

Integrating unsupervised learning and deep reinforcement learning for anti-jamming in backscatter-aided 6G communications

Le Hoang Hiep¹, Ngo Huu Huy¹, Nguyen Ngoc Tan², Nguyen Viet Phuong^{3*}

¹Thai Nguyen University of Information and Communication Technology, Z115 Street, Quyet Thang, Thai Nguyen, Vietnam;

²VNU University of Engineering and Technology, Hoa Lac, Hanoi, Vietnam;

³Academy of Military Science and Technology, 17 Hoang Sam, Nghia Do, Hanoi, Vietnam.

*Corresponding author: vphuongvtl@gmail.com

Received 07 May 2026; Revised 23 Jun. 2026; Accepted 15 Sep. 2026; Published 15 Oct. 2026.

DOI: <https://doi.org/10.54939/1859-1043.j.mst.114.2026.22-29>

ABSTRACT

This paper proposes a leakage-free unsupervised learning-aided D3QN framework for anti-jamming control in backscatter-aided 6G IoT communications. The proposed method first maps physical-layer observations, including SINR, RSSI, and estimated interference, into semantic states using unsupervised clustering, and then employs D3QN to jointly optimize transmission mode, transmit power, and reflection coefficient. To avoid information leakage, the scaler and clustering model are fitted only on training observations and kept fixed during evaluation. Simulation results over multiple independent runs show that the proposed framework converges within about 150-180 episodes and achieves approximately 3.2 bps/Hz average throughput and 0.65 bits/J energy efficiency. Under strong jamming, it improves throughput by about 30% over D3QN without clustering and more than 50% over DQN. Robustness tests under abrupt switching, unseen jammer patterns, noisy observations, and delayed feedback further confirm its effectiveness under challenging interference conditions.

Keywords: Anti-Jamming; Deep reinforcement learning (DRL), Energy efficiency, Frequency diversity, Unsupervised clustering.

1. INTRODUCTION

Sixth-generation (6G) networks are expected to support massive Internet of Things (IoT) connectivity with stringent requirements on energy efficiency, reliability, and large-scale access [1-3]. Backscatter communication is a promising solution for ultra-low-power IoT because devices can transmit information by reflecting ambient RF signals instead of relying on power-hungry active radios [4-7]. However, the weak reflected signal and open wireless medium make backscatter-aided IoT systems highly vulnerable to intentional jamming, which can significantly degrade throughput and link reliability [8]. Machine learning-based anti-jamming has recently attracted increasing attention [9]. In particular, deep reinforcement learning (DRL) can learn adaptive transmission policies in dynamic wireless environments [10-12]. Double and dueling Q-network structures further improve learning stability and reduce value overestimation [13]. Nevertheless, most DRL-based anti-jamming methods still rely on raw physical-layer observations, such as SINR and RSSI, which may be noisy and insufficient to represent hidden jammer dynamics. Meanwhile, unsupervised learning can reveal latent interference patterns, but it is often used only for offline analysis rather than being integrated into the DRL control loop. Existing studies still have several limitations [14, 15]. First, raw-state DRL may suffer from unstable learning under non-stationary jamming. Second, clustering-based perception is rarely evaluated under a leakage-free train/evaluation protocol, which may lead to overly optimistic results if evaluation observations are used during clustering. Third, many backscatter anti-jamming schemes do not jointly optimize transmission mode, transmit power, and reflection coefficient. These limitations motivate a unified framework that combines environment perception, semantic state abstraction, and adaptive anti-jamming control. To address these issues, this paper proposes

