Brain Informatics

RESEARCH Open Access

AI-driven multi-agent reinforcement learning framework for real-time monitoring of physiological signals in stress and depression contexts

Thanveer Shaik[1*], Xiaohui Tao[1†], Lin Li[2†], Haoran Xie[3†], Hong-Ning Dai[4†], Feng Zhao[5†] and Jianming Yong[6†]

Abstract

Purpose Effective patient monitoring is crucial for timely healthcare interventions and improved outcomes, especially in managing conditions influenced by stress and depression, which can manifest through physiological changes. Traditional monitoring systems often struggle with the complexity and dynamic nature of such conditions, leading to delays in identifying critical scenarios. This study proposes a novel multi-agent deep reinforcement learning (DRL) framework to address these challenges by monitoring vital signs and providing real-time decision-making capabilities.

Methods Our framework deploys multiple learning agents, each dedicated to monitoring specific physiological features such as heart rate, respiration, and temperature. These agents interact with a generic healthcare monitoring environment, learn patients' behavior patterns, and estimate the level of emergency to alert Medical Emergency Teams (METs) accordingly. The study evaluates the proposed system using two real-world datasets-PPG-DaLiA and WESAD-designed to capture physiological and stress-related data. The performance is compared with baseline models, including Q-Learning, PPO, Actor-Critic, Double DQN, and DDPG, as well as existing monitoring frameworks like WISEML and CA-MAQL. Hyperparameter optimization is also performed to fine-tune learning rates and discount factors.

Results Experimental results demonstrate that the proposed multi-agent DRL framework outperforms baseline models in accurately monitoring patients' vital signs under stress and varying conditions. The optimized agents adapt effectively to dynamic environments, ensuring timely detection of critical health deviations. Comparative evaluations reveal superior performance in metrics related to decision-making accuracy and response efficiency, highlighting the robustness of the framework.

Conclusions The proposed AI-driven monitoring system offers significant advancements over traditional methods by handling complex and uncertain environments, adapting to varying patient conditions influenced by stress and depression, and making autonomous, real-time decisions. While the framework demonstrates high accuracy and adaptability, challenges related to data scale and future vital sign prediction remain. Future research will focus on extending predictive capabilities to further enhance proactive healthcare interventions.

Keywords Behavior patterns, Decision making, Patient monitoring, Reinforcement learning, Vital signs

[†]Xiaohui Tao, Lin Li, Haoran Xie, Hong-Ning Dai, Feng Zhao and Jianming Yong have contributed equally to this work.

*Correspondence:
Thanveer Shaik
Thanveer.Shaik@unisq.edu.au
Full list of author information is available at the end of the article

1 Introduction

Mental health disorders, particularly depression and stress, are among the most pervasive global health challenges today, significantly impacting individuals' well-being and productivity [1]. These conditions, often referred to as"silent epidemics,"require early detection and timely interventions to mitigate their effects. However, traditional approaches to mental health management have often been reactive, addressing symptoms only after they manifest significantly. This underscores the urgent need for proactive monitoring and assessment frameworks that can identify early warning signs, enabling clinicians to intervene effectively before the conditions escalate.

The physiological and behavioral manifestations of depression and stress, such as changes in heart rate, respiration patterns, and body temperature, provide measurable indicators of an individual's mental health state [2, 3]. Modern advancements in wearable technology and Internet of Things (IoT)-enabled systems now make it possible to continuously monitor these indicators in real time, presenting new opportunities to enhance mental health care. However, the challenge lies in effectively analyzing the complex, multi-dimensional data generated by these systems and deriving actionable insights to guide clinical decisions.

Traditional machine learning (ML) techniques have been extensively employed in this domain to classify physiological signals, identify patterns, and predict health outcomes [4, 5]. These methods have laid a strong foundation for developing monitoring frameworks but are inherently limited by their reliance on static models that do not adapt to changing conditions or learn dynamically from ongoing data streams. They are primarily observational, suggesting potential courses of action without the ability to autonomously adapt or act in response to the observed patterns.

Reinforcement Learning (RL) represents a paradigm shift in this context by enabling autonomous agents to actively interact with their environment, learn from feedback, and optimize their actions to achieve predefined objectives [6]. Unlike traditional ML models, RL agents leverage a reward-driven approach where each action taken by the agent is evaluated through a reward mechanism that reinforces favorable behaviors and discourages undesirable ones. This iterative learning process allows RL agents to adapt dynamically to complex, uncertain, and ever-evolving environments, making RL an ideal candidate for healthcare applications that require precision, adaptability, and responsiveness.

RL has already demonstrated its potential in various domains, including dynamic treatment optimization, diagnostic decision-making, and medication scheduling [7–9]. For instance, RL algorithms have been used to optimize the timing and dosage of medications, ensuring that treatments are administered at the most effective intervals. The analogy of RL agents acting as virtual clinicians, continuously monitoring a patient's state and making decisions based on observed changes, highlights the transformative potential of this technology in healthcare [10].

In this study, we propose a novel monitoring framework that utilizes multi-agent Deep Reinforcement Learning (DRL) to address the complexities associated with monitoring depression and stress. The framework is designed to analyze and interpret real-time physiological data, enabling clinicians to detect deviations from normal patterns and respond proactively. Each DRL agent is dedicated to monitoring a specific physiological parameter, such as heart rate, respiration rate, or body temperature, and learns optimal thresholds based on Modified Early Warning Scores (MEWS) [11]. By introducing a clinically-informed reward mechanism, the framework enables these agents to continuously refine their decision-making capabilities, ensuring timely and accurate alerts to medical teams (Fig. 1).

The proposed framework represents a significant advancement over traditional RL models by employing a multi-agent architecture that allows simultaneous monitoring of multiple vital signs. This distributed approach enhances the system's scalability, enabling it to handle the complexities of real-world healthcare scenarios where multiple parameters must be monitored concurrently. Furthermore, the novel reward system ensures that the agents are aligned with clinically relevant objectives, optimizing their behavior to support timely medical interventions.

The contributions of this study are summarized as follows:

- Introduction of a clinically-informed reward mechanism tailored to support RL agents in learning behavior patterns indicative of depression and stress.
- Development of a generic, multi-agent monitoring environment that enables simultaneous tracking of various physiological parameters.
- Establishment of a novel paradigm for remote monitoring of mental health conditions, leveraging multi-agent DRL to provide actionable insights in real-time.

The rest of this paper is organized as follows: Sect. 2 reviews related literature on RL applications in healthcare and mental health monitoring. Section 3 provides a detailed description of the research problem, technical background, and proposed methodology. Experimental setup and evaluation metrics are discussed in Sect. 4,

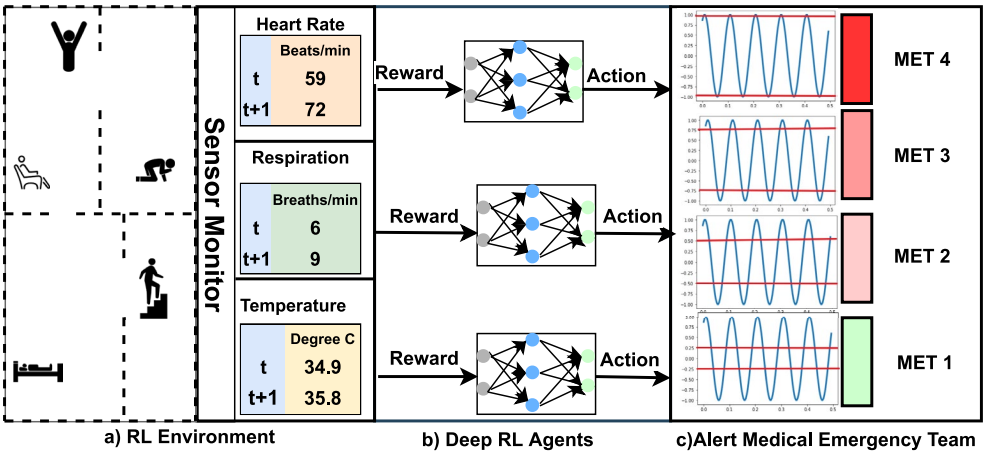

Fig. 1 Human monitoring framework for tracking vital signs and alerting medical teams in emergencies

followed by an analysis of the results and insights in Sect. 5. Applications and implications are discussed in Sect. 6, and Sect. 7 concludes the paper by outlining limitations and future directions.

2 Related works

2.1 Machine learning in healthcare

Machine learning (ML) has transformed healthcare by providing predictive, diagnostic, and monitoring solutions across various domains [12]. Supervised learning algorithms, in particular, leverage labeled datasets to make predictions and classifications based on input features [13, 14]. For instance, ML and deep learning techniques have been employed to predict vital signs like heart rate and classify physical activities [15]. In the context of mental health, ML models have demonstrated efficacy in detecting stress and depression through the analysis of physiological and behavioral data [16, 17]. Stress and depression are critical public health concerns, often linked to chronic conditions like cardiovascular disease and diabetes. Early detection of these conditions can significantly improve outcomes by facilitating timely interventions.

Oyeleye et al. [18] investigated ML and deep learning models to estimate heart rate using wearable devices, comparing various regression algorithms including linear regression, k-nearest neighbor (kNN), decision tree, and LSTM. Similarly, Luo et al. [19] utilized LSTM models to predict heart rate by integrating factors such as gender, age, physical activity [20], and mental state, highlighting the relevance of mental well-being in monitoring overall health.

