E-ISSN: 2964-6219, P-ISSN: 2964-6200
Volume 5 Number 1 February 2027.
pp: 61-75. DOI: https://doi.org/10.58477/cj.v5i1.532.

Computer Journal

RESEARCH ARTICLE Open Access

Evaluation of a Deep Q-Network with a Quadratic Reward Function for Two-Phase Traffic Signal Control

Muhamad Achya Arifudin [1*], **Budiman** [2], **Titan Parama Yoga** [3], **Nur Alamsyah** [4]

[1*,2] Department of Informatics, Universitas Informatika dan Bisnis Indonesia, Bandung City, West Java Province, Indonesia.
[3,4] Department of Information Systems, Universitas Informatika dan Bisnis Indonesia, Bandung City, West Java Province, Indonesia.

Correspondence email:
achyaarifudin@unibi.ac.id.

Article History
Received: 21 September 2026.
Revised: 26 September 2026.
Accepted: 14 November 2026.
Available Online: 5 December 2026
Published: 1 February 2027.

Funding Information
Universitas Informatika dan Bisnis Indonesia.

Abstract
Adaptive traffic signal control can respond to changing traffic demand more effectively than fixed-time operation, but deep reinforcement learning performance depends strongly on environment design, state representation, and reward formulation. This study evaluates a tuned Deep Q-Network (DQN) with a quadratic congestion penalty for two-phase traffic signal control. A dataset-driven environment paired two street segments as conflicting approaches and applied chronological training, validation, and test splits. No external microscopic traffic simulator was used; instead, the observed demand sequences drove a custom queue-based Gymnasium environment. Hyperparameters were tuned using the validation set, while final models were evaluated across three random seeds under normal, peak, incident, and imbalanced-demand scenarios. The quadratic DQN was compared with a linear-reward DQN, a fixed 50:50 controller, and a rule-based adaptive controller. Under normal traffic, the quadratic DQN achieved a mean queue of 22.16 vehicles, a 95th-percentile queue of 35.11 vehicles, throughput of 594.24 vehicles, and a service rate of 95.42%. Compared with the linear DQN, it reduced queue imbalance by 5.15% and action changes by 12.82%, although mean queue and throughput did not improve. The rule-based controller produced the best overall performance. These findings indicate that quadratic reward shaping can improve policy balance and stability, but simpler controllers may remain competitive when training data are limited. Fair baselines and multi-seed evaluation remain essential practices.

Keywords: Deep Q-Network; Reinforcement Learning; Traffic Signal Control; Quadratic Congestion Penalty; Two-Phase Intersection.

Abstrak
Kontrol sinyal lalu lintas adaptif dapat merespons perubahan permintaan lalu lintas lebih efektif dibandingkan operasi waktu tetap, tetapi kinerja deep reinforcement learning sangat bergantung pada desain lingkungan, representasi state, dan formulasi reward. Penelitian ini mengevaluasi tuned Deep Q-Network (DQN) dengan quadratic congestion penalty untuk kontrol sinyal lalu lintas dua fase. Lingkungan berbasis dataset dibangun dengan memasangkan dua segmen jalan sebagai pendekatan yang saling berkonflik serta menerapkan pembagian data training, validation, dan test secara kronologis. Tidak digunakan simulator lalu lintas mikroskopik eksternal; urutan permintaan kendaraan dari dataset digunakan untuk menggerakkan lingkungan antrean kustom berbasis Gymnasium. Hyperparameter ditentukan menggunakan validation set, sedangkan model akhir dievaluasi pada tiga random seed dalam skenario normal, peak, incident, dan imbalanced demand. DQN quadratic dibandingkan dengan DQN linear-reward, fixed 50:50 controller, dan rule-based adaptive controller. Pada kondisi normal, DQN quadratic menghasilkan mean queue 22,16 kendaraan, 95th-percentile queue 35,11 kendaraan, throughput 594,24 kendaraan, dan service rate 95,42%. Dibandingkan DQN linear, model ini menurunkan queue imbalance sebesar 5,15% dan action changes sebesar 12,82%, meskipun mean queue dan throughput tidak membaik. Rule-based controller memberikan kinerja keseluruhan terbaik. Temuan ini menunjukkan bahwa quadratic reward shaping dapat meningkatkan keseimbangan dan stabilitas kebijakan, tetapi controller yang lebih sederhana tetap kompetitif ketika data training terbatas. Penggunaan baseline yang adil, evaluasi multi-seed, dan desain reward transparan tetap penting untuk memastikan interpretasi hasil penelitian yang andal.

Kata Kunci: Deep Q-Network; Pembelajaran Penguatan; Kendali Sinyal Lalu Lintas; Penalti Kemacetan Kuadratik; Persimpangan Dua Fase.

© The Author(s) 2027. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

1. Introduction

Traffic congestion remains a major challenge in urban transportation systems (Adeniran *et al.*, 2026; Huang *et al.*, 2026). The increasing number of vehicles, limited road capacity, and continuously changing traffic demand frequently cause long queues at signalized intersections (Wei *et al.*, 2026). Intersections are particularly important because traffic movements from different directions must share a limited amount of green time (Oguchi *et al.*, 2026). Ineffective signal timing can increase waiting time, reduce traffic flow, and create excessive queue accumulation, especially during peak periods or unexpected disturbances (Chinyanga *et al.*, 2026). Conventional traffic signal systems commonly apply fixed-time control, in which signal durations are determined in advance based on historical traffic patterns (Hassan *et al.*, 2026; Manikandan & Pandian, 2026). This approach is simple, stable, and easy to implement. However, fixed-time control has limited ability to respond to rapid changes in traffic conditions. A timing plan that performs well under normal traffic may become inefficient when the volume of vehicles increases, when an incident occurs, or when traffic demand between conflicting directions becomes unbalanced (Sabbaghian *et al.*, 2026). These limitations have encouraged the development of adaptive traffic signal control methods that adjust signal timing according to current traffic conditions.

Several computational approaches have been applied to adaptive traffic signal control, including rule-based systems, fuzzy logic, optimization algorithms, and machine-learning-based methods (Colpo & Gomes, 2026; Brocchini *et al.*, 2026). Rule-based controllers can respond to traffic conditions by applying predefined decision rules, such as allocating more green time to the direction with the longer queue (Liu *et al.*, 2026). Although these controllers are relatively transparent and computationally efficient, they depend on handcrafted rules, thresholds, or expert-defined logic (Ren *et al.*, 2024). As traffic conditions become more dynamic and the number of indicators or controlled intersections increases, maintaining and extending such rule sets becomes increasingly difficult (Zhao *et al.*, 2024). Reinforcement learning provides a different approach by allowing an agent to learn a traffic control policy through interaction with an environment (Katzilieris *et al.*, 2026; Wang *et al.*, 2026). At each decision step, the agent observes the traffic state, selects an action, receives feedback in the form of a reward, and updates its decision policy. Through repeated learning, the agent is expected to identify actions that improve long-term traffic performance. Deep reinforcement learning extends this concept by using neural networks to estimate the value of possible actions from more complex traffic states.

