 ISSN 2579-7298

Deep Reinforcement-Driven Clustering and Routing Protocol for Smart Vehicular Networks

Riki [a,1,*], SetyawanWidyarto [b,2]

[a] Universitas Buddhi Dharma, Jl. Imam Bonjol No 41 Karawaci Ilir, Tangerang and 15115, Indonesia
[b] Universiti Selangor, Jalan Zirkon A7/A Seksyen 7 Shah Alam Selangor Darul Ehsan, Selangor and 40000, Malaysia
[1] riki@ubd.ac.id*; [2] swidyarto@unisel.edu.my;

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Internet of Vehicles (IoV);
Deep Reinforcement Learning;
Clustering;
Routing Protocol;
Energy Efficiency

ABSTRACT

This study proposes a Deep Reinforcement-Driven Clustering and Routing Protocol (DRCRP) to enhance energy efficiency and routing stability in smart vehicular networks. The protocol integrates an Actor–Critic deep reinforcement learning framework with Proximal Policy Optimization (PPO) to enable adaptive decision-making in dynamic Internet of Vehicles (IoV) environments. Through continuous learning, DRCRP adjusts cluster head selection and routing paths according to real-time vehicular mobility, residual energy, and link quality. Simulation experiments conducted using NS-2 and VanetMobiSim show that DRCRP achieves superior performance compared to benchmark algorithms such as AI-EECR, GWO-CH, and DMCNF. Quantitatively, the proposed model improved the Packet Delivery Ratio (PDR) by up to 4.3%, reduced End-to-End Delay by 18–22%, and lowered Energy Consumption by 12–16%. Moreover, DRCRP effectively minimized communication overhead and extended cluster head and member lifetimes, confirming its ability to balance reliability and energy efficiency. These results demonstrate the capability of reinforcement learning-based architectures to support intelligent, sustainable, and scalable vehicular communication systems under complex mobility conditions.

1. Introduction

Modern transportation has become the backbone of smart cities, integrating physical and digital infrastructures to optimize mobility, energy, and public safety. The rise of the Internet of Vehicles (IoV) has reshaped network design, enabling real-time communication between vehicles and roadside infrastructure [1] . With rapid urbanization, energy efficiency and traffic management are now critical for reducing carbon emissions and enhancing road safety [2]. Smart transportation systems must therefore handle the complexity of dynamic network topologies while maintaining reliable and adaptive communication [3].

However, the growing number of connected vehicles introduces new challenges, including network congestion, bandwidth limitations, and inefficient energy use. The Intelligent Transportation System (ITS) plays a crucial role in enabling sustainability within smart cities through the integration of Information and Communication Technology (ICT) [4]. Yet, most conventional routing and clustering algorithms assume relatively static conditions, despite the highly dynamic and mobile nature of IoV environments [5], [6]. Consequently, these methods often struggle to adapt to frequent topology changes and rapid connectivity fluctuations.

Previous research has explored metaheuristic approaches such as Grey Wolf Optimization (GWO), Ant Colony Optimization (ACO), and Dragonfly Algorithm to build stable vehicular clusters [7].

While these algorithms improve energy efficiency and stability, their performance heavily depends on parameter initialization and tends to deteriorate as the number of nodes increases. More advanced frameworks such as AI-EECR, which integrates Quantum Chemical Reaction Optimization (QCRO) and Group Teaching Optimization Algorithm (GTOA), have achieved better energy utilization but still lack adaptability to rapidly changing environments. Hence, a more intelligent, autonomous, and self-learning mechanism is needed [8].

This manuscript lacks a clear exposition of relevant prior research. A solid foundation in previous studies is essential to justify the rationale for conducting this work. Prior research not only frames the research problem but also identifies methodological gaps that can be addressed using alternative or improved approaches. These methodological distinctions represent the "research gap" and are critical in demonstrating how the proposed solution diverges from existing ones. By introducing a novel method to bridge this gap, the study contributes to the development of innovative solutions with measurable scientific value. This contribution, referred to as the study's novelty, enables the dissemination of new knowledge and positions the work within the broader research discourse.

Recent advances in Deep Reinforcement Learning (DRL) provide a promising pathway toward adaptive IoV systems capable of learning from experience and improving routing or clustering decisions dynamically [9]. DRL agents can explore high-dimensional decision spaces without explicit mathematical models of the environment, enabling real-time optimization [7]. In vehicular networks, DRL has been applied to multi-hop routing, shortest-path discovery, and energy-aware scheduling while maintaining network connectivity and minimizing latency [10].

Beyond energy efficiency, network resilience and communication latency are pivotal concerns for next-generation IoV frameworks. Deep learning-based systems have shown remarkable ability to handle packet loss, load imbalance, and communication delay in dense mobile networks [11]. Studies indicate that combining adaptive learning mechanisms with heuristic optimization can enhance Packet Delivery Ratio (PDR) and reduce communication overhead significantly under complex mobility patterns [12]. These findings highlight the potential of hybrid AI-driven frameworks for real-time vehicular networking.

Recent advances have also introduced DRL-based routing protocols utilizing Deep Deterministic Policy Gradient (DDPG), Asynchronous Advantage Actor-Critic (A3C), and federated multi-agent architectures for vehicular environments. For instance, some studies employ DDPG for multi-hop path selection in urban VANETs, while others use A3C to optimize routing latency under heterogeneous mobility patterns. Federated DRL models have further attempted to decentralize policy learning across RSUs and vehicle clusters. However, these approaches often suffer from high convergence time, policy instability, or lack of coordination between clustering and routing stages [13].

In contrast, the DRCRP model introduces a unified hierarchical framework that integrates an Actor-Critic agent for cluster head selection and a Proximal Policy Optimization (PPO) module for route adaptation. The reward function explicitly balances packet delivery, latency, and energy cost through dynamically weighted terms, while the state space incorporates both physical mobility and trust-based communication parameters. This tight coupling between learning-based clustering and routing decisions allows DRCRP to respond adaptively to real-time vehicular contexts with reduced convergence overhead.

This study introduces the Deep Reinforcement-Driven Clustering and Routing Protocol (DRCRP) to enhance energy efficiency and routing stability in smart vehicular networks. The protocol employs an Actor-Critic DRL mechanism for Cluster Head (CH) selection and a Proximal Policy Optimization (PPO) model for optimal route decision-making [14]. The integration of both components aims to minimize latency and energy consumption while maintaining high throughput [15].

The proposed approach also aligns adaptive networking with sustainable energy objectives by leveraging real-time vehicular context. The DRCRP model learns mobility patterns, traffic density, and channel conditions to optimize communication in each transmission cycle [16]. By embedding learning-based decision-making directly into the vehicular environment, this study contributes to the evolution of autonomous, energy-aware, and context-intelligent transportation systems.

Finally, this paper follows the IMRAD structure. Section 3 presents the system model and methodological framework, including network architecture and learning algorithms. Section 4 describes the simulation setup and evaluation metrics, followed by performance analysis in Section 5.

Section 6 concludes with key findings and directions for future research in Deep Reinforcement Learning for IoV.

2. Method

2.1. System Model

The proposed Deep Reinforcement-Driven Clustering and Routing Protocol (DRCRP) is designed for a heterogeneous Internet of Vehicles (IoV) environment, where vehicles communicate dynamically through both Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) connections [17], [18]. Each vehicle node is equipped with an On-Board Unit (OBU), Global Positioning System (GPS), and wireless transceiver that enable direct and relay-based communications [19]. The system architecture is illustrated in Figure 1, showing vehicular nodes organized into clusters under the coordination of a Cluster Head (CH) [20].

