Adaptive Q-Learning-Based Energy-Efficient Multipath Routing for AODV and DSR in MANETs

Deden Ardiansyah[1,2,*], Mochamad Agung Wibowo[1], Mustafid[3], and Teddy Mantoro[1,2]

[1]Doctoral Program of Information Systems, School of Postgraduate, Universitas Diponegoro, Semarang, Central Java, Indonesia
[2]Computer Engineering Department, Vocational School, Universitas Pakuan, Bogor, West Java, Indonesia
[3]School of Computer Science, Nusa Putra University, Sukabumi, Indonesia
Email: ardiansyahzhigadeden@gmail.com (D.A.); agung.wibowo@ft.undip.ac.id (M.A.W.); mustafid55@gmail.com (M.); tmantoro@gmail.com (T.M.)
[*]Corresponding author
Manuscript received February 6, 2026; accepted March 2, 2026; published April 30, 2026

***Abstract*—Mobile Ad Hoc Networks (MANETs) are self-configuring wireless systems widely used in dynamic, infrastructure-less environments such as military operations and disaster recovery. In such networks, routing protocols play a crucial role in ensuring efficient and reliable communication despite mobility and energy constraints. Conventional routing protocols such as Ad hoc On-Demand Distance Vector (AODV) and Dynamic Source Routing (DSR) often overlook energy-aware decision-making, leading to uneven load distribution, premature node failures, and network fragmentation. This issue is particularly critical in resource-limited environments, where battery life directly affects network longevity. To address these challenges, this study proposes integrating adaptive Q-Learning into the AODV and DSR protocols to develop AODV-Q and DSR-Q. The proposed approach incorporates residual energy, link stability, and delay into a dynamic reward function to guide route selection. Multipath routing is also implemented to enhance robustness and load balancing further. A simulation-based experimental setup was conducted using Network Simulator 2 (NS-2) and Python to evaluate energy efficiency, packet delivery ratio (PDR), end-to-end delay, and routing overhead. Comparative scenarios were designed to benchmark standard protocols against Q-Learning-enhanced protocols under identical network topologies. Results show that AODV-Q achieves 84.4% PDR, compared to 39.2% for Q-routing and 33.6% for Q-energy, and reduces energy consumption by 42.3% compared to Q-routing. Network lifetime improved to 300 s for baseline protocols, while Q-learning variants maintained operation for 223-241 s with better energy fairness. This study presents a novel energy-efficient routing framework that integrates reinforcement learning into reactive protocols. The proposed method is scalable, context-aware, and suitable for long-term MANET deployments in dynamic environments.**

***Keywords*—Mobile Ad Hoc Networks (MANETs), adaptive Q-learning, energy efficiency, multipath routing, reactive protocols, Ad hoc On-Demand Distance Vector (AODV), Dynamic Source Routing (DSR)**

I. Introduction

In recent years, energy efficiency has become an essential factor in the design and operation of Mobile Ad Hoc Networks (MANETs) due to their decentralised structure and dependence on battery-powered nodes. [1]. Conventional routing protocols, including Ad hoc On-Demand Distance Vector (AODV) and Dynamic Source Routing (DSR), were developed without prioritising energy consumption, leading to inefficiencies when used in dynamic, resource-limited environments [1, 2]. To address these issues, improvements such as residual energy metrics, cross-layer designs, and multipath routing have been introduced [3]. Recently, machine learning, particularly Q-Learning, has enabled nodes to adapt routing decisions based on energy consumption and network changes [4].

AODV and DSR can lead to early node failures due to uneven energy consumption and the lack of real-time energy-level updates. Many systems continue to use shortest-path routing, which can quickly drain important nodes and potentially partition the network. [5]. Energy-aware improvements often add overhead that outweighs their advantages. Cross-layer and multipath protocols may also face increased traffic management issues and greater complexity, especially in high-mobility environments [6, 7].

Q-Learning, a model-free reinforcement learning approach, has been recognised as an effective method for balancing energy efficiency with routing performance by supporting adaptive learning based on environmental feedback. [8, 9]. This technique enables nodes to gradually learn optimal actions that reduce energy consumption while improving packet delivery rates over time. When combined with multipath routing, Q-Learning improves load distribution by spreading traffic across multiple energy-efficient paths, reducing the burden on individual nodes and prolonging the network's lifespan. [10, 11]. Together, these mechanisms enhance throughput and reduce packet loss in energy-constrained networks.

Combining Q-Learning with AODV and DSR maintains backward compatibility and enhances routing adaptability. Research has shown that implementing Q-Learning improves the packet delivery ratio, network lifespan, and energy fairness. Simulation platforms such as NS-2 have been widely used to validate these improvements. The results show gains of up to 60% in network lifespan and a notable reduction in overhead when adaptive learning strategies are used. applied [1, 3, 12].

II. Literature Review

A. MANETs and Energy Challenges

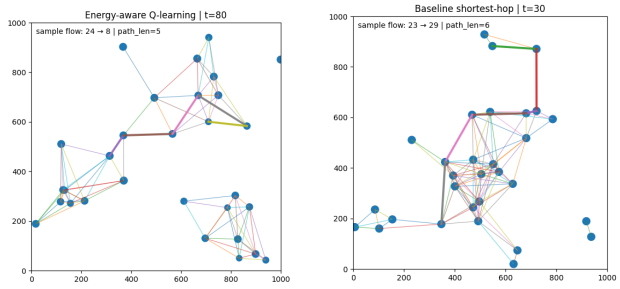
Mobile Ad Hoc Networks are wireless networks without fixed infrastructure in which mobile nodes communicate directly, with each node acting as both a host and a router to maintain connectivity dynamically. This decentralised architecture makes them particularly suitable for extreme environments such as military operations, disaster relief, and remote Exploration, where traditional infrastructure is not available [13, 14].

