IJE TRANSACTIONS B: Applications Vol. 37 No. 11, (November 2024) 2314-2324

International Journal of Engineering

Journal Homepage: www.ije.ir

Energy and Throughput Management in Wireless Body Area Network with Wireless Information and Energy Transfer using Reinforcement Learning

Z. Rashidi, M. Majidi*

Department of Electrical and Computer Engineering, University of Kashan, Kashan, Iran

PAPER INFO

Paper history:
Received 07 October 2023
Received in revised form 16 April 2024
Accepted 10 May 2024

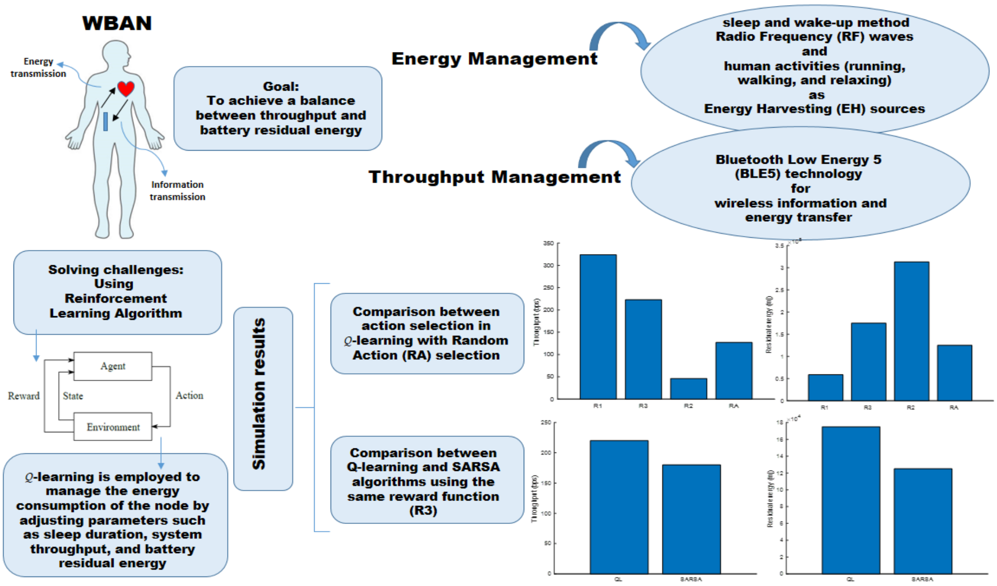
Keywords:
Wireless Information and Energy Transfer
Energy Harvesting
Wireless Body Area Network
Reinforcement Learning
Q-learning

ABSTRACT

In this paper, we address the challenges of energy and throughput management in a Wireless Body Area Network (WBAN) with a focus on a heart rate sensor. Our approach utilizes the sleep and wake-up method to minimize sensor energy consumption while harnessing Radio Frequency (RF) waves and human activities (running, walking, and relaxing) as Energy Harvesting (EH) sources to supplement battery power. Bluetooth Low Energy 5 (BLE5) technology is employed for wireless information and energy transfer. Our goal is to strike a balance between throughput and battery residual energy. The advantages of using Q-learning for action selection in comparison to Random Action (RA) selection are demonstrated through simulations. The results reveal that the reward function in Q-learning, incorporating a balancing parameter, effectively achieves a compromise between throughput and battery residual energy. Additionally, our Q-learning method improves system throughput by 43% compared to RA selection. In addition, we compare the performance of the Q-learning and State- Action- Reward-State- Action (SARSA) algorithms using the same reward function to evaluate their respective abilities in managing system throughput and battery residual energy. These findings have significant implications for developing energy-efficient WBANs, enabling prolonged operation and reliable data transmission.

doi: *10.5829/ije.2024.37.11b.16*

Graphical Abstract



WBAN

Goal: To achieve a balance between throughput and battery residual energy

Energy Management

sleep and wake-up method Radio Frequency (RF) waves and human activities (running, walking, and relaxing) as Energy Harvesting (EH) sources

Throughput Management

Bluetooth Low Energy 5 (BLE5) technology for wireless information and energy transfer

Solving challenges: Using Reinforcement Learning Algorithm

Q-learning is employed to manage the energy consumption of the node by adjusting parameters such as sleep duration, system throughput, and battery residual energy

Simulation results

Comparison between action selection in Q-learning with Random Action (RA) selection

Comparison between Q-learning and SARSA algorithms using the same reward function (R3)

*Corresponding Author Email: m.majidi@kashanu.ac.ir (M. Majidi)

Please cite this article as: Rashidi Z, Majidi M. Energy and Throughput Management in Wireless Body Area Network with Wireless Information and Energy Transfer using Reinforcement Learning. International Journal of Engineering, Transactions B: Applications. 2024; 37(11): 2314-24.

NOMENCLATURE			
P_0^{dB}	Received power reference	P_s	Periods between each measurement
n	Characteristic of path loss	N_b	Number of sampled data
$\mathcal{P}_c$	Power consumption of coordinator	T_s	Sampling time
d_0	Reference distance	T_t	Transmission time
d	Distance from the sensor to the coordinator	**Greek Symbols**	
h^{dB}	Channel gain between the sensor and the coordinator	α	Learning rate
E_0	Initial energy	γ	Discount factor
$E_r(t)$	Battery residual energy at time *t*	μ	Mean
$E_H(t)$	Harvested energy at time *t*	σ	Variance
$\|h_{LN}\|^2$	Lognormal channel gain between sensor and coordinator	ε	Probability of randomly selecting an action
$\|h_{PL}\|^2$	Path-loss channel gain between sensor and coordinator	η	Conversion efficiency
$E_c(t)$	Energy consumption at time *t*	β	Balancing parameter
N	Epochs/Episodes		

1. INTRODUCTION

In recent years, there has been a growing emphasis in engineering on the integration of telemedicine with the Internet of Things (IoT), leading to the increasing prevalence of IoT-enabled structures, environments, and systems (1, 2). Wireless Sensor Networks (WSNs), which have been rapidly growing over the past decade primarily due to their efficiency, comprise many sensor nodes and are widely deployed in sensor environments for data collection and evaluation (3). The WSNs have sparked significant interest in both academia and industry due to their application in health monitoring. Specifically, WSNs leverage Wireless Body Area Networks (WBANs) as a crucial component of the emerging IoT. WBANs have attracted significant attention in recent years, highlighting their importance and potential in various fields (4, 5). WBANs typically comprise multiple sensor nodes that are located on or inside the human body. These sensors periodically transmit physiological and multimedia data to the source, which can then be shared with healthcare providers, emergency services, or family members to monitor the patient's vital signs, take prompt action in case of an emergency, stay informed about the patient's condition, and update medical records accordingly (6, 7). As the most widely adopted protocol in consumer products, Bluetooth Low Energy (BLE) is crucial in facilitating IoT applications and is frequently used for transmitting data from wearable devices. The BLE5, has been designed to meet the diverse needs of IoT use cases and can transfer the same volume of data in less time than previous versions (8, 9).

