemia

A personalized therapeutic strategy with precise hemodynamic monitoring is needed to determine the optimal fluid and vasopressor doses as a complement to SSC in order to improve therapeutic prognosis

Previous Studies

1 Komorowski et al. (2018)

- **Agent**: Tabular Q
- ***State* (S)**: 750 cluster
- (k-means++ *clustering* of clinical features every 4 hours)
- ***Action* (A)**: 25 *actions*, combinations of IV fluid and vasopressor doses
- ***Reward* (R)**: 90-day mortality; +100 (*survive*) and -100 (*not-survive*)
- **Evaluation:** *bootstrapped* **WIS**
- Results: agent policy > clinician's (LB 95% CI > UB 95% CI; IQR 85,1 - 86,0)
- Limitations: *clustered states* → low interpretability, potential overestimation of Q

2 Zhang et al. (2024)

- **Agent**: *Dueling* DDQN
- ***State* (S)**: continuous vector of 46 x 1 clinical features (2 hrs)
- ***Action* (A)**: 25 *actions*, combinations of IV fluid and vasopressor doses
- ***Reward* (R)**: change in SOFA score and lactate levels
- **Evaluasi: WDR,** mortality assessment framework (via SARSA)
- Results: agent policy > clinician's (WDR 12,35 vs. 8,78; estimated mortality decreased by 3%)
- Limitations: transferability has not been confirmed

Wu et al. (2023)

- **Agent**: WD3QNE (Dueling DDQN variant with added *Weighted-Human Expertise*)
- **S:** *windowing* every 4 hrs
- NN: 2 *Hidden layer of* 256 and 128 units, ReLU

Tu et al. (2025)

- **Agent**: CQL (Tabular Q variant with *Conservative function*)
- **S:** continuous vector of 48 x 1 clinical features (4 hrs)
- **Evaluation: FQE**

Bottomline :

	K	Z	W	T
Preprocessing MIMIC-III	●	●		
Agen Architecture		●	●	
MDP Framework	●	●	●	●
Metriks Evaluasi	●	●		●

Problem Statements

Design of a clinical decision-making model (agent) using Dueling DDQN for sepsis management

Evaluation of the Dueling DDQN agent's performance in producing clinical policies

Research Purposes

Designing a clinical decision-making model (agent) using Dueling DDQN

Evaluating the performance of the Dueling DDQN agent in generating clinical policies

Research Benefits

1. An alternative clinical decision-making solution based on retrospective data
2. Supporting efforts to improve patient safety
3. Providing a scientific contribution to the application of RL in the clinical domain
4. Supporting Tri Dharma of Higher Education through tech innovation

Problem Scope

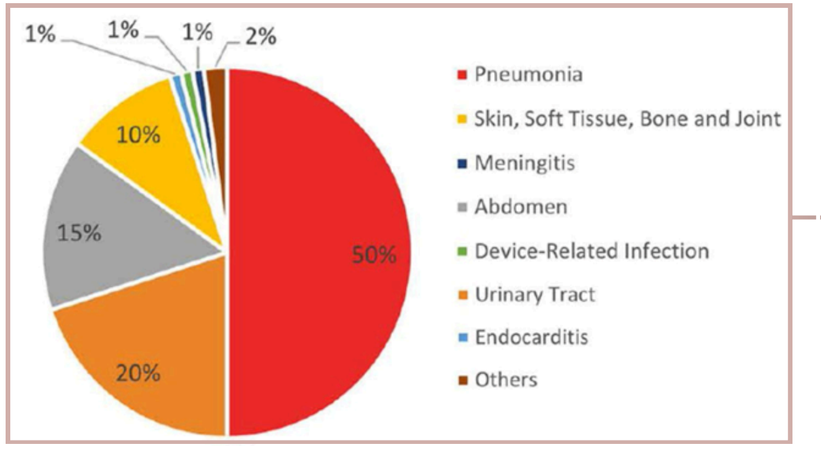
1.	**Dataset**	MIMIC-III; sepsis patients identified using Sepsis-3 criterias
2.	**Data Variability**	Limited to the population and hospital setting represented in MIMIC-III
3.	**Data Extraction**	Relevant subsets obtained using SQL queries
4.	**Vasopressors**	Norepinephrine, epinephrine, dopamine, phenylephrine, and vasopressin; converted to NEE
5.	**IV *fluids***	Includes crystalloid and colloid fluids
6.	**Data Processing**	Training and evaluation performed using Python in Google Colab
7.	**Testing**	Based on historical data and statistical evaluation, without any direct clinical trial

Sepsis

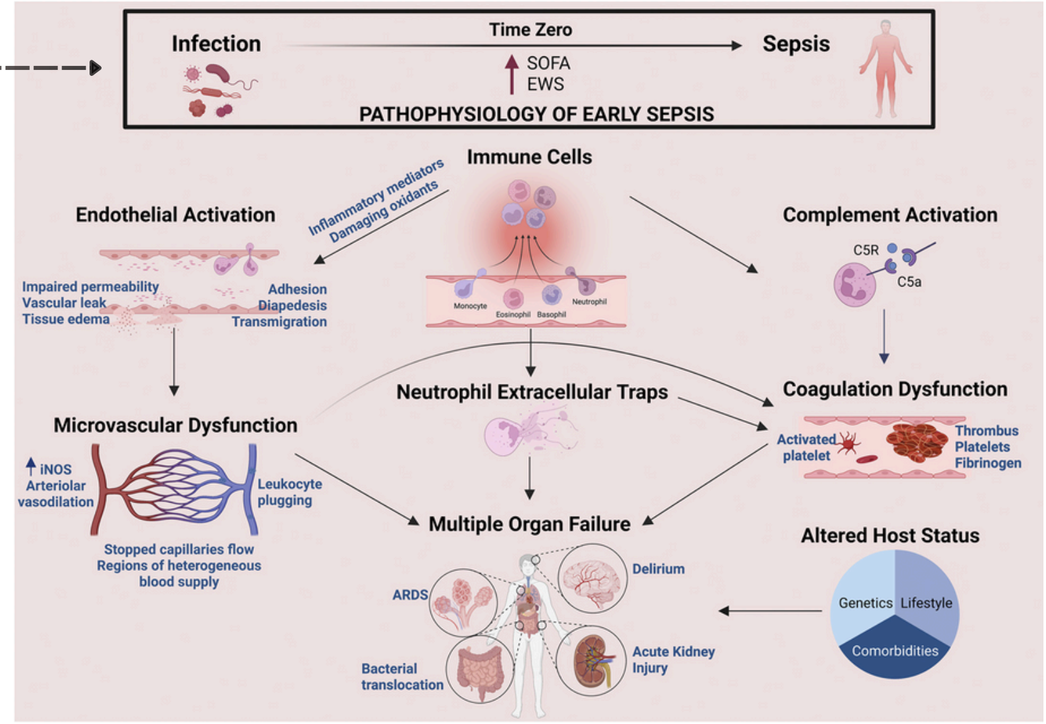
Definition & Epidemiology

Sepsis (Sepsis-3, 2016)	**Life-threatening multiple organ dysfunction** due to uncontrolled dysregulation of the immune response to **infection**
Multiple organ dysfunction	*Sequential Organ Function Assessment* **(SOFA)** → score, quantitative *Quick Sequential Organ Function Assessment* **(qSOFA)** → triage
Septic Shock	Sepsis accompanied by refractory hypotension (MAP <65 mmHg, requiring **vasopressors**) and **hyperlactatemia >2 mmol/L**
Major infection sources	Lungs (pneumonia), abdomen (intra-abdominal infection), urinary tract, skin & soft tissue
Elderly; patients with chronic diseases	often triggered by **urinary tract infections**
Immunocompromised patients; post-surgery	vulnerable to nosocomial infections (e.g., **catheter-associated sepsis, ventilator-associated pneumonia**)

The variation in infections causing sepsis indicates that **sepsis is not a single entity**, but rather a **complex clinical syndrome** → diagnosis and treatment **needs to consider the organs involved,** clinical presentation, and the severity of systemic dysfunction

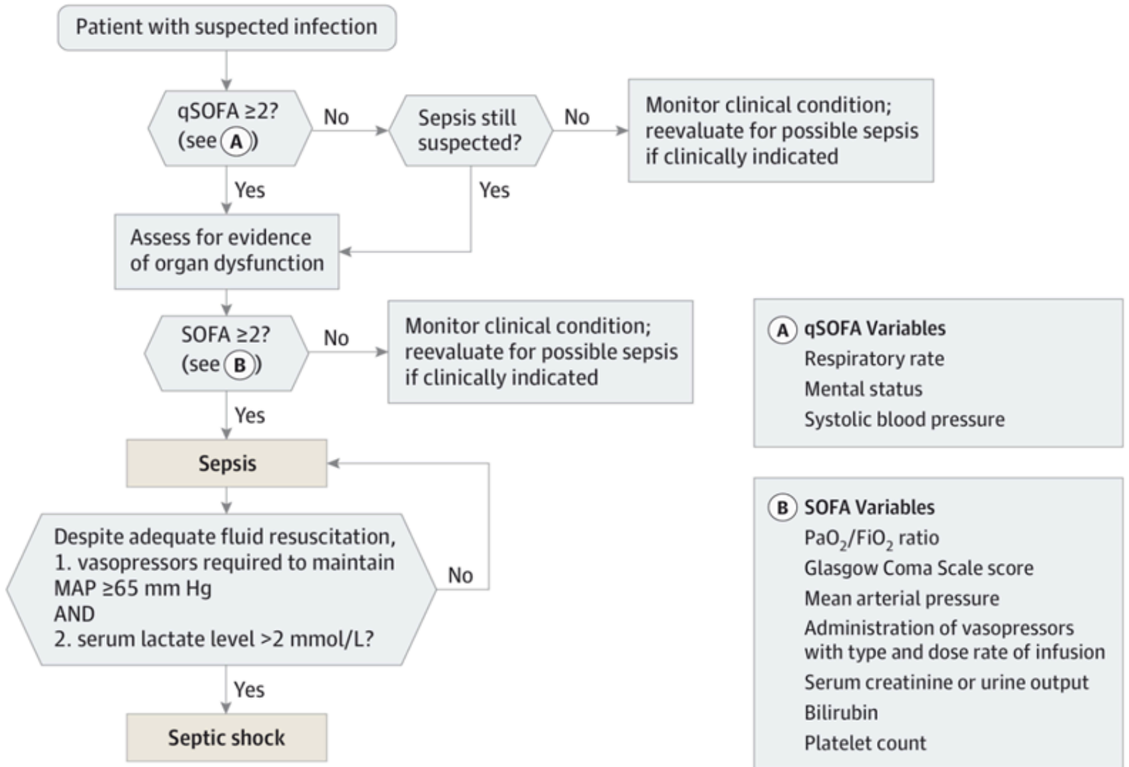


Pathophysiology



1.	Local infection triggers PAMPs recognition by PRRs → innate immune activation
2.	Release of inflammatory mediators (TNF-α, IL-1β, IL-6) → immune cell recruitment
3.	Inflammation → endothelial activation, ↑ capillary permeability, vascular leakage, tissue edema
4.	Microvascular dysfunction: arteriole vasodilation, capillary flow obstruction, tissue hypoperfusion
5.	Complement activation (C5a) + NETs formation → worsening of inflammation, coagulation disorder
6.	Resulting in multiorgan dysfunction → ARDS, AKI, delirium, gut bacterial translocation

