Reinforcement Learning in Fog-based Internet-of-Things: Applications and Research Issues

XXX YYYY, and XXX ZZZZ

***Abstract*—Reinforcement learning (RL) algorithms are the platform for connecting intelligent Internet-of-Things (IoT) devices. The data provided by these intelligent devices can offer more insights about the overall futuristic functionalities in terms of usage and preferences. Machine learning algorithms have been widely used for predictions, with several optimization algorithms using reinforcement learning techniques to improve the performance and interaction of devices within a dynamic IoT ecosystem. In this paper, we present a comprehensive review of state-of-the-art RL techniques that have been applied to solve trending problems in the IoT and cyber-physical systems domain. Furthermore, we give an in-depth analysis into its applications in the emerging fog-based IoT networks. Finally, we propose alternative research directions that may yield better outcomes.**

***Index Terms*—Reinforcement learning (RL), Internet-of-Things (IoT), Fog-based IoT, Markov decision process (MDP), machine learning (ML).**

I. Introduction

REINFORCEMENT learning algorithms are the platform for connecting intelligent Internet-of-Things (IoT) devices. Over the years, there has been several applications of machine learning (ML) in IoTs and Cyber-physical systems (CPS). Nest learning thermostats are good examples of how IoT devices leverage data patterns to predict the preferred temperature in a room during a particular time of day. The prediction of the room temperature can also be on an aggregated neighborhood level, where energy loads can be remotely shifted by the power utility in homes operating Nest devices. Another practical application is the Amazon personal assistant that has the capability of learning voice patterns, the Jaguar's Land Monitoring system, which depends on a complex software that allows the automobile to observe, predict, monitor and notify the car's passengers to assist the driver automatically delegate his tasks and minimize the burden of driving.

Optimization is a very old field with interesting algorithms that has been used to solve simple to complex problems in various areas. Many optimization algorithms use RL techniques to optimize the behaviour of devices in the IoT ecosystem. For instance, in Intelligent Transport Systems (ITS), where cars act smartly, optimality in the interaction between cars and the environment is highly required. Several CPS applications like the Industrial Internet of Things (IIoT), smart grid, and ultimately smart cities, will allow intelligent machine-type interactions. These interactions may often will require robust algorithms that evolve and are adaptable, in order for the IoT devices to be able to achieve desired objectives given some set of operational parameters.

With the inherent heterogeneity in the IoT ecosystem and devices operating in a highly dynamic environment, it becomes necessary to leverage robust and resource-fitting machine learning techniques that will match the resource-constrained nature and stochastic behaviour of the IoT network. Moreover, the magnanimity of the IoT has led to the emergence of the fog-based IoT architecture [1], which promises to run IoT-enabled applications for real-time control and analytics, with millisecond response time. The realization of an optimal real-time control and data analytics is subject to the ability for agents to make optimal decisions that will improve the overall performance in the network. As such, we aim to examine RL techniques as used in previous works to help us understand the intrinsic benefits to the future IoT networks.

The contribution of our work is three-fold, which are listed as follows.

1) We review of some RL algorithms, and highlight key aspects that will drive future fog-based IoT networks.
2) We also perform simulations on some techniques considering some IoT scenarios. We also present detailed RL applications and use cases within the CPS (Intelligent transportation systems, smart grid, smart homes, smart health-care, and smart environment) to demonstrate how different techniques presented in the paper fuse to provide desirable objectives.
3) We also present a variety of open research challenges and suggest possible future trends for building intelligence in IoT, with regard to the latest development in the field.

The remainder of this work is organized as follows. In Section

II. Reinforcement Learning in IoT

Reinforcement Learning (RL) is learning that involves mapping situations to actions with an objective of maximizing a numerical reward. Actions are made by an agent, which have the ability to sense the state of the dynamic environment and consequently take actions that influence its environment. RL facilitates sequential decision making under uncertainty, thereby making it a useful tool in prediction of non-linear phenomenon [6]. Several works have been published in the

Manuscript received March 14, 2019; revised July X, 2019.

area of RL in IoT and CPS, however, there has not been any detailed review that covers their applicability in IoT and the underlying research issues in this field.

