Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 8, No.2, April 2026, pp. 02602036-01 ~ 02602036-012
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v8i2.2080

Reinforcement Learning for Personalised Critical Care Treatment using Scalable Parallel Computing

Chandra Prasetyo Utomo[1], Kohei Ichikawa[2,3], Nashuha Insani[1], Kundjanasith Thonglek[4], Kang Xingyuan[2], Chaerita Maulani[5], Ummi Azizah Rachmawati [1,*]

[1]Department of Informatics, Universitas YARSI, Jakarta 10510, Indonesia

[2]Division of Information Science, Nara Institute of Science and Technology, Ikoma 630-0192, Japan

[3]Faculty of Business Data Science, Kansai University, Suita 565-8585, Japan

[4]Department of Computer Engineering, Kasetsart University, Bangkok 10900, Thailand

[5]Department of Periodontics, Universitas YARSI, Jakarta 10510, Indonesia

*ummi.azizah@yarsi.ac.id

Abstract. Sepsis is one of the leading causes of death in intensive care units. Many patients do not receive timely or effective treatment, which lowers their chances of survival. We developed a reinforcement learning–based framework to provide personalised treatment recommendations for sepsis patients. The model creates simple patient representations from treatment responses, groups patients with similar patterns, and learns the best treatment policy for each group. To reduce long training time, we use parallel and distributed computing. Using the MIMIC-III database and off-policy evaluation with weighted importance sampling, our method achieves a policy value of 79.933, higher than the clinician policy (47.654) and a general AI policy (57.658). A higher policy value indicates a lower mortality risk. These results show that our method can support faster, more accurate, and more effective treatment decisions in the ICU.

Keywords: reinforcement learning, off-policy evaluation, parallel computing, MIMIC-III dataset, sepsis management, ICU decision support, personalised treatment recommendation

(Received 2025-06-04, Revised 2025-11-30, Accepted 2026-01-23, Available Online by 2026-04-30)

1. Introduction

Intensive care unit (ICU) treatments can vary based on the clinical context, provider biases, and local medical practices [1]. To make real-time decisions, ICU physicians must analyze various observations, including medical notes, lab results, and multivariate vital signs. This complexity can lead to suboptimal treatment decisions. For instance, warfarin was improperly dosed in about one-third of patients [2]. Such non-optimal treatments can deplete hospital resources, prolong ICU stays, and pose unnecessary risks to patients [3]. Therefore, it is crucial to derive insights from ICU datasets that include instances of suboptimal treatment.

Sepsis ranks among the leading causes of mortality in the ICU, representing a critical organ dysfunction resulting from an internal infection [4]. Administering effective antimicrobial therapy within 30 minutes of infection onset can raise survival rates to 82%. However, for every 30-minute delay, survival chances decline by roughly 7% [5]. Alarmingly, over 50% of ICU patients do not receive adequate treatment until at least 5 hours have passed since infection, causing the survival rate to drop below 50%. Prompt and effective treatment is essential for improving outcomes. This time-sensitive, high-stakes environment frames the core challenge as a sequential decision-making problem, requiring an optimal policy to guide interventions.

Individual responses to ICU treatments vary significantly due to differences in physiology, medical history, and infection characteristics, making one-size-fits-all approaches ineffective. To address this, we aim to develop a personalised treatment recommendation system for sepsis using reinforcement learning (RL) and quantitatively demonstrate its superiority by achieving a significantly higher estimated policy value compared to baseline policies, enabling timely and tailored care to improve patient outcomes. The main challenges involve achieving real-time personalization and ensuring model scalability. Training such adaptive models requires high computational power and efficient evaluation, making parallel and distributed computing essential. These strategies allow the system to manage complex ICU data effectively, supporting its deployment in real-world clinical settings.

Several studies have investigated the use of AI and RL in ICU treatment. One study used biomarkers to improve policy reliability despite sparse rewards [6]. Offline RL with constrained exploration has been applied to enhance ventilator safety [7]. The SOFA-MDP model was utilized to optimize heparin dosage, employing reliable evaluation methods [8]. It has been demonstrated that ChatGPT can aid in troubleshooting CRRT alarms [9], while the cost-effectiveness of AI in ventilated patients has also been evaluated [10].

AI's role in predicting early sepsis risks, improving planning, and outcomes has been emphasized [11]. RL has also been applied to optimize drug choices and dosages. The use of RL in ICUs with MIMIC-III data was pioneered in [12], focusing on the optimization of IV fluid and vasopressor dosing. A data-driven RL approach using dueling double-deep Q-Networks (Dueling DDQN) was developed to inform IV fluid and vasopressor dosage recommendations and reduce in-hospital mortality [13]. However, integrating Dueling DDQN into clinical workflows is challenging due to the need for real-time data processing.

A Supervised RL model (SRL-RNN) for medication prescription has been introduced, which is hindered by the need for extensive labeled datasets and computational power [14]. Inverse RL based on clinician preferences has also been applied, achieving some success [15]. While RL shows promise for dosing, personalization is still limited. The RL algorithm assists physicians with treatment recommendations [16]. An actor-critic network was created for personalised cancer treatment [17], and offline RL has been recommended to optimize sepsis and diabetes care [18].