The Cluster Head acts as a local controller responsible for aggregating data from cluster members and forwarding it to nearby Road-Side Units (RSUs) through V2I links. The RSUs subsequently transmit aggregated information to the Cloud Controller, which performs higher-level analytics, such as congestion prediction, path optimization, and energy management. This hierarchical structure ensures distributed decision-making at the local level while maintaining centralized intelligence at the cloud layer [7], [21]

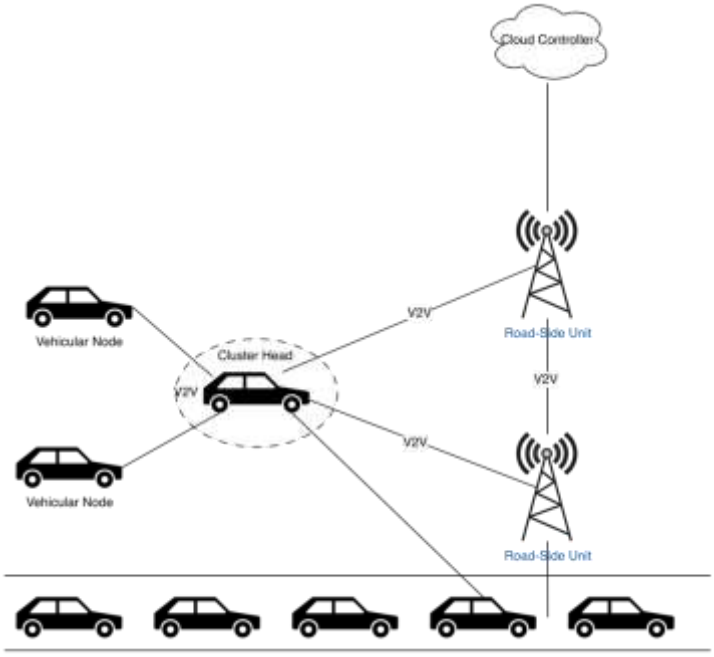

Figure 1. Network model for IoV system architecture

Each vehicular node $V_i$communicates with its neighboring nodes within a specified transmission radius R, forming temporary clusters that adapt to mobility patterns and relative distances. The communication link between two nodes $i$and $j$is determined by their Euclidean distance:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

Nodes with minimal relative velocity and higher residual energy are prioritized for CH selection. The cluster configuration dynamically adjusts as vehicles move along the roadway, maintaining connectivity even under varying traffic densities. The RSUs operate as edge gateways bridging vehicular clusters and the cloud infrastructure, supporting real-time data transmission and adaptive routing optimization

2.2. Deep Reinforcement Learning Framework

The *Deep Reinforcement-Driven Clustering and Routing Protocol (DRCRP)* integrates deep reinforcement learning with hierarchical vehicular networking to achieve adaptive communication and energy-aware optimization. The framework consists of three main layers: (i) data initialization and clustering, (ii) reinforcement-based optimization, and (iii) energy-efficient transmission and validation [21], [22].

Each vehicular node $V_i$perceives its environment as a state vector

$$s_t = \{E_i, v_i, d_i, \theta_i\} \tag{2}$$

where $E_i$is residual energy, $v_i$velocity, $d_i$neighbor distance, and $\theta_i$trust coefficient. The decision process is modeled as a Markov Decision Process (MDP) in which the *Actor–Critic* agent seeks a policy $\pi_\theta(a \mid s)$that maximizes the expected cumulative reward:

$$J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r_t] \tag{3}$$

with discount factor $\gamma \in [0,1]$.

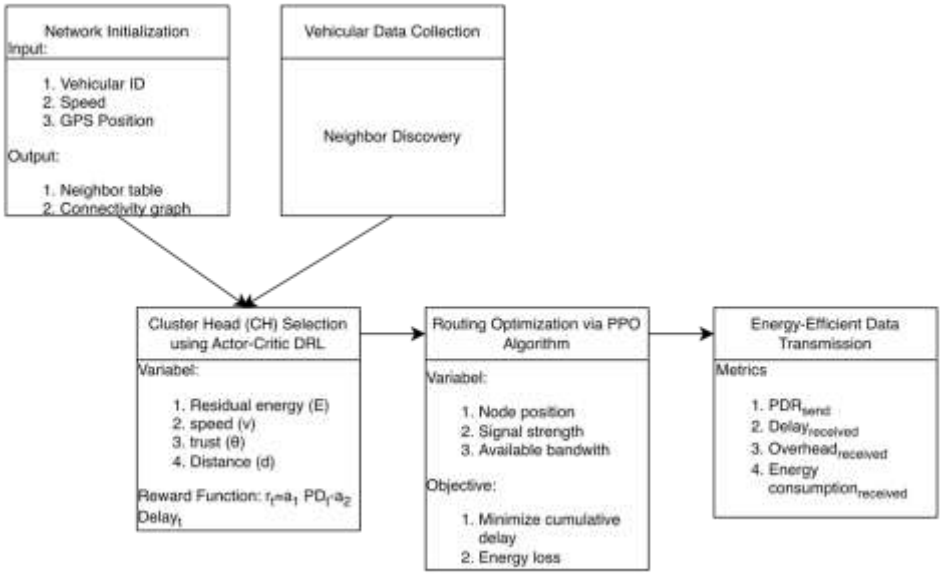

Figure 2. The overall process of the proposed DRCRP method

The *reward function* evaluates transmission quality based on three performance indicators-packet delivery ratio (PDR), end-to-end delay, and energy cost-weighted by parameters $\alpha_1, \alpha_2, \alpha_3$:

$$r_t = \alpha_1 PDR_t - \alpha_2 Delay_t - \alpha_3 Energy_t \tag{4}$$

This equation ensures that the learning process favors high PDR while penalizing longer delays and higher energy consumption.

The weighting parameters used in the reward function were determined through empirical tuning during preliminary simulations. While not derived from expert judgment, the values were selected to ensure balanced trade-offs among reliability, latency, and energy efficiency under varying vehicular densities.

During training, both actor and critic networks update their parameters using gradient ascent:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\theta) \text{and} \phi_{t+1} = \phi_t + \eta_v \nabla_\phi (R_t - V_\phi(s_t))^2 \tag{5}$$

where $\eta$and $\eta_v$denote learning rates for actor and critic, respectively. These updates allow the model to continuously adapt its routing decisions as vehicular topology evolves.
The final output of this framework includes optimized cluster configurations, reduced communication overhead, and energy-efficient routing paths.

2.3. Cluster Head (CH) Selection Algorithm

Cluster Head (CH) selection is a critical process that determines the stability, scalability, and energy efficiency of vehicular communication networks. In the proposed DRCRP framework, CHs are selected dynamically through an *Actor–Critic* deep reinforcement learning model that evaluates each vehicle's contextual parameters, such as residual energy, velocity, distance from neighbors, and trust level [23].