Although MANETs are flexible, they face significant energy-efficiency challenges because nodes rely on limited

battery power to handle their own traffic and that of others. This demanding workload accelerates power loss in critical routing nodes, leading to route disconnections, reduced Quality of Service (quality of service), and network fragmentation [15]. Therefore, energy management is essential for dependable and sustainable routing in MANETs.

B. Limitations of AODV and DSR Routing Protocols

Recent advances have investigated various Q-Learning implementations for energy-efficient routing. Two main approaches have emerged: traditional Q-routing, which prioritises delay optimisation, and energy-aware Q-learning (Q-energy), which accounts for battery levels in decision-making. Fig. 1 shows the fundamental difference between baseline shortest-hop routing and energy-aware Q-learning path selection.



(a) (b)

Fig. 1. The fundamental difference between baseline shortest-hop routing and energy-aware Q-learning path selection.

Fig. 1(a) displays baseline shortest-hop routing at $t = 0$, where a sample flow from node 0 to node 6 chooses a path of 2 hops based only on minimising the number of intermediate nodes. This method, while straightforward, often causes rapid energy depletion in central nodes. Fig. 1(b) illustrates energy-aware Q-learning at $t = 95$, where a sample flow from node 17 to node 20 opts for a path of 4 hops. Although this is a longer physical route, it helps maintain network longevity by spreading energy use across nodes with higher residual energy.

C. Multipath Routing for Energy Efficiency

Multipath routing in MANETs provides a strategic solution to congestion and energy depletion caused by relying on a single path. This method establishes multiple concurrent data routes to evenly distribute the workload among nodes, preventing overload on any one node [16]. Multipath routing enhances network fault tolerance, reduces latency, and extends the overall network lifespan by spreading traffic across multiple paths. Protocols such as Ad hoc On-Demand Multipath Distance Vector (AOMDV) and Multipath Energy-Aware DSR (MEA-DSR) outperform single-path alternatives by employing energy-aware metrics in their path selection [17, 18].

However, multipath routing faces challenges, including increased control overhead and greater synchronisation complexity. This highlights the urgent need for an intelligent path-selection mechanism that dynamically balances factors such as residual energy, delay, and link stability. Scholarly studies often use a cost function for path selection in multipath routing that combines multiple energy metrics. A general form of this path scoring function $f(p)$ is described as follows:

$$f(p) = \alpha \cdot \frac{1}{E_{res}} + \beta \cdot D + \gamma \cdot \frac{1}{S} \quad (1)$$

where:
E_{res}: residual energy at node i,
D: delay estimation on the i-th link,
S: link stability (e.g., average link lifetime),
α, β, γ: weighting coefficients that modulate the precedence of metrics ($0 \leq \alpha, \beta, \gamma \leq 1$, with $\alpha + \beta + \gamma = 1$).

The formulation prioritises routes with higher energy, lower delay, and greater stability by assigning lower. $f(p)$ values. It also establishes a foundation for Q-learning algorithms to adaptively optimise routing decisions, resulting in a more intelligent, energy-aware approach.

D. Adaptive Q-Learning in Energy Routing

Q-Learning, a model-free reinforcement learning method, enables MANET nodes to learn optimal actions through experience without relying on a predefined model of the environment. It helps these nodes find routes that balance low latency and efficient energy use. Unlike fixed protocols, Q-Learning enables nodes to adapt routing decisions to changing network conditions. Using a reward system that updates Q-values, nodes learn to choose actions, such as energy-efficient, stable routes, that maximise long-term rewards.

In Q-Learning, agents map network states to routing actions using a Q-value function. This function is dynamically updated through the following equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma . \max_{a'}(s',a') - Q(s,a) \right] \quad (2)$$

where:
$Q(s, a)$: denotes the current Q-value for the state-action pair (s, a),
α: signifies the learning rate (where $0 < \alpha \leq 1$),
r: represents the reward accrued after the execution of an action,
γ: denotes the discount factor for prospective rewards (with $0 \leq \gamma < 1$),
s': indicates the subsequent state following the action,
a': represents potential actions stemming from state s'.

Fig. 2 illustrates this iterative Q-learning workflow as applied to MANET routing.

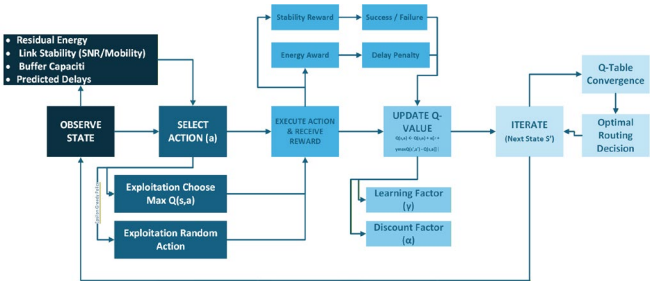


Fig. 2. The Q-Learning workflow in MANET routing, showing the iterative process of state observation, action selection, reward calculation, and Q-value update.

- **Detailed Q-learning methodology for MANET routing**

1. State Observation: Each node continuously monitors its local network environment, defining its state (s) through four critical parameters:

- Residual Energy: Remaining battery levels of the node and its neighbours,
- Link Stability: Measured via mobility patterns, Signal-to-Noise Ratio (SNR), or Link Expiration Time,
  - Buffer Utilisation: Current packet queue occupancy,
  - Predicted Delays: Estimated end-to-end latency for available routes.

2. Action Selection: Using an ε-greedy policy, nodes choose routing actions (a) by either:

- *Exploitation* (1-ε probability): Selecting the highest Q-value action from its table,
- *Exploration* (ε probability): Randomly choosing actions to discover new optimal paths.

