



Deep Reinforcement Learning-Based Resource Scheduling Strategy for Reliability-Oriented Wireless Body Area Networks

Yi-Han Xu¹ , Gang Yu², and Yueh-Tiam Yong³ 

¹College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China

²Department of Electronic and Electrical Engineering, The University of Sheffield, S10 2TN Sheffield, U.K.

³Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Samarahan Campus, Kota Samarahan 94300, Malaysia

Manuscript received November 24, 2020; accepted December 9, 2020. Date of publication December 14, 2020; date of current version January 5, 2021.

Abstract—Reliability is a critical factor in designing of wireless body area networks. In this letter, we propose a resource scheduling strategy and solving an optimization problem to 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. In this strategy, the scheduling decision is made by each body sensor that do not have complete and global network information. Owing to the formulated problem is nonconvex 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—Sensor networks, deep reinforcement learning, reliable transmission, resource scheduling, wireless body area networks (WBANs).

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 noninvasive 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 addition, a novel enhanced deep reinforcement learning approach is proposed in [8] to allocate resource in ultrareliable low latency communication (URLLC-6G) to minimize transmission power under the constraints of reliability, latency, and transmission rate. In this approach, generative adversarial networks is introduced to gain experience in virtual environment and avoiding trial-and-errors, and thus, the proposed solution is more suitable for the URLLC-6G scenarios. In [9], a novel reinforcement learning-based resource allocation scheme to minimize the maximal computational and transmission delay in mobile edge computing networks is proposed. Simulation results illustrated that this algorithm outperforms standard Q-learning algorithm. However, different from the conventional wireless networks, the body

sensors in WBANs are normally deployed as implants, surface contacts, or wearables, the human body plays significant role in channel characteristics. The inhomogeneous structure of the human body and its tissues affect the propagation media resulting in fading models that differ from the commonly deployed in other wireless networks.

Compared to the existing works, in this letter, we aim to maximize the end-to-end reliability of WBANs from the perspective of resource scheduling. In particular, we formulate the resource scheduling problem to be a Markov decision process (MDP) by tactfully designing state space, action space, and reward function. Moreover, owing to the problem is nonconvex, a deep reinforcement learning (DRL) algorithm is proposed to solve the maximization problem and demonstrating how the transmission reliability can be guaranteed. Compared to the conventional DRL algorithm, the proposed algorithm dynamically predicts the transmission mode, relay selection, time slot allocation, and transmit power of each body sensor and then jointly scheduling the resource under specific constraints. 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 [10]. For cooperative transmission mode, only two-hop transmission is allowed to enhance the network throughput and meanwhile avoiding 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 is employed, in which each transmission frame includes K number of time slots and the time slot set is denoted as $\psi = (1, 2, \dots, K)$.

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

Associate Editor: F. Falcone.

Digital Object Identifier 10.1109/LENS.2020.3044337

Moreover, to utilize network resource more efficiently, we allow various body sensors to transmit data at the same time slot under a certain interference threshold.

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 sensor at each time slot and that 2) in each time frame, each sensor only be assigned at most one-time shot for transmission. The purpose of these two assumptions is to maintain the fairness of transmission opportunity of each sensor. Therefore, two constraints can be derived as

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

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

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 $\delta_{S_n \rightarrow S_m}^k \in \{0, 1\}$, ($n, m \in (1, 2, \dots, N), \forall k \in \psi$) 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, $\delta_{S_m \rightarrow H}^k \in \{0, 1\}$ ($m \in (1, 2, \dots, N), \forall k \in \psi$) indicates 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

$$\sum_{m=1, m \neq n}^N \delta_{S_n \rightarrow S_m}^k \leq 1, \sum_{n=1, n \neq m}^N \delta_{S_n \rightarrow S_m}^k \leq 1 \quad (3)$$

$$\sum_{n=1, n \neq m}^N \delta_{S_m \rightarrow H}^k \leq 1, \sum_{m=1, m \neq n}^N \delta_{S_m \rightarrow H}^k \leq 1. \quad (4)$$