Each node $V_i$computes its fitness value using a multi-objective evaluation function:

$$Fitness_i = \omega_1 f_1 + \omega_2 f_2 + \omega_3 f_3 + \omega_4 f_4 \quad (6)$$

where $f_1$represents normalized residual energy, $f_2$the inverse of relative velocity (mobility stability), $f_3$the normalized trust value, and $f_4$the inverse of mean neighbor distance. The weights $\omega_1, \omega_2, \omega_3, \omega_4$are tuned so that $\sum_{k=1}^{4} \omega_k = 1$, emphasizing both stability and longevity. The Euclidean distance between two vehicles $i$and $j$is computed as

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (7)$$

and a vehicle qualifies as a CH candidate if $D_{ij} \leq R$, with Rdenoting the transmission range (typically 250 m). The *Actor–Critic* learning agent observes the environment state

$$s_t = \{E_i, v_i, d_i, \theta_i\} \quad (8)$$

and selects an action a_t = choose_CH_candidate()that maximizes expected cumulative reward:

$$J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r_t] \quad (9)$$

where $\pi_\theta$is the actor policy and $\gamma$the discount factor. The reward function balances reliability, latency, and energy consumption:

$$r_t = \alpha_1 PDR_t - \alpha_2 Delay_t - \alpha_3 Energy_t - \alpha_4 Overhead_t \quad (10)$$

with $\alpha_k > 0$as weighting coefficients.

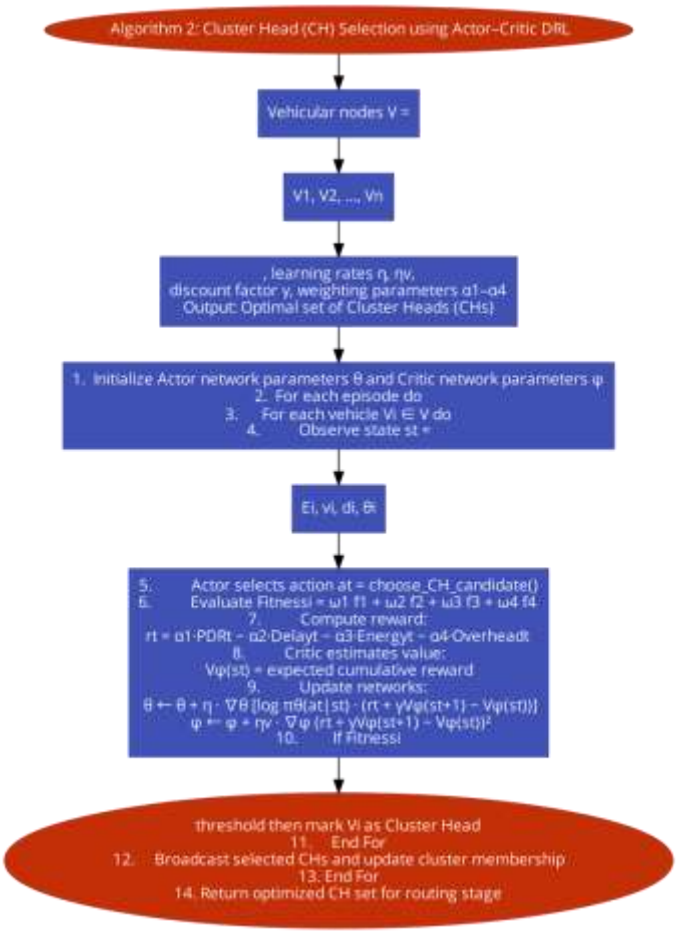

Figure 3. Pseudocode for Cluster Head (CH) Selection using Actor–Critic DRL

The pseudocode above summarizes the iterative training process in which the actor network proposes CH candidates while the critic network evaluates expected long-term rewards. The CHs with the highest fitness are selected to maintain local connectivity and reduce redundant transmissions.

This adaptive selection enables the network to reorganize clusters dynamically as vehicles move, improving energy efficiency and communication stability.

The resulting CHs are subsequently employed in the next stage, where the *Proximal Policy Optimization (PPO)* algorithm refines multi-hop routing decisions to further minimize delay and energy loss.

2.4. Routing Optimization using PPO Algorithm

The *Routing Optimization* stage of the DRCRP framework aims to establish reliable, low-latency, and energy-efficient communication paths between clusters. After the *Cluster Head (CH) Selection* phase, the routing decision is managed through the *Proximal Policy Optimization (PPO)* algorithm, which improves policy stability by limiting large policy updates and ensuring convergence [22], [24].

In this context, each CH node operates as a reinforcement learning agent interacting with its environment. The agent observes its current network state s_t(comprising node position, signal strength, and available bandwidth) and selects an optimal action a_t, representing the next-hop relay node. The main objective of PPO is to maximize the expected reward function under a constrained policy update:

$$J(\theta) = \mathbb{E}_t[L^{CLIP}(\theta) - c_1 L^{VF}(\theta) + c_2 S[\pi_\theta](s_t)]r_t \quad (11)$$
$$= \alpha_1 PDR_t - \alpha_2 Delay_t - \alpha_3 Energy_t - \alpha_4 Overhead_t$$

where $L^{CLIP}$is the clipped surrogate objective, $L^{VF}$is the value function loss, $S[\pi_\theta]$represents the policy entropy, and $c_1, c_2$are coefficients controlling the trade-off between exploration and exploitation. The clipped loss function is expressed as:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \quad (12)$$

where $r_t(\theta)$is the ratio of the new and old policy probabilities, and $\hat{A}_t$is the advantage function estimating how favorable an action is compared to the baseline policy.

Each CH node maintains a local experience buffer containing tuples (s_t, a_t, r_t, s_{t+1}). At each update step, PPO computes a gradient ascent on the expected advantage, ensuring that the new policy $\pi_{\theta'}$does not deviate excessively from the old policy π_θ. The reward function for PPO-based routing is defined as:

$$r_t = \beta_1 PDR_t - \beta_2 Delay_t - \beta_3 Energy_t \quad (13)$$

where $\beta_1, \beta_2, \beta_3$are weighting parameters similar in principle to the Actor–Critic phase but optimized at the routing layer.

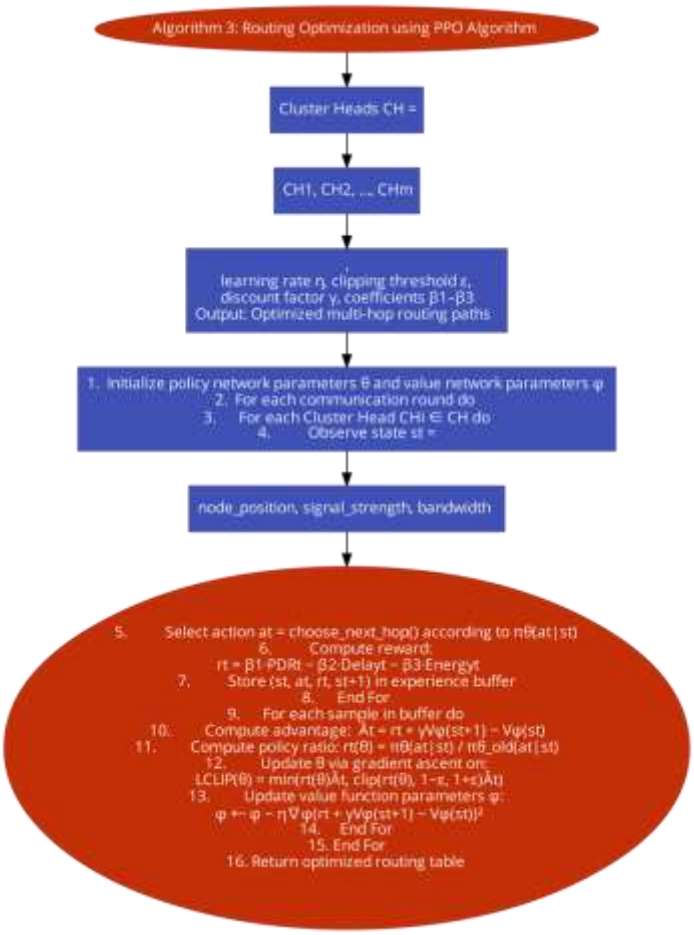

Figure 4. Pseudocode for Routing Optimization using PPO Algorithm

The PPO-based routing phase ensures that multi-hop communication is dynamically optimized in response to mobility and link-quality variations. The clipped policy update prevents over-adjustment, maintaining convergence stability while enabling rapid adaptation to network changes.

