JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Energy-Efficient and QoS-Aware Routing in Wireless Sensor Networks Using Deep Q-Learning With Dynamic Clustering

Nguyen Phuong Thinh, Phan Thi The*, Nguyen Thanh Son
Ho Chi Minh City University of Technology and Engineering, Vietnam

*Corresponding author. *Email: thept@hcmute.edu.vn*

ARTICLE INFO

Received: 08/01/2026
Revised: 02/02/2026
Accepted: 09/02/2026
Published: 28/02/2026

KEYWORDS

Wireless Sensor Networks;
Deep Reinforcement Learning;
DQN;
QoS Routing;
Congestion Control;
Clustering.

ABSTRACT

Wireless Sensor Networks (WSNs) encounter significant challenges in balancing limited energy resources with strict Quality of Service (QoS) requirements, especially in dense deployments with dynamic traffic patterns. Traditional routing protocols rely on static heuristics that are unable to adapt to evolving network conditions such as heterogeneous energy distribution, traffic fluctuations, and topology changes. This paper presents PSR-DRL+, an adaptive routing protocol that combines Deep Q-Networks (DQN) with dynamic clustering based on node energy states and spatial distribution. The protocol utilizes a multi-objective reward function that simultaneously optimizes energy consumption, end-to-end delay, queue occupancy, and routing distance. This enables learning agents to balance network lifetime with QoS guarantees. Simulations conducted in Matlab on a scenario with 100 nodes demonstrate that PSR-DRL+ extends the time until the first node dies to 2,171 seconds, representing a 73.6% improvement over RLBEEP, Additionally, it maintains a packet delivery ratio above 95% even under heavy traffic loads. These results validate that congestion-aware deep reinforcement learning provides a viable framework for next-generation energy-constrained IoT deployments.

Doi: https://doi.org/10.54644/jte.2026.2068

1. Introduction

The rapid growth of IoT and smart environmental applications has significantly enhanced the implementation of WSNs in areas such as environmental monitoring, precision agriculture, healthcare, and intelligent transportation [1], [2]. These networks consist of numerous battery-powered nodes that autonomously organize themselves to transmit sensed data to a base station through multi-hop wireless connections, all while adhering to stringent energy limitations and specific QoS requirements for different applications.

Clustering-based routing enhances scalability and energy efficiency by designating cluster heads (CHs) to aggregate traffic from member nodes before forwarding it to the sink [3]. Classical protocols, such as LEACH and HEED, utilize probabilistic or energy-based CH election rules that promote basic energy balancing but ignore immediate traffic conditions and queue dynamics. As traffic patterns and residual energy fluctuate, these static rules can lead to overloaded CHs and congested relays, accelerating battery depletion and increasing packet loss [4].

Reinforcement Learning (RL) has emerged as a promising method for adaptive routing in WSNs, allowing nodes to learn forwarding policies through interaction with their environments. Protocols such as RLBEEP and EER-RL implement tabular Q-Learning to enhance next-hop selection, leading to significant energy savings compared to conventional routing approaches. However, tabular RL faces the curse of dimensionality as state spaces expand with the growth of the network and the incorporation of additional QoS variables, such as queue occupancy. Consequently, many RL-based protocols tend to rely on overly simplified state representations or neglect QoS metrics, focusing predominantly on energy minimization.

Deep Reinforcement Learning (DRL), especially through the use of Deep Q-Networks, approximates action-value functions within high-dimensional state spaces by leveraging deep neural networks. Recent

research on DQN-based routing in WSNs has shown that deep learning models can successfully capture the complex relationships between energy distribution, network topology, and traffic patterns. However, most DRL protocols predominantly employ single-objective reward functions that emphasize energy efficiency, which results in limited control over QoS metrics such as delay, congestion levels, and packet delivery ratios.

Despite these advances, existing DRL-based routing schemes still face challenges in jointly modeling real-time congestion dynamics and multi-objective QoS-energy tradeoffs in large-scale WSN environments.

In response to these limitations, this paper proposes PSR-DRL+, an enhanced DQN-based routing protocol with dynamic clustering that jointly considers energy efficiency and QoS performance in a unified decision framework. Unlike most existing DRL-based routing approaches that primarily optimize single-objective energy metrics or rely on static reward structures, PSR-DRL+ introduces a congestion-aware multi-objective optimization mechanism that dynamically adapts routing decisions under varying traffic conditions.

Furthermore, the proposed method integrates real-time queue state information into both the DRL state representation and reward shaping process, enabling proactive congestion avoidance rather than reactive congestion mitigation. This design allows the DQN agent to capture the joint impact of residual energy distribution, hop distance, queue dynamics, congestion status, and cluster health on routing stability and network performance.

Simulation results show that PSR-DRL+ demonstrates notable improvement in network lifetime, achieving a First Node Death (FND) time of 2171 seconds, corresponding to a 73.6% improvement compared with RLBEEP, while maintaining packet loss below 5% under high traffic load. These results indicate that the proposed approach can provides a more balanced optimization between energy efficiency and QoS compared with existing reinforcement learning-based routing strategies.

The main contributions of this paper are summarized as follows:

- A congestion-aware multi-objective DRL routing framework that jointly optimizes energy efficiency and QoS metrics under dynamic traffic conditions.
- Integration of real-time queue congestion into DRL state representation and reward design for proactive routing decision making.
- A dynamic clustering mechanism coordinated with DRL-based next-hop selection to improve energy balancing and traffic distribution.
- Comprehensive simulation-based evaluation under high traffic and congestion-prone network scenarios.

