Dynamic Spectrum Access for Internet-of-Things Based on Federated Deep Reinforcement Learning

Feng Li, *Member, IEEE,* Bowen Shen, Jiale Guo, Kwok-Yan Lam, *Senior Member, IEEE,* Guiyi Wei, *Member, IEEE,* and Li Wang

***Abstract*—The explosive growth of Internet-of-Things (IoT) applications such as smart cities and Industry 4.0 have led to drastic increase in demand for wireless bandwidth, hence motivating the rapid development of new techniques for enhancing spectrum utilization needed by new generation wireless communication technologies. Among others, dynamic spectrum access (DSA) is one of the most widely accepted approaches. In this paper, as an enhancement of existing works, we take into consideration of inter-node collaborations in a dynamic spectrum environment. Typically, in such distributed circumstances, intelligent dynamic spectrum access almost invariably relies on self-learning to achieve dynamic spectrum access improvement. Whereas, this paper proposes a DSA scheme based on deep reinforcement learning to enhance spectrum and access efficiency. Unlike traditional Q-learning-based DSA, we introduce the following to enhance the spectrum efficiency in dynamic IoT spectrum environments. First, deep double Q-learning is adopted to perform local self-spectrum-learning for IoT terminals in order to achieve better dynamic access accuracy. Second, to accelerate learning convergence, federated learning (FL) in edge nodes is used to improve the self-learning. Third, multiple secondary users, who do not interfere with each other and have similar operation condition, are clustered for federated learning to enhance the efficiency of deep reinforcement learning. Comparing with the traditional distributed DSA with deep learning, the proposed scheme has faster access convergence speed due to the characteristic of global optimization for federated learning. Based on this, a framework of federated deep reinforcement learning (FDRL) for DSA is proposed. Furthermore, this scheme preserves privacy of IoT users in that FDRL only requires model parameters to be uploaded to edge servers. Simulations are performed to show the effectiveness of the proposed FDRL-based DSA framework.**

***Index Terms*—Federated learning, deep reinforcement learning, Internet of Things (IoT), dynamic spectrum access.**

I. Introduction

THE global emphasis of earth sustainability has motivated large scale adoption of cyber-physical intelligent systems to optimze resource management and minimize urbanization footprint in the natural environment. In this connect, the world is experiencing an explosive growth in Internet-of-Things (IoT) applications such as smart cities, smart manufacturing and intelligent transportation systems, which in turn lead to drastic increase in demand for wireless bandwidth due to the connectivity needs of the massive number of IoT devices being deployed for supporting such intelligent systems. As such, reseacrhers and practitioners have been actively exploring new techniques for enhancing wireless spectrum utilization, which is wide accepted to be a key challenge in new generation wireless communication technologies. Besides, the wireless spectrum challenge is further exacerbated by the emerging era of Internet of Everything, which sees a gigantic number of wireless devices such as autonomous vehicles, medical devices, smart shopping devices being added to the IoT community [1]. Thus, the huge demand for bandwidth has become a serious bottleneck in urbanization development and the need for new generation wireless technologies is one of the most important issues to be addressed by the international community.

As one of the widely accepted approaches to address bandwidth optimization, dynamic spectrum access (DSA) is proposed to improve spectrum utilization of IoT networks. In the process, IoT users are allowed to transmit without causing interference to the primary users (PUs), or IoT users occupy the temporarily idle channel based on the spectrum information in the IoT until the channel is detected to be re-occupied by the licensed users. To improve the efficiency of DSA and sharing, many related techniques and algorithms have been proposed and can be classified into learning-based approaches and non-learning-based approaches. Reinforcement learning (RL) is widely used as the main technique of learning methods [2]-[4]. While most learning-based approaches require users to have high computing power, some DSA strategies based on auction, blockchain, signal feature continue to receive attention [5]-[7]. In [5], the proposal is based on multi-band auction mechanism. In [6], an efficient blockchain network is built by designing a Proof-of-Trust consensus mechanism. In addition, methods for classifying the feature of received signal samples for spectrum sensing are introduced [7].

