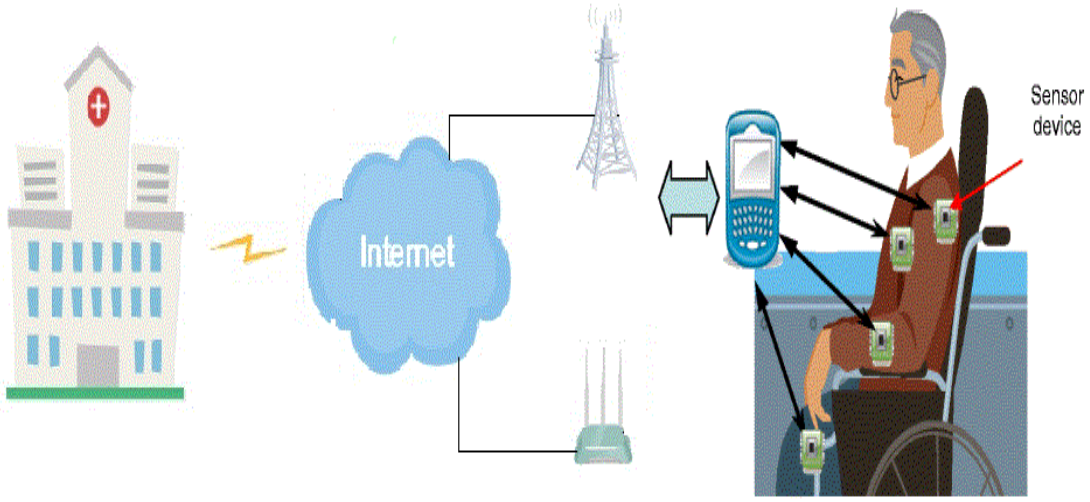

Article:
Xu, Y.-H., Yu, G. and Yong, Y.-T. (2021) Deep reinforcement learning-based resource scheduling strategy for reliability-oriented wireless body area networks. IEEE Sensors Letters, 5 (1). 7500104. ISSN 2475-1472

https://doi.org/10.1109/lsens.2020.3044337

Deep Reinforcement Learning-based Resource Scheduling Strategy for Reliability-oriented Wireless Body Area Networks

Journal:	*IEEE Sensors Letters*
Manuscript ID	SENSL-20-09-RL-0380
Manuscript Subject:	Regular Letter
Date Submitted by the Author:	16-Sep-2020
Complete List of Authors:	Xu, Yi-Han Yu, Gang; The University of Sheffield Yong, Yueh-Tiam; Universiti Teknologi MARA
Keywords:	NET



IEEE Sensors Letters Volume 2(3) (2017)
Sensor Networks

Deep Reinforcement Learning-based Resource Scheduling Strategy for Reliability-oriented Wireless Body Area Networks

Yi-Han Xu[1], Gang Yu[2], and Yueh-Tiam Yong[3]

[1] College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China
[2] Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S10 2TN, UK
[3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Samarahan Campus, Kota Samarahan 94300, Malaysia

Abstract—Reliability is a critical factor in designing of Wireless Body Area Networks (WBANs). In this letter, we propose a resource scheduling strategy and solving an optimization problem that maximize the reliability of the transmission of emergency-critical sensory data. We jointly consider transmission mode, relay selection, time slot allocation and transmit power of each body sensor and formulating the scheduling problem to be a Markov Decision Process (MDP). In this strategy, the scheduling decision is made by each body sensor that do not have complete and global network information. Owning to the formulated problem is non-convex and the high computation complexity, we propose a deep reinforcement learning algorithm to solve the problem. Numerical results reveal that the proposed strategy is capacity of guaranteeing the reliability of transmission with an acceptable convergence speed.