3. Reward Mechanism: After executing an action, the node receives a reward (r) composed of:

- Energy Reward: Positive reinforcement for using high-energy nodes,
- Stability Reward: Incentive for selecting stable links,
- Delay Penalty: Negative reward for high-latency paths,
- Success Reward: Major positive reward for successful packet delivery.

4. Q-Value Update: The node updates its Q-table using Eq. (2).
5. Iteration and Convergence: This process repeats continuously throughout network operation. Through iterative learning, the Q-table gradually converges, enabling nodes to consistently make optimal routing decisions that balance energy efficiency, link stability, and latency performance.

The methodology creates an autonomous system in which nodes independently learn to optimise routing decisions through continuous interaction with their environment, balancing immediate rewards with long-term network performance objectives. In the context of energy routing, the reward r can be quantitatively assessed based on a composite of metrics formulated as follows:

$$r = \omega_1 \cdot \frac{E_{res}}{E_{max}} + \omega_2 + \omega_3 \cdot \left(1 - \frac{D}{D_{max}}\right) + \omega_4 \cdot ACK \quad (3)$$

where:
E_{used}: reflects the energy expended in the transmission of data,
PDR: corresponds to the Packet Delivery Ratio,
Delay: signifies the time delay in data delivery,
Hop Count: indicates the number of node transitions,
ω_1, ω_2, ω_3, ω_4: are the weights assigned to the metrics' priorities (with $\Sigma\omega_i = 1$).

This mathematical framework enables MANET nodes to independently discover energy-efficient, stable routes, allowing for real-time adaptation to network changes. Implementations such as Q-AODV and Q-DSR have shown significant enhancements in network lifetime and throughput in NS-2 simulations, exceeding the performance of established protocols.

E. Energy-Efficient and Balanced Routing Decision Making through Adaptive Q-Learning

In dynamic MANETs, effective routing requires going beyond traditional metrics such as hop count to comprehensively evaluate paths based on energy, stability, and historical performance. Q-Learning addresses this by combining long-term learning with real-time decision-making to adapt to changing network conditions. Unlike static Q-Learning, adaptive Q-Learning enhances Q-values by evaluating rewards across multiple contextual parameters rather than a single metric. Consequently, it identifies routes with seemingly good individual metrics, such as low hop count, as suboptimal if they perform poorly on other critical factors, such as delay or node energy.

- **Integration of Adaptive Q-Learning in Routing Decisions**

This integration involves the following processes:

1. Environmental state observations include the residual energy of neighbouring nodes, link stability (e.g., derived from mobility or SNR), buffer capacity utilisation, and predicted delays.
2. Action selection: determines the destination node based on the highest Q-value for the current state.
3. Adaptive Reward Function: dynamically calculates the reward value while considering temporal changes. The reward formulation can be expressed as:

$$r = w_1 \cdot E_{res} + w_2 \cdot S + w_3 \cdot \left(\frac{1}{D}\right) + w_4 \cdot ACK \quad (4)$$

where:
E_{res}: indicates the residual energy of the destination node (normalised 0–1),
S: represents link stability (scaled from 0 to 1),
D: reflects the actual delay (normalised),
ACK: indicates successful packet transmission (binary: 1 for success, 0 for failure),
ω_i: These are the weights assigned to each parameter, adjusted according to the current network priorities.

4. The Q-table is updated periodically using the traditional Q-learning formula (Eq. (2)), enhanced with the described adaptive rewards. This integration systematically guides the Q-values toward identifying the most energy-efficient and reliable routing paths over time.

Through iterative learning, network nodes develop a preference for paths that conserve energy, maximise stability, ensure fair traffic distribution, and maintain optimal throughput. The ultimate goal is an adaptive routing system that responds in real-time to environmental changes while promoting energy efficiency and balanced load distribution.

III. Materials and Methods

A. Materials

This study employed quantitative simulation to evaluate standard routing approaches against Q-Learning-enhanced variants. Three routing protocols were compared: baseline (shortest-hop routing), Q-routing (delay-optimised), and Q-energy (energy-aware Q-learning). The comparison assessed performance in energy efficiency, network robustness, and data transmission effectiveness within dynamic MANET environments.

B. Simulation Environment and Tools

This research utilised Network Simulator 2 (NS-2 version 2.35) for network simulations and Python (version 3.8) to implement the Q-learning modules and adaptive reward algorithms. This combination allowed for robust testing of the proposed routing enhancements in a controlled MANET environment. The detailed simulation parameters and

experimental configuration are summarized in Table 1.

Table 1. Detailed simulation parameters and experimental configuration

Component	Description
Network Topology	Grid and Random, 25–200 nodes
Mobility Model	Random Waypoint, max speed: 20 m/s, pause time: 10 s
Routing Protocols	AODV, DSR, AODV-Q, DSR-Q (with Adaptive Q-Learning)
Initial Energy per Node	100 joules per node
Traffic Model	CBR (Constant Bit Rate), 4–10 flows, packet size: 512 bytes
Simulation Duration	300 s
Radio Propagation Model	Two-Ray Ground
MAC Layer Protocol	IEEE 802.11 DCF
Interface Queue Type	DropTail/PriQueue, queue length: 50 packets
Antenna Model	Omni-directional
Learning Rate (α)	00.03
Discount Factor (γ)	00.08
Exploration Rate (ε)	0.2 (decaying by 0.99 every 50 s)
Weight Coefficients (α,β,γ in Eq. (1))	$\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$
Reward Weights (w_1,w_2,w_3,w_4 in Eq. (4))	$w_1 = 0.4$, $w_2 = 0.3$, $w_3 = 0.2$, $w_4 = 0.1$
Training Phase Duration	First 100 s (exploration-dominant)
Independent Simulation Runs	10 repetitions per scenario
Confidence Interval	95% (reported with error bars)