a leakage-free unsupervised learning-aided D3QN framework for backscatter-aided 6G IoT anti-jamming. The proposed method first maps the observation vector $\mathbf{x}_t = [\gamma_t, \text{RSSI}_t, \hat{I}_t]$ into a semantic cluster label, where $\hat{I}_t$ is an estimated interference level rather than the true jammer state. The scaler and clustering model are fitted only on training observations and kept fixed during evaluation. The resulting semantic state is then used by a D3QN agent to jointly select the transmission mode, transmit power, and reflection coefficient. The main contributions are summarized as follows: 1) We propose a leakage-free perception–decision framework that integrates unsupervised clustering with D3QN-based anti-jamming control. 2) We design a semantic state abstraction mechanism that uses training-only clustering to reduce observation noise without accessing true jammer labels. 3) We develop a joint mode–power–reflection control policy for hybrid active/backscatter communication under dynamic jamming. 4) We provide clustering-quality analysis, cluster-number sensitivity, fair baseline settings, statistical reporting, and ablation studies to validate the role of each component. 5) We evaluate the proposed method under strong jamming and challenging dynamic scenarios, including abrupt switching, unseen jammer patterns, noisy observations, and delayed feedback.

2. RESEARCH METHODOLOGY

2.1. System model

We consider a backscatter-aided 6G IoT system consisting of a hybrid IoT device, a reader, and a dynamic jammer. The device operates in either active transmission or passive backscatter mode. The received SINR in the two modes is given by:

$$\gamma_{\text{act}}(t) = \frac{P_t |h_d(t)|^2}{I_j(t) + \sigma^2}, \quad (1)$$

and

$$\gamma_{\text{back}}(t) = \frac{\beta_t P_s |h_b(t)|^2}{I_j(t) + \sigma^2}, \quad (2)$$

where P_t is the transmit power, β_t is the reflection coefficient, P_s is the incident RF power, $I_j(t)$ is the jamming power, and σ^2 is the noise power. Since the two modes have different energy and interference characteristics, joint mode–power–reflection control is required. System model of the backscatter-aided 6G IoT network under dynamic jamming.

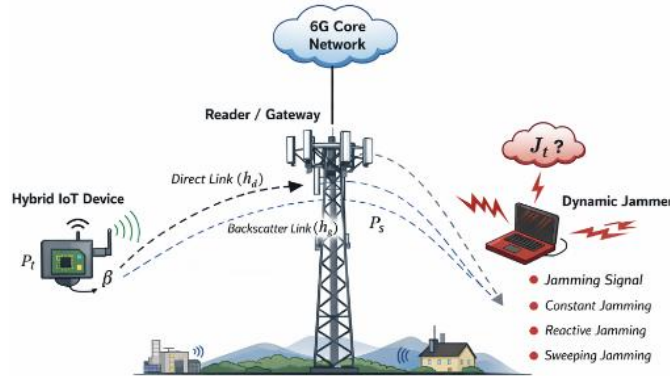


Figure 1. System model of the backscatter-aided 6G IoT network under dynamic jamming, where a hybrid IoT device performs active transmission and backscatter communication toward a reader in the presence of a time-varying jammer.

2.2. Dynamic jamming model

The jammer state J_t is hidden and evolves according to

$$\mathbb{P}(J_{t+1} = j' | J_t = j) = p_{j,j'}. \quad (3)$$

The considered jammer patterns include constant, reactive, sweeping, abrupt-switching, and unseen jamming. Since J_t is not directly observable, the agent relies only on physical-layer measurements, including SINR, RSSI, and estimated interference.

2.3. Unsupervised learning for environment perception

At time slot t , the observation vector is:

$$\mathbf{x}_t = [\gamma_t, \text{RSSI}_t, \hat{I}_t], \quad (4)$$

where $\hat{I}_t$ is receiver-estimated interference rather than the true jammer label. After normalization, K-means/GMM maps each observation to a semantic cluster:

$$c_t = \arg \min_{k \in \{1, \dots, K\}} \|\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_k\|^2. \quad (5)$$

To avoid information leakage, the scaler and clustering model are fitted only on training observations and frozen during validation and testing.

2.4. State abstraction and MDP formulation

The semantic state and joint action are defined as:

$$s_t = \{c_t, h_t, E_t\}, \quad (6)$$

and

$$a_t = \{m_t, P_t, \beta_t\}, \quad (7)$$

where c_t is the inferred semantic cluster, h_t is the channel condition, E_t is the energy state, m_t is the transmission mode, P_t is the transmit power, and β_t is the reflection coefficient.