The Deep Attention Q-Network (DAQN) has been proposed, which personalizes treatments by recalling prior states and actions [19]. An assessment of various RL algorithms found that the enhanced Deep Deterministic Policy Gradient (DDPG) method was better aligned with clinicians' judgments and led to reduced mortality [20]. Despite employing robust methods and datasets, most of these studies lack a focus on model scalability. Furthermore, systematic reviews of AI in the ICU consistently highlight a persistent gap between model development and clinical deployment, noting that most models are not validated for real-world integration or scalability [21], [22].

Scalability is essential for handling AI workloads through distributed training and edge AI to optimize resource use and reduce training time. Foundational work in computer science has established

robust frameworks for distributed RL, such as Ray, which enables massive parallelization [23], while a comprehensive surveys detail numerous techniques for accelerating deep RL [24]. Many studies have showcased and examined their use in AI-driven distributed systems and edge computing in the healthcare domain, transforming AI, big data, and advanced technologies to enhance diagnostics or even to reduce training time for clinical decision support systems [25], [26], [27]. However, these studies highlight a critical disconnect that the scalable tools exist but are not being broadly applied to tackle the personalisation problem.

While RL has been applied to critical care using various models like DDQN, SRL-RNN, inverse RL, DAQN, and actor-critic models, these methods still face challenges in real-time usability, high data requirements, and personalization, revealing two critical, unaddressed research gaps: (1) a lack of personalization, with most models offering non-personalised policies, and (2) a failure to address the computational scalability required for real-world deployment. To address these limitations, we developed a personalised treatment recommendation system using RL that leverages a distributed computing architecture to efficiently train and evaluate tailored policies.

Using real patient data from the MIMIC-III database [28], our primary contribution is the quantitative demonstration that this personalized, scalable model outperforms clinician actions and existing non-personalised AI models, as measured by a statistically significant increase in estimated policy values. In doing so, we bridge the gap between AI model development and the practical implementation challenges, such as workflow integration and clinician trust, that are critical for successful clinical translation [29], [30].

2. Methods

2.1. Problem Formulation

We define our problem as Markov Decision Process (MDP), defined by S represent the *state space* of ICU patients, with $s_t \in S$ indicating the state at time *t*. Define A as the *action space* of medical treatments, where $a_t \in A$ signifies the action at time *t*. The *transition function*, denoted as $T(s, a, s') : S \times A \times S \rightarrow [0,1]$, captures the shift of the patient from state s to state s' following action a. The *reward function*, represented as $R(s, a) : S \times A \rightarrow \mathbb{R}$, provides an immediate scalar reward after action a is implemented in state s. Additionally, let $\gamma \in [0,1]$ denote a *discount factor*.

The complete MDP is defined as a 5-tuple $M=(S,A,T,R,\gamma)$ where each component has been explained earlier. A *policy* $\pi : S \times A \rightarrow [0,1]$ represents the probability distribution that associates a state $s \in S$ with an action $a \in A$. Define $G_t=R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$ as the summation of discounted total rewards from time *t* through the end of ICU stay *T*. The *value function* $V^{\pi}(s)$ denotes the expected total reward for adhering to policy π from state s. The goal is to determine an optimal policy π^* that satisfies the condition $V^{\pi*}(s) \geq V^{\pi}(s)$. This indicates that π^* represents the strategy that maximizes the expected patient outcome G_t. We utilize a Q-learning algorithm to derive this optimal policy.

2.2. Personalised Policy

Creating a personalised computational model involved several steps, as illustrated in Figure 1: identifying state clusters, defining patient representation, clustering patients, and computing optimal treatment policies. The first step focused on discretizing patient health status into a set of physiological states using K-Means clustering on 48 clinical variables; the data preprocessing and extraction details for these variables are described in Section 3.1 Experimental Data. This process aimed to group patient health data into distinct states to simplify pattern recognition. To reflect final patient outcomes, two additional states were added, representing hospital discharge and in-hospital mortality.

To enable personalization, the patient trajectories were then grouped based on similar characteristics. The dimensionality of patient data was first reduced using Principal Component Analysis (PCA). Following this reduction, patients were grouped into clusters using another K-Means application.

After defining both state and patient clusters, we computed optimal treatment policies tailored to each patient group. The final output was the optimal policy π^* and its corresponding state-action values. These policy values, V^{π}, were derived using the Bellman equation and were designed to maximize 90-day survival.