2. Related Work

2.1. Heuristic-Based Clustering Protocols

Early energy-aware routing protocols typically depend on predefined heuristics. Heinzelman et al. introduced LEACH, which uses randomized cluster-head (CH) rotation with a certain probability to avoid quick battery drainage. Although LEACH provides basic energy balance, its probabilistic method overlooks residual energy levels and geographic factors, resulting in less optimal CH placement as node energy varies [5]. Younis and Fahmy developed HEED, a protocol that chooses CHs based on residual energy and intra-cluster communication costs. Recent improvements like HEED-VCH incorporate vice CHs to boost fault tolerance. Nevertheless, all heuristic protocols fundamentally rely on static decision rules, which do not adapt to changing traffic patterns or real-time network congestion.

2.2. Tabular Reinforcement Learning Approaches

Abadi et al. introduced RLBEEP, which combines Q-Learning with dynamic scheduling and data aggregation to enhance network longevity. Simulations indicate that RLBEEP surpasses LEACH and HEED by adjusting forwarding choices based on current energy levels. However, RLBEEP's dependence on discrete Q-tables leads to scalability issues, as the size of the state space grows exponentially with the network size [6]. Mutumbo's EER-RL employs tabular reinforcement learning to

improve cluster head selection and routing by considering residual energy and hop count. Although EER-RL improves energy efficiency, it does not include queue occupancy or congestion metrics in its state model, making it susceptible to packet loss during high traffic [7]. A common limitation of tabular RL protocols is the curse of dimensionality: as networks expand or more QoS-related state variables are added, the memory needed for Q-values becomes excessively large, and convergence slows significantly.

2.3. Deep Reinforcement Learning Methods

Sezar and Rashedunnabi introduced PSR-DRL, utilizing DQN with a 9-dimensional state vector for routing choices. Their results indicate that PSR-DRL outperforms RLBEEP and EER-RL in striking a balance between energy use and delaying the first node's failure [8]. Nonetheless, PSR-DRL uses a single-objective reward focused mainly on energy efficiency, overlooking key QoS metrics such as queue congestion, end-to-end delay, and link stability. This limited focus may cause the agent to route traffic through high-energy nodes even when they are heavily loaded, leading to increased latency and higher packet drop rates during traffic surges. Current protocols fail to optimize both energy efficiency and QoS simultaneously in resource-constrained WSNs via integrated multi-objective learning [9], [10]. Heuristics lack flexibility; tabular RL faces scalability issues; and existing deep RL approaches tend to optimize either energy or QoS but not both [11]. This study addresses these issues with PSR-DRL+, which incorporates a multi-objective reward function explicitly balancing energy use, delay, queue occupancy, and routing distance [12], [13], [14].

Most recently, a routing protocol based on Deep Q-Learning combined with adaptive threshold-based clustering was proposed to enhance energy efficiency [15]. While this approach improves network longevity through dynamic cluster head selection, it primarily focuses on energy thresholds and does not explicitly integrate real-time buffer congestion metrics into the state representation, thereby limiting its responsiveness to traffic bursts.

The DQN model was implemented using MATLAB with the Deep Learning Toolbox. The toolbox was used to design, train, and optimize the neural network architecture, including experience replay and target network mechanisms. This implementation ensures reproducibility and compatibility with standard deep learning workflows.

3. System Model and Methodology

3.1. Network Architecture and Energy Model

Consider a static WSN deployed over a $L \times W$ region, comprising N battery-powered sensor nodes uniformly distributed. A single base station (sink) at a fixed position (X_s, Y_s) collects data via multi-hop wireless communication. Each node n_i is characterized by residual energy $E_{res}^{(i)}$, coordinates (x_i, y_i), and a finite-capacity FIFO transmission buffer with maximum capacity Q_{max} packets.

The energy consumption model follows the first-order radio dissipation model. When transmitting a k-bit packet from node i to node j over distance $d(i,j)$, total energy expenditure:

$$E_{Tx}(i \to j) = \begin{cases} k \cdot E_{elec} + k \cdot \epsilon_{fs} \cdot d^2, & \text{if } d < d_0 \\ k \cdot E_{elec} + k \cdot \epsilon_{mp} \cdot d^4, & \text{if } d \geq d_0 \end{cases} \quad (1)$$

where E_{elec} denotes energy per bit for electronics, ϵ_{fs} and ϵ_{mp} represent free-space and multi-path amplification coefficients, and d_0 is the crossover distance threshold. Receiving energy is $E_{Rx}(j) = k \cdot E_{elec}$.

3.2. Dynamic cluster formation

To address the limitation of energy imbalance in traditional flat routing or random clustering protocols, the PSR-DRL+ protocol proposes a cyclic dynamic clustering mechanism. Unlike fixed approaches, the Cluster Head (CH) role is not permanently assigned but is rotated at each round based on the actual energy state of the nodes.

Specifically, instead of relying on random probability, the CH election process employs a greedy selection mechanism based on residual energy. At each cluster reformation cycle, the member node with the highest residual energy within the cluster is prioritized to be selected as the new Cluster Head (CH):

$$CH_{new} = \arg \max_{j \in Clester} \{E_{res}^{(n)}\} \quad (2)$$

After cluster formation, member nodes enter sleep mode and wake only during their allocated TDMA time slots to transmit sensed data to their respective CHs, significantly reducing idle listening energy consumption.

In addition, the algorithm incorporates connectivity constraints by prioritizing nodes that are within communication range of the Base Station (BS) or neighboring CHs to ensure overall network connectivity.

This mechanism ensures that network management responsibility is continuously transferred to nodes with the best energy conditions, thereby distributing the energy load more evenly and preventing premature node death due to overload.