The Deep Q-Network is one of the reinforcement learning methods that can be applied when the available traffic signal actions are discrete (Jin *et al.*, 2026). In a two-phase signal system, each action may represent a different proportion of green time allocated between two conflicting approaches (Zhao *et al.*, 2026). The network estimates the expected long-term return for each available action and selects the action that is considered most beneficial. Nevertheless, the effectiveness of a DQN controller is not determined only by the neural network architecture or the number of training steps. The design of the environment, the information included in the state, the available actions, and the reward function strongly influence the policy learned by the agent. Reward formulation is especially important because it defines the behavior that the agent is encouraged to learn (ur Rehman *et al.*, 2026). Traffic signal rewards may be based on queue length, waiting time, delay, throughput, pressure, or combinations of several performance indicators (Xiao *et al.*, 2026). A linear queue penalty assigns a proportional cost to each additional queued vehicle (Fokker *et al.*, 2026). This formulation is straightforward, but it treats every increase in queue length with the same rate of penalty. In contrast, a quadratic congestion penalty represents queue accumulation through a nonlinear cost. Such a formulation is intended to examine whether stronger sensitivity to queue growth can improve traffic balance, reduce extreme congestion, and produce a more stable signal-control policy.

However, a more complex reward function does not automatically guarantee better performance. The evaluation of reinforcement-learning-based traffic signal control can be affected by many factors, including the selected traffic scenario, random initialization, training duration, state representation, reward scaling, and baseline controller. Results based on only one random seed may not adequately represent model stability. Similarly, comparison with only a fixed-time controller may provide an incomplete assessment because adaptive rule-based control can also respond effectively to changes in traffic demand. A controlled comparison is therefore required to determine whether a quadratic congestion penalty offers practical benefits over a linear reward and conventional control strategies. This study evaluates a tuned Deep Q-Network with a quadratic congestion penalty for two-phase traffic signal control. A dataset-driven experimental environment is developed by representing two street segments as conflicting traffic approaches. In each control cycle, the selected action determines the proportion of available green time assigned to the first approach, while the remaining green time is allocated to the second approach. This

design creates a direct trade-off because increasing the green allocation for one approach reduces the allocation available to the conflicting approach.

The traffic data are processed and divided chronologically into training, validation, and testing sets. The validation set is used to select the DQN hyperparameter configuration, while the test set is reserved for final evaluation. Two DQN variants are examined using the same model configuration: one with a linear queue penalty and another with a quadratic congestion penalty. Their performances are compared with a fixed 50:50 controller and a rule-based adaptive controller. The evaluation is conducted across multiple random seeds and four traffic-demand scenarios, namely normal, peak, incident, and imbalanced-demand conditions. The performance of each controller is assessed using mean queue length, 95th-percentile queue length, throughput, service rate, queue imbalance, action-switching frequency, and the relationship between traffic demand and the selected green-time allocation. These measurements are intended to provide a broader evaluation than relying on a single traffic indicator. In addition to comparing overall performance, this study examines whether the quadratic reward influences the stability and balance of the learned traffic signal policy. The main contribution of this study is a transparent and controlled evaluation of reward design in DQN-based two-phase traffic signal control. The study compares linear and quadratic queue penalties under an identical learning configuration, evaluates the models across multiple random seeds and traffic scenarios, and includes both fixed-time and adaptive rule-based controllers as baselines. The results are expected to clarify the practical effects of quadratic congestion penalties and to identify the conditions under which a reinforcement learning controller may or may not provide advantages over simpler traffic signal control strategies.

2. Related Work

Recent traffic signal control studies have increasingly adopted deep reinforcement learning to improve the adaptability of signal timing under dynamic traffic conditions. Chu *et al*. (2022) developed an end-to-end off-policy deep reinforcement learning controller that directly processed intersection images and selected traffic signal actions to reduce vehicle waiting time.

Table 1. Related Works

Author(s)	Method Used	Research Focus	Limitation / Research Gap
Chu *et al*. (2022)	End-to-end off-policy DRL	Image-based traffic signal control at an isolated intersection	Relies on visual input and does not isolate the effect of linear and quadratic queue rewards.
Mao *et al*. (2022)	Comparative DRL models	Comparison of learning algorithms for isolated traffic signal control	Focuses on model architectures rather than controlled reward-function ablation.
Bouktif *et al*. (2023)	DDQN with prioritized experience replay	Consistent state and reward design for stable traffic control learning	State, reward, DDQN, and replay improvements are combined, so the effect of a specific nonlinear penalty is not isolated.
Kang *et al*. (2023)	Dueling DQN	Low-carbon and traffic-efficient signal control	Emphasizes emission reduction and does not specifically evaluate two-phase queue balance using linear and quadratic penalties.
Cai & Wei (2024)	Dueling Double DQN with prioritized sampling and noisy networks	Improving convergence, sample use, and robustness	Uses several enhancements simultaneously and does not compare reward formulations against simple adaptive baselines.
Tan *et al*. (2025)	Dueling Double DQN with state fusion and trend reward	Multisensor state representation and dynamic waiting-time reward	Uses a trend-based reward and does not compare linear and quadratic queue penalties under identical configurations.
Haddad *et al*. (2022)	Cooperative multi-agent DRL	Coordinated control of multiple intersections	Focuses on agent cooperation rather than transparent reward ablation for a two-phase isolated controller.

Yu *et al*. (2023)	Decentralized DRL	Multimodal traffic control for private vehicles and buses	Addresses a complex multimodal network and does not focus on queue-penalty design at a two-phase intersection.
Wang *et al*. (2024)	Multi-layer Graph Mask Q-Learning	Large-scale graph-based multi-intersection control	Requires a complex network representation and does not isolate quadratic congestion reward effects.
Bie *et al*. (2024)	Collaborative multi-agent DRL	Traffic signal control under intersection heterogeneity	Concentrates on heterogeneous network coordination rather than fixed, rule-based, linear, and quadratic reward comparison.
This study	Tuned DQN with linear and quadratic congestion penalties	Evaluation of two-phase traffic signal control under four demand scenarios	Provides controlled reward ablation, multi-seed evaluation, chronological data separation, and comparison with Fixed-50 and rule-based controllers.

