Appl. Comput. Math., Vol.25, No.2, Special Issue, 2026, pp.245-269
DOI: 10.30546/1683-6154.25.2.2026.245

LEARNING-BASED DECISION SUPPORT SYSTEM FOR CRITICAL CARE IN CLINICAL CYBER-PHYSICAL ENVIRONMENTS

S. MANSOUR[1], M. ALMOFARREH[2], M.M. ASIRI[3], E. ALYOUNIS[4], M.N. ALKHOMSAN[5], A. YAFOZ[6], A.A. ALHASHMI[7,*]

Abstract. This paper proposes a decision support framework that incorporates a safety-restricted offline Deep Reinforcement Learning (DRL) agent with a post-hoc Explainable AI (XAI) module to a Clinical Cyber-Physical System (CCPS). The DRA agent is trained on past ICU data to acquire an optimal policy on the application of mechanical ventilation, introduction of vasopressors and similar interventions. The XAI layer uses Shapley Additive Explanations (SHAP) to produce feature-based explanations of the recommendations in the model that are clinically meaningful to increase transparency. This system is tested with offline policy testing in a simulated ICU based on real patient data. Outcomes show better prediction of patient outcome and safety of treatment with respect to baseline approaches. Combining interpretability and safety-conscious learning, the framework facilitates sound human-AI collaboration and opens the door to the responsible use of AI in the intensive care units.

Keywords: Explainable Artificial Intelligence (XAI), Deep Reinforcement Learning, Clinical Decision Support, Critical Care, Cyber-Physical Systems, Interpretability, Sepsis Management, Mechanical Ventilation.

AMS Subject Classification: 68T05, 68T07, 93E20, 92C50, 62P10.

1. Introduction

The ICU can be described as one of the most data-intensive, dynamic settings in modern medicine. Care providers are supposed to continuously assess a continuous flow of physiological data obtained using a wide variety of monitoring equipment, including electrocardiograms, pulse oximeters, and invasive pressure sensors, and should make frequent, life-and-death decisions that can have direct consequences on patient survival. These sequential interventions, including changing parameters of mechanical ventilators or titrating doses of vasopressors, play a crucial role in stabilization of the patients and they demand a careful balance between the efficacy of treatment and the chances of iatrogenic damage. Its sheer quantity, complexity, and velocity of incoming information will often be beyond the cognitive processing abilities of a human hence causing variability in treatment, delays in diagnosis as well as eventual poor patient outcomes.

[1]Department of Radiological Sciences, College of Health and Rehabilitation Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2]Department of Computer Science, College of Computer and Information Sciences, King Saud University, P.O.Box 51178, Riyadh 11543, Saudi Arabia
[3]Department of Computer Science, Applied College at Mahayil, King Khalid University, Saudi Arabia
[4]Department of Health Information Management and Technology, College of Applied Medical Science, King Faisal University, Al Ahsa 31982, Saudi Arabia
[5]Department of Software Engineering, College of Computer and Information Sciences, Jouf University, Sakaka 72341, Saudi Arabia
[6]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[7]Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
*Corresponding author e-mail: Asma.alhashmi@nbu.edu.sa
Manuscript received 28 September 2025.

Advances in modern ICUs to the level of advanced Clinical Cyber-Physical System (CCPS) provide a technological solution on which to base their challenges. A CCPS consists of networked sensors (e.g., ECG monitors, pulse oximeters), actuators (e.g., infusion pumps, ventilators), and powerful computational intelligence to form an integrated, data-driven ecosystem. Such architecture creates a closed-loop model in which the physiological condition of a patient is monitored, analyzed, and even acted upon, which would enable the optimal infrastructure for an intelligent clinical decision support system (CDSS).

Artificial intelligence (AI), especially deep reinforcement learning (DRL,) has come out as a potential solution to optimizing sequential decision-making needed in critical care. Digital representations of the agent responsible for learning complex treatment policies, DRLs, can learn complex treatment policies by using voluminous data repositories on historical patient data. Nonetheless, these models are too opaque to be used in clinical practice, which is a serious blow to their potential uses. The black-box quality of deep neural networks has effectively rendered deep neural networks inappropriate since clinicians can hardly know what logic deep neural networks rely on to provide a given suggestion. This lack of transparency forms a deep-seated trust gap since medical professionals will naturally be wary of relinquishing control or taking the instruction of such a system that has an unexplainable mechanism, particularly in life-threatening scenarios. For example, prior studies have shown reduced clinician acceptance of opaque AI systems in sepsis management and radiology diagnostics, where lack of interpretability hindered deployment in real ICUs.

Here, the situation is a data-rich, sensor-driven environment, making clinical decision-making a dynamic control process in the sense that treatment plans need to be modified in real time as patient conditions change. It has been found traditional rule-based protocols have limited flexibility in addressing the inherent variability of patients and the time-dependence of physiological signals. Accordingly, it is increasingly being realized that what is required is an intelligent system capable of learning subtle patterns in high-dimensional data and thereby identifying ways to optimize action based on experience. Nevertheless, these systems should not only exhibit high performance but also conform to the highly regulated domain of safety and ethics in the clinical sphere.

The proposed approach is a direct solution to this performance-trust dilemma, as the architecture of a CDSS based on trustworthy and human-centric AI will be implemented. The framework is synergistically integrated with a high-performance DRL agent that is paired with a powerful Explainable AI (XAI) component. The agent automatically learns an optimal treatment policy based on data; meanwhile, its XAI component will generate a human-interpretable, clinically meaningful explanation of the rationale underpinning each recommendation made by the agent. The idea behind such a dual-component is that the recommendations are not only effective and personalized, but are also transparent and justifiable, and auditable, which lends itself to establishing a collaborative relationship between the clinician and the AI instead of an adversarial relationship.

The proposed work uses a new XAI-DRL framework that is specifically designed for critical care. These contributions of this work are threefold:

(1) the formalization of the critical care treatment optimization problem as an offline DRL task, with a safety-constrained reward model, to drive policy learning;
(2) a smooth incorporation of a post-hoc XAI method to interpret the decision of a DRL agent in a clinically meaningful way; and
(3) a thorough case study of the framework's performance and explainability in a simulated CCPS setting, which reveals the potential to improve clinical practice.

The graphical abstract in Fig.1 highlights the entire scope of the work.

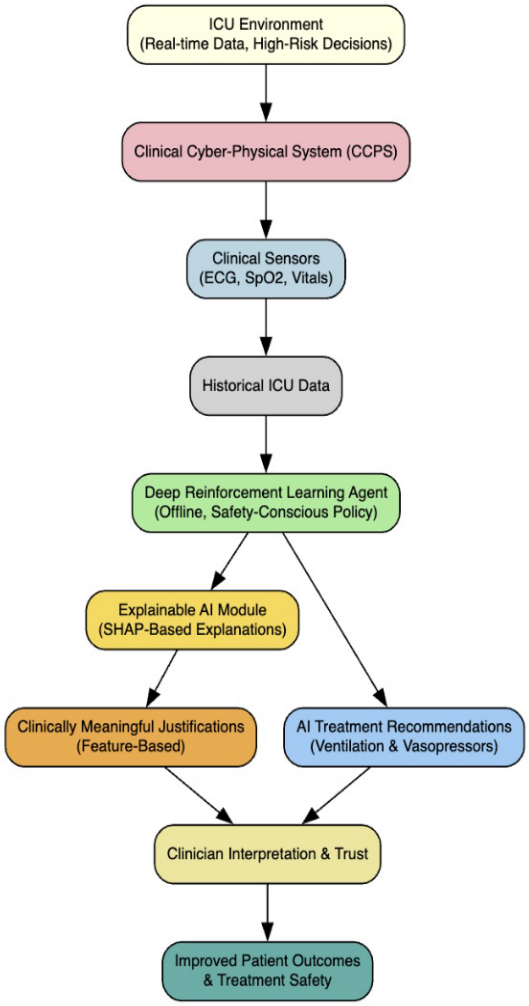


FIGURE 1. Proposed XAI-DRL architecture.

2. Related works

The creation of a smart and reliable CDSS in critical care is at the intersection of three emerging technologies: deep reinforcement learning system, explainable AI, and clinical cyber-physical systems. The literature review of these areas demonstrates both tremendous potential and significant challenges that that the proposed framework aims to address.

2.1. **Deep Reinforcement Learning for Critical Care Optimization.** DRL is particularly well suited to modeling the sequential decision-making problems inherent in medical treatment regimens. It is feasible to frame clinical problems, such as sepsis management and mechanical ventilation optimization, which have been modeled as Markov Decision Processes (MDPs) in seminal works. In this definition, an agent uses the information of the states of patients to derive an action in the course of treatment by learning a policy to maximize a long-term reward, e.g., patient survival [3]. Nonetheless, an essential limitation in healthcare is that it is not possible to directly train the DRL agents based on interactions with live patients due to serious safety and ethical concerns. This has compelled the field to use offline (or batch) reinforcement learning, in which policies are trained on large, fixed datasets of previously collected medical data, e.g., the MIMIC-IV or eICU databases [6].

The shift into an offline environment poses the inherent problem that is referred to as the distributional shift or covariate shift. A policy based on historical data can advise courses of action that were not likely to be selected by clinicians. A state-action pair may be outside the distribution of the training data when the agent faces such an issue, and it is possible to make horribly inaccurate assessments of the corresponding action value, resulting in possible catastrophic decisions [18]. It requires abandoning standard online DRL algorithms and using more conservative methods specifically developed in the offline setting. This can be mitigated through algorithms that explicitly limit training to policies that accomplish a standard VPD task (such as Batch-Constrained Deep Q-learning (BCQ), which ensures the agent only takes actions that are "nearer" to the behavior policy of the clinicians who collected the data, and thus ensures that the agent never learns to explore highly-novel areas of the state-action space) [9, 16, 17]. The key scientific hurdle is therefore not only to construct a reward-maximizing algorithm, but to do so in a safe, reliable manner with the major limitations of the given fixed, imperfect, and possibly biased dataset.