WBANs are primarily used for data collection, storage, processing, and transmission. Replacing batteries frequently can be highly inconvenient, particularly for implanted devices, and the limitations of both size and battery power in a WBAN can have a notable impact on its usability and user satisfaction. To address this issue, Energy Harvesting (EH) technology has emerged as a popular area of research (10).

EH involves gathering energy from the surrounding environment and converting it into electrical power. The development of EH technology holds the potential to substantially extend the lifetime of wireless networks (11, 12). Wireless Power Transfer (WPT) can be combined with Wireless Information Transfer (WIT) to utilize radio waves for both communication and energy transfer. A unified Wireless Information and Power Transfer (WIPT) design could leverage the Radio Frequency (RF) spectrum and network infrastructure for efficient communication and power delivery, enabling greater patient mobility (13, 14).

Reinforcement learning (RL) involves the dynamic learning process of adjusting actions based on continuous feedback from the environment to maximize a reward (15). RL algorithms have gained traction in recent years for managing energy and throughput. These algorithms adjust a node's behavior by incentivizing good decisions through a reward function. Since RL is designed to make decisions in uncertain environments, it is well-suited for energy management in systems that utilize EH technologies (16).

Energy efficiency is a critical consideration in designing WBANs, and many studies are currently focused on optimizing energy consumption.

To prolong the lifespan of sensor nodes in WBANs, various energy management methods have been developed, such as EH, sleep and wake-up mechanism. Gupta and Chaurasiya (17) presented an energy management system based on RL algorithms and explored health surveillance in WBANs. The paper investigates several EH models, including vibration, solar, and thermal energy sources, to enhance energy efficiency. Implementing a sleep and wake-up mechanism can help conserve energy by allowing the sensor to access data only when necessary. In this mechanism, the sensor transitions to a wake-up state when data transmission is required, and returns to a sleep state after transmitting the data (17). Also, Wang et al. (18) introduced the blood pressure sensor node that reduces energy consumption by switching between sleep and wake-up modes to optimize the duty cycle. Xu et al. (19) proposed an EH technology as a solution to address the issue of energy efficiency, with the ability to collect energy from surrounding or environmental sources using RL algorithms.

In RL methods, the agent strives to acquire the optimal policy by engaging with the environment and gaining experience through iterative experimentation. Using EH-WBAN networks, Mohammadi and Shirmohammadi (20) proposed a Q-learning-based sleep and wake-up scheduling method. This method enhanced energy efficiency, reduced network delay, and maintained network connectivity. Rioual et al. (21) proposed a method of supplying energy to sensor nodes using the body's biomechanical energy sources, specifically piezoelectric energy. Various reward functions have been developed for the sensor node to evaluate its efficiency in determining the optimal duty cycle. This approach enables the sensor node to learn and adapt to uncertain EH conditions (21, 22). Rioual et al. (22) investigated various reward functions to identify the most appropriate variables for achieving the desired performance. Experimental results were obtained by comparing the different functions, highlighting the impact of reward functions on energy consumption and their potential to optimize energy management.

Only a limited number of papers within the field of WBANs have explored the combination of EH and RL techniques for effectively managing throughput across the entire network node. Ge et al. (23) proposed a dynamic clustering protocol for WSNs that incorporates considerations of both the remaining energy of sensor nodes and the predicted levels of harvestable energy in each iteration to create several uneven nodes. To enhance throughput in a clustered network, the Cooperative Q-Learning and State- Action- Reward- State- Action (SARSA) algorithm is employed. To achieve this goal, the Cooperative Reinforcement Learning (CRL) algorithm is utilized to ascertain the ideal quantity of packets to transmit from each sensor at every time step. This approach aims to maximize the throughput of the network. Roy et al. (24) proposed an EH protocol based on RL, which is designed to optimize resource allocation and maximize throughput while minimizing delay.

In this paper, a one-tier WBAN is considered with a sensor node and a coordinator node. The sensor node transmits essential data extracted from the body to the coordinator node. The required energy for the sensor is provided with the help of the WIPT technique, and harvesting wireless energy from RF sources and body energy sources. The paper brings four several key contributions, which can be outlined as follows:

i. The wireless information transfer from the sensor to the coordinator and the wireless energy transfer from the coordinator to the sensor is done with BLE5.

ii. The energy consumption of node components in the sleep and wake-up modes for energy efficiency is calculated and considered in RL.

iii. Selecting an appropriate reward function is a complex task. As the system's behavior is determined by this function, it becomes crucial for system designers to make this choice carefully. Therefore, in this research, we have investigated the impact of various reward functions employed in a widely used RL algorithm, namely Q-learning.

iv. Three distinct reward functions ($\mathcal{R}_1 - \mathcal{R}_3$) are compared and evaluated to identify the optimal variables for designing a function that effectively balances management between system battery residual energy and throughput.

v. Finally, the Q-learning algorithm is compared to the SARSA algorithm.

The subsequent sections of the paper are organized as follows. Section 2 presents our use case, and section 3 introduces the RL mechanism. Section 4 presents our simulation results wherein, three reward functions are compared and evaluated. Finally, section 5 provides a summary and concludes the paper.

2. SYSTEM MODEL

The system model in Figure 1, is a one-tier WBAN, comprising a sensor node (S) and a coordinator node (C). The sensor node is positioned on the chest of a human to observe heart activity and incorporates a heart rate detection sensor, a low-power microcontroller unit (MCU), a BLE5 transceiver, and a battery. Additionally, as illustrated in Figure 1, the coordinator is positioned on the body's waist, specifically on the right side. The power consumption of each component is summarized in Table 1. Once the data related to heart rate is collected, it is transmitted promptly to the coordinator for processing.

In addition to a battery-powered sensor, the sensor is equipped with RF and Body Energy Harvesting (BEH) system. Although their values are not large, they still can extend the node's lifespan.

2. 1. The Process of Information and Energy Transfer in The System Model

According to Table 2, the wireless transfer of both information and energy occurs in two phases, namely WIT and Wireless Energy Transfer (WET). The sensor is awake in the WIT phase, and it measures and stores heart rate information and then sends it to the coordinator node. During the WET phase, the sensor is asleep and it receives an RF energy signal from the coordinator node. Also, the sensor extracts energy from the body to power the sensor during periods of both activity and sleep. In Figure 1 the solid line represents the transmission of information from the sensor to the coordinator ($s \rightarrow c$), and the dashed line represents the transmission of RF energy from the coordinator to the sensor ($c \rightarrow s$).