Diagnosis & Supporting Examination



qSOFA is used in the Emergency Department for rapid screening of suspected sepsis using 3 parameters easily observed during **the triage process** to represent SOFA

qSOFA score ≥ 2 → patient is considered high-risk for sepsis → **requires further evaluation using SOFA**

SOFA score ≥ 2 → **indicates organ dysfunction** → supports sepsis diagnosis according to Sepsis-3

In the initial phase of sepsis evaluation, several supportive examinations are performed to review the severity of organ dysfunction and to search for the source of infection, including:

Blood Specimen

- **Used for CBC; coagulation analysis and clinical chemistry**
- **Aids in diagnosis; monitoring severity & therapy**

- Findings in sepsis patients :
  - Leukocytes ↑ / ↓ (inflammation)
  - Creatinine/BUN ↑ (renal)
  - Bilirubin ↑ (hepatic)
  - Lactate ↑ (hypoperfusion)
  - Hyperglycemia (metabolic stress)
  - Albumin ↓, abnormal K^+ (nutrition & electrolytes)

Blood Culture

- **Detection of microorganisms (bacteria, fungi, viruses); identification of pathogens + antibiotic susceptibility testing**

- Performed before empirical antibiotic administration (SSC guideline)
- Detection success rate :
  - 50–57% before;
  - Drops to 19–28% after antibiotics

Arterial Blood Gas

- **Arterial blood test to assess the patient's oxygenation, ventilation, and acid-base status**

- Findings in sepsis patients :
  - pH < 7.35 (acidosis)
  - $PaCO_2$ > 45 (respiratory acidosis); < 35 (respiratory alkalosis)
  - HCO_3^- < 22 (metabolic acidosis)
  - PaO_2 < 80 (hypoxemia)
  - SaO_2 < 90% (oxygenation impairment)
  - Base Excess negatif (metabolic acidosis)

Respiratory Sample

- **An important examination performed on sepsis patients with suspected lung infection**

- Findings in sepsis patients :
  - Horowitz ratio decrease (PaO_2/FiO_2 < 300 mmHg) → severe pneumonia/ARDS
  - The Horowitz ratio is also used in the respiratory dysfunction score in sepsis
  - The use of a ventilator can cause *Ventilator-associated Pneumonia* (VAP)

Common Handling Implementation Order

***One-Hour Sepsis Bundle* in "*Surviving Sepsis Campaign* (SSC)"**

1. Measure blood lactate level, repeat if > 2 mmol/L
2. Obtain blood cultures before administering antibiotics
3. Administer broad-spectrum antibiotics immediately
4. Begin rapid administration of **30 mL/kg BW crystalloid** for hypotension or lactate level > 4 mmol/L
5. **Apply vasopressors** if hypotension persists, aiming for a target MAP > 65 mmHg

Intravenous Fluid Resuscitation (IV fluids)

Types of fluids are divided into two groups:

- **Crystalloids**→ small-molecule solutions, inexpensive, readily available, rapidly increase intravascular volume
- **Colloids** → large molecules, faster intravascular volume expansion, more expensive, and have risks of certain side effects

The Surviving Sepsis Campaign (SSC) **recommends crystalloids** as the primary choice for initial resuscitation due to a better safety profile and wide accessibility

Colloids (e.g., Human Albumin Solution / HAS) may be considered if crystalloid resuscitation **is insufficient** or in specific conditions

Vasopressor Management Recommendations

!! Use norepinephrine as first-line vasopressor

For patients with septic shock on vasopressor

!! Target a MAP of 65mm Hg

! Consider invasive monitoring of arterial blood pressure

If central access is not yet available

! Consider initiating vasopressors peripherally*

If MAP is inadequate despite low-to-moderate-dose norepinephrine

! Consider adding vasopressin

If cardiac dysfunction with persistent hypoperfusion is present despite adequate volume status and blood pressure

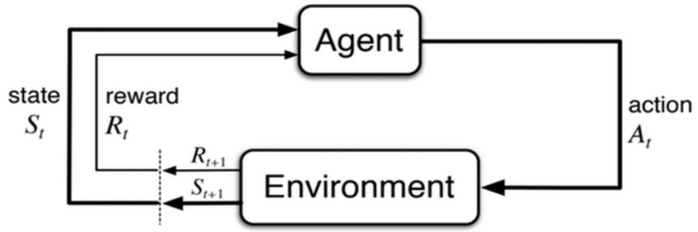
! Consider adding dobutamine or switching to epinephrine

!! Strong recommendations ! Weak recommendations

*When using vasopressors peripherally, they should be administered only for a short period of time and in a vein proximal to the antecubital fossa.

Dueling DDQN

Markov Decision Process (MDP)



- **MDP** → a mathematical framework of Reinforcement Learning that formalizes the agent's and environment's interaction at each discrete time-step t = 0, 1, 2, ... in the following sequence :

$$S_0,\ A_0,\ R_1,\ S_1,\ A_1,\ R_2,\ S_2,\ A_2,\ R_3,\ \ldots$$

- **Policy** is agent's strategy for choosing actions based on the state, defined as :

$$\pi(a \mid s) = P[A_t = a \mid S_t = s]$$

- **Value** is the expected accumulated long-term reward (return) when the agent follows a specific policy.
  - **Return** is defined as :

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$$

  - **Action-value function (Q)** → the expected return from a specific action in a specific state :

$$q_\pi(s,a) = \mathbb{E}_\pi\left[G_t \mid S_t = s,\ A_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s,\ A_t = a\right]$$

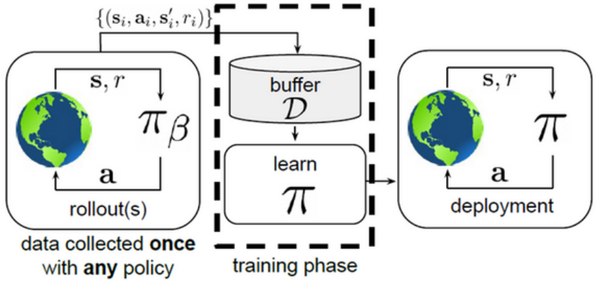
- in the form of the **recursive Bellman equation :**

$$q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \sum_{a'} \pi(a' \mid s')\, q_\pi(s',a')\right]$$

Legends:

$q_\pi(s,a)$	:	the expected long-term return from taking action a in state s while following policy π
γ	:	the discount factor that determines how much future rewards are taken into account (0 < γ ≤ 1)
R_{t+k+1}	:	the reward received at time t + k + 1
$p(s',r \mid s,a)$	:	the probability of transitioning from state s to state s′ and receiving reward r, after the agent takes action a in state s
$\pi(a' \mid s')$	:	the probability that the agent will choose action a′ in state s′ according to policy π

Offline Reinforcement Learning



- Offline RL → trains an agent entirely **from historical data** without direct interaction with the environment
- The agent learns from stored transitions (s,a,r,s′) in a replay buffer (**memory bank**) → does not require real-time interaction → suitable for clinical domains
- To mitigate the risk of learned policy deviating from the available historical actions → **Kullback-Leibler (KL) regularization**, controlled by coefficient beta_KL, as:

$$L_{\text{Total}}(\theta) = L(\theta) + \beta_{\mathbf{KL}} \cdot L_{KL}$$
$$L_{KL} = \mathbb{E}_{s\sim\mathcal{D}}\left[\text{KL}(\pi(\cdot|s)||\pi_b(\cdot|s))\right]$$

Q-Learning

- Q-Learning is an off-policy and model-free algorithm that learns the **optimal action values** through **sampling historical data** without requiring a transition model
- The value is updated iteratively using experience tuples:

$$(s_t, a_t, r_{t+1}, s_{t+1})$$

according to the following Q-Learning equation:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

α : the learning rate, which determines how large the update is relative to the previous Q-value ($0 < \alpha \leq 1$)

Dueling + Double DQN

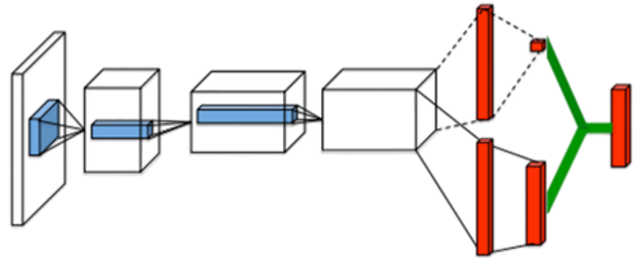
Dueling Double Deep Q-Network (**Dueling DDQN**) → **Dueling + Double DQN**

Double DQN

- *Double DQN reduces overestimation bias by using two identical networks*:
  - ***Main network*** (parameter θ) → selects the optimal action (a*)
  - ***Target network*** (parameter θ') → evaluates the value of a* as :

$$Y_t^{DoubleQ} = R_{t+1} + \gamma\, Q\left(S_{t+1}, \arg\max_a Q(S_{t+1}, a; \theta)\ ;\ \theta^-\right)$$

Dueling DQN



- *Dueling* DQN separates the network into two streams:
  - ***Value stream*** → *Value* V(s), estimates how good a state is
  - ***Advantage stream*** → *Advantage* A(s, a), estimates how good a specific action is when taken in a given state

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

$|\mathcal{A}|$: the number of possible actions

MDP in Sepsis Treatment

State (S)