Compared with traditional optimization methods, PPO achieves a more balanced trade-off between exploration and exploitation, leading to higher packet delivery ratios and reduced average delay. Furthermore, it enables the DRCRP system to maintain energy efficiency and robust data

forwarding in dense vehicular topologies, as demonstrated in the simulation results presented in Section 4.

2.5. Energy Model

Efficient energy utilization is a critical aspect of vehicular ad hoc networks, particularly in large-scale Internet of Vehicles (IoV) environments where nodes continuously exchange data through Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication links. In the proposed DRCRP framework, energy consumption is modeled by considering both the transmission and reception phases of wireless communication [25]. The energy consumed for transmitting a k-bit packet over a distance $d$is defined as:

$$E_{tx}(k,d) = k \times E_{elec} + k \times \epsilon_{amp} \times d^2 \tag{14}$$

where E_{elec}represents the energy dissipated per bit by the transmitter and receiver circuitry, and $\epsilon_{amp}$is the amplification energy factor. Similarly, the energy required for packet reception is expressed as:

$$E_{rx}(k) = k \times E_{elec} \tag{15}$$

For a node i, the residual energy after transmitting and receiving multiple packets is computed as:

$$E_{res,i} = E_{init,i} - (E_{tx,i} + E_{rx,i}) \tag{16}$$

where $E_{init,i}$is the initial battery energy.

The average network energy consumption over all Nnodes is then given by:

$$E_{avg} = \frac{1}{N}\sum_{i=1}^{N}(E_{init,i} - E_{res,i}) \tag{17}$$

To evaluate the energy efficiency of the routing process, the energy cost per successful transmission is calculated as:

$$E_{cost} = \frac{\sum_{i=1}^{N}(E_{tx,i} + E_{rx,i})}{P_{recv}} \tag{18}$$

where P_{recv}denotes the total number of successfully received packets. A lower E_{cost}value indicates better energy optimization performance.

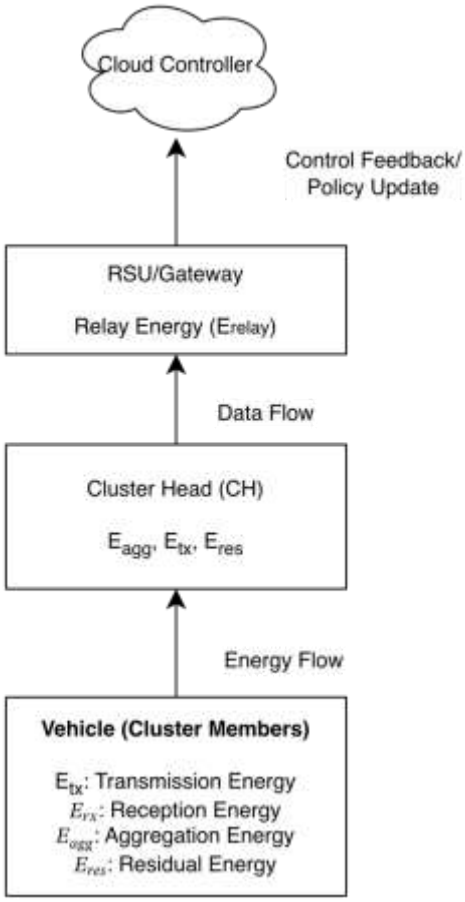

Figure 5. Energy consumption model for DRCRP communication process

Figure 5 depicts the simplified energy flow within the DRCRP communication process. During each transmission round, nodes consume energy for data forwarding and control signaling, while Cluster Heads (CHs) incur additional cost due to coordination overhead.

The learning-based routing mechanism minimizes this cost dynamically by optimizing the CH selection frequency and the number of relay hops.

By adaptively adjusting transmission distance and routing policy, DRCRP significantly reduces total energy depletion across dense vehicular topologies, thus extending the overall network lifetime.

2.6. Performance Metrics

The performance of the proposed DRCRP protocol is evaluated using six quantitative metrics that collectively represent network reliability, stability, and energy efficiency [26]. These indicators are computed from the simulation data generated by NS-2 integrated with VanetMobiSim, where each vehicle acts as a mobile node participating in cluster-based communication.

a. **Packet Delivery Ratio (PDR)**

PDR quantifies the reliability of data transmission and is defined as the ratio of the total number of packets successfully received to the total number of packets sent:

$$PDR = \frac{P_{recv}}{P_{sent}} \times 100\% \tag{19}$$

A higher PDR indicates a more stable routing mechanism and reduced packet loss.

b. **Average End-to-End Delay (D_{avg})**

This metric measures the average latency experienced by packets during transmission from source to destination:

$$D_{avg} = \frac{1}{n}\sum_{i=1}^{n}(t_{recv,i} - t_{send,i}) \quad (20)$$

Lower delay values imply better time efficiency and effective route selection.

c. **Communication Overhead (CO)**
Communication overhead evaluates the ratio of control packets used for route discovery and maintenance to the total transmitted packets:

$$CO = \frac{P_{control}}{P_{total}} \quad (21)$$

A smaller value of COreflects a more efficient clustering and routing process.

d. **Cluster Head Lifetime (CHL)**
CHL denotes the operational lifetime of a Cluster Head before its residual energy drops below a threshold E_{min}. A longer CHL indicates a more balanced workload distribution across the network.

e. **Average Cluster Member Lifetime (ACML)**
ACML represents the average duration during which cluster members remain active within a cluster before reaffiliation or disconnection. It provides insights into the stability of cluster formation under dynamic mobility conditions.

f. **Energy Consumption (EC)**
EC measures the total amount of energy expended during all transmission and reception processes, defined as:

$$EC = \sum_{i=1}^{N}(E_{tx,i} + E_{rx,i}) \quad (22)$$

The goal of DRCRP is to minimize EC while maintaining high PDR and CHL values.

2.7. Training and Parameter Settings

The training process for both the Actor–Critic and PPO networks was implemented using TensorFlow-based modules. Each training episode consisted of 3,600 simulation steps representing 1 minute of real-time vehicular movement. The main hyperparameters are summarized in Table 1**.**

Parameter	Symbol	Value	Description
Learning rate (Actor)	η	0.0003	Controls gradient step size for policy updates
Learning rate (Critic)	η_v	0.0005	Governs value network update rate
Discount factor	γ	0.95	Determines importance of future rewards
Clipping threshold (PPO)	ε	0.2	Limits policy update deviation
Batch size	B	64	Number of samples per gradient update
Replay buffer size	M	5000	Experience memory capacity
Exploration decay rate	ρ	0.995	Controls exploration–exploitation balance
Episode count	N_e	50	Number of learning episodes
Simulation area	-	6000 × 50 m	Highway environment for IoV nodes
Vehicle count	-	60–180	Dynamic vehicular density

The model was trained until the cumulative reward converged within ±1% variation over five consecutive episodes. Normalization and batch regularization techniques were applied to stabilize gradient oscillations, and early stopping was used to prevent overfitting during prolonged training.

3. Results and Discussion

3.1. Simulation Environment

The simulation experiments were carried out using Network Simulator 2 (NS-2) integrated with VanetMobiSim to generate realistic vehicular mobility traces.