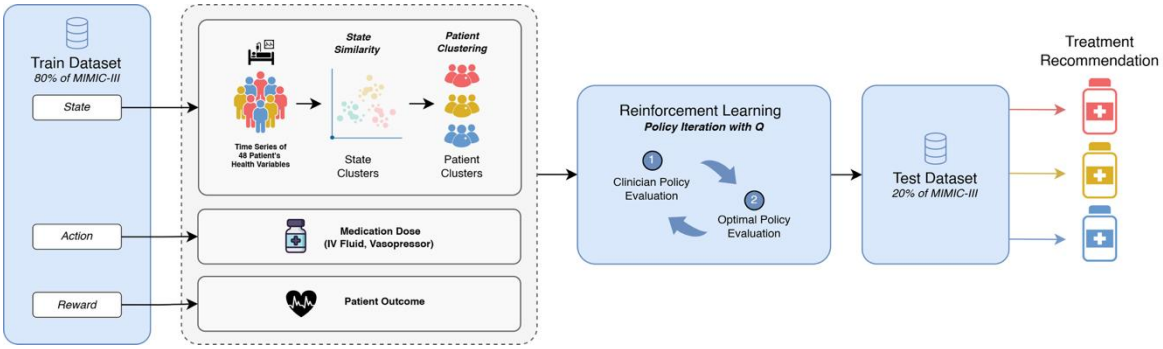


Figure 1. Personalised Policy

2.3. Off-Policy Evaluation

The performance of the AI policies was estimated using off-policy evaluation (OPE), which allows for the assessment of a new policy using historical data generated by a different policy (the clinicians' policy). This method avoids the costs and risks associated with a prospective clinical trial. Specifically, we employed the High-Confidence Off-Policy Evaluation (HCOPE) method from Komorowski et al. [12], which utilizes Weighted Importance Sampling (WIS) with 2,000 bootstrap resamples. This technique provides a statistically robust and conservative estimate of a real-world performance by generating a distribution of potential policy values from which 95% confidence lower and upper bounds (LB and UB) are derived. While WIS may be a biased (though consistent) policy estimator, bootstrapping is accepted as producing accurate confidence intervals for such applications [12], [31], [32], [33]. This approach is suggested in reinforcement learning research as a safe and effective method for high-risk applications, such as healthcare, and allows for a highly conservative comparison between policies [12], [31], [34].

2.4. Parameter Selection

A comprehensive parameter search was conducted to find the optimal configuration for the model. To identify an effective number of physiological state clusters, we systematically varied the number of states from 10 to 1,500. The suitability of each configuration was evaluated using the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and Total Within-Cluster Sum of Squares (WSS), with the full analysis presented in Figure 2.

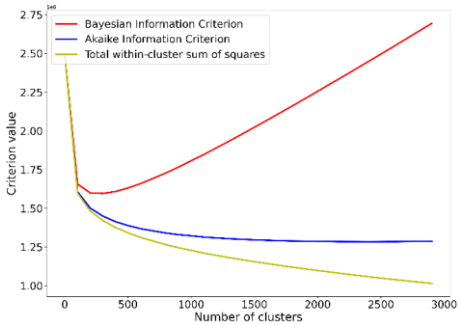


Figure 2. Analysis of the Number of State Clusters

We also investigated the patient clustering parameters, testing configurations from two to five groups. To assess the feature reduction step, we evaluated multiple Principal Component Analysis (PCA) solvers (arpack, auto, randomized, and full) and varied the number of principal components from 2 to 10. The impact of these parameters on the final estimated policy value was measured using two representative state cluster sizes (750 and 1250 states), as detailed comparative results shown in Figure 3. Due to its significantly higher computational cost without a corresponding performance benefit, the 'full' PCA solver was excluded from the final model set.