Unsupervised learning algorithms further contribute by deriving patterns from unlabeled data, employing clustering and association techniques [21, 22]. Sheng and Huber [21] proposed an encoder-decoder framework to cluster physical activity data, achieving high accuracy by learning behavioral embeddings. Despite their strengths, these traditional ML techniques face limitations in dynamically adapting to uncertain environments or integrating diverse data sources.

Reinforcement Learning (RL) addresses these gaps by enabling systems to learn through interaction with their environment [23]. Unlike supervised approaches, RL relies on rewards or penalties to optimize decision-making processes, making it particularly suited for real-time and sequential decision-making tasks [24]. This capability is critical for monitoring complex conditions like stress and depression, where continuous data-driven interventions can prevent deterioration.

2.2 Mimicking human behavior patterns

Understanding human behavior is vital for developing personalized healthcare solutions, especially for stress and depression management. Stressful events and depressive episodes often manifest through changes in physical activity, sleep patterns, and physiological responses [16]. Tirumala et al. [25] explored probabilistic trajectory models to analyze human movement and interactions, proposing a hierarchical reinforcement learning (HRL) framework for identifying behavior patterns. Janssen et al. [26] extended this concept by segmenting complex biological behaviors into manageable subtasks using HRL, which organizes sequential actions into logical options.

Tsiakas et al. [27] proposed a human-centric cyber-physical systems (CPS) framework for personalized human-robot collaboration and training, which focused on minimally intrusive methods to predict human attention. Similarly, Kubota et al. [28] examined robots'

adaptability to cognitive impairments, exploring therapeutic and assistive applications. Such frameworks emphasize the importance of understanding both high-level behaviors (e.g., emotional and cognitive states) and low-level behaviors (e.g., speech, gestures, and physiological signals) [29]. This research forms the foundation for developing systems that can effectively address mental health challenges like stress and depression.

2.3 Reinforcement learning in healthcare

Reinforcement Learning (RL) has emerged as a transformative tool in healthcare for its ability to handle complex, dynamic, and uncertain environments. Lisowska et al. [30] demonstrated how RL could optimize the timing of interventions for cancer patients, employing models such as Deep Q-Learning (DQL), Advantage Actor-Critic (A2C), and Proximal Policy Optimization (PPO) to develop virtual coaches for personalized prompts. Personalized interventions, including messaging for diabetes patients, have shown efficacy in increasing physical activity and improving mental health [31].

Li et al. [32] leveraged RL to analyze electronic health records (EHRs) for sequential decision-making, employing a model-free Deep Q-Networks (DQN) algorithm for clinical decision support. Guo et al. [33] proposed a dynamic weight assignment network inspired by advanced RL algorithms, demonstrating its application in human activity recognition. RL's ability to integrate multi-agent frameworks further enhances its potential for mental health monitoring by enabling concurrent learning across multiple parameters.

Despite RL's success in areas like gaming and assistive robotics, its deployment in healthcare, especially for mental health conditions, poses unique challenges. Traditional approaches struggle with the safety and uncertainty inherent in dynamic healthcare environments. Stress and depression monitoring, for example, require systems that can adapt to fluctuating physiological and behavioral data. Our study introduces a multi-agent reinforcement learning (MARL) framework designed specifically for these challenges. Unlike single-agent systems, MARL allows for concurrent monitoring of multiple physiological parameters, each modeled by a specialized agent. This framework is particularly suited for stress and depression monitoring, where indicators such as heart rate variability, sleep disruptions, and activity levels must be continuously analyzed.

By incorporating a clinically-informed reward mechanism, our MARL framework aligns agent behavior with healthcare objectives, ensuring timely interventions. This approach not only addresses safety concerns but also enhances the scalability and adaptability of mental health monitoring systems, providing a novel contribution to AI-driven healthcare.

3 DRL monitoring framework

In this section, the design of a human behavior monitoring system, DRL monitoring framework, that uses R is presented in detail. The aim of the system is to monitor vital signs to learn human behavior patterns and ensure clinical safety in an uncertain environment. The proposed framework involves a multi-agent system where each vital sign state is observed by an individual agent, as shown in Fig. 2. A DRL algorithm, DQN, is used to learn effective strategies in the sequential decision-making process without prior knowledge through trial-and-error interactions with the environment[34, 35].

3.1 Technical background

The challenge addressed in this research is the development of a multi-agent framework for real-time health status monitoring by learning and interpreting patterns in vital signs through wearable sensors. The agents must detect deviations from normal vital sign patterns that exceed Modified Early Warning Scores (MEWS) thresholds and alert the emergency team accordingly.

To formulate this problem, we leverage the framework of Markov Decision Processes (MDP), expressed as a 5-tuple $M = (S, A, P, R, \gamma)$. Here, S represents the finite state space, where each state $s_t \in S$ corresponds to a distinct combination of vital sign readings at time t. The action set A comprises potential alerts the agents can issue based on the observed vital signs. The transition function $P(s, a, s')$ models the probability of moving from state s to state s' upon taking action a, reflecting the dynamic nature of human vital signs.

Central to our approach is the reward function $R(s, a)$, which is defined to prioritize actions that lead to the early detection of potential health risks, thereby enabling timely intervention. This is mathematically represented as:

$$R(s_t, a_t) = \sum_{t=0}^{\infty} \gamma^t r_t, \tag{1}$$

where γ is the discount factor that balances the importance of immediate versus future rewards, ensuring the agents' actions are aligned with long-term health monitoring objectives.

The goal is to discover an optimal policy $\pi(s_t)$ that maximizes the expected reward by selecting the most appropriate action a_t in any given state s_t. This optimization is achieved through the iterative update of the Q-function, as outlined in the Bellman equation:

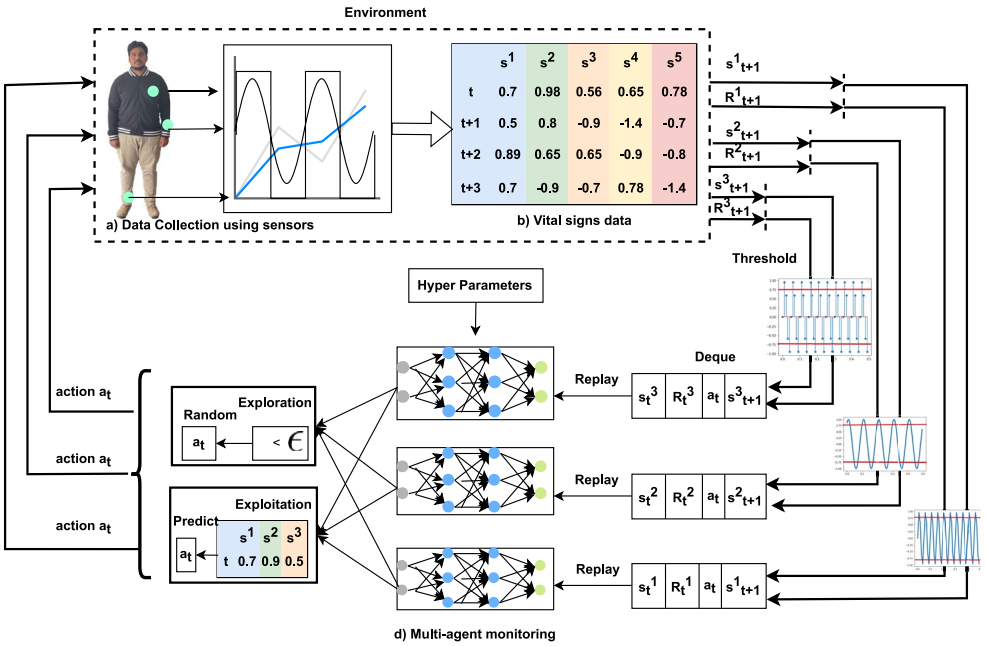

Fig. 2 Multi-agent monitoring framework

$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left(r_t + \gamma \max_a Q(s_{t+1}, a)\right), \tag{2}$$

where α represents the learning rate, influencing the integration of new information into the Q-function. Through this process, the agents continually refine their decision-making strategy, enhancing the system's capability to monitor and respond to emerging health risks effectively.

3.2 Monitoring environment

A custom RL monitoring system based on MDP has been created to have human vital signs data serve as the observation space S, action space A for learning agents to make decisions, and rewards R for the agents' actions as depicted in Fig. 2. This study introduces a novel isolated multi-agent MDP framework that allows multi-agents to share the same environment and make decisions based on the health parameters they are monitoring, receiving rewards without being influenced by the decisions of other agents. The goal of all agents in this environment is to monitor the health of patients using the predefined MEWS, as shown in Tab. 1. In healthcare, each vital sign plays a critical role in determining a person's clinical safety.

In the current framework, we have implemented three RL agents to monitor heart rate, respiration, and temperature. These agents operate primarily in cooperative mode, sharing information about the patient's health status and working together to ensure timely interventions. Cooperation allows the agents to pool rewards from collective actions, improving overall system learning. However, when multiple patients are being monitored or resource constraints arise (e.g., limited access to medical personnel), the agents may enter competitive mode. In this mode, agents prioritize the most critical health states and may compete for resources by adjusting the urgency of alerts based on the patient's condition.