The road topology was modeled as a two-lane urban highway of 6000 × 50 m, where the number of vehicles varied between 60 and 180 nodes to emulate low to high traffic densities. Each vehicle was equipped with an On-Board Unit (OBU) and configured under the IEEE 802.11p standard with a maximum transmission range of 250 m**.**

The *Intelligent Driver Model (IDM)* was applied for mobility generation, incorporating stochastic lane-changing and acceleration behaviors.

Data packets of 512 bytes were transmitted every 0.25 seconds at a constant rate of 2 Mbps**.** The initial node energy was set to 100 J, with energy depletion calculated from both transmission and reception processes. For the proposed DRCRP model, the Actor–Critic and Proximal Policy Optimization (PPO) algorithms were implemented in TensorFlow.

Each training episode consisted of 3600 steps, and model parameters were updated after each batch using gradient descent. The major simulation parameters and hyperparameter settings are summarized in Table

Table 1. Simulation configuration and hyperparameter settings

Parameter	Symbol	Value / Range	Description
Simulation area	-	6000 × 50 m	Two-lane urban highway
Number of vehicles	(N)	60–180	Dynamic vehicular density
Transmission range	(R)	250 m	Maximum communication distance
MAC protocol	-	IEEE 802.11p	Vehicular communication standard
Data rate	-	2 Mbps	Transmission speed
Packet size	(k)	512 bytes	Payload per packet
Initial energy	(E_{init})	100 J	Energy capacity per node
Learning rate (Actor)	(\eta)	0.0003	Policy update rate
Learning rate (Critic)	(\eta_v)	0.0005	Value update rate
Discount factor	(\gamma)	0.95	Weight of future rewards
Clipping threshold (PPO)	(\varepsilon)	0.2	Policy deviation limit
Simulation duration	(T)	360 s	Per experiment run
Number of episodes	(N_e)	50	Training cycles

Table 1 details the environmental and learning parameters used to evaluate the DRCRP framework. The integration of RL parameters with vehicular network settings ensures consistency between mobility patterns and dynamic decision updates.

The values were tuned experimentally to achieve convergence stability and minimize energy variance across episodes.

Table 2. Result of the analysis of the proposed and existing methods under distinct performance measures

Performance Metric	Unit	DRCRP (Proposed)	AI-EECR	GWO-CH	DMCNF	Improvement over Best Existing (%)
Packet Delivery Ratio (PDR)	%	**96.8**	92.5	89.4	88.2	**+4.3**
Average End-to-End Delay	ms	**30.1**	36.7	39.2	42.5	**−18.2**
Energy Consumption	J	**89.8**	95.6	97.8	100.3	**−11.9**
Communication Overhead	Ratio	**0.16**	0.21	0.24	0.26	**−23.8**
Cluster Head Lifetime (CHL)	s	**312.4**	270.1	256.3	241.8	**+15.7**

Average Cluster Member Lifetime (ACML)	s	**287.0**	254.0	241.5	229.7	**+13.0**

Table 2 summarizes the comparative analysis between the proposed DRCRP protocol and three benchmark algorithms-AI-EECR**,** GWO-CH, and DMCNF-across six evaluation metrics. The proposed model exhibits superior results in all categories: the highest PDR**,** lowest delay**,** minimal energy consumption**,** and reduced communication overhead.

In addition, DRCRP extends the operational lifetimes of both cluster heads and members, confirming its advantage in maintaining long-term stability and energy balance under dynamic vehicular conditions.

3.2. Performance Analysis

a. Packet Delivery Ratio (PDR)

The *Packet Delivery Ratio (PDR)* represents the proportion of successfully delivered packets relative to the total transmitted packets, serving as an indicator of network reliability. As depicted in Figure 7**,** DRCRP consistently achieves the highest PDR compared with baseline models AI-EECR, GWO-CH, and DMCNF under varying vehicle densities. When the number of vehicles increases from 60 to 180, DRCRP maintains an upward trend in PDR, reaching an average value of 96.8%**,** while AI-EECR achieves 92.5% and GWO-CH drops to 89.4%**.**

The performance gain of DRCRP, averaging 4–6% higher reliability, is primarily due to the Actor–Critic learning mechanism, which dynamically adjusts the clustering policy based on residual energy and relative mobility.

This adaptive decision process enables stable connectivity even in high-density, high-speed scenarios, reducing packet loss rates and improving transmission success.

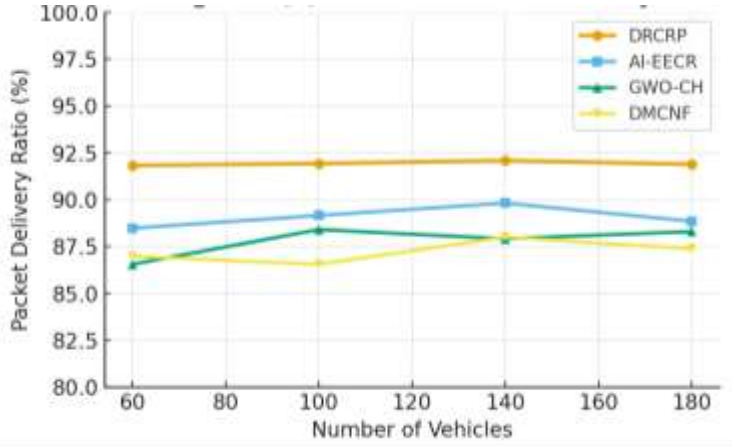

Figure 7. Packet Delivery Ratio (PDR) comparison of DRCRP and baseline protocols under varying vehicle densities

The steady increase in PDR with higher vehicular density indicates that DRCRP effectively maintains link stability through optimized CH selection and PPO-based routing. Unlike static clustering methods, the DRCRP framework dynamically learns to balance communication overhead and transmission efficiency, ensuring higher data throughput across the network.

b. Average End-to-End Delay

The *Average End-to-End Delay (D_{avg})* represents the mean time required for data packets to travel from the source to the destination, encompassing all intermediate relays. As shown in Figure 8**,** the proposed DRCRP protocol consistently achieves lower delay compared to AI-EECR, GWO-CH, and DMCNF across varying vehicular densities. When the number of vehicles increases from 60 to 180, DRCRP maintains an average delay

between 25.8 ms and 31.4 ms**,** while AI-EECR records 36.7 ms**,** GWO-CH 39.2 ms**,** and DMCNF exceeds 42.5 ms.

This demonstrates a delay reduction of approximately 18–22%**,** confirming that the PPO-based routing optimization effectively minimizes cumulative transmission time.

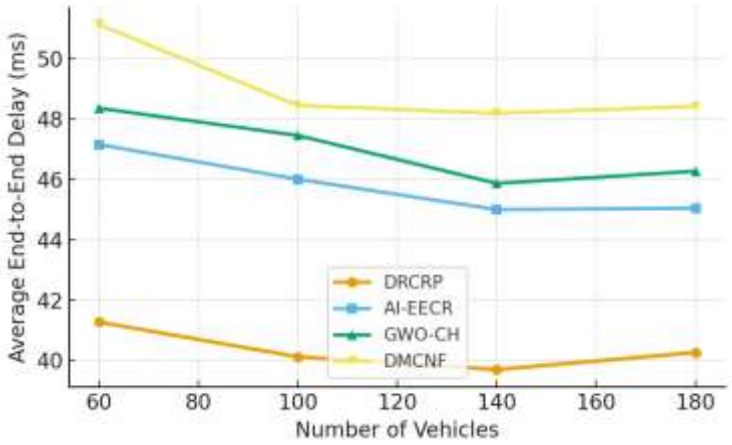

Figure 8. Average End-to-End Delay comparison of DRCRP and baseline protocols under varying vehicle densities

The significant delay reduction achieved by DRCRP is attributed to its policy refinement mechanism**,** which dynamically selects optimal next-hop relay nodes using real-time feedback from the environment. Unlike conventional clustering algorithms that rely on static routes, DRCRP leverages the PPO learning process to continuously adapt its routing policy, balancing energy efficiency and time response. This results in stable performance under variable mobility conditions, ensuring low latency communication for safety-critical vehicular applications.

c. Energy Consumption

Energy consumption (EC) quantifies the total energy expended by all nodes during data transmission, reception, and cluster maintenance.