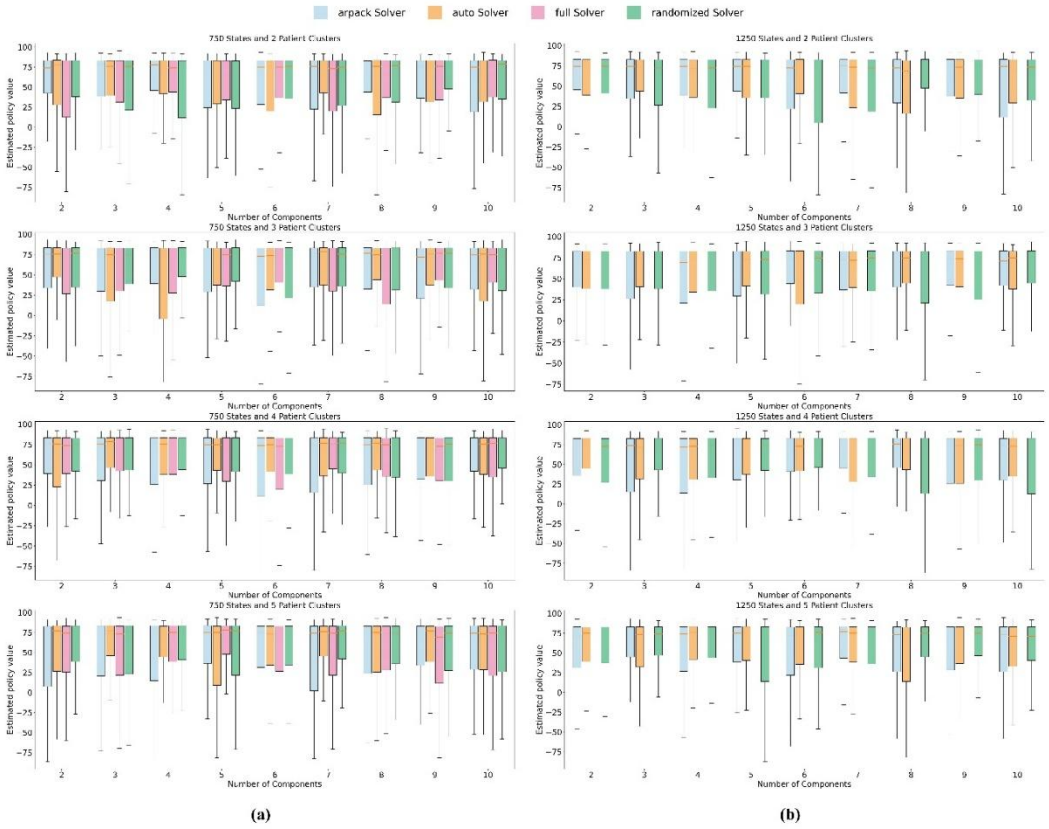


Figure 3. Comparison of estimated policy values across different PCA solvers (arpack, auto, randomized, and full) and components (2 to 10) for 750 states (a) and 1250 states (b), showing no significant differences between solvers except for the full solver, which had lower values and higher computation time in the 750-state scenario.

2.5. Efficient Parallelization Method for Model Building

Building optimal machine learning models involves exploring many parameter combinations, which can be time-consuming. To speed up this process, we used parallel computing to build multiple models at once. Our compute cluster consists of 40 nodes (see Table 1) and supports multiprocessing, enabling faster model building by leveraging multiple CPU cores. As shown in Figure 4, building a model on a single compute node using one core takes approximately 1,525 seconds, reduced to 831 seconds with two cores, and 211 seconds with all 52 cores. However, the number of cores gains diminishing returns due to parallelization overhead.

Based on these findings, distributing model building across multiple nodes and running them concurrently is more efficient than heavily parallelizing the construction of a single model. We therefore decomposed our program into sub-programs, each responsible for building a single model. These sub-programs are distributed and executed concurrently across multiple nodes, with specified CPU core allocations. This architecture allows flexible resource use while enabling parallel model training, greatly improving the overall efficiency of the model-building process.

It is important to clarify that our parallelization strategy was selected to address the specific computational challenge of this study. That is building and evaluating many distinct, independent models for robust statistical analysis. This parallel workload differs significantly from that required to accelerate the training of a single, computationally intensive model.

Each model-building task, involving K-Means clustering and Q-learning on a discrete state space, is relatively lightweight and CPU-bound. We therefore prioritized high-throughput, task-level parallelization (distributing jobs horizontally across many CPU nodes) over model-level acceleration (e.g., GPU acceleration). The latter approach would provide negligible performance benefits for these specific, non-GPU-intensive tasks and was not implemented, as it is fundamentally unsuited to our problem.

Table 1. Specifications of a single compute node in the 40-node compute cluster

Component	Specification
CPU per Node	2 x Intel Xeon Gold 6230R (52 cores in total)
Memory per Node	384GB
NIC	Mellanox ConnectX-5 Ex (Dual port 100GbE)

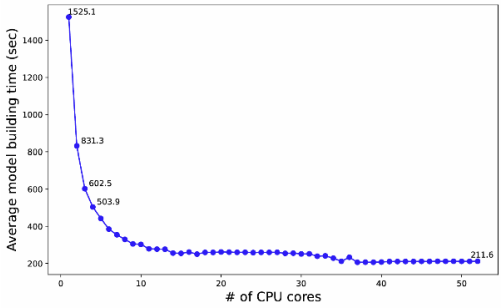


Figure 4. Average model building time relative to the number of CPU cores used for a Personalised Policy model with 750 states and five patient clusters.

3. Results and Discussion

3.1. Experimental Data

Our AI policy was developed and evaluated using the Medical Information Mart for Intensive Care III (MIMIC-III) database, a comprehensive medical dataset from Beth Israel Deaconess Medical Center between 2001 and 2012 [28]. We applied strict inclusion and exclusion criteria to ensure data quality. Patients included had records from 24 hours before to 48 hours after sepsis onset. We excluded individuals under 18, those missing death status, receiving high vasopressor doses, lacking IV fluid data, deceased during the period, or with incomplete vasopressor records. After filtering, 20,913 ICU stays were selected for analysis.

We extracted 48 variables covering demographics, comorbidities, vital signs, labs, fluid, and vasopressor data. The data was formatted as a multidimensional time series with 4-hour intervals. The primary treatment variables were the cumulative intravenous fluid volume and peak vasopressor dose administered per interval. The dataset was then split 80:20 into training and testing sets, with 80% for training and 20% for testing. To ensure robustness, the split was randomized in each modeling iteration, causing variation in ICU stay distribution between iterations. The complete data preprocessing is illustrated in Figure 5.