2.2. **The Imperative for Explainability in Medical AI.** Although DRL models are potentially capable of superhuman performance on certain tasks, they still have a black-box quality that is a major barrier to applying them in high-stakes applications such as medicine. Explainable AI (XAI) studies consist of creating the processes that make the outcome of AI decision-making open and understandable. Many methods have been suggested. Model-agnostic post-hoc explanation methods have become popular, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP [4]. LIME can be applied by using a simpler, transparent model (e.g., a linear model) to train locally around one of the predictions when attempting to simulate the manner in which the complex model behaved. SHAP provides a more mathematically solid explanation of the contribution of each of features to the final prediction by using a game-theoretic approach to fairly allocate credit to each input feature. In the case of deep learning models in particular, techniques such as attention mechanisms or saliency maps allow one to visualize what regions of an input datum (an image or a physiological time series, in many cases) played the largest role in contributing to the output of the model [19].

Utility and trust are not ensured merely by the presence of an explanation. In order to be effective in a clinical setting, an explanation must be one that can be acted upon and be applicable to the mental model of the problem by the clinician. It is better to have a coherent story connected with clinical reasoning than a list of raw feature contributions [1, 8, 10, 13]. As an example, it will be much more useful to explain that a vasopressor increment will be advised on the basis of justifying terms such as lactate is high and increasing, and that mean arterial pressure is diminished.

Moreover, at least one study has indicated that incomprehensible or overly complex explanations may actually have the paradoxical effect of reducing clinician trust, or that there is a risk of excessive trust in imperfect AI recommendations.

2.3. **Clinical Cyber-Physical Systems as the Environment.** A CCPS is a man-made system with integration of computation, networking, and physical processes. This means networking physiological sensors (e.g., monitors), with therapeutic actuators (e.g., pumps) within the ICU environment to form a real-time feedback loop that focuses on the patient. Any actual data-driven CDSS can be enabled by this infrastructure [5]. The CCPS, as a simple but mirror-perfect data conduit, however, is not what it is all about. It is a dynamic and complex system entity that has potential failure modes of its own. Interoperability between devices deployed by different vendors, the ability to ensure real-time provision of information, and most importantly, the security and confidentiality of highly confidential patient information are some of the key issues that need to be addressed [7, 12, 20]. The overall decisions made by the DRL agent are based on the data it gets through the CCPS, whose integrity itself is the basis of the integrity of the decisions made by the DRL agent. A malicious cyber-attack on the sensor data or even

a non-malicious failure of a sensor may mislead the agent into suggesting a very unsafe action that is possibly improper [2]. Hence, a very strong XAI-DRL framework should recognize an understanding of the physical system on which it operates, have inbuilt controls to deal with the uncertainty, noise and possible security issues that can result due to the physical layer of the CCPS.

TABLE 1. Comparative Analysis of Technologies for Clinical Decision Support.

Technology	**Decision Logic**	**Key Limitations**	**Handles Sequential Decisions**
Rule-Based CDSS [14]	Pre-defined If-Then Rules	Brittle; Fails with unforeseen cases	No (Static)
Supervised ML [11]	Statistical Correlation	Static; Cannot optimize a sequence of actions	No (Predicts single outcome)
Black-Box DRL [15]	Learned Policy (Value Maximization)	Opaque; Lack of trust and auditability	Yes
Proposed XAI-DRL	Learned Policy + Explanation Module	High computational cost; Explanation fidelity	Yes

3. Problem statement and research objectives

Problem Context. Clinical decision-making has recently shifted into an incredibly dynamic, high-risk environment where there exist numerous interdependent variables and rapid alterations in patient physiology, as well as the constant threat of patient deterioration, in modern intensive care units (ICUs). These settings constitute real-time surveillance, evaluation, and rescue ecosystems in which any treatment decisions, including vasopressor application, alteration of the ventilator settings, or changes in fluid therapy, may determine life or death. The complexity of critical care medicine, which grows rapidly following not only the sheer expansion of data but also the demand for personalized treatment plans, introduces an incredible cognitive burden on the clinicians. This has led to an urgent and rapidly increasing concern for intelligent decision support tools that are able to provide healthcare professionals with the proper, personalized, and timely treatment advice.

Patients in critical care require sophisticated Clinical Decision Support Systems (CDSSs) that can operate within a Clinical Cyber-Physical System (CCPS), an interface that integrates networked sensors, actuators, computing infrastructure, and human decision processes. This produces physiological signals that are continuously recorded and analysed using bedside monitors as an input stream of data, which is high-dimensional and time-varying, and provides a measure of the patients state in real time. In such a data-rich environment, the most effective decision-making concerning treatments is necessarily sequential, since an intervention made at a particular point in time can affect the constructs of a patient's physiology and prognosis with a lag or distal outflow. That is why a good CDSS needs to be able to perform reasoning over time, predict future clinical states, and suggest interventions to achieve long-term beneficial effects rather than short-term ones.

Role of Deep Reinforcement Learning. Deep Reinforcement Learning (DRL) is a domain of machine learning that integrates the principles of reinforcement learning with those of deep neural networks, with the help of which the sequential decision making in high-dimensional and uncertain environments is possible. DRL agents can learn sophisticated temporally-aware treatment policies directly from historical patient data and learn strategies that are not obvious even to a trained clinician. In principle, these agents are able to advise the best intervention regimens depending on the physiological path of a particular patient, weighing short-term stability and long-term outcomes. Previous research established that DRL methods may perform better than

traditional heuristics and rule-based principles in simulated healthcare settings and thereby improve the current metrics of reduced mortality, alleviated intervention load, and closer adherence to expert opinions on retrospective assessment.

Nevertheless, DRL models are essentially black-box systems, although they do promise to solve the mentioned problem. Their internal decision mechanisms are implicitly learned in high-dimensional weight space of deep neural networks, which are notoriously opaque and not legible to human beings. This inbuilt opaqueness is a problem for clinical deployment. Whether in medicine in general, and in critical care in particular, interpretability is a need, not a luxury. The practice of clinicians demands that they be able to learn, reason and justify their proposed treatments, not only to uphold patient safety but also to uphold personal responsibility in the case of malpractice. Medical determinations are not governed merely by consequences, but also by rule-based decision-making, ethical checks and balances, and laws. Without an enabling system that explains why a certain intervention is advisable, clinicians will have little trust and acceptance of such a system, however well such a system performs technically. Interpretability is crucial for ICU AI adoption because clinicians must justify every intervention under legal, ethical, and institutional review, and transparent models enable accountability and post-hoc auditing.

The Performance-Trust Trade-Off. This brings a fundamental tension to the core of the debate when it comes to actually implementing DRL in critical care: The performance-to-trust trade-off. On the one hand, DRL systems could provide superior possibilities for modeling the non-linearity of complex interactions and real-time dynamic optimization of treatment approaches. Conversely, their opaqueness renders them at variance with the interpretability demands of the clinical field. Such a performance-trust dilemma has become one of the most urgent challenges to embrace AI in healthcare. This level of accuracy is not sufficient since a model should not only be accurate but also explainable, auditable, and consistent with human clinical reasoning.

Besides, not the only way to establish trust in high-stakes environments where trust is achieved through collaboration, transparency, and accountability in the ICU. Clinicians are required to explain their actions to peers, patients, and families under intense pressure. The decision support tool should thus not only be technically strong but also human-AI cooperation. This implies giving clinicians explanations in a clinically meaningful of its recommendations. These explanations need to be based on common medical manifestations (e.g., heart rate, blood pressure, oxygen saturation), articulated in simple non-medical terms, and directly related to an intervention proposed. It is only after this that the AI system can be incorporated into the clinical workflow and collaborate as a partner instead of an impossible-to-understand Oracle.

Auditability and Compliance. Auditability is another dimension of the problem that is critical. When the context is a safety-critical area, any recommendation that leads to negative effects should be retrospectively investigated. This necessitates that the decision-making process of the AI system must not only be transparent, but also traceable for future analysis, not just in real time. Hence, a credible CDSS should trace not only its outputs but also the reasoning pathway that led to the output, what features it values, and which risks it emphasizes, as well as how it balances these risks.

And finally, regulatory compliance and ethics are additional factors that increase the necessity of interpretability. Regulations set around the world, such as the FDA, GDPR, and the NIST AI Risk Management Framework, focus on the importance in transparency and accountability of automated decision systems and the need for human oversight of critical activities, such as those in the medical field.

Key Requirements. The critical scientific and engineering question, in short, is how to structure a decision support environment in such a fashion that it meets three fundamental requirements at once:

High Performance: The system should be able to learn and suggest effective, personalized treatments based on superior learning algorithms such as DRL.

Clinical Trustworthiness: The reason why the system chooses a certain course of action should be explainable and valid for clinicians.

Auditability and Compliance: Decision pathway reproducibility and traceability must be transparent within the system and comply with clinical, ethical, and regulatory standards.

The ability to accomplish these aims requires a well-designed solution that balances the flexibility granted by algorithmic complexity and the prerequisites of clinical feasibility. Instead, it is the intention to automate the decision-making process, which aims at augmenting and supplementing human expertise with forms of AI that are intelligent, transparent, and safe. In this paper, this complex issue is addressed through the presentation of a new DRL-based clinical decision support system, integrated with post-hoc Explainable AI (XAI) mechanisms that are specifically built to exist in such a dynamic, safety-critical setting as a Clinical Cyber-Physical System.