Most of the users in the IoT are based on different operating platforms, with different computing power and different data. Thus, it is still a great challenge to take the advantage of learning-based approach to go for intelligent DSA despite the low computing power of the users. In recent years, federated learning (FL) as an emerging learning-based method has

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Also, this work was also supported by the "Fundamental Research Funds for the Central Universities" (3132021335).

F. Li, B. Shen and G. Wei are with School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, 310018, China. F. Li is also at School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore. (fengli2002@yeah.net, bwshen@outlook.com, weigy@mail.zjgsu.edu.cn)

J. Guo and K. Y. Lam are with School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore. (jiale001@e.ntu.edu.sg, kwokyan.lam@ntu.edu.sg)

L. Wang is with College of Marine Electrical Engineering, Dalian Maritime University, Dalian, 116026, China. (liwang2002@dlmu.edu.cn)

received rapid growth of attention in the distributed field [8][9]. The algorithm that the model parameters are further aggregated after users learned locally to form a global model effectively alleviate the computing pressure on the IoT user terminals [10]-[13].

In this paper, we propose a scheme for DSA based on federated deep reinforcement learning (FDRL). The scheme describes multi-channel DSA as a Markov decision process and implements intelligent DSA through a neural network model derived from Federated deep reinforcement learning. Typically, in order to achieve distributed DSA in IoT networks, users often need a large amount of accurate IoT environment information and spectrum state information. This leads to the communication overhead in the whole network is always high and the heavy computing burden for IoT users, which reduce the network efficiency. Meanwhile, the large amount of communication between different devices makes the security of terminal data and personal privacy data particularly important. The proposed scheme is based on federated learning, which achieves faster learning while protecting user data security by aggregating the neural network parameters of the users local model to form a global model and distributing it to the users for use.

When performing FDRL in DSA, two stages of deep learning will be conducted. In the process of local deep learning, we compare three kinds of self-learning methods including Q-learning, neutral networks and deep double Q-learning and choose the last one as our solution which has better dynamic access accuracy. In this process, FedAvg [14] is applied to reduce the number of communications in the IoT networks. Besides, considering the different modes of operation and data types among users, we divide the users into clusters for FL to improve the learning efficiency. Compared with traditional FL in edge nodes, cluster-based FL in distributed IoT can achieve better convergence speed. In addition, we design a reward function based on the interference between different IoT users to better match deep RL to DSA scenarios of IoT. The rest of this paper is organized as follows. Section II introduces the system model. In Section III, the detailed operation of the proposed framework of FDRL for DSA is presented. Numerical results are presented in Section IV to justify the performance of our proposed solution. At last, this paper is concluded in Section V.

II. System Model

In this paper, the DSA problem in the case of dense multi-cell users is considered and the system model is shown in Fig. 1. The scenario considered is that SUs in all cells participate in FDRL process. Taking the SUs who do not interfere with each other and have similar operation conditions as a cluster, the SUs in each cluster upload their own local model parameter to the same edge parameter server for neural network model parameter aggregation, so as to quickly form a global model of the cluster. Then, the parameter server distributes the weight parameters of the global model to all SUs of the corresponding cluster for the next round of learning and intelligent DSA. Fig. 2 shows the FL framework of each SU. The proposed deep

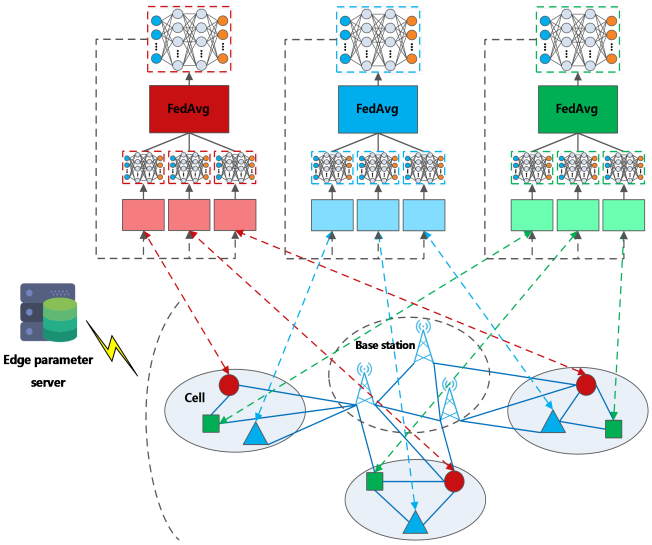

Fig. 1. System model of federated learning for dynamic spectrum access in IoT.

