Reinforcement Learning Based Communication Protocols for Industrial WSNs: A Critical Mini-Review.

Ismahene Alem, Samira Yessad, Louiza Bouallouche-Medjkoune

▶ **To cite this version:**

Ismahene Alem, Samira Yessad, Louiza Bouallouche-Medjkoune. Reinforcement Learning Based Communication Protocols for Industrial WSNs: A Critical Mini-Review.. Colloque sur les Objets et Systèmes Connectés 2023, Institut Supérieur des études technologiques de Sfax; Institut Supérieur des études technologiques de Mahdia, Jun 2023, Mahdia, Tunisie. hal-04219667

HAL Id: hal-04219667
https://hal.science/hal-04219667

Submitted on 27 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reinforcement Learning Based Communication Protocols for Industrial WSNs: A Critical Mini-Review

Ismahene Alem
LAMOS Research Unit
Faculty of Exact Sciences
University of Bejaia
06000 Bejaia, Algeria
ismahene.alem@univ-bejaia.dz

Samira Yessad
LAMOS Research Unit
Faculty of Exact Sciences
University of Bejaia
06000 Bejaia, Algeria
samira.yessad@univ-bejaia.dz

Louiza Bouallouche-Medjkoune
LAMOS Research Unit
Faculty of Exact Sciences
University of Bejaia
06000 Bejaia, Algeria
louiza.medjkoune@univ-bejaia.dz

***Abstract*—Reinforcement Learning (RL) is a promising technology that has made a significant improvement in Industrial Internet of Things (IIoT) based applications by providing a high level of accuracy and optimal resource utilization while handling complicated problems efficiently. This paper aims to help future researchers to design efficient communication protocols for IIoT-based networks by exploiting the power of the RL technology. We provide a critical mini review on the recent RL-based MAC and routing protocols intended for IIoT-based applications, focus on Wireless Sensor Networks (WSN), and show through in-depth analysis how RL is exploited to address their related issues. The comparison between the reviewed works shows that the learning model's complexity, real-time communications, and security should deserve more consideration during the development of communication protocols to meet the specific needs of IIoT applications.**

***Index Terms*—Reinforcement Learning, IIoT, WSN, Routing Protocols, MAC Protocols.**

I. Introduction

Internet of Things (IoT) is a promising paradigm in the era of industry 4.0 that uses wireless technologies (e.g. Wi-Fi, BLE, Zigbee, 6LoWPAN, LTE/5G, WiMAX, LRWPAN, etc.) to interconnect smart objects allowing their management [1]. These smart objects like smartphones, RFID tags, sensors and actuators need minimal human intervention to communicate, collect, and process data. Industrial IoT is a subset of IoT which employs the IoT concept in manufacturing to connect industrial things, such as engines, machines, and robots over the industrial network and focus on real-time monitoring, scalability, reliability, security and efficiency in industries with low cost and high business profits [2]. Storage Tank monitoring, Smart Grid monitoring, pipeline monitors, Steam Distribution Lines and Fuel Supply Systems are just few examples of the numerous Industrial systems that adopt IIoT concept to satisfy their high-level requirements [3].

Today, most IIoT based applications use WSN technology to control and monitor the functioning of the machines. Therefore, large number of IoT sensors are deployed in the smart factories and industrial environments to establish communication between various machines, collect and process a large amount of observation data, and transfer it to the cloud for analysis. This relationship between IIoT and WSN provides to manufactories tremendous potentials that reduce operational costs, predict malfunction, minimize downtime, and even take measure in dangerous situations [4]. However, the small memory size and the short battery life of sensor nodes make it limited in processing power and communication capacity leading to several challenges whenever designing these networks such as; low complexity, real-time communications, low-latency, low power consumption, high reliability, heterogeneity, mobility, scalability, and security [5]. Addressing such challenges is often started with the design of MAC and routing protocols because it radically reduces the energy consumption and enhances the data transport efficiency of the designed network.