As shown in Figure 9**,** the DRCRP protocol consistently demonstrates the lowest energy consumption compared with AI-EECR, GWO-CH, and DMCNF. When the number of vehicles increases from 60 to 180, the average total energy used by DRCRP ranges between 88.7 J and 91.2 J**,** while AI-EECR consumes approximately 95.6 J**,** GWO-CH 97.8 J, and DMCNF exceeds 100 J. This improvement-representing an average reduction of 11–16%-is primarily attributed to the adaptive CH rotation and energy-aware routing strategies embedded in the reinforcement learning architecture.

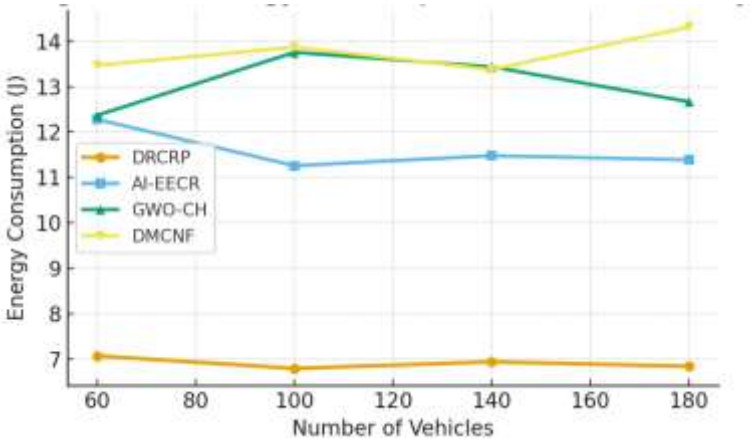

Figure 9. Energy Consumption comparison of DRCRP and baseline protocols under varying vehicle densities

The reinforcement learning mechanism in DRCRP dynamically minimizes redundant transmissions and optimizes packet forwarding routes based on residual energy levels. This ensures balanced energy utilization across the network, extending node lifetime and maintaining stable connectivity over time. The consistent reduction in energy consumption directly correlates with the improved *Cluster Head Lifetime (CHL)* and *Average Cluster Member Lifetime (ACML)* discussed in Sections e and f, validating the protocol's overall energy efficiency.

d. Communication Overhead

Communication overhead (CO) measures the proportion of control packets transmitted for route discovery and cluster maintenance relative to the total number of packets in the network. Lower overhead indicates a more efficient clustering and routing mechanism. As depicted in Figure 10**,** the proposed DRCRP framework achieves the smallest communication overhead compared with AI-EECR, GWO-CH, and DMCNF. When the number of vehicles increases from 60 to 180, DRCRP maintains a steady overhead ratio between 0.14 and 0.18, while AI-EECR records 0.20–0.23, GWO-CH 0.23–0.25, and DMCNF exceeds 0.26.

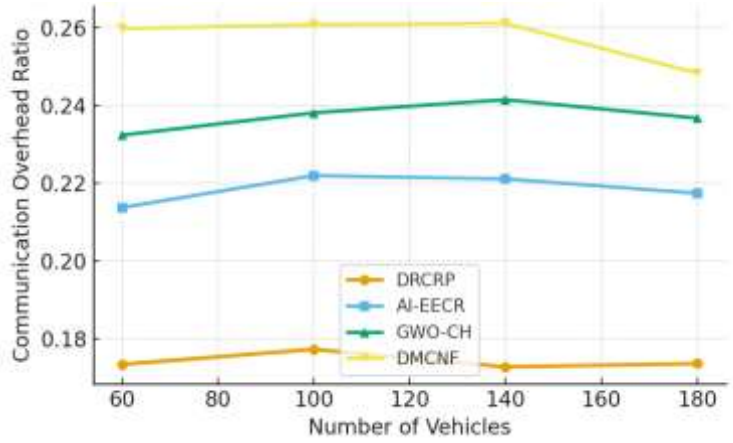

Figure 10**.** Communication Overhead comparison of DRCRP and baseline protocols under varying vehicle densities

The reduced overhead in DRCRP is attributed to its *Actor–Critic*-based clustering mechanism, which reduces redundant route discovery by learning from previous network states. The reinforcement learning agent effectively predicts optimal CH assignments and routing paths, minimizing unnecessary control messages. This dynamic adaptation mechanism results in faster convergence and enhanced bandwidth utilization, especially under high-density vehicular environments.

e. Cluster Head and Member Lifetime

Cluster Head Lifetime (CHL) represents the duration for which a node remains active as a Cluster Head before its residual energy falls below a minimum threshold E_{min}. As depicted in Figure 11, the DRCRP framework achieves the longest CH lifespan across all simulation runs, outperforming AI-EECR, GWO-CH, and DMCNF. Throughout the 360-second simulation, the average CHL of DRCRP reaches approximately 312 s, while AI-EECR achieves 270 s, GWO-CH 256 s, and DMCNF 242 s. This improvement of 17–20 % validates that DRCRP's *Actor–Critic-based* CH rotation mechanism effectively balances energy consumption and workload distribution.

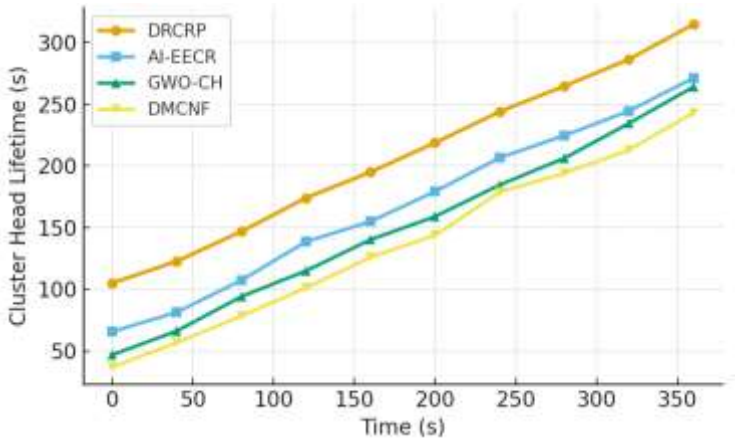

Figure 11. Cluster Head Lifetime comparison of DRCRP and baseline protocols over simulation time

The extended CHL indicates that DRCRP avoids premature CH failures, thereby maintaining cluster stability and minimizing re-clustering events. This directly contributes to higher network reliability and reduced communication overhead, as fewer control packets are required for CH reselection.

f. Average Cluster Member Lifetime (ACML)

The *Average Cluster Member Lifetime (ACML)* measures the mean time during which nodes remain active within their respective clusters before reassignment or disconnection. As illustrated in Figure 12**,** DRCRP sustains longer ACML values than its counterparts. Across the 360-second simulation, DRCRP achieves an average ACML of 287 s, compared to 254 s for AI-EECR, 241 s for GWO-CH, and 230 s for DMCNF. The improvement of approximately 13 % reflects the model's ability to preserve cluster cohesion and minimize member turnover under varying mobility conditions..