2. 2. Problem Formulation For the channel model being discussed, a Lognormal distribution is assumed to describe the link between the sensor and coordinator. The Lognormal distribution is defined by its mean value, denoted by μ, and its variance, represented by σ (Equation 1). These parameters are used to characterize the channel gain between the sensor and the coordinator (25).

$$h^{dB} \sim \mathcal{N}(\mu, \sigma) \tag{1}$$

In addition to the Lognormal distribution, we also utilize the channel path loss model between the sensor and the coordinator (Equation 2). In this equation, d is the distance from the sensor to the coordinator, n is characteristic of the path loss, and P_0^{dB} is the received power reference at distance d_0 (26).

$$P^{dB} = P_0^{dB} - 10n \log_{10}(\frac{d}{d_0}) \tag{2}$$

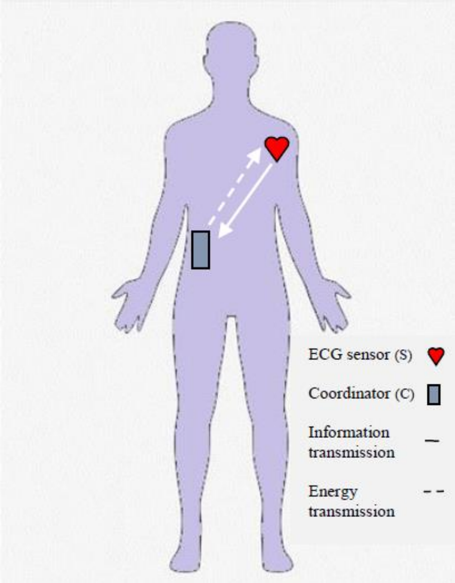


Figure 1. System model

TABLE 1. The power consumption of sensor components (9, 22)

Sensor components	Wake-up mode	Sleep mode
Heart rate monitor sensor	5.28 mW	0.396 mW
BLE5	29 W	0 mW
MCU	1.7 mW	0.00257 mW

TABLE 2. The process of information and energy transfer

	Wake-up	Sleep
WIT	$S \rightarrow C$	
WET		$C \rightarrow S$
BEH	√	√

The amount of sensor energy is initially considered equal to E_0, which is equal to the initial energy of the battery as follows:

$$E_r(0) = E_0 \tag{3}$$

Battery residual energy at time t $(E_r(t))$ is calculated according to Equation 4:

$$E_r(t) = E_r(t-1) + E_H(t) - E_c(t), \tag{4}$$

where the harvested energy ($E_H(t)$) includes BEH and the harvested RF signal energy in time t, given by:

$$E_H(t) = (60 * \mathcal{P}_{EH}(s_t)) + (\eta\, \mathcal{P}_c |h_{LN}|^2 |h_{PL}|^2 * sleep\ time\ (a_t)), \tag{5}$$

$\mathcal{P}_{EH}(s_t)$ is the power harvested from the body by the heart node, and its amount depends on the state or activity of the person at time t. Since we have the BEH during all times of sleep and wake-up, and the amount of battery residual energy is calculated for every minute, the harvested power is multiplied by 60. $|h_{LN}|^2$ is Lognormal channel gain, and $|h_{PL}|^2$ is path-loss channel gain between sensor and coordinator. η is the efficiency of energy conversion and, $\mathcal{P}_c$ denotes the power consumed by the coordinator to transmit the RF energy signal. Also, $sleep\ time\ (a_t)$ represents the sleep time duration during which the sensor receives RF energy, and its quantity depends on the action at time t, i.e., a_t.

The energy consumption ($E_c(t)$) includes the energy consumed by the sensor components during sleep and wake-up time as follows:

$$E_c(t) = E_{ECG}^{W}(a_t) + E_{BLE}^{W}(a_t) + E_{MCU}^{W}(a_t) + E_{ECG}^{S}(a_t) + E_{MCU}^{S}(a_t) \tag{6}$$

where $E_{ECG}^{W}(a_t)$ and $E_{ECG}^{S}(a_t)$ are energy consumed by the heart rate sensor during wake-up and sleep time respectively, $E_{MCU}^{W}(a_t)$ and $E_{MCU}^{S}(a_t)$ are energy consumed by the MCU during wake-up and sleep time respectively, and $E_{BLE}^{W}(a_t)$ is the energy consumed by BLE5 in wake-up time. All of these consumed energies depend on the action at that specific time. Since BLE5 is not active during sleep time, its energy consumption is considered zero, hence not included in the formula.

3. REINFORCEMENT LEARNING

RL algorithms (Q-learning and SARSA) possess the capability to learn the optimal policy for interacting with an environment by utilizing rewards. In each step, the agent selects the best possible action based on a policy, given the current state. This action leads to a change in the environment's state, and the agent receives an immediate reward signal in an ideal scenario (Figure 2). Despite lacking prior knowledge of the environment, the agent acquires an optimal policy, which is a mapping from states to actions. The policy is learned through a process of trial and error, or exploration and exploitation, where the agent explores different actions to determine the best course of action in a given state. In general, The primary objective of the agent is to maximize its cumulative reward over an extended period (27).

In RL, there is a trade-off between exploration and exploitation. Exploration involves randomly selecting an action to investigate the usefulness of that action. On the other hand, exploitation involves using actions that were previously learned to be useful based on their utility (28).

The epsilon-greedy method is a strategy employed to strike a balance between exploration and exploitation while training RL policies. For instance, when utilizing the epsilon greedy method, the parameter ε determines the probability of randomly selecting an action from the action space. With a probability of 1-ε, the output action is chosen greedily based on the argmax (Q) function. A variant of the epsilon-greedy method that exhibits improvement is known as the decayed-epsilon-greedy method. In this approach, for instance, during the training process consisting of N epochs/episodes (depending on the specific problem), the algorithm begins by setting $\varepsilon = p_{init}$. Subsequently, throughout n_{step} training epochs/episodes, ε gradually decreases until it reaches $\varepsilon = p_{end}$. In the initial stages of the training process, the model is given greater freedom to explore with a high probability. Subsequently, the value of ε is gradually reduced with a rate ζ (Equation 7) over successive training epochs/episodes, as described in Equation 8 (29):