Table 1 Modified Early Warning Scores [36]

MEWS	4/MET	3	2	1	0	1	2	3	4/MET
Respiratory Rate	≤4	5–8			9–20	21–24	25–30	31–35	≥36
Oxygen Saturation	≤84	85–89	90–92	93–94	≥95				
Temperature		≤34.0	34.1–35.0	35.1–36.0	36.1–37.9	38.0–38.5	≥38.6		
Heart Rate	≤39			40–49	50–99	100–109	110–129	130–139	≥140
Sedation Score					Awake		Mild	Moderate	Severe

As the number of agents increases, the framework is designed to scale effectively. Each additional agent monitors new physiological parameters or additional patients, with the system adjusting the reward mechanism and communication strategy to maintain efficient performance. The system remains modular, enabling easy expansion without significantly impacting computational load or decision-making speed. Importantly, the system's ability to operate in both cooperative and competitive modes ensures flexibility, allowing it to adapt to various healthcare scenarios, including large-scale monitoring in hospitals.

3.2.1 Observation space

The environment in Fig. 2 has a state, represented by $s_t^i \epsilon S$, where $i = 0, 1, 2, ...n$, refers to observations at time t. The aim is to divide the state into observations and allocate them to multi-agents. Suppose S represents the state of the human body, and there are three observations, $s_t^0, s_t^1, s_t^2 \epsilon S$, that represent different internal vital signs of the human body at time t. The human health status is controlled by multiple internal vital signs, each with a different threshold as shown in MEWS Tab. 1. Using a single agent to monitor all the vital signs can result in a sparse rewards challenge [16], where the environment might produce few useful rewards and hinders the learning of an agent. Therefore, multi-agents need to be deployed for each human to monitor the critical vital signs. The expected return E_π of a policy π in a state s can be defined by state-value Eq. 3 in the multi-agent setting, where $i = 0, 1, 2, 3, ...n$ is a finite number of observations n in the state.

$$V^\pi(s^i) = E_\pi \left\{ \sum_{t=0,i=0}^{\infty,n} \gamma^t R(s_t, \pi(s_t)) | s_0^0 = s \right\} \tag{3}$$

3.2.2 Action space

The action space of the monitoring environment is defined based on the MEWS [36] as shown in Tab. 1. The table presents early warning scores of adults' vital signs with the appropriate Medical Emergency Team (MET) to contact if any escalations in the health parameters. Based on the MEWS as threshold values, the action space has been segmented to have five discrete actions to communicate the vital signs to MET–0, MET-1, MET-2, MET-3, and MET-4. Each of these actions will be taken by agents based on the current state of the vital signs they are monitoring. The expected return E_π for taking an action a in a state s under a policy π can be measured using the action-value function $Q_\pi(s, a)$ defined in Eq. 4.

$$Q^\pi(s, a) = E_\pi \left\{ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, \pi(s_t)) | s_0 = s, a_0 = a \right\} \tag{4}$$

3.2.3 Rewards

The reward policy is designed to incentivize accurate monitoring and timely alerts. Agents are positively rewarded for actions aligned with MEWS thresholds (Table 2), ensuring critical conditions like stress-induced hyperthermia or depression-related bradycardia are prioritized. Rewards are categorized for each action, as shown in Eq. 6. This encourages agents to maximize cumulative rewards and learn behavior patterns, crucial for addressing mental health risks.

The goal of RL is to maximize cumulative rewards obtained through the actions of learning agents in an environment. In traditional RL, an agent is rewarded based on its action that leads to a transition from state s_t to s_{t+1}. In this study, the objective of the learning agent is to learn patterns in human vital signs. This is achieved through the design of an effective reward policy. The reward policy, as defined in this study, is calculated using Eq. 5. The agents are positively rewarded if they monitor vital signs in a state and take the correct action from the action space to communicate with the correct MET as defined in MEWS Tab. 1. On the other hand, if the agent takes the wrong action, it is negatively rewarded. The rewards are split into five categories for the five actions in the action space based on the MET from MEWS Tab. 1. The full rewards for each action selected by the agents are presented in Tab. 2. The reward policy utilizes the DRL agents' desire to maximize rewards in each learning iteration, making them learn the behavior patterns. Under each category, different levels of rewards were configured. For example, an observation $s_t^1 \epsilon S$ at the time t is related to heart rate falling under MET–4, the rewards are shown in Eq. 6.

$$R(s_t, a_t) = \begin{cases} +reward & \text{if } action = MET \\ -reward & \text{if } action \neq MET \end{cases} \tag{5}$$

Table 2 Rewards Policy

MEWS	4	3	2	1	0
Action 0	− 4	− 3	− 2	−1	10
Action 1	− 4	− 3	− 2	10	− 1
Action 2	− 4	− 3	10	−1	− 2
Action 3	− 4	10	−1	−2	− 3
Action 4	10	− 3	− 2	−1	− 4

$$R(s_t^1, a_t) = \begin{cases} 10 & \text{if } MET = 4 \& action = 4 \\ -1 & \text{if } MET = 4 \& action = 3 \\ -2 & \text{if } MET = 4 \& action = 2 \\ -3 & \text{if } MET = 4 \& action = 1 \\ -4 & \text{if } MET = 4 \& action = 0 \end{cases} \quad (6)$$

Correctness Determination in Reward Design The clinically-informed reward mechanism in our framework is designed to reflect the accuracy of agent decisions with respect to established triage protocols. Each physiological observation is assigned a severity band based on the Modified Early Warning Scores (MEWS), which are widely used in clinical settings to determine escalation levels. If the agent selects the correct Medical Emergency Team (MET) level that corresponds to the MEWS-derived threshold (for instance, selecting MET-3 when the heart rate exceeds 130 bpm), it receives a high positive reward (+10). In contrast, if the agent overestimates or underestimates the appropriate escalation level, it is penalized proportionally (e.g., −1 to− 4) based on the deviation from the correct action. This graded reward policy allows agents to learn both clinical accuracy and escalation sensitivity, supporting a balance between safety (avoiding false negatives) and efficiency (avoiding false positives).

3.3 Learning agent

In this study, a game learning agent DQN algorithm is employed. The DQN algorithm was first introduced by DeepMind, a subsidiary of Google, for playing Atari games. It allows the agent to play games by simply observing the screen, without any prior training or knowledge about the games. The DQN algorithm approximates the Q-Learning function using neural networks, and the learning agent is rewarded based on the neural network's prediction of the best action for the current state. For this research, the reward policy is described in more detail in Sect. 3.2.3.

3.3.1 Function approximation

The neural network used in this study to estimate the Q-values for each action has three layers: an input layer, a hidden layer, and an output layer. The input layer has a node for each vital sign in a state and the output layer has a node for each action in the action space. The model is configured with a relu activation function, mean square error as the loss function, and an Adam optimizer. The model is trained on the states and their corresponding rewards and, once trained, it can predict the accumulated reward.

The learning agent takes an action $a_t \in A$ in a transition from state s_t to $s_t^{'}$ and receives a reward R. In this transition, the maximum Q-function value is calculated according to Eq. 4, and the calculated value is discounted by a discount factor γ to prioritize immediate rewards over future rewards. The discounted future reward is combined with the current reward to obtain the target value. The difference between the prediction from the neural network and the target value forms the loss function, which is a measure of the deviation of the predicted value from the target value and can be estimated using Eq. 7. The square of the loss function penalizes the agent for large loss values.

$$loss = (\underbrace{R + \gamma \cdot max(Q^{\pi^*}(s,a))}_{target\_value} - \underbrace{Q^{\pi}(s,a)}_{predicted\_value})^2 \quad (7)$$

3.3.2 Memorize and replay

The basic neural network model has a limitation in its memory capacity and can forget previous observations as they are overwritten by new observations. To mitigate this issue, a memory array that stores the previous observations including the current state s_t, action a_t, reward R, and next state $s_t^{'}$ is used. This memory array enables the neural network to be retrained using the replay method, where a random sample of previous observations from the memory is selected for training. In this study, the neural network model was retained by using a batch size of 32 previous observations.

3.3.3 Exploration and exploitation

The exploration-exploitation trade-off in RL refers to the balancing act between trying out new actions to gather information and exploiting the actions that lead to the highest rewards. This balance can be modeled mathematically using the ϵ-greedy algorithm, which defines a probability ϵ of choosing a random action and a probability $1 - \epsilon$ of choosing the action believed to lead to the highest reward based on the current knowledge of the action-value function $Q(s_t, a)$. The equation to determine the action taken at time t is as follows:

$$a_t = \begin{cases} random(a_t) & \text{with probability } \epsilon \\ greedy(a_t) & \text{with probability } 1 - \epsilon \end{cases} \quad (8)$$

where the greedy action is defined as:

$$a_t = \arg\max_a Q(s_t, a) \quad (9)$$

The value of ϵ determines the level of exploration versus exploitation, with smaller values leading to more exploitation and larger values leading to more exploration. Over time, as the action-value function becomes more accurate, ϵ can be decreased to allow for more exploitation and convergence to the optimal policy.

In this study, we emphasize the importance of balancing exploration and exploitation for effective patient monitoring. Exploration allows agents to discover better monitoring strategies, while exploitation ensures timely alerts by acting on learned knowledge. Through empirical testing, we found that an exploration rate ϵ between 0.1 and 0.2 provided the optimal balance in our healthcare environment. This range ensured that agents could adapt to changing patient conditions while still providing timely and accurate interventions. In critical situations with frequent health deviations, a higher exploitation rate proved beneficial, whereas environments with fewer critical events required more exploration to discover new monitoring patterns.

3.3.4 Hyper parameters

Other than the parameters defined for the neural networks, a set of hyperparameters has to supply for the RL process. They are as follows:

- episodes ($\mathcal{M}$): This is a gaming term that means the number of times an agent has to execute the learning process.
- learning_rate(α): Learning rate is to determine much information neural networks learn in an iteration.
- discount_factor(γ): Discount factor ranges from 0 to 1 to limit future rewards and focus on immediate rewards.