RL algorithm and the aggregation phase of FL are performed in each SU and edge parameter server respectively.

As shown in Fig. 1, we assume that there are g IoT cells and u SUs. In the case, we take both the large scale and small scale fading channels into consideration for the uplink transmission and assume that the average power of noise is τ^2 and the path loss parameter is δ. Thus, the Signal to Interference plus Noise Ratio (SINR) on the uplink transmission for SU u can be expressed as

$$SINR^u = \frac{p^u h^u d^{u^{-\delta}}}{\sum_{v \neq u} p^v h^v d^{v^{-\delta}} + \tau^2} \tag{1}$$

where p^u, h^u, d^u denote the transmit power, small scale fading for SU u and the distance between SU u and the base station respectively.

III. Proposed Framework for DSA

The whole framework can be divided into three phases, namely local double deep Q-learning network (DDQN) model training phase, central model aggregation phase, and model parameter release phase. The process and details of these three phases in proposed framework of FDRL of each training round are introduced in Section A, Section B and Section C respectively.

A. Local DDQN Training Phase

We use DDQN and ϵ-greedy policy as the basis for the deep reinforcement learning to solve the problem, which can avoid the defect of overestimation of standard deep Q-learning network algorithm. Each user device includes the following components.

- **State space**: The state space of user device u at training round i can be described as $s_t^u = [C^u, I_t^u]$, where C^u, I_t^u denote the channel selected for access and state of the channel in each step t.

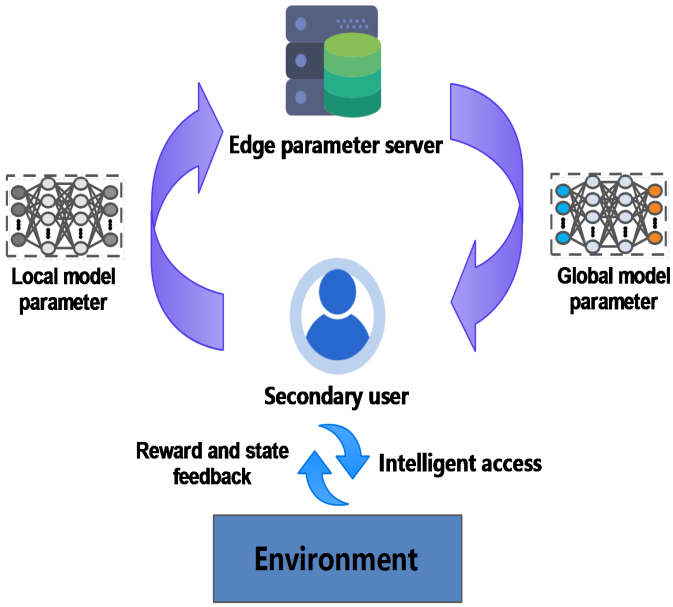

Fig. 2. Federated learning framework for secondary user.

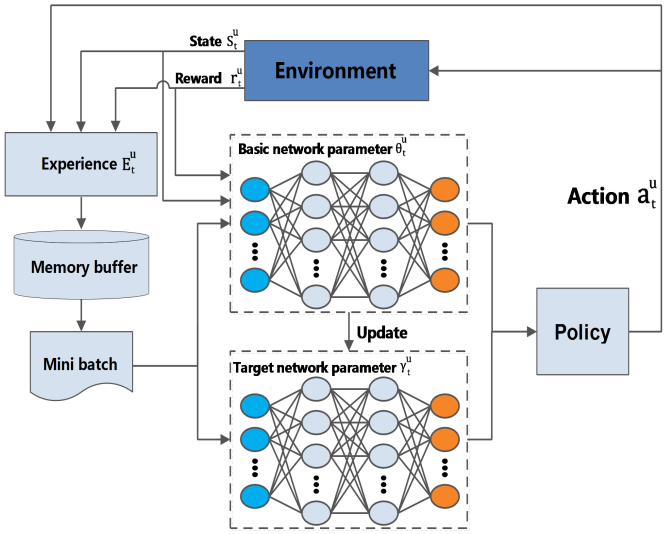

Fig. 3. Process of local DDQN learning.