Although the approach reduced dependence on manually extracted traffic features, it focused on an isolated intersection and did not specifically investigate how different queue-based reward formulations affected the learned policy. Mao *et al*. (2022) compared several deep reinforcement learning models for isolated traffic signal control and demonstrated that controller performance was influenced by the selected learning architecture. However, their study emphasized algorithm comparison rather than a controlled evaluation of linear and nonlinear congestion penalties. Reward and state design have also been recognized as important components of reinforcement-learning-based traffic signal control. Bouktif *et al*. (2023) proposed a consistent state and reward formulation combined with Double DQN and prioritized experience replay to improve training stability and convergence. Their results indicated that the interaction between state representation and reward design substantially influenced controller performance. However, several model components were introduced simultaneously, making it difficult to isolate the individual contribution of a specific penalty form. Kang *et al*. (2023) applied a dueling DQN to low-carbon traffic signal control by considering both traffic efficiency and environmental objectives. The method extended the control objective beyond congestion reduction, but it did not focus on the comparative effect of linear and quadratic queue penalties in a two-phase signal environment. Cai and Wei (2024) enhanced a deep reinforcement learning controller through dueling networks, double Q-learning, prioritized sampling, and noisy network parameters. The combination improved sample utilization and robustness, although the contribution of each component was not independently evaluated against simpler traffic controllers. More recently, Tan *et al*. (2025) introduced a Dueling Double DQN model using multisource state fusion and a waiting-time trend reward to address volatile traffic flow and reward sparsity. Their approach demonstrated the importance of dynamic reward information, but it did not examine whether a quadratic queue penalty could improve queue balance and policy stability compared with a linear reward under identical model settings. Several studies have expanded reinforcement learning from isolated intersections to larger and more complex traffic networks. Haddad *et al*. (2022) proposed a cooperative multi-agent deep reinforcement learning approach in which neighboring intersection agents shared state, action, and reward information. Yu *et al*. (2023) developed a decentralized controller for multimodal traffic networks that simultaneously considered private vehicles and bus operations. Although these studies demonstrated the potential of cooperative and decentralized learning, their objectives and environmental structures were more complex than the two-phase isolated control problem considered in the present study. Wang *et al*. (2024) proposed a multi-layer graph-mask Q-learning model for large-scale traffic signal control by representing road-network relationships at both network and intersection levels. Similarly, Bie *et al*. (2024) developed a collaborative multi-agent deep reinforcement learning method that accounted for differences in intersection structures and traffic demand. These approaches improved scalability and coordination across heterogeneous intersections, but they required more complex graph or multi-agent architectures and did not provide a controlled comparison between linear and quadratic congestion penalties using conventional fixed-time and rule-based baselines.

The reviewed studies show that recent research has mainly concentrated on improving network architecture, state representation, multi-agent coordination, multimodal control, and scalability. However, limited attention has been given to a transparent ablation study that isolates the influence of a quadratic congestion penalty while retaining the same DQN configuration. In addition, evaluations frequently emphasize comparisons with fixed-time control without simultaneously considering an adaptive rule-based controller, multiple random seeds, and several traffic-demand scenarios. Therefore, this study evaluates a tuned DQN

using a quadratic congestion penalty in a two-phase traffic signal environment. The quadratic-reward model is compared with a DQN using a linear queue penalty, a fixed 50:50 controller, and an adaptive rule-based controller. The experiments use identical DQN configurations for the two reward variants, chronological training, validation, and testing data, multiple random seeds, and normal, peak, incident, and imbalanced-demand scenarios. This design is intended to clarify whether quadratic reward shaping provides measurable benefits in queue reduction, throughput, balance, and signal-policy stability.

3. Methodology

The methodology of this study consists of six sequential stages, as illustrated in Figure 1. The process begins with traffic data collection, followed by data preprocessing to improve data quality and prepare the variables for model development. The preprocessing stage includes clipping and imputing missing values, extracting temporal features, encoding categorical variables, and applying MinMax normalization. The processed data are then used to construct a two-phase traffic signal environment, referred to as TrafficSignalEnv, which defines the traffic state, signal-control actions, and reward function. Within this environment, a tuned Deep Q-Network is developed to learn the allocation of green time between two conflicting traffic approaches. The model is trained and evaluated using chronologically separated training, validation, and testing data, with comparisons against linear-reward DQN, fixed-time, and rule-based controllers. Finally, the experimental results are visualized and analyzed using queue length, throughput, service rate, queue imbalance, and action-selection behavior to assess the effectiveness and stability of the proposed control approach.

3.1 Data Collection and Preprocessing

The dataset used in this study was obtained from the Kaggle portal under the title *Intelligent Streetlight Dataset*. It contains 1,000 observations from 10 street identifiers, with variables describing traffic, temporal, environmental, energy, and contextual conditions. The main traffic attributes include traffic count, traffic density, and traffic speed, supported by information such as weather, ambient light, power state, dim level, special events, and holiday or weekend status. The preprocessing stage began with duplicate removal, chronological sorting, and validation of numerical values. Invalid values were clipped to reasonable ranges, missing numerical values were replaced with the median, and missing categorical values were filled using the most frequent category. Categorical variables were transformed using one-hot encoding, while binary attributes were converted into zero-one values. Temporal features, including hour, weekday, and month, were extracted from the timestamp, with sine and cosine transformations applied to hour and weekday to preserve their cyclical patterns. To construct the two-phase traffic environment, Street ID 9 and Street ID 3 were selected because they contained the largest numbers of observations. Both sequences were arranged chronologically and limited to 106 records, then treated as two conflicting traffic approaches. The paired data were divided chronologically into 74 training observations, 15 validation observations, and 17 testing observations. Traffic count was converted into estimated arrivals per signal cycle using a scaling factor derived from the training data. Finally, all numerical features were normalized to the range of zero to one using MinMax scaling. The scaler was fitted only on the training set and then applied to the validation and testing sets to prevent data leakage.

3.2 Environment Design (TrafficSignalEnv)

A two-phase traffic signal environment was developed using the Gymnasium framework and implemented as TwoPhaseTrafficSignalEnv. The environment represents two conflicting traffic approaches, denoted as Approach A and Approach B. At each decision step, the agent allocates the available green time between the two approaches. Therefore, increasing the green-time proportion for one approach automatically reduces the green time available for the other approach. In this study, "dataset-driven" denotes that no external microscopic traffic simulator such as SUMO or VISSIM was used. Traffic demand was replayed from the paired dataset observations, while queue evolution and service dynamics were computed inside the custom Gymnasium environment. Accordingly, the environment is a simplified queue-based experimental model rather than a calibrated microscopic representation of a specific real intersection. The state observed by the agent consists of the preprocessed traffic and contextual features, the normalized queue lengths of both approaches, and the green-time allocation selected in the previous step. The state vector is expressed as:

$$s_t = [x_t, \bar{q}_{A,t}, \bar{q}_{B,t}, g_{A,t-1}] \in \mathbb{R}^{43} \quad (1)$$

Here, $x_t \in \mathbb{R}^{40}$is the 40-dimensional normalized contextual feature vector at decision step t. It contains 20 variables for Approach A and 20 for Approach B: traffic density, traffic speed, ambient light, energy consumption, power state, dim level, special-event indicator, holiday/weekend indicator, hour, weekday, month, hour-sine, hour-cosine, weekday-sine, weekday-cosine, two one-hot Day/Night indicators, and three one-hot weather indicators (Clear, Cloudy, and Rainy). Traffic Count itself is excluded from $x_t$and is converted separately into $\lambda_{A,t}$and $\lambda_{B,t}$. With the normalized queues $\bar{q}_{A,t}$and $\bar{q}_{B,t}$and the previous effective green allocation $g_{A,t-1}$appended to x_t, the complete observation is $s_t = [x_t, \bar{q}_{A,t}, \bar{q}_{B,t}, g_{A,t-1}] \in \mathbb{R}^{43}$. Queue values are normalized by the reference queue of 100 vehicles and clipped to $[0, 1]$. The action space is discrete and consists of five possible green-time proportions for Approach A:

$$a_t \in \{0.20, 0.35, 0.50, 0.65, 0.80\} \quad (2)$$

The remaining proportion is assigned to Approach B. A base lost-time ratio of 0.10 was applied in every signal cycle. An additional lost-time ratio of 0.03 was imposed whenever the selected action differed from the previous action, representing the efficiency loss caused by signal switching.