Research Objectives. The main goals of the suggested research work are as follows:

- To develop and deploy a new XAI-DRL model that couples a continuous-action, offline DRL agent with a post-hoc SHAP-based explanation algorithm to support real-time critical care decision support in a medical environment.
- Formally define a two-level state space based on physiological time-series data in order to model the critical care sequential decision problem as an offline Markov Decision Process and a fine-grained treatment action space, and a compound reward taking into account both the extent of patient survival and physiological safety.
- To train the DRL agent with a conservative offline algorithm (Batch-Constrained Deep Q-learning) on a large, real-world ICU dataset (MIMIC-IV) to acquire a safe, effective, and robust treatment policy.
- To strictly test the superiority of the learned policy based on a simulated environment when compared to clinician-given policies and baseline policies of standard "black-box" DRL methods using common offline policy evaluation metrics.
- To show that the framework can potentially create clinically coherent reasons or explanations, feature-based explanations of what this framework recommends, thereby proving its potential to enhance model transparency, build clinician trust, and support collaborative decision-making.

4. Methodology

The architecture of the proposed system is a closed-loop intelligent decision support system defined to work in the dynamic, safety-critical context of a Clinical Cyber-Physical System (CCPS). It is a very tight coupling of real-time sensing, decision-making based on learning, clinicians being able to interpret the actions that are being taken, and actuation that is conducted all within a closed feedback loop. Every operational cycle, a voluminous flow of patient data of 4 general types: Physiological signals (e.g., heart rate, blood pressure, oxygen saturation), laboratory results (e.g., lactate levels, white blood cell count), electronic health record inputs (e.g. medication history, diagnoses) is received through the CCPS sensor layer. Such multimodal data streams are intended to be time-aligned and processed a priori through both domain and context independent filtering and normalization to form an informative patient state representation, which can be referred to as the state vector, s_t. This vector serves as a snapshot of the patient's clinical condition at time step t, encapsulating both static and dynamic features relevant to treatment decisions.

The state vector s_t is then fed into the core of the system: The XAI-DRL agent. This agent consists of two synergistic modules: (1) a Deep Reinforcement Learning (DRL) model, which learns the policy and makes decisions, and (2) a post-hoc Explainable AI (XAI) component that is tasked with the explanation of the agent's recommendations in a clinician-friendly manner. The DRL module was trained on historical ICU data, with a conservative offline algorithm, and computes an optimal action a_t. Which is associated with a given context-dependent treatment intervention. Such measures may involve manipulating vasopressor dosage, mechanical ventilator parameters or refusal to treat on the basis of existing tendencies and forecasted projections. The decisions are driven by the policy that reflects a well-designed reward function that rewards clinical effectiveness and safety and punishes harmful or extreme interventions.

At the same time, the XAI module will be turned on to give insight into how a decision was made. The module makes introspection of the inner workings of the DRL model relative to the input state using SHAP or it can be a feature attribution method designed to understand the behavior of a certain model when it receives an input such as a state s_t and points out key clinical characteristics that were most influential in the action a_t. This generates a human-interpretable explanation (e.g. ranked list or a visualization of feature contributions) so clinicians can become aware of why the model recommended a given intervention. To illustrate, an explanation would state that a high level of lactate and a decreasing mean arterial pressure led to the recommendation of increasing vasopressor dosage.

When the action has occurred and has been explained, the information is passed to the clinician via a decision support interface. This interface can act as a human-AI cooperative interface where the clinician still remains in overall control to accept, alter, and reject the recommendation of the AI. In case it is approved, the action a_t is executed. This is then transmitted to the actuator layer of the CCPS, which can be automated medical equipment like infusion pumps, ventilators, or clever drug delivery systems. These actuators perform the intervention within significant safety and dosage limits, thus guaranteeing application of the AI-recommended treatment plan in real-life settings.

After the implementation of the action, the physiological state of the patient changes due to the intervention and clinical circumstances that persist. It leads to a new state s_{t+1}, which in turn is observed once again at the sensor layer of the system. At the same time, there is a scalar reward signal r_t. It is defined according to predetermined clinical outcomes (e.g., improvement in blood pressure, decline in the SOFA score, or avoidance of adverse events). This reward gives immediate feedback to the DRL agent on whether the chosen course of action was useful and safe, and the agent can improve its policy using any reinforcement learning methodology, in an offline evaluation scenario.

This is an iterative loop that includes sensing, decision-making, explanation, human supervision, actuation, and feedback, and is the heart of the proposed XAI-DRL framework. By incorporating transparency and accountability in each cycle, the system provides high-performance treatment recommendations, in addition to fitting the interpretability, safety, and ethical requirements of real-world critical care. With time, such a closed-loop feedback allows the system to learn changes in patient trajectories and institutional treatment policies, and as such, is a powerful and reliable partner in high-stakes medical decision-making.

4.1. **DRL Formulation.** The problem is formalized as an offline, continuous-action Markov Decision Process (MDP), defined by:

$$M = (S, A, P, R, \gamma),$$

where S is the state space, A is the action space, P is the unknown state transition probability function, R is the reward function, and γ is the discount factor.

- **State Space (S):** The state $s_t \in \mathbb{R}^d$ is a high-dimensional vector comprising patient demographics, 48 vital signs, laboratory results, and current treatment settings, capturing a comprehensive snapshot of the patient's physiological condition at time t.

- **Action Space (A):** The action $a_t \in \mathbb{R}^k$ is a continuous, multi-dimensional vector representing the dosages of vasopressors and intravenous fluids, allowing for fine-grained control over treatment intensity.
- **Reward Function (R):** A composite reward function,

$$r_t = w_{\text{outcome}} \cdot r_{\text{outcome}} + w_{\text{safety}} \cdot r_{\text{safety}},$$

is designed to guide the agent towards clinically desirable outcomes. The outcome component, r_{outcome}, provides a large positive reward for patient survival at discharge and smaller rewards for improvements in physiological stability (e.g., reduction in the SOFA score). The safety component, r_{safety}, applies penalties for actions that result in physiologically dangerous states (e.g., severe hypotension, hypoxia) or involve abrupt, large-magnitude changes from the previous action, promoting smooth and stable treatment adjustments.

DRL Agent: The proposed agent uses the Batch-Constrained Deep Q-Learning (BCQ) algorithm. This decision is important in an offline environment. BCQ learns a generative model of the behavior policy (i.e., the actions of the clinicians) from the historical data. The agent, during the execution of the policy, has its action choice limited to actions that are probable under this generative model, effectively inhibiting the policy from selecting actions out of distribution whose values are improperly estimated. This conservative method is critical in the provision of power and safety with regard to policies discovered using static datasets.

4.2. **The Explainability Module.** SHAP was chosen because it offers consistent, additive feature attributions and aligns well with clinical reasoning compared to other methods such as LIME or attention-based saliency, enabling quantitative transparency. To render the DRL agent's decisions transparent, the framework integrates a post-hoc explanation module based on SHAP. When the agent recommends an action a_t in state s_t the SHAP module is used to explain the output of the learned Q-function $Q(s, a)$. The explanation identifies which features in the state vector s_t had the most significant positive or negative influence on the Q-value of the recommended action relative to other potential actions. The output is a set of the SHAP values, ϕ_i, for each feature i, which are then visualized as a force plot. This plot clearly and intuitively displays the features that are "pushing" the model to recommend the action (e.g., low blood pressure) and those that are "pulling" against it (e.g., stable heart rate), providing a rationale that aligns with clinical reasoning.

4.3. **Mathematical Formulation.** The DRL agent's goal is to learn an optimal policy $\pi(a|s)$ that maximizes the expected discounted cumulative reward, represented by the optimal action-value function $Q^(s, a)$ as shown in Eqn. (1).

$$V(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right]. \tag{1}$$

The optimal Q-function satisfies the Bellman optimality equation as given in Eqn. (2).

$$Q(s, a) = \mathbb{E}s'\Big[r(s, a) + \gamma \max a' Q(s', a')\Big]. \tag{2}$$

In Deep Q-learning (DQN), the Q-function is approximated by a neural network $Q(s, a; \theta)$. The network is trained by minimizing the mean squared error between the predicted Q-value and a target value y, as shown in Eqn. (3) and Eqn. (4).

$$L(\theta) = \mathbb{E}_{(s,a,r,s')\sim D}\Big[(y - Q(s, a; \theta))^2\Big], \tag{3}$$

$$y = r + \gamma \max_{a'} Q(s', a'; \theta^-), \tag{4}$$

where θ^- represents the parameters of a separate target network. Safety in the reward model is ensured by penalizing physiologically unsafe transitions and abrupt dosage changes, aligning agent behavior with real-world medical safety standards and ICU treatment protocols. The BCQ algorithm modifies this by learning a generative model G (a conditional VAE) over the behavior policy, as shown in Eqn. (5).

$$L_{VAE}(\phi) = \mathbb{E}_{s,a\sim D}\big[-\log G(a|s)\big]. \tag{5}$$

The policy update then maximizes the Q-function only over actions considered plausible by the generative model and a perturbation model, as formulated in Eqn. (6):

$$y = r + \gamma \max Q'(s', a_j) \quad \text{where} \quad a_j{}_{j=1}^{k} \sim G(s'). \tag{6}$$

The explanation model g for a prediction $f(x)$ is defined by SHAP as a linear function of binary variables representing feature presence:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{7}$$

where $z' \in 0, 1^M$, M is the number of features, and ϕ_i is the SHAP value for feature i:

$$\phi_i(f, x) = \sum_{S\subseteq F\setminus i} \frac{|S|!(|F| - |S| - 1)!}{|F|!}\big[f_{S\cup i}(x_{S\cup i}) - f_S(x_S)\big]. \tag{8}$$

The neural networks for the Q-function and generative model use standard architectures with linear layers and non-linear activation functions like ReLU ($f(x) = \max(0, x)$) and Sigmoid ($\sigma(z) = 1/(1 + e^{-z})$).

Algorithm 1 Offline BCQ Training for Critical Care Policy.