Index Terms—wireless body area networks, deep reinforcement learning, resource scheduling, reliable transmission

I. INTRODUCTION

The emergence of Wireless Body Area Networks (WBANs) is encouraging new innovative function to make the daily healthcare more efficient, thus paving the path to intelligent medical services in the forthcoming smart city [1]. Different with the traditional wired healthcare devices, WBANs consist of a number of heterogeneous invasive and/or non-invasive body sensors and one hub with the communication function in the form of wireless to continuously monitor the physiological signals of the human body and transmitting the real-time data to the doctors without any interruption [2]. The utilization of WBANs brings various benefits to daily life. However, it also faces one tremendous challenge in the practical deployment that is how to guarantee the reliability of the transmission as the data in a healthcare-oriented WBANs is emergence-critical in nature. To the best of our knowledge, the investigation on transmission reliability for WBANs is in its infancy to date, despite having some pioneering studies intended to study the network performance of WBANs in terms of throughput, packet loss and energy efficiency [3-7].

In [3], a fair and efficient radio resource allocation scheme is proposed to maximize the overall throughput of WBANs. The problem is formulated as a sum-utility maximization problem and a dual decomposition method is proposed to solve the optimal solution. However, the proposed method involves high computation complexity and the authors did not give the evaluation. In [4], a transmission power allocation scheme is proposed to improve the capacity and the outage probability of WBANs. Simulation results shown that the proposed scheme outperforms the water-filling and truncated-inversion approaches. Authors of [5] formulated the transmission power and time slot allocation optimization problem to be a Markov Decision Process to provide a high-quality service of WBANs. However, in this work, the model of channel fading is not discussed.

Compared to the existing works, we aim to maximize the end-to-end reliability of WBANs based on Deep Reinforcement Learning (DRL) algorithm. In particular, we formulate the resource scheduling problem to be a Markov Decision Process (MDP) by tactfully designing the state space, action space and reward function. Moreover, owning to the problem is non-convex, we propose a DRL algorithm to solve the maximization problem maximize and demonstrating how the transmission reliability can be guaranteed. Finally, we verify the proficiency and performance merits of the proposed resource scheduling strategy through numerical simulations.

II. Network Model

In this letter, we consider a single scenario of WBAN, in which a hub and multiple heterogeneity body sensors are deployed. We denote hub as H and N body sensors as $S_n, n \in (1, 2, \dots, N)$. In order to strength the utilization of network resource, both direct and cooperative transmission modes are supported by network layer as recommended by IEEE 802.15.6 standard [8]. For cooperative transmission mode, only two-hop transmission is allowed to avoid the stability issue and the redundant of signaling overhead. We use a binary indicator $\alpha_{S_n} \in \{0, 1\}\, n \in (1, 2, \dots, N)$ to denote which transmission mode is utilized by n-th body sensor. In MAC layer, Time Division Multiple Address (TDMA) is employed, in which each transmission frame includes K number of time slots and the time slot set is denoted as $\psi = (1, 2, \dots, K)$. We set $t_0 = 0$ and $t_K = T$. Thus, the duration of each time slot can be represented as $\tau_k = t_k - t_{k-1}\ \forall k \in \psi$.

In case of direct mode, we define a binary indicator $\beta_{S_n}^k \in \{0, 1\}, (n \in (1, 2, \dots, N), \forall k \in \psi)$ to denote which time slot is assigned to a specify body sensor. In this model, we reasonably assume that: 1) the hub can only receive data from one body sensor at each time slot; 2) in each time frame, each body sensor only be assigned at most one time shot for transmission. The purpose of these two assumptions is to maintain the fairness of transmission

Corresponding author: Yi-Han Xu (xuyihan@njfu.edu.cn)."

Article # Volume 2(3) (2017) IEEE Sensors Letters

opportunity of each body sensor. Therefore, two constraints can be derived as Equations 1 and 2:

$$\sum_{n=1}^{N} \beta_{S_n}^k \le 1, k \in \psi \qquad (1)$$

$$\sum_{k=1}^{K} \beta_{S_n}^k \le 1, n \in (1,2,\dots,N) \qquad (2)$$

In case of cooperative mode, we assume that the K number of time slots are allowed to allocate to both source-relay (*S-R*) and relay-hub (*R-H*) links. Similarly, we define an indicator $\delta_{S_n \to S_m}^k \in \{0,1\}, (n,m \in (1,2,\dots,N), \forall k \in \psi)$ to denote that the k-th time slot is allocated to n-th body sensor for transmitting data to the m-th body sensor, which is the relay of the n-th body sensor. Meanwhile, $\delta_{S_m \to H}^k \in \{0,1\} (m \in (1,2,\dots,N), \forall k \in \psi)$ to indicate the m-th body sensor forwards the data from n-th body sensor to the hub at the k-th time slot. Practically, we believe that each source sensor can select one relay sensor during any time slot and each relay sensor can only forward data from one source sensor at any time slot. Thus two constraints are obtained as Equations 3 and 4:

$$\sum_{m=1,m\ne n}^{N} \delta_{S_n \to S_m}^k \le 1, \quad \sum_{n=1,n\ne m}^{N} \delta_{S_n \to S_m}^k \le 1 \qquad (3)$$

$$\sum_{n=1,n\ne m}^{N} \delta_{S_m \to H}^k \le 1, \quad \sum_{m=1,m\ne n}^{N} \delta_{S_m \to H}^k \le 1 \qquad (4)$$

Moreover, we believe that each link can only be assigned at most one time slot and the transmission of *S-R* link should be prior to the transmission of *R-H* link. Thus, we can obtain another two constraint as Equations 5 and 6:

$$\sum_{k=1}^{K} \delta_{S_n \to S_m}^k \le 1, \sum_{k=1}^{K} \delta_{S_m \to H}^k \le 1 \ \ n \ne m \qquad (5)$$

$$\sum_{k=1}^{x} \delta_{S_n \to S_m}^k - \sum_{k=x+1}^{K} \delta_{S_m \to H}^k \ge 0, x \in (1,2,\dots,K-1) \qquad (6)$$

From the above analysis, the instantaneous Signal to Interference plus Noise Ratio (SINR) of direct transmission mode in the k-th time slot can be derived as Equation 7:

$$SINR_{n,k}^d = \frac{p_{n,k}^d \cdot g_{S_n \to H}}{\sum_{n_1=1,n_1\ne n}^{N} \sum_{m=1,m\ne n,n_1}^{N} \delta_{S_{n_1} \to S_m}^k \cdot p_{n_1,m,k}^s \cdot g_{S_{n_1} \to S_m} + n_0} \qquad (7)$$

Furthermore, the instantaneous SINR of cooperative transmission mode in the k-th time slot includes two parts: the SINR of *S-R* link and the *R-H* link, which are given as Equations 8 and 9:

$$SINR_{n,m,k}^{s\to r} = \frac{p_{n,m,k}^{s\to r} \cdot g_{S_n \to S_m}}{I_{n,m,k}^{s\to r} + n_0} \qquad (8)$$

$$SINR_{n,m,k}^{r\to H} = \frac{p_{n,m,k}^{r\to H} \cdot g_{S_m \to H}}{I_{n,m,k}^{r\to H} + n_0} \qquad (9)$$

Where:

$$I_{n,m,k}^{s\to r} = \sum_{\substack{n_1=1 \\ n_1\ne n,m}}^{N} \sum_{\substack{m_1=1 \\ m_1\ne n,n_1,m}}^{N} \delta_{S_{n_1} \to S_{m_1}}^k \cdot p_{n_1,m_1,k}^{s\to r} \cdot g_{S_{n_1} \to S_m} + \sum_{\substack{n_1=1 \\ n_1\ne n,m}}^{N} \beta_{S_{n_1}}^k \cdot p_{n_1,k}^d \cdot g_{S_{n_1} \to S_m}$$

$$+ \sum_{\substack{n_1=1 \\ n_1\ne n,m}}^{N} \sum_{\substack{m_1=1 \\ m_1\ne n,n_1,m}}^{N} \delta_{S_{m_1} \to H}^k \cdot p_{n_1,m_1,k}^{r\to H} \cdot g_{S_{m_1} \to S_m}$$

and

$$I_{n,m,k}^{r\to H} = \sum_{\substack{n_1=1 \\ n_1\ne m}}^{N} \sum_{\substack{m_1=1 \\ m_1\ne m,n_1}}^{N} \delta_{S_{n_1} \to S_{m_1}}^k \cdot p_{n_1,m_1,k}^{s\to r} \cdot g_{S_{n_1} \to S_m}$$