The reward balances throughput, energy efficiency, and jamming penalty:

$$r_t = \omega_R R_t + \omega_E \xi_t - \omega_C C_t, \quad (8)$$

where

$$\xi_t = \frac{R_t}{P_t + P_c}. \quad (9)$$

The objective is to find the policy

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} [\sum_{t=0}^{\infty} \rho^t r_t]. \quad (10)$$

2.5. Proposed D3QN-based anti-jamming scheme

The D3QN agent estimates the action value as

$$Q(s, a; \theta) = V(s; \theta_V) + \left(A(s, a; \theta_A) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} A(s, a'; \theta_A) \right). \quad (11)$$

The Double Q-learning target is

$$y_t = r_t + \rho Q(s_{t+1}, \arg \max_{a'} Q(s_{t+1}, a'; \theta); \theta^-), \quad (12)$$

and the network is trained by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}[(y_t - Q(s_t, a_t; \theta))^2]. \quad (13)$$

Unlike raw-state DQN methods, the proposed scheme uses leakage-free semantic states generated by the fixed clustering model, enabling D3QN to learn robust anti-jamming decisions without true jammer labels.

2.6. Proposed UL-aided D3QN algorithm

Algorithm 1 summarizes the proposed framework. Training observations are first collected to fit the scaler and clustering model. The frozen clustering model then converts physical-layer observations into semantic states for D3QN training and evaluation, ensuring a leakage-free anti-jamming protocol.

Algorithm 1: Leakage-Free UL-Aided D3QN Anti-Jamming Framework

Input: K , replay buffer $\mathcal{D}$, discount factor ρ , learning rate α_Q , exploration rate ϵ , target update period C

Output: Anti-jamming policy π^*

- 1 Collect training observations $\mathbf{x}_t = [\gamma_t, \text{RSSI}_t, \hat{I}_t]$ from training episodes only;
- 2 Fit the scaler and K-means/GMM model with K clusters using only training observations;
- 3 Freeze the scaler and clustering model for validation and testing;
- 4 Initialize online network θ , target network $\theta^- \leftarrow \theta$, and replay buffer $\mathcal{D}$;
- 5 **for each training episode do**
- 6 Reset the environment and observe $\mathbf{x}_0$;
- 7 **for each time slot t do**
- 8 Normalize $\mathbf{x}_t$ and assign $c_t = \arg \min_k \|\tilde{\mathbf{x}}_t - \boldsymbol{\mu}_k\|^2$;
- 9 Construct semantic state $s_t = \{c_t, h_t, E_t\}$;
- 10 Select $a_t = \{m_t, P_t, \beta_t\}$ using ϵ -greedy policy;
- 11 Execute a_t , observe r_t and $\mathbf{x}_{t+1}$;
- 12 Assign c_{t+1} using the frozen clustering model and form s_{t+1} ;
- 13 Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$;
- 14 Sample a mini-batch from $\mathcal{D}$;
- 15 Compute $y_t = r_t + \rho Q_{\theta^-}(s_{t+1}, \arg \max_{a'} Q(s_{t+1}, a'; \theta))$;
- 16 Update θ by minimizing $\mathcal{L}(\theta) = \mathbb{E}[(y_t - Q(s_t, a_t; \theta))^2]$;
- 17 **if $t \bmod C = 0$ then**
- 18 $\theta^- \leftarrow \theta$;
- 19 **end if**;
- 20 Decay ϵ ;
- 21 **end for**;
- 22 **end for**;
- 23 Evaluate the learned policy using the frozen scaler and clustering model without refitting;
- 24 **return** $\pi^*(s) = \arg \max_a Q(s, a; \theta)$

3. RESULTS AND DISCUSSION**3.1. Simulation setup**

The main simulation and training parameters are summarized in Table 1. The considered setup includes the wireless channel model, jammer configurations, observation features, clustering settings, D3QN hyperparameters, compared baselines, and evaluation metrics. To ensure a leakage-free evaluation, the scaler and K-means/GMM clustering model are fitted only on observations collected from the training episodes. After training, both models are kept fixed. During evaluation, each new observation is assigned to the nearest pre-trained cluster without updating the scaler, centroids, or clustering parameters. For fair comparison, all learning-based schemes use the same channel realizations, jammer dynamics, action space, reward function, training budget, and statistical protocol whenever applicable. Each method is evaluated over at least five independent random seeds, and the results are reported as mean $\pm$ standard deviation or mean $\pm$ 95% confidence interval.