Patient physiological conditions in *window time-step* every 4 hrs

- Demographics
- Lab values
- Vital signs
- Fluids
- Ventilation

→ ***vector 48 x 1***

Action (A)

Combination of IV fluid and vasopressor

- IV *fluids* (4 hrs) — **5 bins**
- *Max.* Vasopressor (4 hrs) — **5 bins**

→ ***25 actions***

Reward (R)

Evaluates the impact of each dosage on the patient's condition every 4 hrs

- **Intermediate reward :**

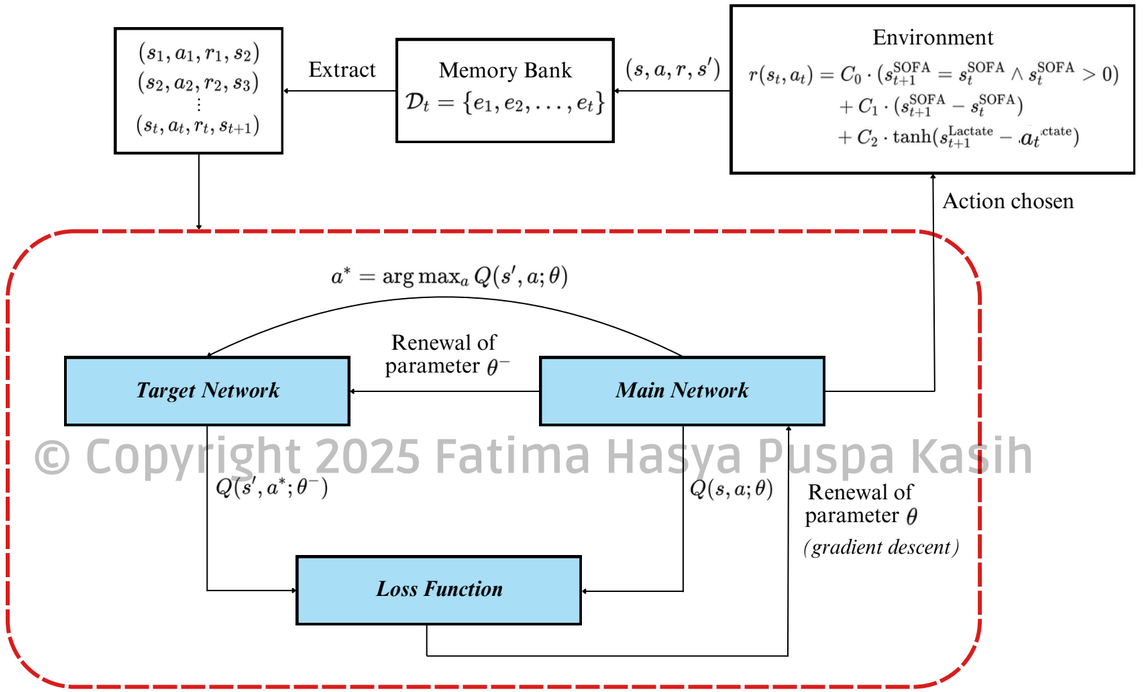
$$r(s_t, a_t) = C_0 \cdot (s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}} \wedge s_t^{\text{SOFA}} > 0)$$ → stagnation penalty

$$+ C_1 \cdot (s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}})$$ → SOFA *reward*/penalty

$$+C_2 \cdot \tanh\left(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}}\right)$$ → Lactate *reward*/penalty

- **Terminal reward :**
  - Survivor = +15
  - Non-survivor = -15

Dueling DDQN Agent in Sepsis Treatment



- The MDP components are constructed in a **memory bank** $\mathcal{D}_t = \{e_1, e_2, \ldots, e_n\}$ containing clinical transition tuples of the patient $e_t = (s_t, a_t, r_t, s_{t+1})$
- The target network is used to evaluate a⋆ from the main network by calculating the loss at each iteration i according to the following function:

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma\, Q\left(s',\, a^*; \theta_i^-\right) - Q\left(s, a; \theta_i\right)\right)^2\right]$$

- The obtained loss is then regularized with:
  - a penalty if Q value > Q threshold
  - a penalty if the change in vasopressor dosage > 0.5 µg/kg/min

$$L(\theta) = L_i(\theta_i) + \max(Q(s, a; \theta) - Q_{\text{thresh}},\, 0) + \max(V_{\text{change}} - 0.5,\, 0)$$

Off-Policy Evaluation (OPE)

Weighted Importance Sampling (WIS)

Attributable Success

Estimates the value of a new policy by **re-weighting** past clinical outcomes based on how **often** the new policy would have chosen the actions in the data

- **WIS** estimates the value of the target policy π1 by assigning importance weights to historical rewards based on the action-probability ratio between π1 and π0 :

$$\rho_t = \frac{\pi_1(a_t|s_t)}{\pi_0(a_t|s_t)}$$

- The WIS estimate for each trajectory H is **:**

$$V_{\text{WIS}}^{(i)} = \frac{\rho_{1:H}^{(i)}}{w_H} \left(\sum_{t=1}^{H} \gamma^{t-1} r_t \right)$$

- The final WIS value is obtained by averaging over all trajectories in D

$$V_{\text{WIS}} = \frac{1}{|D|} \sum_{i=1}^{|D|} V_{\text{WIS}}^{(i)}$$

Fitted Q Evaluation (FQE)

Intrinsic Value Potential

Estimates the value of a new policy by building a **direct prediction** model (Q-function) from the data

- **FQE** estimates the value of the target policy through **iterative regression on the Q-function**
- Direct method → relies on historical transitions but does not reference the clinician policy's action probabilities
- At each iteration, the target value is computed as :

$$y_i = r_i + \gamma\, Q_{k-1}(s_i',\, \pi(s_i'))$$

- The target yi is then used to train the Q-function Qk until convergence
- The final FQE estimate of the target policy value is calculated as :

$$V^{\text{FQE}} = \frac{1}{N} \sum_{i=1}^{N} Q_K(s_i, \pi(s_i))$$

Where :

- π(si') : the action selected by the target policy at next state si'
- Q(k-1) : the Q-function from iteration k-1
- QK : the final Q-function after the iterative FQE process
- N : the total number of transitions (states) in the historical dataset

Weighted Doubly Robust (WDR)

Verified Robustness

Estimates the value by **combining** the prediction (FQE) and the re-weighting (WIS)

- **WDR** combines Importance Sampling (IS) and the Direct Method (DM) to estimate the value of the target policy
- Each transition is assigned an importance weight based on the ratio of the target policy's action probability (πe) to that of the clinician policy (πb)
- The estimate is **bias-corrected using the predicted Q-value** of the **target policy** (as a control variate)
- The cumulative importance ratio is computed as :

$$\rho_t^i = \prod_{t'=0}^{t} \frac{\pi_e(A_{t'}^{H_i} \mid S_{t'}^{H_i})}{\pi_b(A_{t'}^{H_i} \mid S_{t'}^{H_i})}$$

- The weights are then normalized :

$$w_t^i = \frac{\rho_t^i}{\sum_{j=1}^{n} \rho_t^j}$$

- **The final WDR value estimate is given by :**

$$\text{WDR}(D) := \sum_{i=1}^{I} \sum_{t=0}^{T} \gamma^t\, w_t^i\, r_t^{H_i} + \sum_{i=1}^{I} \sum_{t=0}^{T} \gamma^t \left(w_t^i\, \hat{Q}_{\pi_e}(s_t^{H_i}, a_t^{H_i}) - w_{t-1}^i\, \hat{V}_{\pi_e}(s_t^{H_i}) \right)$$