Over the two last decades, advanced technologies (e.g. Edge/Fog computing, Big data, Cloud computing, Blockchain, RL, Cyber physical systems, etc.) has been broadly investigated to address the above mentioned challenges [6]. This survey paper focuses on RL-based communication protocols and covers both RL based MAC and Routing protocols including scheduling, routing, clustering, and data aggregation protocols. We provide a classification of the main issues of WSN based communication protocols and their studied RL-based solutions. We study some works of the last five years and show if they fit the strict requirements of IIoT applications in term of low computational complexity (time and space complexity), real-time communications, high reliability, low power consumption, heterogeneity, mobility, and security.

The rest of the paper is organized as follows; an overview on RL technique is provided in section 2. Section 3 presents the recent proposed RL-based approaches by classifying them into two main classes; RL-based MAC protocols and RL-based routing protocols. Section 4 provides a summary analysis of the studied works based on several features. Finally, section 5 concludes the survey.

II. Reinforcement Learning

Reinforcement learning (RL) is a biologically inspired method and an important class of machine learning in the artificial intelligence system [7]. It helps to solve sequential decision making and large-scale optimization problems by enforcing the system learning capabilities based on historical experience [8]. The RL model is composed of the following components; the agent, the environment, the reward, and the policy. The agent or decision-maker can efficiently select its best actions in the future by learning a policy through interactions with its environment, which can be modeled as a Markov decision process (MDP) [8], [9]. As a result of a selected action, the agent will receive a reward from its environment. The reward value can be either positive or negative according to the goodness of the recent taken action. The policy is the strategy used by the learning agent to select the best action. The goal of RL is to let the agent learning suitable actions and optimize the policy to get the maximum cumulative reward in order to achieve a goal by interacting over time with its environment [7].

Basic reinforcement learning is based on Markov chains property where the expected reward value is independent of the past rewards and the agent does not have any initial idea of the next transition state. At each time step $t \in T = \{0, 1, 2 \ldots\}$, the environment is in a current state $S_t \in S$ with current reward $R_t \in R$. The agent takes an action $A_t \in A$ from the set of available actions, then the environment changes its state to another state S_{t+1}, obtains a reward R_{t+1} after the transition from S_t to S_{t+1} under A_t, and the process is repeated as shown in Fig. 1 [7], [9]. In WSN, a state may represent for example the residual energy of sensor node, the selected action represents for example the next hop node for routing packets, and the reward evaluates the network performances (such as throughput, delay, latency) when taking an action in a particular state at a particular time instant [10].

The RL methods are categorized as follows:

A. Model based RL methods

The agent learns with a model of its environment which provides much faster convergence of the algorithm. This model gives predictions about the outcomes of any (state, action) pair. However, this kind of methods are less popular because they depend on the initial environment model and its accuracy [7].

B. Model free RL methods

In these methods, the agent learns from its experiences explicitly by trail-and-error rule and adjusts its policy without any model of the environment.

Q-learning is the most promising form of model-free RL techniques, proposed by Watkins in 1989 [7]. In WSN, Q-learning algorithm has been frequently used to solve the related challenges and it has made a significant contribution in the development of WSN based systems. In Q-learning algorithm, the agent tries to optimize its policy trough trial and error, selects the optimal action given the current state at the given time step and obtains delay reward (i.e reward can be received far in the future). The expected future total reward of a particular (state-action) pair, also known as Q-value (quality value) is estimated using temporal differences (TD) and updated at every time step using the Bellman equation (1) and stored in a table called Q-table through an iterative approach [9], [10].

$$Q_{t+1}(S_t, A_t) \leftarrow (1-\alpha) Q_t(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \max_{a \in A} Q_t(S_{t+1}, a) \Big]. \quad (1)$$

Where : $\alpha \in [0, 1]$ is the learning rate (LR) that controls the updating of the Q-value and plays an important role in the convergence rate of the algorithm.