- **Action space**: The user devices will select a channel to access according to the state s_i^u, which can be expressed as $a_t^u (a \in A)$. The action space of user devices is denoted by A, which contains all optional access channels.
- **System reward**: During the local training of user devices, the reward mechanism has a great impact on the selection of appropriate channels, which guides the system performance towards the desired objective, so it needs to be carefully designed. In this paper, our objective is to maximize channel utilization while improving as much as possible the channel quality of each user device after access. The reward function is based on SINR, which can be a good indicator of DSA performance. And to enhance the variability of different users accessing different channels, a quality of service (QoS) threshold is introduced. The reward function can be expressed as

$$r_t^u = \frac{p^u h^u d^{u^{-\delta}}}{\sum_{j \in n} p^j h^j d^{j^{-\delta}} + \tau^2 - \mu^u} \tag{2}$$

where n denotes the set of all the user device, μ^u denotes the QoS threshold required by the user device u.

When the users device is trained locally, past experience is stored in the device for the neural network to train the appropriate model parameters. In DDQN algorithm, each user device has two neural networks, which are basic network denoted by θ_t^u and target network denoted by γ_t^u. In each training step t, the basic network parameter θ_t^u of user device u is updated in real time while the target network parameter γ_t^u is updated much less frequently. After training step f, γ_t^u is updated by setting the value equal to θ_t^u to avoid overestimating Q value. And in order to find an optimal policy, the action-value function obeys the Bellman optimality

Algorithm 1 Local DDQN Training Process

Initialize the basic network parameter θ_t^u, target network parameter γ_t^u, target network updating frequency f, standard channel threshold μ^u learning rate α, and discount factor β.
for episode $i = 1$ **to** N **do**
  Select an action randomly and obtain an initial state s^u.
  for step $t = 1$ **to** T **do**
    Select an action a_t^u according to the ϵ-greedy policy.
    Execute action a_t^u to access the channel.
    Obtain the reward r_t^u and the new state s_{t+1}^u.
    Update the action-value function $Q(s_t^u, a_t^u)$.
    Store the experience $E_t = [s_t^u, a_t^u, s_{t+1}^u, r_t^u]$ to the memory buffer M^u.
    Draw randomly a mini-batch $\hat{M}^u$ from memory buffer M^u.
    Update the basic network parameter θ_t^u.
    if $t \mod f == 0$ **then**
      Update the target network parameter γ_t^u by setting γ_t^u equal to θ_t^u.
    end if
    if $r_t^u \geq \mu^u$ **or** the selected access channel is being used by primary users **then**
      End step.
    end if
  end for
end for

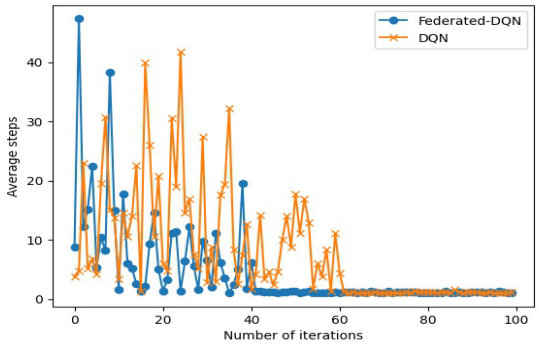

Fig. 4. Comparison of average steps for federated learning and non-federated learning.