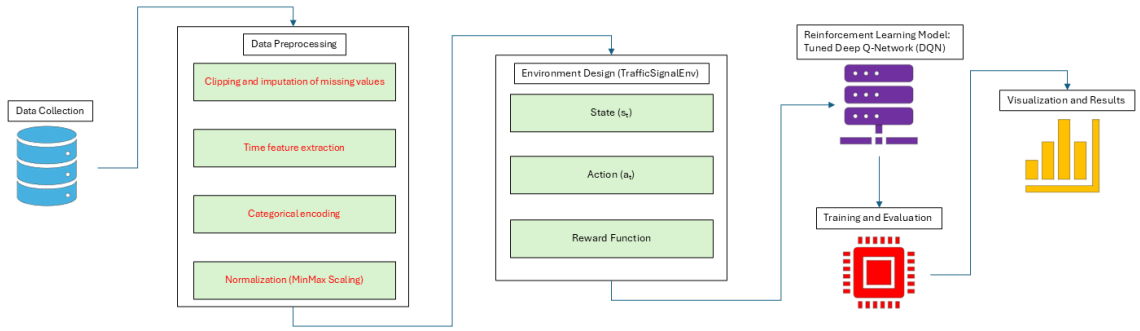


Figure 1. Research Methodology for Two-Phase Traffic Signal Control Using Tuned DQN

The effective green allocations are calculated as:

$$g_{A,t} = G_t a_t \quad (3)$$

$$g_{B,t} = G_t(1 - a_t) \quad (4)$$

where $G_t = 1 - 0.10 - 0.03 \cdot \mathbb{I}(a_t \neq a_{t-1})$is the effective green-time ratio after accounting for the base lost time and the additional action-change lost time. Vehicle arrivals for both approaches were obtained from the scaled traffic-count variables. During training, random arrival noise was added to improve policy robustness. The traffic demand before service is defined as the sum of the current queue and the incoming vehicles:

$$d_{i,t} = q_{i,t} + \lambda_{i,t},\ i \in \{A, B\} \quad (5)$$

where $q_{i,t}$is the existing queue and $\lambda_{i,t}$is the number of vehicle arrivals. The service capacity of each approach depends on its effective green allocation and a saturation flow of 45 vehicles per cycle:

$$c_{i,t} = 45 g_{i,t} \quad (6)$$

The number of served vehicles and the updated queue are then calculated as:

$$v_{i,t} = \min(d_{i,t}, c_{i,t}) \quad (7)$$

$$q_{i,t+1} = \max(0, d_{i,t} - v_{i,t}) \quad (8)$$

Two reward formulations were implemented to support a controlled comparison. The linear formulation penalizes the sum of the normalized queues, whereas the quadratic formulation penalizes their squared values. Both formulations also include queue-imbalance and action-switching costs. The reward is defined as:

$$r_t = -(w_c C_t + w_b B_t + w_s S_t) \quad (9)$$

where w_c, w_b, and $w_s$are the weights for congestion, queue imbalance, and switching costs, respectively. The congestion cost is calculated as:

$$C_t = \begin{cases} \bar{q}_{A,t+1} + \bar{q}_{B,t+1}, & \text{linear} \\ \bar{q}^2_{A,t+1} + \bar{q}^2_{B,t+1}, & \text{quadratic} \end{cases} \quad (10)$$

The queue imbalance and switching costs are calculated as:

$$B_t = | \bar{q}_{A,t+1} - \bar{q}_{B,t+1} | \quad (11)$$

$$S_t = | g_{A,t} - g_{A,t-1} | \quad (12)$$

A scaling consideration is important for interpreting the quadratic formulation. Queue lengths are normalized to the interval $[0, 1]$before the congestion term is calculated; therefore, for any moderate normalized queue $0 < q < 1$, $q^2$is numerically smaller than q. Consequently, the quadratic formulation does not impose a larger absolute penalty on moderate queues. Instead, it changes the curvature of the cost and places relatively greater emphasis on states that approach the reference queue limit. The same normalization was retained for both reward variants to preserve a controlled ablation, and this interaction is considered when interpreting the results. This environment design enables the DQN agent to learn the trade-off between serving competing traffic approaches while considering total congestion, queue balance, and signal stability. An episode ends when the specified episode length is reached or when the available traffic observations have been fully processed.

3.3 Reinforcement Learning Model: Tuned Deep Q-Network

A Deep Q-Network was employed to learn the green-time allocation policy in the two-phase traffic signal environment. The DQN receives the current traffic state $s_t$and estimates the expected cumulative reward for each available action. The selected action represents one of five discrete green-time proportions assigned to Approach A, while the remaining green time is allocated to Approach B. The action-value function is approximated by a neural network parameterized by θ:

$$Q(s_t, a_t, \theta) \quad (13)$$

During training, the target value is calculated as:

$$y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a', \theta^-) \quad (14)$$

where $r_t$is the reward obtained after selecting action a_t, $\gamma$is the discount factor, and θ^-represents the parameters of the target network. The network parameters are updated by minimizing the temporal-difference loss:

$$L(\theta) = [y_t - Q(s_t, a_t, \theta)]^2 \quad (15)$$

An experience replay buffer was used to store transitions in the form of (s_t, a_t, r_t, s_{t+1}). Random mini-batches were sampled from this buffer to reduce correlation between consecutive observations and improve training stability. A separate target network was updated periodically to provide more stable learning targets.

Hyperparameter tuning was performed using the validation set. Four candidate configurations were evaluated by varying the learning rate, discount factor, batch size, and neural network architecture. The configuration with the highest validation episode return was selected as the final DQN setting. The selected model used two hidden layers with 128 neurons each, a learning rate of 3×10^{-4}, a discount factor of 0.99, and a batch size of 128. The final DQN was trained using a replay buffer containing 100,000 transitions. Learning began after 5,000 initial interactions, and the model was updated every four environment steps using one gradient update. The target network was refreshed every 1,000 steps. An epsilon-greedy strategy was applied to balance exploration and exploitation. The exploration rate decreased from 1.0 to 0.02 during the first 25% of the training process. The same tuned DQN configuration was applied to both the linear-reward and quadratic-reward variants. This controlled setting ensured that differences in performance were primarily associated with the reward formulation rather than changes in network architecture or training parameters.