$\gamma \in [0, 1]$ is the discount factor that sets the preference for either immediately receiving rewards or deferring them.

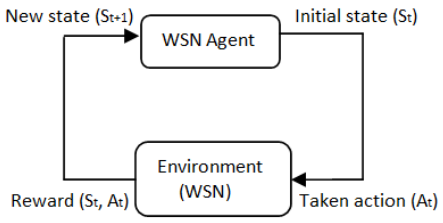

Fig. 1. The reinforcement learning model.

The biggest limitation of applying basic Q-learning algorithm is the curse of system dimensionality, i.e. when the size of Q-table increases due to the large actions/states space. In this case, the search time will increase exponentially and more storage space will be required hence the tabular Q-learning will not be practical since the insufficient memory of most IIoT based systems including WSNs. To address these issues, researchers tend to use Deep Q-learning Network (DQN) to extend the scalability of RL by combining the classical Q-learning and the Convolutional Neural Networks (CNN).

In IIoT based systems, we generally deal with multiple agents (sensors, machines, robots, etc.) which interact within a same environment. The behavior of agents can be cooperative, competitive or neutral according to the application requirements. More improved variants of the Q-Learning algorithm which are suitable for Multi-agent systems like Nash Q-Learning, Modular Q-Learning, and Ant Q-Learning are well described in [9].

III. RL based communication protocols in Wireless Sensor Networks

Reinforcement learning has been tailored in various algorithms of wireless sensor networks such as scheduling, routing, clustering and data aggregation. For the best understand of these algorithms, we categorize the selected works into two groups; RL-based MAC protocols and RL-based Routing protocols. Also, the space/time and the communication overhead of each proposed algorithm are discussed and a summary

analysis is given. Fig. 2 summarizes the major applications of reinforcement learning used to solve several issues in WSNs.

A. RL-based MAC protocols

In [11], Li et al. proposed a Q-learning-based dynamic spectrum access method to improve spectrum efficiency and help unlicensed users (also called secondary users) to dynamically access the sensed spectrum hole and exploit the unused channels while reducing the collision probabilities in mobile networks. The channel selection strategy is performed in distributed manner and lies in two cases. In the case of only single idle channel is available, unlicensed user is assisted to access spectrum by using a self-learning-based MAC protocol with memory function. Otherwise, in the case of accessing multi-channels simultaneously, unlicensed users use Q-learning algorithm to select the idle channels for data transmission. In this case, each cognitive user migrating to another mesh cell selects the channel with the most Q-value based on the most successful access witch is obtained through ACK packets after each successful data transmission at every slot. In this algorithm, the space overhead is related to the number of channels and users which means that when these latter increase, the space overhead will increase and greater size storage will be needed. The time overhead refers to the waiting time of unlicensed users to access channel. A higher number of users increases the time overhead to find the best solution and decelerates the convergence of the algorithm. For the communications overhead, the high frequency of ACK packets may generate interferences and collisions, which degrades the network performances. The performance evaluation of the proposed protocol shows that it outperforms other existing schemes in terms of channel usage rate and conflict probability. However, this protocol does not support query prioritization between unlicensed users and the associated energy consumption affected by the learning is not considered. In addition, the complexity of this algorithm is relatively high in large-scale networks with high traffic load and its efficiency depends on the number of channels and users which have a significant impact on the speed of convergence.

QL-MAC protocol has been proposed in [12] to optimize the radio sleeping and active periods of network's nodes by introducing an intelligent and predictive radio scheduling strategy through RL techniques. The proposed approach allows each sensor node to independently predict the most suitable sleep/active scheduling policy during each time slot by learning the traffic condition of the node itself and the properties of its neighborhood. The space and time overhead are related to the number of slots within a frame which can be considered as very low. For the communication overhead, packets are transmitted in broadcast mode and propagated hop by hop to reach the sink which increases the amount overhead packets, decreases the energy efficiency and generates interference in dense networks. Also, sensor nodes exchange the value of expected received packets used to calculate the amount of packets a specific neighboring node has sent to another during a slot time including packets not successfully received due to collisions. The proposed protocol outperforms CSMA-CA MAC protocol both in small and large scale scenarios as well as in real and simulated environments by optimizing the Packet Delivery Ratio (PDR) and reducing the energy consumption of each sensor node. However, the high communication overhead in dense networks may decrease the algorithm efficiency and deteriorate its performances.