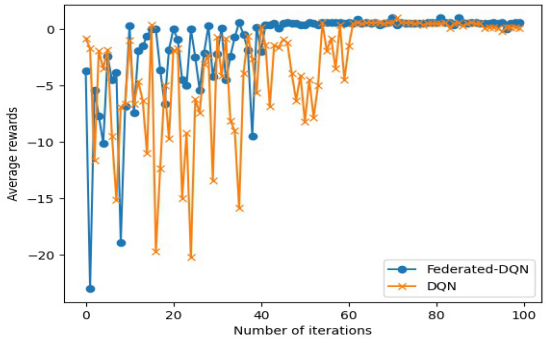

Fig. 5. Comparison of average rewards for federated learning and non-federated learning.

equation which can be expressed as

$$Q(s_t^u, a_t^u) = r_t^u + \beta Q(s_{t+1}^u, argmax_{a_t^u \in A} Q(s_{t+1}^u, a_t^u; \theta_t^u); \gamma_t^u) \tag{3}$$

where β denotes the discount factor. The larger the value of β, the more the system values the past experience during training. And the process of the Q-value updating is as follows:

$$\begin{aligned} Q_{t+1}(s_t^u, a_t^u) = (1-\alpha) Q_t(s_t^u, a_t^u) + \alpha(r_t^u + \beta Q(s_{t+1}^u, \\ argmax_{a_t^u \in A} Q(s_{t+1}^u, a_t^u; \theta_t^u); \gamma_t^u)) \end{aligned} \tag{4}$$

where $\alpha \in (0, 1]$ denotes the learning rate. $Q(s_t^u, a_t^u)$ is an accumulation of reward starting with s_t^u and a_t^u In the training process, the maximum function value of a_t^u is obtained by the basic network parameter θ_t^u, and then the function value based on the target network parameter γ_t^u is obtained after performing this action. A Q-learning table is gradually formed and completed during the training process. Following the ϵ-greedy policy in each training step, the user device randomly selects an action from the action space A with probability ϵ, and select the action with maximum Q-learning value with probability $1-\epsilon$, which can be formulate as

$$a_{t+1}^u = \begin{cases} a_{random}, & P = \epsilon \\ argmax_{a_{t+1}^u \in A} Q(s_{t+1}^u, a_{t+1}^u), & P = 1-\epsilon \end{cases} \tag{5}$$

The process of local DDQN training of each user device is presented in Algorithm 1 and Fig. 3 in detail.

B. Central Model Aggregation Phase

Due to the differences in data types, hardware performance and model performance of different user devices, aggregating the model parameters of all user devices in a simple way instead has detrimental effects on the system. In this paper, we consider a federated learning scenario in which model parameters of user devices in multiple cells that do not interfere with each other and have similar operation condition are aggregated to address the problem. With the advantages of lower communication costs and shorter training time, FedAvg is utilized to perform model parameter aggregation and form a global model, which can be expressed as

$$\gamma_{global}^o = \frac{\sum_{l \in C_o} \gamma^l}{|C_o|} \tag{6}$$

where C_o denotes the set of user device of federated cluster o.

C. Model Parameter Release Phase

After the edge parameter server completes the global model aggregation of the federated cluster o in each training round to form the model parameter γ_{global}^o, γ_{global}^o will be distributed to each user device u in cluster o for the next round of FL and intelligent DSA. And the next round of local deep RL model training will be based on the distributed global model parameter γ_{global}^o.

IV. Numerical Results

In this section, the performance of proposed scheme is presented based on simulations. We set that a cluster have 5 SUs. And we also set a total of 10 channels, of which 5 channels are occupied by the PU randomly, 3 channels are occupied by other SUs randomly, and 2 channels are idle channels for access. The parameters for FL method are set as: batch size as 30 in Fig. 4, Fig. 5 and Fig. 8, learning rate $\alpha = 1$, discount factor $\beta = 0.95$, updating frequency $f = 10$, episodes number as $[100, 100, 100, 100, 100]$, experience pool size as $[50, 100, 100, 50, 100]$. Besides, the transmit power for each SUs are as $[25, 25, 30, 30, 25]$ dBm, background noise as $[-40, -40, -95, -95, -40]$ dBm, QoS threshold as -95 dBm. We also simulate the difference in computing power of SUs in the IoT by making them participate in aggregation at different frequencies.