Table 2. Selected DQN Hyperparameters

Parameter	Selected Value
Learning rate	3×10^{-4}
Discount factor	0.99
Batch size	128
Hidden layers	[128, 128]
Replay buffer size	100,000
Learning starts	5,000 steps
Training frequency	Every 4 steps
Gradient steps	1
Target update interval	1,000 steps
Initial exploration rate	1.00
Final exploration rate	0.02
Exploration fraction	0.25

3.4 Training, Evaluation, and Result Visualization

The final DQN models were trained using the best hyperparameter configuration obtained from the validation stage. Two variants were evaluated under identical settings: DQN with a linear congestion penalty and DQN with a quadratic congestion penalty. Each model was trained for 100,000 timesteps using three random seeds, namely 11, 42, and 73. Random starting points and arrival noise were applied during training to improve policy robustness. The training split contained 74 paired observations, so the 100,000 training timesteps should not be interpreted as 100,000 independent traffic samples. Episodes repeatedly revisited this finite training sequence from randomized starting points, while arrival noise introduced stochastic perturbations. Thus, the long training horizon increased optimization exposure to the available patterns but did not increase the underlying diversity of observed traffic states. Because only 74 paired observations form the training sequence, the 100,000 timesteps represent repeated off-policy interaction with a finite set of base arrival patterns rather than 100,000 independent traffic observations. Random episode starts and 5% multiplicative arrival noise introduce stochastic variation, but they do not create new traffic regimes. The experiment is therefore interpreted as a controlled reward-ablation study; broader generalization requires substantially larger data or dynamic microscopic simulation (e.g., SUMO or CityFlow), which is identified as future validation rather than claimed in the present work. The models were compared with a Fixed-50 controller and a rule-based adaptive controller. Evaluation was conducted on separate test data under four scenarios: normal, peak, incident, and imbalanced demand. Performance was measured using mean queue length, 95th-percentile queue length, throughput, service rate, queue imbalance, action-switching frequency, episode return, and the correlation between traffic demand and selected green-time allocation.
The four test scenarios were generated from the same 17-observation held-out test sequence (16 decision steps) without evaluation noise. Normal used the original scaled arrivals ($A \times 1.00$, $B \times 1.00$). Peak increased arrivals on both approaches by 25% for the entire episode ($A \times 1.25$, $B \times 1.25$). Incident kept the normal demand outside the disruption window but, during 35%–65% of episode progress (five middle decision steps for the 16-step test horizon), multiplied Approach A arrivals by 1.50 and Approach B by 1.10. Imbalanced demand applied $A \times 1.35$and $B \times 0.80$ for the entire episode. All controllers were evaluated on exactly the same scenario transformations. The DQN results were reported as the mean and standard deviation across the three seeds, while the deterministic baselines were evaluated once per scenario. The results were presented through tables and figures showing queue and throughput comparisons, queue

evolution, scenario robustness, action distribution, demand-responsive behavior, and hyperparameter-tuning performance. All figures were generated at 330 dpi for publication quality.

4. Result and Discussion

This section presents the implementation and evaluation results of the two-phase traffic signal control experiment. The analysis covers hyperparameter selection, controller performance under normal traffic conditions, robustness across different demand scenarios, and the behavior of the learned DQN policies. The DQN with a quadratic congestion penalty was compared with a linear-reward DQN, a Fixed-50 controller, and a rule-based adaptive controller.

4.1 Hyperparameter Tuning Results

Four DQN configurations were evaluated using the validation dataset. The configurations differed in learning rate, discount factor, batch size, and neural network architecture. Since the reward represents a traffic-control cost, a validation return closer to zero indicates better performance. As shown in Table 3, configuration C2 achieved the highest validation return of −0.4427, the lowest mean queue of 14.23 vehicles, and the highest service rate of 95.69%. Therefore, this configuration was selected for the final DQN models. The same configuration was applied to both the linear and quadratic reward variants so that differences in performance were primarily influenced by the reward formulation.

Table 3. Hyperparameter Tuning Results

Configuration	Learning Rate	Gamma	Batch Size	Hidden Layers	Validation Return	Mean Queue	Service Rate
C1	0.0001	0.990	64	[128, 128]	−0.7744	19.75	93.89%
C2	0.0003	0.990	128	[128, 128]	−0.4427	14.23	95.69%
C3	0.0001	0.995	128	[256, 256]	−1.0104	24.88	92.26%
C4	0.0003	0.995	64	[256, 256]	−0.5446	17.04	94.43%

4.2 Controller Performance Under Normal Traffic Conditions

Table 4 summarizes the performance of the four controllers under the normal test scenario. The DQN results are reported as the mean and standard deviation across three random seeds, while Fixed-50 and rule-based are deterministic controllers. The rule-based controller achieved the best overall performance, with a mean queue of 6.91 vehicles, a P95 queue of 19.46 vehicles, and a throughput of 609.49 vehicles. The Fixed-50 controller ranked second, producing a mean queue of 12.05 vehicles and a service rate of 97.58%.

Table 4. Controller Performance Under the Normal Test Scenario

Controller	Mean Queue	P95 Queue	Throughput	Service Rate	Queue Imbalance	Action Changes	Demand Green Correlation
DQN-Quadratic	22.16 ± 4.83	35.11 ± 6.10	594.24 ± 7.06	95.42 ± 1.13%	11.10 ± [SD]	11.33 ± [SD]	0.168
DQN-Linear	22.05 ± 4.42	35.21 ± 5.67	595.42 ± 4.86	95.61 ± 0.78%	11.71 ± [SD]	13.00 ± [SD]	0.154
Fixed-50	12.05	23.49	607.68	97.58%	7.71	0.00	0.000
Rule-Based	6.91	19.46	609.49	97.87%	2.27	14.00	0.955

The two DQN variants produced similar results. Compared with DQN-linear, DQN-quadratic reduced the P95 queue slightly, from 35.21 to 35.11 vehicles. However, its mean queue was marginally higher and its throughput slightly lower. These differences were smaller than the standard deviations across the random seeds, indicating that the quadratic penalty did not provide a substantial improvement in average traffic performance under normal demand. Nevertheless, DQN-quadratic reduced queue imbalance by approximately 5.15% and action changes by approximately 12.82% compared with DQN-linear. This indicates that the quadratic formulation produced a slightly more balanced and stable policy, although the improvement did not translate into lower average queues or higher throughput. Figure 2 presents the queue evolution throughout the normal test scenario. The rule-based controller maintained the lowest queue for

most decision steps and responded effectively to sudden increases in demand, particularly around decision steps 9 and 14. Fixed-50 generally produced higher queues than the rule-based controller but remained below both DQN variants. The trajectories of DQN-linear and DQN-quadratic followed similar patterns, with their highest queue occurring around decision step 9. DQN-quadratic produced a slightly lower peak than DQN-linear at this point, but the difference remained limited.

