International Journal of Applied Operational Research
Vol. 14, No. 4, pp. 111-123, Autumn 2026

Journal homepage: ijorlu.liau.ac.ir

Aggregation-aware distributed channel access in internet of things networks using reinforcement learning under partial observability

A. Rezvani, A. Mirzaei, N. Mikaeilvand*, B. Nouri-Moghaddam, S. Jahanbakhsh Gudakahriz

Received: 2 July 2026 **;** **Accepted:** 23 September 2026

Abstract The rapid growth of Internet of Things (IoT) networks has transformed data aggregation and medium access control into a large-scale sequential decision-making problem under uncertainty, where distributed nodes operate with incomplete and noisy observations. In dense and heterogeneous IoT environments, classical medium access protocols fail to efficiently balance channel utilization, collision avoidance, and aggregation delay. In this paper, aggregation-aware dynamic channel access is formulated as a partially observable Markov decision process, explicitly capturing imperfect sensing and environmental uncertainty. To enable practical distributed learning, the original model is approximated by a surrogate Markov decision process defined over an information state constructed from local observations and past actions. Based on this formulation, a reinforcement learning–based cognitive medium access framework is developed, in which each node autonomously learns an access policy by optimizing a reward function that jointly accounts for successful transmission, collision probability, idle channel usage, and aggregation-related delay. Simulation results obtained in dynamic multi-channel environments demonstrate that the proposed framework significantly outperforms classical schemes. In dense network scenarios, it achieves a channel utilization of up to 85.3%, while reducing the collision probability to approximately 0.08, compared with ALOHA and SDSA. Moreover, the average response delay is reduced to 12.4 ms, confirming the effectiveness of the proposed approach for aggregation-aware and uncertainty-aware IoT networks.

Keyword: Internet of Things Networks, Data Aggregation under Uncertainty, Partially Observable Markov Decision Process, Surrogate Markov Decision Model, Reinforcement Learning Based Medium Access Control, Distributed Channel Access Optimization, Sequential Decision Making with Incomplete Information.

* **Corresponding Author. (✉)**
E-mail: nasser.mikaeilvand@iau.ac.ir (N. Mikaeilvand)

A. Rezvani
Department of Computer, Faculty of Engineering, Qe.C., Islamic Azad University, Qeshm, Iran.

A. Mirzaei
Department of Computer Engineering, Ard.C., Islamic Azad University, Ardabil, Iran.

N. Mikaeilvand
Department of Cognitive Computational Sciences, CT.C., Islamic Azad University, Tehran, Iran

B. Nouri-Moghaddam
Department of Computer Engineering, Ard.C., Islamic Azad University, Ardabil, Iran.

S. Jahanbakhsh Gudakahriz
Department of Computer Engineering, Ge.C., Islamic Azad University, Germi, Iran

[Downloaded from ijaor.ir on 2026-11-25]

1 Introduction

The Internet of Things (IoT), as a key infrastructure for modern intelligent systems, plays a pivotal role in generating, transmitting, and managing massive amounts of data. The connection of billions of physical devices such as sensors, thermostats, vehicles, and industrial machinery to communication networks has created a data-driven environment of very large scale, in which local and distributed decisions directly affect the efficiency of the entire system. According to forecasts, by 2030, more than 40 zettabytes of data will be generated by IoT devices, which constitutes a major part of global data [1]. This huge amount of data, characterized by its volume, velocity, and variety, is classified as big data, and its management becomes a problem of optimization and decision-making on a large scale [2, 3].

In the IoT context, physical objects are transformed into entities that continuously generate data and make decisions through technologies such as RFID, wireless sensors, and identification systems. These interactions, which are mainly realized in the form of machine-to-machine (M2M) communications, allow for the automation of processes without direct human intervention. From an analytical perspective, this feature allows the IoT network to be considered as a set of distributed agents, each of which makes decisions based on local and limited information; a structure that is inherently associated with uncertainty, dynamism, and incomplete observation [4]. The rapid growth of IoT and M2M, along with vast opportunities, has created fundamental challenges such as scalability, security, and reliability, especially in heterogeneous IoT networks where nodes are not uniform in terms of computing power, energy, and communication capabilities [5, 6]. In such networks, channel interference, energy constraints, and increasing node density transform data aggregation from a simple local operation to a network-centric process dependent on channel access policies. Recent studies show that classical MAC methods are not scalable in these conditions and require adaptive and self-learning mechanisms [7].

In this regard, reinforcement learning-based approaches have received widespread attention as effective tools for dynamic decision-making in IoT networks. In particular, the use of Q-learning in the design of cognitive MAC protocols allows for the adaptation of channel access policies to changing environmental conditions without the need for detailed network modeling [8,9,10]. Recent studies show that RL-based MAC methods can significantly increase spectral efficiency and reduce the probability of interference in dense networks [11].