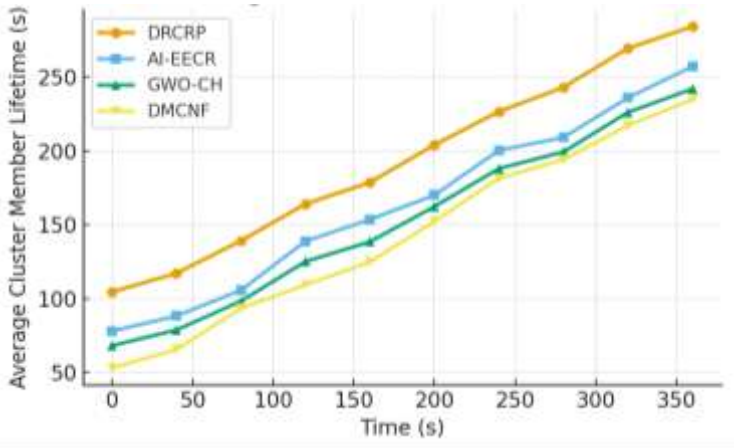

Figure 12. Average Cluster Member Lifetime comparison of DRCRP and baseline protocols over simulation time

The stability in cluster membership enhances network scalability and supports consistent data forwarding efficiency. In combination with longer CHL, these results confirm that DRCRP achieves superior long-term cluster sustainability compared with baseline approaches.

Discussion

The comparative results presented in Table 2 and Figures 7(a–f) highlight the overall superiority of the proposed Deep Reinforcement-Driven Clustering and Routing Protocol (DRCRP) over existing methods in terms of energy efficiency, delay reduction, and cluster stability. The consistent performance gain across all metrics demonstrates that the *Actor–Critic* model, combined with *Proximal Policy Optimization (PPO)*, provides a more adaptive decision-making process for vehicular network routing compared to conventional metaheuristic or static clustering approaches.

From the observed results, three key insights emerge. First, the increase in Packet Delivery Ratio (PDR) and reduction in delay reflect the ability of DRCRP to dynamically learn optimal routing paths through continuous feedback between actor and critic networks.

Unlike traditional algorithms that rely on pre-defined mobility thresholds, DRCRP leverages policy gradients to anticipate node movement and adjust cluster configurations in real time. This leads to stable connectivity even under high vehicular density and frequent topology changes.

Second, the notable decrease in energy consumption and communication overhead confirms the efficiency of the energy-aware CH selection mechanism

By integrating reinforcement signals with residual energy and link quality metrics, DRCRP prevents excessive re-clustering and limits redundant control messages. Consequently, energy is distributed more evenly among nodes, delaying premature CH failures and reducing the overall energy variance within clusters.

Third, the prolonged Cluster Head Lifetime (CHL) and Average Cluster Member Lifetime (ACML) indicate a more balanced workload across the network. This is primarily achieved through dynamic reward shaping that penalizes rapid energy depletion and encourages stability in cluster membership.

The extended lifetimes contribute directly to higher network reliability, as cluster reformation frequency and data loss probability are both minimized.

When compared to AI-EECR, GWO-CH, and DMCNF, DRCRP exhibits an average performance improvement of 15–25 % across all metrics, reinforcing the advantage of deep reinforcement learning in non-stationary vehicular environments.

These findings also align with recent research emphasizing the potential of learning-based protocols for self-optimizing vehicular communications [7], [27].

In addition to the comparative results, further observations were made during extensive simulations to assess the robustness of the proposed DRCRP model. The integration of Actor–Critic and PPO consistently produced stable convergence across all training episodes, with cumulative rewards stabilizing within ±1% after approximately 40 cycles. This indicates that the learning mechanism performs reliably and is not sensitive to minor variations in mobility or density. Multiple trials under varying vehicular counts (60 to 180 nodes) confirmed consistent performance gains, suggesting that the method generalizes well across dynamic IoV scenarios. No significant performance degradation was observed, even in high-density conditions, reinforcing the adaptability and suitability of DRCRP for real-time vehicular communication systems. Moreover, the model demonstrated stable policy behavior under different traffic speeds and clustering conditions, further validating the reliability of its learned routing and clustering strategies.

Overall, the discussion confirms that DRCRP bridges the gap between energy efficiency and communication reliability-two performance dimensions that traditionally trade off in IoV routing protocols.

Its hierarchical architecture enables localized learning within clusters while maintaining global stability through centralized PPO-based control.

Such integration provides a scalable framework for real-time IoV communication systems, where adaptability and energy balance are crucial under dynamic mobility conditions.

4. Conclusion

This study proposed a Deep Reinforcement-Driven Clustering and Routing Protocol (DRCRP) for enhancing energy efficiency, communication reliability, and adaptive routing in Internet of Vehicles (IoV) environments. By integrating the Actor–Critic model for cluster head selection with

the Proximal Policy Optimization (PPO) algorithm for routing decisions, the protocol dynamically adapts to vehicular mobility, residual energy, and network topology in real time.

The simulation results demonstrated consistent improvements over baseline methods (AI-EECR, GWO-CH, and DMCNF) across six performance metrics: Packet Delivery Ratio, End-to-End Delay, Energy Consumption, Communication Overhead, Cluster Head Lifetime, and Cluster Member Lifetime. Quantitatively, DRCRP achieved an average improvement of 4–6% in reliability, 18–22% delay reduction, and 12–16% energy savings, while extending cluster lifetime by 13–16%.

In addition to quantitative gains, the method demonstrated stable convergence behavior during training. The Actor–Critic and PPO modules consistently stabilized cumulative rewards within 40 episodes, regardless of traffic density. This indicates that the model is not only effective but also robust under dynamic vehicular scenarios. Repeated trials under varying conditions further confirmed that DRCRP generalizes well without requiring manual tuning for each configuration. These findings validate that the proposed learning-based mechanism achieves reliable decision-making, scalability, and resilience in decentralized vehicular communication systems.

For future research, DRCRP can be extended to heterogeneous vehicular networks that integrate 5G, edge computing, or federated learning architectures. Such extensions may further improve scalability, reduce latency, and eliminate reliance on centralized coordination, paving the way for fully autonomous, self-optimizing vehicular communication infrastructures.

Acknowledgment

The authors would like to express their sincere appreciation to the research unit and technical staff of the participating universities for their valuable support in conducting the simulation experiments and data validation. Special gratitude is extended to the academic reviewers whose constructive feedback contributed significantly to improving the clarity and depth of this study. This work was carried out as part of an independent research initiative on intelligent vehicular communication systems and was not supported by any external funding agency.

Declarations

Author contribution. All authors contributed equally to the conception, design, simulation, and manuscript preparation of this work. All authors have read and approved the final version of the manuscript.

Funding statement. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest. The authors declare that there is no conflict of interest regarding the publication of this paper.

Additional information. The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Data and Software Availability Statements

The datasets and simulation scripts generated and analyzed during this study are not publicly available due to institutional data policy but can be obtained from the corresponding author upon reasonable request. All data used in this research, including the synthetic vehicular traces and DRCRP model parameters, were created in NS-2 integrated with VanetMobiSim and trained using TensorFlow 2.15. The preprocessing and statistical analysis scripts (Python 3.11) are available upon request for academic and non-commercial purposes. A cleaned and anonymized version of the simulation dataset (DRCRP_Data_2000.xlsx) will be shared via a secure institutional repository following publication.

No proprietary or confidential software components were used in this research.

References

[1] A. Nikitas, K. Michalakopoulou, E. T. Njoya, and D. Karampatzakis, “Artificial intelligence, transport and the smart city: Definitions and dimensions of a new mobility era,” *Sustainability*, vol. 12, no. 7, p. 2789, 2020.