$$\zeta = \max\left(\frac{N-n_{step}}{N}, 0\right) \tag{7}$$

$$\varepsilon = (p_{init} - p_{end})\zeta + p_{end} \tag{8}$$

By adopting this more adaptable approach, particularly when the exploration probability reaches the low threshold value p_{end} after n_{step}, the training process can prioritize exploitation (i.e., greedy) while still maintaining a minimal capacity for exploration. After selecting the action, the next state and the corresponding reward function are obtained, and the current state Q(s, a) in the Q-Table of Q-learning algorithm is updated as:

$$\begin{gathered} Q(s_t, a_t) = Q(s_t, a_t) + \\ \alpha\left[r_{t+1}\gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right], \end{gathered} \tag{9}$$

and the current state Q(s, a) in the Q-Table of the SARSA algorithm is updated as:

$$\begin{gathered} Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \\ \gamma\, Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)], \end{gathered} \tag{10}$$

where $Q(s_t, a_t)$ denotes the estimated Q-value of taking action a in state s at time t, r_{t+1} represents the immediate reward obtained at time t+1. The maximum Q-value of taking the optimal action a in state s at time t+1 is represented by $\max_a Q(s_{t+1}, a_{t+1})$. $Q(s_{t+1}, a_{t+1})$ denotes the Q-value of taking the action a in state s at time t+1. The learning rate, $0 \leqslant \alpha \leqslant 1$, defines the speed at which new information replaces old information and the discount factor, $0 \leqslant \gamma \leqslant 1$, determines the significance of future rewards, which are used in the equations. The outcomes of the equations are stored in a policy table referred to as the Q-table. In this table, the rows correspond to the available states, the columns represent the feasible actions, and the cells contain the expected total reward (30).

3. 1. Framework of Reinforcement Learning

The environment in our system is a WBAN, where the sensor node (agent) interacts with the environment through its actions.

State space: The state space is partitioned into three different states, each associated with the activity of the individual wearing the device (Table 3). This table displays the amount of power that can be harvested based on the activity. In our work, it is assumed that the activity or state changes randomly every 30 minutes.

Action space: According to Table 4 we establish a collection of actions with different periods between each measurement (P_s), sampled data (N_b), duration time of sampling (T_s), and transfer data time (T_t). The heart rate is determined based on the collected data bits over 5, 10, and 15-second intervals, which represent the T_s. Kwon et al. (31) demonstrated that sampling frequency F_s=250 Hz provides excellent results for the examination of heart rate variability. So, the rate of sending data bits is R_b=2$Kbps$. For instance, action 1 entails collecting 30 Kb of sampled data with a sampling duration of 15 seconds and measurements taken every minute. Each action possesses unique energy consumption levels due to its reliance on sampled data in wake-up mode and varying sleep times. Our agent selects an action from Table 4 every 20 minutes.

Reward: The first reward $\mathcal{R}_1$ considers system throughput given by:

$$\mathcal{R}_1 = \frac{\mathcal{N}_b}{Ps * 60} \tag{11}$$

The second reward $\mathcal{R}_2$ (Equation 12) considers the system energy. $E_r(t)$ is the residual energy in the node's battery at time t.

$$\mathcal{R}_2 = \frac{E_r(t)}{E_0} \tag{12}$$

The third reward (Equation 13) is the combination of the two above rewards of system throughput and energy.

$$\mathcal{R}_3 = (\beta)\,\frac{\mathcal{N}_b}{\mathcal{P}_s * 60} + (1-\beta)\,\frac{E_r(t)}{E_0} \tag{13}$$

The primary aim of this system is to minimize energy consumption while simultaneously maximizing throughput, although these goals are conflicting with one another. The primary objective of the RL algorithm is to find the balance point that satisfies these constraints. For this purpose, the third reward function incorporates the parameter β to balance the throughput and battery residual energy. Rewards and Q-Table are updated every 10 minutes.

4. SIMULATION RESULTS

In this section, the simulation results are examined. The first part focuses on comparing the rewards in Q-learning and RA selection, while the second part compares the Q-learning algorithm with the SARSA algorithm reported by Ge et al. (23).

The simulation of this system is conducted using MATLAB, with a total simulation time of 4500 minutes, equivalent to three full days. For the simulations, a single electrocardiogram (ECG) sensor with a battery capacity of 100 mAh has been considered. The Q-Table is structured as a (3×5) matrix, where the columns represent the number of states, and the rows represent the number of actions. The parameters used in the simulations are described in Table 5.

4. 1. The Comparison of Rewards in Q-learning with RA Selection

Algorithm 1 shows the steps of our Q-learning method. In the RA selection, the environment, states, and actions remain the same as in Q-learning. However, unlike Q-learning, there are no rewards or updates to the Q-Table. Actions are chosen randomly without any consideration of past experiences or rewards.

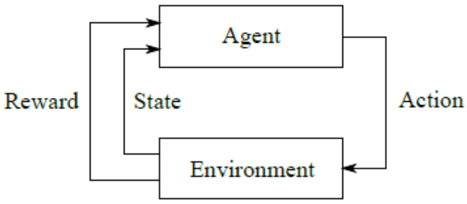


Figure 2. Interaction between an agent and its environment

TABLE 3. Kinetic motion harvested power for three different activities (32)

Activity	Power harvested
Relaxing	2.4 μW
Running	180.3 μW
Walking	678.3 μW

TABLE 4. Set of actions with different periods between each measurement (P_s) and sampled data (N_b) at different sampling and transmission times

Action	P_s (min)	N_b (Kb)	T_s (s)	T_t (s)
1	1	30	15	0.03
2	1	20	10	0.02
3	5	20	10	0.02
4	20	20	10	0.02
5	60	10	5	0.01

TABLE 5. Simulation parameters

Symbol	Value
E_0 (21)	3600 mAh
α	0.1
γ	0.9
β	0.7
P_0^{dB}(25)	-16.6
n (25)	1.29
d_0 (25)	10 cm
μ (25)	-0.72
σ (25)	2.67
$\mathcal{P}_c$ (7)	1 mW
η (7)	0.8

Algorithm 1: Q-learning algorithm

```
Initialize $Q$-table as $Q(s_t, a_t)$ = 0, with size ($s$=3 ,$a$=5)
The agent observes the initial state $s_0$
for each time do
    Calculate the $E_r(t)$ in the battery every minute
    Choose an action $a$ every 20 minutes as follows:
    if rand (0,1) > $\varepsilon$
        Select the action that has the maximum $Q$-value in the current state of $Q$-table
    else
        Select an action randomly in the action space
    end if
    Choose randomly a state every 30 minutes
    Get a reward $r$ every 10 minutes
    Update the related $Q$-value every 10 minutes as follows
        $Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1}\gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$
end for
```

In Figure 3(a), (b), and (c) represent the actions taken under reward functions $\mathcal{R}_1$, $\mathcal{R}_3$, and $\mathcal{R}_2$, respectively, and illustrate how the agent selects actions over time. Initially (as indicated by the dashed line), the agent lacks any prior knowledge of its environment. Consequently, it adopts a strategy of selecting random actions to explore and gain a better understanding of the environment. After 2000 minutes, the agent's behavior adapts more, and it learns to take actions that optimize the reward. The reward function $\mathcal{R}_1$ determines the reward by considering both sleep time and system throughput. The highest possible reward is achieved when sleep time is minimized and throughput is maximized.