Table 1. Main simulation and training parameters.

Parameter	Setting
IoT devices/channels	$N_d = 5\text{--}20$, $N_c \in \{4, 8, 12, 16\}$
Channel model	Rayleigh fading, $\sigma^2 = -100$ dBm
Jammer model	Constant, reactive, sweeping, abrupt, and unseen patterns; 0–20 dBm
Observation	$\mathbf{x}_t = [\gamma_t, \text{RSSI}_t, \hat{I}_t]$
Clustering	K-means/GMM, $K \in \{2, 3, 4, 5, 6, 8\}$
State/action	$s_t = \{c_t, h_t, E_t\}$, $a_t = \{m_t, P_t, \beta_t\}$
D3QN settings	Replay 10^5 , batch 64, $\rho = 0.99$, $\alpha_Q = 10^{-3}$, $C = 10$
Network/optimizer	Two 128-neuron hidden layers, ReLU, Adam
Training/testing	500–1000 training episodes, 100 evaluation episodes, ≥ 5 seeds
Baselines/metrics	Random, Q-learning, DQN, raw-D3QN, D3QN-KM; reward, throughput, EE, robustness

3.2. Clustering quality and cluster-number sensitivity

Before DRL evaluation, we assess whether clustering provides useful semantic states for the D3QN agent. Different values of (K) are tested using standard clustering metrics to select a compact yet informative state representation. As shown in Table 2, a moderate (K) offers the best balance between cluster separability and learning stability; hence, this setting is used for subsequent anti-jamming evaluation.

Table 2. Clustering quality under different numbers of clusters.

K	Silhouette	DBI	CHI	Conv. Episodes
2	0.41	1.32	185.6	245
3	0.48	1.08	231.4	215
4	0.56	0.86	278.9	180
5	0.58	0.82	291.7	165
6	0.55	0.91	276.3	190
8	0.49	1.14	248.5	225

3.3. Results and analysis

A paired t-test over independent seeds is used, with $p < 0.05$ indicating statistical significance. Figure 2 shows that leakage-free clustering forms distinguishable semantic regions from $\mathbf{x}_t = [\gamma_t, \text{RSSI}_t, \hat{f}_t]$, without using true jammer labels. The clustering metrics further confirm that the selected configuration provides a meaningful state abstraction.

As shown in Figure 3, the proposed UL-aided D3QN converges within about 150–180 episodes and reaches the highest stable reward, outperforming D3QN with K-means states, raw-state D3QN, DQN, Q-learning, and Random Policy. This indicates that leakage-free semantic abstraction improves both convergence speed and learning stability.

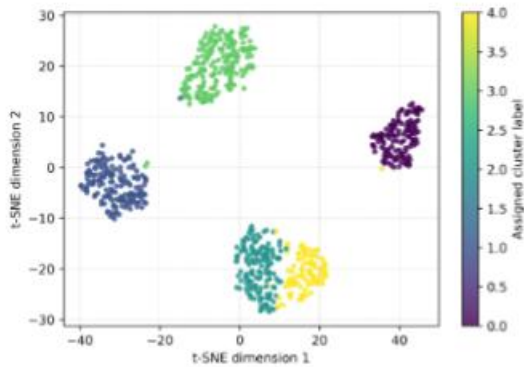


Figure 2. t-SNE visualization of leakage-free clustered jammer-related observations, where cluster labels are assigned using the pre-trained clustering model without using evaluation data for refitting.

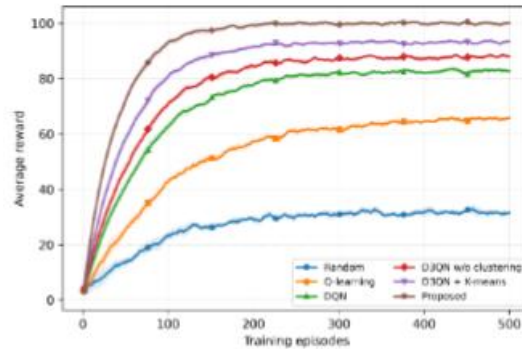


Figure 3. Average reward versus training episodes for different schemes, reported as mean $\pm$ 95% confidence interval over multiple independent runs.