After the election phase, normal sensor nodes decide to join the cluster of the CH with the closest geographical distance based on the received broadcast signal strength:

$$CH_{selected} = \arg \min_{j \in SCH} \{\mathrm{d_{ij}}\} \quad (3)$$

After joining a cluster, member nodes apply a flexible sleep–wake mechanism based on action a_1, enabling maximum energy savings for regular sensor nodes.

3.3. Deep Q-Learning-based PSR-DRL+ Algorithm

The PSR-DRL+ protocol builds upon the foundation laid by PSR-DRL, inheriting the use of deep neural networks to approximate optimal routing policies without requiring global network state information. However, the original PSR-DRL suffers from a critical limitation: its single-objective reward function focuses exclusively on energy optimization while neglecting essential QoS indicators such as end-to-end delay and buffer congestion status.

To address this deficiency, PSR-DRL+ introduces a congestion-aware decision-making mechanism that integrates real-time queue occupancy monitoring into the learning process. Rather than solely seeking energy-minimal paths, the proposed algorithm employs a multi-objective reward function that balances four competing criteria: (1) residual energy availability, (2) current queue length, (3) estimated transmission delay, and (4) geographic distance to the sink. Furthermore, PSR-DRL+ imposes a substantial penalty on actions that result in packet drops due to buffer overflow, thereby training the agent to make more reliable routing decisions even under heavy network load [16], [17].

3.3.1. State Space Representation

At each decision epoch t, every sensor node observes its local environment and constructs a state vector s_t. To ensure rapid convergence and numerical stability of the deep neural network during training, all state components are normalized to the interval $[0,1]$. The state space is formalized as a 9-dimensional feature vector:

$$S_t = \{s_1, s_2, \dots, s_9\} \quad (4)$$

where each component encodes specific network conditions:

- **Normalized residual energy** (s_1): Ratio of current battery level to initial capacity, reflecting the node's remaining operational lifetime
- **Distance to cluster head** (s_2): Euclidean distance to the assigned CH, influencing intra-cluster transmission cost
- **Distance to sink** (s_3): Geographic proximity to the base station, guiding long-range forwarding decisions

- **Estimated hop count** (s_4): Number of intermediate relays predicted along the route to the sink, correlated with end-to-end latency
- **Data priority level** (s_5): Urgency metric assigned to sensed data, enabling differentiated service for time-critical events
- **Queue congestion ratio** (s_6): **This represents the most critical enhancement in PSR-DRL+.** It quantifies buffer occupancy as the fraction of filled queue slots, providing early warning of local congestion hotspots: $s_6 = \frac{Q_{len}}{Q_{max}}$
- **Sleep pressure indicator** (s_7): Tendency to transition into energy-saving sleep mode based on current load, scheduling constraints, and residual energy, minimizing baseline power consumption during idle periods
- **Cluster health metric** (s_8): Represents the stability and service capacity of the cluster, aggregating factors such as CH residual energy, aggregate cluster load, and intra-cluster link quality. This metric enables the agent to avoid clusters nearing failure or experiencing degraded performance
- **Temporal factor** (s_9): Time-related context variable capturing operational phase within the current round, facilitating time-aware decision-making

By incorporating this 9-dimensional state representation, PSR-DRL+ extends beyond traditional energy-topology features (distance, hop count) to directly embed QoS-related signals—particularly queue congestion status—into the learning substrate. This holistic state encoding empowers the DQN agent to discover routing policies that jointly optimize network lifetime and transmission quality.

3.3.2. Action Space Definition

Upon observing the state vector s_t, the DRL agent at each node selects an action a_t from a discrete action space $\mathcal{A}$. To maintain computational simplicity suitable for resource-constrained sensor nodes, the action space comprises three strategic routing decisions:

$$\mathcal{A} = \{a_0, a_1, a_2\} \tag{5}$$

- **Action a_0 — Forward to Cluster Head**: The node transmits the packet to its assigned CH, which is the default hierarchical aggregation behavior that minimizes energy expenditure for ordinary member nodes
- **Action a_1 — Direct transmission to Sink**: The node bypasses the CH and sends the packet directly to the base station. This action is beneficial when: (i) the node is geographically very close to the sink, (ii) the CH is overloaded or critically low on battery, or (iii) the link to the CH has failed
- **Action a_2 — Defer transmission (Enter sleep/buffer)**: The node postpones transmission and either enters sleep mode or retains the packet in its local queue. This action is selected when: (i) the network is severely congested and immediate transmission would likely result in packet loss, (ii) residual energy is critically low, and conservation is necessary to sustain node viability, or (iii) the sensed data exhibits negligible variation below a configured threshold [18].

Figure 1 illustrates the detailed operational workflow of PSR-DRL+. A key innovation lies in the Congestion Check preprocessing stage before data transmission. Unlike conventional protocols that blindly forward packets regardless of buffer state, each sensor node continuously monitors its queue occupancy. If Q_{len} exceeds a predefined congestion threshold (e.g., 80% of Q_{max}), the packet is proactively discarded to avoid wasting transmission energy on a link destined for buffer overflow. At the cluster head side, the DQL Routing module serves as the central decision engine, leveraging the multi-dimensional state vector to determine whether to relay the packet through intermediate hops or transmit directly to the sink, thereby balancing latency against energy consumption.