In [13], Sharma et al. addressed the coverage redundancy problem by proposing the coverage-connectivity maintenance based on Nash Q-learning algorithm. Its aim is to maximize the coverage rate, minimize the total energy consumption, and maintain network connectivity provided by sensor nodes in the WSN. In the suggested protocol, sensor nodes learn independently about each other and then each node performs its best action $\{$active, hibernate, sleep, adjust the sensing range$\}$ so that the total number of activated nodes in each scheduling round becomes minimum. The proposed protocol consists of two phases; learning phase for coverage maintenance where the sensing range of the sensor nodes is customized after learning the best action to remove coverage redundancy; and learning phase for connectivity maintenance where the optimal subset of active nodes is selected. The authors classified rewards into local and global rewards. The space overhead is related to the number of sensor nodes because each sensor node needs to store Q-values of all sensors nodes in Q-table. The convergence time of this algorithm increases with the rise of the number of sensor nodes. The communication overhead of this algorithm is low because sensor nodes don't need to exchange notification messages. Performance evaluation of this protocol in both small and large scale WSN has showed that it helps to reduce the energy consumption of sensor nodes by customizing their sensing range and minimizing the number of active sensor nodes. However, latency between active sensor nodes is not considered in this protocol. In addition, energy parameter needs to be introduced because sensor nodes nearer to the base station may require more energy to route the data from distant nodes.

X. Fu et al. [14] introduced a Q-learning-based scheduling algorithm for Bluetooth Low Energy (BLE) technology to satisfy the requirement of energy efficiency and QoS by optimizing the length of the Connection Interval (CI) and the number of packets to transmit during each CI in the connection mode. Specifically, the master, which is the agent, intelligently selects the most suitable CI length and the number of packets to transmit per CI using the information provided by the slave including the number of packets waiting for transmission and their remaining delays. Reward function is designed to indicate whether or not selected action meets the delay requirement and helps to minimize packet losses. For the space overhead, authors used reduced action/states space for experimentation but a larger number of packets waiting for transmission at the slave increases the size of the state space and provides different results. In this work, authors fixed the maximum number of waiting packets in the queue to 5 and only 32 different CI out of 3195 are used. Hence, this experimentation does not reflect the reality. The time overhead is very high in