A. Basic Overview

1) Q-Learning: Q-learning is also known as an off-policy temporal difference (TD) learning algorithm, which allows the agent to learn about an optimal policy using an exploratory policy [7], [8], [9]. The Q-learning algorithm can be simply defined by

$$
\begin{aligned}
Q_{t+1}(s_t, a_t) &:= Q_t(s_t, a_t) \\
&+ \alpha\Big[r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)\Big].
\end{aligned} \tag{1}
$$

2) W-Learning: Unlike Q-learning which is a single-agent, single-policy learning technique, W-learning is a multi-agent, multi-policy learning technique that has been used on non-cooperating agents [13]. The W-learning algorithm is given by

$$
\begin{aligned}
W_i(s) &:= (1-\alpha)W_i(s) \\
&+ \alpha\Big[Q_i(s, a_i) - (r_i + \gamma \max_a Q(s', a_i'))\Big].
\end{aligned} \tag{2}
$$

3) SARSA: State-action-reward-state-action (SARSA) is an On-policy TD learning algorithm, which the agent learns an action-value function instead of a state-value function. SARSA always converges to an optimal policy so long as all state-action pairs are visited an infinite number of times.

$$
\begin{aligned}
Q_{t+1}(s_t, a_t) &:= Q_t(s_t, a_t) \\
&+ \alpha\Big[r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)\Big].
\end{aligned} \tag{3}
$$

4) Deep Q Net (DQN):

5) Deep Deterministic Policy Gradients (DDPG):

6) Normalized Advantage Functions (NAF):

7) Asynchronous Advantage Actor-Critic (A3C):

B. Related Works in IoT

Over the years, many perceived difficult sub-problems have began to receive research attention, some of which are categorized as follows.

1) *Energy allocation and utilization within the network*: IoT devices deplete energy in several ways, such as when transmitting and receiving data, listening to the medium, performing computation or even changing position. There exist several works that have used RL and its variants to optimize energy usage in different domains, some of which could be modified to suit the resource-limitation in the typical IoT domain. Zhong *et al.* in [2] designed sensor energy utilization and power allocation in IoT sensors to maximize the total data throughput of the system. They proposed a low-complexity algorithm to optimally solve the two formulated sub-problems. The energy utilization sub-problem was formulated as a MDP, while the power allocation sub-problem was approached using discrete optimization. The proposed algorithm performed better than some simple heuristics when simulated. However, the work considered a scenario where the channel state is perfectly known.

   A Markov-based analytical model was integrated with a RL process in Conti *et al.* [21] to optimize the server activation policy, where optimal control of an energy storage system in a green fog-computing node is needed to improve the system performance, hereby allowing the system to bear high job arrivals even at low-power generation periods. The fog-computing nodes used in this work are not typical of those used in a typical fog-based IoT environment. A Q-learning-based adaptive power management was proposed by Debizet *et al.* in [27] to optimize the power consumed during the suspend state of an Internet-of-Things (loT) System-on-Chip (SoC) with eight embedded power states. The algorithm out-performed some expert-based solution for most suspend sequences under 12000 cycles, however, with an additional power cost. Just like the traditional IoT architecture, the fog-based IoT networks are often power-constrained and as such, require low complexity learning algorithms to optimize the energy utilization.

2) *Latency Minimization*: Mission-critical and delay-sensitive IoT application can effectively leverage the intelligence of some RL techniques. An RL approach that made use of evolution strategies for real-time task assignment among fog servers was introduced by Mai *et al.* in [22] to minimize the total computation latency during a long-term period. The paper claimed that the proposed model is scalable when the number of IoT devices increases, however, no experimentations were carried out on a real-world IoT scenario. A dynamic programming based duty cycle control technique was employed by Li *et al.* in [25] to provide an optimal solution to an inventory control problem. The focus was on optimizing the duty cycle by jointly considering energy efficiency, end-to-end delay and reliability of the network. A two-hop cluster tree network model was used in the work. A distributed multi-agent Q-learning algorithm was proposed by Liu *et al.* in [30] to optimize both the motorized and non-motorized traffic. The work considered several constraints to help depict real operational scenarios. The paper applied AI powered Internet-of-Things (AIoT) technologies for traffic light control, which is a key to the success of ITS. However, the approach was limited to stationary agents. A dynamic programming approach was deployed by Routray *et al.* in [29] to provide near optimal results under high traffic conditions in an IoT network scenario with consideration for node mobility. A DP approach was used in routing in a dynamically changing network. However, all the nodes in the network were assumed to be routers, which is often not the case in a typical IoT environment.