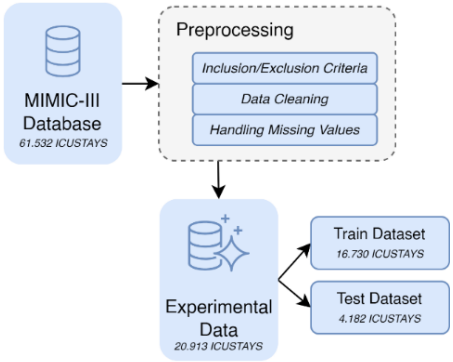


Figure 5. Experimental Data

3.2. Experimental Design

We built 500 distinct reinforcement learning models to conduct a thorough evaluation of treatment policies. The performance of the best-performing AI Personalised Policies was compared against two

baselines: the observed Clinician Policy and a non-personalized AI General Policy. For this experiment, the action space was defined by discretizing treatment into a 5x5 grid representing five levels of intravenous fluid and five levels of vasopressor dosages administered in a 4-hour window. The reward function was directly tied to the 90-day patient outcome, defined as a terminal reward of +100 for survival and –100 for mortality. The Clinician Policy baseline was established by analyzing all historical treatment decisions in the test set and calculating their average return, scaled to this [-100, 100] range. The overall process for model development is illustrated in Figure 1.

3.3. Policy Values Result

In this section, we thoroughly compare the performance of three policies: Clinician Policy, AI General Policy, and AI Personalised Policy. Detailed in Table 2, our analysis focused on a conservative estimate of performance. Following the methodology in [12], our optimal personalised model was selected by identifying the configuration that maximized the 95% confidence lower bound (LB) of the policy value. As shown in Figure 6, the Clinician Policy maintained relatively stable performance across all state sizes, with median values hovering around 47 (≈ 47), with 95% confidence upper bound of 77,95. The AI General Policy initially underperformed at smaller state sizes but gradually improved as the number of states increased, surpassing the Clinician Policy at 750 states.

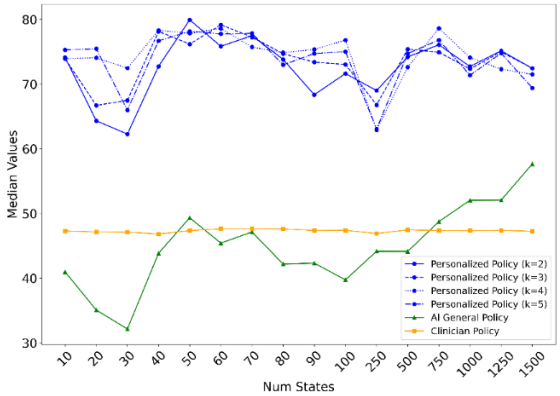


Figure 6. Median of the mean of estimated policy values over 500 models.

The AI Personalised Policy, represented by four curves corresponding to different numbers of patient clusters, consistently outperformed the other two policies at all state sizes. The highest performing configuration (k=2 at 50 states) achieved a median value of 79.93, with a 95% confidence lower bound of 89,59. Statistical significance was established by demonstrating that the 95% LB of the best-performing personalised policy consistently exceeded the 95% UB of the Clinician Policy**.** This non-overlapping bound is a highly conservative metric indicating that our model's worst-case expected performance is statistically superior to the clinician's best-case expected performance. These findings highlights the ability of the personalised approach to adapt effectively to the state space complexity by leveraging tailored clustering strategies. The relatively minor performance variations among different cluster configurations further indicate the robustness of this personalized method.

To translate this statistical finding into a clinical context, the policy value serves as a robust proxy for alignment with patient survival, as the underlying reward function was defined as +100 for 90-day survival and -100 for mortality. A higher policy value, therefore, indicates that a policy more consistently recommends actions that are historically associated with positive patient outcomes. The significant increase from a median value of ≈ 47 to 79.93 is not just a statistical abstraction. We can translate this improvement into a more concrete clinical metric by leveraging the analysis from Komorowski et al. [12], which established a direct correlation between this action-return value and observed 90-day mortality. Based on their findings, the clinician policy value of ≈ 47 corresponds to an observed mortality risk of approximately 23%. In contrast, our personalized policy's median value of 79.93 corresponds to an observed mortality risk of approximately 10%. This suggests that the personalized policy is more adept at identifying optimal treatment paths, and this enhanced alignment

with favorable outcomes indicates a strong potential to improve clinical decision-making and guide clinicians toward interventions with a higher likelihood of leading to patient survival.