Research Methods

Time & Place

Research Place

- Laboratorium Teknik Biomedis Universitas Airlangga
- Online, WFA

Research Time

3 Months,
Agustus - November

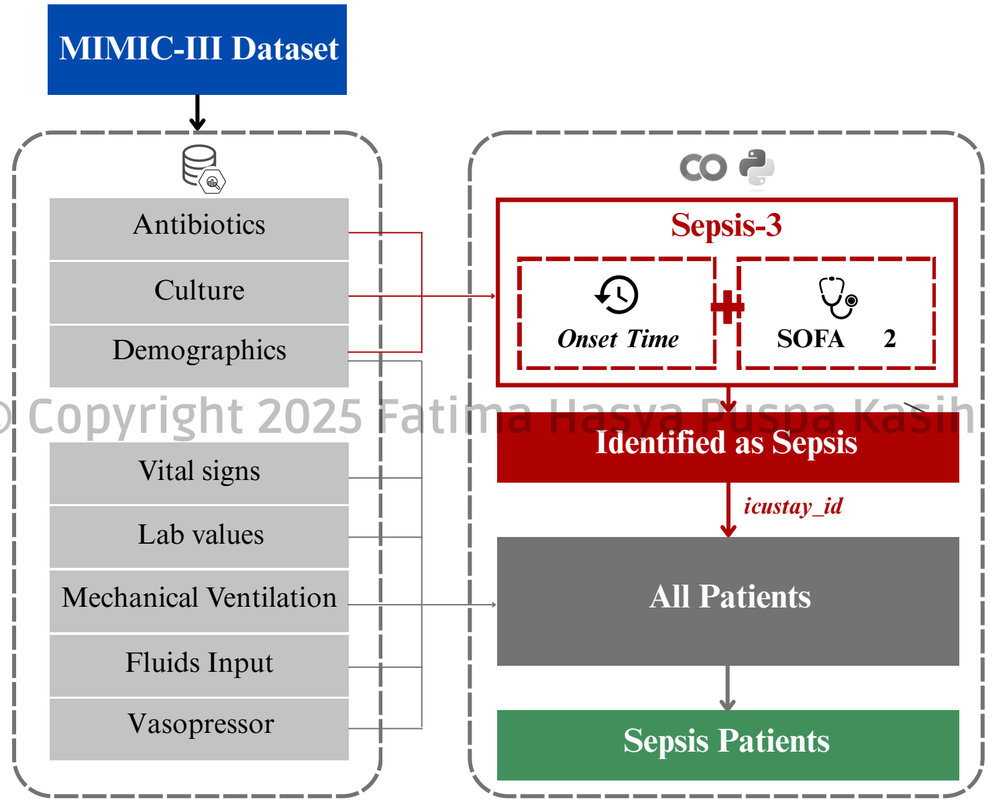
Tools & Materials

- **Tools**

- **Material**

MIMIC

Initial Data Extraction

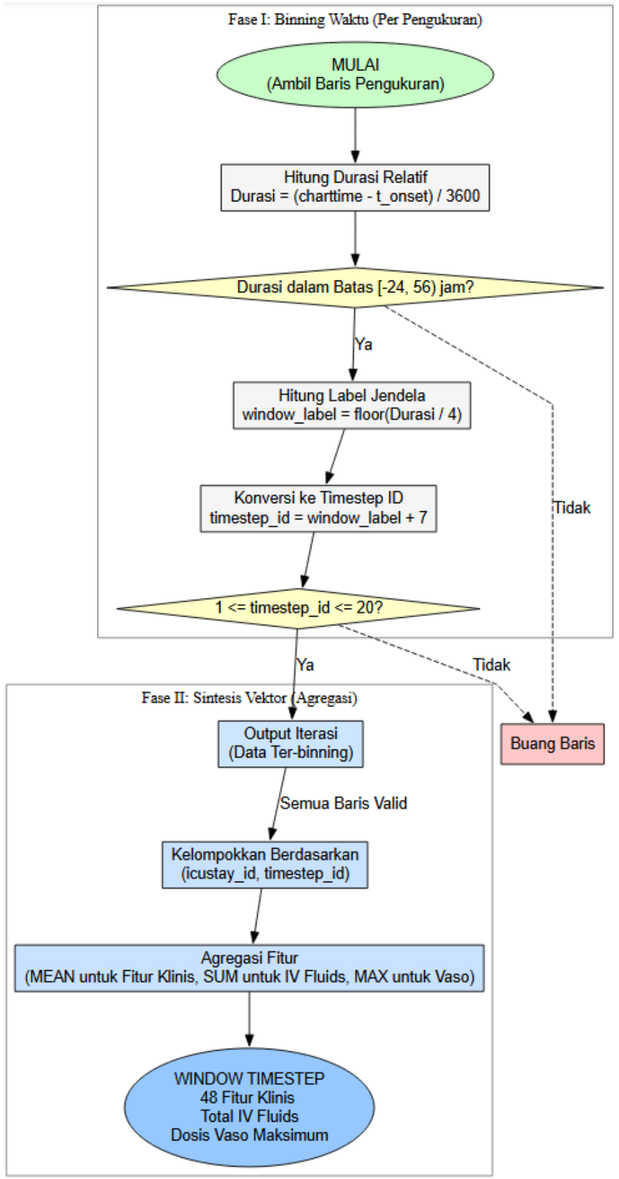


$$\text{Onset Time} = \begin{cases} t_{ABx}, & \text{if } t_{ABx} < t_{Culture} \text{ and } |t_{ABx} - t_{Culture}| \leq 24 \text{ hours}, \\ t_{Culture}, & \text{if } t_{Culture} < t_{ABx} \text{ and } |t_{ABx} - t_{Culture}| \leq 72 \text{ hours}. \end{cases}$$

Literature Review and Data Access Handling

Data Preprocessing

Windowing Time-Step every 4 hours



1. Compute the relative time difference between each charttime and the onset timestamp → **retain only data within the -24 to +56 hour interval (80-hours)**

2. Assign measurements to 4-hour windows using floor-division and **map them to timestep_id (1–20)**

3. **Group** measurements by patient and time window using the **key pair (icustay_id, timestep_id**) to create consistent state representations across patients

4. **Aggregate features**: mean for vitals/labs, sum for IV fluids, max for vasopressors

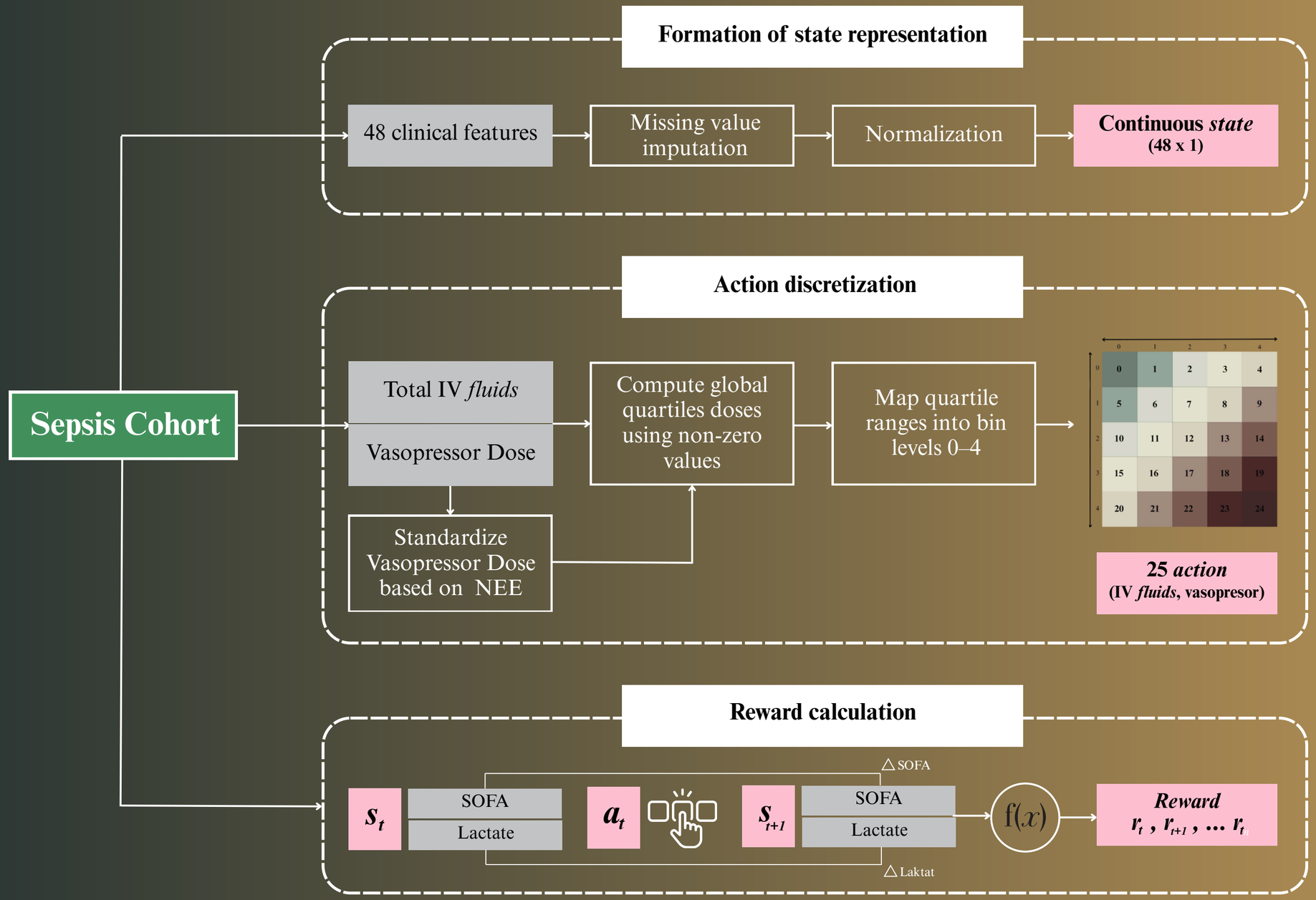
Finalization of Sepsis Cohort

- The final cohort was determined by applying several exclusion criteria adapted from previous studies to ensure that the data represented clinically valid and relevant treatment episodes for sepsis policy learning

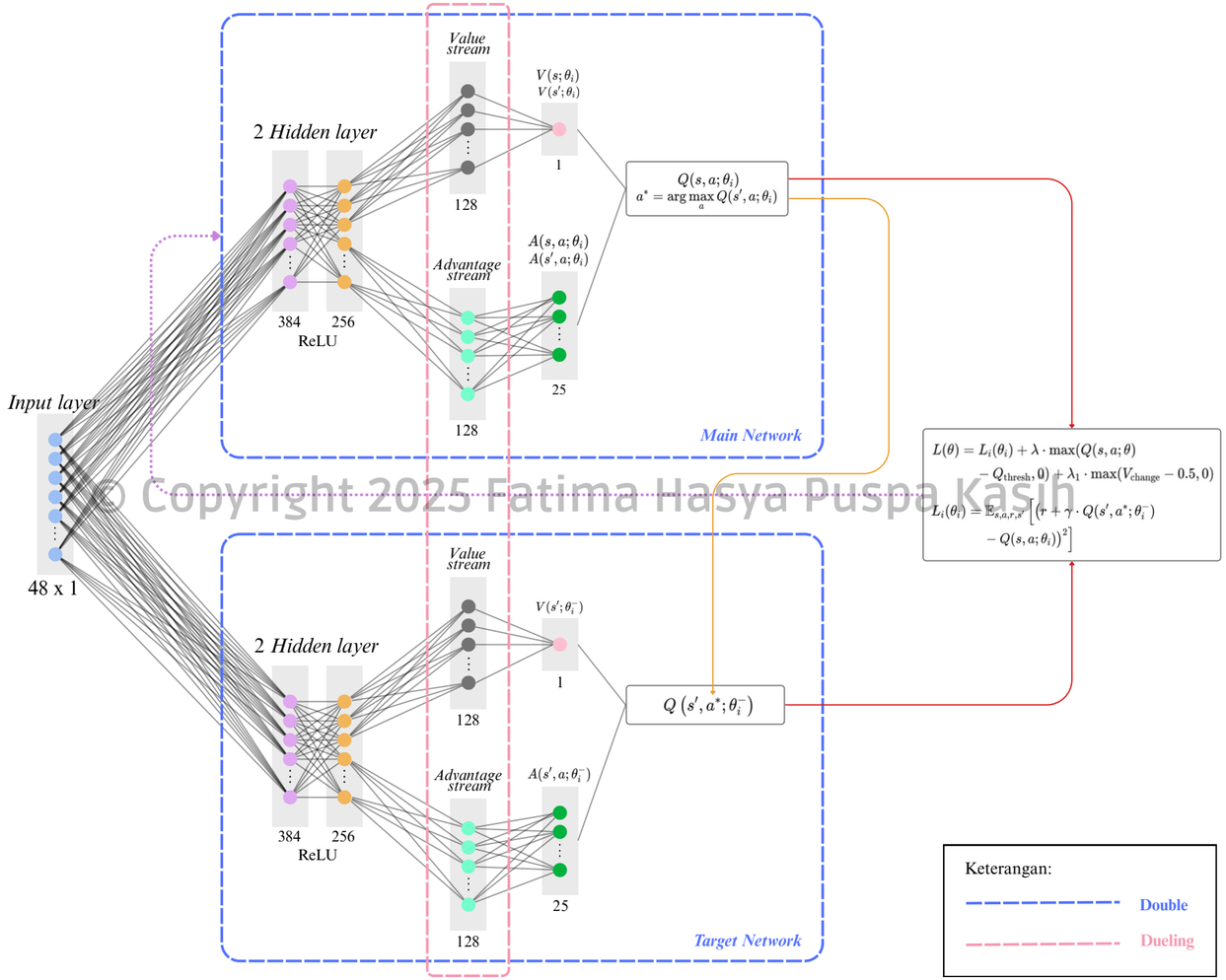
- **Exclusion criterias :**

Adults patients only (≥18 years old)	**Pediatric** sepsis was excluded because its physiology, clinical presentation, and therapeutic management **differ fundamentally from adults**
	Sepsis-3/SOFA frameworks are validated only in **adult populations**
Missing mortality records	Patients **without documented mortality status were removed** to avoid bias in calculating terminal rewards and evaluating treatment outcomes
Treatment withdrawal cases	Excluded when no clinical data were recorded for ≥24 hours after sepsis onset and mortality occurred within ≤24 hours (<6 time windows), indicating termination of therapy. This exclusion is due to prevent confusion in agent training process.