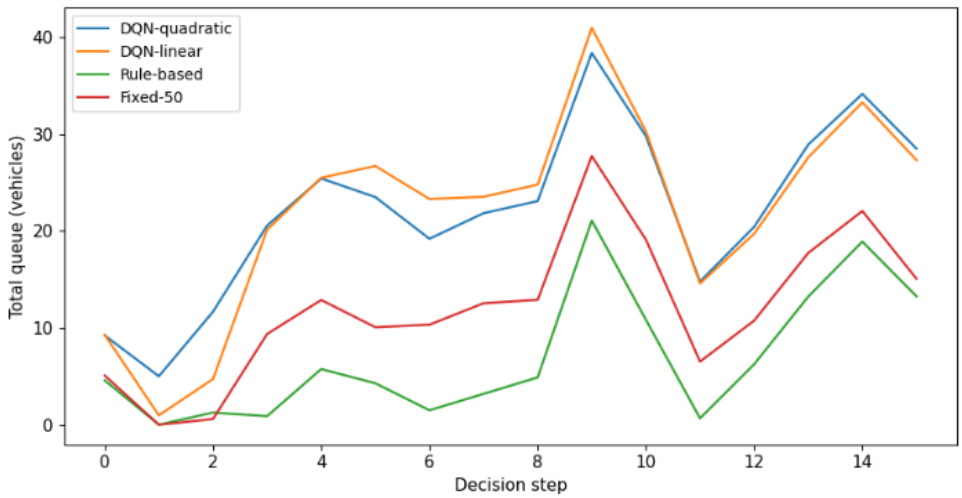


Figure 2. Queue Evolution Under the Normal Test Scenario

Figure 3 compares the total throughput obtained by the four controllers. The rule-based controller achieved the highest throughput of 609.49 vehicles, followed closely by Fixed-50 with 607.68 vehicles. DQN-linear and DQN-quadratic processed 595.42 and 594.24 vehicles, respectively. The relatively small error bars of the DQN models indicate that throughput was reasonably stable across the three random seeds. However, the quadratic reward did not increase throughput compared with the linear reward. Its main contribution was therefore associated with policy balance and switching stability rather than the total number of served vehicles.

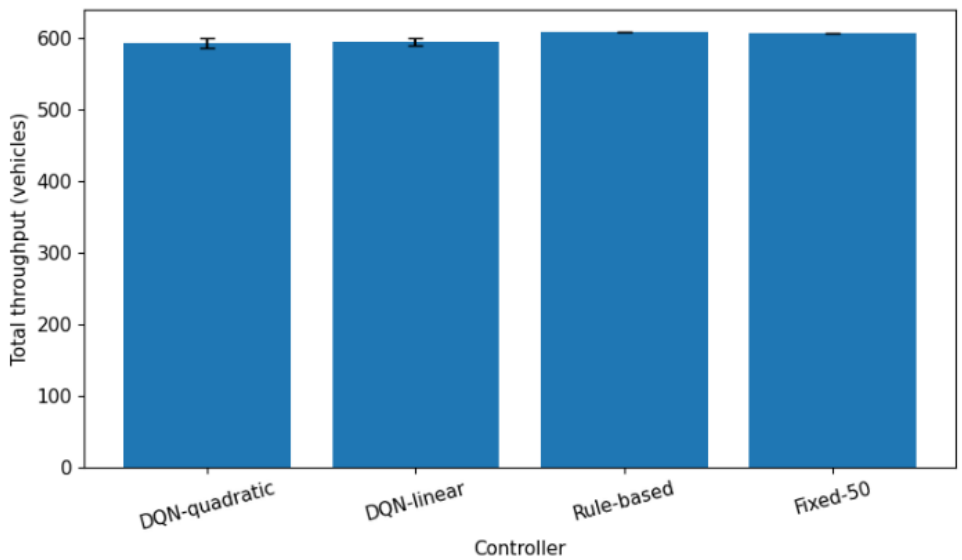


Figure 3. Throughput Comparison Under the Normal Test Scenario

Figure 4 shows the mean queue comparison under normal traffic conditions. The rule-based controller generated the lowest average queue, followed by Fixed-50. Both DQN models produced substantially higher average queues, with only a small difference between their results. The error bars indicate variation across the random seeds used to train the DQN models. DQN-quadratic showed slightly greater variation than DQN-linear, suggesting that the quadratic reward did not improve the consistency of average queue performance.

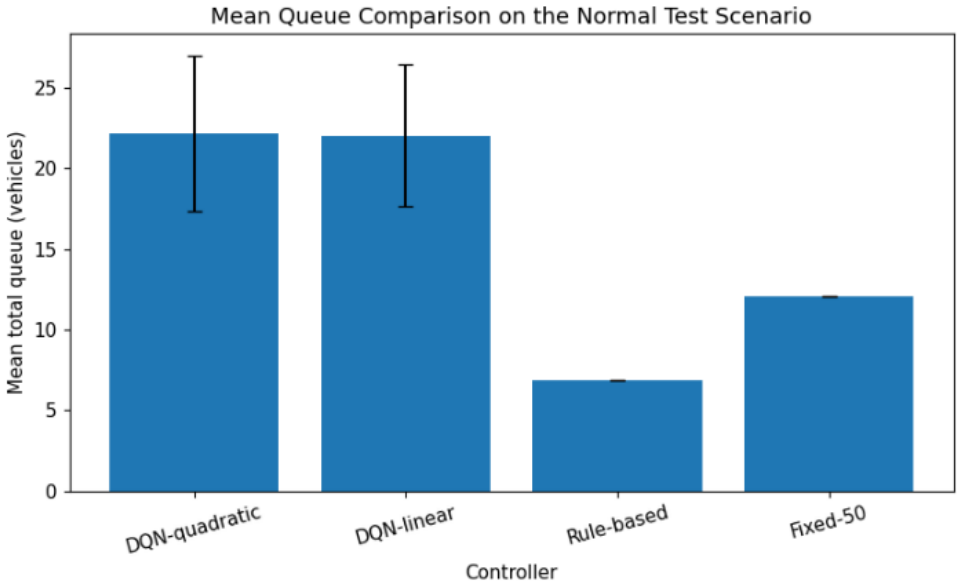


Figure 4. Mean Queue Comparison Under the Normal Test Scenario

4.3 Robustness Across Traffic Demand Scenarios

The four controllers were further evaluated under normal, peak, incident, and imbalanced-demand conditions. Figure 5 shows that all controllers experienced their highest mean queues under the peak scenario, demonstrating the difficulty of maintaining sufficient service capacity when traffic demand increased simultaneously on both approaches. Under peak demand, Fixed-50 achieved the lowest mean queue. This result suggests that an equal green-time allocation remained effective when the increase in traffic was approximately balanced between the two approaches. The rule-based controller achieved the best performance under normal, incident, and imbalanced-demand conditions because it adjusted the green allocation directly according to the current demand distribution. The performance of Fixed-50 declined substantially under imbalanced demand because equal allocation could not sufficiently accommodate the dominant traffic approach. In contrast, both DQN models adapted their actions and produced lower mean queues than Fixed-50 in this scenario. DQN-quadratic consistently generated slightly lower mean queues than DQN-linear under peak, incident, and imbalanced conditions. This suggests that the quadratic penalty became more useful as congestion increased or demand became uneven. However, its performance remained below that of the best conventional controller in each scenario.