According to Shannon's theorem, the transmission rate of the direct mode ${R_n}^d$ can be obtained by Equation 11:

$${R_n}^d = \sum_{k=1}^{K} \beta_{S_n}^k \cdot B \cdot log_2(1 + SINR_{n,k}^d) \qquad (11)$$

Whereas, the transmission rate of the cooperative mode ${R_n}^c$ includes both the transmission rate of *S-R* link ${R_n}^{c,s\to r}$ and *R-H* link ${R_n}^{c,r\to H}$, which are can be given by Equations 12 and 13:

$${R_n}^{c,s\to r} = \sum_{\substack{m=1 \\ m\ne n}}^{N} \sum_{k=1}^{K} \delta_{S_n \to S_m}^k \cdot B \cdot log_2(1 + SINR_{n,m,k}^{s\to r}) \qquad (12)$$

$${R_n}^{c,r\to H} = \sum_{\substack{m=1 \\ m\ne n}}^{N} \sum_{k=1}^{K} \delta_{S_m \to H}^k \cdot B \cdot log_2(1 + SINR_{n,m,k}^{r\to H}) \qquad (13)$$

However, it should be noted that the overall transmission rate of the cooperative mode ${R_n}^c$ is limited by the smaller rate of *S-R* link and *R-H* link. Thus, the ${R_n}^c = \min\ ({R_n}^{c,s\to r},\ {R_n}^{c,r\to H})$.

Hence, the transmission rate of the n-th body sensor can be expressed as Equation 14:

$$R_n = \alpha_{S_n} \cdot {R_n}^d + \left(1 - \alpha_{S_n}\right) \cdot {R_n}^c, n \in (1,2,\dots,N) \qquad (14)$$

As we mentioned earlier, different with other networks, WBAN concentrates mainly on the reliable transmission of the emergency-critical information. Hence, the transmission rate may not significant important. We define a novel metric: delivery probability, to indicate the reliability level of transmission link. The delivery probability is the probability of successfully deliver the payload of sensory data with the size of B bits within an acceptable time T_{cct}. The delivery probability can be expressed as Equation 15:

$$Prb\left\{\sum_{k=1}^{K}\sum_{n=1}^{N} R_n \ge \frac{B}{T_{cct}}\right\}, \forall n \in N \qquad (15)$$

Where, T_{cct} is the channel coherence time and B is definitely depends on the advancement of the monitoring/detecting services supported by the body sensor.

To this end, the resource scheduling strategy can be formulated as:

$$\underset{\alpha_{S_n}, \beta_{S_n}^k, \delta_{S_n}^k, p_{n,k},}{maximize} \quad Prb \qquad (16)$$

s.t.

$$\sum_{n=1}^{N} \beta_{S_n}^k \le 1, k \in \psi, \quad \sum_{k=1}^{K} \beta_{S_n}^k \le 1, n \in (1,2,\dots,N)$$

$$\sum_{m=1,m\ne n}^{N} \delta_{S_n \to S_m}^k \le 1, \quad \sum_{n=1,n\ne m}^{N} \delta_{S_n \to S_m}^k \le 1$$

$$\sum_{n=1,n\ne m}^{N} \delta_{S_m \to H}^k \le 1, \quad \sum_{m=1,m\ne n}^{N} \delta_{S_m \to H}^k \le 1$$

$$\sum_{k=1}^{K} \delta_{S_n \to S_m}^k \le 1, \sum_{k=1}^{K} \delta_{S_m \to H}^k \le 1 \ \ n \ne m$$

$$\sum_{k=1}^{x} \delta_{S_n \to S_m}^k - \sum_{k=x+1}^{K} \delta_{S_m \to H}^k \ge 0, \forall x \in (1,2,\dots,K-1)$$