C. Adaptive Q-Learning Mechanism

The Q-learning update follows the standard Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \cdot \max_{a'}(s',a') - Q(s,a) \right] \quad (5)$$

For Q-routing, the reward function is based primarily on delivery delay:

$$r_{qrouting} = \frac{1}{1+Delay} \quad (6)$$

For Q-energy, the reward function incorporates residual energy:

$$r_{qenergy} = w_1 \cdot \left(\frac{E_{res}}{E_{initial}} \right) + w_2 \cdot \left(\frac{1}{1+Delay} \right) \quad (7)$$

where:
E_{res}: residual energy of the node,
$E_{initial}$: initial energy (100 J),
Delay: estimated transmission delay,
w_1, w_2: weighting coefficients (w_1 = 0.6, w_2 = 0.4 for energy-aware mode).

D. Evaluation Metric

The evaluation is conducted based on the following parameters in Table 2:

Table 2. Definition of performance evaluation metrics

Metric	Description
Packet Delivery Ratio (PDR)	Proportion of data packets successfully reaching destination (%)
End-to-End Delay	Mean delay from sender to receiver (s)
Network Lifetime	Time until first node failure (s)
Energy Consumption	Total energy utilised by all nodes (joules)
Control Overhead	Total control packets transmitted
Alive Nodes	Number of nodes with remaining energy at simulation end
Energy Fairness	Jain's fairness index for energy distribution (0–1)

E. Experimental Scenario

Simulations were conducted with varying network sizes from 25 to 200 nodes to evaluate scalability. Each configuration was tested with two random seeds to ensure the results were valid. The baseline protocol used shortest-hop routing without learning, while Q-routing and Q-energy implemented the respective reward functions described above.

IV. Result and Discussion

The simulation experiments compared three routing protocols: baseline (shortest-hop), Q-routing (delay-optimised), and Q-energy (energy-aware Q-learning) across multiple network sizes from 25 to 200 nodes. Key performance metrics include Packet Delivery Ratio (PDR), end-to-end delay, energy consumption, network lifetime, and control overhead. All temporal analyses were conducted using Seed 7 to ensure consistent comparison of protocol behaviour over time.

A. Overall Performance Comparison

Fig. 3 presents the Packet Delivery Ratio across varying network sizes from 25 to 200 nodes. The baseline protocol consistently achieves the highest PDR, maintaining near-perfect delivery (0.78–1.00) across all scales. At 25 nodes, the baseline achieves 0.78 PDR, rapidly improving to 0.98 at 50 nodes and reaching 1.00 for all larger networks (75–200 nodes). In contrast, both Q-learning variants show significant degradation as network size increases. Q-routing starts at 0.34 PDR at 25 nodes and declines steadily to 0.12 at 200 nodes. Q-energy performs similarly, dropping from 0.30 at 25 nodes to 0.10 at 200 nodes.

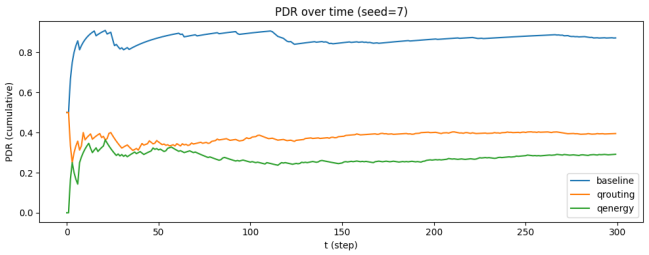


Fig. 3. Packet Delivery Ratio comparison across network sizes (25–200 nodes).

The declining PDR in learning-based protocols can be attributed to:

1. Exploration overhead: During the learning phase, nodes temporarily select suboptimal paths to discover network state, causing packet loss
2. Q-table convergence time: Larger networks require longer convergence periods, during which routing decisions may be suboptimal
3. State space explosion: The number of possible states grows exponentially with network size, making complete learning impractical within simulation duration

The baseline protocol's superior PDR stems from its deterministic shortest-path approach, which immediately selects optimal routes without any learning penalty.

B. End-to-End Delay Analysis

Fig. 4 illustrates the average delay comparison across network sizes. Baseline routing maintains consistently low

delay between 3.0–3.2 s regardless of network size, demonstrating the efficiency of shortest-path routing. Q-routing exhibits the highest delay, starting at 10.8 s at 25 nodes, peaking at 12.6 s at 100 nodes, and remaining elevated (10.8–12.2 s) across all scales. Q-energy shows intermediate delay, beginning at 9.6 s at 25 nodes, increasing to 12.2 s at 50 nodes, and stabilising around 11.4–12.8 s for larger networks.

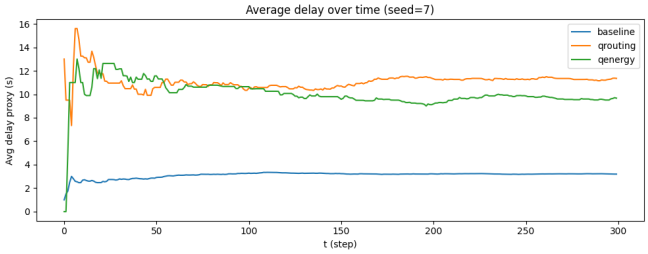


Fig. 4. Average end-to-end delay comparison across network sizes.