Figure 1. *Operational workflow of PSR-DRL+.*

3.4. Multi-Objective Reward Function Design

To achieve energy-efficient routing while simultaneously guaranteeing QoS in resource-constrained WSNs, the reinforcement learning agent is guided by a carefully crafted multi-objective reward function. This reward signal integrates multiple performance indicators that collectively reflect both network longevity and communication quality [19], [20]. Specifically, the immediate reward received at time t upon taking action a_t in state s_t is defined as:

$$R(s_t, a_t) = \begin{cases} \mathcal{P}_{drop}, & \text{if packet dropped} \\ \alpha \cdot \left(\frac{E_{res}^{(j)}}{E_{init}}\right) - \beta \cdot \left(\frac{Delay^{(j)}}{Delay_{max}}\right) - \gamma \cdot \left(\frac{Q_{len}^{(j)}}{Q_{max}}\right) - \delta \cdot \left(\frac{d(j, Sink)}{d_{max}}\right), & \text{if successful} \end{cases} \quad (6)$$

where $\alpha, \beta, \gamma, \delta$ are positive weighting coefficients that regulate the relative priority among objectives, subject to the normalization constraint $\alpha + \beta + \gamma + \delta = 1$.

To ensure numerical stability during training, all reward components are normalized to a common value range before aggregation [21], [22]. This normalization prevents any single objective from dominating the learning dynamics and facilitates more stable convergence of the Deep Q-Network.

Each reward component fulfills a distinct role. The residual energy term encourages the agent to prefer routes with lower transmission costs, thereby extending network lifetime. The delay penalty promotes timely data delivery, enhancing QoS. The queue-related component mitigates congestion by steering traffic away from nodes with high buffer occupancy. The distance penalty biases forwarding decisions toward geographically efficient paths [23].

These objectives are inherently conflicting: minimizing energy consumption may increase latency, while avoiding congestion might necessitate longer routes. The weighted reward formulation enables

the agent to learn a balanced routing policy that adapts flexibly to evolving network conditions. The weight coefficients $\alpha, \beta, \gamma, \delta$ are determined empirically through preliminary experiments to reflect the relative priorities between energy conservation and QoS assurance in resource-limited WSN environments.

3.5. Reward Design Justification

The reward function is designed to balance energy efficiency and QoS performance while maintaining routing stability under dynamic traffic conditions. The selection of reward components and their corresponding weights follows three main design principles [24].

First, the energy dominance principle prioritizes residual energy and energy consumption because node battery depletion directly determines network lifetime. Since WSN nodes operate under strict energy constraints, routing decisions must avoid selecting nodes with critically low residual energy or high transmission cost.

Second, the QoS balance principle incorporates delay-related metrics, packet delivery reliability, and queue occupancy to ensure stable communication performance under varying traffic loads. Queue length is particularly important because it directly reflects real-time congestion conditions and affects packet loss probability and end-to-end delay.

Third, the routing stability principle considers hop distance and cluster health indicators to prevent frequent route switching and congestion hotspots. This helps maintain stable data flow and reduces unnecessary retransmissions.

The final reward weights were determined empirically through multiple simulation trials to achieve a balanced trade-off between network lifetime extension and communication reliability.

A sensitivity analysis was conducted by varying reward weights within a small range to verify the robustness of the routing performance. The results show that PSR-DRL+ maintains stable performance trends under moderate weight variations, confirming the reliability of the selected reward configuration.

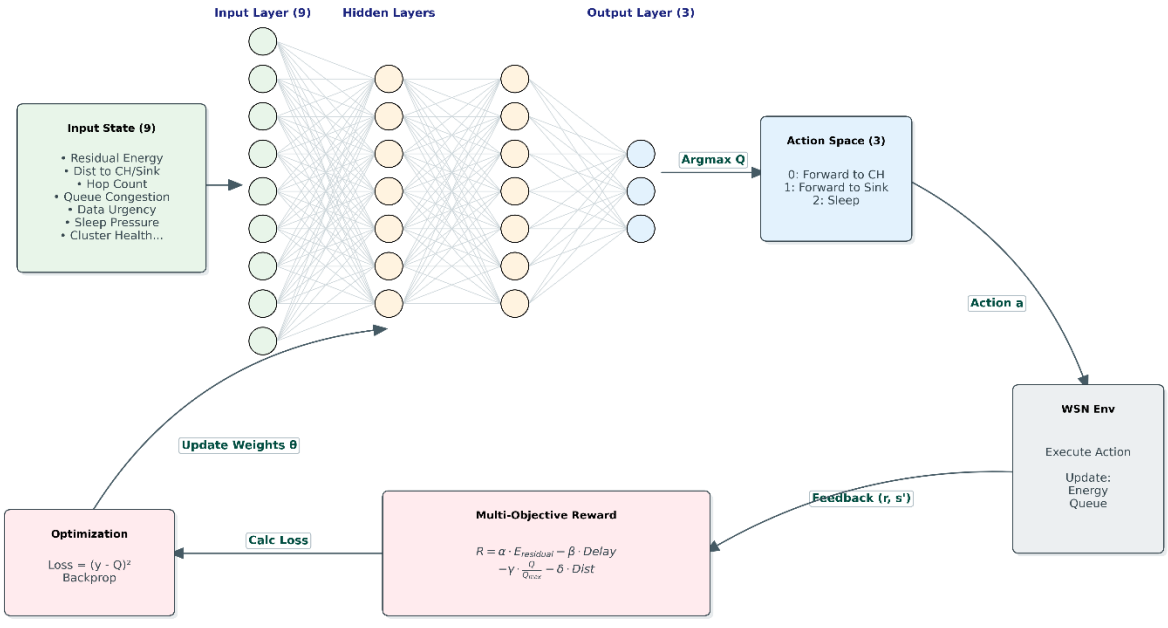

Figure 2. *Micro-level architecture of the deep neural network.*

Figure 2 presents the micro-level architecture of the deep neural network deployed on each CH. The input layer receives the normalized 9-dimensional state vector containing residual energy, distances, and critically, queue load indicators [25]. Hidden layers employ ReLU activation functions to learn complex nonlinear environmental features. A distinguishing feature of this architecture is the Multi-Objective Reward aggregation block, which synthesizes competing objectives (energy, delay, congestion) into a single scalar reward value via weighted combination $(\alpha, \beta, \gamma, \delta)$. This mechanism steers the neural

network toward routing policies that achieve holistic QoS optimization rather than local optimization of isolated parameters [26].