In Fig. 4 and Fig. 5, the federated and non-federated learning cases are compared. The non-federated learning method denotes the deep Q-learning network (DQN) as shown in Fig. 4. We can achieve that the average number of steps required and the rewards obtained for both the federated and non-federated approaches are less optimal at the beginning of the iterations, but the convergence starts to occur with an increasing number of iterations and the FL-based method converges faster. In this test, the model of each iteration is the result of multiple SUs' local models aggregated after local training. Then, based on FedAvg algorithm and cluster classification, the proposed scheme requires less communication and has faster convergence in model aggregation. In

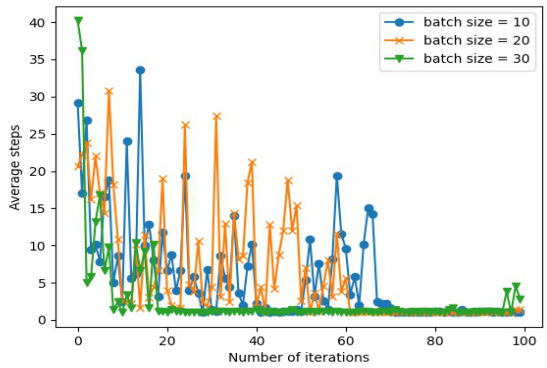

Fig. 6. Average steps in different batch size.

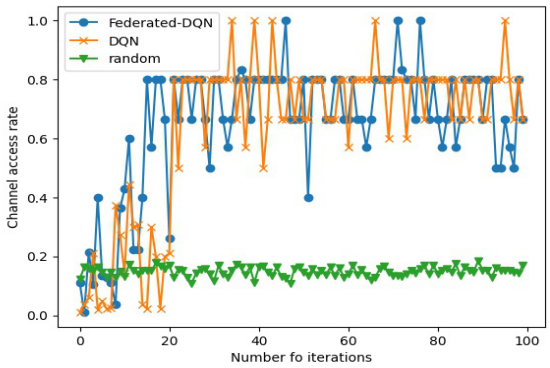

Fig. 8. Channel access rate.

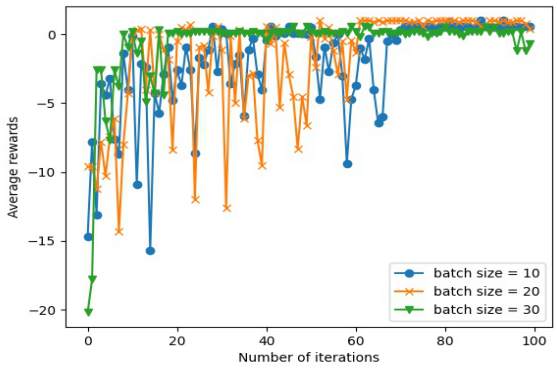

Fig. 7. Average rewards in different batch size.

contrast, the models based on the DQN approach lack the cooperation between various nodes, which are the SUs' own training models. Moreover, most importantly, since the FL-based approach only needs to upload SUs' own locally trained model parameters during model aggregation, it can better protects the user's data privacy.

In Fig. 6 and Fig. 7, we compare the effect of the batch size on the convergence speed of the model. A larger batch size means that more data are involved in one training. Thus, we can get that the larger the batch size is, the faster the convergence speed is in the proposed algorithm.

Whats more, the performance of the success rate of suitable channel access with increasing iterations number for the three cases is given in Fig. 8. The success rate of the reinforcement learning based algorithm increases rapidly after several iterations compared to the random channel access, and the scheme applying FL converges faster. However, due to the influence of ϵ-greedy strategy, even if the current channel environment state does not change, SUs still have a very small probability to randomly select a channel for access, so they can not always maintain a 100% successful access rate.