Algorithm 1 Multi-agents monitoring

Require: Input: a set of subjects $\mathcal{C} = \{1, 2, \ldots, C\}$;a set of vital signs $\mathcal{V} = \{1, 2, \ldots, V\}$; Episodes $\mathcal{M} = \{1, 2, \ldots, M\}$;
Ensure: Output: Rewards achieved by agentss in each episode.
Initialization : $observation\_space = s_t^i \epsilon S, action\_space = a_t \epsilon A, reward\ R, \gamma, \epsilon, \epsilon_{decay}, \epsilon_{min}, memory = \emptyset, batch\_size$
$Set\ monitor\_length = N$
if action is appropriate **then**
    $R \leftarrow +reward$
else
    $R \leftarrow -reward$
end if
Define $model \leftarrow NeuralNetworkModel$
$memory \leftarrow memory \cup (s_t, a_t, R, s_{t+1})$
if $np.random.rand < \epsilon$ **then** ▷ Exploration
    $action\_value \leftarrow random(a_t)$
else ▷ Exploitation
    $action\_value \leftarrow greedy(a_t)$
end if
for episode $m \in \mathcal{M}$ **do**
    $score = 0$
    for time in range(monitor_length) **do**
        $a_t \leftarrow action(s_t)$
        $s_{t+1}, R, done \leftarrow step(a_t)$
        $memory \leftarrow memory \cup (s_t, a_t, R, s_{t+1})$
        $s_t \leftarrow s_{t+1}$
        if done **then**
            $displaym, score$
            $break$
        end if
    end for
    $replay \leftarrow batch\_size$
end for

Algorithm 1 implements the proposed multi-agent human monitoring framework. It takes as input a set of subjects $\mathcal{C} = 1, 2, \ldots, C$ and a set of vital signs $\mathcal{V} = 1, 2, \ldots, V$, along with the number of episodes $\mathcal{M} = 1, 2, \ldots, M$. The algorithm outputs the rewards achieved by agents in each episode. Lines 1–2 initializes all the parameters needed for monitoring the environment and learning agent. Lines 3–7 present the reward policy. Lines 8–14 present the function approximation using the neural networks model, memorize & replay, exploration & exploitation of the DRL agent. Lines 15–28 are nested for loops with conditional statements to check if the episode is completed or not. The outer loop is to iterate each episode while resetting the environment to initial values and score to zero. The inner loop is to iterate timesteps which denote the time of the current state and calls the methods.

The patient monitoring system operates with multiple agents, each tasked with monitoring specific vital signs such as heart rate, respiration rate, and temperature. The agents are initialized with a basic action set, which includes triggering alerts, adjusting monitoring intervals, and taking no action based on the patient's condition. At each time step, agents receive vital sign data as input and evaluate the patient's state. Based on the current state and the agent's policy, an action is selected. The reward function provides feedback based on the timeliness and accuracy of the action: positive rewards are given for correct, timely interventions, while penalties are applied for false alarms or missed emergencies. Over time, the agents improve their performance through continuous learning and collaboration, ensuring that vital signs are monitored comprehensively and interventions are timely (Fig. 3).

4 Experiment

In this study, the proposed multi-agent framework was evaluated by deploying an agent for each physiological feature of a different set of subjects. The aim of the learning agents was to monitor their respective vital signs, communicate with the corresponding MET based on the estimated level of emergency, and learn the subjects' behavior patterns. All the experiments were conducted using Python programming language version 3.7.6 and related libraries such as TensorFlow, Keras, Open Gym AI, and stable_baselines3.

4.1 Dataset

- PPG-DaLiA [37]: The dataset contains physiological and motion data of 15 subjects, recorded from both a wrist-worn device and a chest-worn device while the subjects were performing a wide range of activities under conditions close to real life.
- WESAD [38]: The WESAD (Wearable Stress and Affect Detection) dataset includes multimodal physiological signals such as ECG, PPG, GSR, respiration, and body temperature, recorded from 15 subjects while they performed a series of stress-inducing and affective tasks under laboratory conditions.

4.2 Baseline models

- WISEML [39]: Mallozzi et al. proposed an RL framework for runtime monitoring to prevent dangerous and safety-critical actions in safety-critical applications. In this framework, runtime monitoring is used to enforce properties to the agent and shape its reward during learning.
- CA-MQL [40]: Chen et al. proposed constrained action-based MQL (CA-MQL) for UAVs to autonomously make flight decisions that consider the uncertainty of the reference point location.
- MADDPG [41]: Lowe et al. introduced a deep reinforcement learning framework for multi-agent environments. This framework uses an adaptation of actor-critic methods to coordinate agents in both

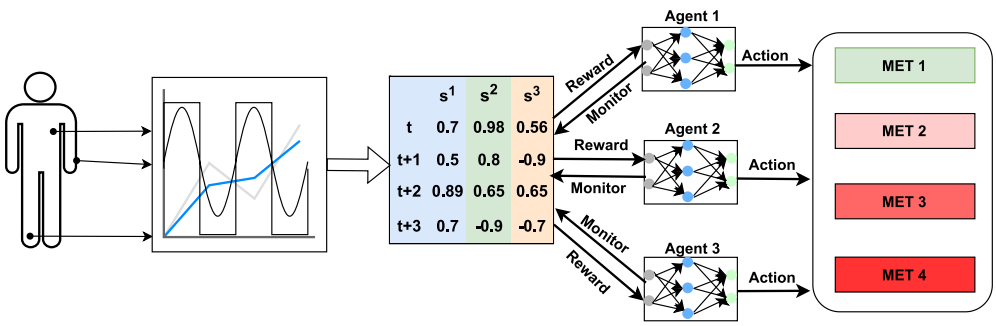

Fig. 3 Experiemental Design

cooperative and competitive settings by accounting for other agents' policies. It highlights the difficulty of traditional algorithms in multi-agent scenarios and introduces policy ensembles for more robust learning.

- QMIX [42]: Rashid et al. developed QMIX, a value-based multi-agent RL algorithm that factors joint action-values into per-agent values, allowing for decentralised policies while training in a centralised manner. QMIX demonstrated superior performance on challenging StarCraft II tasks by ensuring consistency between centralised and decentralised learning.
- Existing RL baseline models by Li et al. [32] were deployed to optimize sequential treatment strategies based on Electronic Health Records (EHRs) for chronic diseases using DQN. The multi-agent framework results were compared with Q-Learning and Double DQN.
- Similarly, RL was deployed to recognize human activity using a dynamic weight assignment network architecture with TD3 (a combination of Deep Deterministic Policy Gradient (DDPG), Actor-Critic, and DQN) by Guo et al. [33].
- Yom et al. [31] used Advantage Actor-Critic (A2C) and Proximal Policy Optimization (PPO) algorithms to act as virtual coaches in decision-making and send personalized messages.

4.3 Performance measures

In the initial phase, Cumulative Rewards were selected as the primary performance metric because they offer a direct reflection of the RL agents' success in achieving healthcare objectives. These cumulative rewards quantify the agents' ability to make correct decisions based on real-time physiological data, which is essential for ensuring timely medical interventions. Given the critical nature of healthcare systems, focusing on cumulative rewards allowed for the evaluation of how well the agents were trained to detect early signs of health deterioration.

To provide a more holistic evaluation, we introduced additional performance metrics:

- Learning Rate: This metric evaluates how quickly the agents converge to an optimal policy, which is vital in healthcare applications where rapid adaptation to changing patient conditions is crucial. Faster learning ensures that the agents can respond to emergencies in real time, improving the effectiveness of the system.
- Computational Complexity: This metric assesses the system's processing demands, particularly in terms of CPU/GPU time. Minimizing computational complexity is essential in healthcare settings with resource constraints, such as hospitals or wearable monitoring devices. Lower complexity ensures that the system can run efficiently without causing delays in decision-making.
- Memory Usage: As the system scales to monitor multiple physiological parameters across various agents, memory usage becomes a key factor. Efficient memory utilization is critical for deploying the framework on resource-constrained devices like wearables, ensuring scalability and adaptability without compromising performance.

Incorporating these metrics provides a more comprehensive evaluation of the proposed framework, ensuring not only its effectiveness in terms of rewards but also its efficiency, scalability, and real-world deployment potential in healthcare environments.

5 Experiment results and analysis

The advantage of RL for monitoring systems is that it can learn to handle complex, dynamic environments. Many monitoring tasks involve making decisions based on incomplete, uncertain information, and the optimal decision may depend on the context of the situation [43]. RL can learn to make decisions in these types of problems by considering the current state of the system and past experience. In this study, the aim is to leverage the RL capability to optimize the decision-making process in patient monitoring.

5.1 DRL agents performance

The performance of the proposed DRL framework was evaluated using two datasets, with a focus on cumulative rewards, learning rate, computational complexity, and memory usage. Additionally, we expanded our comparison to include multi-agent RL frameworks, MADDPG and QMIX, to assess how well these frameworks handle the complexities of real-time health monitoring tasks.

The results are summarized in Tab. 3, which includes the performance of single-agent RL methods (Q-Learning, PPO, A2C, and DDPG) and multi-agent RL models (MADDPG and QMIX). The proposed DRL framework consistently outperforms all other models in terms of cumulative rewards, with significant improvements over the baseline methods.

As shown in Tab. 3, the proposed DRL framework surpasses both MADDPG and QMIX in cumulative rewards for both datasets, particularly excelling in agent 1's performance on the PPG-DaLia dataset. This indicates that our framework's design, which includes a tailored reward mechanism based on Modified Early Warning Scores

(MEWS), enables more efficient learning in healthcare environments. Additionally, the exploration-exploitation trade-off in our system is better optimized for the variability of physiological data.