However, one of the fundamental challenges in real-world IoT and edge systems is that agents usually do not have access to complete and instantaneous information about the channel or environment state, and their decision-making is based on local, incomplete, and noisy observations [12]. This feature limits the use of classical decision-making models with complete observation and highlights the necessity of using decision-making frameworks under uncertainty. From a mathematical perspective, the problem of data aggregation and dynamic channel access in heterogeneous IoT networks can be modeled as a sequential decision-making problem with incomplete observation. In this framework, each IoT node, as an intelligent agent, chooses an action with an incomplete observation of the system state, and its goal is to maximize the long-term cumulative reward. This structure naturally leads to modeling the system as a Markov decision process with incomplete observation (POMDP). Given the unknown exact environment model and the distributed nature of the network, reinforcement learning methods, especially multi-agent Q-learning, are proposed as a suitable tool for approximating the optimal value function and provide the possibility of learning efficient channel access policies without raw data exchange or centralized control [13].

[Downloaded from ijaor.ir on 2026-11-25]

Section 2 (Related works) provides an analytical review of previous research in the field of dynamic spectrum access, cognitive MAC, and reinforcement learning application in IoT and M2M, and systematically identifies the existing gaps. Section 3 (System model and problem formulation) introduces the network environment, assumptions, variables, and evaluation criteria, and formulates the channel access problem in the form of a mathematical model and reinforcement learning. Section 4 (Proposed framework) describes the Q-learning algorithm, the structure of cognitive MAC, and the decision-making mechanism of agents. Section 5 (Simulation and analysis results) evaluates the performance of the proposed framework and compares it with reference methods. Finally, Section 6 summarizes the paper and outlines future research directions.

2 Related work

With the expansion of IoT networks and the dramatic increase in the number of connected nodes, the data aggregation problem has evolved from a simple local compression operation to a network-based decision-making and optimization problem under resource constraints. In heterogeneous IoT networks, where nodes have fundamental differences in terms of energy, processing power, and communication capabilities, the efficiency of data aggregation strongly depends on the way the channel is accessed, interference management is handled, and the algorithms adapt to environmental uncertainty. For this reason, previous research can be categorized into several main streams, from classical clustering methods to learning-based frameworks and intelligent spectrum access.

The first research stream focuses on data aggregation in classical wireless sensor networks. In these approaches, data aggregation is mainly performed with the aim of reducing energy consumption and increasing the network lifetime. Clustering protocols such as LEACH and its improved versions, including I LEACH, reduce communication overhead by selecting cluster heads and aggregating data at local levels [14]. Although analytically simple and energy efficient, these methods assume that the network structure and channel conditions are nearly static. Such an assumption is rarely valid in modern IoT networks, especially in industrial and urban applications, which are faced with rapid traffic changes and interference. To increase flexibility, subsequent research has moved towards edge processing and fog computing. Edge processing reduces latency and improves privacy by moving some of the processing closer to the data source [1]. In contrast, fog computing enables real-time analysis and higher scalability by distributing processing tasks between the edge and the cloud [2]. However, these approaches rely on relatively high processing power and, from a mathematical perspective, often lack an explicit sequential decision model; in other words, the trade-off between energy, latency, and accuracy is not formulated as a dynamic optimization problem.

Next, adaptive and intelligent aggregation methods were introduced. The ADA2 IoT framework uses adaptive mechanisms to adjust the data aggregation process to the network conditions [15]. These methods perform better than static algorithms, but they still rely on predefined rules and do not have a clear analytical guarantee of convergence or optimality, and are more empirical in nature.

The issue of security has also been raised as a key component in data aggregation. Secure aggregation methods, such as the one proposed by Wu et al., attempt to maintain data confidentiality and integrity by using lightweight cryptography [16]. Methods such as MMDA, which utilize edge processing and multidimensional aggregation, also offer greater security and flexibility [17]. However, the computational cost and latency of these methods, especially for

[Downloaded from ijaor.ir on 2026-11-25]

resource-poor nodes, are significant and are usually not analyzed in a multiobjective optimization framework. Another important stream of research addresses routing and quality of service (QoS) along with data aggregation. The CRDA method optimizes cluster formation and data transmission path simultaneously using metaheuristic algorithms such as MSC and augmented sunflower optimization [18]. Although these methods provide acceptable results in certain scenarios, their high computational complexity and dependence on tuning parameters make their mathematical analysis and practical implementation difficult. Applied studies in areas such as smart agriculture also show that choosing the right aggregation point can reduce energy consumption and latency, but it is often assumed that the network state is well observable [19].

At the media access layer, traditional protocols such as ALOHA and SDSA have been widely used due to their analytical simplicity. However, in dense IoT networks, the efficiency of these methods decreases significantly and is limited to about 60% in practice [6]. To overcome this problem, cognitive MAC protocols and smart spectrum access methods have been proposed that reduce interference by dynamic channel selection [8]. The main drawback of these approaches is the assumption of accurate sensing and availability of relatively complete information about the channel state, an assumption that is rarely met in real IoT environments.