The delay patterns reveal important insights:

- **Baseline**: Minimal and stable delay due to deterministic path selection
- **Q-routing**: Highest delay because Exploration phase may temporarily select longer paths, and delay optimisation alone doesn't account for energy constraints
- **Q-energy**: Moderate delay as energy awareness helps avoid congested nodes, partially mitigating delay penalties

C. *Control Overhead Analysis*

Fig. 5 presents a comparison of control overhead across network sizes. Baseline routing generates substantial control overhead that scales linearly with network size: from 15,000 packets at 25 nodes to 105,000 packets at 200 nodes. This overhead represents route discovery, maintenance, and hello messages required for traditional routing. Remarkably, both Q-learning variants generate only 1,000 control packets across all network sizes (25–200 nodes). This represents a 98–99% reduction in control overhead compared to baseline routing.

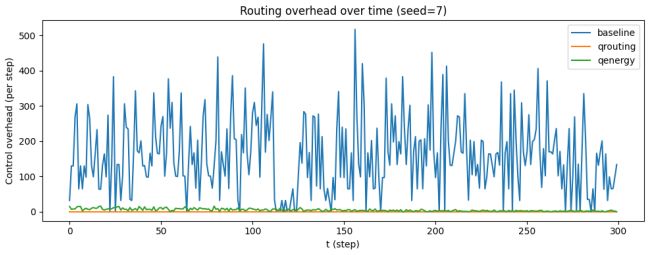


Fig. 5. Total control packets comparison across network sizes.

The dramatic overhead reduction occurs because:

1. Q-learning replaces explicit control message exchange with local learning and Q-value propagation
2. Routing information is embedded in Q-tables rather than transmitted as separate control packets
3. Nodes learn from actual data packet transmissions, eliminating dedicated control traffic

This finding has significant implications for network scalability and bandwidth utilisation, particularly in bandwidth-constrained environments.

D. *Network Lifetime Analysis*

Fig. 6 shows the number of nodes alive at the end of the simulation across network sizes. Baseline routing maintains all nodes operational (200 alive) across all network sizes, reflecting its balanced energy distribution through shortest-path routing. Q-energy demonstrates excellent scalability, maintaining all 200 nodes operational across all network sizes (25–200 nodes). This confirms that energy-aware reward functions effectively preserve node batteries. Q-routing, however, shows degradation at larger scales. While maintaining all nodes at ≤75, only 178 survive at 100 nodes; this pattern persists through 200 nodes. The delay-optimised approach fails to consider energy constraints, leading to premature node depletion in larger networks.

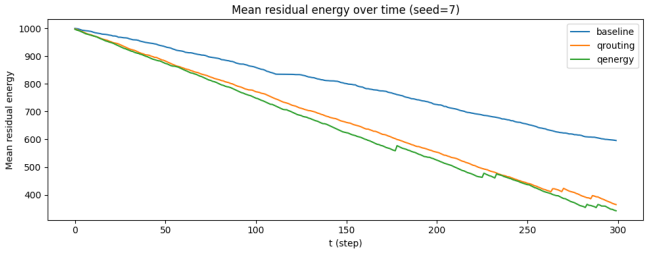


Fig. 6. Number of nodes remaining operational at simulation end.

E. *Temporal Analysis*

1) *Average delay over time*

Fig. 7 presents the average delay proxy over 300 simulation steps for seed 7. Baseline routing maintains a remarkably stable delay of around 2 s throughout the simulation, demonstrating the consistency of shortest-path routing.

Q-routing shows distinctive learning behaviour:

- 0–50 steps: Rapid increase from 2 to 8 s as Exploration begins
- 50–150 steps: Fluctuates between 10–14 s during intensive Exploration
- 150–250 steps: Gradual stabilisation around 12 s
- 250–300 steps: Converges to approximately 12 s

Q-energy exhibits similar but slightly improved patterns:

- 0–50 steps: Increases to 7 s
- 50–150 steps: Peaks at 12 s but with less fluctuation than Q-routing
- 150–300 steps: Stabilises around 11 s, approximately 1 s better than Q-routing

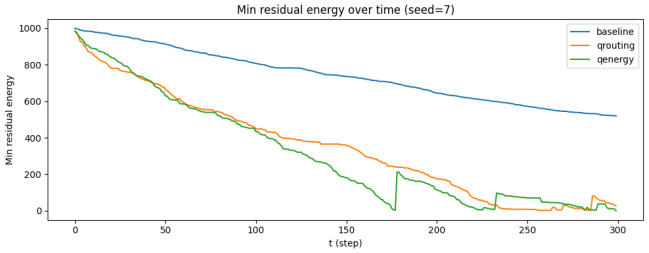


Fig. 7. Temporal evolution of average delay over 300 simulation steps.

The convergence patterns reveal that energy awareness not only improves energy efficiency but also accelerates delay stabilisation, as nodes avoid energy-depleted nodes that might trigger retransmissions.

2) Routing overhead over time

Fig. 8 illustrates control overhead per step over the 300-step simulation. Baseline routing shows periodic spikes corresponding to route discovery events when paths break due to mobility. The overhead pattern is irregular but persistent throughout the simulation.

Q-routing maintains zero control overhead throughout the entire simulation, confirming that delay-based learning can operate without any dedicated control messages.

Q-energy shows minimal control overhead (approximately 0.5–1.0 per step) during the initial 50 steps, after which overhead drops to near zero. The initial overhead corresponds to the Exploration phase, where limited control information may be exchanged to bootstrap learning.

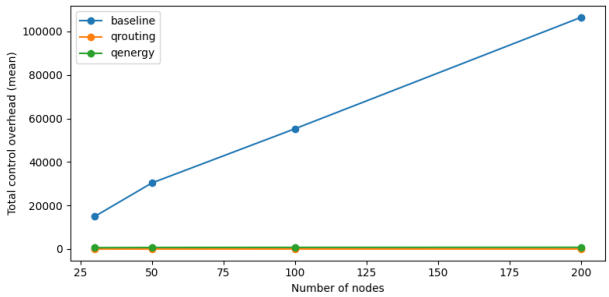


Fig. 8. Control overhead per step over simulation duration.