To provide a more intuitive assessment of the framework's performance, we evaluated the classification accuracy of the agents by comparing their actions against MEWS-derived ground truth escalation levels. Accuracy was computed as the ratio of correct escalation actions (e.g., MET-2 chosen when MEWS score corresponds to MET-2) to the total number of decisions made across episodes. This metric offers a clinically relevant view of agent performance, especially for practitioners accustomed to discrete outcome measures. The proposed DRL agents achieved an average decision accuracy of 88.3% across all episodes and subjects, outperforming baseline models such as PPO (79.1%), A2C (76.4%), and Double DQN (81.6%). These results demonstrate that the agents not only maximize cumulative rewards but also maintain high decision accuracy in real-time physiological monitoring tasks.

Beyond cumulative rewards, we evaluated the proposed DRL framework against baseline models using additional performance metrics, including learning rate, computational complexity, and memory usage, as shown in Tab. 4. The proposed DRL framework showed superior performance across all these metrics, indicating its suitability for real-time applications in resource-constrained healthcare environments.

In terms of learning rate, the proposed DRL framework converged after 850 epochs, outperforming all baseline models, including Q-Learning (1200 epochs) and Double DQN (1100 epochs). This faster convergence demonstrates the DRL framework's enhanced efficiency in learning complex healthcare scenarios. Faster learning is especially critical in healthcare, where timely interventions directly impact patient outcomes. The use of multiple agents, each dedicated to a specific physiological metric, accelerates policy optimization and enhances responsiveness in dynamic, real-world environments.

For computational complexity, the proposed DRL framework exhibited a significantly lower iteration time of 0.70 s, outperforming more complex multi-agent models like CA-MQL (1.30 s) and PPO (1.10 s). This indicates that the framework is computationally efficient, making it ideal for real-time healthcare monitoring where decision delays could compromise patient safety. This improved efficiency is due to an optimized reward structure and action space, which reduces the time required for decision-making without compromising accuracy.

In terms of memory usage, the DRL framework consumed 110MB, which is lower than all other baseline models, such as DDPG (160MB) and CA-MQL (175MB). This low memory footprint is crucial for deploying the framework on resource-constrained hardware like wearable devices or low-power hospital systems. The efficient memory usage ensures the system can scale with additional agents without overloading system resources, making the framework suitable for large-scale healthcare applications.

All three learning agents were fed with physiological features such as heart rate, respiration, and temperature, respectively, from the PPG-DaLiA dataset. Based on the observation space, action space, and reward policy defined for a customized gym environment for human behavior monitoring, the learning agents were run for 10 episodes, as shown in Fig. 4. In the results, agent 1 refers to the heart rate monitoring agent, which showed a constant increase in scores for each episode for most of the subjects except subjects 5 and 6. The intermittent low scores in agent 1 performance are due to the exploration rate in DQN learning, where the algorithm tries exploring all the actions randomly instead of relying on neural networks' predictions. Similarly, agent 2 and agent 3 monitor two other physiological features, respiration and temperature, respectively. agent 2 performed better than the other two agents and achieved consistent scores for all subjects. Out of all agents, agent 3, temperature monitoring performance, was poor. This issue was traced back to the data level, where the units of the temperature thresholds in the MEWS table and the input body temperature data from the dataset were different. Still, agent 3 achieved high scores in monitoring subjects 9, 8, 4, and 10.

The reward policy designed in the proposed multi-agent framework enables agents to learn the human physiological feature patterns. For example, if a subject's heart rate is 139 beats per minute, agent 1 takes Action 3 to communicate the message to MET-3. The agent will get rewarded with +10 points only if Action 3 is taken; otherwise, the agent gets negatively rewarded according to the reward policy (Table 2). With this example, the results in Fig. 4 can be interpreted better. An increase in scores episode by episode, with the exception of the exploration rate, actually infers an increase in the learning curve of the agents in terms of human physiological patterns.

5.2 Hyper-parameters optimization

The DRL agents were further evaluated by hyperparameters optimization. Out of all the hyperparameters discussed in this study, two hyperparameters, learning rate (α) and discount factor (γ), were optimized for all three agents, and the results are shown in Figs. 5 and 6. The learning rate determines how much information neural networks learn in an iteration to predict action and

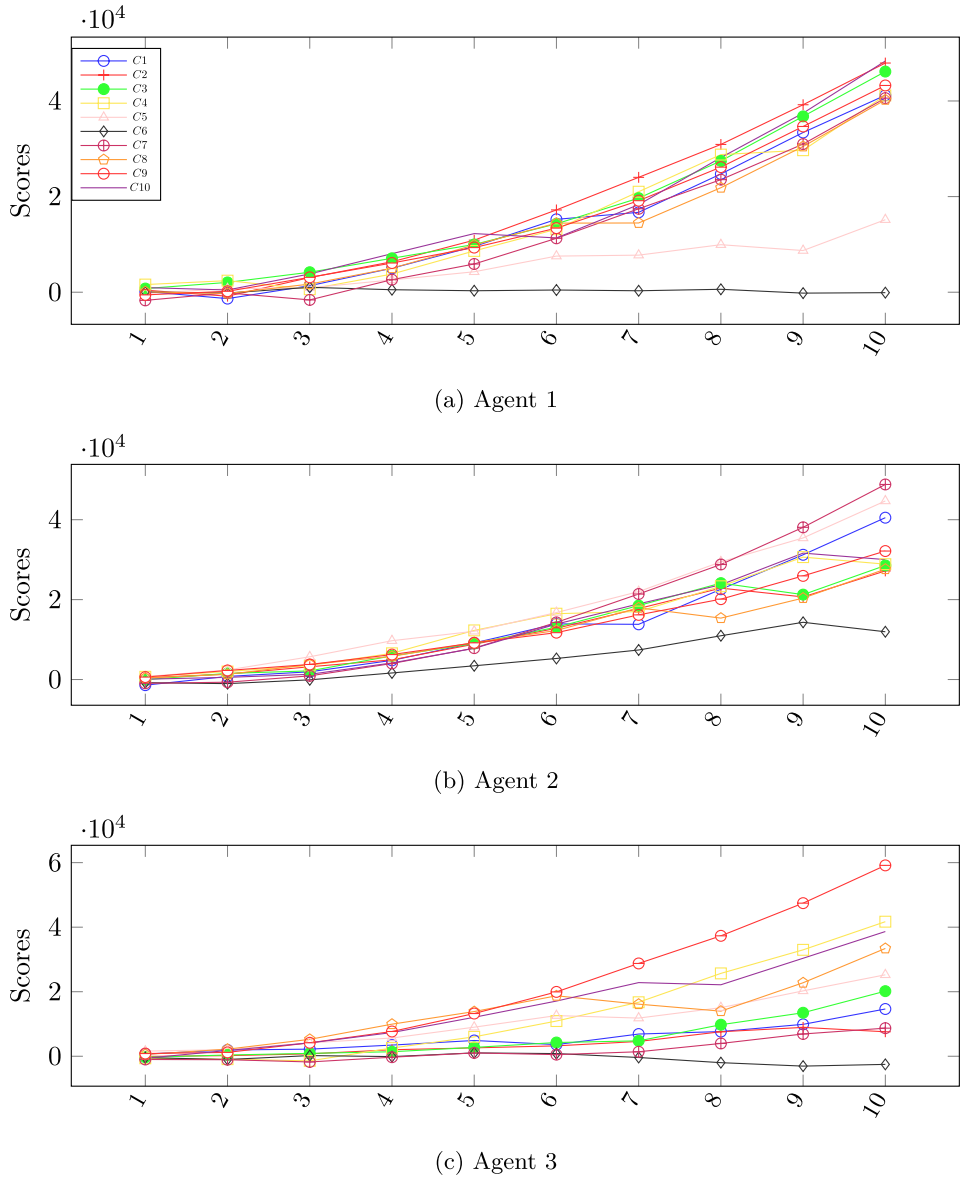

Fig. 4 DQN Agents Performance

approximate the rewards. The discount factor measures how much RL agents focus on future rewards relative to those in the immediate rewards. In Fig. 5, Fig. 5a show the agents' performance while optimizing α of neural networks. The x-axis of the plots represents scores (cumulative rewards) achieved by an agent in each episode shown on the y-axis. The bar plots show that the learning rate $\alpha = 0.01$ is a more optimized value in all the monitoring agents. Similarly, Figs. 6a present the γ optimization of agent 1, agent 2, and agent 3, respectively. The discount factors $\gamma = 0.9$ and $\gamma = 0.75$ are the more optimized values for agents 2 and 3, respectively, after 10 episodes of training.

Convergence Visualization and Hyperparameter Effectiveness. To provide a clearer view of how different hyperparameters affect model performance and convergence speed, we conducted additional experiments and visualized the episode-wise cumulative rewards under varying values of learning rate (α) and discount factor (γ). As shown in Figs. 5 and 6, learning rate $\alpha = 0.01$ led to faster and more stable convergence compared to higher or lower values, which either caused slower learning or higher variance across episodes. Similarly, $\gamma = 0.9$

Table 3 Comparison of DRL and MARL Frameworks on Cumulative Rewards

Method	PPG-DaLiA Dataset			WESAD Dataset		
	Agent 1	Agent 2	Agent 3	Agent 1	Agent 2	Agent 3
Q-Learning	25878	17304	23688	25318	16341	22823
PPO [31]	23688	20367	17688	23128	19404	16823
A2C [31]	24717	13707	24369	24157	12744	23504
Double DQN [32]	25569	15360	20367	25009	14397	19502
DDPG [33]	26760	20754	23967	26200	19791	23102
WISEML [39]	28654	25789	33669	28094	24826	32804
CA-MQL [40]	32985	27856	34685	32425	26893	33820
MADDPG [41]	42500	29870	36015	41200	28560	35345
QMIX [42]	44800	30520	37600	43200	29230	36980
Proposed DRL	48354	30019	38651	47794	29056	37786

resulted in optimal long-term reward accumulation, balancing future reward consideration with immediate decision-making. These visualizations offer intuitive insights into the convergence dynamics of the proposed DRL framework and reinforce our hyperparameter selection strategy.