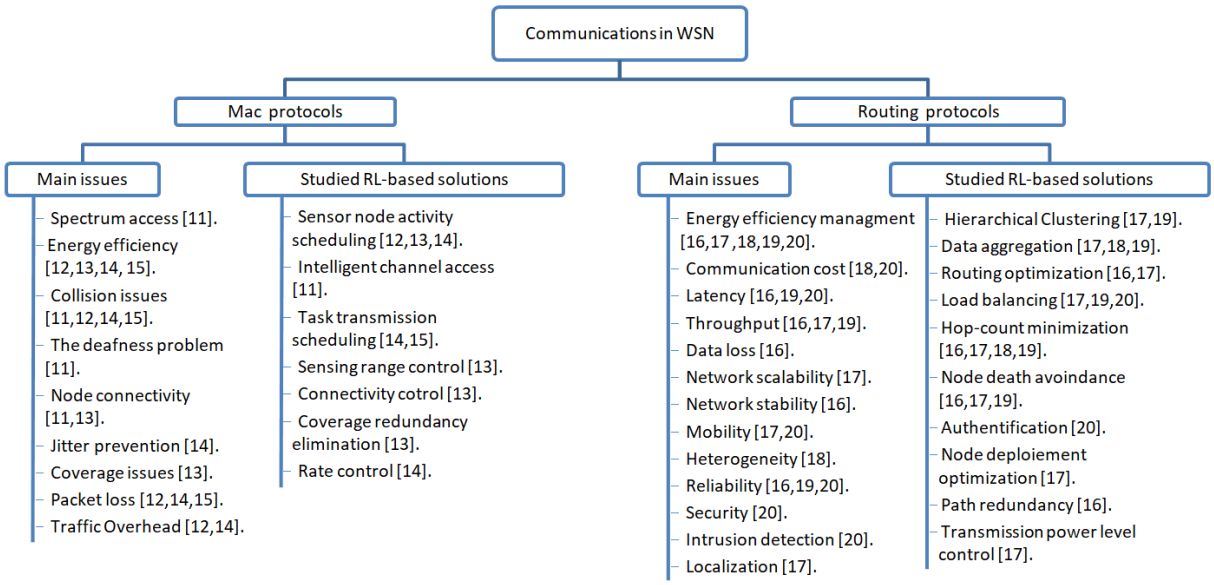

Fig. 2. Taxonomy of WSN communication issues and their existing RL-based solutions.

high-traffic applications where the number of waiting packets increases exponentially. For the communication overhead, the master receives notifications from the slave about the latency of the packets waiting for transmission on each CI and selects the optimal transmission parameters of the slave. This behaviour increases the communication overhead in high-traffic applications. The results obtained show that the proposed scheme increases the network lifetime compared to other scheduling schemes while meeting delay requirement. The proposed work may be well suited for real-time application because it takes into consideration the latency of the waiting packets to minimize packet losses. However, this proposal does not fit for high-traffic applications because of its sensitivity to the increase in the traffic to transmit which deteriorates the network performances.

In [15], Zhang et al. combined the multi-layer stacked auto-encoding network model with Q-learning and proposed a solution for the problem of concurrent data transmission scheduling in industrial WSNs. They consider the influence of many factors on data transmission process, such as interference between nodes, remaining deadline of packets, and remaining hop count to the destination while determining the data urgency and minimizing the number of lost packets. For the exploration strategy, the Metropolis criterion in the simulated annealing algorithm is used based on the ϵ-$greddy$ strategy to solve the problem of too fast convergence. Rewards are assigned to the actions which led to less packet loss. The space overhead is reduced significantly using the stacked auto-encoder model in the Q-learning phase. However, the agent requires to calculate and store the matrix of interference between nodes which is very costly in term of storage space. For the time overhead, combining deep learning with Q-learning improves significantly the learning speed in systems with large state/action spaces. For the communication overhead, nodes require to communicate at least the remaining cutoff time of their packets to calculate the urgency data on the network at each time slot. Also, they need to receive the network topology from the base station after every topology update, so, the DQN needs to be re-trained for several times especially in high dynamic networks. The experimental results show that the proposed scheme significantly improves the average number of lost packets compared with existing solutions. However, only 25 nodes were used for experimentation which does not refer to a large-scale network.

B. RL-based Routing protocols

Künzel et al. in [16] presented Q-Learning Reliable Routing with a Weighting Agent (QLRR-WA) algorithm for WirelessHART protocols. In this work, the Network Manager (NM), which is the learning agent, builds an uplink routing graph using the information received from sensor nodes (the Received Signal Level RSL, the number of hops from the gateway, and the residual energy). Q-learning algorithm is adopted to iteratively adjust the weight values of a cost equation used to choose for each node at least two best neighbors to forward messages toward the gateway. Rewards are received when the agent decreases the average network latency or increases the expected network lifetime. For the space overhead, the states refer to a set of weights (w_h, w_p, w_s) which have a significant impact on the accuracy and the convergence time of the algorithm by reducing the distance in hops from nodes to the gateway (w_h), avoiding the use of battery-powered nodes as successors (w_p), and reducing the probability of packet transmission failures by choosing nodes with greater RSL as successors (w_s). The time overhead is influenced by the changes of network topology which require re-configurations. Also, the state/action space which requires