   Despite the interesting proximity feature of IoT end-devices to the network edge in the fog-based IoT network, delay is prevalent due to several bottlenecks, which may lead system failure or operational hazard for critical applications. These techniques may be used to minimize the delay especially for high priority communications.

3) *Channel assessment and intelligent routing*: Prior to communicating via a link, it is very useful for nodes within the network to assess the communication medium. Learning by experience using some RL techniques will go a long way in minimizing number of loss packets in highly stochastic IoT environment. A Q-learning relay selection algorithm was introduced by Jadoon *et al.* in [3] to maximize the throughput in a typical cooperative network with relays supporting communication between source and destination. The relays were assumed to be the states, with three possible actions of remaining in the present state with relay r or choosing $r+1$ or $r-1$. However, the outage analysis was carried out without a closed-form expression for the outage probability within the network. Furthermore, in an attempt to reduce the complexity of the formulated problem, the problem algorithm did not consider the relays (states) which do not satisfy the minimum mutual information constraint. A scalable parallel Q-learning algorithm was introduced by Camelo *et al.* in [32] to minimize communication cost. The work considered distributed and resource-constrained environment. A Q-learning-based duty cycle control technique was introduced by Li *et al.* in [24] to provide improved performance and reliable M2M communication for IoT applications. However, the proposed Q-learning based duty cycle control only considered a two-hop cluster tree network in evaluating the performance of the system. Despite system dynamics and heterogeneity within the ultra-distributed IoT environment, it is important for devices which are mobile to move seamlessly without degradation of the quality of service (QoS) when communicating. Moreover, the communication should be ubiquitous irrespective of technology or domain within the system. As such, the RL techniques will play a very important role in areas of IoT device/service discovery, which will support adaptability and interfacing between devices and sub-networks within the IoT domain.
4) *Scheduling*: Wen *et al.* in [8] formulated an automated energy management system (EMS) rescheduling problem as a reinforcement learning (RL) problem. Simulations were carried out using Q-learning technique on a specific scenario with good results. The paper considers the EMS to act as an agent for energy users. A variant of Q-learning, Deep Q-learning, was introduced by Zhu *et al.* in [10] to maximize system throughput by applying appropriate scheduling strategy for cognitive radio-based IoT networks. This work performed poorly when compared with the strategy iteration algorithm, though with a reduced computational complexity.
5) *Resource Allocation and Services*: An RL-based resource allocation in fog radio access networks was presented by Nassar *et al.* in [33] for various IoT environments. A cooperative reinforcement learning algorithm was introduced by Khan *et al.* in [28] for adaptive power allocation in D2D communication. When compared with the Distributed RL algorithm, the proposed approach provides better system throughput as well as D2D throughput with less interference to pave way for the massive MTC in IoT. However, the experiments were conducted in a single cell. The RL-based mapping table (RLMT) and RL-based resource allocation algorithms were introduced by Gai *et al.* in [12] to handle cost mapping tables creation and for optimal resource allocation in IoT content-centric services. The work was limited to the use QoE in examining service level, which allowed a dynamic resource allocation and avoided a fixed task table. A semi-supervised DRL model that suits IoT and smart city services was introduced in Mohammadi *et al.* in [18], using both labeled and unlabeled data for performance improvement and accuracy in the learning agent. The work employed an indoor localization based on the Bluetooth low energy signal strength, and limited its findings to a single floor of a building.

TABLE I
SOME IOT RELATED WORKS ADDRESSING KEY ISSUES

Issues	*Some related works*
Energy allocation and utilization	[2], [21], [27]
Latency Minimization	[22], [25], [30], [29]
Channel assessment and intelligent routing	[3], [32], [24]
Scheduling	[8], [10]
Resource Allocation and Services	[33], [28], [12], [18]

TABLE II
CATEGORIES OF DIFFERENT OBJECTIVES WITHIN IOT

Objective	*Work*	*Category*
Maximize data throughput	[2]	SA-based
	[3]	MA-based
Optimize server activation policy	[21]	SA-based
Minimize computational latency	[22]	MA-based
Optimize duty cycle	[25]	MP-based
Optimize traffic	[29]	MA-based
	[30]	MA-based
Minimize communication cost	[4]	SA-based
	[24]	SA-based
	[32]	MA-based
Optimize scheduling strategy	[8]	SA-based
	[11]	SA-based
Optimize power allocation	[28]	MA-based
Optimize resource allocation	[12]	SA-based
	[18]	SA-based