Clinical Relevance of Cumulative Rewards The cumulative rewards obtained by the DRL agents are not arbitrary metrics but are directly linked to the agents' ability to make timely and clinically relevant decisions. Each reward is assigned based on how well an agent's action aligns with the MEWS-defined threshold for a given vital sign. For instance, if an agent detects an elevated heart rate indicative of stress-induced tachycardia and correctly escalates the condition to the appropriate MET level, it receives a positive reward. Conversely, a delayed or incorrect escalation results in a penalty. Over time, higher cumulative rewards indicate that the agents are successfully learning to respond to physiological deviations in ways that mirror clinical priorities. Thus, cumulative rewards in this framework serve as a quantitative proxy for the agents' effectiveness in the proactive monitoring and assessment of stress- and depression-linked health indicators.

Generalization Across Heterogeneous Conditions The ability to generalize across varying physiological patterns is essential for any real-world stress and depression monitoring system. While this study uses Modified Early Warning Scores (MEWS) to establish clinically informed reward boundaries, the reinforcement learning agents are not bound by fixed rules. Instead, they learn adaptive policies by interacting with dynamically evolving input states. To assess generalization, we employed two publicly available and heterogeneous datasets-PPG-DaLiA and WESAD-which differ in sensor configurations, experimental settings, and stress elicitation protocols. The consistent performance of our DRL agents across both datasets suggests promising generalizability. However, we acknowledge that additional validation on datasets encompassing richer behavioral modalities and more diverse populations is necessary to further substantiate this claim. Future extensions will focus on integrating multimodal data sources and deploying the framework in cross-domain learning environments to evaluate transferability and robustness under real-world conditions.

6 Discussion

This study introduces an innovative approach to patient monitoring within the unpredictable environment of healthcare settings, employing adaptive multi-agent deep reinforcement learning (DRL) to ensure timely healthcare interventions. The fluctuating nature of vital signs, crucial indicators of patient health, necessitates a robust system capable of real-time analysis and decision-making. Stress and depression, increasingly prevalent in modern healthcare contexts, are known to significantly impact vital signs such as heart rate, respiration, and temperature [16, 17]. By addressing these conditions, the proposed framework enhances early detection and intervention capabilities, which are critical for mitigating the physical and mental health risks associated with stress-induced tachycardia or depression-related bradycardia.

By leveraging the sequential decision-making prowess of RL algorithms, we have established a framework where each vital sign is monitored by a dedicated DRL agent. These agents operate within a cohesive monitoring environment, guided by meticulously defined reward policies to identify and respond to potential health emergencies based on MEWS and MET standards. This approach extends traditional patient monitoring by integrating the capacity to dynamically adapt to physiological changes

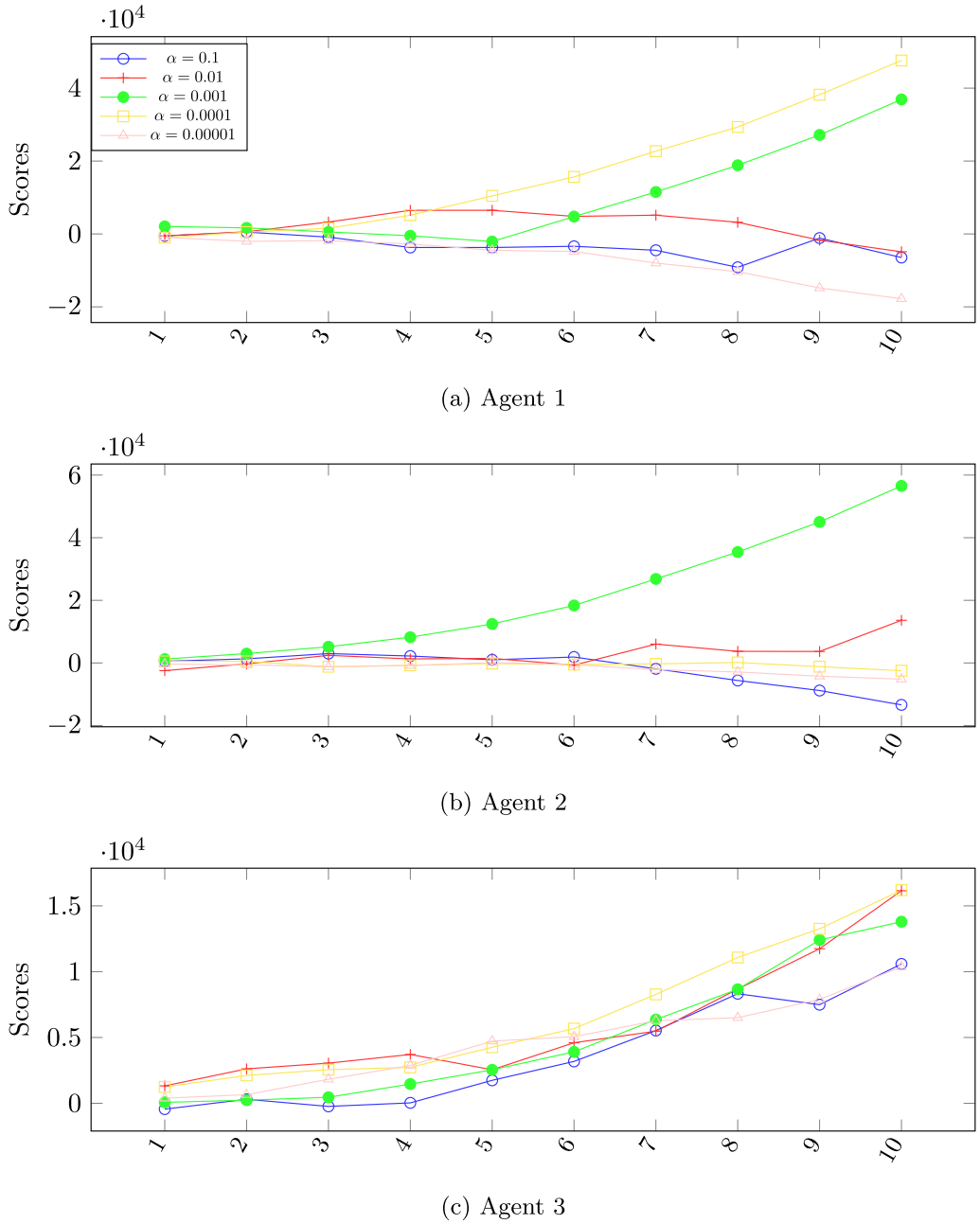

Fig. 5 Hyper Parameters - α optimization

influenced by mental health stressors, thereby providing a more comprehensive solution.

A notable aspect of our research is the emphasis on the design of the observation space for each DRL agent. This design is pivotal in ensuring the accuracy and effectiveness of the learning process, as it directly impacts the agent's ability to interpret vital sign data and make informed decisions. The challenge encountered with DRL agent 3, responsible for monitoring body temperature, underscores the importance of data consistency and the need for a harmonized observation space. The discrepancy between the temperature units in the MEWS table and the dataset highlighted a critical area for improvement, emphasizing the need for standardized data inputs to enhance agent performance and ensure reliability in detecting stress or depression-related anomalies.

The autonomous decision-making capability inherent in RL represents a significant advancement in