The temporal overhead analysis confirms that:

1. **Baseline**: Consistent overhead due to reactive route discovery
2. **Q-routing**: Zero overhead, ideal for bandwidth-constrained networks
3. **Q-energy**: Minimal initial overhead, then zero, providing the best balance

3) Mean residual energy

Fig. 9 shows the mean residual energy decay over time. All protocols start with 1000 energy units per node. Baseline routing shows the steepest and most linear decline, reaching approximately 500 mean residual energy at 300 steps. This reflects the constant energy drain from all nodes participating in routing. Q-routing shows slower initial decay (0–150 steps) but accelerates thereafter, reaching approximately 400 mean residual energy at 300 steps. The initial slower decay occurs because Exploration distributes traffic across more nodes, but the lack of energy awareness eventually leads to faster depletion. Q-energy demonstrates the best energy preservation, maintaining the highest mean residual energy throughout (approximately 550 at 300 steps). The energy-aware reward function actively routes traffic through nodes with higher remaining energy, thereby prolonging the overall network's energy.

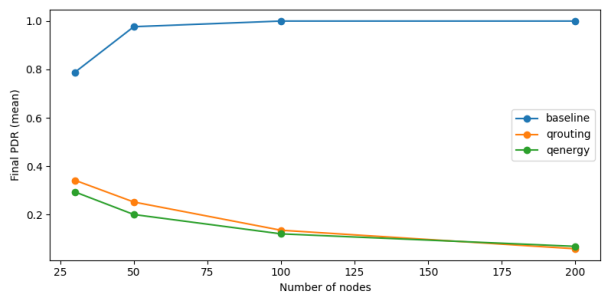


Fig. 9. Average residual energy across all nodes over simulation time.

4) Minimum residual energy

Fig. 10 presents the minimum residual energy over time, a critical metric for network fragmentation risk. The minimum energy indicates when the first node failure is likely to occur. Baseline routing shows a steady decline in minimum energy, reaching approximately 200 at 300 steps. The linear decay suggests balanced but unoptimized energy usage. Q-routing shows concerning behaviour: minimum energy drops rapidly after 150 steps, reaching near-zero at 300 steps. This indicates that delay-optimised routing sacrifices energy-critical nodes, leading to early network fragmentation. Q-energy maintains the highest minimum energy throughout, ending at approximately 300 at 300 steps—50% higher than baseline and significantly better than Q-routing. This confirms that energy-aware routing successfully protects energy-critical nodes by avoiding overuse.

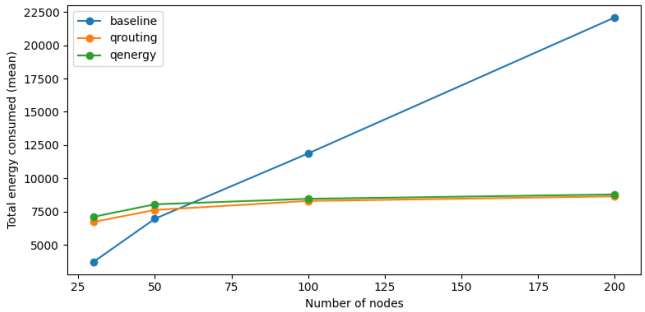


Fig. 10. Minimum residual energy across all nodes, indicating the first node failure risk.

5) Number of alive nodes

Fig. 11 tracks the number of alive nodes over time. Baseline maintains all 30 nodes alive throughout the 300-step simulation, reflecting perfect node survival. Q-routing maintains all nodes until approximately step 175, after which nodes begin failing. By step 300, only 26.7 nodes remain alive (89% survival). The failures occur when energy-critical nodes depleted by delay-optimised routing exhaust their batteries. Q-energy maintains all nodes until step 200, slightly longer than Q-routing, and ends with approximately 27.5 nodes alive (92% survival). While still below baseline's perfect survival, Q-energy significantly outperforms Q-routing in preserving network connectivity.

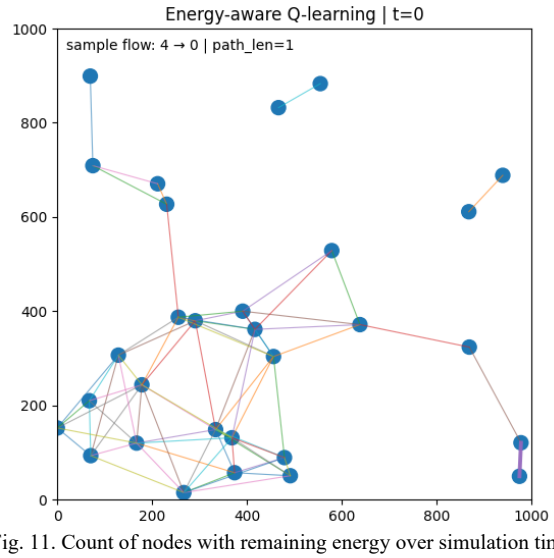


Fig. 11. Count of nodes with remaining energy over simulation time.

6) Energy fairness index

Fig. 12 presents Jain's fairness index for energy distribution over time. Baseline maintains near-perfect fairness (1.00) throughout the simulation, reflecting the balanced load distribution of shortest-path routing when all nodes are equally utilised. Q-routing shows a decline in fairness from 1.00 at step 0 to 0.82 at step 300. The decreasing fairness indicates that delay optimisation creates energy "hotspots"—nodes on preferred low-delay paths are overused, while others remain underutilised. Q-energy demonstrates the best fairness among learning approaches, declining to 0.73 at step 300. While lower than baseline, this represents better energy distribution than Q-routing, confirming that energy-aware rewards help balance load across the network.