Construction of Memory Bank



Development of Dueling DDQN Agent



Training and Validation of Dueling DDQN Agent

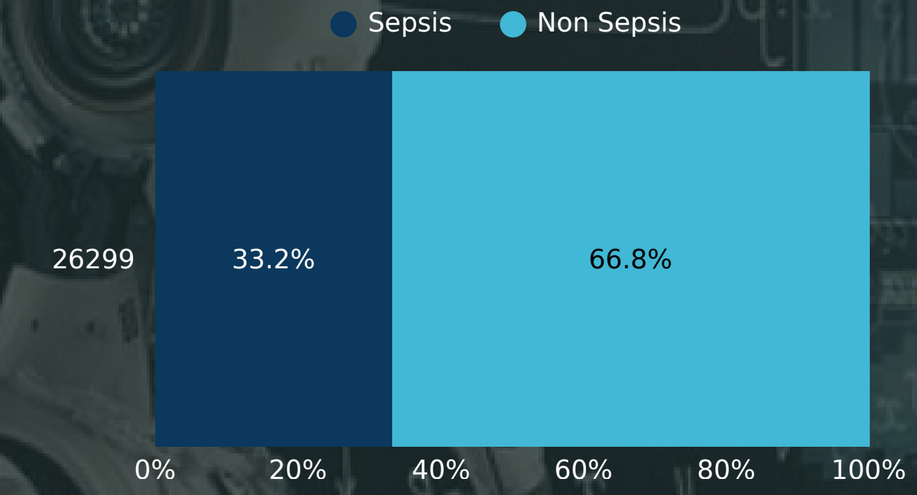
- The dataset is split into **80% training, 10% validation, and 10% testing**
- The validation set is used to select the most stable agent variant by comparing expected returns, OPE metrics, loss stability, and progress consistency of Q-value estimation
- Model selection involves tuning the KL-regularization parameter β_KL using values 0.005, 0.001, 0.00075, 0.0005, and 0.00025
- Training and validation are performed for 100 epochs and batch size 128 with:
  - Python 3.8
  - PyTorch
  - Adam optimizer
  - Learning rate (α) = 0.0001
  - Discount factor (γ) = 0.97
- The main network is updated via backpropagation
- The target network uses soft updates → tau = 0.005

Testing and Performance Analysis

- The Dueling DDQN agent is evaluated using the best hyperparameter configuration identified using validation set
- Agent performance is assessed by comparing its expected return against the clinician policy using three Off-Policy Evaluation (OPE) methods: WIS, FQE, and WDR
- To assess the contribution of each state feature to the agent's action selection → Permutation Feature Importance (PFI)
- PFI measures the drop in model performance when a feature is randomly permuted → features causing larger performance degradation are deemed more influential

Results & Discussions

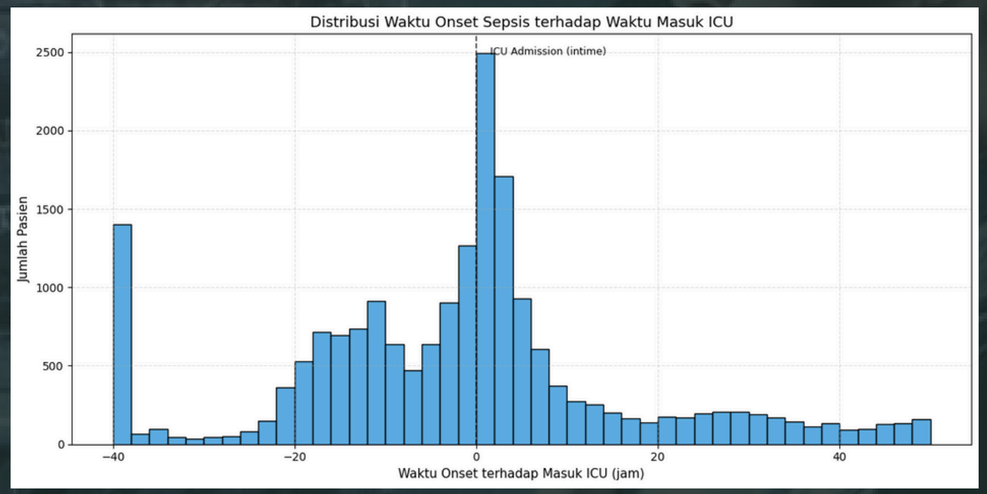
Extracted Sepsis Population



From a total of 52,953 adult ICU patients, **26,299 patients (33.2%)** were identified as having **sepsis** based on Sepsis-3 criteria

Approximately one-third of the ICU population exhibited infection accompanied by organ dysfunction, forming a clinically meaningful cohort for learning sepsis treatment policies

Histogram of Onset Sepsis Distribution



Most patients experienced onset within the first **2–4 hours** after ICU admission

Second peak of onset was observed in the **3–4-hour range**, followrd by additional concentration occurring approximately 2 hours before ICU admission

Implication → sepsis manifestations are already present in the pre-ICU phase, but confirmation occurs only after initial diagnostic workup or early interventions in the ICU

Data Preprocessing

- From 26,299 sepsis patients, a total of **24,948 patients had sufficient observational data within the -24 to +56 hour interval (80-hours)**
- The windowing process produced **448,521 window time-steps**, with an average of 9.9 (SD 5.6) windows per patient and a minimum of 1 window
- After applying **exclusion criteria** → a final sepsis cohort of **18,830 patients** was obtained
- A total of **364,563 window time-steps** were generated for these patients

Kategori	Fitur	(Mean, SD)	Fitur	(Mean, SD)
Demografi	Jumlah Pasien	18830	% *Survivor*	73,6
	Usia	65,16 (16,27)	Skor *Elixhauser*	4,06 (2,15)
	% Laki-Laki	55,70	% Re-admisi ke ICU	33,93
	Berat Badan	76,58 (19,92)		
Tanda Vital	SOFA	7,49 (3,15)	Indeks Syok	0,745 (0,18)
	SIRS	1,93 (1,04)	SpO2	96,93 (2,38)
	GCS	12,48 (3,44)	Suhu Tubuh	36,5 (1,13)
	SysBP	121,29 (21,2)	HR	87,76 (15,9)
	DiaBP	56,29 (12,7)	RR	21 (5,58)
	MeanBP	77,85 (12,93)		
Ventilasi	% *MechVent*	56,64	FiO2	0,41 (0,09)
Cairan Tubuh	Output Urin	2018		
Input Cairan	Total IV *Fluids*	2753		
Vasopressor	Dosis VP (*max*)	0,035 (0,51)		

Hasil Laboratorium	Potasium	4,09 (0,57)	Kalsium	8,3 (0,7)
	Sodium	138,68 (5)	Kalsium Terionisasi	1,1 (0,07)
	Klorida	104,84 (6,7)		
	Glukosa	133,2 (40)	SGOT	46,7 (35,5)
	Kreatinin	1,34 (0,88)	SGPT	35,28 (26)
	Magnesium	2,01 (0,33)	BUN	31,36 (20,3)
	Albumin	2,85 (0,63)	Bilirubin	2,69 (4,60)
	WBC *Count*	10,83 (5,8)	pH	7,39 (0,07)
	Trombosit	209,4 (120)	PaO2	114,7 (41,9)
	PT	14,74 (2,45)	PaCO2	41,78 (9,47)
	PTT	36,84 (13,9)	Laktat	1,81 (1,17)
	Hb	9,97 (1,54)	*Base Excess*	3,1 (3,28)
	HCT	29,96 (4,4)	HCO3	24,55 (5,37)
	INR	1,43 (0,4)	PaO2/FiO2	294 (136,5)

- Demographics → Cohort dominated by **older adults** (65.16 ± 16.27) with a **73.6% survival rate**
- Severity → **SOFA** variability (7.49 ± 3.15) **indicates fluctuating clinical states** across ICU time-windows
- **Low dispersion** → Most **vital-sign and lab features show CV < 30%**, meaning measurements cluster well around the mean
- **High dispersion** → Several **lab values** (WBC, platelets, SGOT/SGPT, PaO_2/FiO_2, bilirubin) show **CV > 30%**; **extreme variability (e.g., bilirubin CV 171%)** supports the need for **normalization** before agent training

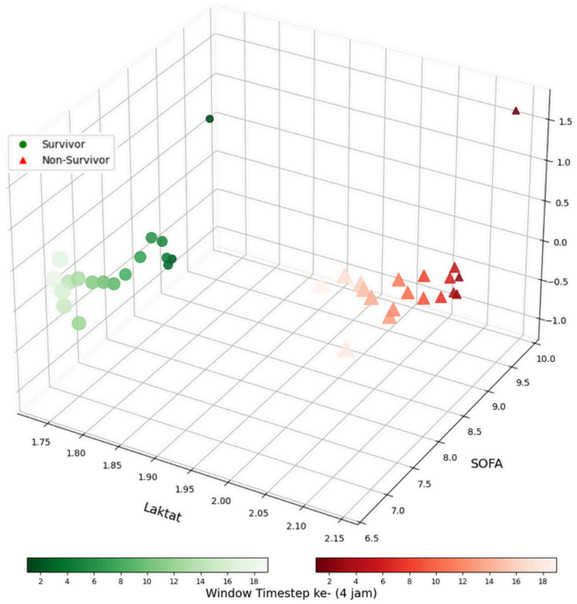
State Representation

- Features with the **largest separation** include: SOFA, MeanBP, bilirubin, creatinine, lactate, platelet count, GCS, and vasopressor dose history
- **Non-survivors** typically exhibit **higher organ dysfunction markers** (e.g., SOFA, bilirubin, lactate) and **lower physiologic stability** (e.g., MeanBP, GCS, platelets)