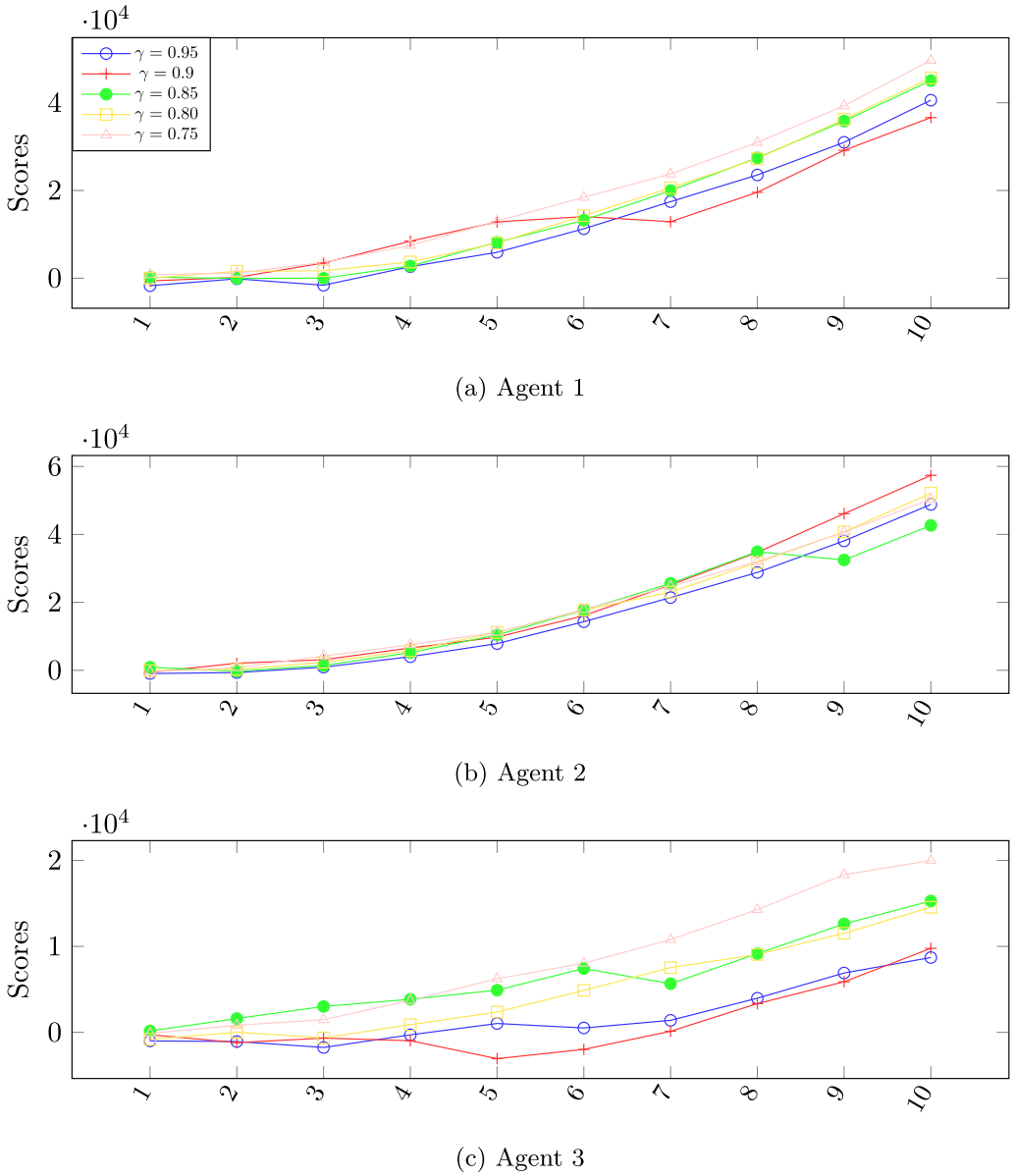

Fig. 6 Hyper Parameters - γ optimization

supporting clinicians. By providing real-time updates on patient health, the DRL framework facilitates a proactive approach to patient care, extending its applicability beyond hospital settings to include home monitoring and specialized care environments. This adaptability is further enhanced by the strategic optimization of hyperparameters, which fine-tunes the learning process of DRL agents to achieve optimal performance. Our investigation into hyperparameters such as the learning rate and discount factor reveals the critical balance between immediate and future rewards, a balance that is essential for the effective monitoring of patient health, particularly in cases where stress or depression can cause delayed yet significant physiological effects.

Comparatively, traditional supervised learning algorithms, while accurate in predicting vital signs, fall short in dynamic healthcare environments due to their reliance on extensive labeled datasets and external supervision. The DRL approach, free from the constraints of labeled data, offers a more flexible and efficient solution for patient monitoring. However, it is essential to acknowledge the considerable effort required in data preparation and model tuning within supervised learning frameworks, which, despite their limitations, contribute

Table 4 Evaluation of DRL Framework and Baseline Models on Additional Metrics

RL Method	Learning Rate (Epochs to Converge)	Computational Complexity (Time in Seconds)	Memory Usage (MB)
Q-Learning	1200	0.85s per iteration	120MB
PPO [31]	900	1.10s per iteration	150MB
A2C [31]	1000	1.05s per iteration	140MB
Double DQN [32]	1100	0.95s per iteration	135MB
DDPG [33]	950	1.20s per iteration	160MB
WISEML [39]	900	1.15s per iteration	145MB
CA-MQL [40]	1000	1.30s per iteration	175MB
MADDPG [41]	950	1.25s per iteration	155MB
QMIX [42]	1100	1.20s per iteration	165MB
Proposed DRL	850	0.70s per iteration	110MB

significantly to the development of informed clinical decisions.

The adaptive multi-agent DRL framework proposed in this study represents a paradigm shift in patient monitoring, offering a dynamic, efficient, and scalable solution for timely healthcare interventions [44]. By addressing both the physical and mental health challenges posed by stress and depression, this framework introduces a holistic approach to patient monitoring. The challenges and insights gleaned from this research pave the way for future advancements in the field, promising to enhance the quality of patient care through innovative technological solutions.

Scope of Physiological Monitoring. We acknowledge that stress and depression are highly complex psychophysiological conditions that cannot be comprehensively diagnosed through the monitoring of only three physiological parameters. In this study, the use of heart rate, respiration rate, and body temperature was intended as a proof-of-concept for evaluating the feasibility and performance of the proposed multi-agent DRL framework in a controlled setting. These variables were selected due to their well-documented correlation with acute stress responses and their widespread availability in wearable sensor systems. However, they serve as proxies for physiological arousal rather than definitive indicators of mental health status. The modular nature of our framework allows for the seamless integration of additional biosignals (e.g., GSR, HRV, EEG) or behavioral indicators (e.g., sleep disruption, speech features) in future work. As such, the current implementation should be viewed as a foundational step toward building a more comprehensive and multimodal system for mental health monitoring.

Explainability

While the proposed multi-agent DRL framework demonstrates strong adaptability and decision-making performance in patient monitoring, ensuring explainability remains a vital aspect for clinical adoption. To this end, we suggest incorporating agent-specific decision traceability as a foundational mechanism. Each agent can log transitions in Q-values alongside corresponding MEWS thresholds and selected actions, providing a transparent record of decision rationale over time. Such traceability supports retrospective audits by clinicians and aligns with the interpretability expectations of healthcare AI systems. Furthermore, future extensions of this work will explore the integration of model-agnostic interpretability techniques, such as SHapley Additive exPlanations (SHAP), to assess the contribution of each physiological feature to the agents' actions in real time. This dual approach-combining Q-value trajectory logging with post-hoc feature attribution-has the potential to enhance clinician trust, uncover failure points, and guide improvements in agent design. Emphasizing explainability is particularly important in sensitive contexts such as stress and depression monitoring, where transparent and accountable AI systems are essential for safe and ethical deployment.

Dataset Size and Generalizability. Although the proposed framework was evaluated using two widely recognized datasets-PPG-DaLiA and WESAD-each comprising 15 subjects, the size of these cohorts reflects an ongoing challenge in stress-related physiological research. Collecting high-quality, multimodal data under controlled conditions involving stress and affect remains inherently complex and resource-intensive, often limiting sample sizes across benchmark studies in this domain. Despite this constraint, the framework consistently demonstrated reliable policy convergence and adaptive learning across multiple agents and subjects, providing strong evidence of its robustness and effectiveness in modeling temporal patterns in physiological signals.

Importantly, the controlled nature of the datasets allowed for reproducible experimentation and precise evaluation of the technical capabilities of the multi-agent DRL system. Nonetheless, future work will aim to expand validation efforts using larger and more diverse datasets, potentially integrating synthetic data augmentation and transfer learning techniques to improve generalizability. These steps will ensure broader applicability of the proposed monitoring framework in real-world healthcare settings, while preserving the methodological rigor established in this study.

7 Conclusion

This study has pioneered an adaptive framework for healthcare interventions using multi-agent DRL to dynamically monitor vital signs, establishing a novel approach in patient care. By considering the significant influence of stress and depression on vital signs, this research underscores the importance of addressing mental health challenges in conjunction with physical health monitoring. Through the development of a generic monitoring environment coupled with a strategic reward policy, the DRL agents were empowered to learn from and adapt to vital sign fluctuations, enabling timely interventions by healthcare professionals. The ability of these agents to detect stress-induced or depression-related anomalies demonstrates the potential of this system to provide a comprehensive and proactive approach to healthcare.

Despite its innovative contributions, the research faced challenges, such as discrepancies in body temperature data scales and the absence of predictive capabilities for future vital sign trends, which limited the effectiveness of one DRL agent and the overall predictive potential of the system. These limitations highlight the need for enhanced data standardization and the integration of predictive analytics to anticipate trends in vital signs influenced by mental health conditions. Future research will focus on overcoming these challenges by augmenting the framework with predictive modeling capabilities, enabling DRL agents to forecast vital sign trends and anticipate health emergencies.

This advancement aims to revolutionize patient monitoring by facilitating proactive healthcare measures, significantly reducing the risk of critical health episodes associated with stress and depression. The future direction of this research will extend the scope to include multi-agent DRL frameworks capable of predicting future health trajectories, thereby enhancing the integration of mental and physical health monitoring in adaptive patient care systems.

Funding
Not applicable. No specific funding was received for this research.

Data Availability
The datasets used and analyzed during the current study are publicly available. Specifically, the PPG-DaLiA dataset [37] and the WESAD [38] were used in this research. Detailed instructions on accessing these datasets are provided in their respective publications.

Materials Availability
Not applicable. This study does not rely on specialized materials requiring Conflict of interest.

Code availability
he source code used to implement the multi-agent DRL monitoring framework, including the environment configuration and training routines, is publicly available at: https://github.com/Thanveer-Analyst/multi-agent-health-monitoring.git

Declarations

Ethics approval and consent to participate
Not applicable. This study does not involve human participants or animal studies requiring ethical approval.

Consent for Publication
All authors have provided their consent for the publication of this manuscript.