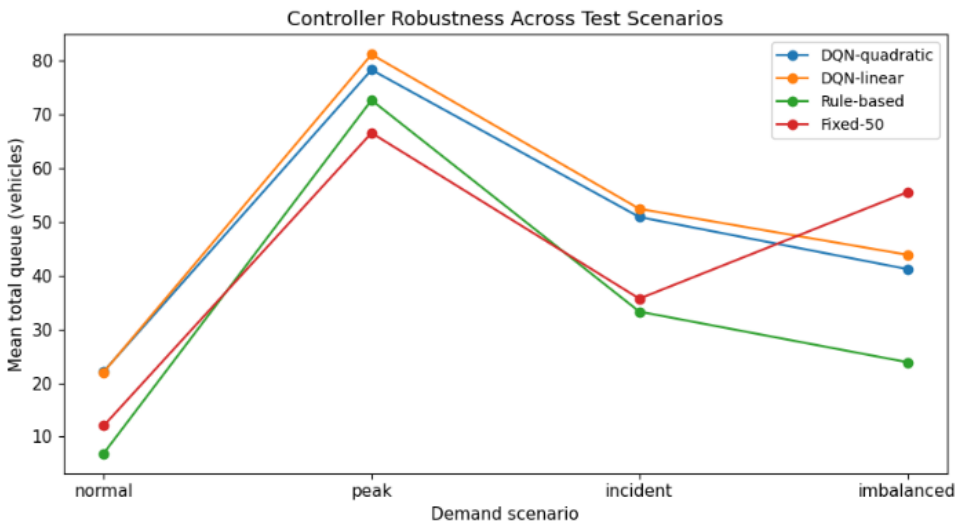


Figure 5. Controller Robustness Across the Test Scenarios

These results demonstrate that controller effectiveness depends strongly on traffic conditions. A more complex learning model does not automatically outperform conventional controllers, especially when

demand patterns are relatively simple or training data are limited.

4.4 Policy Behavior Analysis

The relationship between traffic demand and the selected green-time allocation was examined to determine whether the DQN policies behaved adaptively. DQN-quadratic achieved a demand–green correlation of 0.168, which was slightly higher than the 0.154 obtained by DQN-linear. However, both values were considerably lower than the rule-based controller's correlation of 0.955. The relatively low correlations indicate that the DQN actions were only weakly associated with the current traffic-demand proportion. The learned policy may have considered other state information, such as historical queue conditions, contextual variables, and the previous green allocation. However, the results also suggest that the model did not respond as directly to current demand as the rule-based controller. The action distribution showed that DQN-quadratic did not repeatedly select a single green-time allocation. Moderate allocations of 0.35, 0.50, and 0.65 were selected more frequently, while the extreme actions of 0.20 and 0.80 were used less often. This confirms that the two-phase environment successfully prevented the agent from obtaining an automatic advantage by continuously selecting the maximum green allocation. However, action diversity alone does not necessarily indicate strong adaptability. A responsive controller should select green-time allocations consistently according to changes in vehicle arrivals, queue pressure, and the relative demand of the conflicting approaches.

4.5 Overall Discussion

The results show that the quadratic congestion penalty provided modest improvements in P95 queue, queue balance, and action stability, but did not improve mean queue length, throughput, or service rate compared with the linear DQN. One possible explanation is the reward scaling used in the environment. Queue lengths were normalized before being squared. For normalized queue values below one, the squared value becomes smaller than the corresponding linear value. Consequently, moderate queues received a weaker penalty under the quadratic formulation, while stronger penalization only occurred when the normalized queue approached its upper limit. The superior performance of the rule-based controller may also be related to the information used in its decision process. The rule-based controller directly used the current queue and vehicle arrivals to determine green-time allocation. In contrast, the DQN inferred its action from the preprocessed feature vector, normalized queues, and previous action. Explicitly incorporating current arrivals and demand proportions into the DQN state could improve its responsiveness. The limited number of traffic observations may also have restricted DQN generalization. Although each model was trained for 100,000 timesteps, the training dataset contained only 74 paired observations. The agent therefore revisited the same finite set of demand patterns many times, with randomized episode starts and arrival noise providing variation but not new independent traffic regimes. In this setting, additional training steps primarily refine the policy on repeated observations and may increase the risk of overfitting to the limited demand support. The multi-seed design evaluates sensitivity to stochastic initialization, but it does not compensate for limited data diversity. Consequently, the reported results should be interpreted as evidence from a controlled dataset-driven experiment rather than as proof of generalization to unseen real-world intersections. In practical terms, the replay buffer can store many transitions generated during repeated passes, but those transitions remain anchored to the same 74 base demand records. Consequently, buffer size and timestep count increase optimization opportunities, not the empirical diversity of the traffic process. This distinction limits claims about out-of-sample policy generalization and motivates future evaluation with longer temporal data and microscopic simulation. Despite these limitations, the two-phase environment provides a more realistic evaluation than a single-queue setting. Allocating additional green time to one approach automatically reduces the allocation available to the conflicting approach, requiring the agent to learn an actual control trade-off. The use of separate training, validation, and testing data, multiple random seeds, and both fixed and adaptive baseline controllers also provides a more transparent evaluation of the learned policy. Overall, the findings indicate that quadratic reward shaping should not be assumed to improve traffic signal control automatically. Its effectiveness depends on reward scaling, state representation, traffic variability, and training-data availability. In this experiment, its main benefit was the production of a slightly more balanced and less frequently changing policy rather than superior queue or throughput performance.

5. Result and Discussion

This study evaluated a tuned Deep Q-Network with a quadratic congestion penalty for two-phase traffic signal control using a dataset-driven environment. The findings show that quadratic reward shaping produced a slightly more balanced and stable control policy, particularly in reducing extreme queue growth, queue imbalance, and frequent action changes. However, it did not consistently improve average queue length, throughput, or service rate compared with the linear-reward DQN. The rule-based controller remained more effective under most evaluated scenarios, indicating that model complexity does not automatically lead to superior control performance when training data and state information are limited. These findings highlight that the effectiveness of DQN-based traffic signal control depends not only on hyperparameter tuning but also on reward scaling, state representation, data availability, and traffic-demand characteristics. Because the study used a custom dataset-driven queue environment without a microscopic traffic simulator and relied on a limited set of repeated training patterns, the findings should be interpreted as controlled comparative evidence rather than direct field-level validation. Future research should incorporate current vehicle arrivals and demand proportions into the state, refine the quadratic reward formulation, use larger and temporally aligned traffic datasets, and evaluate the controller in more realistic traffic simulations or operational environments.

Referensi

Adeniran, A. O., Adeniran, A. A., Ogieva, M. O., & Ogwuche, G. (2026). Adoption of Intelligent Transport Systems (ITS) in urban transportation planning. *Discover Global Society*, *4*, Article 13. https://doi.org/10.1007/s44282-025-00312-3

Bie, Y., Ji, Y., & Ma, D. (2024). Multi-agent deep reinforcement learning collaborative traffic signal control method considering intersection heterogeneity. *Transportation Research Part C: Emerging Technologies*, *164*, 104663. https://doi.org/10.1016/j.trc.2024.104663

Bouktif, S., Cheniki, A., Ouni, A., & El-Sayed, H. (2023). Deep reinforcement learning for traffic signal control with consistent state and reward design approach. *Knowledge-Based Systems*, *267*, 110440. https://doi.org/10.1016/j.knosys.2023.110440

Brocchini, L., Wang, C., & Pratelli, A. (2026). Intelligent traffic control strategies for road networks: A taxonomy-based perspective on methods, applications, and future directions. *Applied Sciences*, *16*(13), 6341. https://doi.org/10.3390/app16136341

Cai, C., & Wei, M. (2024). Adaptive urban traffic signal control based on enhanced deep reinforcement learning. *Scientific Reports*, *14*, 14116. https://doi.org/10.1038/s41598-024-64885-w