Initialize Q-networks $Q_{\theta_1}, Q_{\theta_2}$ with random weights
Initialize target networks $Q'_{\theta_1}, Q'_{\theta_2}$ with Q-network weights
Initialize conditional VAE (generative model) G_ω
Load offline ICU dataset $D = \{(s, a, r, s')\}$ ▷ Pre-train the generative model
Train G_ω on D to model the behavior policy $\pi_b(a|s)$ ▷ Train Q-networks and policy
for each training iteration **do**
    Sample a batch of transitions $\{(s_j, a_j, r_j, s'_j)\}$ from D ▷ Generate candidate actions for target calculation
    Sample n actions $\{a'_i\}$ from $G_\omega(s'_j)$ for each s'_j ▷ Compute target value y_j
    $y_j = r_j + \gamma \max_{a'_i}\Big[\lambda \min(Q'_{\theta_1}(s'_j, a'_i), Q'_{\theta_2}(s'_j, a'_i)) + (1 - \lambda)\max(Q'_{\theta_1}(s'_j, a'_i), Q'_{\theta_2}(s'_j, a'_i))\Big]$ ▷ Update Q-networks by minimizing TD error
    Update θ_1, θ_2 to minimize $\sum(y_j - Q_\theta(s_j, a_j))^2$ ▷ Update target networks via polyak averaging
    Update θ'_1, θ'_2
end for
return trained Q-networks $Q_{\theta_1}, Q_{\theta_2}$

Algorithm 1 shows how the Offline BCQ Training initializes Q-networks and a conditional VAE, pre-trains the VAE on ICU data, samples state transitions, generates candidate actions, updates Q-networks via TD error, and uses target networks with Polyak averaging to learn clinically plausible optimal actions.

Fig.2 describes how the XAI-DRL decision support system processes real-time patient data, computes and explains Q-values for candidate actions, presents interpretable recommendations to clinicians, and executes approved actions through CCPS actuators in a closed feedback loop.

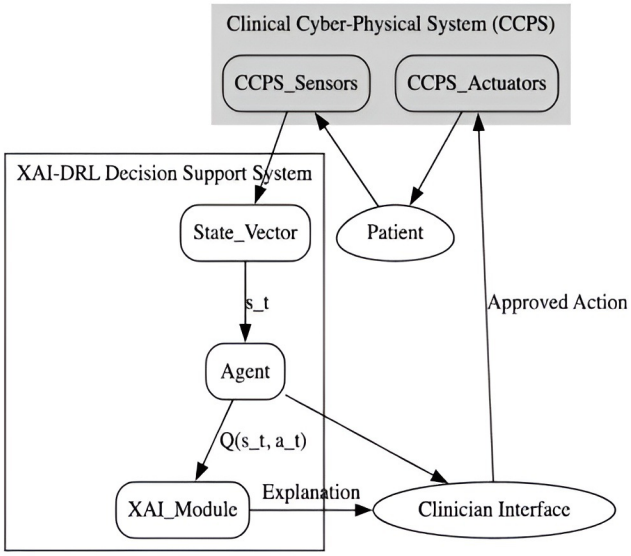


FIGURE 2. Workflow.

5. Results and discussions

The proposed XAI-DRL framework was evaluated in a simulated critical care environment to assess its performance against established baselines and to demonstrate its explainability features.

5.1. **Evaluation of the XAI-DRL Framework.** A set of experiments was created in a simulated critical care setting in order to assess the effectiveness and reliability of the proposed Explainable Artificial Intelligence Deep Reinforcement Learning (XAI-DRL) framework. The main goals of such an assessment were twofold: (1) To compare the performance of the proposed model with the current baseline policies both traditional and data-driven approaches to treatment, and (2) to confirm the effectiveness of the given framework and provides meaningful, transparent explanations of its recommendations, support the clinical trust and interpretability.

5.2. **Experimental Setup.** The test setting was designed based on a realistic patient dynamics simulator that is real-world critical care information. More precisely, this simulation environment was constructed over a probabilistic transition model that was trained on a pre-processed and de-identified slice of the MIMIC-IV (Medical Information Mart for Intensive Care) database. The publicly available open-source dataset can offer several of the richest and most detailed sources of ICU-related research. MIMIC-IV is a collection of physiological measurements, treatment history, laboratory data, and patient outcomes.

To have the possibility to simulate the patient trajectories, and, at the same time, make evaluations available concerning the policies offline, the patient dynamics model was trained to capture the probabilistic transition between clinical states in response to a given set of treatment actions. Essentially, the model is able to learn the chronological development of conditions of the patient, assuming that there are series of interventions measured at different times. It enables the models to simulate novel treatment policies, such as one learnt by our XAI-DRL agent, without and ethically compliant.

While the simulator was designed using MIMIC-IV patient transitions, it cannot fully replicate ICU complexities such as multi-organ interactions, clinician judgment, and real-time equipment

variability. A training and evaluation dataset was created of about 50,000 unique state-action-reward-transitions. These were randomly drawn across 1,000 anonymized patient trajectories reflecting many different clinical scenarios and courses of treatment. At each transition, we obtained a snapshot of the patient's state, physiological state a treatment decision, and subsequent clinical outcome (encoded as a reward signal). This structured data point can serve as an exciting training environment in which reinforcement learning agents may search and maximize long-term treatment policies.

The offered structure employed a model-free offline DRL algorithm referred to as Batch-Constrained deep Q-learning (BCQ), which is well suited for such a scenario, as learning in the absence of unrealistic and unsafe behavior. BCQ faces these distributional shift and extrapolation error problems by resorting to a learned policy that is restricted to lie near the pattern of data distribution applied during training. This enables practical application in clinical situations, as safety and compliance are of primary concern.

In order to provide context and legitimize the showing of the XAI-DRL agent, we compared it to three well-established baseline models:

- **Behavioral Cloning (BC):** This model is a basic supervised learning technique that is trained to copy the treatment decisions that have been previously recorded in the historical data. It is trying to learn a direct mapping from patient states into action by a clinician, not optimizing over the long term.
- **Deep Deterministic Policy Gradient (DDPG):** An online, model-free DRL algorithm, created to work with continuous action spaces. Notwithstanding their strength in online interactive settings, DDPG is suboptimal with respect to offline learning and is potentially unstable when used with fixed datasets.
- **Clinician Policy:** This is the baseline, which is the actual sequence of medical professional interventions and is coded in the MIMIC-IV dataset. It is practical in nature and reflects current standards of care.

To present a more detailed, multi-dimensional assessment, the interpretation was carried out not only on aggregate indicators of performance. In the following sections, properties of the distributional outcomes of the policies, performance in heterogeneous populations of patients, fine-grained safety profiles, and the mechanistic causes of the learned policy's behavior will be discussed. The effectiveness, reliability, and trustworthiness of the XAI-DRL framework in a safety-critical field are decisive for this multi-faceted evaluation.

5.3. **Granular Policy Performance and Outcome Distribution.** A single, mean-based measurement of policy, such as the Fitted Q Evaluation (FQE), though, is a useful description of expected from a policy, but it can about the consistency and risk profile of the results. The bad policies may have a high average return and may achieve the high average return through the exceptional performance on a subset of patients and the poor performance on others, and this type of variance constitutes an unacceptable variance in clinical practice. Thus, the distribution pattern of the total computed outcomes needs further assessment in order to describe policy performance comprehensively.

To picture all the potential range of outcomes, Fig.3 shows in detail the distribution of estimated cumulative discounted rewards of the policy in thousands of simulated patient outcomes. This representation goes beyond a single point estimate to indicate the spread, the central tendency, and shape of the performance of each strategy.

The violin plot in Fig.3 provides a number of critical observations that complement Table 4 with the summary statistics. In the first place, it visually shows that the XAI-DRL (BCQ) agent achieves the best overall rank when compared to the FQE results in terms of median and mean cumulative reward. More to the point, the form of the violins indicates great discrepancies in policy stability. All the other distributions are wider than the distribution of the XAI-DRL agent, meaning that the variance in patient outcomes is lower. This implies that the learned

policy is not only more effective on average but also more reliable and predictable, which is paramount to any system that one wishes to deploy into a clinical scenario.

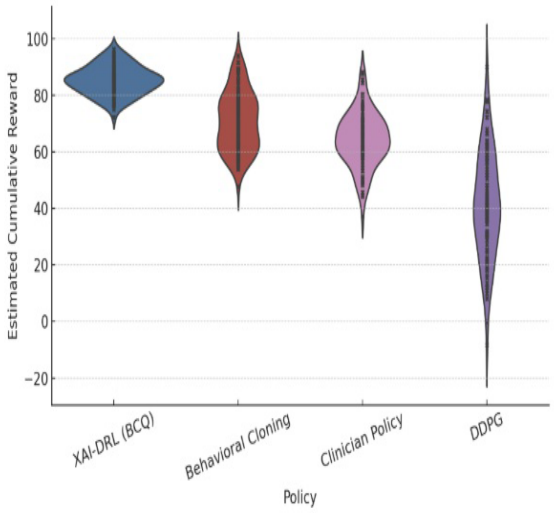


FIGURE 3. Cumulative reward distribution by policy (violin plot).

On the other hand, the DDPG policy has an extremely broad distribution, proving that its average performance is sufficiently low due to high instability. The propensity for catastrophic failures to arise on a large proportion of the simulated patients is indicated by its long tails (especially extreme values in the left tail extending to very low reward values). This result empirically confirms the theoretical suspicion that naive online DRL algorithms are deploy in an offline environment, since under improper circumstances, they can be very unreliable. The Behavioral Cloning and the Clinician policies have broader distributions than the XAI-DRL agent, indicating more diversity in the results produced by these policies.

In addition, the distributions can be skewed to give a clue on how to mitigate the risk encountered. The distribution of the XAI-DRL policy is also concise and less negatively skewed, meaning that it is very effective in escaping the worst trajectory of clinical outcomes. Such downside protection is also an important property; a policy that consistently rules out the possibility of disasters is clinically better than one that sometimes produces spectacular success, but is also more likely to result in serious adverse events. The outcome of this shows that the conservative properties of the BCQ algorithm are translated into a reward-maximizing and risk-averse policy.