Action Discretization

Bin	IV *fluids* (mL/4 jam)	Vasopresor (µg/4 jam)
0	0	0
1	(0 , 1112,73]	(0 , 0,08]
2	(1112,73 , 2956,74]	(0,08 , 0,20]
3	(2956,74 , 6009,85]	(0,20 , 0,45]
4	> 6009,85	> 0,45

Reward Calculation

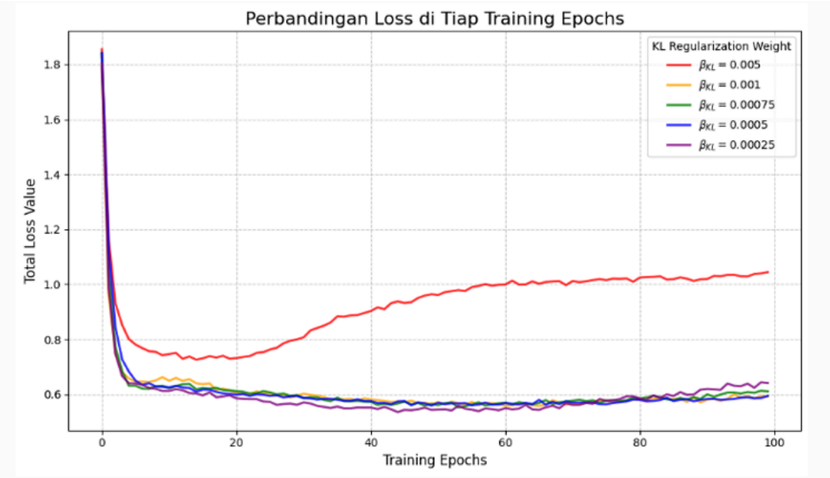


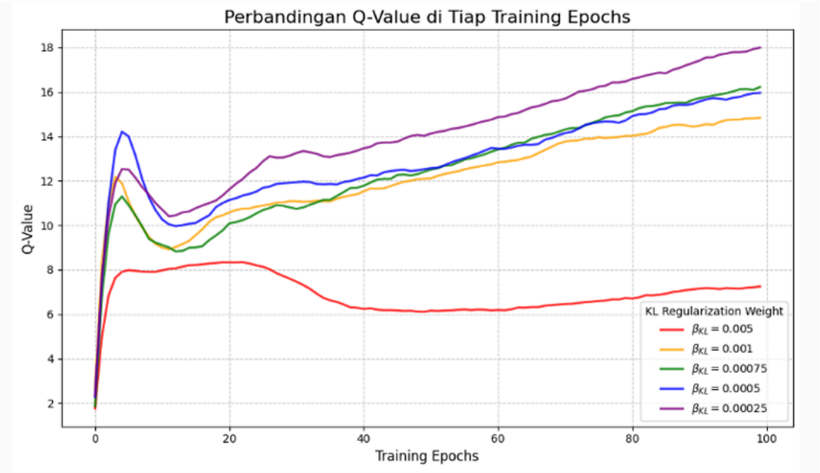
Survivors show **increasing intermediate rewards** with a positive **mean** net reward **(+0.949)** → clinical improvement

Non-survivors exhibit **declining intermediate rewards** with a negative **mean** net reward **(-0.685)** → worsening organ dysfunction and poor outcomes

Clear separation → **strong discriminative feedback** → helping the agent learn which therapeutic trajectories lead to successful or failed outcomes

Comparison of Total Loss and Q-Value during Agent Training





- **Loss** represents **prediction error between target and predicted Q-values**
- **Q-values** reflect the agent's **estimated expected return** under its learned policy

A large KL-regularization weight (β_{kl} = 0.005) produced the **worst performance**, characterized by:

- unstable and rising loss after epoch 20 (final loss 1.038)
- persistently low Q-values (~7), indicating under-learning due to excessive regularization

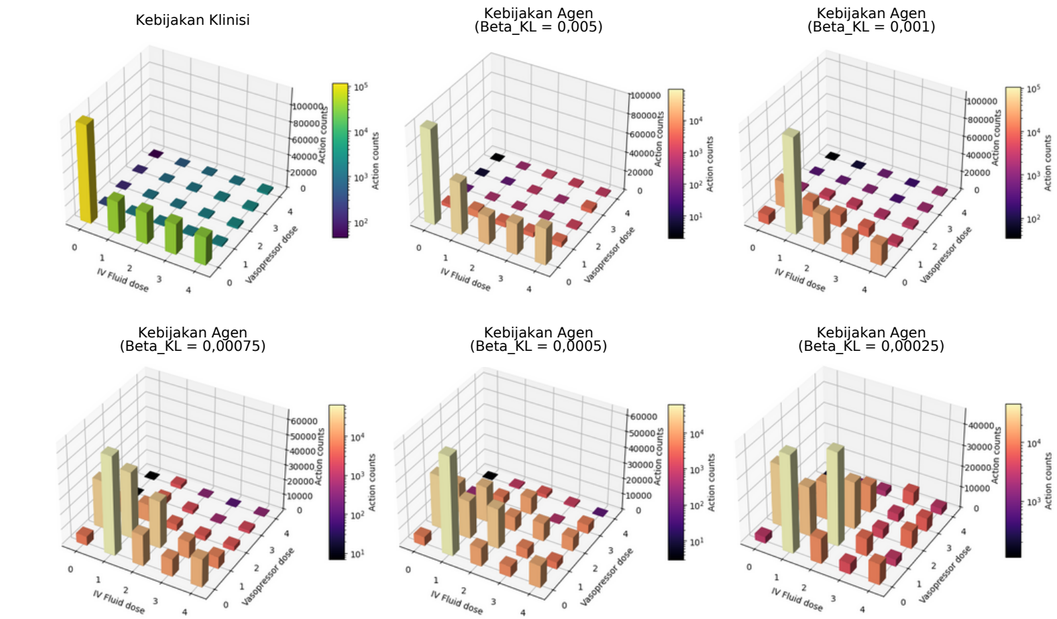
Intermediate β_{kl} values (0.001, 0.00075, 0.0005) showed **stable, fast loss convergence** and steadily increasing Q-values (final range 14.8–17.2)

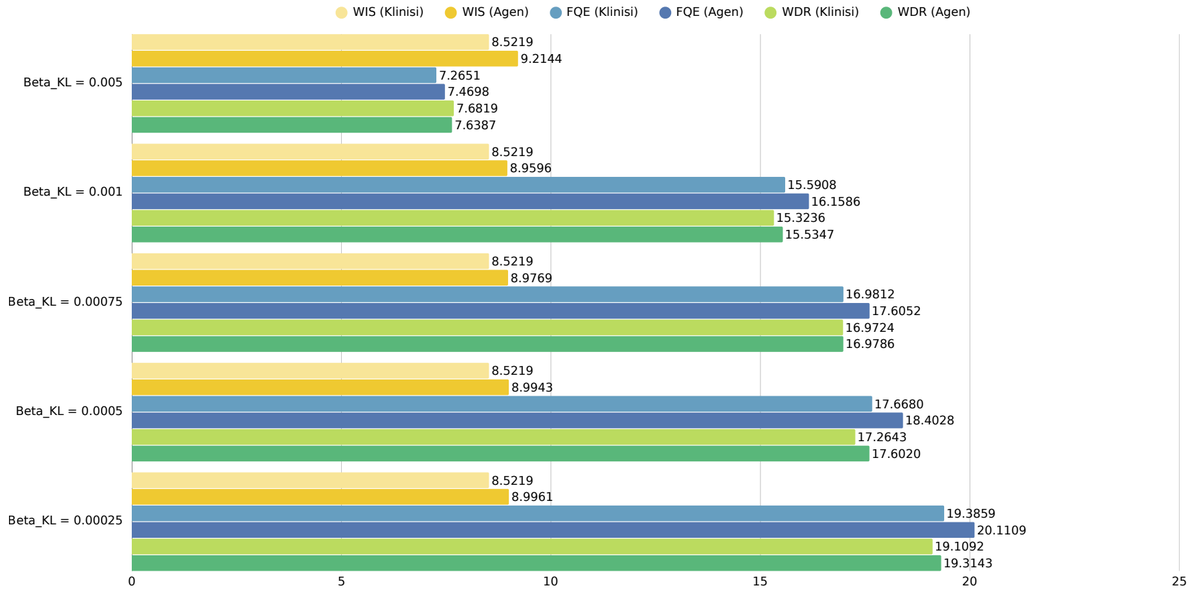
Among them, **β_{kl} = 0.0005** achieved **the most stable learning dynamics**, with the lowest and most stable final loss (0.592) and strong Q-value performance (~16.999)

The **smallest regularization (β_{kl} = 0.00025) achieved the highest Q-value** (18.998) but exhibited **late-epoch loss instability (final loss 0.642)**, suggesting over-optimism and reduced reliability

Overall, **β_{kl} = 0.0005 provided the best balance** between stable loss convergence and high Q-value estimates, making it the **most reliable configuration** for downstream evaluation

Policy Comparison in Selecting Optimal Action





Larger β_{kl} (0.005) produced **high-fidelity**, clinician-like policies, **dominated by action bins 0–1 with minimal exploration**

S**maller β_{kl} (0.00025) resulted in more exploratory policies**, with increased use of intermediate-dose vasopressors and IV fluids **(bins 2–3)**, diverging more strongly from clinician behavior

Trade-offs between Exploration & Stability

- β_{kl} = 0.005 failed to explore effectively, showing high, stagnant loss and low, stagnant Q-values, indicating inability to learn higher-value policies.
- β_{kl} = 0.00025 achieved the highest Q-values (≈18.99) but showed late-epoch loss instability, raising concerns about long-term reliability.