Table 2. Median of mean of estimated policy values

Num of States	Clinician Policy	AI General Policy	Personalised Policy			
			k=2	k=3	k=4	k=5
10	47.315	40.987	74.126	73.916	73.940	**75.296**
20	47.153	35.081	64.332	66.721	74.073	**75.469**
30	47.128	32.160	62.277	67.480	**72.459**	65.984
40	46.836	43.856	72.749	78.153	**78.319**	76.689
50	47.352	49.348	**79.933**	76.179	77.853	78.146
60	47.646	45.423	75.850	**79.172**	78.597	77.769
70	47.654	47.153	77.484	77.253	75.737	**77.859**
80	47.625	42.185	73.813	74.697	**74.869**	72.982
90	47.373	42.333	68.360	73.394	**75.352**	74.722
100	47.407	39.749	71.630	73.038	**76.814**	75.018
250	46.915	44.162	**69.001**	66.787	62.902	63.072
500	47.459	44.147	74.205	**75.400**	72.625	74.726
750	47.371	48.744	76.077	74.928	**78.631**	76.820
1000	47.341	52.047	72.727	72.344	**74.094**	71.367
1250	47.391	52.070	**75.152**	74.980	72.313	74.786
1500	47.263	57.658	72.437	**72.454**	71.512	69.406

These findings underscore the importance of personalization and extend prior work in this domain. Research by Komorowski et al. [12] first established the viability of reinforcement learning for sepsis treatment, demonstrating a significant improvement over clinician policies. Our work builds directly on this by introducing an explicit personalization layer, providing evidence that a tailored approach can yield greater performance gains than a single, general policy. This highlights that while reinforcement learning holds promise for clinical decision support, its true potential is substantially amplified through personalization. Stratifying patients into distinct health states allows the model to move beyond a one-size-fits-all approach and learn more nuanced, effective treatment strategies.

3.4. Model Robustness and Sensitivity Analysis

To ensure our findings were not artifacts of specific hyperparameter choices, we conducted a sensitivity analysis on key components of the modeling pipeline. We first assessed the patient representation step, as shown in Figure 3. The estimated policy values remained stable when varying the number of principal components from 2 to 10 and across multiple PCA solvers (arpack, auto, randomized). This stability demonstrates that the superior performance of the personalized approach is a robust finding, not dependent on the fine-tuning of this dimensionality reduction step.

Additionally, we analyzed the sensitivity of the personalization effect to the size of the state space (Figure 6). With a smaller number of state clusters (e.g., 10-20 states), the performance differences among models with varying numbers of patient clusters were more pronounced. This is likely because the data is more densely distributed across fewer state-action pairs, allowing the model to better capture the benefits of personalization. Conversely, as the number of state clusters increased, the performance gap between different patient cluster configurations narrowed. This is attributed to data sparsity, as the available patient trajectories become insufficient to robustly learn distinct policies for multiple subgroups across a vast number of states. These analyses together strengthen the validity of our conclusions by demonstrating their robustness across a range of key parameters.

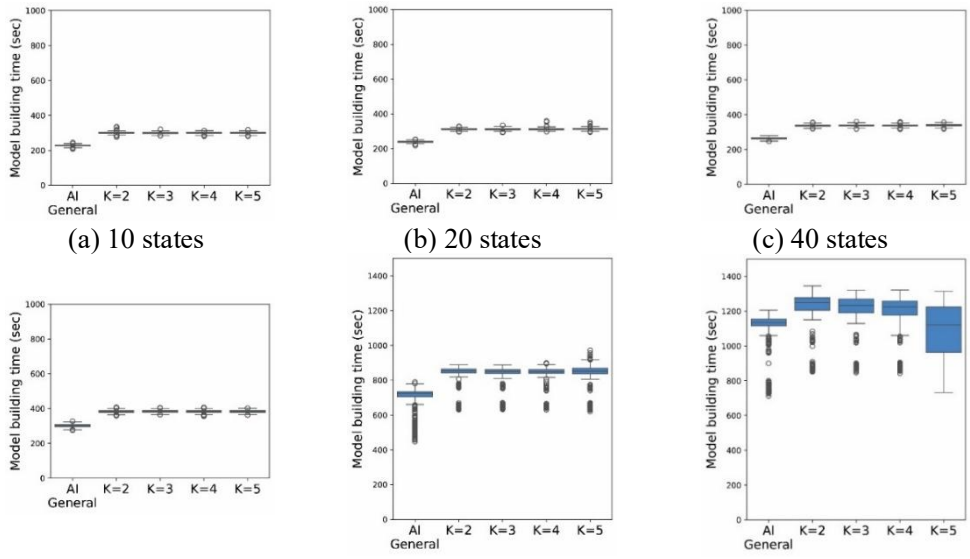
3.5. Parallelized Model Building Time

As discussed in Section III, efficient model building depends on optimizing CPU core allocation and enabling parallel execution of model-building tasks. In this study, each model was assigned 2 CPU cores to build 500 models distributed across multiple compute nodes. When sufficient computational resources are available, these tasks can run concurrently, significantly reducing the total processing

time. Although increasing the number of CPU cores per model could speed up individual builds, it would increase overall resource consumption and potentially lengthen job queue times, extending the total processing time.