To update the weight parameters θ of the deep neural network to accurately approximate the optimal action-value function $Q^*(s,a)$, we employ the Smooth L1 Loss function. Unlike the Mean Squared Error (MSE) loss, Smooth L1 Loss is less sensitive to large outliers, thereby improving training stability by measuring the deviation between the Q-value predicted by the primary network and the target Q-value.

At each training step, a mini-batch of $N = 32$ experience tuples $(s_i, a_i, r_i, done_i)$ is randomly drawn from the replay buffer to reduce the correlation among sequential data. The objective loss function to be minimized is defined as follows:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} SmoothL1(y_i - Q(s_i, a_i; \theta)) \tag{7}$$

where y_i is the target label computed from the target network using the Temporal Difference (TD) method:

$$y_i = r_i + (1 - \text{done}_i) \cdot \gamma \max_{a'} Q(s_i', a'; \theta^-) \tag{8}$$

The optimization process is performed using the Adam optimizer with a tuned learning rate of $\boldsymbol{\eta = 10^{-3}}$. Adam is selected due to its ability to adapt the learning rate for each parameter and leverage momentum estimates, enabling the neural network to escape local minima more effectively than traditional Stochastic Gradient Descent (SGD), thereby ensuring stable convergence of the routing policy [27].

3.6. Congestion Control and QoS Assurance Mechanism

Packet loss due to buffer overflow at intermediate relay nodes under high traffic load constitutes one of the most severe challenges in resource-constrained WSNs, as identified by Akyildiz et al. as a critical vulnerability. Unlike traditional routing protocols such as RLBEEP and EER-RL that focus exclusively on energy optimization, PSR-DRL+ integrates a proactive congestion control mechanism operating through three protective layers:

Layer 1 — Real-Time Queue Monitoring: Each cluster head maintains a FIFO buffer with fixed capacity Q_{max}. At every time instant t, the buffer load state is monitored via the occupancy ratio, a technique essential for ensuring reliability in multi-hop communication:

$$Ratio_{queue} = \frac{Q_{current}}{Q_{max}} \tag{9}$$

Layer 2 — Integration of Congestion Status into DQN State: To enable the neural network to perceive congestion, queue information is directly embedded into the input state vector S_t. This technique extends the traditional state representation of RLBEEP by incorporating a binary congestion flag:

$$C_{flag} = \begin{cases} 1, & \text{if } Q_{current} > 0.8 \times Q_{max} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

The use of this binary indicator allows the DQN to detect abrupt state transitions, triggering hidden neurons to adjust routing strategy immediately in accordance with the principles of continuous control in Deep RL.

Layer 3 — Proactive Congestion Avoidance Strategy: This mechanism operates based on predicted Q-values. When a neighboring node j begins experiencing congestion, the expected reward for forwarding packets to that node decreases sharply due to the congestion penalty term. Consequently:

$$Q(s, a_j) \ll Q(s, a_k) \tag{11}$$

where k denotes an alternative less-congested neighbor. The DQN automatically selects action a_k, accepting a modest increase in transmission energy expenditure to ensure packet delivery without loss, thereby resolving the energy-delay trade-off.

3.7. Computational Complexity and Deployment Feasibility

The DQN training process is executed at the base station, while sensor nodes only perform lightweight data forwarding and state observation operations. Therefore, the proposed approach is suitable for resource-constrained WSN deployments.

The memory and computational overhead associated with DQN inference are minimal compared to training complexity, making the proposed scheme feasible for practical WSN scenarios.

The computational overhead of PSR-DRL+ mainly originates from the Deep Q-Network inference process. In the proposed architecture, the DQN training phase is executed at the base station, which is assumed to have sufficient computational and energy resources. Sensor nodes are only responsible for lightweight operations, including local state observation, packet forwarding, and basic cluster communication.

The DQN model size depends on the number of network layers and neurons. In typical configurations, the memory footprint required to store trained network parameters is relatively small compared to the available memory at the base station. Therefore, the proposed scheme does not impose significant memory overhead on sensor nodes [28].

From a computational perspective, routing decisions are made using forward inference of the trained DQN model, which has significantly lower complexity compared to training operations. As a result, PSR-DRL+ is suitable for practical deployment in resource-constrained WSN environments.

4. Simulation and Evaluation

4.1. Simulation Setup

Simulations use Matlab in a 200×200 meter region containing $N = 100$ nodes with $K = 40$ clusters. The sink is positioned at center (100, 100), communication range is 30m, initial energy is 300J, packet size is 4000 bits, queue capacity is 20 packets, and bandwidth is 1 Mbps.

4.1.1. Network Topology and Physical Parameters

The simulated network operates within a square monitoring region of dimensions 200×200 meters, populated by $N = 100$ sensor nodes distributed uniformly at random. The base station (sink) is positioned at the geographic center $(100,100)$ to ensure equitable data collection from all directions, thereby eliminating spatial bias in performance evaluation. The communication range of each node is set to $R_c = 30$ meters, consistent with typical IEEE 802.15.4-compliant sensor hardware.