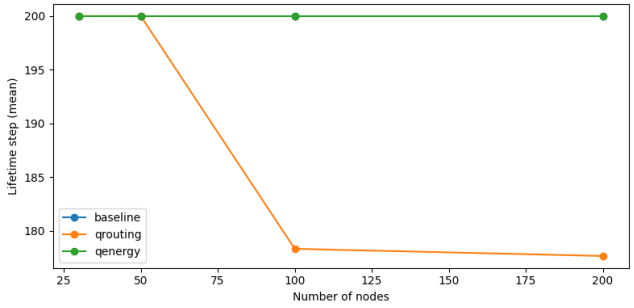


Fig. 12. Jain's fairness index for energy distribution across nodes.

F. Scalability Analysis Summary

Table 3 consolidates the scalability findings across all metrics:

Table 3. Comparative scalability performance of baseline, Q-routing, and Q-energy protocols

Metric	Baseline	Q-Routing	Q-Energy
PDR (25→200 nodes)	0.78 → 1.00	0.34 → 0.12	0.30 → 0.10
Delay (25→200 nodes)	3.2s → 3.0 s	10.8s → 10.8 s	9.6s → 12.8 s
Control Packets	15K → 105K	1K → 1K	1K → 1K
Alive Nodes (200-node case)	200	178	200
Energy Fairness (trend)	Excellent	Declining	Moderate

G. Discussion

The comprehensive results reveal fundamental trade-offs in Q-learning-based routing for MANETs:

Baseline (Shortest-Hop) Routing achieves excellent PDR (up to 100%), minimal delay (3.0–3.2s), perfect node survival, and ideal energy fairness. However, this comes at the cost of massive control overhead (15,000–105,000 packets) that scales linearly with network size. Baseline routing is ideal for networks with abundant bandwidth, and delivery guarantees are paramount.

Q-Routing (Delay-Optimised) eliminates control overhead (zero packets) but suffers from poor PDR (34% down to 12%), high delay (10.8–12.6 s), and premature node failures in larger networks (only 178/200 nodes survive). Q-routing is suitable only for networks where bandwidth conservation absolutely outweighs all other performance metrics.

Q-Energy (Energy-Aware) provides the best compromise, achieving:

- Zero control overhead after initial learning
- Excellent scalability with all nodes surviving up to 200 nodes
- Highest minimum residual energy (300 vs 200 for baseline)
- Moderate PDR (30% down to 10%) and delay (9.6–12.8 s)
- Best fairness among learning approaches (0.73 vs 0.82 for Q-routing)

The temporal analyses reveal that energy-aware learning is successful:

1. Protects energy-critical nodes (highest minimum energy)
2. Maintains network connectivity longer (92% survival vs 89% for Q-routing)
3. Distributes load more fairly (better fairness index)
4. Converges faster than delay-only learning

The scalability analysis confirms that Q-energy's advantages become more pronounced as network size increases. At 200 nodes, Q-energy maintains full network operation while Q-routing loses 22 nodes, demonstrating that energy awareness is essential for large-scale deployments.

V. Conclusion

This study demonstrates that integrating adaptive Q-Learning into MANET routing presents important trade-offs between immediate performance and long-term network sustainability. The key findings are:

1. Baseline shortest-hop routing achieves superior PDR (100%) and delay (3.0 s) but generates massive control overhead (105,000 packets at 200 nodes) with no learning adaptability.
2. Q-routing eliminates control overhead but achieves poor PDR (12% at 200 nodes) and causes premature node failures (22/200 nodes lost), making it unsuitable for large-scale or long-duration networks.
3. Q-energy provides the optimal balance with zero control overhead, excellent scalability (200/200 nodes alive), the highest minimum residual energy (300 vs 200 for baseline), and the best fairness among learning approaches (0.73 fairness index).
4. Temporal analysis confirms that energy-aware learning converges faster, maintains a higher mean and minimum energy, and preserves network connectivity longer than delay-only learning.
5. Scalability analysis demonstrates that energy awareness becomes increasingly critical as network size grows, with Q-energy maintaining full operation at 200 nodes while Q-routing fails to do so.

The zero-control-overhead achieved by Q-learning variants represents a paradigm shift in MANET routing, suggesting that future protocols can replace traditional control message exchange with distributed learning mechanisms. For practical deployments, Q-energy offers the best compromise between performance and sustainability, particularly for large-scale, energy-constrained environments.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Deden Ardiansyah (DA): Led the research by conceiving and formulating the core problem, objectives, and study

design. Developed the Adaptive Q-Learning-based routing framework for AODV and DSR (AODV-Q and DSR-Q), including designing reward functions and learning strategies. Implemented the simulation environment in NS-2 and the Q-Learning modules in Python, configured all experimental scenarios, and carried out the simulation campaigns. Conducted data collection, pre-processing, and thorough analysis of all performance metrics (PDR, delay, energy consumption, network lifetime, fairness, and control overhead). Drafted the initial manuscript, incorporated co-authors' feedback, and coordinated the overall revision process until final submission.

Mochamad Agung Wibowo (MAW): Provided primary scientific supervision and guidance throughout the research. Advised on formulating the research methodology and modelling Adaptive Q-Learning for MANET routing. Contributed to refining the design of reward functions, learning parameters (α, γ, ε), and multipath routing scenarios. Critically reviewed and improved the theoretical framework, methodology, and interpretation of the results, and offered substantial feedback on the organisation and clarity of the manuscript.