C. Works in WSN

In [34], a reliable and energy-efficient routing (REER) protocol was proposed using a geographic routing approach. The work considered the idea of a central entity, called a reference node (RN), which is assumed to be situated at an ideal location between source and destination. Several other cooperative nodes, which contend to relay data, are assumed to be situated around the RN. The work was able to examine the trade-off between reliability and energy-efficiency when the distances between RNs was adjusted. The work in [35] proposed an improvement to [34], addressing the issue of scalability in a dynamic network. As such, a novel distributed multi-hop cooperative communication scheme (DMC) was proposed to improve certain QoS metrics such as the communication energy, hop count, and end-to-end delay. However, both works assumed complete prior knowledge of the network environment. In [36], a multi-agent reinforcement learning-based multi-hop mesh cooperative (MRL-CC) mechanism for the improvement of some QoS metrics such as the end-to-end delay, packet delivery ratio in the WSNs. The work outperformed the work in [37], which incurred higher delay, and was distance-based, which is not suitable for a dynamic network environment. Though the mobility of the cooperative nodes were taken into account when learning the optimal policy, the MRL-CC failed to consider the power dissipated by the power-constrained devices.

A self-organizing RL technique for scheduling the wake-up cycles of nodes was proposed in [38] to effectively ensure the successful delivery of messages along a routing tree while minimizing interference. A node synchronization /desynchronization strategy was used to minimize interferences, as well as decrease the data retrieval and energy consumption. Experimentation was carried using three static topologies (line, mesh and grid) with results averaged over 50 runs, however, this does not depict a typical dynamic IoT environment. In [39], a distributed independent reinforcement learning (DIRL) techniques, which is based on Q-learning, was introduced to allow autonomous self-learning applications with inherent support for efficient task management. The DIRL approach showed significant performance improvement when compared to traditional static scheduling schemes. A resource management framework based on a two-tier RL scheme was proposed in [40] to enable autonomous self-learning and adaptive applications with inherent support for efficient resource management. However, the works [39], [40] only considered task scheduling.

A RL-based Q-routing technique, the Feedback Routing for Optimizing Multiple Sinks (FROMS), was proposed in [43], [44] to enable efficient routing to multiple sinks without overhead. Simulations were done under scenarios where the sink nodes are mobile and with consideration for node failure. However, the mobility pattern and speed was bounded. Moreover, an important cost metric such as the energy consumed by the mobile nodes was not considered. In [45], an evaluation was carried out on the previously proposed FROMS framework using real WSN hardware. The objective was to show that machine learning algorithms could be efficiently deployed on resource-constrained devices. However, during the implementation phase, certain bottlenecks were encountered such as dynamic memory allocation, interference in the wireless channel, poor data gathering, and difficulty in handling lost or corrupted packets. As such, a data clustering and aggregation approach, CLIQUE, was proposed in [46] to minimize the energy expended by the selecting a cluster head using Q-learning. The approach was able to save significant amount of energy when compared to the traditional and random cluster head selection approach.

TABLE III
SOME WSN RELATED WORKS ADDRESSING KEY ISSUES

Issues	*Some related works*
Energy allocation and utilization	[34], [35]
Latency Minimization	[36], [37]
Channel assessment and intelligent routing	[43], [44], [45], [46]
Scheduling	[38]
Resource Allocation and Services	[39], [40]

TABLE IV
SUMMARY OF WORKS IN REINFORCEMENT LEARNING.