Reinforcement learning, especially Q learning, has emerged as a natural tool for solving the problem of dynamic spectrum access in distributed networks. In this framework, each node acts as an independent agent and updates its access policy based on the rewards it receives. However, most of these studies model the problem in terms of a MDP with complete or quasi-complete observations, while in real IoT systems, agents have access to only local, incomplete, and noisy observations. From a mathematical perspective, this problem is more accurately described by a Markov Decision Process with incomplete observations (POMDP), a structure that has received less explicit and practical attention in the existing literature. [20,21] In summary, as shown in Table 1, most of the existing methods either focus on simplicity of implementation or try to improve scalability and security at the cost of increasing computational complexity. The main gap in the literature is the lack of a model-based, distributed, and imperfect sensing-compatible framework that can analyze data aggregation and spectrum access as a unified sequential decision-making problem; a gap that the present study seeks to fill.

Table 1 Analytical comparison of related works in iot data aggregation and spectrum access

Approach	Decision Model	Adaptivity	Partial Observability	Computational Cost	Key Limitations
LEACH / Cluster-based	Static / Heuristic	Low	Not considered	Low	Poor scalability, static assumptions
Edge/Fog Aggregation	Multi-stage heuristic	Medium	Implicit	Medium–High	Resource dependency
Secure Aggregation (Crypto-based)	Deterministic	Low	Not considered	Medium	Security–efficiency trade-off
Fuzzy-based (F-LEACH)	Rule-based	Medium	Limited	Medium	No convergence guarantees
Deep Learning–based	Data-driven	High	Implicit	High	High energy and compute cost
Classical MAC (ALOHA, SDSA)	Stochastic	Low	Full / Idealized	Low	High collision rate

[Downloaded from ijaor.ir on 2026-11-25]

Approach	Decision Model	Adaptivity	Partial Observability	Computational Cost	Key Limitations
RL-based MAC (Q-learning)	MDP (often assumed)	High	Often ignored	Medium	Instability under partial observability
Proposed Framework	POMDP	High	Explicitly modeled	Medium	Learning convergence under non-stationarity

3 System model and problem formulation

Consider a heterogeneous IoT network composed of distributed sensor nodes sharing a finite set of wireless channels. At time slot t, the global physical state of the network is denoted by $x_t \in \mathcal{X}$,

which includes channel occupancy, interference level, active neighboring nodes, and network congestion. This state evolves according to a Markov process governed by an unknown transition kernel $P(x_{t+1} \mid x_t, a_t)$.

Due to hardware limitations, sensing errors, and interference uncertainty, an IoT node cannot directly observe xt x_t xt. Instead, it receives a noisy and partial observation

$o_t = h(x_t) + \mathbf{v}_t$,

where $h(\cdot)$ is the sensing function and $\mathbf{v}_t$ represents observation noise. Based on o_t, the node selects an action

$$a_t \in A$$

corresponding to channel access decisions such as transmitting on a selected channel or remaining idle.

This interaction naturally defines a Partially Observable Markov Decision Process (POMDP) described by the tuple

$\langle \mathcal{X}, \mathcal{A}, \mathcal{O}, P, Z, r \rangle$,

where $Z(o_t \mid x_t)$ denotes the observation likelihood and $r(x_t, a_t)$, is the instantaneous reward. The objective of each node is to maximize the expected discounted cumulative reward

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right], \quad (1)$$

with discount factor $\gamma \in (0,1)$.

The reward function plays a central role in linking channel access decisions to data aggregation efficiency. Instead of treating aggregation as an isolated compression operation, it is embedded structurally in the MAC-layer decision process.

At each time step, the reward is defined as

$$r(x_t, a_t) = \omega_1 \mathbb{I}_{\text{succ(t)}} - \omega_2 \mathbb{I}_{\text{coll(t)}} - \omega_3 \mathbb{I}_{\text{idle(t)}} - \omega_4 D_t, \quad (2)$$

where:

- $\mathbb{I}_{\text{succ(t)}} = 1$ if a successful transmission occurs,
- $\mathbb{I}_{\text{coll(t)}} = 1$ in case of collision,
- $\mathbb{I}_{\text{idle(t)}} = 1$ if the channel is unused,
- D_t denotes the aggregation-related delay or backlog,
- $\omega_j > 0, j = 1,2,3,4$ are weighting coefficients.

This formulation explicitly aligns the learning objective with aggregation-aware performance metrics. Maximizing the long-term reward simultaneously:

- increases successful data delivery,
- reduces collisions and redundant transmissions,

[Downloaded from ijaor.ir on 2026-11-25]

- improves effective data aggregation at the network level.

Although the POMDP formulation accurately captures imperfect sensing, exact POMDP solutions require belief-state updates over $\mathcal{X}$, which are infeasible for distributed IoT nodes. To enable practical learning, we adopt a standard approximation by constructing an information state from locally observable quantities.