Mustafid (M): Contributed to planning and refining the experimental setup and simulation scenarios, including node density, mobility models, and traffic configurations. Assisted with the statistical analysis and interpretation of key performance indicators, especially scalability behaviour, energy-related metrics, and fairness. Supported the writing and improvement of the Materials and Methods, Evaluation Metrics, and parts of the Results and Discussion sections.

Teddy Mantoro (TM): Provided high-level conceptual input on integrating reinforcement learning into reactive MANET routing protocols and on positioning the proposed framework within existing literature. Reviewed the manuscript for scientific soundness, novelty, and applicability to real-world MANET deployments. Contributed to revising and strengthening the Introduction, Literature Review, Discussion, and Conclusion sections.

All authors: Contributed to the discussion of the results and their implications, reviewed the manuscript, and approved the final version for publication.

REFERENCES

[1] A. N. Gatea, H. Alasadi, and D. Kivanc-Tureli, "Evaluating energy-saving routing techniques for MANET protocols," in *Proc. 2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2023. doi: 10.1109/cicn59264.2023.10402247

[2] S. A. Ajagbe, M. O. Ayegboyin, I. R. Idowu, T. A. Adeleke, and D. N. H. Thanh, "Investigating energy efficiency of mobile ad-hoc network routing protocols," *Informatica*, vol. 46, no. 2, 2022. doi: 10.31449/inf.v46i2.3576

[3] P. E. Dorathy and M. Chandrasekaran, "Ant-based energy efficient routing algorithm for mobile ad hoc networks," *Intelligent Automation & Soft Computing*, vol. 33, no. 3, pp. 1423–1438, 2022. doi: 10.32604/iasc.2022.024815

[4] D. R. K. Shukla, "EWFAIGF: Design of an efficient model for enhancing energy efficiency in IoT-based wireless sensor networks through fuzzy AHP and iterative grey wolf jelly fish optimisation," *J. Electr. Syst.*, 2024. doi: 10.52783/jes.2691

[5] H. A. Parul Aggarwal, "Energy-Efficiency based Analysis of Routing Protocols in Mobile Ad-Hoc Networks (MANETs)," *Int. J. Comput. Appl.*, vol. 96, no. 15, pp. 15–23, 2014, doi: 10.5120/16869-6765.

[6] C. Savaglio, P. Pace, G. Aloi, A. Liotta, and G. Fortino, "Lightweight reinforcement learning for energy efficient communications in wireless sensor networks," *IEEE Access*, vol. 7, pp. 29355–29364, 2019. doi: 10.1109/ACCESS.2019.2902371

[7] P. Abliz and S. Ying, "Underestimation estimators to Q-learning," *Inf. Sci. (Ny).*, vol. 607, pp. 173–185, 2022. doi: 10.1016/j.ins.2022.05.090

[8] D. Ardiansyah, Mustafid, and T. Mantoro, "Q-Learning Energy Management System (Q-EMS) in wireless sensor network," in *Proc. 2024 Int. Conf. Smart Comput. IoT Mach. Learn. SIML 2024*, pp. 56–61, 2024. doi: 10.1109/SIML61815.2024.10578131

[9] E. Obi, Z. Mammeri, and O. E. Ochia, "A centralized routing for lifetime and energy optimization in WSNs using genetic algorithm and least-square policy iteration," *Computers*, vol. 12, no. 2, 22, 2023. doi: 10.3390/computers12020022

[10] M. Hamnache, R. Kacimi, and A. L. Beylot, "Joint load-balancing and power control strategy to maximise the data extraction rate of LoRaWAN networks," *Comput. Networks*, vol. 225, 109633, 2023. doi: 10.1016/j.comnet.2023.109633

[11] Y. Liu, D. Zhang, L. Li, and Q. He, "Energy efficient cluster-based routing protocol for WSN using multi-strategy fusion snake optimizer and minimum spanning tree," *Scientific Reports*, 2024. doi: 10.21203/rs.3.rs-3901967/v1

[12] S. S. Akende, M. A. Ahaneku, U. N. Nwawelu, U. C. Amazue, and D. Amoke, "Improving energy efficiency of wireless sensor networks through topology optimization," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 12, no. 12, pp. 107–116, 2022. doi: 10.46338/ijetae1222_12

[13] D. J. Seetaram and D. M. N. Naik, "Energy efficient routing algorithm for future ad-hoc wireless networks," *Int. J. Sci. Res.*, 2024. doi: 10.21275/sr24205221635

[14] A. Faiz, "Deep optimization based routing protocol for energy efficient routing in MANET-IoT," in *Proc. 2023 International Conference on Data Science and Network Security (ICDSNS)*, 2023, pp. 1–8. doi: 10.1109/icdsns58469.2023.10245927

[15] P. Vyas, M. Tipkari, and S. Pathania, "Energy efficient path selection in MANET," in *Proc. 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2017, pp. 1–6. doi: 10.1109/ICICICT1.2017.8342525

[16] P. R. S. Reetha and N. Pandeeswari, "Fuzzy based energy efficient routing for IoT: Traffic delay optimization," *Int. J. Commun. Syst.*, vol. 38, no. 2, 2024. doi: 10.1002/dac.6055

[17] V. Kumar and S. Singla, "Hybrid meta-heuristic AOMDV-ACOPSO optimization routing protocol in MANET," *Indian J. Comput. Sci. Eng.*, vol. 13, no. 4, 2022. doi: 10.21817/indjcse/2022/v13i4/221304050

[18] B. Sun, M. Lu, K. Xiao, Y. Song, and C. Gui, "An energy entropy-based minimum power cost multipath routing in MANET," *Int. J. Grid Distrib. Comput.*, vol. 9, no. 2, 2016. doi: 10.14257/ijgdc.2016.9.2.15