Reinforcement learning techniques	Details	Limitations	Potential contributions
RL-based mapping table (RLMT) and RL-based resource allocation	These algorithm was introduced Gai *et al.* in [12] to handle cost mapping tables creation and for optimal resource allocation in IoT content-centric services.	The work was limited to the use QoE in examining service level, which allowed a dynamic resource allocation and avoided a fixed task table.	To apply RL-based approach on several other KPIs that may improve IoT services.
Deep-Q-learning	This variant of Q-learning was introduced by Zhu *et al.* in [10] to maximize system throughput by applying appropriate scheduling strategy for cognitive radio-based IoT networks.	This work performed poorly when compared with the strategy iteration algorithm, though with a reduced computational complexity.	To adapt the concept of cognition to our proposed IoT network and with the consideration of several network dynamics.
Deep Reinforcement Learning (DRL)	A semi-supervised DRL model that suits IoT and smart city applications was introduced in Mohammadi *et al.* in [18] for performance improvement and accuracy in the learning agent.	The work employed an indoor localization based on the Bluetooth low energy signal strength, and limited its findings to a single floor of a building.	To consider multi-agent environment and conduct some outdoor experimentation.
RL using a Markov-based analytical model	A Markov-based analytical model was integrated with a RL process in Conti *et al.* [21] to optimize the server activation policy, where optimal control of an energy storage system in a green fog-computing node is needed to improve the system performance, hereby allowing the system to bear high job arrivals even at low-power generation periods.	The fog-computing nodes used in this work is fixed and has a large energy source, which fails to depict the resource-constrained nature of fog devices.	We will consider a multi-tier fog architecture with lots of heterogeneity, having static and mobile fog nodes which may or may not be power-constrained.
Q-learning	Wen *et al.* in [8] formulated an automated energy management system (EMS) rescheduling problem as a reinforcement learning (RL) problem. Simulations were carried out using Q-learning technique on a specific scenario with good results.	The paper considers the EMS to act as an agent for energy users. This approach is not practical in a real IoT scenario, where nodes are mobile with dynamic system requirements.	We will consider a decentralized agent-based system to support for a realistic IoT scenario.
Evolutionary strategies and RL	A RL approach that made use of evolution strategies for real-time task assignment among fog servers was introduced by Mai *et al.* in [22] to minimize the total computation latency during a long-term period.	The paper claimed that the proposed model is scalable when the number of IoT devices increases, however, the approach proposed failed to examine real-world IoT-scaled scenario.	We aim to apply a variant of these techniques to realistic IoT scenarios.
Q-learning-based duty cycle control	This RL-based duty cycle control technique was introduced by Li *et al.* in [24] to provide improved performance and reliable M2M communication for IoT applications.	The performance evaluation of the proposed Q-learning based duty cycle control only considered a two-hop cluster tree network.	The proposed approach can be enhanced to capture a larger network size, and DP approach employed may not be suitable.
Dynamic programming based duty cycle control	This technique used by Li *et al.* in [25] to provide an optimal solution to an inventory control problem. The focus was on optimizing the duty cycle by jointly considering energy efficiency, end-to-end delay and reliability of the network.	A two-hop cluster tree network model was used in the work, and the problem was evaluated using the DP approach	The proposed approach can be enhanced to capture a larger network size, and the DP approach used may not be suitable for a realistic IoT scenario.
Predictive and Resilient Q-learning	A variant of Q-learning algorithm employed by Grammatopoulou *et al.* in [26], which considers historical data about irregular operations such as faults and attacks by malicious agents in an IoT network (a smart water supply system).	The work relied upon historical data about irregular operations such as faults and attacks by malicious agents which may not be readily available in a typical IoT network.	We will attempt to depict a realistic IoT scenario where historical data will not be available, and the agent will be required to learn the optimal policy in a new and dynamic environment.

Summary of works in Reinforcement Learning. (Contd. 1)