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

$$\sum_{k=1}^K \delta_{S_n \rightarrow S_m}^k \leq 1, \sum_{k=1}^K \delta_{S_m \rightarrow H}^k \leq 1, n \neq m \quad (5)$$

$$\sum_{k=1}^x \delta_{S_n \rightarrow S_m}^k - \sum_{k=x+1}^K \delta_{S_m \rightarrow H}^k \geq 0, x \in (1, 2, \dots, K-1). \quad (6)$$

From the abovementioned analysis, the instantaneous signal to interference plus noise ratio (SINR) of direct transmission mode in the k th time slot can be derived as

$$\text{SINR}_{n,k}^d = \frac{P_{n,k}^d \cdot g_{S_n \rightarrow H}}{\sum_{n_1=1, n_1 \neq n}^N \sum_{m=1, m \neq n, n_1}^N \delta_{S_{n_1} \rightarrow S_m}^k \cdot P_{n_1, m, k}^s \cdot g_{S_{n_1} \rightarrow S_m} + n_0}. \quad (7)$$

Remarkably, the channel model used in this letter includes both large-scale path loss and small-scale fading. Specifically, the large-scale path loss consists of path loss and body shadowing, and small-scale fading is modeled as Rayleigh fading with unit mean. Since all these components can be expressed by channel coefficient in a mathematic way. Therefore, $g = \sqrt{10^{-\frac{(PL(d_0) + \varphi_{n,m})}{10}}} d_{n,m}^{-\epsilon} \xi_{n,m}$, where $PL(d_0)$ is the path loss at the reference distance, $\varphi_{n,m}$ is the body

shadowing loss margin between sensors, ϵ is the path loss exponent, and $\xi_{n,m}$ is the small scale fading power gain.

Similarly, the instantaneous SINR of cooperative transmission mode includes two parts: the SINR of S - R link and R - H link, which are given, respectively, as

$$\text{SINR}_{n,m,k}^{s \rightarrow r} = \frac{P_{n,m,k}^{s \rightarrow r} \cdot g_{S_n \rightarrow S_m}}{I_{n,m,k}^{s \rightarrow r} + n_0} \quad (8)$$

$$\text{SINR}_{n,m,k}^{r \rightarrow H} = \frac{P_{n,m,k}^{r \rightarrow H} \cdot g_{S_m \rightarrow H}}{I_{n,m,k}^{r \rightarrow H} + n_0} \quad (9)$$

where

$$\begin{aligned} I_{n,m,k}^{s \rightarrow r} = & \sum_{n_1=1}^N \sum_{m_1=1}^N \delta_{S_{n_1} \rightarrow S_{m_1}}^k \cdot P_{n_1, m_1, k}^{s \rightarrow r} \cdot g_{S_{n_1} \rightarrow S_{m_1}} \\ & + \sum_{n_1=1}^N \beta_{S_{n_1}}^k \cdot P_{n_1, k}^d \cdot g_{S_{n_1} \rightarrow S_m} \\ & + \sum_{n_1=1}^N \sum_{m_1=1}^N \delta_{S_{m_1} \rightarrow H}^k \cdot P_{n_1, m_1, k}^{r \rightarrow H} \cdot g_{S_{m_1} \rightarrow S_m} \end{aligned}$$

and

$$I_{n,m,k}^{r \rightarrow H} = \sum_{n_1=1}^N \sum_{m_1=1}^N \delta_{S_{n_1} \rightarrow S_{m_1}}^k \cdot P_{n_1, m_1, k}^{s \rightarrow r} \cdot g_{S_{n_1} \rightarrow S_m}. \quad (10)$$

According to Shannon's theorem, the transmission rate of the direct mode R_n^d can be obtained by

$$R_n^d = \sum_{k=1}^K \beta_{S_n}^k \cdot B \cdot \log_2 (1 + \text{SINR}_{n,k}^d). \quad (11)$$

However, the transmission rate of the cooperative mode R_n^c includes both the transmission rate of S - R link $R_n^{c, s \rightarrow r}$ and R - H link $R_n^{c, r \rightarrow H}$, which can be given by

$$R_n^{c, s \rightarrow r} = \sum_{m=1}^N \sum_{k=1}^K \delta_{S_n \rightarrow S_m}^k \cdot B \cdot \log_2 (1 + \text{SINR}_{n,m,k}^{s \rightarrow r}) \quad (12)$$

$$R_n^{c, r \rightarrow H} = \sum_{m=1}^N \sum_{k=1}^K \delta_{S_m \rightarrow H}^k \cdot B \cdot \log_2 (1 + \text{SINR}_{n,m,k}^{r \rightarrow H}). \quad (13)$$

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 $R_n^c = \min(R_n^{c, s \rightarrow r}, R_n^{c, r \rightarrow H})$.