several time to explore the environment in each state and many iterations to converge. For the communication overhead, the NM needs to collect information messages from sensor nodes to run the algorithm which increases the communication overhead especially in dense networks. Performance evaluation shows that QLRR-WA algorithm outperforms other existing solution in term of low latency and reliability. However, re-configurations of the network topology in high-dynamic networks require many iterations to converge and increase the waiting time of nodes for the new configuration to transmit their messages which decelerate the network performance.

In [17], Cho and Lee proposed Q-LEACH protocol based on reinforcement learning and F-LEACH protocol based on the Fuzzifier method within both static and dynamic topology consideration to improve the clustering phase of the well-known LEACH protocol. In Q-LEACH algorithm, the total area is divided into several units and Q-table is calculated for each unit dimension. The current state consists of the Signal Interference Noise Rate (SINR) between the sensor node and the CH node, the action is the selection of a transmit power, and the reward function considers the SINR and the agent is rewarded when node arrives to the best CH. For the space overhead, a large subset of SINR values increases the accuracy of the algorithm but may lead to a high space overhead. For the communication overhead, Q-LEACH protocol does not need to exchange information messages for clustering. The results of simulation indicate that Q-LEACH improves the network lifetime by minimizing the total dissipated energy and achieves the best throughput compared to the LEACH protocol and other existing routing schemes. However, this protocol does not have an effective selection method for cluster head nodes which can provide better performance. Also, the packet losses caused by nodes mobility were not examined.

In [18], the authors proposed RL-based energy-aware routing algorithm (Q-DAEER) to dynamically find an optimal routing path that optimizes the overall energy consumption and improves the lifetime of heterogeneous WSN. It considers several types of sensors with multiple queue management and three different data aggregation models (representative, lossy compressive and lossless aggregation model). Sensor nodes can determine the best next hop node using their updated Q-values based on the reward function which considers the data aggregation level of the neighbor node, the residual energy, the communication cost, and the hop count to the sink. For the space overhead, each sensor node requires a separated Q-table to learn and a queue to store data for each sensor type of the network. The sensor type-queue overload depends on the the packet sizes, and the Q-table size depends on the number of neighbor nodes. The time overhead refers to the waiting time for data aggregation and the convergence time of Q-learning algorithm which is related to the network density. For the communication overhead, sensors periodically exchange rewards and Q-values which increases the communication overhead of the network, overloads the sensor memory and decelerates the network performances. Simulation results show that Q-DAEER protocol outperforms other existing routing algorithms in terms of energy consumption, network lifetime, average hop count, and the number of transmissions. However, the waiting time for data aggregation is not managed, which increases the latency for data delivery to the sink node.

In [19], Sathyamoorthy et al. proposed a centralized approach and blended the Q-learning technique with the K-Means algorithm to enhance clustering and node balancing in WSN. Each cluster is partitioned into 'k' partitions and a Partition Head (PH) for each partition will be chosen. The PH with the most residual energy and the closest distance to the sink will be elected as CH (Cluster Head). As well for the node balancing algorithm, Q-learning has been used for evenly distributing the sensors in each partition. The time overhead refers to the waiting time of sensor nodes for the new coordinates and the elected PH. Thus, the sink requires more learning time to find the best solution in dense networks. The performance evaluation of the Q-K-means protocol indicated that this later increases the throughput, reduces the-end-to end delay, and produces a significant improvement in network lifetime with a reliable data packet delivery ratio. However, the learning agent needs to be aware about nodes positions which is not possible in most Industrial networks.

