International Journal of Research in Engineering, Science and Management
Volume 6, Issue 10, October 2023
https://www.ijresm.com | ISSN (Online): 2581-5792

Traffic Signal Control Systems Incorporate Reinforcement Learning Techniques

Najat Benchelha[1*], Mohamed Bezza[2], Noureddine Belbounaguia[3], Taoufik Benchelha[4]

[1,2,3]*Electronic, Automatic Energy and Information Processing Laboratory, Faculty of Technical Sciences, Mohammedia, Morocco*
[4]*Geosciences Laboratory, Hassan II University of Casablanca, Faculty of Sciences, Ain Chock, Morocco*

***Abstract*: An effective transportation system must include intelligent traffic light regulation. An intelligent traffic light control system should dynamically respond to real-time traffic, unlike traditional traffic lights, which are typically operated using manual instructions. Q-reinforcement learning is a technology that is increasingly being applied to traffic light regulation, and recent experiments have yielded promising results. In this study, an adaptive traffic signal scheduling strategy is designed utilizing Q-learning (QL) to minimize the number of vehicles blocking an intersection.**

***Keywords*: Internet of Things, Q-Learning, Signal control, Traffic, Vissim.**

1. Introduction

Every day, Urban crossroads are becoming increasingly congested as traffic demand rises, which is a significant challenge in the transportation system (Alberto 2005). Recently, research has focused on intelligent transportation systems (ITS), which aim to enhance the safety, efficiency, and eco-friendliness of traffic management systems [1].

Among various traffic management systems, Traffic Signal Control (TSC) is considered a vital component of ITS, serving as a fundamental tool for traffic management. It's can be divided into: Fixed-time signal control and adaptive traffic signal control [1]. First at a junction is optimized offline based on historical traffic data (not real-time traffic demands) in a practice known as fixed-time traffic signal control. But occasionally, changes in traffic conditions could render outdated predetermined settings for traffic light timing. As a result, traffic congestion results from the inability of fixed-time traffic signal control to adjust to dynamic and bursty traffic needs [1].

Contrarily, the second has been demonstrated to be a successful strategy to lessen traffic congestion. This technique modifies traffic signal timing in accordance with real-time traffic demand [2], [3]. Artificial intelligence has recently offered a fresh approach to resolving this issue. For example, (A Deep Reinforcement Learning Approach to Adaptive Traffic Lights Management) proposes a micro-agent-based simulator and a Reinforcement Learning method for controlling traffic lights.

Artificial intelligence has recently offered a fresh approach to resolving this issue. To create signal control systems, for i, several academics suggested using fuzzy theory [1], [4], fuzzy neural networks [10], and fuzzy control models [5], [6]. Reinforcement learning is another method that can be used to learn the best signal control approach [7]. The static temporal optimization problems can also be solved using evolutionary algorithms (EAs) [8], such as genetic algorithms, particle swarm optimization, etc.

The best approach for determining the appropriate course of action in a dynamic environment among the various AI strategies is reinforcement learning. The stochastic challenges in the traffic environment can be addressed using Q-learning (QL), a values-based reinforcement learning algorithm. The Q-learning (QL) technique is used in this study to improve the performance of the proposed optimal TSC system, which aims to minimize the number of vehicles blocking an intersection.

2. Literature Review

Many research projects have been conducted to enhance the intelligence of traffic management systems for intersection traffic management. However, conventional methods cannot address the complex traffic signal optimization problems. Therefore, current traffic lights utilize AI methods for example fuzzy logic [9], Q-learning (QL) [10], and deep Q-learning [11]. In the fuzzy logic approach [9], the optimal signal extension time is determined using a membership function and two sensors to assess traffic flow. When compared to fixed-time controllers, this technology exhibits greater adaptability. The fuzzy system continuously adjusts to the dynamic environment by incorporating and employing fuzzy control phrases tailored to the changing conditions. Nevertheless, these tasks demand a substantial amount of processing power, eventually resulting in reduced system performance. Consequently, researchers have explored the use of QL for traffic control.

The reinforcement learning technique Q-learning (QL) can be applied to traffic control with dynamic changes. QL does not require predefined models, making it suitable for real-time traffic management. In a study [10], the stop delay was minimized by finding the optimal green light duration using QL with the assistance of a fuzzy rule set that classified the traffic

*Corresponding author: benchelha.doc@gmail.com

environment. However, the performance was not optimal for low traffic volumes. Another study [7] aimed to lessen traffic congestion by minimizing road waiting times. The traffic signal sequence used to activate the green light was pre-defined, and the duration of the green light was adjusted based on the length of the queue. Although this method was more adaptable than fixed-time traffic signals, it couldn't change the sequence of the traffic lights. To reduce the waiting time for vehicles, a cluster-based Q-learning approach was proposed. The parameters used to decide the sequence of green lights were the queue length and the duration of time the queue had been waiting. However, there was a risk associated with exclusively focusing on congested areas.

Deep Q-learning (deep QL) is a method for handling inputs with many states and high dimensions compared to QL. It's employed to learn the Q-function, and the value-function-based agent subsequently selects the optimal control action. In [6], the objective is to minimize queue length differences in all directions using a parameter. Similarly, the aim of [12] to minimize the variation in overall cumulative delays between the previous and current time. [13], applied a reinforcement learning technique to control a Manhattan-like traffic network with nine intersections. Although frequently include numerous layers, the complexity required for our signal management challenge is not necessary. Therefore, the primary focus of this work will be on investigating light control systems using QL.