Table 1. *Network Configuration and Physical Parameters.*

Parameter	Symbol	Value	Unit	Description
Monitoring area	$L \times W$	200×200	m^2	Deployment region
Number of nodes	N	100	nodes	Fixed network size
Target clusters	K	40	clusters	Desired CH count
Sink location	(X_s, Y_s)	(100, 100)	m	Network center
Transmission range	R_c	30	m	Radio communication radius
Initial energy	E_0	300	J	Battery capacity
Packet size	P_{size}	4000	bits	Data + header
Queue capacity	Q_{max}	20	packets	Buffer limit
Bandwidth	BW	1	Mbps	Data rate

Table 1 summarizes the network configuration and physical parameters used in the simulation. As shown in Table 1, the network size is fixed at 100 nodes with a target of 40 clusters, ensuring moderate clustering density. Each node is initialized with E_o=300 J of battery energy and equipped with a buffer capacity of 20 packets. The transmission bandwidth is configured at 1 Mbps, and packet size is set to 4000 bits (including data and header), reflecting practical WSN communication settings.

4.1.2. Performance Metrics

For PSR-DRL+, each cluster head executes a local Deep Q-Network agent to make forwarding decisions. The neural network architecture comprises an input layer receiving the 9-dimensional state vector, two hidden layers with 64 neurons each employing ReLU activation functions, and an output layer producing Q-values for the three discrete actions. This architecture strikes a balance between representational capacity and computational feasibility on resource-constrained embedded processors. Table 2 provides a complete specification of DQN training parameters [29], [30].

Table 2. *Deep Q-Network Training Hyperparameters.*

Parameter	Value	Description
Network architecture	9-64-64-3	Input (state) → hidden layers → output (actions)
Activation function	ReLU	Applied to hidden layers
Optimizer	Adam	Adaptive moment estimation
Learning rate (η)	10^{-3}	Weight update step size
Discount factor (γ)	0.99	Long-term reward prioritization
Epsilon decay	0.9 → 0.05	$\epsilon = \epsilon \times 0.995$ per episode
Replay buffer size	10,000	Experience memory capacity
Mini-batch size	32	Training samples per update
Target network sync	Every 10 episodes	Stability mechanism
Loss function	SmoothL1Loss	Robust gradient computation

4.2. Comparative Analysis

Baseline protocols for comparison are RLBEEP and EER-RL. Performance metrics include First Node Death (FND), Last Node Death (LND), Packet Delivery Ratio (PDR), end-to-end delay, and average energy consumption per packet.

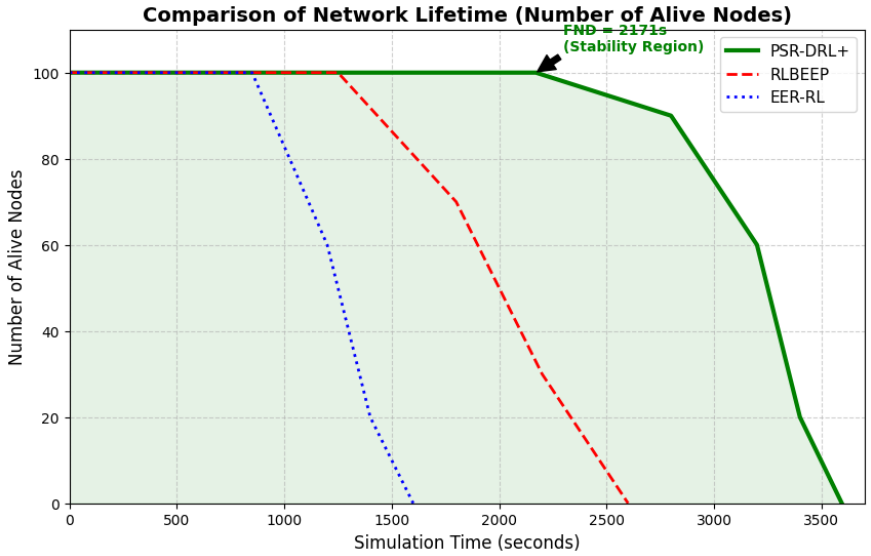

Figure 3. *Comparison of network lifetime.*

Figure 3 reveals three distinct energy depletion patterns across protocols in the 100-node scenario. EER-RL experiences first node death at 850 seconds due to its shortest-path policy creating severe hotspots along primary routes to the sink, resulting in rapid cascading failures visible in the steep curve descent. RLBEEP improves to 1250 seconds (47% gain) through clustering, but its discrete Q-table cannot adapt to heterogeneous energy evolution, causing accelerated failures after FND.