Figure 4 shows that throughput decreases as jamming power increases. Nevertheless, the proposed UL-aided D3QN consistently achieves the highest throughput. At 20 dBm, it reaches about 2.7–2.8 bps/Hz, improving performance by roughly 15%, 30%, and over 50% compared with D3QN-KM, raw-state D3QN, and DQN, respectively.

Figure 5 further confirms the robustness of the proposed method under abrupt switching, unseen jammer patterns, noisy observations, delayed feedback, and mixed-stress conditions.

Although performance degrades in harder scenarios, the proposed method maintains the best throughput, achieving about 2.4 bps/Hz in the mixed-stress case.

Figure 6 shows that energy efficiency first increases with transmit power and then saturates due to higher power consumption. The proposed UL-aided D3QN achieves the best energy efficiency, about 0.45–0.48 bits/J, confirming the benefit of joint mode–power–reflection control.

Figure 7 separately compares average throughput and energy efficiency. The proposed method achieves about 3.2 bps/Hz and 0.65 bits/J, improving both metrics over DQN by approximately 55%–65%.

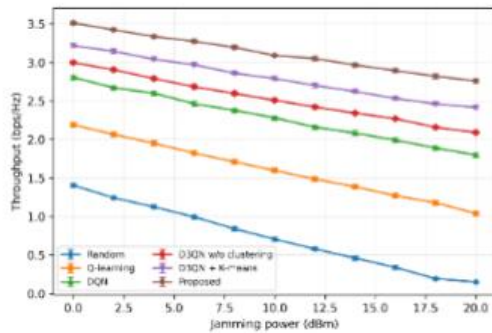


Figure 4. Throughput versus jamming power for different anti-jamming schemes, reported as mean $\pm$ 95% confidence interval over multiple independent runs.

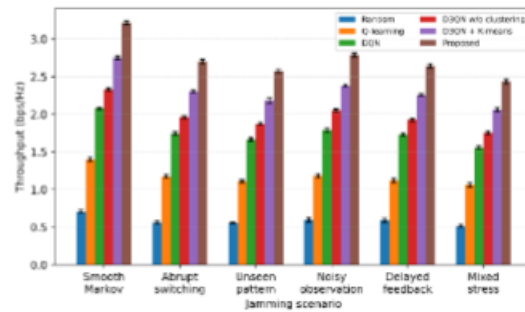


Figure 5. Robustness comparison under challenging jamming scenarios, including abrupt switching, unseen jammer patterns, noisy observations, delayed feedback, and mixed stress conditions.

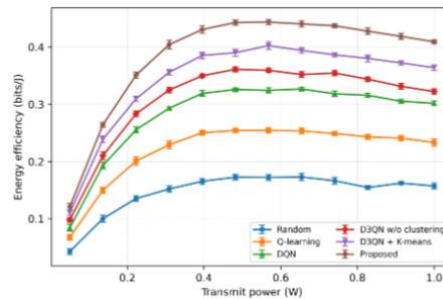


Figure 6. Energy efficiency versus transmit power for different schemes, reported in bits/Joule with mean $\pm$ 95% confidence interval.

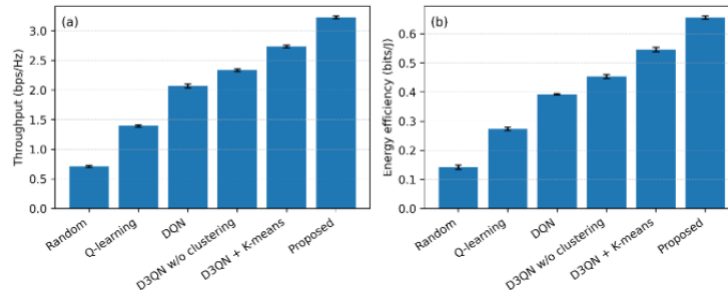


Figure 7. Comparative performance evaluation: (a) Average throughput and (b) Energy efficiency, where the two metrics are reported separately to avoid mixing different units.