In [20], Device to Device (D2D) multi-criteria reinforcement learning algorithm was proposed for smart cities to improve the performance of packet delivery ratio, disturbances, latency, and energy consumption using mobile IoT devices that support authentication to ensure security. After establishing secure sessions for direct communication, each device exploits RL technique to choose the most optimal route for forwarding the data towards the sink node using the obtained information from its neighbors such as residual energy, the speed, the radio coverage, and the link cost information to calculate the route rank. A reward is assigned to the neighbor with the highest route rank value which will be selected as next hop device. High communication overhead is generated in this algorithm due to the excessive exchange of control packets between devices which requires also more convergence time to find optimal routes. The suggested D2D multi-criteria algorithm is tested and compared with other existing solutions (CTEER and QL-MAC algorithms) to demonstrate that the proposed algorithm leads to lightweight complexity by balancing the resources consumption among the mobile nodes, reducing the cost of communication by finding optimal routes, and identifying the malicious nodes which generate excessive false traffic.

C. Summary analysis

A reinforcement learning-based communication protocol might perform poorly when the agents and the reward functions are badly setting up, which leads to high space, time, or communication overhead and even convergence insufficiency. These later decelerate the learning process and negatively affect the basic requirements needed to develop industrial systems, such as computational complexity, latency, power consumption, reliability, and real-time communications. In table I, we summarize and compare the discussed communica-

TABLE I
SUMMARY ANALYSIS OF RL-BASED COMMUNICATION PROTOCOLS IN WSN

	RL-Based MAC Protocols					RL-Based Routing Protocols				
	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
Low-computational complexity			✓		✓			✓		✓
Low-latency				✓		✓	✓		✓	✓
Low-energy consumption		✓	✓	✓		✓	✓	✓	✓	✓
Reliability	✓	✓	✓		✓	✓		✓	✓	✓
Real-time communications				✓	✓					
Heterogeneity	✓			✓	✓	✓	✓	✓		✓
Mobility	✓						✓			✓
Scalability		✓	✓				✓	✓	✓	✓
Security										✓

tion schemes assisted by RL according to several requirements of WSNs and IIoT-based systems. Based on this table, the learning model's complexity, real-time communications, and security should be better considered. Further enhancements on these protocols are expected to meet the specific needs of the IIoT-based applications while achieving convergence and reducing overhead and complexity.

IV. CONCLUSION

This paper reviewed some recent RL-based communication protocols including MAC, scheduling, clustering, data aggregation, and routing protocols. RL has the capability to provide high performance to IIoT based applications with involved WSNs while mitigating their related issues. However, the computational complexity of RL-based protocols has an important impact on the application it is used for. Also, the seek for balancing the trade-off between cost, delay, energy efficiency, accuracy, and security remains a challenging issue. In our future work, we will extend our study to cover more Intelligence Artificial techniques and address other IIoT challenges.

REFERENCES

[1] P. K. Malik, R. Sharma, R. Singh, A. Gehlot, S. C. Satapathy, W.S. Alnumay, D. Pelusi, U. Ghosh, and J. Nayak, "Industrial Internet of Things and its Applications in Industry 4.0: State of The Art," Computer Communications, vol. 166, pp. 125–139, 2021.

[2] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial Internet of Things: Challenges, Opportunities, and Directions," IEEE transactions on industrial informatics, vol. 14, no. 11, pp. 4724–4734, 2018.

[3] M.Y. Aalsalem, W. Z. Khan, W. Gharibi, M. K. Khan, and Q. Arshad, "Wireless Sensor Networks in oil and gas industry: Recent advances, taxonomy, requirements, and open challenges," Journal of network and computer applications, vol. 113, pp. 87–97, 2018.

[4] S. Messaoud, S. Bouaafia, A. Bradai, M. A. Hajjaji, A. Mtibaa, and M. Atri, "Network slicing for industrial IoT and industrial wireless sensor network: Deep federated learning approach and its implementation challenges," in Emerging Trends in Wireless Sensor Networks, IntechOpen, 2022.