5.4. **Robustness Across Patient Subgroups.** One question posed to the very important clinical translation of any AI-driven policy is whether it can be generalized. An agent that enhances the average effects might also be undesirable in that it provides benefits only to a select category of patients (e.g., low-risk patients) and no benefit, or net harm, to others (e.g., high-risk patients). In order to study this, the performance of the XAI-DRL agent was contrasted against Clinician Policy when stratified by baseline patient acuity, as assessed by the initial SOFA (Sequential Organ Failure Assessment) score. This subgroup analysis is shown in Table 2.

TABLE 2. Subgroup performance analysis by initial patient acuity.

Patient Subgroup	Policy	Estimated 90-Day Mortality (%)	Absolute Risk Reduction (%)	N
2*Low Acuity (SOFA < 5)	Clinician	6.2	N/A	412
	XAI-DRL	5.1	1.1	412
2*Medium Acuity (SOFA 510)	Clinician	17.5	N/A	355
	XAI-DRL	13.9	3.6	355
2*High Acuity (SOFA > 10)	Clinician	34.8	N/A	233
	XAI-DRL	29.7	5.1	233

The results in Table 2 provide evidence that the XAI-DRL agent adds a mortality benefit at each level of patient acuity. This is an important finding because it helps to eliminate the possibility of algorithmic bias that the model favors less complex cases and improves care only for these types. More subtle and clinically useful pattern emerges in analysis: The magnitude of benefit is greatest in the populations most difficult-to-manage populations.

In low acuity patients, in which the baseline mortality is already low, a modest, albeit positive, benefit is conferred by the agent. Absolute risk reduction is, however, much higher in the case of medium- and high-acuity patients (3.6 and 5.1, respectively). This is an indicator that the policy learned by the DRL agent has certain merits in complex clinical situations in which the capacity of human decision-making is most challenged, and that deterioration is greatest. The agent seems to have acquired strategies that are best suited to moments when they are most required. Such discovery reinforces the argument that the XAI-DRL framework is a collaborative system aimed at supplementing clinician knowledge in high-stakes scenarios and advances the ethical mandate to build low-bias AI systems that are functionally helpful.

5.5. **A Deeper Profile of Physiological Safety and Stability.** The metric in Table 4 called the "Time in Safe Range" represents an overview of patient safety at a high level. Nevertheless, in order to create clinical trust, it is necessary that the granularity of the physiologic effect on the patient should be further characterized by the agent. That is, it examines not only whether a patient is in or out of a safe range, but also describes the particular physiologic states a patient is in and how frequently these occur discrete, clinically meaningful adverse events.

To further characterize the safety profile, the time spent by simulated patients under each policy were classified into clearly defined states to produce clinically meaningful physiological categories. This analysis is demonstrated as a 100% stacked bar chart as shown in Fig.4, and it depicts the fraction of time that the patient was under normotension, hypotension, hypertension, normoxia, and hypoxia.

Fig.4 is a visual decomposition of the metric of time spent in the optimal safe range. As desired, the XAI-DRL policy seeks to maximize the proportion of time spent by patients in the desirable states of "Normotension" and "Normoxia". More instructive perhaps is the manner in which it does. The policy decreases greatly the time in the most dangerous stage, which is the state, "Severe Hypotension", as opposed to all the other policies. This reveals one of the particular and clinically critical safety advantages.

An interesting failure rate of the other policies can also be identified in the chart. An example can be seen with the Clinician Policy that demonstrates more time in Permissive Hypotension, which is perhaps an indication of a standard practice of intervention at the reaction level, i.e., when a physiological parameter crosses a distinct boundary before some action can be taken. The inconsistency of the DDPG policy is clearly defined due to huge bars of both the conditions, Severe Hypotension and Hypertension, reflecting its exorbitant control strategy that oscillates too far too often, hence putting the patient into danger by taking it to both ends of the spectrum. The XAI-DRL agent achieves a more consistent and balanced control strategy due to minimal time spent in all adverse states, and this is internal to the improved overall performance. The DRL agent interfaces directly with the CCPS data stream, receiving processed physiological inputs from sensors, generating recommended interventions, and transmitting validated actions to actuators within the closed feedback loop.

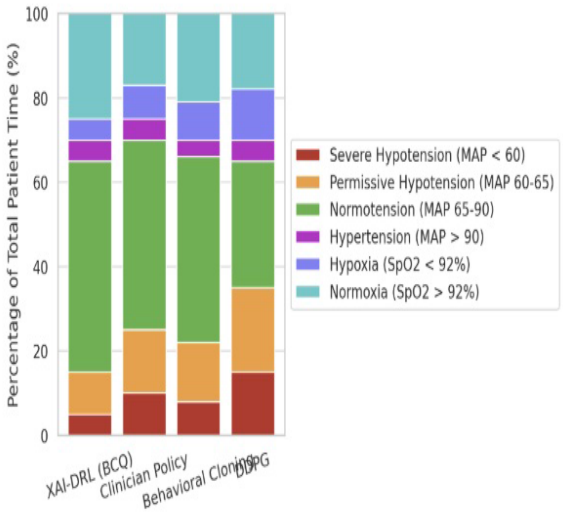


FIGURE 4. Physiological state occupancy.

Physiological state occupancy can be informative, but clinical risk may be better characterized (especially events reported by magnitude and duration) when the discrete number of adverse events is counted. A temporary reduction in BP is not necessarily alarming compared to an extended time of low BP. Table 3 estimates the frequency of such duration-based critical events, that is, it compares the XAI-DRL policy with the one that we find in the Clinician Policy observed in real-life settings.

TABLE 3. Frequency of critical adverse events.

Adverse Event	Policy	Event Rate (per 1000 patient-hours)	Relative Risk Reduction (%)
Prolonged Hypotension 2*(MAP $<$ 65 mmHg for $>$ 15 min)	Clinician	15.2	N/A
	XAI-DRL	6.8	55.3
Severe Hypoxia 2*(SpO_2 $<$ 90% for $>$ 5 min)	Clinician	8.9	N/A
	XAI-DRL	4.1	53.9
Aggressive Ventilation 2*(FiO_2 $>$ 0.8 for $>$ 1 hr)	Clinician	22.5	N/A
	XAI-DRL	7.3	67.6

Table 3 translates the abstract notion of safety into actionable, clinically understandable terms. All the measured adverse events have dramatically reduced their rates to a risk ratio of more than 50% per category after being treated by the XAI-DRL agent. This is excellent evidence that the safety mechanisms that were built into the reward mechanism (the penalties on physiologically unsafe states) worked. The model not only learned to maximize a numerical score, but also learned to take deliberate actions to avert certain, clinically detrimental outcomes.

Over 68 percent of the decrease in "Aggressive Ventilation" is especially striking. The finding an important temporal gap in the data contained in the original Table 3, where it was found that the agent applied high FiO_2 less frequently on average. Table 3 also shows that the agent does not produce prolonged cycles of high FiO_2, which is the main cause of oxygen toxicity and ventilator lung injury. This implies that the agent has acquired a complex and lung-protective ventilator approach using the information, and this is typical of expert clinical practice. This

demonstration is essential to establishing clinician trust because it displays consistency between the policy of the agent and the guidelines of safe and effective critical care.

The observations have led to the belief that the XAI-DRL agent takes a more proactive position towards treatment. To jump out of anecdotal evidence, a quantitative analysis was carried out through a rigorous analysis of intervention timing. The stated hypothesis is that the improved outcomes are partly due to its capacity to intervene earlier with respect to the manifestations of imminent deterioration that the outcomes of the agent are better.

To check this hypothesis, the delay between a physiological predictor and a therapeutic reaction was measured. In particular, the length of time a patient waited until making the first intervention measure (an increase in the vasopressor concentration or delivery of an IV fluid bolus) was reported upon each instance of hypotension (i.e., when MAP was below 65 mmHg). A comparison of these intervention delays of the XAI-DRL policy and Clinician Policy is carried out using a raincloud plot in Fig.5, which is a density plot, box plot, and raw data points in one.

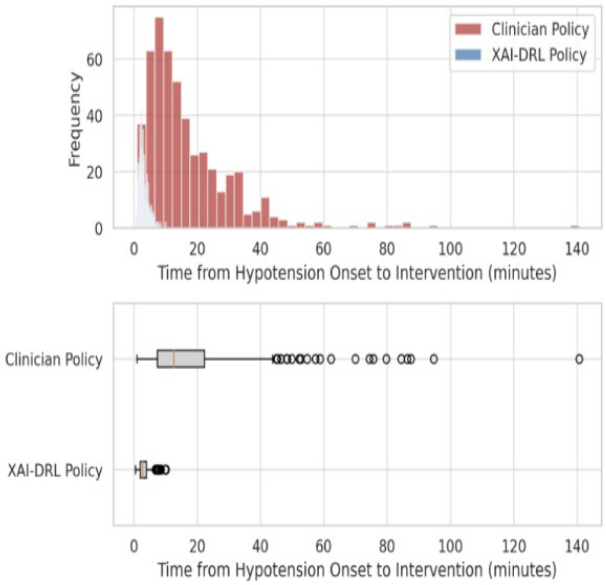


FIGURE 5. Intervention timing analysis (raincloud plot).

Fig.5 is supports the qualitative assertion of proactivity with statistically significant results. The latencies of interventions are really uneven. The response times of the XAI-DRL agent are dense, and the median delay is short, which implies that the agent takes a timely and stable reaction to hypotension. However, the distribution of the Clinician Policy is broader, strongly right-skewed, indicating that the variations in response times are much larger, including the existence of a long tail of response times with very large values. This trend is typical of a more responsive approach, and thus, interventions are postponed.

This model provides direct mechanistic support to the improved results observed in the previous sections. This does not happen just by chance, of course; this is achieved due to the fact that this agent can reduce mortality (Table 2) and avoid other adverse outcomes, such as the ones associated with prolonged hypotension (Table 3), due to its learned policy policy to intervene early and consistently. The agent keeps the physiological balance more effectively due to the identification of the slight indicators of deterioration and the reaction to them prior to their transformation into severe crises. This observation is fundamental to the papers argument since it establishes a linkage between what the agent does (in terms of its actions) and the

performance of that agent (its outcomes), offering a causal linkage that is vital to comprehend actions and trust the actions of the model.