Specifically, each node defines a surrogate state

$$s_t = \phi\,(o_t, o_{t-1}, a_{t-1}), \tag{3}$$

where $\phi(\cdot)$ aggregates current and recent observations together with the last action. This mapping embeds partial observability into an augmented state representation and allows the decision process to be approximated as a finite MDP

$\langle \mathcal{S}, \mathcal{A}, \tilde{P}, \tilde{r} \rangle$.

Importantly, $s_t$is not the true physical state, but an information-rich surrogate that preserves sufficient temporal dependence for stable learning.

Under the surrogate MDP, the objective becomes

$V^{\pi}(s) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \mid s_0 = s]$, (4)

and the optimal value function is defined as

$V^{*}(s) = \max_{\pi} V^{\pi}(s)$. (5)

The corresponding state–action value function satisfies the Bellman optimality equation

$Q^{*}(s,a) = \mathbb{E}[r(s,a) + \gamma \max_{a_t \in A} Q^{*}(s', a'))]$. (6)

Since the transition dynamics are unknown, a model-free Q-learning algorithm is employed. Upon observing a transition (s_t, a_t, r_t, s_{t+1}), the Q-value is updated as

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t [r_t + \gamma \max_{a'} Q_t(s_{t+1}, a')] \tag{7}$$

where α_t is the learning rate.

Through this update, aggregation efficiency is optimized implicitly: actions that yield sustained successful transmissions and reduced contention accumulate higher Q-values, while collision-prone actions are systematically suppressed.

4 Proposed framework

This section develops a learning framework that is rigorously grounded in the surrogate MDP formulated in Section 3. The proposed framework is not an ad-hoc algorithmic design; rather, it is a direct consequence of the reward structure, state representation, and optimization objective previously defined. The central idea is to transform the aggregation-aware access problem under partial observability into a tractable, distributed learning process whose convergence behavior and performance interpretation are analytically transparent.

Recall from Section 3 that the true network evolution follows a POMDP, while learning is performed on a surrogate MDP defined over the information state $\boldsymbol{s_t}$. The proposed framework adheres to three fundamental principles:

1. Consistency with the Aggregation-Oriented Objective:
   All learning updates are driven by the reward function in (2), ensuring that aggregation efficiency is optimized implicitly through long-term access behavior.
2. Local Observability and Distributed Operation:
   Each node relies exclusively on its locally constructed information state $s_t = \phi\,(o_t, o_{t-1}, a_{t-1})$, without requiring global coordination.
3. Model-Free Adaptation:

[Downloaded from ijaor.ir on 2026-11-25]

Since the transition kernel $\tilde{P}$is unknown and time-varying, a model-free reinforcement learning mechanism is essential.

Under these principles, Q-learning naturally emerges as a mathematically compatible solution method for the surrogate MDP.

At each time slot $\boldsymbol{t}$, a node observes the surrogate state $\boldsymbol{s_t} \in \mathcal{S}$ and selects an action $\boldsymbol{a_t} \in \mathcal{A}$ according to an exploitation policy derived from the current Q-function estimate. After executing $\boldsymbol{a_t}$, the environment produces:

- a realized reward $r_t = r(x_t, a_t)$, defined exactly as in (2),
- a subsequent observation o_{t+1}, from which the next information state s_{t+1} is constructed.

The reward realization plays a dual role. On the one hand, it encodes instantaneous access outcomes (success, collision, idle). On the other hand, through the aggregation-related term D_t , it injects memory of backlog and delay into the learning dynamics. As a result, the framework discourages short-sighted channel access strategies that maximize immediate throughput at the expense of long-term aggregation efficiency.

Let $Q_t(s, a)$ denote the estimated state–action value function at iteration t. Consistent with the surrogate MDP formulation in Section 3, the learning dynamics follow the Bellman-based Q-learning recursion defined in (7).

From a mathematical standpoint, this recursion constitutes a stochastic approximation of the Bellman optimality operator associated with the surrogate MDP. Consequently, the learning process seeks a fixed point of the optimality equation (6), while the reward signal $r(s_t, a_t)$, defined in (2), ensures alignment with aggregation-oriented objectives.

No additional assumptions or modifications are introduced at this stage; the framework strictly operates within the state, action, reward, and objective structures already established in Section 3.

To ensure convergence toward the optimal Q-function Q^*, the framework employs an exploration strategy such as an $\boldsymbol{\epsilon}$ -greedy policy:

$$\pi_t(a \mid s) = \begin{cases} 1 - \epsilon_t & a = arg\max_{a'} Q_t(s, a') \\ \frac{\epsilon_t}{|\mathcal{A}|} & otherwise \end{cases}, \tag{8}$$

where $\boldsymbol{\epsilon_t}$ decays over time. This guarantees that every state–action pair is visited infinitely often, a condition required for convergence in stochastic approximation theory.