IEEE Sensors Letters Volume 2(3) (2017) Article #

$$p_{n,k}^{d} \le p_n^{max}\ \forall n \in (1,2,\dots,N), \forall k \in \psi$$
$$p_{n,m,k}^{s\to r} \le p_n^{max}\ n,m \in (1,2,\dots,N), n \ne m, \forall k \in \psi$$
$$p_{n,m,k}^{r\to H} \le p_n^{max} n,m \in (1,2,\dots,N), n \ne m, \forall k \in \psi$$

From the Equation 16, we can found that the scheduling problem is a mixed integer nonlinear programming problem, which cannot be directly solved by convex optimization methods. Therefore, we intend to formulate the problem to be a DFMDP by tactfully designing the state space, action space and the reward function. In this DFMDP, each body sensor acts as agent and exploring the unknown communication environment to obtain experiences, and then iteratively learned to get its optimal policy. Therefore, the state space, action space and the reward function can be designed as follows:

1) The state of each individual body sensor could include the global channel information and its own observation. Therefore, the state of each body sensor contains its own channel power gain and the interfering channel from other links, for all $n \in N$.
2) The action in this scenario should be the resource scheduling variables which including transmission mode α_{S_n}, time slot allocation $\beta_{S_n}^k$, relay selection $\delta_{S_n}^k$ and power control $p_{n,k}$. In summary, the resource scheduling strategy can be expressed as a set of $\{\alpha_{S_n}, \beta_{S_n}^k, \delta_{S_n}^k, p_{n,k}\}$.
3) The reward function in this DFMDP is the average delivery probability of links, which is expressed as Equation 16.

To solve the problem, despite the classical Q-learning algorithm can be a candidate tool, but herein, we intend to exploit the deep reinforcement learning algorithm to find the optimal scheduling strategy. The reason for this intention is because the deep reinforcement learning algorithm employs the Deep Q-Network (DQN) instead of the Q-table in Q-learning algorithm to train and improve the learning process [9]. Therefore, the approximate value of $Q(s^k, a^k)$ in classical Q-learning can be rewritten as $Q(s^k, a^k, \omega)$, where ω is the weight of Deep Neural Network (DNN). After the approximation, the optimal policy $\pi^*(s)$ can be obtained by Equation 17:

$$\pi^*(s) = \arg\max_{a^k} Q^*(s^k, a^{k+1}, \omega) \qquad (17)$$

Where, $Q^*(s,a)$ is the optimal Q-value via DNN approximation. DQN will choose the approximated action $a^{k+1} = \pi^*(s^{k+1})$. Then the approximated $\tilde{Q}(s^k, a^k)$ can be given as Equation 18:

$$\tilde{Q}(s^k, a^k, \omega) = r(s^k, a^k, \omega) + \gamma max_{a^{k+1}}[Q(s^{k+1}, a^{k+1}, \omega)] \qquad (18)$$

The value of ω is updated by minimizing the loss as expressed in Equation 19:

$$\text{Loss} = E\left[\left(\tilde{Q}(s^k, a^k, \omega) - Q(s^{k+1}, a^{k+1}, \omega)\right)^2\right] \qquad (19)$$

The pseudo code of the proposed deep reinforcement learning resource scheduling strategy is given in Algorithm 1.

Algorithm 1. The deep reinforcement learning resource scheduling strategy

1. initialize replay memory D to the number of body sensors N
2. initialize the Q-network Q with random weights ω
3. **for** *episode = 1 to M* **do**
4. Initialize the WBAN scenario, receive initial observation state s_1
5. **for** *k = 1 to K* **do**
   select a random action a^k (α_{S_n}, $\beta_{S_n}^k$, $\delta_{S_n}^k$, $p_{n,k}$) with the probability ε
6. Otherwise select $a^k = argmaxQ^*(s^k, a^k, \omega)$
7. perform action a^k and observe immediate reward r^k (Prb^k) and
   next state s^{k+1} (g_{S_n})
8. store transition (s^k, a^k, r^k, s^{k+1}) in D
9. select randomly samples $c(s_i, a_i, r_i, s_{i+1})$ from D
10. the weights of the DNN are updated by using stochastic gradient descent with respect to the ω to minimize the loss as Equation 19
11. update the policy $\pi(s^k) = \arg\max_{a^{k+1}} Q^*(s^k, a^{k+1}, \omega)$ after every
    a fixed number of steps
12. **end for**
13. **end for**