Chinyanga, F., Sibiya, M., & Sumbwanyambe, M. (2026). Improving urban traffic efficiency through IoT-based intelligent traffic control systems and machine learning. *IEEE Open Journal of Intelligent Transportation Systems*, *7*, 1552–1562. https://doi.org/10.1109/OJITS.2026.3705979

Chu, K.-F., Lam, A. Y. S., & Li, V. O. K. (2022). Traffic signal control using end-to-end off-policy deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 7184–7195. https://doi.org/10.1109/TITS.2021.3067057

Colpo, L. R., & Gomes, H. M. (2026). An adaptative fuzzy logic control with road class classifications based on machine learning approaches. *Mechanics Based Design of Structures and Machines*, *54*(1), Article 2523577. https://doi.org/10.1080/15397734.2025.2523577

Fokker, E. S., van der Mei, R. D., & Dugundji, E. R. (2026). From models to mobility systems: A survey of multimodal dynamic traffic assignment. *Operations Research Forum*, *7*, Article 58. https://doi.org/10.1007/s43069-026-00642-1

Haddad, T. A., Hedjazi, D., & Aouag, S. (2022). A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control. *Engineering Applications of Artificial Intelligence*, *114*, 105019. https://doi.org/10.1016/j.engappai.2022.105019

Hassan, M., Turjo, T. D., Rambe, A. H., Jusoh, M., Islam, M. K., Hafizur Rahman, M. M., Alam, M. S., & Rahman, M. M. (2026). A comprehensive survey of adaptive traffic signal control: Methods, applications, challenges, and future research. *Archives of Computational Methods in Engineering*. Advance online publication. https://doi.org/10.1007/s11831-026-10574-y

Huang, G., Wang, H., Zhang, L., & Xu, D. (2026). The impact of ride-hailing regulations on traffic congestion: Evidence from China. *Transportation Research Part A: Policy and Practice*, *203*, 104765. https://doi.org/10.1016/j.tra.2025.104765

Jin, J., Wu, S., Zhang, P., Zhang, X., Yin, C., & Dong, C. (2026). Hybrid action space approach to traffic signal optimization using deep reinforcement learning. *Computers & Operations Research*, *189*, 107391. https://doi.org/10.1016/j.cor.2026.107391

Kang, L., Huang, H., Lu, W., & Liu, L. (2023). A dueling deep Q-network method for low-carbon traffic signal control. *Applied Soft Computing*, *141*, 110304. https://doi.org/10.1016/j.asoc.2023.110304

Katzilieris, K., Kampitakis, E., & Vlahogianni, E. I. (2026). A multi-agent reinforcement learning framework for integrated traffic signal control and dynamic bus lane access management. *Transportation Research Part C: Emerging Technologies*, *182*, 105391. https://doi.org/10.1016/j.trc.2025.105391

Liu, H., Li, X., Zhang, L., & Cheng, R. (2026). Bridging phase and timing: A joint Q-value learning framework for synergistic traffic signal control at consecutive arterial road intersections. *Physica A: Statistical Mechanics and Its Applications*, *688*, 131421. https://doi.org/10.1016/j.physa.2026.131421

Manikandan, R., & Pandian, J. A. (2026). A systematic review of traditional and reinforcement learning-based traffic signal control methods. *Archives of Computational Methods in Engineering*. Advance online publication. https://doi.org/10.1007/s11831-026-10692-7

Mao, F., Li, Z., & Li, L. (2022). A comparison of deep reinforcement learning models for isolated traffic signal control. *IEEE Intelligent Transportation Systems Magazine*, *15*(1), 160–180. https://doi.org/10.1109/MITS.2022.3144797

Oguchi, T., Iryo, M., & Toriumi, A. (2026). Traffic engineering. In *Pioneering the future for traffic and safety sciences: Transdisciplinary wisdom* (pp. 77–101). Springer.

Ren, F., Dong, W., Zhao, X., Zhang, F., Kong, Y., & Yang, Q. (2024). Two-layer coordinated reinforcement learning for traffic signal control in traffic network. *Expert Systems with Applications*, *235*, 121111. https://doi.org/10.1016/j.eswa.2023.121111

Sabbaghian, M. H., Molan, A. M., Hummer, J. E., & Osafo-Gyamfi, S. (2026). Traffic operation and safety performance of the new roundabout and median U-turn combination design in comparison to four roundabout designs. *International Journal of Transportation Science and Technology*. Advance online publication. https://doi.org/10.1016/j.ijtst.2025.09.007

Tan, X., Zhou, Y., & Jiao, X. (2025). Traffic signal control based on deep reinforcement learning using state fusion and trend reward. *Engineering Applications of Artificial Intelligence*, *159*, 111701. https://doi.org/10.1016/j.engappai.2025.111701

ur Rehman, H. M. R., Gul, M. J., Younas, R., Jhandir, M. Z., Alvarez, R. M., Miro, Y., & Ashraf, I. (2026). End-to-end emergency response protocol for tunnel accidents augmentation with reinforcement learning. *Scientific Reports*, *16*, 6226. https://doi.org/10.1038/s41598-026-37191-w

Wang, R., Zhang, J., Wang, X., He, H., & Zou, Y. (2026). Multi-modal and multi-agent reinforcement learning framework for urban traffic flow prediction and signal control optimization. *Scientific Reports*, *16*, 7612. https://doi.org/10.1038/s41598-026-37722-5

Wang, T., Zhu, Z., Zhang, J., Tian, j., & Zhang, W. (2024). A large-scale traffic signal control algorithm based on multi-layer graph deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, *162*, 104582. https://doi.org/10.1016/j.trc.2024.104582

Wei, W., Zheng, L., & El Esawey, M. (2026). Model-agnostic meta-learning-based real-time traffic conflict prediction with limited sample at heterogeneous signalized intersections. *IEEE Transactions on Intelligent Transportation Systems*, *27*(4), 4811–4823. https://doi.org/10.1109/TITS.2025.3648403

Xiao, H., Li, H., Wang, Z., Yang, Z., Qi, S., Zhang, J., Cai, D., & Yin, J. (2026). Intelligent traffic signal control based on reinforcement learning: A survey. *Artificial Intelligence Review*, *59*, Article 128. https://doi.org/10.1007/s10462-026-11530-9

Yu, J., Laharotte, P.-A., Han, Y., & Leclercq, L. (2023). Decentralized signal control for multi-modal traffic network: A deep reinforcement learning approach. *Transportation Research Part C: Emerging Technologies*, *154*, 104281. https://doi.org/10.1016/j.trc.2023.104281

Zhao, H., Dong, C., Cao, j., & Chen, Q. (2024). A survey on deep reinforcement learning approaches for traffic signal control. *Engineering Applications of Artificial Intelligence*, *133*, 108100. https://doi.org/10.1016/j.eswa.2024.108100

Zhao, M., Zou, Y., Wang, J., Li, D., & Yang, Q. (2026). Safe and scalable multi-agent optimization for autonomous electric taxi dispatching via a two-stage reinforcement learning framework. *IEEE Transactions on Automation Science and Engineering*. Advance online publication. https://doi.org/10.1109/TASE.2026.3651287