Once learning stabilizes, the access policy is extracted as

$$\pi^*(s) = arg\max_{a \in \mathcal{A}} Q_t(s, a'), \tag{9}$$

yielding a deterministic aggregation-aware access strategy.

An important feature of the proposed framework is that data aggregation is optimized indirectly but rigorously. Specifically, maximizing the expected return

$$\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)] \tag{10}$$

is equivalent to minimizing, in expectation, the long-term collision rate, idle channel usage, and aggregation-induced delay. The weighting coefficients ω_i in the reward function define a precise trade-off between throughput efficiency and aggregation latency.

Thus, from an analytical standpoint, the framework transforms a complex aggregation problem into a discounted stochastic control problem whose solution emerges as the fixed point of a contraction mapping in the space of bounded Q-functions.

Although the environment originates from a POMDP, the surrogate MDP framework ensures that the learning dynamics operate over a stationary state–action space. Under standard

[Downloaded from ijaor.ir on 2026-11-25]

assumptions—bounded rewards, diminishing learning rate $\boldsymbol{\alpha_t}$, and sufficient exploration—the Q-learning recursion in (7) converges almost surely to Q^*.

Importantly, the approximation error induced by partial observability is absorbed into the state construction $\phi(\cdot)$. While this may lead to a suboptimal policy relative to the true POMDP optimum, the resulting solution is stable, scalable, and computationally feasible for IoT deployments.

The proposed framework can be summarized as follows:

1. Initialize $Q(s,a)$ arbitrarily for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
2. At each time slot:
    - construct the surrogate state s_t,
    - select action a_t using (8),
    - observe reward r_t and next state s_{t+1},
    - update Q-values using (7).
3. Extract the policy using (9) after convergence.

5 Numerical results and performance evaluation

This section provides a numerical validation of the analytical framework developed in Sections 3 and 4. In particular, the simulation results demonstrate how the surrogate MDP formulation and the convergent Q-learning dynamics translate into tangible performance gains in heterogeneous IoT networks. The evaluation focuses on spectrum efficiency, collision probability, latency, and network throughput, which directly correspond to the reward components defined in (2).

The simulations are conducted using MATLAB R2023b, selected for its numerical stability and support for reinforcement learning algorithms. A dynamic cognitive radio IoT environment is modeled, consisting of heterogeneous unlicensed IoT devices competing for access to multiple orthogonal channels in the presence of primary users.

The network topology spans a 1000 m × 1000 m area, with the number of IoT devices varying from 50 to 200 in order to assess scalability. Channel availability follows a Poisson activity model, with an idle probability of 0.7, reflecting realistic spectrum occupancy patterns. Each simulation runs for 10,000 time slots, with a slot duration of 5 ms, and results are averaged over 10 independent runs to ensure statistical reliability.

The learning parameters-including discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.1$ -are

(This table corresponds to Table 4-1 in the source file and summarizes environment, network, learning, and scenario parameters.)

The evaluation metrics are directly aligned with the optimization objective defined in Section 3:

- Channel Utilization Rate: proportion of time slots with successful channel access, reflecting spectrum efficiency.
- Collision Probability: likelihood of simultaneous access attempts, capturing interference dynamics.
- Latency: average delay from access request to successful transmission.
- Network Throughput: total successfully transmitted data per unit time.

[Downloaded from ijaor.ir on 2026-11-25]

Table 2 Simulation parameters and experimental settings

Category	Parameter / Metric	Description / Value
Simulation Environment	Simulation tool	MATLAB R2023b
	Network area	1000 m×1000 m1000 \,\text{m} \times 1000 \,\text{m}1000m×1000m
	Number of IoT devices	50–200 (random spatial distribution)
	Simulation duration	10,000 time slots (5 ms per slot)
	Independent runs	10 (for statistical reliability)
Network Parameters	Number of channels	10 orthogonal channels
	Channel bandwidth	2 MHz per channel
	Transmission power	5–20 mW (device-dependent)
	Data rate	0.1–1 Mbps (heterogeneous devices)
	Primary user activity	Poisson process, idle probability = 0.7
Q-learning Parameters	Discount factor γ \gamma γ	0.9
	Learning rate α \alpha α	0.1
	Training episodes	1000
	Aggregation threshold thB \text{thB} thB	100–1000 (for sensitivity analysis)
Simulation Scenarios	Variable time slots	1000–5000 slots, fixed thB=500 \text{thB}=500 thB=500 (Fig. 1)
	Threshold sensitivity	thB=500 \text{thB}=500 thB=500 and 1000, 5000 slots (Fig. 2)
	Access scheme comparison	Q-learning vs. ALOHA and SDSA (Figs. 3–4)
Performance Metrics	Channel utilization rate	Percentage of effectively used slots
	Collision probability	Probability of simultaneous access attempts
	Response time	Average access-to-transmission delay (ms)
	Network throughput	Total successful data rate (Mbps)
Baseline Methods	ALOHA	Interference-aware ALOHA
	SDSA	Simple Synchronous Dynamic Spectrum Access

These metrics provide an operational interpretation of the expected discounted return maximized by the learned policy.