3.4. Ablation study

Table 3 reports the ablation results of the proposed framework. The comparison verifies the individual effects of semantic clustering, D3QN learning, leakage-free state abstraction, and joint mode–power–reflection control. The results show that meaningful clustering is more effective than raw states or random clusters, while D3QN improves learning stability over DQN under the same semantic representation. The oracle-state variant serves only as an upper-bound reference, since true jammer labels are unavailable in practice. Overall, the proposed UL-aided D3QN achieves a favorable balance between performance and deployability without relying on ground-truth jammer information.

Table 3. Ablation study of the proposed framework.

Variant	State Type	Throughput (bps/Hz)	Energy Eff. (bits/J)	Conv. Ep.	Gain (%)
D3QN with raw states	Raw	2.45	0.48	245	–
DQN with K-means states	Semantic	2.62	0.51	225	6.9
D3QN with random clusters	Random semantic	2.30	0.44	270	–6.1
D3QN with K-means states	Semantic	2.88	0.56	195	17.6
D3QN with online clustering	Online semantic	2.78	0.54	210	13.5
D3QN with oracle jammer states	Oracle	3.35	0.67	145	36.7
Proposed w/o joint action	Semantic	2.74	0.50	205	11.8
Proposed UL-aided D3QN	Leakage-free semantic	3.22	0.65	165	31.4

4. CONCLUSIONS

This paper presented a leakage-free unsupervised learning-aided D3QN framework for anti-jamming control in backscatter-aided 6G IoT communications. The proposed method uses unsupervised clustering to transform raw physical-layer observations, including SINR, RSSI, and estimated interference, into semantic states for DRL-based decision-making. To avoid information leakage, the scaler and clustering model are fitted only on training observations and kept fixed during evaluation. The simulation results show that the proposed framework improves convergence stability, throughput, and energy efficiency compared with Random Policy, Q-learning, DQN, and D3QN without clustering. Clustering-quality analysis and cluster-number sensitivity evaluation indicate that a moderate number of clusters provides a suitable trade-off between state abstraction and learning stability. The ablation study further confirms that meaningful clustering, D3QN learning, and joint mode–power–reflection optimization all contribute to the observed performance gains. Robustness evaluation under abrupt jammer switching, unseen jammer patterns, noisy observations, and delayed feedback shows that the proposed method remains effective under challenging interference conditions. However, performance degradation is still observed in highly non-stationary scenarios. Future work will consider online clustering adaptation, multi-device coordination, and experimental validation using software-defined radio platforms.

REFERENCES

- [1]. Walid Saad, Mehdi Bennis, and Mingzhe Chen, “A vision of 6G wireless systems: Applications, trends, technologies, and open research problems”, *IEEE Network*, 34, 3, 134–142, (2020).
- [2]. Shuping Dang, Osama Amin, Basem Shihada, and Mohamed-Slim Alouini, “What should 6G be?”, *Nature Electronics*, 3, 1, 20–29, (2020).
- [3]. L. H. Hiep and H. H. Ngo, “AI-driven jamming detection and DRL-based anti-jamming for ambient backscatter communications in 6G networks”, *VNU Journal of Science: Computer Science and Communication Engineering*, 42, 2, 98–114, (2026).