According to Figure 3(a), the agent selects the high-consumption action (the first action) because it has P_s =1 and the highest throughput. This action results in increased node consumption due to its high energy usage. The reward function $\mathcal{R}_2$ calculates the reward based on the residual energy in the battery. This function produces the maximum reward value when the residual energy in the battery is high. According to (c), the agent selects the least energy-consuming action (the fifth action) because it has the highest sleep time (P_s = 60 minutes) and the lowest throughput. These actions contribute to preserving battery energy, which results in increased residual battery energy after 4500 minutes. The reward function $\mathcal{R}_3$ considers sleep time, system throughput, and battery residual energy to compute the reward. The highest reward value is achieved when both the battery residual energy and throughput are high. However, since we aim to balance the conflicting goals of throughput and energy consumption, the β parameter can be used to manage the node's behavior. In this case, the β parameter is set to a fixed value of 0.7 to achieve good management of the system. (b) indicates that the node selects actions 2 and 5. In the RA scenario, actions are chosen randomly without any consideration of rewards. As a result, it can be observed in (d) that action selection does not progress in a way that achieves the intended purpose of the system.

Figure 4 represents the system throughput under $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ reward functions compared to RA selection. As previously mentioned, the agent is trained to take actions that maximize the reward. When using the reward function $\mathcal{R}_1$, the system throughput is nearly 7 times higher than when using the reward function $\mathcal{R}_2$. Our goal is to manage system throughput so that it is neither too high nor too low. For this reason, the agent's behavior is more acceptable when using the reward function $\mathcal{R}_3$. The RA scenario shows the system throughput when actions are randomly selected. Compared to reward function $\mathcal{R}_3$, it can be seen that Q-learning leads to a 43% increase in system throughput. This demonstrates the superiority of the ε-greedy action selection method in Q-learning over RA selection.

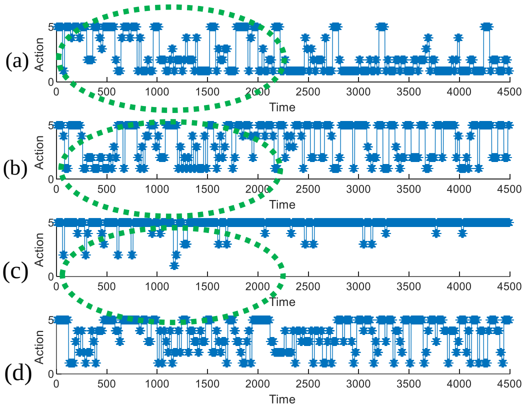


Figure 3. The selected actions when the rewards are (a) $\mathcal{R}_1$, (b) $\mathcal{R}_3$, and (c) $\mathcal{R}_2$ in terms of time compared to (d) RA selection

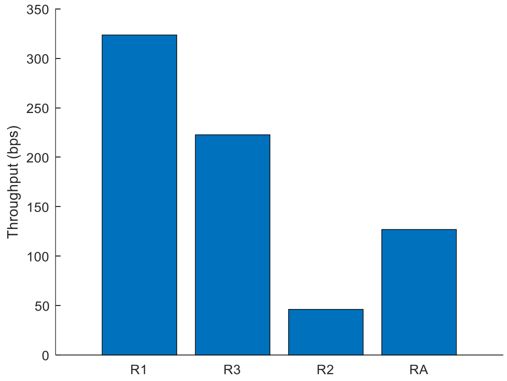


Figure 4. System throughput when the rewards are $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ compared to RA selection

Figure 5 represents the battery residual energy under $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ reward functions compared to RA selection. The figure shows that the reward function $\mathcal{R}_1$ depletes the battery energy quickly, while reward function $\mathcal{R}_2$ minimizes energy consumption, resulting in a battery charge of over 75% at the end of the simulation. However, $\mathcal{R}_2$'s low throughput indicates that it may not be a desirable reward function. According to the figure, the reward function $\mathcal{R}_3$ results in a battery charge of around 50% at the end of the simulation. The battery charge in the reward function $\mathcal{R}_3$ is similar to that in the RA scenario, and there is no significant improvement.

Based on the system throughput in Figure 4 and the battery residual energy in Figure 5, it can be concluded that the reward function $\mathcal{R}_1$ has low battery residual energy and high throughput, in contrast, reward function $\mathcal{R}_2$ has high battery residual energy and low throughput. However, neither of these reward functions aligns with our goal of managing both battery residual energy and throughput to be neither too high nor too low. According to the simulation results, the node exhibits more acceptable behavior with the reward function $\mathcal{R}_3$.

Figure 6 represents the sum of transmitted bits under $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ reward functions compared to RA selection. The figure shows that the sum of transferred bits in reward function $\mathcal{R}_1$ is high, while it is low in reward function $\mathcal{R}_2$. For example, at the 4500th minute, the total count of transferred bits in reward function $\mathcal{R}_1$ is 18 times higher than the total count of transmitted bits in reward function $\mathcal{R}_2$.

In contrast, $\mathcal{R}_3$ achieves a more balanced total count of transferred bits compared to the RA scenario. For example, it is observed that about 134 * 10^2 Kb more bits are transmitted by $\mathcal{R}_3$ in 4500 minutes. This result further demonstrates the superiority of Q-learning with the $\mathcal{R}_3$.