Reinforcement learning techniques	Details	Limitations	Potential contributions
Q-learning-based adaptive power management	A Q-learning-based adaptive power management was proposed by Debizet *et al.* in [27] to optimize the power consumed during the suspend state of an Internet-of-Things (loT) System-on- Chip (SoC) with 8 embedded power states.	The algorithm out-performed some expert-based solution for most suspend sequences under 12000 cycles, however, with an additional power cost.	The parameters used in the study may be better tuned to produce more satisfactory performance.
Cooperative reinforcement learning	A cooperative reinforcement learning algorithm was introduced by Khan *et al.* in [28] for adaptive power allocation in D2D communication.	When compared with the Distributed RL algorithm, the proposed approach provides better system throughput as well as D2D throughput with less interference to pave way for the massive MTC in IoT. However, the experiments were conducted in a single cell.	Experiments using the proposed algorithm may be conducted to in a heterogenous multi-cell environment to depict a realistic IoT scenario.
Dynamic programming	A dynamic programming approach was deployed by Routray *et al.* in [29] to provide near optimal results under high traffic conditions in an IoT network scenario with consideration for node mobility.	The DP approach was used in routing in a dynamically changing network. However, all the nodes in the network were assumed to be routers, which is often not the case in a typical IoT environment.	We will consider heterogeneity in IoT device, with some devices IoT sensors, and fog devices, which may be static or dormant, active or inactive.
Multi-agent Q-learning	A distributed multi-agent Q learning algorithm was proposed by Liu *et al.* in [30] to optimize both the motorized and non-motorized traffic. The work considered several constraints to help depict real operational scenarios.	The paper applied AI powered Internet-of-Things (AIoT) technologies for traffic light control, which is a key to the success of ITS. However, the agents considered are static and may not be suitable for other IoT applications.	This approach may be applied in scenarios of non-stationary agents to examine the overall system performance in a dynamic IoT environment.
On-line RL	A Q-learning algorithm was introduced by Dias *et al.* in [31] to adjust sensors' sampling interval in real-time, with respect to environmental conditions and application requirements.	This work was able to learn the most suitable sampling intervals under different conditions, without an a-priori model of the environment's evolution. However, it considered only static sensor devices, of which no clear indication on where the learning computation will take place.	Considering the resource-limitation of IoT sensor nodes, we leverage on the fog/edge computing paradigm to provide some learning and computational capability.
Parallel Q-learning	A scalable parallel Q-learning algorithm was introduced by Camelo *et al.* in [32] to minimize communication cost.	The work considered distributed and resource-constrained environment.	y.
Parallel Q-learning	A scalable parallel Q-learning algorithm was introduced by Camelo *et al.* in [32] to minimize communication cost.	The work considered distributed and resource-constrained environment.	yyyyy.

REFERENCES

[1] B. Omoniwa, R. Hussain, M. A. Javed, S. H. Bouk and S. A. Malik, "Fog/Edge Computing-based IoT (FECIoT): Architecture, Applications, and Research Issues," in IEEE Internet of Things Journal.

[2] S. Zhong and X. Wang, "Energy Allocation and Utilization for Wirelessly Powered IoT Networks," in IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2781-2792, Aug. 2018.

[3] M. A. Jadoon and S. Kim, "Relay selection Algorithm for wireless cooperative networks: a learning-based approach," in IET Communications, vol. 11, no. 7, pp. 1061-1066, 11 5 2017.

[4] B. Omoniwa et al., "An Optimal Relay Scheme for Outage Minimization in Fog-based Internet-of-Things (IoT) Networks," in IEEE Internet of Things Journal.

[5] S. Earley, "Analytics, Machine Learning, and the Internet of Things," *IT Professional,* vol. 17, no. 1, pp. 10-13, Jan.-Feb. 2015.

[6] D. Zhang, X. Han and C. Deng,"Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems,* vol. 4, no. 3, pp. 362-370, September 2018.

[7] C. J. C. H. Watkins and P. Dayan, Machine Learning (1992), Kluwer Academic Publishers, vol. 8: 279.

[8] Z. Wen, D. O'Neill and H. Maei, "Optimal Demand Response Using Device-Based Reinforcement Learning," *IEEE Transactions on Smart Grid,* vol. 6, no. 5, pp. 2312-2324, Sept. 2015.

[9] R. S. Sutton, and A. G. Barto, *Reinforcement learning - an introduction.* Cambridge, MA: MIT Press, 1998.

[10] J. Zhu, Y. Song, D. Jiang and H. Song, "A New Deep-Q-Learning-Based Transmission Scheduling Mechanism for the Cognitive Internet of Things," *IEEE Internet of Things Journal,* vol. 5, no. 4, pp. 2375-2385, Aug. 2018.

[11] J. Zhu, Z. Peng and F. Li, "A transmission and scheduling scheme based on W-learning algorithm in wireless networks," *2013 8th International Conference on Communications and Networking in China (CHINACOM),* Guilin, 2013, pp. 85-90.

[12] K. Gai, M. Qiu, "Optimal resource allocation using reinforcement learning for IoT content-centric services," *Applied Soft Computing,* vol. 70, pp. 12-21, Sept. 2018.

[13] I. Dusparic and V. Cahill, "Distributed W-Learning: Multi-Policy Optimization in Self-Organizing Systems," 2009 Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems, San Francisco, CA, 2009, pp. 20-29.