Competing interests
Author Xiaohui Tao is a member of the Editorial Board of the Journal Brain Informatics. The paper was handled by another Editor and has undergone a rigorous peer review process. Author Xiaohui Tao was not involved in the journal's peer review of, or decisions related to, this manuscript.

Author details
[1]School of Mathematics, Physics & Computing, University of Southern Queensland, Toowoomba, Australia. [2]School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China. [3]Division of Artificial Intelligence, School of Data Science, Lingnan University, Hong Kong, China. [4]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China. [5]Huazhong University of Science and Technology, Wuhan, China. [6]School of Business, University of Southern Queensland, Toowoomba, Australia.

Received: 17 January 2025 Accepted: 25 May 2025
Published online: 09 June 2025

References

1. Khalid A, Syed J (2024) Mental health and well-being at work: A systematic review of literature and directions for future research. Human Resour Manag Rev 34(1):100998
2. El-Rashidy N, El-Sappagh S, Islam SR, El-Bakry M, H, Abdelrazek S. (2021) Mobile health in remote patient monitoring for chronic diseases: principles, trends, and challenges. Diagnostics 11(4):607
3. Shaik T, Tao X, Higgins N, Xie H, Gururajan R, Zhou X (2022) Ai enabled rpm for mental health facility. In: Proceedings of the 1st ACM Workshop on Mobile and Wireless Sensing for Smart Healthcare, pp. 26–32
4. Pattanayak RK, Kumar VS, Raman K, Surya MM, Pooja MR (2022) E-commerce application with analytics for pharmaceutical industry. In: Advances in Intelligent Systems and Computing, pp. 291–298. Springer, ??? . https://doi.org/10.1007/978-981-19-3590-9_22
5. Thirunavukarasu R, C GPD, R G, Gopikrishnan M, Palanisamy V, (2022) Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review. Comput Biol Med 149:106020
6. Kiran BR, Sobh I, Talpaert V, Mannion P, Sallab AAA, Yogamani S, Perez P (2022) Deep reinforcement learning for autonomous driving: A survey. IEEE Trans Intell Transp Syst 23(6):4909–4926. https://doi.org/10.1109/tits.2021.3054625
7. Watts J, Khojandi A, Vasudevan R, Ramdhani R (2020) Optimizing individualized treatment planning for parkinson's disease using deep reinforcement learning. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, ??? . https://doi.org/10.1109/embc44109.2020.9175311
8. Naeem M, Paragliola G, Coronato A (2021) A reinforcement learning and deep learning based intelligent system for the support of impaired patients in home treatment. Exp Syst Appl 168:114285
9. Hong N, Liu C, Gao J, Han L, Chang F, Gong M, Su L (2022) State of the art of machine learning-enabled clinical decision support in intensive care units: Literature review. JMIR Med Inf 10(3):28781. https://doi.org/10.2196/28781
10. Chen IY, Joshi S, Ghassemi M, Ranganath R (2021) Probabilistic machine learning for healthcare. Ann Rev Biomed Data Sci 4(1):393–415. https://doi.org/10.1146/annurev-biodatasci-092820-033938
11. Khezeli K, Siegel S, Shickel B, Ozrazgat-Baslanti T, Bihorac A, Rashidi P (2023) Reinforcement learning for clinical applications. Clin J Am Soc Nephrol 18(4):521–523

12. Rastogi M, Vijarania DM, Goel DN (2022) Role of machine learning in healthcare sector. SSRN Electr J. https://doi.org/10.2139/ssrn.4195384
13. Mahesh B (2020) Machine learning algorithms-a review. Int J Sci Res (IJSR) 9:381–386
14. Palanisamy V, Thirunavukarasu R (2019) Implications of big data analytics in developing healthcare frameworks-a review. J King Saud Univ-Comput Inf Sci 31(4):415–425
15. Alsareii SA, Awais M, Alamri AM, AlAsmari MY, Irfan M, Aslam N, Raza M (2022) Physical activity monitoring and classification using machine learning techniques. Life 12(8):1103
16. Wang C, Wang J, Wang J, Zhang X (2020) Deep-reinforcement-learning-based autonomous uav navigation with sparse rewards. IEEE Int Things J 7(7):6180–6190
17. Mentis A-FA, Lee D, Roussos P (2024) Applications of artificial intelligence-machine learning for detection of stress: a critical overview. Mol Psychiatr 29(6):1882–1894
18. Oyeleye M, Chen T, Titarenko S, Antoniou G (2022) A predictive analysis of heart rates using machine learning techniques. Int J Environ Res Pub Health 19(4):2417. https://doi.org/10.3390/ijerph19042417
19. Luo M, Wu K (2020) Heart rate prediction model based on neural network. IOP Conf Series: Mater Sci Eng 715(1):012060. https://doi.org/10.1088/1757-899x/715/1/012060
20. Dang LM, Min K, Wang H, Piran MJ, Lee CH, Moon H (2020) Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognit 108:107561
21. Sheng T, Huber M (2020) Unsupervised embedding learning for human activity recognition using wearable sensor data. In: The Thirty-Third International Flairs Conference
22. Norgaard S, Saeedi R, Sasani K, Gebremedhin AH (2018) Synthetic sensor data generation for health applications: A supervised deep learning approach. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1164–1167 . IEEE
23. Sarker IH (2021) Machine learning: Algorithms, real-world applications and research directions. SN Computer Sci 2(3):160
24. Lou R, Lv Z, Dang S, Su T, Li X (2021) Application of machine learning in ocean data. Multimed Syst. https://doi.org/10.1007/s00530-020-00733-x
25. Tirumala D, Galashov A, Noh H, Hasenclever L, Pascanu R, Schwarz J, Desjardins G, Czarnecki WM, Ahuja A, Teh YW, et al. (2020) Behavior priors for efficient reinforcement learning
26. Janssen M, LeWarne C, Burk D, Averbeck BB (2022) Hierarchical reinforcement learning, sequential behavior, and the dorsal frontostriatal system. J Cognit Neurosci 34(8):1307–1325. https://doi.org/10.1162/jocn_a_01869
27. Tsiakas K, Papakostas M, Theofanidis M, Bell M, Mihalcea R, Wang S, Burzo M, Makedon F (2017) An interactive multisensing framework for personalized human robot collaboration and assistive training using reinforcement learning. In: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments. ACM, ??? . https://doi.org/10.1145/3056540.3076191
28. Kubota A, Riek LD (2022) Methods for robot behavior adaptation for cognitive neurorehabilitation. Ann Rev Control Robot Auton Syst 5(1):109–135. https://doi.org/10.1146/annurev-control-042920-093225
29. Pourpanah F, Etemad A (2024) Exploring the landscape of ubiquitous in-home health monitoring: a comprehensive survey. ACM Trans Comput Healthc 5(4):1–43
30. Lisowska A, Wilk S, Peleg M (2021) From personalized timely notification to healthy habit formation: a feasibility study of reinforcement learning approaches on synthetic data. In: SMARTERCARE@ AI* IA, pp. 7–18
31. Yom-Tov E, Feraru G, Kozdoba M, Mannor S, Tennenholtz M, Hochberg I (2017) Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. J Med Int Res 19(10):338. https://doi.org/10.2196/jmir.7994
32. Li T, Wang Z, Lu W, Zhang Q, Li D (2022) Electronic health records based reinforcement learning for treatment optimizing. Inf Syst 104:101878. https://doi.org/10.1016/j.is.2021.101878
33. Guo J, Liu Q, Chen E (2022) A deep reinforcement learning method for multimodal data fusion in action recognition. IEEE Signal Process Lett 29:120–124. https://doi.org/10.1109/lsp.2021.3128379
34. Yu C, Liu J, Nemati S, Yin G (2021) Reinforcement learning in healthcare: A survey. ACM Comput Surv 55:1. https://doi.org/10.1145/3477600
35. Shaik T, Tao X, Xie H, Li L, Zhu X, Li Q (2023) Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. arXiv preprint arXiv:2305.06360
36. Signs V (2021) Canberra hospital and health services clinical procedure
37. Reiss A, Indlekofer I, Schmidt P, Laerhoven KV (2019) Deep PPG: Large-scale heart rate estimation with convolutional neural networks. Sensors 19(14):3079. https://doi.org/10.3390/s19143079
38. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K (2018) Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 400–408
39. Mallozzi P, Castellano E, Pelliccione P, Schneider G, Tei K (2019) A runtime monitoring framework to enforce invariants on reinforcement learning agents exploring complex environments. In: 2019 IEEE/ACM 2nd International Workshop on Robotics Software Engineering (RoSE). IEEE, ??? . https://doi.org/10.1109/rose.2019.00011
40. Chen Y-J, Chang D-K, Zhang C (2020) Autonomous tracking using a swarm of UAVs: A constrained multi-agent reinforcement learning approach. IEEE Trans Veh Technol 69(11):13702–13717
41. Lowe R, Wu YI, Tamar A, Harb J, Pieter Abbeel O, Mordatch I (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. Adv Neural Inf Process Syst 30:1
42. Rashid T, Samvelyan M, De Witt CS, Farquhar G, Foerster J, Whiteson S (2020) Monotonic value function factorisation for deep multi-agent reinforcement learning. J Mach Learn Res 21(178):1–51
43. Schippers MC, Rus DC (2021) Optimizing decision-making processes in times of covid-19: using reflexivity to counteract information-processing failures. Front Psychol 12:650525
44. Shaik T, Tao X, Xie H, Li L, Yong J, Li Y (2024) Graph-enabled reinforcement learning for time series forecasting with adaptive intelligence. IEEE Transactions on Emerging Topics in Computational Intelligence

Publisher's Note