V. Conclusion

In this paper, we proposed a multi-channel DSA strategy based on federal deep reinforcement learning in IoT networks. Specifically, we formulated the multi-channel dynamic spectrum as a Markovian decision process based on its characteristics, and proposed deep reinforcement learning algorithms to achieve intelligent access to multi-channel dynamic spectrum. Based on this, we apply the federated learning algorithm to design a cluster of SUs with similar operating conditions that do not interfere with each other in multiple small areas according to their data and their own characteristics, and each SUin the cluster uploads its own training model to the same edge parameter server for neural network model parameter aggregation to quickly form a global model of the cluster. Thereafter the parameter server distributes the weight parameters of the global model to all SUs in the corresponding cluster for the next round of learning, in order to achieve efficient, fast and secure intelligent access to reasonable spectrum resources by SUs. We also performed comparative simulations of different cases, numerical results demonstrated the efficiency of the DSA strategy proposed in this paper.

References

[1] F. Li, K.Y. Lam, Z. Ni, D. Niyato, et al., "Cognitive Carrier Resource Optimization for Internet-of-Vehicles in 5G-Enhanced Smart Cities," *IEEE Network*, September 2021, DOI: 10.1109/MNET.211.2100340.

[2] F. Tang, Y. Zhou and N. Kato, "Deep Reinforcement Learning for Dynamic Uplink/Downlink Resource Allocation in High Mobility 5G HetNet," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, pp. 2773-2782, Dec. 2020.

[3] X. Shen, D. Liu, C. Huang, L. Xue, H. Yin, W. Zhuang, R. Sun, and B. Ying, "Blockchain for Transparent Data Management towards 6G," *Engineering* (Elsevier), to appear, 2021.

[4] W. Wu, C. Zhou, M. Li, H. Wu, H. Zhou, N. Zhang, X. Shen, and W. Zhuang, "AI-Native Network Slicing for 6G Networks," *IEEE Wireless Communications Magazine*, to appear, 2021.

[5] U. Abhishek and S. J. Darak, "Bayesian multi-armed bandit framework for multi-band auction based dynamic spectrum access in multi-user decentralized networks," *in Proc. URSI GASS*, pp. 1-4, 2017.

[6] Ye, Jingwei et al., "A Trust-Centric Privacy-Preserving Blockchain for Dynamic Spectrum Management in IoT Networks," *arXiv preprint* arXiv: 2106.13958, 2021.

[7] A. Ivanov, K. Tonchev, V. Poulkov and A. Manolova, "Probabilistic Spectrum Sensing Based on Feature Detection for 6G Cognitive Radio: A Survey," *IEEE Access*, vol. 9, pp. 116994-117026, 2021.

[8] J. Yang, Y. Duan, T. Qiao, et al., "Prototyping federated learning on edge computing systems," *Frontiers of Computer Science*, vol. 14, pp. 1-3, 2020.

[9] S. Wang, T. Tuor, T. Salonidis, et al., "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205-1221, 2019.

[10] Z. M. Fadlullah and N. Kato, "HCP: Heterogeneous Computing Platform for Federated Learning Based Collaborative Content Caching Towards 6G Networks," *IEEE Transactions on Emerging Topics in Computing*, 2020, DOI: 10.1109/TETC.2020.2986238.

[11] K. Lam, S. Mitra, F. Gondesen, X. Yi, "ANT-Centric IoT Security Reference Architecture-Security-by-Design for Satellite-Enabled Smart Cities," *IEEE Internet of Things Journal*, April 2021, DOI: 10.1109/JIOT.2021.3073734.

[12] Z. Shi, J. Liu, S. Zhang and Nei Kato, "Multi-Agent Deep Reinforcement Learning for Massive Access in 5G and Beyond Ultra-Dense NOMA System," *IEEE Transactions on Wireless Communications*, 2021, DOI: 10.1109/TWC.2021.3117859

[13] X. Liu, K. Lam, F. Li, et al., "Spectrum Sharing for 6G Integrated Satellite-Terrestrial Communication Networks Based on NOMA and CR," *IEEE Network*, vol. 35, no. 4, pp. 28-34, Jul. 2021, DOI: 10.1109/MNET.011.2100021.

[14] D. R. H. B. McMahan, E. Moore and B. A. Y. Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint* arXiv: 1602.05629, 2016.