[14] L. Zhou, A. Swain and A. Ukil, "Q-learning and Dynamic Fuzzy Q-learning Based Intelligent Controllers for Wind Energy Conversion Systems," *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia),* Singapore, 2018, pp. 103-108.

[15] T. Park, N. Abuzainab and W. Saad, "Learning How to Communicate in the Internet of Things: Finite Resources and Heterogeneity," *IEEE Access,* vol. 4, pp. 7063-7073, 2016.

[16] A. Hans and S. Udluft, "Ensembles of Neural Networks for Robust Reinforcement Learning," *2010 Ninth International Conference on Machine Learning and Applications,* Washington, DC, 2010, pp. 401-406.

[17] D. D. Nguyen, H. X. Nguyen and L. B. White, "Reinforcement Learning With Network-Assisted Feedback for Heterogeneous RAT Selection," *IEEE Transactions on Wireless Communications,* vol. 16, no. 9, pp. 6062-6076, Sept. 2017.

[18] M. Mohammadi, A. Al-Fuqaha, M. Guizani and J. Oh, "Semisupervised Deep Reinforcement Learning in Support of IoT and Smart City Services," *IEEE Internet of Things Journal,* vol. 5, no. 2, pp. 624-635, April 2018.

[19] S. Alletto et al., "An Indoor Location-Aware System for an IoT-Based Smart Museum," *IEEE Internet of Things Journal,* vol. 3, no. 2, pp. 244-253, April 2016. doi: 10.1109/JIOT.2015.2506258

[20] K. Kolomvatsos, and C. Anagnostopoulos, "Reinforcement Learning for Predictive Analytics in Smart Cities," *Informatics,* vol. 3, no. 16, June 2017.

[21] S. Conti, G. Faraci, R. Nicolosi, S. A. Rizzo and G. Schembra, "Battery Management in a Green Fog-Computing Node: a Reinforcement-Learning Approach," *IEEE Access,* vol. 5, pp. 21126-21138, 2017.

[22] L. Mai, N.-N. Dao, and M. Park, "Real-time task assignment approach leveraging reinforcement learning with evolution strategies for long-term latency minimization in fog computing," *Sensors,* vol. 18, no. 2830, Aug. 2018.

[23] C. Kwok and D. Fox, "Reinforcement learning for sensing strategies," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566),* Sendai, 2004, pp. 3158-3163 vol.4.

[24] Y. Li, K. K. Chai, Y. Chen and J. Loo, "Smart duty cycle control with reinforcement learning for machine to machine communications," *2015 IEEE International Conference on Communication Workshop (ICCW),* London, 2015, pp. 1458-1463.

[25] Y. Li, K. K. Chai, Y. Chen and J. Loo, "Optimised delay-energy aware duty cycle control for IEEE 802.15.4 with cumulative acknowledgement," *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC),* Washington, DC, 2014, pp. 1051-1056.

[26] M. Grammatopoulou, A. Kanellopoulos, and K. G. Vamvoudakis, "A multi-step and resilient predictive Q-learning algorithm for IoT: a case study in water supply networks", *ACM Proceedings of the 8th International Conference on the Internet of Things,* Santa Barbara, California, 2018, pp. 1-8.

[27] Y. Debizet, G. Lallement, F. Abouzeid, P. Roche and J. Autran, "Q-Learning-based Adaptive Power Management for IoT System-on-Chips with Embedded Power States," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), Florence, 2018, pp. 1-5.

[28] M. I. Khan, M. M. Alam, Y. Le Moullec and E. Yaacoub, "Cooperative reinforcement learning for adaptive power allocation in device-to-device communication," 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 2018, pp. 476-481.

[29] S. K. Routray and Sharmila K. P., "Routing in dynamically changing node location scenarios: A reinforcement learning approach," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 458-462.

[30] Y. Liu, L. Liu and W. Chen, "Intelligent traffic light control using distributed multi-agent Q learning," 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, 2017, pp. 1-8.

[31] G. M. Dias, M. Nurchis and B. Bellalta, "Adapting sampling interval of sensor networks using on-line reinforcement learning," 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, 2016, pp. 460-465.