Hence, the transmission rate of n th sensor is given as

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

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

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

where T_{cct} is the channel coherence time, and B 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

$$\begin{aligned} & \text{maximize} \quad Prb \\ & \alpha_{S_n}, \beta_{S_n}^k, \delta_{S_n \rightarrow S_m}^k, p_{n,k}, \\ & \text{s.t.} \\ & \alpha_{S_n}, \beta_{S_n}^k, \delta_{S_n \rightarrow S_m}^k, \delta_{S_m \rightarrow H}^k \in \{0, 1\}, (n, m \in (1, 2, \dots, N), \forall k \in \psi) \\ & \sum_{n=1}^N \beta_{S_n}^k \leq 1, k \in \psi, \sum_{k=1}^K \beta_{S_n}^k \leq 1, n \in (1, 2, \dots, N) \\ & \sum_{m=1, m \neq n}^N \delta_{S_n \rightarrow S_m}^k \leq 1, \sum_{n=1, n \neq m}^N \delta_{S_m \rightarrow S_n}^k \leq 1 \\ & \sum_{n=1, n \neq m}^N \delta_{S_m \rightarrow H}^k \leq 1, \sum_{m=1, m \neq n}^N \delta_{S_m \rightarrow H}^k \leq 1 \\ & \sum_{k=1}^K \delta_{S_n \rightarrow S_m}^k \leq 1, \sum_{k=1}^K \delta_{S_m \rightarrow H}^k \leq 1, n \neq m \\ & \sum_{k=1}^x \delta_{S_n \rightarrow S_m}^k - \sum_{k=x+1}^K \delta_{S_m \rightarrow H}^k \geq 0 \quad \forall x \in (1, 2, \dots, K-1) \\ & p_{n,k}^d, p_{n,m,k}^{s \rightarrow r}, p_{n,m,k}^{r \rightarrow H} \leq p_{n,m,k}^{\max} \quad \forall n, m \in (1, 2, \dots, N), \forall k \in \psi. \quad (16) \end{aligned}$$

From (16), we can found that the problem is a mixed integer nonlinear programming problem, which cannot be directly solved by convex optimization methods. Therefore, we formulate the problem to be a MDP, in which each body sensor acts as agent and exploring the unknown communication environment to obtain experiences, and then iteratively learned to get its optimal policy. The state, action, and the reward function can be designed as follows.

- 1) The state of each individual body sensor should include the global channel information and its own observation. Therefore, the state of each body sensor S_n^k contains its own channel power gain and the interfering channel from other links, for all $n \in N$. $S_n^k = (g_{S_n}^k, I_{S_n}^k)$, where $g_{S_n}^k = [g_{S_n}^1, g_{S_n}^2, \dots, g_{S_n}^{K-1}, g_{S_n}^K]$ and $I_{S_n}^k = [I_{S_n}^1, I_{S_n}^2, \dots, I_{S_n}^{K-1}, I_{S_n}^K]$.
- 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 $p_{n,k}$. Therefore, action $a_{S_n}^k = [\alpha_{S_n}^k, \beta_{S_n}^k, \delta_{S_n}^k, p_{n,k}]$, where $\alpha_{S_n}^k = [\alpha_{S_n}^1, \alpha_{S_n}^2, \dots, \alpha_{S_n}^{K-1}, \alpha_{S_n}^K]$, $\beta_{S_n}^k = [\beta_{S_n}^1, \beta_{S_n}^2, \dots, \beta_{S_n}^{K-1}, \beta_{S_n}^K]$, $\delta_{S_n}^k = [\delta_{S_n}^1, \delta_{S_n}^2, \dots, \delta_{S_n}^{K-1}, \delta_{S_n}^K]$, and $p_{n,k}$ is divided into χ power levels, and thus, $p_{n,k} = [\frac{p_{n,k}}{\chi}, \frac{2p_{n,k}}{\chi}, \dots, p_{n,k}]$.
- 3) The reward function in this MDP is the average delivery probability of links, which is expressed as (16).