Figure 7 shows the build times for models with 10, 20, 40, 80, 750, and 1,250 states using two CPU cores. The "AI General" label indicates the building time for the AI General Policy, while "K=2" to "K=5" represent the building times for Personalised Policy with 2 to 5 patient clusters. Although build times vary based on patient clustering, Personalised Policy models generally take longer than the AI General Policy model. Additionally, as the number of states increases, the time required to build the model tends to increase. However, even for models with 1,250 states, the average building time is approximately 1,200 seconds. With sufficient resources to execute all 500 models in parallel, the entire set can be built in about 20 minutes, illustrating a highly efficient process.

This 20-minute completion time for the entire 500-model experiment validates the effectiveness of our chosen horizontal scaling strategy. It confirms that, for our goal of generating a large cohort of models for statistical evaluation, distributing many lightweight, CPU-bound tasks across a cluster is substantially more efficient than attempting to accelerate individual tasks with specialized hardware such as GPUs.



(a) 10 states (b) 20 states (c) 40 states

(d) 80 states (e) 750 states (f) 1250 states

Figure 7. Comparison of individual model building time.

4. Conclusion

This study proposed a personalized ICU treatment system for sepsis using reinforcement learning, addressing personalization and scalability. Using the MIMIC-III database, it outperformed clinician actions and general AI models, indicating better treatments and lower mortality. Scalability was improved via parallel computing, reducing training time and handling complex states efficiently.

While promising, the model hasn't undergone clinical validation, and real-time deployment is a future goal. Future work will focus on integrating it with electronic health record (EHR) platforms, validating recommendations with prospective or multicenter ICU data, and extending to continuous state and action spaces. Ethical issues like bias, privacy, and transparency will guide development. To address privacy, especially in multicenter collaborations, we plan to explore federated learning, allowing the model to learn from diverse datasets without centralizing sensitive data. These efforts aim to ensure reliable, fair, and actionable AI support for critical care decision-making.

Declaration of AI and AI assisted technologies in the writing process
During the preparation of this work, the authors used ChatGPT and Gemini in order to enhance language and grammar and improve readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements
This project is partially funded by the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek), Republic of Indonesia, with Grant ID: 105/E5/PG.02.00.PL/2024, JSPS KAKENHI Grant Number 25K15138, and ROIS NII Open Collaborative Research 252S5-23676.

References
[1] S. Alban, "Adverse Effects of Heparin," in *Heparin - A Century of Progress*, R. Lever, B. Mulloy, and C. P. Page, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 211–263. doi: https://doi.org/10.1007/978-3-642-23056-1_10.
[2] Health Quality Ontario, "Point-of-Care International Normalized Ratio (INR) Monitoring Devices for Patients on Long-term Oral Anticoagulation Therapy: An Evidence-Based Analysis.," *Ont Health Technol Assess Ser*, vol. 9, no. 12, pp. 1–114, 2009, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23074516
[3] S. Nemati, M. M. Ghassemi, and G. D. Clifford, "Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach," in *Proceeding of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'16)*, Orlando, FL, USA, Aug. 2016, pp. 2978–2981. doi: https://doi.org/10.1109/EMBC.2016.7591355
[4] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, p. 801, Feb. 2016, doi: https://doi.org/10.1001/jama.2016.0287.
[5] A. Kumar *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*," *Crit Care Med*, vol. 34, no. 6, pp. 1589–1596, Jun. 2006, doi: https://doi.org/10.1097/01.CCM.0000217961.75225.E9.
[6] A. Shirali, A. Schubert, and A. Alaa, "Pruning the Way to Reliable Policies: A Multi-Objective Deep Q-Learning Approach to Critical Care," *IEEE J Biomed Health Inform*, vol. 20, no. 10, pp. 6268 – 6279, 2024, doi: https://doi.org/10.1109/JBHI.2024.3415115.
[7] B. Zhang, X. Qiu, and X. Tan, "Balancing therapeutic effect and safety in ventilator parameter recommendation: An offline reinforcement learning approach," *Eng Appl Artif Intell*, vol. 131, p. 107784, May 2024, doi: https://doi.org/10.1016/J.ENGAPPAI.2023.107784.
[8] J. Liu *et al.*, "Value function assessment to different RL algorithms for heparin treatment policy of patients with sepsis in ICU," *Artif Intell Med*, vol. 147, p. 102726, Jan. 2024, doi: https://doi.org/10.1016/J.ARTMED.2023.102726.
[9] M. S. Sheikh *et al.*, "Personalized Medicine Transformed: ChatGPT's Contribution to Continuous Renal Replacement Therapy Alarm Management in Intensive Care Units," *J Pers Med*, vol. 14, no. 3, p. 233, Mar. 2024, doi: https://doi.org/10.3390/JPM14030233/S1.
[10] L. R. Zwerwer *et al.*, "The value of artificial intelligence for the treatment of mechanically ventilated intensive care unit patients: An early health technology assessment," *J Crit Care*, vol. 82, p. 154802, Aug. 2024, doi: https://doi.org/10.1016/J.JCRC.2024.154802.
[11] D. O'Reilly, J. McGrath, and I. Martin-Loeches, "Optimizing artificial intelligence in sepsis management: Opportunities in the present and looking closely to the future," *Journal of Intensive Medicine*, vol. 4, no. 1, pp. 34–45, Jan. 2024, doi: https://doi.org/10.1016/j.jointm.2023.10.001.