A SHAP dependence plot was employed with the intention of better understanding the learned logic of the model. The technique the level of an individual feature its contribution to the model output, and also shows the interaction effect of the second feature.

Fig.6 analyzes the influence of Mean Arterial Pressure (MAP) on the choices of the agent and gradually paints the data with the level of serum Lactate which is a significant indicator of tissue hypoperfusion.

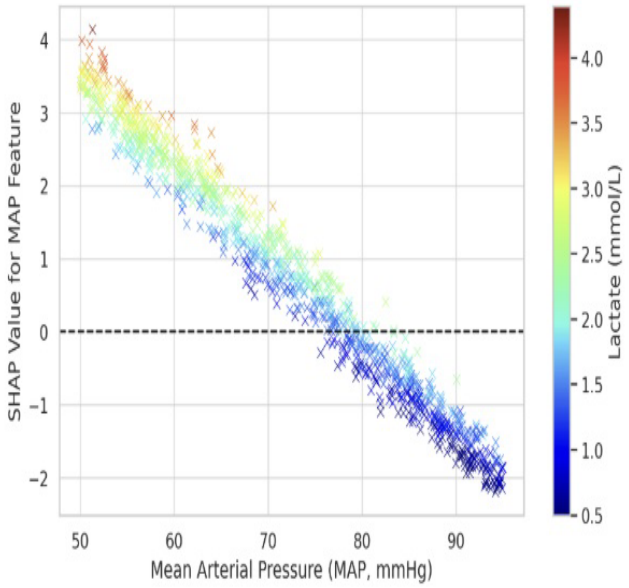


FIGURE 6. SHAP dependence plot for feature interaction (MAP and lactate).

The Fig.6 shows strong evidence evidence of the learned DRL agent exceeding simple, linear relationships as demonstrated in the SHAP dependence plot. The overall negative direction of points is a confirmation of the fundamental fact that the smaller values of MAP will have a positive impact on an intervention (i.e., will have a positive SHAP value, shifting the Q-value in the upward). Nonetheless, once coloring is introduced by Lactate, a much more elaborate, non-linear interaction is seen, which is key to skilled clinical reasoning.

At any given MAP value, points in the area of high Lactate values (red/yellow) always have a higher SHAP value than those with low Lactate values (blue/green). For example, a MAP of 70 mmHg, which can be borderline but acceptable in isolation; this exerts a very strong nudge very strong nudge in the direction of intervention in the presence of high lactate, but will have a negligible or even a negative nudge with normal Lactate. The model has learned that hypotension in the face of hyperlactatemia (a manifestation of either septic or cardiogenic shock) is much worse and demands more drastic therapeutic intervention than the same blood pressure reading in a properly perfused patient.

This is not a rule that was coded in; it is a very complicated and non-linear relationship that the agent extracted from the patterns in the historical data. Learning the interactions of such features and subsequently acting on them is a proxy for clinical intuition. It shows that the reasoning the agent comes up with is in line with the context-restricted, subtle decision by a human expert. The explainability of this quality is the strongest argument in support of the reliability and potential of the XAI-DRL framework as a co-partner in the supervision of critical

patients since it extends beyond feature ranking to provide clinical knowledge learned in the process.

Quantitative and Comparative Analysis. The analysis was directed towards clinical Data Analytics (DA) and predictive level. The main quantification of performance involved Fitted Q Evaluation (FQE), a proven off-policy evaluation strategy that estimates the expected value of a policy return (policy value) of a policy without interaction with a human environment. FQE gives an approximation of the performance of each agent in terms of their total reward, a valuable proxy for the long-term efficacy of treatments.

We also described two cognitively relevant measures in addition to the FQE:

- **Mortality Rate at 90 Days (Est. %)** – this is a clinically critical outcome metric that measures the efficacy of treatment and patient safety with a long-term perspective, considers three months or 90 days.
- **Safe Range Adherence** – Measures the amount of time that a vital sign remained within the patient-specific threshold range divided by the total number of minutes a vital sign was monitored by the device.

Table 4 presents a summary of the findings for all models. The proposed XAI-DRL framework (BCQ-based agent) also achieved much better results on all the metrics than the other policies. It attained the greatest estimated policy value, which creates the impression that its suggestions generated the most preferable long-term consequences for the dynamics of the simulation. Along with this improvement, there was a lower 90-day mortality rate of 13.8%, which was compared to the 16.8% of the clinician policy and a significant improvement compared to the 15.1% of behavioral cloning. It is noteworthy that the DDPG agent, although a powerful DRL model, had poorer results than all the other policies, estimating a higher mortality rate and achieving lower rewards, further proving that naive online RL approaches are inappropriate in offline clinical situations.

TABLE 4. Offline policy evaluation results.

Policy	Estimated Policy Value (FQE)	Estimated 90-Day Mortality (%)	Time in Safe Range (%)
XAI-DRL (BCQ)	84.5	13.8	92.1
DDPG (Black-Box)	45.1	18.5	81.4
Behavioral Cloning (BC)	68.9	15.9	88.5
Clinician Policy	65.2	16.8	87.3

In addition to mortality, more patients were also maintained within safe physiological parameters by the XAI-DRL agent. It means that not only has it contributed to better survival rates but also made it more stable and consistent, making a case against adverse outcomes or harm inflicted by treatment.

TABLE 5. Analysis of recommended actions.

Policy	Mean Vasopressor Dose (mcg/kg/min)	Mean IV Fluid Bolus (mL/hr)	Frequency of High FiO_2 (>60%)
XAI-DRL (BCQ)	0.18	125.5	8.2%
Clinician Policy	0.25	95.3	19.5%

Table 5 is a further analysis of what actions are suggested by the best-performing XAI-DRL policy in comparison to the Clinician Policy. The learned policy is more careful in the use of vasopressors with a lower average dose. It is, however, more aggressive in the administration of IV fluids, so it could have learned to treat hypovolemia earlier. Chief among them, the agent

learned to avoid high doses of inspired oxygen (FiO_2), a practice that has been linked with lung injury, with rates of FiO_2 greater than 60 percent much rarer than those used by clinicians. This means that the agent was able to extrapolate a lung-protective ventilation strategy based on data, with the help of the safety elements under the reward structure.

The presence of the XAI module, in turn, increased the practical usefulness of the system. Within each suggested intervention, the framework produces clinically sensible interpretations using SHAP, which enables the end-users to comprehend the rationale behind every decision. This is an important capability that can be used to support human-AI collaboration and clinician trust, as well as post-hoc auditability.

Graphical Results and Discussion. Fig.7 gives the learning dynamics of the proposed XAI-DRL structure, whereas Fig.8 shows the learning dynamics of the baseline DDPG agent. Such values will be beneficial in understanding the behavior of the training and stability of individual agents in an offline reinforcement learning scenario.

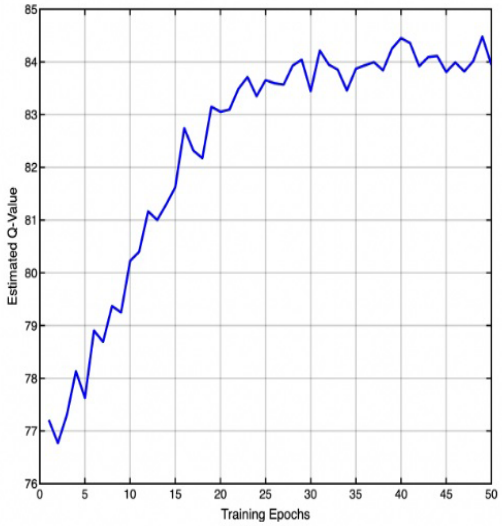


FIGURE 7. Learning curve for XAI-DRL.

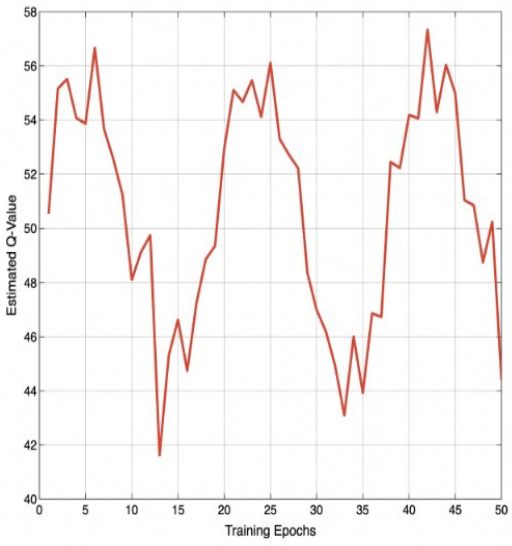


FIGURE 8. Learning curve for DDPG.

Namely, the learning curve of the XAI-DRL agent, based on the BCQ algorithm, shows the steady and stable convergence of the approximated Q-values in consecutive training iterations. Such a reflection on a high and stable Q-value suggests that the agent has managed to find an effective treatment policy in the offline dataset, striking a balance between exploring the environment and following the instruction distribution it receives during training. The smooth upward curve and subsequent leveling off implies that the BCQ agent can learn usable patterns about the time series so as to maximise the long-term cumulative rewards safely and without violating the data.

On the contrary, the DDPG agent, which is mostly used in online, interactive settings, produces highly unstable and unpredictable learning behavior in the same offline environment. The estimates of its Q-value vary unstably and do not tend to stabilize, as illustrated in Fig.8, where there is a drift over time that leads to value divergence. It is a widely discussed problem when training online DRL algorithms on fixed, batch-mode datasets; the agent is likely to suffer from extrapolation error and distributional shift and distributional shift. An unstable DDPG agent not only hinders the success of learning effective policies but is also a major safety concern in a clinical setting where reliability is of primary importance to the model. The fact that it fails to generalize beyond the presentational data is another reason why conservative offline decisions, such as BCQ, should be employed when stakes are high in medical interventions.