To solve the problem, despite Q-learning algorithm can be a candidate tool, but herein, we exploit the DRL algorithm to find the optimal policy. The reason for this intention is because the DRL algorithm employs the deep Q-network (DQN) instead of the Q-table in Q-learning to train and improve the learning process [11]. 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).

Algorithm 1: The proposed DRL 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**
6. select a random action $a^k (\alpha_{S_n}^k, \beta_{S_n}^k, \delta_{S_n}^k, p_{n,k})$ with the probability ε
7. Otherwise select $a^k = \text{argmax} Q^*(s^k, a^k, \omega)$
8. perform action a^k and observe immediate reward $r^k (Prb^k)$ and next state $s^{k+1} (g_{S_n})$
9. store transition (s^k, a^k, r^k, s^{k+1}) in D
10. select randomly samples $c(s_i, a_i, r_i, s_{i+1})$ from D
11. the weights of the DNN are updated by using stochastic gradient descent with respect to the ω to minimize the loss as Equation 19
12. update the policy $\pi(s^k) = \text{argmax}_{a^{k+1}} Q^*(s^k, a^{k+1}, \omega)$ after every a fixed number of steps
13. **end for**
14. **end for**

After the approximation, the optimal policy $\pi^*(s)$ can be obtained by

$$\pi^*(s) = \text{argmax}_{a^k} Q^*(s^k, a^{k+1}, \omega) \quad (17)$$

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

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

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

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

The pseudocode of our algorithm is given in Algorithm 1.

III. SIMULATION RESULTS AND ANALYSIS

We evaluate the performance of the proposed DRL 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 10 m. Each body sensor is randomly placed with the distance range from 2 to 5 m. Each sensor acts as an agent and independently run the proposed DRL to find the optimal policy to maximize the delivery probability. We set 200 time instants for one episode and the delivery probability is averaged to reduce the instability. The DNN deployed in DQN framework contains two fully connected hidden layers, in which 64 neurons and 32 neurons are set, respectively. We evaluate the performance on PC Intel Core (TM) i7-8700 CPU @ 3.2 GHz. The implementation of all algorithms is carried out by using Tensorflow 1.13.1 with Python 3.6.5. For initial values of weights and biases, we set small random values based on zero-mean Gaussian distribution with a standard deviation of 0.1. Additionally, the Adam optimizer is used for training.

In WBAN scenario, the reliable transmission is vital. Therefore, Fig. 1 compares the optimization process for average delivery probability achieved by DRL algorithm with Q-learning algorithm. From the

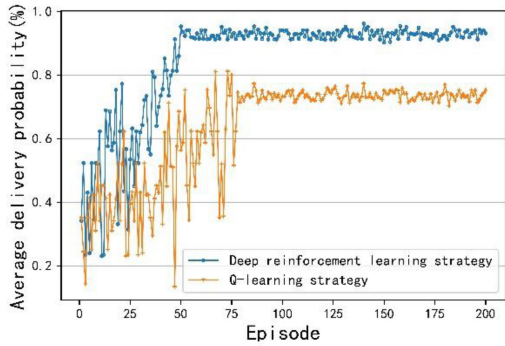


Fig. 1. Optimization process for delivery probability of body sensors.

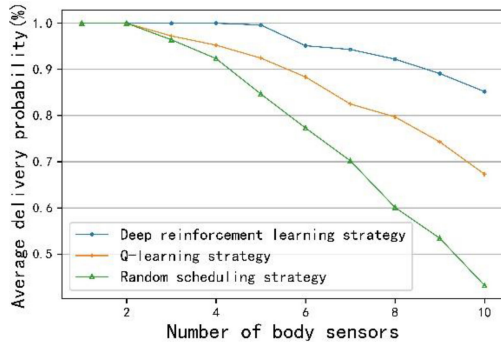


Fig. 2. Average delivery probability against number of body sensors.