[12] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care," *Nat Med*, vol. 24, no. 11, pp. 1716–1720, 2018, doi: https://doi.org/10.1038/s41591-018-0213-5.

[13] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, "Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach," in *Proceedings of the 2nd Machine Learning for Healthcare Conference*, F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., in Proceedings of Machine Learning Research, vol. 68. PMLR, Oct. 2017, pp. 147–163. [Online]. Available: https://proceedings.mlr.press/v68/raghu17a.html

[14] L. Wang, X. He, W. Zhang, and H. Zha, "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2447–2456, 2018, doi: https://doi.org/10.1145/3219819.3219961.

[15] C. Yu, J. Liu, and H. Zhao, "Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units," *BMC Med Inform Decis Mak*, vol. 19, Apr. 2019, doi: https://doi.org/10.1186/s12911-019-0763-6.

[16] H. Zheng, I. O. Ryzhov, W. Xie, and J. Zhong, "Personalized Multimorbidity Management for Patients with Type 2 Diabetes Using Reinforcement Learning of Electronic Health Records," *Drugs*, vol. 81, no. 4, pp. 471–482, Mar. 2021, doi: https://doi.org/10.1007/s40265-020-01435-4.

[17] M. Liu, X. Shen, and W. Pan, "Deep reinforcement learning for personalized treatment recommendation," *Stat Med*, vol. 41, no. 20, pp. 4034–4056, Sep. 2022, doi: https://doi.org/10.1002/sim.9491.

[18] M. Nambiar, S. Ghosh, P. Ong, Y. E. Chan, Y. M. Bee, and P. Krishnaswamy, "Deep Offline Reinforcement Learning for Real-world Treatment Optimization Applications," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2023, pp. 4673–4684. doi: https://doi.org/10.1145/3580305.3599800.

[19] S. Ma, J. Lee, N. Serban, and S. Yang, "Deep Attention Q-Network for Personalized Treatment Recommendation," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, Dec. 2023, pp. 329–337. doi: https://doi.org/10.1109/ICDMW60847.2023.00048.

[20] T. Lin *et al.*, "A dosing strategy model of deep deterministic policy gradient algorithm for sepsis patients," *BMC Med Inform Decis Mak*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12911-023-02175-7.

[21] D. van de Sande, M. E. van Genderen, J. Huiskens, D. Gommers, and J. van Bommel, "Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit," *Intensive Care Med*, vol. 47, no. 7, pp. 750–760, Jul. 2021, doi: https://doi.org/10.1007/s00134-021-06446-7.

[22] M. R. Pinsky *et al.*, "Use of artificial intelligence in critical care: opportunities and obstacles," *Crit Care*, vol. 28, no. 1, p. 113, Apr. 2024, doi: https://doi.org/10.1186/s13054-024-04860-z.

[23] P. Moritz *et al.*, "Ray: a distributed framework for emerging AI applications," in *Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation*, in OSDI'18. USA: USENIX Association, 2018, pp. 561–577.

[24] Z. Liu, X. Xu, P. Qiao, and D. Li, "Acceleration for Deep Reinforcement Learning using Parallel and Distributed Computing: A Survey," *ACM Comput Surv*, vol. 57, no. 4, pp. 1–35, Apr. 2025, doi: https://doi.org/10.1145/3703453.

[25] F. Al-Turjman, "AI-powered cloud for COVID-19 and other infectious disease diagnosis," *Pers. Ubiquitous Comput.*, vol. 27, no. 3, pp. 661–664, 2023.

[26] S. Aminizadeh *et al.*, "The applications of machine learning techniques in medical data processing based on distributed computing and the Internet of Things," *Comput Methods Programs Biomed*, vol. 241, p. 107745, Feb. 2023, doi: https://doi.org/10.1016/j.cmpb.2023.107745.

[27] Z. Xue *et al.*, "A Resource-Constrained and Privacy-Preserving Edge-Computing-Enabled Clinical Decision System: A Federated Reinforcement Learning Approach," *IEEE Internet Things J*, vol. 8, no. 11, pp. 9122–9138, Jun. 2021, doi: https://doi.org/10.1109/JIOT.2021.3057653.
[28] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci Data*, vol. 3, no. 1, pp. 1–9, 2016.
[29] M. Pinsky, A. Dubrawski, and G. Clermont, "Intelligent Clinical Decision Support," *Sensors*, vol. 22, no. 4, p. 1408, Feb. 2022, doi: https://doi.org/10.3390/s22041408.
[30] S. Helman *et al.*, "Engaging Multidisciplinary Clinical Users in the Design of an Artificial Intelligence–Powered Graphical User Interface for Intensive Care Unit Instability Decision Support," *Appl Clin Inform*, vol. 14, no. 04, pp. 789–802, Aug. 2023, doi: https://doi.org/10.1055/s-0043-1775565.