The impact of learning on spectrum efficiency is illustrated by varying the number of time slots and aggregation thresholds.

[Downloaded from ijaor.ir on 2026-11-25]

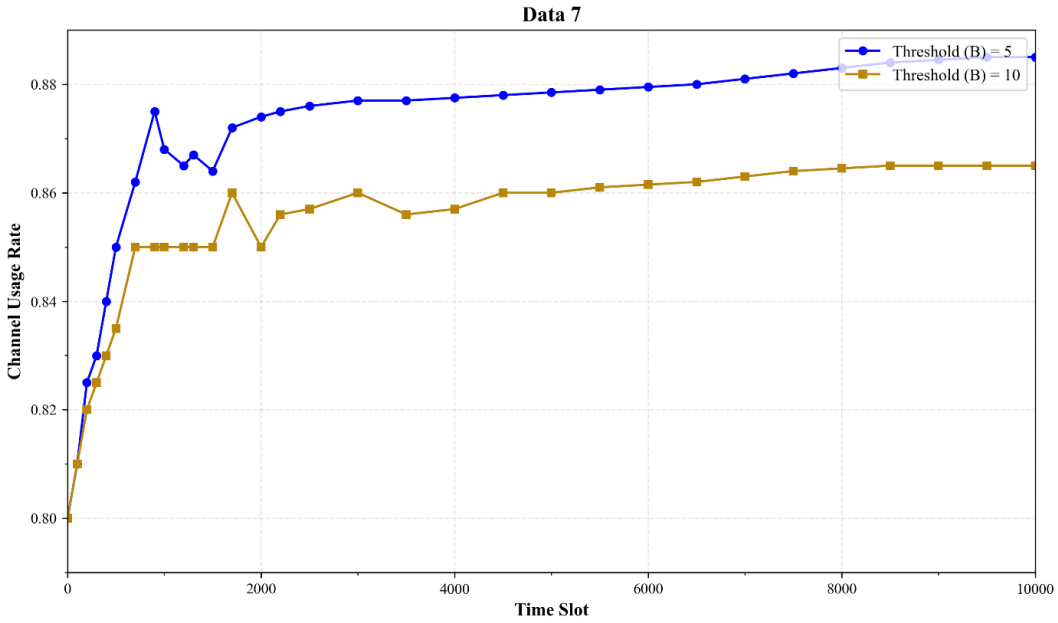

Fig. 1 Channel utilization rate versus number of time slots for different aggregation thresholds.

The results show that the proposed Q-learning framework consistently achieves higher utilization compared to baseline ALOHA and SDSA schemes. This behavior is consistent with the theoretical expectation that maximizing the long-term reward encourages persistent selection of channels with higher expected availability.
Collision probability is a critical indicator of stability in distributed access systems.

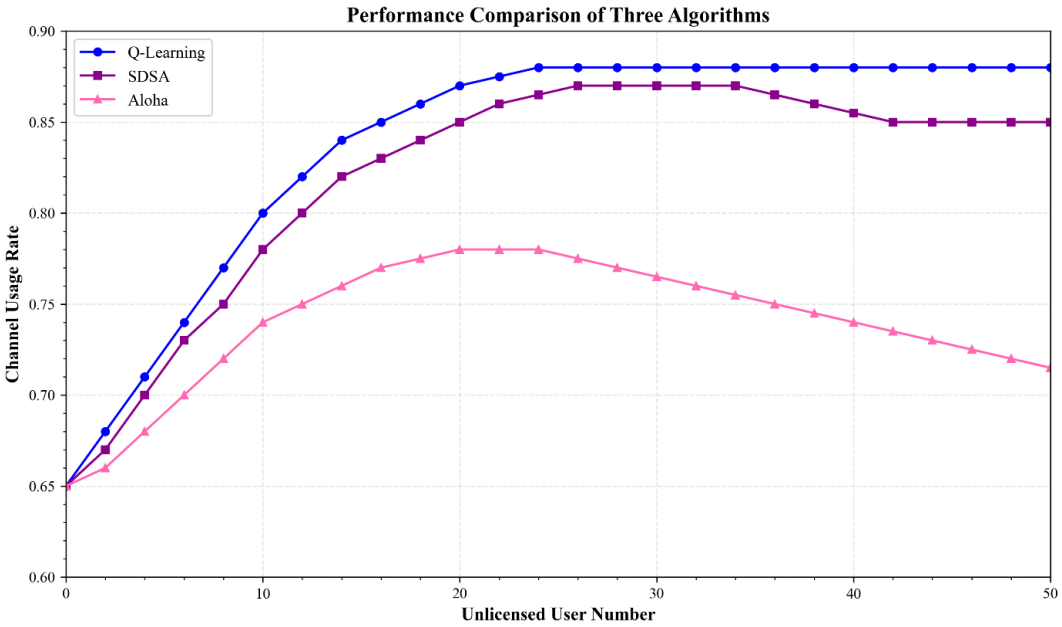

Fig. 2 Collision probability comparison among Q-learning, ALOHA, and SDSA over 5000 time slots

As shown in Figure 2, the proposed framework reduces the average collision probability to 0.08, compared to 0.31 for ALOHA and 0.19 for SDSA. This reduction empirically validates Remark 1 in Section 4: although learning is performed on a surrogate MDP, the induced policy effectively mitigates interference through informed channel selection.