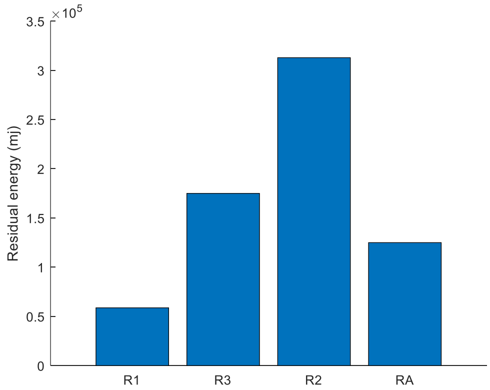


Figure 5. Battery residual energy when the rewards are $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$ compared to RA selection

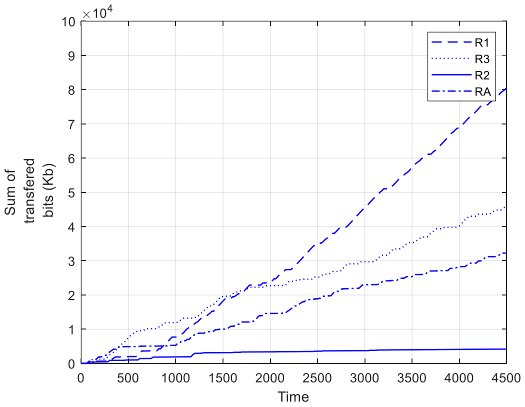


Figure 6. Transferred bits when the rewards are $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, and RA selection in terms of time

4. 2. The Comparison between Q-learning and SARSA

We compared the SARSA method introduced by Ge et al. (23) with our method. Algorithm 2 shows the steps of the SARSA method. Figures 7 and 8 respectively show the system throughput and battery residual energy between $\mathcal{Q}$-learning and SARSA algorithms in $\mathcal{R}_3$ with 1 iteration. Figure 7 compares the SARSA algorithm, an on-policy approach, with $\mathcal{Q}$-learning, an off-policy approach. SARSA considers the current policy during learning, leading to a cautious exploration of actions and potentially lower output. On the other hand, $\mathcal{Q}$-learning explores a wider range of actions, potentially discovering more optimal policies and resulting in higher throughput. In Figure 8, SARSA's on-policy nature allows it to adapt to the current policy and make decisions based on the current state-action pairs. This adaptability can lead to more energy-efficient decisions by avoiding actions that consume excessive energy, resulting in lower energy consumption.

In Figures 9 and 10, we can see that SARSA results with 60 iterations are better in both. As the iterations progress, both algorithms exhibit a gradual improvement in reward accumulation. However, SARSA consistently outperforms $\mathcal{Q}$-learning, showcasing its ability to make better policy decisions and optimize the reward over time. The graph demonstrates that SARSA achieves a higher cumulative reward compared to $\mathcal{Q}$-learning, indicating its effectiveness in maximizing the system's overall performance.

Figures 11 and 12, demonstrate reward $\mathcal{R}_3$ achieved by $\mathcal{Q}$-learning and SARSA during the time for 1 and 60

Algorithm 2: SARSA algorithm

```
Initialize $\mathcal{Q}$-table as $\mathcal{Q}(\mathcal{s}_t, \mathcal{a}_t) = 0$, with size ($\mathcal{s}$=3 ,$\mathcal{a}$=5)
The agent observes the initial state $\mathcal{s}_0$
for each time do
    Calculate the $E_r(t)$ in the battery every minute
    Choose the action $\mathcal{a}_0$ every 20 minutes from $\mathcal{Q}$-table with $\varepsilon$-greedy policy
    Get a reward $r$ every 10 minutes
    Update the related $\mathcal{Q}$-value every 10 minutes as follows
        $\mathcal{Q}(\mathcal{s}_t, \mathcal{a}_t) = \mathcal{Q}(\mathcal{s}_t, \mathcal{a}_t) + \alpha[r_{t+1}\gamma\mathcal{Q}(\mathcal{s}_{t+1}, \mathcal{a}_{t+1}) - \mathcal{Q}(\mathcal{s}_t, \mathcal{a}_t)]$
end for
```

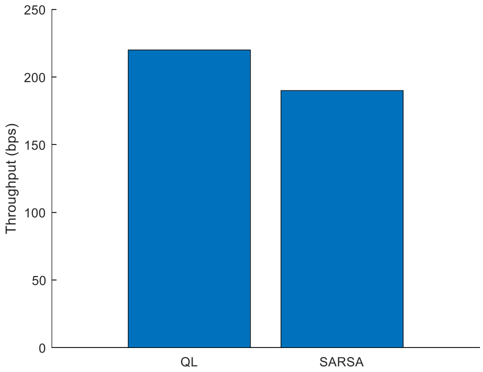


Figure 7. System throughput between $\mathcal{Q}$-learning and SARSA algorithms in $\mathcal{R}_3$ with 1 iteration

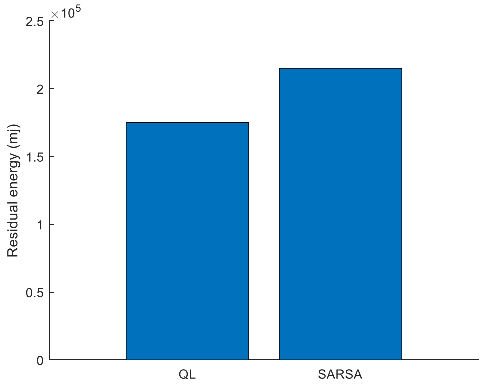


Figure 8. Battery residual energy between $\mathcal{Q}$-learning and SARSA algorithms in $\mathcal{R}_3$ with 1 iteration

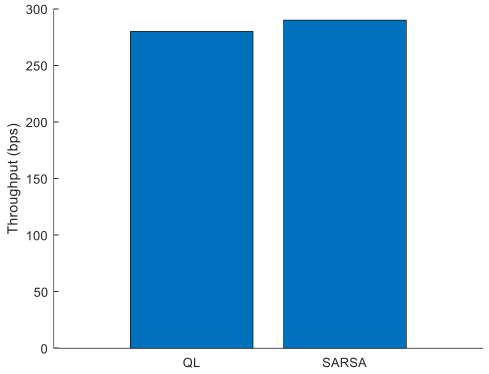


Figure 9. System throughput between $\mathcal{Q}$-learning and SARSA algorithms in $\mathcal{R}_3$ with 60 iterations

iterations, respectively. According to Figure 11, $\mathcal{Q}$-learning achieves higher rewards with fewer iterations and exhibits superior performance compared to SARSA. As depicted in Figure 12, clearly it indicates that SARSA surpasses $\mathcal{Q}$-learning, exhibiting an improvement rate of up to 6%. As the number of iterations increases, the

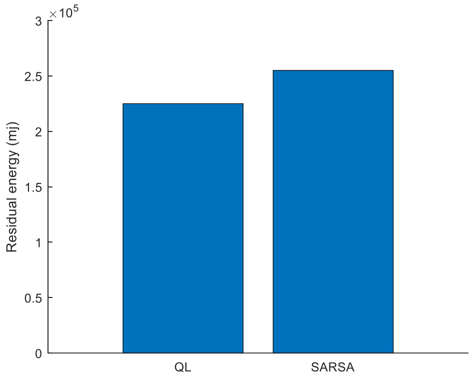


Figure 10. System throughput between $\mathcal{Q}$-learning and SARSA algorithms in $\mathcal{R}_3$ with 60 iterations