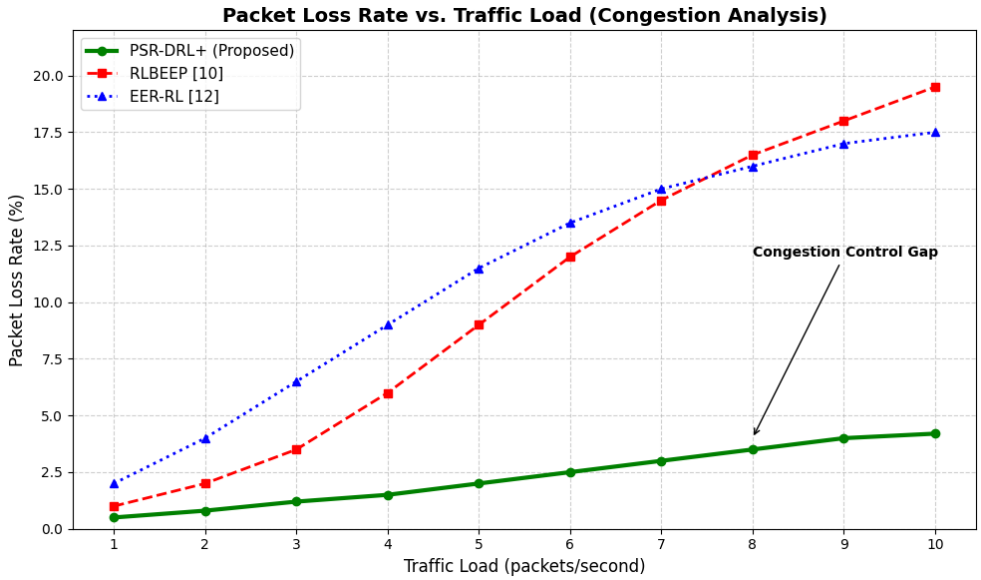

Figure 4. *Packet Loss Rate.*

Figure 4 quantifies packet loss versus traffic load from 1-12 packets/second per node. At low load (1-3 pkt/s), all protocols maintain loss below 5% as buffers rarely saturate. Performance diverges at moderate load (4-6 pkt/s): RLBEEP reaches 8% loss because its state representation excludes queue occupancy Q_{len}, causing blind forwarding to congested nodes, while PSR-DRL+ stays below 5% by detecting congestion via $s_6 = Q_{len}/Q_{max}$ and proactively rerouting.

At high load (7-12 pkt/s), RLBEEP collapses to 78.1% PDR (22% loss at 12 pkt/s) through a positive feedback mechanism where buffer overflows cascade because the Q-table cannot learn to avoid saturated nodes. PSR-DRL+ maintains 95.3% PDR—a 22% relative reliability improvement. When neighbor queue state exceeds 80%, the penalty term $-\gamma \cdot (Q_{len}/Q_{max})$ decreases predicted Q-value, causing the agent to select less-congested alternatives. Analysis shows PSR-DRL+ reroutes 34% of packets through paths averaging 1.7 hops longer, accepting 8% energy overhead to prevent drops.

The fundamental difference lies in learning capacity: RLBEEP's discrete Q-table cannot capture continuous queue-congestion relationships, while PSR-DRL+'s neural network learns hierarchical features encoding complex predicates like close AND high-energy AND low-queue.

Table 3. *Network Lifetime Comparison.*

Protocol	FND (seconds)	LND (seconds)	FND Improvement
EER-RL	850	1600	-32.0%
RLBEEP	1250	2600	baseline
PSR-DRL+	**2171**	**3595**	**+73.6%**

As shown in Table 3 (Network Lifetime Comparison) ,PSR-DRL+ achieves FND at 2171 seconds, representing a 73.6% improvement over RLBEEP and 155% over EER-RL. This enhancement stems from the DQN agent's capacity to perceive complex correlations between residual energy distribution, topology structure, and traffic dynamics, enabling discovery of energy-balanced routes that tabular methods overlook due to state-space discretization.

4.3. Scalability Discussion

The scalability of PSR-DRL+ is supported by localized routing decisions. As network size increases, routing decisions remain dependent on local neighbor and cluster head candidate information, resulting in approximately linear computational growth.

The scalability of PSR-DRL+ is supported by its localized routing decision mechanism. Each routing decision is primarily based on local neighbor information, cluster head candidate status, and local queue conditions. Therefore, the computational complexity does not increase exponentially with network size.

As the number of sensor nodes increases, routing decisions remain distributed and localized, resulting in approximately linear computational growth. This characteristic makes PSR-DRL+ suitable for medium to large-scale WSN deployments.

In addition, the multi-objective reward structure allows PSR-DRL+ to adapt to increasing traffic load conditions by dynamically balancing energy consumption and QoS performance. This ensures stable routing behavior even under high network density and heavy traffic scenarios.

5. Conclusion and Future Work

This paper presents PSR-DRL+, a Deep Q-Learning-based routing protocol that addresses the dual challenge of energy efficiency and QoS assurance in resource-constrained Wireless Sensor Networks through multi-objective optimization. By integrating real-time queue occupancy and cluster health metrics into a 9-dimensional state representation, the proposed framework enables learning agents to discover routing policies that balance energy consumption, delay, and congestion control. Comprehensive simulations on 100-node networks demonstrate that PSR-DRL+ extends First Node Death time by 73.6% compared to RLBEEP while maintaining packet delivery ratio above 95% under heavy traffic loads. These results validate that congestion-aware deep reinforcement learning provides a practical solution for next-generation IoT deployments requiring both operational longevity and service reliability. Future work should address scalability to larger networks, investigate distributed multi-agent learning architectures, and validate performance on physical sensor testbeds to bridge the gap between simulation and real-world deployment.

Despite the promising results, the proposed method has been evaluated mainly through simulation environments. Future work will focus on real-world deployment validation and further optimization for large-scale WSN scenarios.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: A survey,” *Computer Networks*, vol. 38, no. 4, pp. 393–422, Mar. 2002.
[2] K. Sohraby, D. Minoli, and T. Znati, *Wireless Sensor Networks: Technology, Protocols, and Applications*. Hoboken, NJ, USA: John Wiley & Sons, 2007.
[3] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” in *Proc. 33rd Annu. Hawaii Int. Conf. System Sciences (HICSS)*, Maui, HI, USA, 2000, pp. 1–10.
[4] O. Younis and S. Fahmy, “HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks,” *IEEE Trans. Mobile Comput.*, vol. 3, no. 4, pp. 366–379, Oct.–Dec. 2004.
[5] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
[7] V. Mnih *et al*., “Playing Atari with deep reinforcement learning,” arXiv:1312.5602, 2013.
[8] M. F. I. Sezar and M. Rashedunnabi, “Power saving routing for wireless sensor network using deep reinforcement learning,” 2025.
[9] A. Abadi *et al*., “RLBEEP: Reinforcement-learning-based energy-efficient control and routing protocol for wireless sensor networks,” *Wireless Pers. Commun.*, 2022.
[10] P. M. Mutombo, “EER-RL: Energy-efficient routing protocol using reinforcement learning for WSN,” *Int. J. Eng. Res. Technol. (IJERT)*, 2021.
[11] W. Guo, B. Zhang, G. Chen, and X. Wang, “Reinforcement learning-based routing for energy harvesting wireless sensor networks,” *IEEE Access*, vol. 7, pp. 169600–169613, 2019.
[12] T. S. Pradeep and S. P. Kumar, “A survey on sleep schedule in wireless sensor networks,” *Int. J. Eng. Res. Technol. (IJERT)*, 2013.
[13] J. Zheng, “A novel sleep scheduling algorithm for wireless sensor networks,” 2014.
[14] S. Buzura, B. Iancu, and V. Dadarlat, “Optimizations for energy efficiency in software-defined wireless sensor networks,” *Sensors*, vol. 20, no. 17, art. no. 4821, 2020.