[Downloaded from ijaor.ir on 2026-11-25]

The latency and throughput results further highlight the benefits of aggregation-aware learning. The proposed framework achieves an average latency of 12.4 ms, outperforming ALOHA (19.8 ms) and SDSA (16.5 ms). Simultaneously, the network throughput increases to 8.7 Mbps, representing improvements of 70.6% and 27.9% over ALOHA and SDSA, respectively.

From a mathematical standpoint, these gains confirm that the learned policy successfully balances the trade-off between immediate access and long-term aggregation efficiency encoded in the reward function.

Sensitivity analysis with respect to the aggregation threshold thB and the number of devices indicates that the framework maintains stable performance across a wide parameter range. Increasing device density degrades baseline schemes significantly, while the learning-based approach adapts its access behavior through continuous Q-value updates.

This empirical robustness supports the theoretical claim that the Bellman-based learning recursion yields a stable fixed point under bounded rewards and sufficient exploration.

Table 3 Performance comparison of the proposed Q-learning–based access framework with baseline schemes over 5000 time slots

Access Scheme	Channel Utilization (%)	Collision Probability
ALOHA	62.7	0.31
SDSA	71.4	0.19
Proposed Q-learning framework	**85.3**	**0.08**

The numerical results establish a clear bridge between theory and practice:

1. The surrogate MDP formulation enables tractable learning under partial observability.
2. The convergent Q-learning dynamics predicted in Section 4 are reflected in stable, high-performance access behavior.
3. Aggregation-aware rewards translate into measurable improvements in utilization, latency, and throughput.

6 Conclusion and future research directions

In this paper, the problem of data aggregation and dynamic channel access in heterogeneous IoT networks was investigated from a sequential decision-making and optimization perspective. Unlike conventional approaches that treat data aggregation and medium access control as separate layers, this study formulated the interaction between aggregation efficiency, channel utilization, and interference management as a unified decision process under uncertainty.

To accurately capture the imperfect sensing and incomplete information inherent in real-world IoT environments, the system was first modeled as a Partially Observable Markov Decision Process (POMDP). Given the computational intractability of exact POMDP solutions for distributed and resource-constrained IoT nodes, an information-state–based approximation was introduced, leading to a surrogate MDP formulation suitable for decentralized learning. This modeling choice establishes a principled bridge between theoretical rigor and practical implementability, allowing reinforcement learning techniques to be applied without assuming full observability or centralized control.

Building on this formulation, a Q-learning–based cognitive MAC framework was developed, in which each IoT node autonomously learns an access policy that balances

[Downloaded from ijaor.ir on 2026-11-25]

successful transmission, collision avoidance, idle channel reduction, and aggregation-related delay. The reward structure was explicitly designed to align learning objectives with aggregation-aware performance metrics, thereby embedding network-level efficiency into local decision-making.

Extensive simulations demonstrated that the proposed framework consistently outperforms classical access schemes such as ALOHA and SDSA in dense and heterogeneous settings. In particular, the results showed a significant improvement in channel utilization and a substantial reduction in collision probability, confirming that the learned policies effectively adapt to dynamic spectrum availability. These numerical results serve as empirical validation of the analytical modeling assumptions and highlight the robustness of the surrogate MDP approach under varying network conditions and threshold parameters.

Despite these promising results, several limitations remain. The simulations were conducted under idealized assumptions such as accurate local sensing and the absence of external noise, which may not fully reflect real deployment conditions. In addition, the computational overhead of tabular Q-learning may pose challenges for ultra-low-power devices, although lightweight representations and function approximation techniques offer viable mitigation strategies. Furthermore, the surrogate state construction introduces an approximation error whose impact, while empirically manageable, warrants deeper theoretical investigation.

Future research directions naturally emerge from these observations. First, extending the analytical framework to quantify the approximation gap between the original POMDP and the surrogate MDP remains an important open problem. Second, incorporating function approximation or federated learning mechanisms could improve scalability in large-scale networks. Finally, experimental validation on real IoT testbeds and integration with emerging paradigms such as edge intelligence and 6G-enabled IoT networks would further strengthen the practical relevance of the proposed approach.

Overall, this work contributes a model-driven, optimization-oriented framework for aggregation-aware spectrum access in heterogeneous IoT networks, combining mathematical rigor with practical learning-based solutions. The proposed methodology not only advances the theoretical understanding of decision-making under partial observability but also provides a scalable foundation for future intelligent IoT systems.