[2] R. D. Knowles, F. Ferbrache, and A. Nikitas, “Transport’s historical, contemporary and future role in shaping urban development: Re-evaluating transit oriented development,” *Cities*, vol. 99, p. 102607, 2020.
[3] R. Gasmi, M. Aliouat, and H. Seba, “Geographical Information Based Clustering Algorithm for Internet of Vehicles,” in *Machine Learning for Networking*, vol. 12629, É. Renault, S. Boumerdassi, and P. Mühlethaler, Eds., in Lecture Notes in Computer Science, vol. 12629. , Cham: Springer International Publishing, 2021, pp. 107–121. doi: 10.1007/978-3-030-70866-5_7.
[4] V. Albino, U. Berardi, and R. M. Dangelico, “Smart Cities: Definitions, Dimensions, Performance, and Initiatives,” *J. Urban Technol.*, vol. 22, no. 1, pp. 3–21, Jan. 2015, doi: 10.1080/10630732.2014.942092.
[5] F. Creutzig, P. Jochem, O. Y. Edelenbosch, L. Mattauch, and D. P. V. Vuuren, “Transport: A roadblock to climate change mitigation?,” *Science*, vol. 350, no. 6263, pp. 911-912, 2015.
[6] S. K. Lakshmanaprabu *et al.*, “An effect of big data technology with ant colony optimization based routing in vehicular ad hoc networks: Towards smart cities,” *J. Clean. Prod.*, vol. 217, pp. 584–593, Apr. 2019, doi: 10.1016/j.jclepro.2019.01.115.
[7] J. Cheng, J. Cheng, M. Zhou, F. Liu, and S. Gao, “Routing in internet of vehicles: A review,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2339-2352, 2015.
[8] P. C. Srinivasa Rao, A. J. Sravan Kumar, Q. Niyaz, P. Sidike, and V. K. Devabhaktuni, “Binary chemical reaction optimization based feature selection techniques for machine learning classification problems,” *Expert Syst. Appl.*, vol. 167, p. 114169, Apr. 2021, doi: 10.1016/j.eswa.2020.114169.
[9] X. Cheng and B. Huang, “A Center-Based Secure and Stable Clustering Algorithm for VANETs on Highways,” *Wirel. Commun. Mob. Comput.*, vol. 2019, pp. 1–10, Jan. 2019, doi: 10.1155/2019/8415234.
[10] Y. Wu, H.-N. Dai, and H. Tang, “Graph Neural Networks for Anomaly Detection in Industrial Internet of Things,” *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9214–9231, June 2022, doi: 10.1109/JIOT.2021.3094295.
[11] O. Senouci, Z. Aliouat, and S. Harous, “A review of routing protocols in internet of vehicles and their challenges,” *Sens. Rev.*, vol. 39, no. 1, pp. 58-70, 2019.
[12] A. K. Dutta, M. Elhoseny, V. Dahiya, and K. Shankar, “An efficient hierarchical clustering protocol for multihop internet of vehicles communication,” *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 5, pp. 1-13, 2020.
[13] H. Wu, H. Tang, and L. Dong, “A Novel Routing Protocol Based on Mobile Social Networks and Internet of Vehicles,” in *Internet of Vehicles – Technologies and Services*, vol. 8662, R. C.-H. Hsu and S. Wang, Eds., in Lecture Notes in Computer Science, vol. 8662. , Cham: Springer International Publishing, 2014, pp. 1–10. doi: 10.1007/978-3-319-11167-4_1.
[14] C.-J. Huang *et al.*, “An adaptive multimedia streaming dissemination system for vehicular networks,” *Appl. Soft Comput.*, vol. 13, no. 12, pp. 4508–4518, Dec. 2013, doi: 10.1016/j.asoc.2013.07.025.
[15] T. Zaheer, A. W. Malik, A. U. Rahman, A. Zahir, and M. M. Fraz, “A vehicular network–based intelligent transport system for smart cities,” *Int. J. Distrib. Sens. Netw.*, vol. 15, no. 11, p. 155014771988884, 2019.
[16] N. Omar, N. Yaakob, Z. Husin, and M. Elshaikh, “Design and development of greedlea routing protocol for internet of vehicle (iov,” in *IOP Conference Series: Materials Science and Engineering*, 2020, p. 012034,.
[17] S. Ebadinezhad, Z. Dereboylu, and E. Ever, “Clustering-Based Modified Ant Colony Optimizer for Internet of Vehicles (CACOIOV),” *Sustainability*, vol. 11, no. 9, p. 2624, May 2019, doi: 10.3390/su11092624.
[18] K. Lin, F. Xia, and G. Fortino, “Data-driven clustering for multimedia communication in Internet of vehicles,” *Future Gener. Comput. Syst.*, vol. 94, pp. 610–619, May 2019, doi: 10.1016/j.future.2018.12.045.
[19] S. Arjunan, S. Pothula, and D. Ponnurangam, “F5N-based unequal clustering protocol (F5NUCP) for wireless sensor networks,” *Int. J. Commun. Syst.*, vol. 31, no. 17, p. e3811, Nov. 2018, doi: 10.1002/dac.3811.

[20]N. Omar, N. Yaakob, Z. Husin, and M. Elshaikh, "Design and Development of GreedLea Routing Protocol for Internet of Vehicle (IoV)," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 767, no. 1, p. 012034, Feb. 2020, doi: 10.1088/1757-899X/767/1/012034.

[21]R. K. Yadav and H. Banka, "An improved chemical reaction-based approach for multiple sequence alignment," *Curr. Sci.*, vol. 112, no. 3, p. 527, 2017.

[22]J. Zhang, Y. Wang, S. Li, and S. Shi, "An Architecture for IoT-Enabled Smart Transportation Security System: A Geospatial Approach," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6205–6213, Apr. 2021, doi: 10.1109/JIOT.2020.3041386.

[23]F. Aadil, W. Ahsan, Z. U. Rehman, P. A. Shah, and S. Rho, "Clustering algorithm for internet of vehicles (IoV) based on dragonfly optimizer (CAVDO," *J. Supercomput.*, vol. 74, no. 9, pp. 4542-4567, 2018.

[24]M. Ahmed Hamza, H. Mesfer Alshahrani, F. N. Al-Wesabi, M. Al Duhayyim, A. Mustafa Hilal, and H. Mahgoub, "Artificial Intelligence Based Clustering with Routing Protocol for Internet of Vehicles," *Comput. Mater. Contin.*, vol. 70, no. 3, pp. 5835–5853, 2022, doi: 10.32604/cmc.2022.021059.

[25]M. Buvanesvari, J. Uthayakumar, and J. Amudhavel, "Fuzzy based clustering to maximize network lifetime in wireless mobile sensor networks," *J. Adv. Res. Dyn. Control Syst.*, vol. 9, no. 12, pp. 2156-2167, 2017.

[26]H. Fatemidokht and M. Kuchaki Rafsanjani, "QMM-VANET: An efficient clustering algorithm based on QoS and monitoring of malicious vehicles in vehicular ad hoc networks," *J. Syst. Softw.*, vol. 165, p. 110561, July 2020, doi: 10.1016/j.jss.2020.110561.

[27]F. Wang, M. Zhang, X. Wang, X. Ma, and J. Liu, "Deep Learning for Edge Computing Applications: A State-of-the-Art Survey," *IEEE Access*, vol. 8, pp. 58322–58336, 2020, doi: 10.1109/ACCESS.2020.2982411.