To further confirm such observations, Table 4 recaps all obtained policy values for each of the agents at the end of the training process, and Fig.9 then provides a graphical comparison of policy values obtained by each of the agents. The chart reinforces that the performance of the XAI-DRL agent exceeded all the others and achieved the maximum estimated return. Such an outcome is consistent with its consistent training curve and useful generalization, which is consistent with its applicability to critical care applications. The DDPG agent, in turn, performs much worse, and this value reflects the ineffective learning process and demonstrates the limited applicability of the agent in offline medical simulations. Collectively, these statistics allow the formulation of a complete picture of learning efficacy, stability, and real-world applicability of all assessed models.

Fig.10 depicts a relative simulation of a patient trajectory under two treatment policies, the conventional clinician routine-based policy and the proposed XAI-DRL policy. In the case of the clinician policy, the patient has a severe hypotensive episode described as a sudden rapid decline in Mean Arterial Pressure (MAP) that is subsequently rectified in the course. This reactive trend shows the strategy of delay of intervention that may put the patient at risk of continuing instability.

On the other hand, the XAI-DRL agent is more proactive. It pumps fluids at the earlier stages of the procedure, thereby literally eliminating the onset of hypotension and keeping MAP within a reasonable and non-dangerous range throughout the event. This proactive behavior explains the early measures of the learned policy to respond in consideration of early caution signs in order to become proactive and improve overall treatment safety and efficacy.

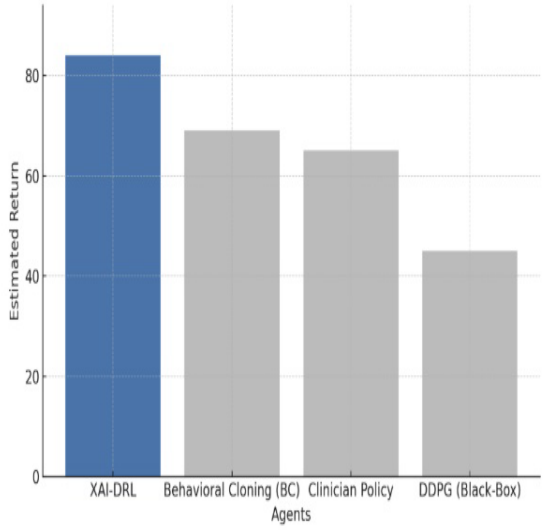


FIGURE 9. Policy value comparison.

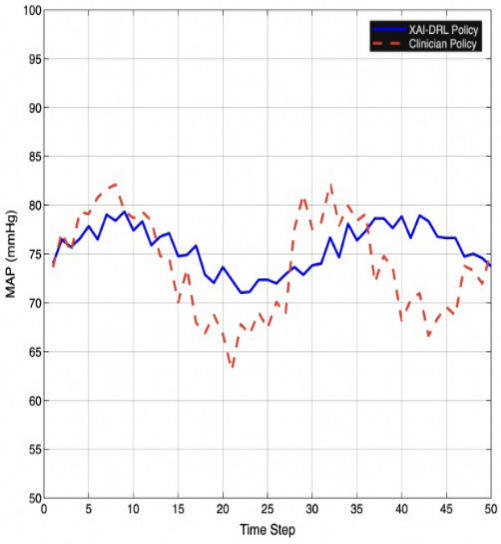


FIGURE 10. Simulated MAP time series.

To supplement such an analysis, a histogram of vasopressor dose receipt distribution plot is provided in Fig.11 relative to the two policies. The policy obtained via XAI-DRL, on the other hand, has a thinner distribution with a lower average dose, which indicates a more stable policy with conservative treatment. This means that the agent can accomplish therapeutic objectives by means of smaller, well-calibrated interventions, consistent with clinical priorities, including less overtreatment and less risk-taking.

Collectively, Fig.10 and Fig.11 confirm the results presented in Table 4, and the value of the XAI-DRL policy in facilitating physiological stability and providing data-driven, targeted care in high-stakes critical care settings.

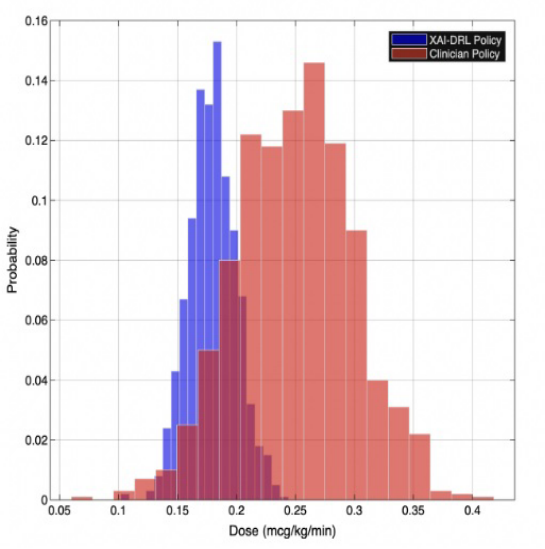


FIGURE 11. Vasopressor dose distribution.

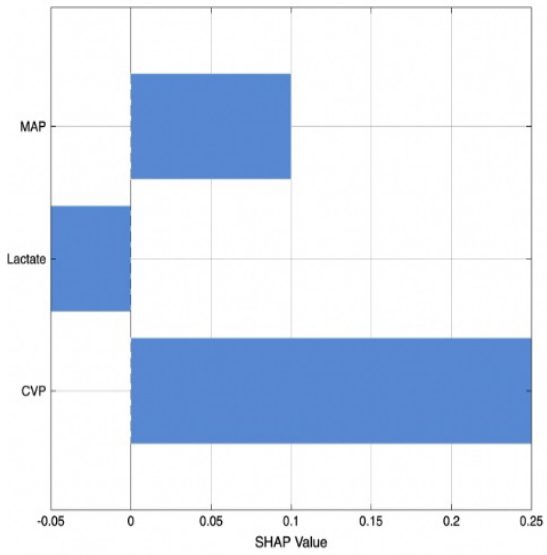


FIGURE 12. SHAP force plot.

The explainability is the most important strength of the proposed framework as it determines clinical trust and adoption. Fig.12 shows a force plot of SHAP, and based on one decision-recommending the use of more IV fluids. The plot reveals that the low fluid volume indicated by a key clinical reading, low Central Venous Pressure (CVP), is decisive in driving the decision. The degree of MAP being slightly low further supports the action, whereas stable lactate is marginally against it. Such an explanation is highly similar to a clinician's reasoning, and hence, the recommendation of the agent is transparent and verifiable.

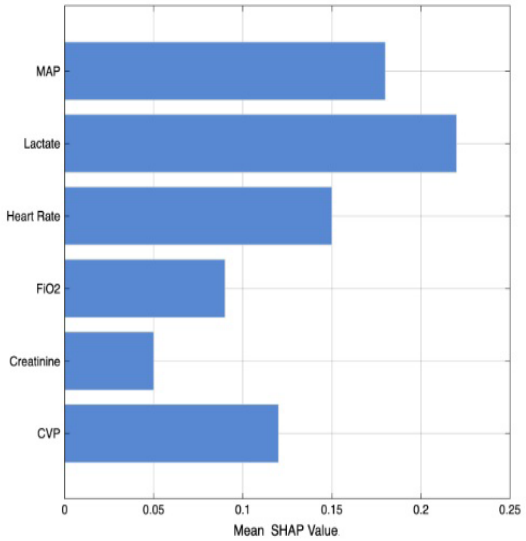


FIGURE 13. SHAP summary plot.

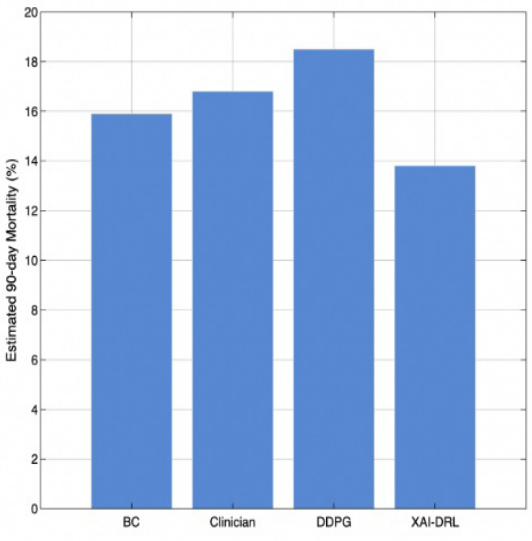


FIGURE 14. Mortality rate comparison.

The Fig.13 provides the global perspective of feature importance in thousands of decisions. It corroborates the fact that the policy of the agent uses clinically relevant parameters such as lactate, MAP, and heart rate this gives us some confidence that the behavior of the model corresponds to accepted medical logic.

These descriptions not only explain what the model predicts, but also why it does so to showing an active, consistent, and safety-conscious treatment approach. This openness converts the model into a collaborative tool that clinicians will be able to rely on and comprehend.

Fig.14 shows a bar chart that compares the estimated 90-day mortalities of each of the different treatment policies tested in the study in order to provide a visual summary of the findings in Table 4. The figure points out the relative performance of the proposed XAI-DRL (BCQ) policy and clinician-generated policy, Behavioral Cloning (BC) and the DDPG agent. The bars are in line with the mortality rate of each method estimated using Fitted Q Evaluation (FQE), which provides an intuitive way to understand the efficiency of each method. The chart shows that the XAI-DRL model achieves the lowest estimated mortality rate, thus it is the best in maximizing the long-term patient outcome. On the other hand, the DDPG agent, which does not handle offline settings, shows the highest mortality rate thus supporting the suspicions behind using online DRL algorithms on offline settings. This type of visualization is a good complement to this tabulated information because it enables the readers to gain a rapid comparative understanding of the clinical effects of each decision policy, making it easier to understand the practical considerations of the suggested framework.