β_{kl} = 0.00025 produced the **highest estimated return** (FQE ≈ 20.11) but **lacked stability**

β_{kl} = 0.0005 achieved strong, **reliable performance** (FQE ≈ 18.40; WDR ≈ 17.60) and consistently outperformed the clinician baseline across OPE metrics

Final Selection

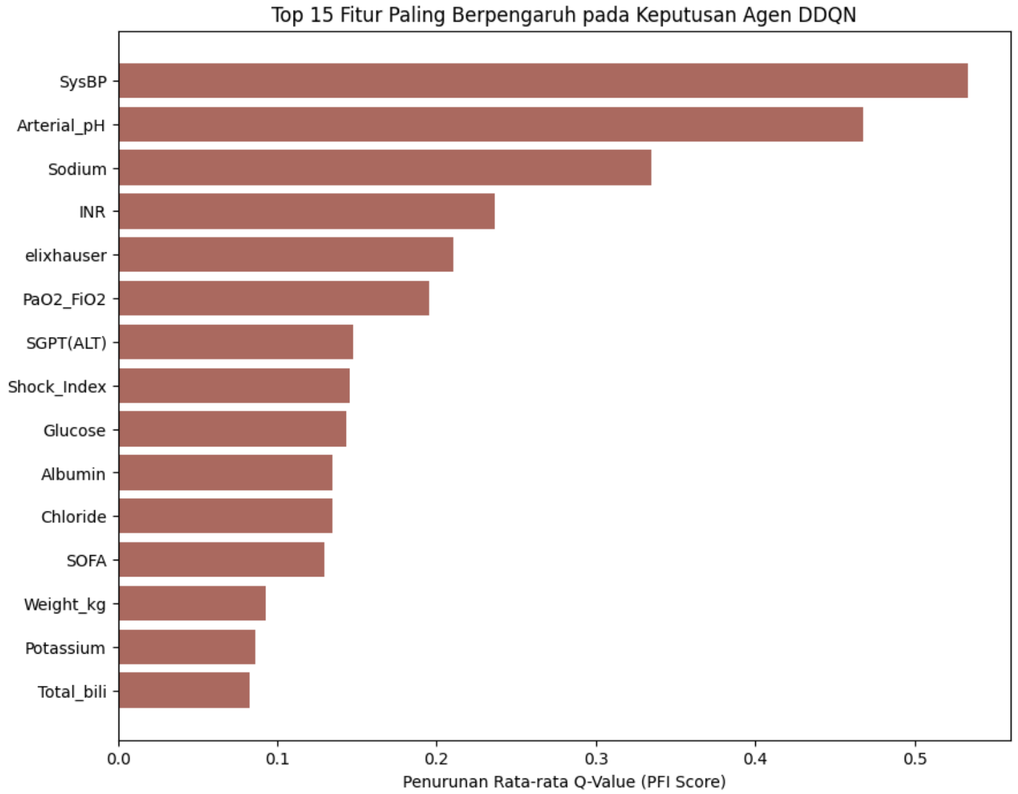
The agent with **β_{kl} = 0.0005 was selected as the optimal variant**, offering the best balance between exploration, stability, and high expected return, making it the most clinically reliable configuration

DDQN Dueling Agent Performance Testing

Policy (π)	WIS	FQE	WDR
Klinisi	8,5987	17,7555	17,5044
Agen *Dueling* DDQN	9,1763	18,4194	17,8040

- The selected agent (β_{kl} = 0.0005) was evaluated on the held-out test set using WIS, FQE, and WDR
- The **agent consistently outperformed the clinician policy across all metrics**
  - **Higher WIS** indicates the agent directly achieves **greater empirical reward** on unseen data
  - **Higher FQE and WDR** confirm the agent learned a policy with s**tronger long-term expected value** even after accounting for Q-function estimation

Feature Importance Analysis (PFI Score)



- **SysBP** showed the highest importance (~0.55), indicating **hemodynamic stability** is the dominant driver of the agent's decisions
- **pH** (~0.47) ranked second, highlighting the **relevance of metabolic status** and **acidosis in sepsis**

- **Key biochemical markers** such as Sodium (~0.35) and INR (~0.25) also had high importance, reflecting **fluid balance and hepatic dysfunction** as critical signals
- Overall, the top 15 features centered on organ dysfunction and hemodynamic instability, **aligning with findings from Wu et al. (2023)** and validating that the agent learned clinically meaningful decision cues

Closing

Conclusions

Based on the results of the research and the analyses conducted, it can be concluded that the design of a clinical decision-making agent for sepsis management using Dueling DDQN involves several key stages:

1. **Extraction of a sepsis cohort** from the MIMIC-III database using Sepsis-3 criteria and temporal rules for determining onset time.
2. **Preprocessing of clinical time-series data**, including windowing into 4-hour intervals, imputation of missing values, normalization, and construction of 48-feature state vectors.
3. **Action discretization** through quartile-based binning of IV fluids and vasopressor doses, resulting in 25 clinically interpretable action combinations.
4. **Design of reward functions**, capturing intermediate clinical progress (SOFA and lactate dynamics) and terminal outcomes (survival vs. mortality).
5. **Training and validation of multiple Dueling DDQN variants**, with systematic exploration of KL-regularization values to identify the most stable and optimal configuration.
6. **Evaluation** of learned policies using Off-Policy Evaluation metrics (WIS, FQE, WDR) and visualization of policy behavior.
7. Interpretation of model decision-making using Permutation Feature Importance (PFI) to identify the most influential clinical features.

The experimental findings confirm that the **agent with β_{kl} = 0.0005 provides the best balance between stability and performance**, achieving consistently higher expected returns than the clinician policy on both validation and test datasets. Off-Policy Evaluation confirmed superior performance, with the agent achieving higher expected return than clinicians (e.g., FQE ≈ 18.40 on validation). Furthermore, final test-set evaluation showed consistent improvement across all OPE metrics (e.g., WIS 9.1763 vs. 8.5987 for clinicians)

PFI analysis identified **SysBP (~0.55), pH (~0.47), Sodium (~0.35), and INR (~0.25)** as most influential features, reflecting hemodynamic instability and metabolic status as primary determinants of agent decisions.

Suggestions

1. Improve data preprocessing by exploring advanced outlier handling and machine-learning–based imputation to enhance state quality and temporal consistency
2. Investigate time-aware RL architectures (e.g., RNN, LSTM) to better capture the sequential nature of clinical trajectories
3. Explore refined reward formulations that incorporate additional indicators of sepsis progression
4. Conduct external validation using sepsis cohorts from other hospitals to assess generalizability and clinical reliability of the learned policy

UNIVERSITAS AIRLANGGA | FAKULTAS SAINS & TEKNOLOGI
SI TEKNIK BIOMEDIS

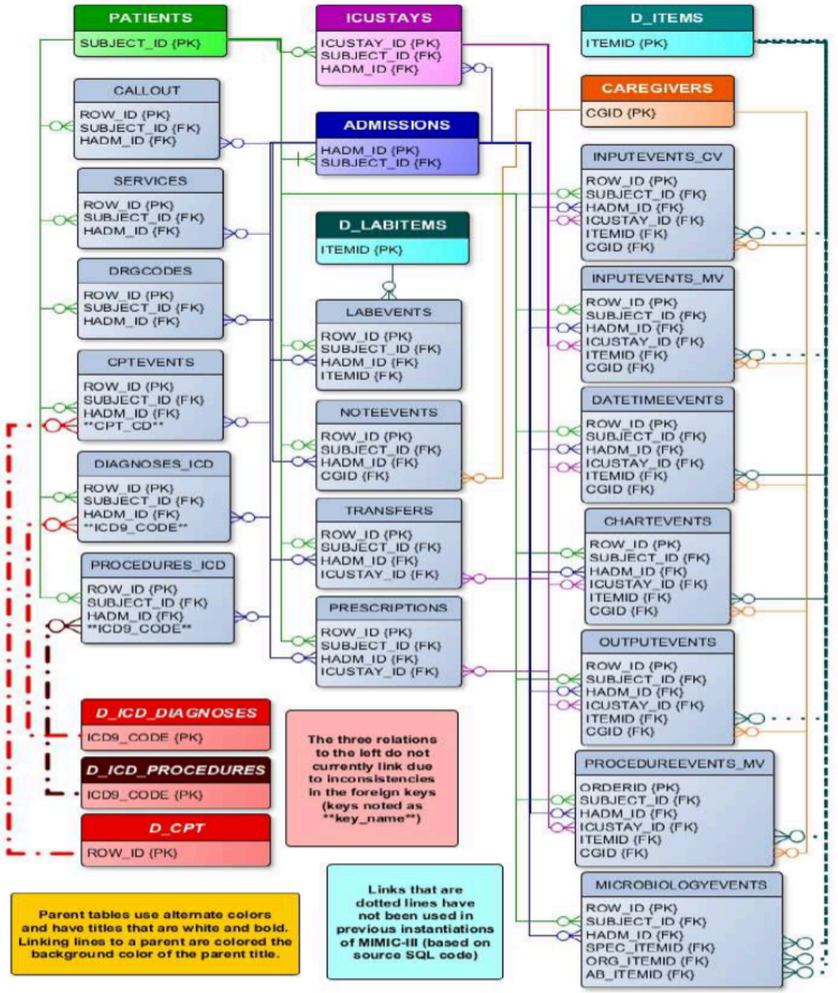
Thank You

SMART UNIVERSITY
TOP #308 QS WORLD UNIVERSITY RANKINGS
ZONA INTEGRITAS menuju wilayah bebas dari korupsi