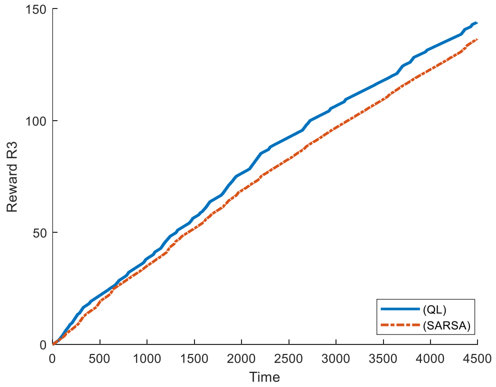


Figure 11. Performance of SARSA and $\mathcal{Q}$-Learning with 1 iteration

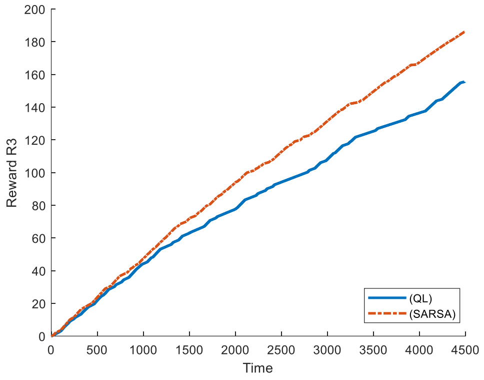


Figure 12. Performance of SARSA and $\mathcal{Q}$-Learning with 60 iterations

degree of improvement also escalates. However, for 60 iterations, SARSA consistently outperforms $\mathcal{Q}$-learning. It can be concluded that with an increase in training iterations, the disparity between SARSA and $\mathcal{Q}$-learning widens, leading to amplified reward gains.

5. CONCLUSIONS

In this paper, a one-tier WBAN including a sensor and a coordinator was presented. In this system model, the coordinator receives the data from the sensor located on the body and subsequently transmits RF energy to it. To meet the challenge of energy scarcity, the feasibility of harvesting energy from the human body was considered for the sensor. The goal of this work is to manage energy and throughput using $\mathcal{Q}$-learning. The results suggest that the reward function's design is a crucial aspect of the system's performance. This work tested three different reward functions in simulations to identify the best approach. The results indicate that a balancing parameter that adjusts the trade-off between throughput and energy consumption is the most effective solution. The $\mathcal{R}_1$ and $\mathcal{R}_2$ reward functions did not permit effective management of throughput and energy consumption. However, the reward function $\mathcal{R}_3$ enabled the node to adjust its behavior more effectively and perform better than the previous two rewards. In addition, $\mathcal{R}_3$ compared in the $\mathcal{Q}$-learning and SARSA algorithms with 1 iteration and 60 iterations, and the results were analyzed. The results indicated that SARSA can achieve better performance in exchange for higher iteration costs. Our proposal for the future is to investigate energy and throughput management in WBAN by utilizing multiple sensors, and this can be examined through the use of multi-agent RL algorithms. Furthermore, deep $\mathcal{Q}$-learning can be employed for scalability in a larger environment to handle large-scale Markov decision processes effectively.