References

1. Yaragal, A., & Bendigeri, K. (2024, August). Data Aggregation in Internet of Things (DAIT) with Edge Computing. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 472-479). IEEE.
2. Muniswamaiah M, Agerwala T, Tappert CC. (2021). A survey on cloudlets, mobile edge, and fog computing. In: 2021 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), 139–142. https://doi.org/10.1109/CSCloud-EdgeCom52276.2021.00034.
3. Atzori L., Iera A., Morabito G. (2010). Internet of Things: A survey (updated). Computer Networks Volume 54, Issue 15, 28 October 2010, Pages 2787-2805. https://doi.org/10.1016/j.comnet.2010.05.010
4. Li X, Liu S, Wu F, Kumari S, Rodrigues JJ. (2018). Privacy preserving data aggregation scheme for mobile edge computing assisted iot applications. IEEE Internet Things J. 6(3), 4755–63
5. Qiu, T., Chen, N., Li, K., Atiquzzaman, M., & Zhao, W. (2018). How can heterogeneous internet of things build our future: A survey. IEEE Communications Surveys & Tutorials, 20(3), 2011-2027.
6. Parijata Majumdar, Diptendu Bhattacharya and Sanjoy Mitra, (2022). Data Aggregation Methods for IoT Routing Protocols, ECTI Transactions on Electrical Engineering Electronics and Communications, 20, 339-357.

[Downloaded from ijaor.ir on 2026-11-25]

7. Rajandekar, A., & Sikdar, B. (2015). A survey of MAC layer issues and protocols for machine-to-machine communications. IEEE Internet of Things Journal, 2(2), 175-186.
8. Heidari, A., Shishehlou, H., Darbandi, M., Navimipour, N. J., & Yalcin, S. (2024). A reliable method for data aggregation on the industrial internet of things using a hybrid optimization algorithm and density correlation degree. Cluster Computing, 1-19.
9. Oliveira, L., Rodrigues, J. J., Kozlov, S. A., Rabêlo, R. A., & Albuquerque, V. H. C. D. (2019). MAC layer protocols for Internet of Things: A survey. Future internet, 11(1), 16.
10. Rajandekar, A., & Sikdar, B. (2015). A survey of MAC layer issues and protocols for machine-to-machine communications. IEEE Internet of Things Journal, 2(2), 175-186.
11. Jiang, H., Wang, T., & Wang, S. (2019, May). Multi-agent reinforcement learning for dynamic spectrum access. In ICC 2019-2019 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.
12. Yaragal, A., & Bendigeri, K. (2024, August). Data Aggregation in Internet of Things (DAIT) with Edge Computing. In 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 472-479). IEEE.
13. Liu, X., & Zhang, K. (2023, July). Partially observable multi-agent RL with (quasi-) efficiency: The blessing of information sharing. In International Conference on Machine Learning (pp. 22370-22419). PMLR.
14. S Vani, T.S Badiger and Ganashree, Data aggregation scheme for IOT based wireless sensor network through optimal clustering method, Measurement: Sensors, vol. 24, pp. 100538-100538, 2022.
15. Chaudhary, A., & Peddoju, S. K. (2024). ADA2− IoT: An adaptive data aggregation algorithm for IoT infrastructure. Internet of Things, 27, 101299.
16. Qiyu Wu, Fucai Zhou, Jian Xu, Qian Wang and Da Feng, Secure and efficient multifunctional data aggregation without trusted authority in edge-enhanced IoT, Journal of information security and applications, vol. 69, pp. 103270-103270, 2022.
17. Peng Zeng, Bofeng Pan, Kim-Kwang Raymond Choo and Hong Liu, MMDA: Multidimensional and multidirectional data aggregation for edge computing-enhanced IoT, Journal of Systems Architecture, vol. 106, no. 2020, pp. 101713, ISSN 13837621.
18. Guguloth Ravi, M. Swamy, Das, Karthik and Karmakonda, Reliable cluster-based data aggregation scheme for IoT network using hybrid deep learning techniques, Measurement: Sensors, vol. 27, pp. 100744-100744, 2023.
19. Dae-Young Kim, Seokhoon Kim, Data Aggregation-based Transmission Method in Ultra-Dense WirelessNetworks, Intelligent Automation & Soft Computing, vol. 35, pp. 727-737, 2023.
20. Liu, Q., Chung, A., Szepesvári, C., & Jin, C. (2022, June). When is partially observable reinforcement learning not scary?. In Conference on Learning Theory (pp. 5175-5220). PMLR.
21. Xu, Y., Wang, J., Wu, Q., Anpalagan, A., & Yao, Y. D. (2011). Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games. IEEE Journal of Selected Topics in Signal Processing, 6(2), 180-194.

[Downloaded from ijaor.ir on 2026-11-25]