III. Simulation Results And Analysis

We evaluate the performance of the proposed deep reinforcement learning resource scheduling strategy in this section. The WBAN scenario considered includes a hub and multiple heterogeneous body sensors are deployed in different positions for various detection purposes. The hub is located at the center of the topology with the communication range of 10m. Each body sensor is randomly placed with the distance range from 2 to 5 m. We set 200 time instants for one episode and the delivery probability is averaged to reduce the instability. The DNN contains two fully connected hidden layers, in which 64 neurons and 32 neurons are set respectively. For each setting, we generate 100 independent runs and average the performance.

In WBAN scenario, the reliable transmission of emergence-critical information is vital. Therefore, Fig. 1 compares the optimization process for average delivery probability achieved by deep reinforcement learning algorithm with classical Q-learning algorithm. From the result, we can observed that the deep reinforcement learning algorithm tends to stable after 50 episodes rather than 80 episodes for the Q-learning algorithm, which indicates that the deep reinforcement learning algorithm has the higher convergence speed than the Q-learning algorithm. Another important finding is that the deep reinforcement learning algorithm exceeds the Q-learning algorithm approximate 18% after 80 episodes.

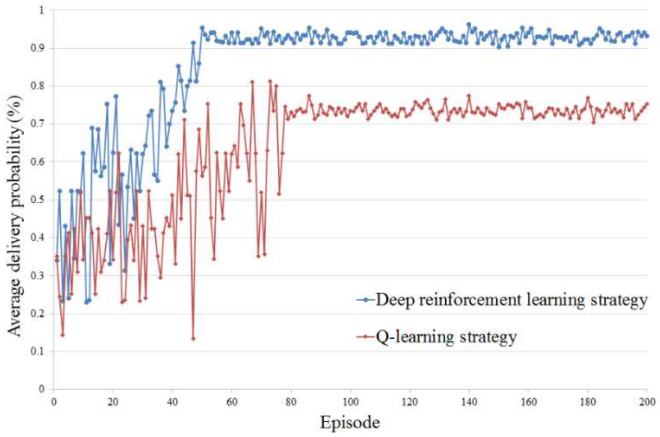


Fig. 1. The optimization process for delivery probability of body sensors

Fig. 2 depicts the average delivery probability against the number of deployed body sensors while the payload size B of each body sensor is randomly distributed between 0.4 Mbit and 1.6 Mbit. It can be observed that as the body sensors increase, the delivery probability decreases for all strategies. This is because the increase of the deployed body sensor will cause more mutual interference among

Article # Volume 2(3) (2017) IEEE Sensors Letters

each other. However, we still can find that the deep reinforcement learning strategy achieves the best delivery probability and the random scheduling strategy has the worst value. This is due to the fact that the random scheduling strategy schedules network resource randomly which generates catastrophic mutual interference among body sensors.

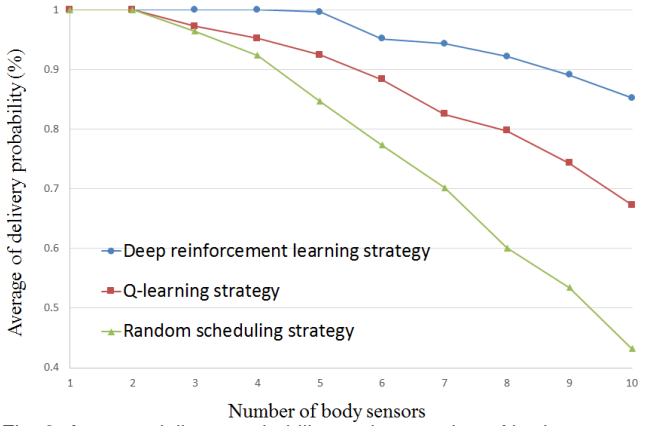


Fig. 2. Average delivery probability against number of body sensors

Fig. 3 presents the average delivery probability against the different payload size B of each body sensor while the number of body sensor is constantly set to 6. It is clear that the deep reinforcement learning strategy achieves the highest desirable delivery probability throughout all the cases. This is because the deep reinforcement learning strategy always enables to find the optimal scheduling strategy to guarantee the delivery probability. Remarkably, even in the worst case that B is set to the maximum size of 1.6Mb, the proposed deep reinforcement learning strategy still achieves 91.4% of average delivery probability.