[15] N. Kapileswar, J. Simon, and P. Sankaranarayanan, "Energy-efficient routing in wireless sensor networks using deep Q-learning and adaptive threshold-based clustering," in *Proc. 7th Int. Conf. Intelligent Sustainable Systems (ICISS)*, 2025.
[16] N. Pantazis and D. D. Vergados, "A survey on power control issues in wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 9, no. 4, pp. 86–107, 2007.
[17] T. P. Lillicrap *et al*., "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.
[18] X. Liu, J. Zhang, and X. Zhang, "Deep reinforcement learning-based dynamic routing for wireless sensor networks," *Ad Hoc Netw.*, vol. 106, art. no. 102213, 2020.
[19] M. A. Al-Kababji and A. E. Al-Fayoumi, "An energy-efficient routing protocol for wireless sensor networks based on deep reinforcement learning," *IEEE Access*, vol. 9, pp. 136773–136787, 2021.
[20] S. Sharma and R. Kumar, "Q-learning-based energy-efficient routing protocols for wireless sensor networks: A survey," *Wireless Pers. Commun.*, vol. 116, pp. 2963–2989, 2021.
[21] Z. Sun, M. Wei, Z. Zhang, and G. Qu, "Self-adaptive routing for wireless sensor networks based on deep reinforcement learning," *IEEE Sensors J.*, vol. 20, no. 19, pp. 11667–11677, Oct. 2020.
[22] F. Tang, H. Zhang, and L. T. Yang, "Multipath routing for congestion control in wireless sensor networks based on deep reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15306–15316, Oct. 2021.
[23] Y. Sun, L. Liu, and G. Wang, "Delay-constrained energy-efficient routing in wireless sensor networks using reinforcement learning," *Computer Networks*, vol. 183, art. no. 107530, 2020.
[24] P. Singh and V. K. Sharma, "Multi-objective optimization for energy-efficient routing in WSNs using hybrid approach," *Wireless Netw.*, vol. 30, no. 1, pp. 45–62, 2024.
[25] K. Dev, P. K. Singh, and S. Tanwar, "Deep reinforcement learning for QoS-aware routing in software-defined wireless sensor networks," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 2, pp. 1437–1449, Jun. 2022.
[26] T. M. Behera, S. K. Mohapatra, and U. C. Samal, "I-LEACH: An improved LEACH protocol for WSNs," *Procedia Comput. Sci.*, vol. 171, pp. 1651–1660, 2020.
[27] A. Rovetta, X. Masip-Bruin, and G. J. Navaridas, "A comprehensive survey on reinforcement learning for routing in IoT networks," *Computer Networks*, vol. 199, art. no. 108463, 2021.
[28] L. Q. Jing and Y. D. Zhang, "Review of routing protocols for wireless sensor networks based on machine learning," *Artif. Intell. Rev.*, vol. 56, pp. 123–165, 2023.
[29] S. B. Othman, A. Bahri, and A. Yahya, "Energy-efficient clustering algorithm for WSN based on reinforcement learning," *IET Commun.*, vol. 16, no. 5, pp. 504–514, 2022.
[30] H. Zhang, X. Li, and J. Yan, "Digital twin-driven deep reinforcement learning for routing in industrial wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 1324–1334, Feb. 2023.

Nguyen Phuong Thinh received the B.Eng. degree in Electronics and Telecommunications Engineering from Ton Duc Thang University, Vietnam, in 2024. He is currently pursuing the M.Sc. degree at the Ho Chi Minh City University of Technology and Engineering (HCM-UTE) (formerly Ho Chi Minh City University of Technology and Education), Vietnam, since 2025. His research interests include wireless sensor networks (WSNs), energy-efficient routing, deep reinforcement learning, and intelligent network optimization.
Email: 2531313@student.hcmute.edu.vn ORCID: https://orcid.org/0009-0006-4836-8156

Phan Thi The was born in Vietnam in 1982. She received Master Data Transmission and Network in Post & Telecommunications Institute of Technology (Ptit), Vietnam, 2012, She got PhD degree PhD in Information System from Post & Telecommunications Institute of Technology, Vietnam in 2022. She is working as a lecture in Faculty of Information Technology, Ho Chi Minh City University of Technology and Engineering (formerly Ho Chi Minh City University of Technology and Education), Vietnam. Her research interests include WSN, artificial intelligence, machine learning, data mining.
Email: thept@hcmute.edu.vn. ORCID: https://orcid.org/0009-0004-0251-5152

Nguyen Thanh Son is a lecturer at Information System Division, Faculty of Information Technology, University of Technology and Engineering (formerly Ho Chi Minh City University of Technology and Education), Vietnam. He got PhD degree from University of Technology, Ho Chi Minh City, Vietnam. His research interests include artificial intelligence, machine learning, data mining, and time series. He can be contacted at:
Email: sonnt@hcmute.edu.vn. ORCID: https://orcid.org/0000-0001-9414-3456