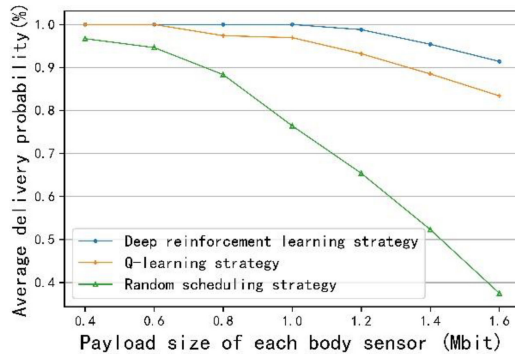


Fig. 3. Average delivery probability against different payload size B .

result, we can observe that the DRL algorithm tends to stable after 50 episodes rather than 80 episodes for the Q-learning algorithm, which indicates that DRL algorithm has the higher convergence speed than the Q-learning algorithm. Another important finding is that the performance from DRL algorithm outperforms the Q-learning algorithm approximately 18% after 80 episodes.

Fig. 2 depicts the average delivery probability against the number of deployed sensors while the payload size B of each sensor is randomly distributed between 0.4 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 increasing will cause more mutual interference. However, we still find that DRL strategy achieves the best performance.

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 proposed DRL strategy achieves the highest desirable delivery probability throughout all the cases. This is because DRL 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 1.6 Mb, the DRL strategy still achieves 91.4% of average delivery probability.

IV. CONCLUSION

In this letter, we introduce a DRL-based resource scheduling strategy for WBANs. We first jointly consider the transmission mode, relay selection, time slot allocation, and transmit power of each sensor, and formulating the resource scheduling problem to be a MDP. After that, we propose a DRL 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 demonstrated the effectiveness of the proposed strategy.

ACKNOWLEDGMENT

This work was supported by Introduction of high-level talents and overseas returnee's scientific fund in Nanjing Forestry University under Grant GXL015.

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, no. 1, pp. 1–38, Mar. 2017.
- [2] C. Dagdeviren, Z. Li, and Z. L. Wang, "Energy harvesting from the animal/human body for self-powered electronics," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, pp. 85–108, Jun. 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 Syst. J.*, vol. 13, no. 1, pp. 635–646, Mar. 2019.
- [5] B. Liu, S. Yu, and C. W. Chen, "Optimal resource allocation in energy harvesting-powered body sensor networks," in *Proc. 2nd Int. Symp. Future Inf. Commun. Technol. Ubiquitous Health Care*, 2015, pp. 1–5.
- [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 Commun.*, vol. 11, no. 6, pp. 925–929, Apr. 2017.
- [7] O. Amjad, E. Bedeer, and S. Ikki, "Energy efficiency maximization of self-sustained wireless body sensor network," *IEEE Sens. Lett.*, vol. 3, no. 12, Dec. 2019, Art. no. 7501204.
- [8] A. T. Z. Kargari, W. Saad, M. Mozaffari, and H. V. Poor, "Experienced deep reinforcement learning with generative adversarial networks (GANs) for model-free ultra reliable low latency communication," *IEEE Trans. Commun.*, to be published, doi: [10.1109/TCOMM.2020.3031930](https://doi.org/10.1109/TCOMM.2020.3031930).
- [9] S. Wang, M. Chen, X. Liu, C. Yin, S. Cui, and H. V. Poor, "A machine learning approach for task and resource allocation in mobile edge computing-based networks," *IEEE Int. Things J.*, to be published, doi: [10.1109/JIOT.2020.3011286](https://doi.org/10.1109/JIOT.2020.3011286).
- [10] IEEE Standard for local and metropolitan area networks - part 15.6: wireless body area networks, 2012, doi: [10.1109/IEEESTD.2012.6161600](https://doi.org/10.1109/IEEESTD.2012.6161600).
- [11] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, and 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, Fourth Quarter 2019.