[32] M. Camelo and J. F. a. S. Latré, "A Scalable Parallel Q-Learning Algorithm for Resource Constrained Decentralized Computing Environments," 2016 2nd Workshop on Machine Learning in HPC Environments (MLHPC), Salt Lake City, UT, 2016, pp. 27-35.

[33] A. T. Nassar, and Y. Yilmaz, "Reinforcement-Learning-Based Resource Allocation in Fog Radio Access Networks for Various IoT Environments," Arxiv

[34] M. Chen, T. Kwon, S. Mao, Y. Yuan and V. C. M. Leung, "Reliable and energy-efficient routing protocol in dense wireless sensor networks," Int. J. Sen. Netw. vol. 4, no. 1/2, pp. 104-117, July 2008.

[35] M. Chen, M. Qiu, L. Liao, J. Park and J. Ma, "Distributed multi-hop cooperative communication in dense wireless sensor networks", *The Journal of Supercomputing,* vol. 56, no. 3, pp. 353-369, June 2011.

[36] Xuedong Liang, Min Chen, Yang Xiao, I. Balasingham and V. C. M. Leung, "A novel cooperative communication protocol for QoS provisioning in wireless sensor networks," 2009 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities and Workshops, Washington, DC, 2009, pp. 1-6.

[37] M. Chen, Xuedong Liang, V. Leung and I. Balasingham, "Multi-hop mesh cooperative structure based data dissemination for wireless sensor networks," 2009 11th International Conference on Advanced Communication Technology, Phoenix Park, 2009, pp. 102-106.

[38] M. Mihaylov and Y. L. Borgne, "Decentralised reinforcement learning for energy-efficient scheduling in wireless sensor networks," Int. J. Communication Networks and Distributed Systems, Vol. 9, Nos. 3/4, 2012

[39] K. Shah and M. Kumar, "Distributed Independent Reinforcement Learning (DIRL) Approach to Resource Management in Wireless Sensor Networks," 2007 IEEE International Conference on Mobile Adhoc and Sensor Systems, Pisa, 2007, pp. 1-9.

[40] K. Shah, M. D. Francesco and M. Kumar, "Distributed resource management in wireless sensor networks using reinforcement learning," Wireless Networks, vol. 19, no. 5, pp. 705-724, July 2013.

[41] K.-L. A. Yau, P. Komisarczuk, P. D. Teal, "Reinforcement learning for context awareness and intelligence in wireless networks: Review, new features and open issues," Journal of Network and Computer Applications, vol. 35, no. 1, pp. 253-267, Jan. 2012.

[42] K.-L. A. Yau, H. G. Goh, D. Chieng and K. H. Kwong, "Application of reinforcement learning to wireless sensor networks: models and algorithms," Computing, vol. 97, no. 11, pp. 1045–1075, Nov. 2015.

[43] A. Egorova-Forster and A. L. Murphy, "Exploiting Reinforcement Learning for Multiple Sink Routing in WSNs," 2007 IEEE International Conference on Mobile Adhoc and Sensor Systems, Pisa, 2007, pp. 1-3.

[44] A. Forster and A. L. Murphy, "FROMS: Feedback Routing for Optimizing Multiple Sinks in WSN with Reinforcement Learning," 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, Melbourne, Qld., 2007, pp. 371-376.

[45] A. Forster, A. L. Murphy, J. Schiller and K. Terfloth, "An Efficient Implementation of Reinforcement Learning Based Routing on Real WSN Hardware," 2008 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, Avignon, 2008, pp. 247-252.

[46] A. Forster and A. L. Murphy, "CLIQUE: Role-Free Clustering with Q-Learning for Wireless Sensor Networks," 2009 29th IEEE International Conference on Distributed Computing Systems, Montreal, QC, 2009, pp. 441-449.

[47] N. Vucevic, J. Perez-Romero, O. Sallent and R. Agusti, "Reinforcement Learning for Active Queue Management in Mobile All-IP Networks," 2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, 2007, pp. 1-5.

[48] W. T. B. Uther and M. M. Veloso, "Tree Based Discretization for Continuous State Space Reinforcement Learning," *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence,* 1998, Madison, Wisconsin, USA, pp. 769-774.

[49] H. Cuayahuitl, S. Renals, O. Lemon, and H. Shimodaira. "Learning multi-goal dialogue strategies using reinforcement learning with reduced state-action spaces," In Int. Journal of Game Theory, pp. 547–565, 2006.