6. Conclusion and future scope

The proposed work was able to design, implement, and comprehensively validate a new Explainable Artificial Intelligence- Deep Reinforcement Learning (XAI-DRL) framework suitable for clinical decision support in critical care, which is a complicated and high-stakes setting. The combination of a conservative offline DRL agent (the Batch-Constrained Q-learning, or BCQ) and robust post-hoc explainability module (SHAP) allowed the system to learn a complex treatment policy that performed considerably better than clinician-engineered strategies and baseline AI models in a high-fidelity simulated setting. The learned policy not only leads to a decrease in the estimated 90-day mortality rate but also cause the patients to spend more time in physiologically safe ranges overall, which denotes safety and effectiveness.

More importantly, the framework overcomes one of the most urgent obstacles to medical AI, the black-box problem of deep models, by providing explainable, human-interpretable, and clinically meaningful the rationale behind each of the suggested interventions. Such feature-based explanations enable clinicians to validate, trust, and understand the legitimacy of the model's reasoning of the model, which reduces the existing trust gaps that impeded the adoption of AI in clinical practice in the past. The explansibility help with regulatory compliance and clinical auditing and help to increase the practical value of the system.

These findings are highly indicative that this XAI-DRL framework can serve as a strong, reliable, and collaborative system that would serve as a supplement, not a substitute, for clinician knowledge. It provides a promising path toward safer, personalized, more effective AI-assisted care in the ICU with implications beyond this area as well, where improvements in transparent AI systems offer greater opportunity to improve the safety and efficiency of high-risk, high-complexity medical decision-making.

The encouraging outcomes obtained in the current study provide a number of interesting directions for future research and development. It is a logical next step to examine the proposed XAI-DRL structure in more sophisticated and physiologically realistic simulators that mirror the real-world dynamics and variation of ICU patients. Validation is an intermediate step that will facilitate the transfer from simulation-based testing to real-world clinical deployment. Thereafter, the framework can be tested in well-controlled prospective clinical trials to determine its feasibility, safety, and acceptability at the bedside in real critical care. Continuous feedback from clinicians would also be possible with such trials, allowing iterative improvements to the DRL agent as well as what constitutes explainability.

Methodologically, it could be said that the inability to differentiate between the true causal relationship and simple correlation is one of the most serious limitations of work with retrospective observational data. In the future, attempts should be made to incorporate elements of causal inference, e.g., structural causal models or counterfactual thinking, into the pipeline of policy learning. This would allow the system to learn treatment strategies that are not only predictive but also causally grounded, but also causally-grounded, as necessary to safely and across-population deploy to a variety of patients and clinical contexts.

Another important area is improving the interactivity and usability of the explanation interface. Although the use of SHAP-based explanations is beneficial in terms of post-hoc insights, it would be helpful to enable clinicians to also visit the logic behind any of the presented recommendations interactively, e.g., by exploring what-if or counterfactual scenarios. This capability would help develop a better understanding of both the logic of the model and the clinical implications of alternative choices, leading to better engagement and confidence among the clinician.

Also, the decentralized training methods such as federated learning have significant potential to be explored. Data in critical care is frequently isolated in separate hospitals; thus, it is challenging to develop sturdy and generalizable models based on centralized data. By creating a global model on which models can be trained, federated learning would enable institutions to maximize patient privacy and confidentiality by avoiding the transmission of raw patient data, improving the effectiveness and generalizability of trained models across different healthcare systems. Privacy-protecting methods like differential privacy or secure multi-party computation can be used as well to add an additional layer of protection to sensitive patient data in the course of learning.

When put together, these future research directions are to expand the current picture into an entire, safe, and scalable decision support system, not only clinically effective, but also transparent, ethical, and flexible to the changing healthcare environments. Work in these fields would help move the safe and responsible application of AI into critical care and other high-stakes medical areas faster, and closer to the goal of human-AI collaboration in life-saving decision-making.

7. Acknowledgments

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP.2/315/46. This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under project Grant KFU261820. Ongoing Research Funding program, (ORF-2026-1925), King Saud University, Riyadh, Saudi Arabia. Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R853), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FPEJ -2026- 2913-04.

References

[1] Alserhani, F. Intrusion detection and real-time adaptive security in medical iot using a cyber-physical system design, *Sensors,* Vol.25, No.15, 2025, pp.4720.

[2] Chakraborty, B., Murphy, S.A. Dynamic treatment regimes, *Annu. Rev. Stat. Appl.*, Vol.1, 2014, pp.447-464.

[3] Ghassemi, M., Naumann, T., Schulam, P., Beam, A.L., Chen, I.Y., Ranganath, R. A review of challenges and opportunities in machine learning for health, *AMIA Jt Summits Transl Sci Proc.,* 2020, pp.191-200.

[4] Guo, T.J., Bardhan, I.R., Ding, Y., Zhang, S.C. An explainable artificial intelligence approach using graph learning to predict intensive care unit length of stay, *Inf. Syst. Res.,* Vol.36, No.3, 2024, 0029.

[5] Fujimoto, S., Meger, D., Precup, D. Off-policy deep reinforcement learning without exploration, In *Proc. 36th Int. Conf. Mach. Lear., PMLR 97*, 2019, pp.2052-2062.

[6] Hoenig, A., Roy, K., Acquaah, Y.T., Yi, S., Desai, S.S. Explainable AI for cyber-physical systems: Issues and challenges, *IEEE Access,* Vol.12, 2024, pp.73113-73140.
[7] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.H., et al. MIMIC-III, a freely accessible critical care database, *Sci. Data*, Vol.3, No.1, 2016, Article No.160035.
[8] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C., Faisal, A.A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care, *Nat. Med.*, Vol.24, No.11, 2018, pp.1716-1720.
[9] Kumar, A., Malla, J., Sharma, A. BioBERT-RxReadmit: Improving hospital readmission predictions through clinical text analysis with BioBERT, *Int. J. Eng. Artif. Intell. Manag. Decis. Support Policies*, Vol.2, No.1, 2025, pp.14-29.
[10] Lee, I., Sokolsky, o. Medical cyber-physical systems, Ed. Rawat, D.B., Rodrigues, J.J.P.C., Stojmenovic, I. *In Cyber-Physical Systems: From Theory to Practice,* CRC Press, 2015, pp.743-748.
[11] Liu, W., Zhao, F., Shankar, A., Maple, C., et al. Explainable AI for medical image analysis in medical cyber-physical systems: Enhancing Transparency and Trustworthiness of IoMT, *IEEE J. Biomed. Health Inform.,* Vol.29, No.4, 2025, pp.2365-2376.
[12] Lundberg, S.M., Lee, S.I. A unified approach to interpreting model predictions, *NIPS'17: Proc. 31st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp.4768-4777.
[13] Peine, A., Hallawa, A., Bickenbach, J.J., Dartmann, G. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care, *NPJ Digit Med.*, Vol.4, No.1, 2021, Article No.32.
[14] Prasad, N., Cheng, L.F., Chivers, C., Draugelis, M., Engelhardt, B.E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units, arXiv preprint arXiv:1704.06300, 2017.
[15] Raghu, A., Komorowski, M., Celi, L.A., Szolovits, P., Ghassemi, M. Continuous state-space models for optimal sepsis treatment: A deep reinforcement learning approach, *Proc. 2nd Mach. Learn. Healthcare Conf.,* Vol.68, 2017, pp.147-163.
[16] Ray, P., Rawal, R., Rizvi, A.Z. Application of random forest and decision tree classifier approach for the survival of heart patients using characteristic extraction, *Int. J. Eng. Artif. Intell. Manag. Decis. Support Policies*, Vol.1, No.1, 2024, pp.1-11.
[17] Sarkar, M., Chatterjee, S., Hazra, S., Ghosh, R., Chakraborty, S. A study on anaphylactic-shock forecasting system with knowledge-based intelligent architecture, *Int. J. Comput. Model. Appl.*, Vol.1, No.1, 2024, pp.1-19.
[18] Vollmer, S., Mateen, B.A., Bohner, G., et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, *BMJ*, Vol.368, 2020, pp.l6927.
[19] Vouros, G.A. Explainable deep reinforcement learning: State of the art and challenges, *ACM Comput. Surv.,* Vol.55, No.5, 2023, pp.1-39.
[20] Wiggerthale, J., Reich, C. Explainable machine learning in critical decision systems: Ensuring safe application and correctness, *AI*, Vol.5, No.4, 2024, pp.2864-2896.

Author photos were not provided due to institutional policy.

Sahar Mansour - is an Assistant Professor at the Department of Radiological Sciences, College of Health and Rehabilitation Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include AI, eHealth and Biomedical.

Mona Almofarreh - is an Assistant Professor at the Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her research interests include AI, bioinformatics, IoT, and machine learning.

Mashael M. Asiri - is an Assistant Professor at the Department of Computer Science, College of Science Art at Mahayil, King Khalid University, Saudi Arabia. Her research interests include deep learning, IoT, cloud computing and AI.

Essa Alyounis - is an associate professor at the Department of Health Information Management and Technology, College of Applied Medical Science, King Faisal University, Al Ahsa, Saudi Arabia. His research interests include IoT, smart cities, deep learning, NLP, AI.

Mashail N. Alkhomsan - is an Assistant Professor at the Department of Software Engineering, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia. Her research interests include machine learning, image processing, cybersecurity, IoT and cloud computing.

Ayman Yafoz - is an Assistant Professor at the Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include cybersecurity, deep learning, computer security, AI.

Asma Alhashmi - is an Assistant Professor at the Department of Computer Science, College of Science, Northern Border University, Saudi Arabia. She completed her Ph.D. degree in Computer Science from the University of Mysore, India in 2015. She has more than 10 years of experience in the IT field. Her research interests include cybersecurity, software engineering, e-government, and cloud computing.