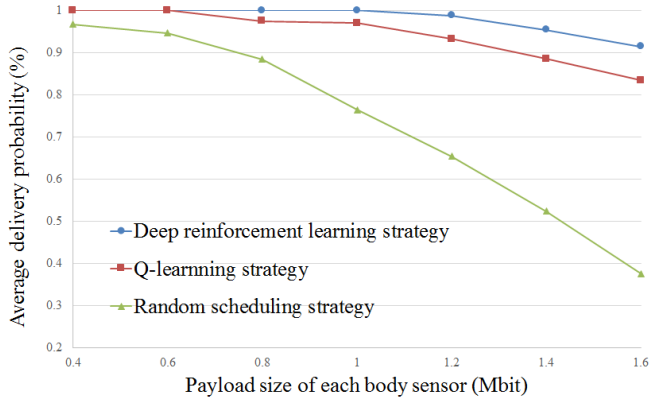


Fig. 3. Average delivery probability against different payload size B

IV. CONCLUSION

In this letter, we introduce a deep reinforcement learning based resource scheduling strategy for WBANs. We first jointly consider the transmission mode, relay selection, time slot allocation and transmit power of each body sensor, and formulating the resource scheduling strategy to be a DFMDP by designing the state space, action space and the reward function. After that, we propose a deep reinforcement learning algorithm to find the optimal strategy of maximizing the average delivery probability of each body sensor to guarantee the reliability of the transmission of emergency-critical sensory data. Finally, simulation results shown the effectiveness of the proposed strategy.

REFERENCES

[1] S. Marwa, A.D. Ahmed, and R. Imed, "Wireless Body Area Network (WBAN): A survey on reliability, fault tolerance, and technologies coexistence," *ACM Comput. Surv,* vol. 50, 3, 2017.

[2] C. Dagdeviren, Z. Li, Z. L. Wang, "Energy harvesting from the animal/human body for self-powered electronics," *Annu. Rev. Biomed. Eng.,* vol.19, pp. 85–108, 2017.

[3] S. Shen, J. Qian, D. Cheng, K. Yang, and G. Zhang, "A sum-utility maximization approach for fairness resource allocation in wireless powered body area networks," *IEEE Access.,* vol. 7, pp. 20014-20022, 2019.

[4] A. Razavi, and M. Jahed, "Capacity-outage joint analysis and optimal power allocation for wireless body area networks", *IEEE Systems Journal.,* vol. 13, no. 1, pp. 635-646, 2019.

[5] B. Liu, S. Yu and C. W. Chen, "Optimal resource allocation in energy harvesting-powered body sensor networks," in *Future Information and Communication Technologies for Ubiquitous HealthCare*, 2nd International Symposium on, pp. 1-5, IEEE, 2015.

[6] F. Y. Hu, X. L. Liu, D. Sui, M. Q. Shao and L. H. Wang, "Performance analysis of reliability in wireless body area networks," *IET Communications.,* vol. 11, no. 6, pp. 925-929, 2017.

[7] O. Amjad, E. Bedeer, S. Ikki, "Energy efficiency maximization of self-sustained wireless body sensor network," *IEEE Sensors Letters*., vol. 3, no. 12, 2019.

[8] IEEE standard for local and metropolitan area networks - part 15.6: wireless body area networks, 2012.

[9] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, & Y. C. Liang, "Applications of deep reinforcement learning in communications and networking: a survey". *IEEE Commun Surv & Tut.* vol. 21, no. 4, pp. 3133–3174, May. 2019.