6. REFERENCES

1. Siavashi A, Majidi M, editors. Sensing, wireless transmission, and smart processing of heart signals. 2021 5th International Conference on Internet of Things (IoT) and Applications 2021: IEEE. 10.1109/IoT52625.2021.9469710
2. Sridher T, Sarma A, Naveen Kumar P. Performance evaluation of onboard wi-fi module antennas in terms of orientation and position for iot applications. International Journal of Engineering, Transactions A: Basics. 2022;35(10):1918-28. 10.5829/ije.2022.35.10a.11
3. Babu T, Jayalakshmi V. Conglomerate energy efficient elgamal encryption based data aggregation cryptosystems in wireless sensor network. International Journal of Engineering, Transactions B: Applications. 2022;35(2):417-24. 10.5829/ije.2022.35.02b.18
4. Zhang R, Li X, editors. Joint power control and time allocation for WBANs with RF energy harvesting. 2020 IEEE/CIC International Conference on Communications in China (ICCC); 2020: IEEE. 10.1109/ICCC49849.2020.9238815
5. Sagar AK, Banda L, Sahana S, Singh K, Singh BK. Optimizing quality of service for sensor enabled internet of healthcare systems. Neuroscience Informatics. 2021;1(3):1-16. 10.1016/j.neuri.2021.100010
6. Li S, Hu F, Mao Z, Liu H, Ling Z, editors. Joint power allocation for energy harvesting to maximize throughput in classified WBAN. 2019 IEEE Global Communications Conference (GLOBECOM); 2019: IEEE. 10.1109/GLOBECOM38437.2019.9013828
7. Khatami N, Majidi M. Resource allocation for full-duplex wireless information and power transfer in wireless body area network. Journal of Electrical and Computer Engineering Innovations (JECEI). 2021;10(2):1-11. 10.22061/jecei.2021.8112.485
8. Badihi B, Sheikh MU, Ruttik K, Jäntti R, editors. On performance evaluation of BLE5 in indoor environment: An experimental study. 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications; 2020: IEEE. 10.1109/PIMRC48278.2020.9217132
9. Bulić P, Kojek G, Biasizzo A. Data transmission efficiency in bluetooth low energy (BLE) versions. Sensors. 2019;19(17):3746. 10.3390/s19173746
10. He M, Hu F, Ling Z, Mao Z, Huang Z. A dynamic weights algorithm on information and energy transmission protocol based on WBAN. IEEE Transactions on Vehicular Technology. 2021;70(2):1528-37. 10.1109/TVT.2021.3053964
11. Hasan K, Biswas K, Ahmed K, Nafi NS, Islam MS. A comprehensive review of wireless body area network. Journal of Network and Computer Applications. 2019;143:178-98. 10.1016/j.jnca.2019.06.016
12. Liu H, Hu F, Qu S, Li Z, Li D. Multipoint wireless information and power transfer to maximize sum-throughput in WBAN with energy harvesting. IEEE Internet of Things Journal. 2019;6(4):7069-78. 10.1109/JIOT.2019.2914147
13. Clerckx B, Kim J, Choi KW, Kim DI. Foundations of wireless information and power transfer: theory, prototypes, and experiments. Proceedings of The IEEE. 2022;110(1):8-30.
14. Rabby MKM, Alam MS, Shawkat MSA. A priority based energy harvesting scheme for charging embedded sensor nodes in wireless body area networks. PloS One. 2019;14(4):1-22. 10.1371/journal.pone.0214716
15. Jameii S, Khanzadi K. A latency reduction method for cloud-fog gaming based on reinforcement learning. International Journal of Engineering, Transactions C: Aspects. 2022;35(9):1674-81. 10.5829/ije.2022.35.09c.01
16. Aoudia FA, Gautier M, Berder O. RLMan: An energy manager based on reinforcement learning for energy harvesting wireless sensor networks. IEEE Transactions on Green Communications and Networking. 2018;2(2):408-17. 10.1109/TGCN.2018.2801725
17. Gupta A, Chaurasiya VK, editors. Reinforcement learning based energy management in wireless body area network: A survey. IEEE Conference on Information and Communication Technology; 2019: IEEE. 10.1109/CICT48419.2019.9066260
18. Wang L, Xi S, Liu W, Zhou Q, editors. Duty cycle optimization for blood pressure sensors in wireless body area networks based on reinforcement learning. 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS); 2021: IEEE. 10.1109/ICPS49255.2021.9468195
19. Xu Y-H, Xie J-W, Zhang Y-G, Hua M, Zhou W. Reinforcement learning -based energy efficient resource allocation for energy harvesting-powered wireless body area network. Sensors. 2020;20(1):1-22.
20. Mohammadi R, Shirmohammadi Z. RLS2: An energy efficient reinforcement learning-based sleep scheduling for energy harvesting WBANs. Computer Networks. 2023;229:109781. 10.1016/j.comnet.2023.109781
21. Rioual Y, Moullec YL, Laurent J, Khan MI, Diguet J-P, editors. Design and comparison of reward functions in reinforcement learning for energy management of sensor nodes. Biennial Baltic Electronics Conference (BEC); 2021: arXiv. 10.48550/arXiv.2106.01114
22. Rioual Y, Le Moullec Y, Laurent J, Khan MI, Diguet J-P, editors. Reward function evaluation in a reinforcement learning approach for energy management. 2018 16th Biennial Baltic Electronics Conference (BEC); 2018: IEEE.
23. Ge Y, Nan Y, Guo X. Maximizing network throughput by cooperative reinforcement learning in clustered solar-powered wireless sensor networks. International Journal of Distributed Sensor Networks. 2021;17(4):1-19. 10.1177/15501477211007411
24. Roy M, Biswas D, Aslam N, Chowdhury C. Reinforcement learning based effective communication strategies for energy harvested WBAN. Ad Hoc Networks. 2022;132:102880. 10.1016/j.adhoc.2022.102880
25. Van Roy S, Quitin F, Liu L, Oestges C, Horlin F, Dricot J-M, De Doncker P. Dynamic channel modeling for multi-sensor body area networks. IEEE Transactions on Antennas and Propagation. 2012;61(4):2200-8. 10.1109/TAP.2012.2231917
26. Ling Z, Hu F, Li D, editors. Optimal resource allocation in point-to-point wireless body area network with backscatter communication. 2020 International Conference on Computing, Networking and Communications (ICNC); 2020: IEEE. 10.1109/ICNC47757.2020.9049666
27. Al-Turjman F, Baali I. Machine learning for wearable IoT-based applications: A survey. Transactions on Emerging Telecommunications Technologies. 2019:1-16. 10.1002/ett.3635
28. Hu C, Xu M. Adaptive exploration strategy with multi-attribute decision-making for reinforcement learning. IEEE Access. 2020;8:32353-64. 10.1109/ACCESS.2020.2973169
29. Demir S, Stappers B, Kok K, Paterakis NG. Statistical arbitrage trading on the intraday market using the asynchronous advantage actor-criti method. Applied Energy. 2022;314:118912. 10.1016/j.apenergy.2022.118912
30. Frikha MS, Gammar SM, Lahmadi A, Andrey L. Reinforcement and deep reinforcement learning for wireless internet of things: A survey. Computer Communications. 2021;178:98-113. 10.1016/j.comcom.2021.07.014
31. Kwon O, Jeong J, Kim HB, Kwon IH, Park SY, Kim JE, Choi Y. Electrocardiogram sampling frequency range acceptable for heart

rate variability analysis. Healthcare Informatics Research. 2018;24(3):198-206. 10.4258/hir.2018.24.3.198

32. Gorlatova M, Sarik J, Grebla G, Cong M, Kymissis I, Zussman G. Movers and shakers: Kinetic energy harvesting for the internet of things. IEEE Journal on Selected Areas in Communications. 2015;33(8):1624-39. 10.1109/JSAC.2015.2391690

COPYRIGHTS

©2024 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, as long as the original authors and source are cited. No permission is required from the authors or the publishers.

Persian Abstract

چکیده

در این مقاله، به چالش‌های مدیریت انرژی و گذردهی در یک شبکه بی‌سیم ناحیه بدن (WBAN) با تمرکز بر حسگر ضربان قلب می‌پردازیم. رویکرد ما از روش خواب و بیداری برای به حداقل رساندن مصرف انرژی حسگر استفاده می‌کند، در حالی که از امواج فرکانس رادیویی (RF) و فعالیت‌های انسانی (دویدن، پیاده‌روی و استراحت) به عنوان منابع برداشت انرژی (EH) برای تأمین انرژی باتری استفاده می‌کند. فناوری بلوتوث کم انرژی نسخه ۵ (BLE5) برای انتقال بی‌سیم اطلاعات و انرژی استفاده می‌شود. هدف ما ایجاد تعادل بین گذردهی و انرژی باقیمانده باتری است. مزایای استفاده از یادگیری-Q برای انتخاب عمل، در مقایسه با انتخاب تصادفی عمل (RA) از طریق شبیه‌سازی نشان داده شده‌است. نتایج نشان می‌دهد که تابع پاداش در یادگیری-Q، با ترکیب یک پارامتر متعادل‌کننده، به طور موثر بین گذردهی و انرژی باقی‌مانده باتری بهتر عمل می‌کند. علاوه بر این، روش یادگیری-Q ما، در مقایسه با RA گذردهی سیستم را تا ٤٣ درصد بهبود می‌بخشد. همچنین، ما عملکرد الگوریتم‌های یادگیری-Q و SARSA را با استفاده از همان تابع پاداش مقایسه می‌کنیم تا توانایی‌های مربوطه آن‌ها را در مدیریت گذردهی سیستم و انرژی باقی‌مانده باتری ارزیابی کنیم. این یافته‌ها پیامدهای مهمی برای توسعه WBANهای کارآمد در انرژی دارند که امکان عملیات طولانی‌مدت و انتقال داده قابل اعتماد را فراهم می‌کند.