[5] S. Yessad, L. Bouallouche-Medjkoune, and D. Aïssani, "A Cross-Layer Routing Protocol for Balancing Energy Consumption in Wireless Sensor Networks," Wireless Pers Commun vol. 81, pp. 1303–1320, 2015.

[6] W. Z. Khan, M. Rehman, H. M. Zangoti, M. K. Afzal, N. Armi, and K. Salah, "Industrial Internet of Things: Recent advances, enabling technologies and open challenges," Computers & Electrical Engineering, vol. 81, Art. no. 106522, 2020.

[7] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," MIT press Cambridge, 1998.

[8] T. Kegyes, Z. Süle, and J. Abonyi, "The applicability of reinforcement learning methods in the development of industry 4.0 applications," Complexity, vol. 2021, pp. 1–30, 2021.

[9] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-Learning Algorithms: A Comprehensive Classification and Applications," IEEE Access, vol. 7, pp. 133 653–133 667, 2019.

[10] Z. Mammeri, "Reinforcement Learning Based Routing in Networks: Review and Classification of Approaches," IEEE Access, vol. 7, pp. 55 916–55 950, 2019.

[11] F. Li, K.-Y. Lam, Z. Sheng, X. Zhang, K. Zhao, and L. Wang, "Q-Learning-Based Dynamic Spectrum Access in Cognitive Industrial Internet of Things," Mobile Networks and Applications, vol. 23, no. 6, pp. 1636–1644, 2018.

[12] C. Savaglio, P. Pace, G. Aloi, A. Liotta, and G. Fortino, "Lightweight Reinforcement Learning for Energy Efficient Communications in Wireless Sensor Networks," IEEE Access, vol. 7, pp. 29 355–29 364, 2019.

[13] A. Sharma and S. Chauhan, "A distributed reinforcement learning based sensor node scheduling algorithm for coverage and connectivity maintenance in wireless sensor network," Wireless Networks, vol. 26, no. 6, pp. 4411–4429, 2020.

[14] X. Fu, L. Lopez-Estrada, and J. G. Kim, "A Q-Learning-Based Approach for Enhancing Energy Efficiency of Bluetooth Low Energy," IEEE Access, vol. 9, pp. 21 286–21 295, 2021.

[15] A. Zhang, M. Sun, J. Wang, Z. Li, Y. Cheng, and C. Wang, "Real-Time Data Transmission Scheduling Algorithm for Wireless Sensor Networks Based on Deep Q-Learning," Electronics, vol. 11, no. 12, Art. no. 1877, 2022.

[16] G. Künzel, L. S. Indrusiak, and C. E. Pereira, "Latency and Lifetime Enhancements in Industrial Wireless Sensor Networks: A Q-Learning Approach for Graph Routing," IEEE Transactions on Industrial Informatics, vol. 16, no. 8, pp. 5617–5625, 2019.

[17] J. H. Cho and H. Lee, "Dynamic Topology Model of Q-Learning LEACH Using Disposable Sensors in Autonomous Things Environment," Applied Sciences, vol. 10, no. 24, Art. no. 9037, 2020.

[18] W. K. Yun and S. J. Yoo, "Q-Learning-Based Data-Aggregation-Aware Energy-Efficient Routing Protocol for Wireless Sensor Networks," IEEE Access, vol. 9, pp. 10 737–10 750, 2021.

[19] M. Sathyamoorthy, S. Kuppusamy, R. K. Dhanaraj, and V. Ravi, "Improved K-Means Based Q Learning Algorithm for Optimal Clustering and Node Balancing in WSN," Wireless Personal Communications, vol. 122, no. 3, pp. 2745–2766, 2022.

[20] K. Haseeb, A. Rehman, T. Saba, S. A. Bahaj, and J. Lloret, "Device-to-Device (D2D) Multi-Criteria Learning Algorithm Using Secured Sensors," Sensors, vol. 22, no. 6, Art. no. 2115, 2022.