- [4]. Carsten Bockelmann, Nuno Pratas, Hosein Nikopour, Kelvin Au, Tommy Svensson, Cedimir Stefanovic, Petar Popovski, and Armin Dekorsy, “Massive machine-type communications in 5G: Physical and MAC-layer solutions”, IEEE Communications Magazine, 54, 9, 59–65, (2016).
- [5]. L. H. Hiep, H. M. Viet, and T. D. Minh, “RIS-assisted deep reinforcement learning for intelligent anti-jamming communications in 6G networks”, TNU Journal of Science and Technology, 231, 6, 419–428, (2026).
- [6]. Nguyen Hong Kiem, Tran Manh Hoang, Nguyen Tuan Minh, Le Thi Thanh Huyen, “Outage probability analysis of RIS-assisted RSMA systems over Nakagami-m fading channels”, JMST, IITE, 72–82, (2025).
- [7]. L. H. Hiep, V. T. Hoang and H. H. Ngo, “RLIoT: Reinforcement Learning based network resource optimization using IoT sensor data”, International Journal of Computer Networks & Communication (IJCNC), 18, 4, 1-16, (2026). DOI: 10.5121/ijcnc.2026.18401
- [8]. Xiao Lu, Dusit Niyato, Hai Jiang, Dong In Kim, Yong Xiao, and Zhu Han, “Ambient backscatter assisted wireless powered communications”, IEEE Wireless Communications, 25, 2, 170–177, (2018).
- [9]. L. H. Hiep, “Probabilistic Analysis of Anti-Jamming Frequency Hopping in Backscatter IoT Networks”, Journal of Science & Technology on Information and Communications – PTIT, 11, 01, 19-33, (2026).
- [10]. Xin Liu, Yuhua Xu, Luliang Jia, Qihui Wu, and Alagan Anpalagan, “Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach”, IEEE Communications Letters, 22, 5, 998–1001, (2018). Journal of Science and Technology on Information security, 1.CS (27), 1-17, (2026).
- [11]. L. H. Hiep and H. H. Ngo, “D3QN-based joint power control and reflection adaptation for anti-jamming ambient backscatter communications with mobile UAV jammers”, International Journal of Robotics and Control Systems, 6, 4, (2026). DOI: 10.31763/ijrcs.v6i5.2872
- [12]. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning”, Nature, 521, 7553, 436–444, (2015).
- [13]. Nguyen Duy Thai, “Segmentation and reconstruction of wireless signals under interference using the FoT-UNET neural network”, JMST, 110, 45–54, (2026).
- [14]. Hado van Hasselt, Arthur Guez, and David Silver, “Deep reinforcement learning with double Q-learning”, Proc. AAAI Conference on Artificial Intelligence, 2094–2100, (2016).
- [15]. Liang Xiao, Yan Li, Guoan Han, Guolong Liu, and Weihua Zhuang, “PHY-layer spoofing detection with reinforcement learning in wireless networks”, IEEE Transactions on Vehicular Technology, 65, 12, 10037–10047, (2016).

TÓM TẮT

Tích hợp học không giám sát và học sâu tăng cường cho phương pháp chống gây nhiễu trong truyền thông 6G hỗ trợ backscatter

Bài báo này đề xuất một khung D3QN hỗ trợ học không giám sát theo cơ chế không rò rỉ dữ liệu nhằm điều khiển chống nhiễu trong truyền thông IoT 6G hỗ trợ backscatter. Phương pháp đề xuất trước hết ánh xạ các quan sát tầng vật lý, bao gồm SINR, RSSI và mức nhiễu ước lượng, thành các trạng thái ngữ nghĩa thông qua phân cụm không giám sát; sau đó sử dụng D3QN để đồng tối ưu chế độ truyền, công suất phát và hệ số phản xạ. Để tránh rò rỉ thông tin, bộ chuẩn hóa dữ liệu và mô hình phân cụm chỉ được huấn luyện trên các quan sát huấn luyện và được giữ cố định trong giai đoạn đánh giá. Kết quả mô phỏng trên nhiều lần chạy độc lập cho thấy khung đề xuất hội tụ sau khoảng 150-180 episode, đạt thông lượng trung bình xấp xỉ 3.2 bps/Hz và hiệu quả năng lượng 0.65 bits/J. Trong điều kiện gây nhiễu mạnh, phương pháp cải thiện thông lượng khoảng 30% so với D3QN không sử dụng phân cụm và hơn 50% so với DQN. Các kiểm thử độ bền vững dưới các kịch bản chuyển đổi jammer đột ngột, mẫu jammer chưa quan sát, quan sát có nhiễu và phản hồi trễ tiếp tục xác nhận hiệu quả của phương pháp trong các điều kiện nhiễu phức tạp.

Từ khoá: Chống gây nhiễu; Học sâu tăng cường (DRL); Hiệu quả năng lượng; Đa dạng tần số; Phân cụm không giám sát.