Supporting Slides

MIMIC-III table map

Figure 2: MIMIC-III Entity Relationship Class Diagram



Reconstruction of the MIMIC-III Database for Data Analytics at SUNY Oswego

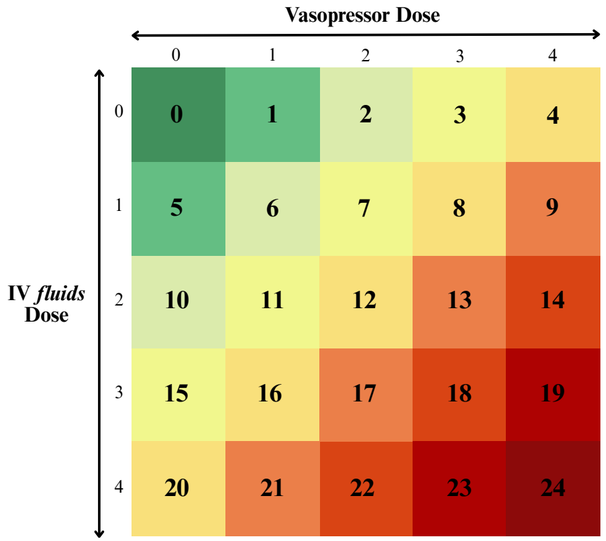
Supporting Slides

48 clinical indicators as state representation

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

Supporting Slides

Discretized actions



a_n	Interpretation
0	No Vasopressor and No IV Fluids Administered
1	Given 0 - 0,08 µg Vasopressor without IV Fluids Administration
2	Given 0,08 – 0,20 µg Vasopressor without IV Fluids Administration
3	Given 0,20 - 0,45 µg Vasopressor without IV Fluids Administration
4	Given > 0,45 µg Vasopressor without IV Fluids Administration
5	Given 0 - 1112,73 mL IV fluids without Vasopressor Administration
6	Given 0 - 0,08 µg Vasopressor and 0 - 1112,73 mL IV fluids
7	Given 0,08 – 0,20 µg Vasopressor and 0 - 1112,73 mL IV fluids
8	Given 0,20 - 0,45 µg Vasopressor dan 0 - 1112,73 mL IV fluids
9	Given > 0,45 µg Vasopressor dan 0 - 1112,73 mL IV fluids
10	Given 1112,73 - 2956,74 mL IV fluids without Vasopressor Administration
11	Given 0 - 0,08 µg Vasopressor and 1112,73 - 2956,74 mL IV fluids
12	Given 0,08 – 0,20 µg Vasopressor and 1112,73 - 2956,74 mL IV fluids
13	Given 0,20 - 0,45 µg Vasopressor and 1112,73 - 2956,74 mL IV fluids
14	Given > 0,45 µg Vasopressor and 1112,73 - 2956,74 mL IV fluids
15	Given 2956,74 – 6009,85 mL IV fluids without Vasopressor Administration
16	Given 0 - 0,08 µg Vasopressors and 2956,74 – 6009,85 mL IV fluids
17	Given 0,08 – 0,20 µg Vasopressor and 2956,74 – 6009,85 mL IV fluids
18	Given 0,20 - 0,45 µg Vasopressors and 2956,74 – 6009,85 mL IV fluids
19	Given > 0,45 µg Vasopressors and 2956,74 – 6009,85 mL IV fluids
20	Given > 6009,85 mL IV fluids without Vasopressor Administration
21	Given 0 - 0,08 µg Vasopressor and > 6009,85 mL IV fluids
22	Given 0,08 – 0,20 µg Vasopressor and > 6009,85 mL IV fluids
23	Given 0,20 – 0,45 µg Vasopressor and > 6009,85 mL IV fluids
24	Given > 0,45 µg Vasopressor and > 6009,85 mL IV fluids

Supporting Slides

Dueling DDQN Learning per Sequence

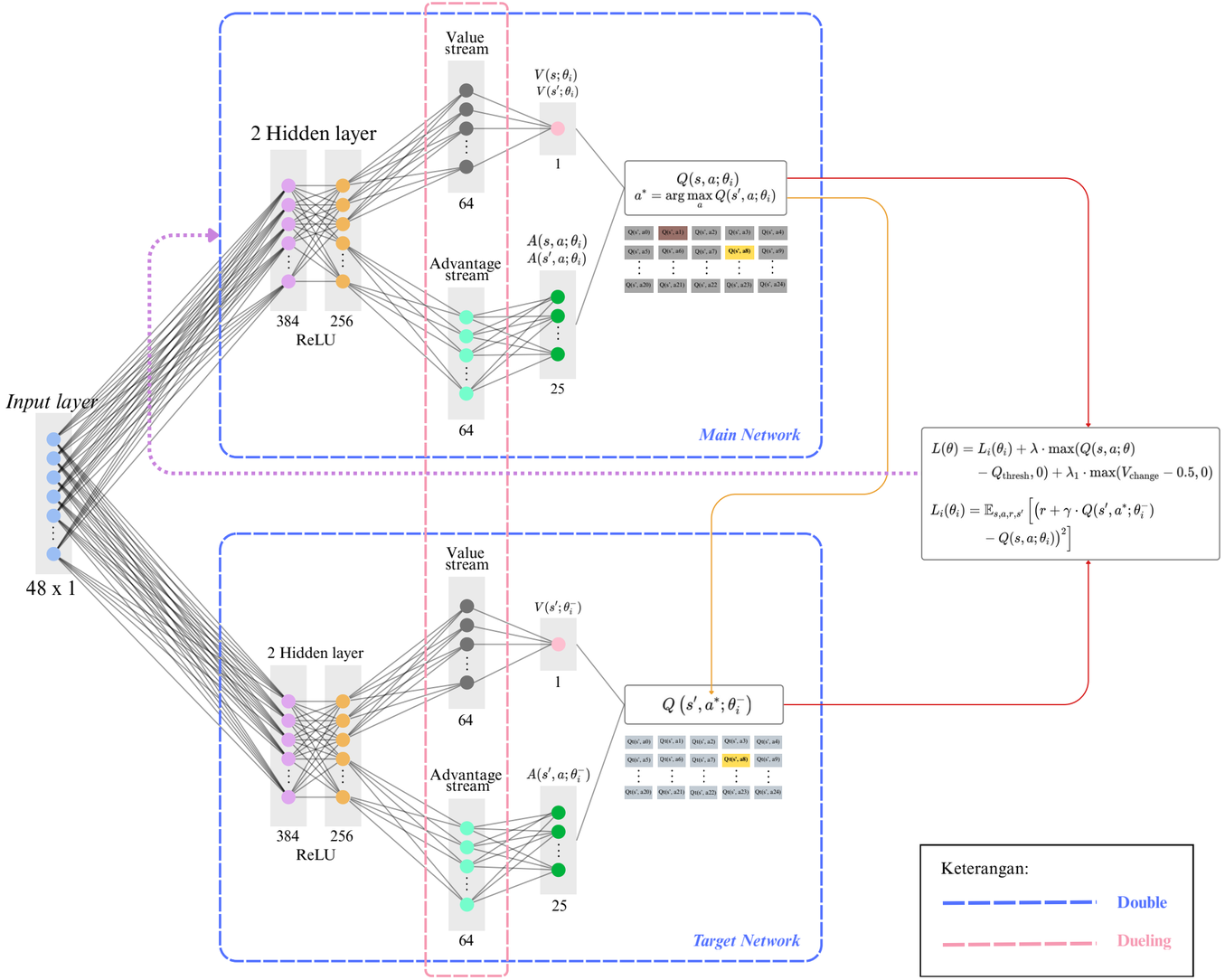
$$e_t = (s_t, a_t, r_t, s_{t+1})$$

$$s_t = [67,\ 101,\ 85,\ 7,\ 1.3,\ 2.8, \ldots]^\top$$

$$a_t = 1$$

$$r_t = 0.15$$

$$s_{t+1} = [67,\ 95,\ 89,\ 6,\ 1.2,\ 2.5, \ldots]^\top$$



$$L(\theta) = L_i(\theta_i) + \lambda \cdot \max(Q(s,a;\theta) - Q_{\text{thresh}}, 0) + \lambda_1 \cdot \max(V_{\text{change}} - 0.5, 0)$$

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma \cdot Q(s', a^*; \theta_i^-) - Q(s,a;\theta_i)\right)^2\right]$$

Keterangan:

Double

Dueling

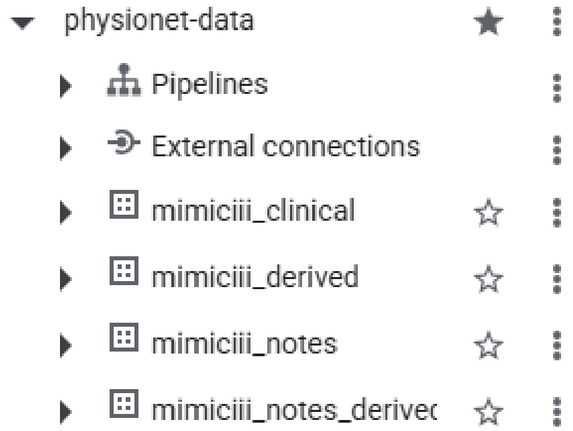
Backpropagation

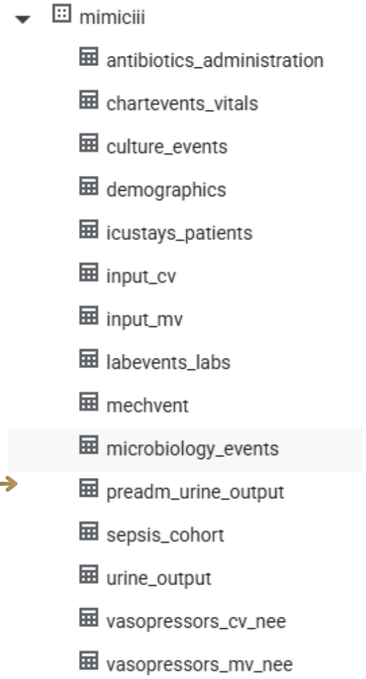
soft-update

$$\theta^- \leftarrow (1-\tau)\,\theta^- + \tau\,\theta$$

Supporting Slides

Data extraction dan state windowing






```
antibiotics_administration 84491
chartevents_vitals 55560167
culture_events 19343
demographics 61532
icustays_patients 61532
input_cv 3135918
input_mv 957959
labevents_labs 10612611
mechvent 2217611
microbiology_events 631726
preadm_urine_output 12853
sepsis_cohort 17763
urine_output 3400332
vasopressors_cv_nee 923686
vasopressors_mv_nee 159318
```

Patient cohorts. In both datasets, MIMIC-III and eRI, we included adult patients fulfilling the sepsis-3 criteria[7]. Sepsis was defined as a suspected infection (prescription of antibiotics and sampling of bodily fluids for microbiological culture) combined with evidence of organ dysfunction, defined by a SOFA score ≥2 (refs [7,31]). We adhered to the original temporal criteria for diagnosis of sepsis: when the antibiotic was given first, the microbiological sample must have been collected within 24 h; when the microbiological sampling occurred first, the antibiotic must have been administered within 72 h[31]. The earlier event defined the onset of sepsis. In line with previous research, we assumed a baseline SOFA of zero for all patients[31,32].

Sepsis Cohort → (icustays_id; charttime)