In this paper, QL method is presented that focuses on two main parameters: throughput and the standard derivation of queue length. By modifying traffic lights with QL, the intention to increase the volume of vehicles passing through an intersection while maintaining the balance between the roadways. The action set, however, is defined in this study based on driving instructions, so even if the intersection structure changes, the action set remains unchanged. As a result, this approach can be readily applied to various n-way intersections.

3. The Suggested Method

A. Definition of the Problem

Optimizing intersection traffic light control is a challenging problem that requires taking into account many factors to achieve the best possible outcome. Minimizing traffic delay is a critical objective since delays lead to congestion, longer commute times, and increased air pollution. However, optimizing traffic signals is not just about reducing delays; It's crucial to ensure that signals are distributed fairly to all intersecting sides.

One way to ensure that signals are distributed equally is to use the standard deviation of queue lengths as a parameter. This parameter represents the variation in the number of vehicles parked on each side of the intersection. A low standard deviation implies that the queue lengths are the same on all sides, which indicates that the distribution is fair. On the other hand, a high standard deviation indicates that the traffic is not distributed evenly, and some drivers may face longer waiting times.

To minimize traffic delay and ensure a fair signal distribution, throughput can also be used as a parameter. The throughput of an intersection is the total number of cars that can pass through in a predetermined amount of time. By enabling more vehicles to pass through the intersection, throughput is enhanced, reducing delays and improving traffic flow.

By optimizing traffic signals based on both the throughput and the standard deviation of queue lengths, it is possible to achieve a balance between reducing delays and ensuring a fair distribution of signals. Traffic simulation models and machine learning techniques are used to enhance the timing of traffic signals and improve the efficiency of traffic flow.

d_{ql} - standard deviation of the queue lengths
t_{inter} - time until the signal is back
l_{signal} - signal length

The model is presented in the following way:

$$\max \mathit{through\ put} \tag{1}$$

subject to:

$$d_{\text{ql}} \leq \varphi \tag{2}$$

$$l_{\text{signal} = } \text{c} \tag{3}$$

$$t_{\text{inter}} < \varphi'' \tag{4}$$

In Eq. (1), the model is being maximized. Throughput represents the number of vehicles passing through an intersection per hour. Equation (2) imposes the condition that the standard deviation of queue lengths must be less than or equal to the threshold value (φ) to ensure a fair distribution of signals. Equation (3) expresses the constancy of the signal duration. Eq. (4) indicates that the time interval between the end of the green light in a given direction and the start of the subsequent green light in the same direction is less than the predetermined threshold value (φ'').

B. Q-learning

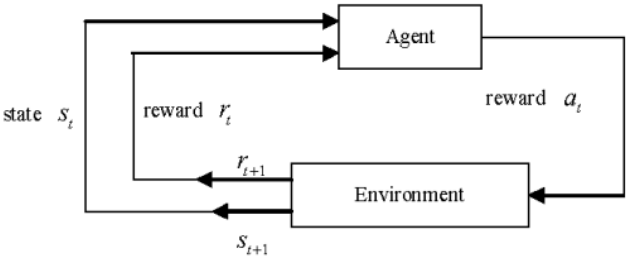


Fig. 1. Agent-environment interaction in reinforcement learning

Reinforcement learning is a method that can enhance its performance through previous learning experiences [14]. QL, a type of reinforcement learning, utilizes a trial-and-error method to navigate a complex and unpredictable environment and determine the optimal behavior based on past experiences [15]. QL involves three key components: state, action, and reward. The environment's current situation is the state, the behavior taken is the action, and the result of that behavior is the reward.

The interaction between agent and environment is illustrated as Fig. 1.

Eq. (5) illustrates how an action (a_t) in a state (s_t) leads to a transition to the next state (s_{t+1}).

$$S_t \xrightarrow{at} S_{t+1} \qquad (5)$$

The Q-table gets revised based on the previous value of (Q(s_t, a_t)) for the current state (st), action (at), reward (r_{t+1}), and the maximum values (max$a$ Q (s_{t+1}, a_{t+1}) from the new state (s_{t+1}) using a learning rate (η), as depicted in Equation (6).

$$Q(s_t,a_t) \leftarrow Q(s_t, a_t)+ \eta\cdot(r_{t+1}+\gamma\cdot \max Q(s_{t+1}, a_{t+1})-Q(s_t,a_t) \qquad (6)$$

The convergence and speed of algorithms can be influenced by the learning rate and discount factor (γ). During each iteration, the learning rate specifies how much the movement in the ideal direction will shift, and the discount factor reflects how important the next state will be. Typically, these parameters are set within a range of 0 to 1. A reduced learning rate results in more refined learning, but the convergence process is longer. In contrast, a larger learning rate leads to faster convergence, but it can overshoot the optimal solution. The discount factor determines the balance between past and new information, with values closer to 0 emphasizing past knowledge and those closer to 1 emphasizing new data. For instance, a discount factor of 0 indicates that no new learning takes place, relying solely on previous knowledge for decision-making. On the other hand, a discount factor of 1 means that only the most recent information is used to make a decision.

The QL algorithm employs two different methods for decision-making: exploitation and exploration. Exploitation involves selecting an action that maximizes the reward based on previously learned information. This method aims to make the best possible choice using available knowledge. However, exploitation has limitations since it is prone to local optimization, making it difficult to achieve global optimization. To overcome this, exploration is used as it is related to global search and aims to find more diverse options for decision-making. Exploration randomly selects an action to acquire new experiences that can enhance the decision-making process. The ε-greedy selection is used for exploration, and the ε parameter, ranging between 0 and 1, controls the randomness. In this method, utilizing exploitation, to determine the direction to drive when the next signal is green based on learned information, while exploration selects a random action to receive a signal, enabling the algorithm to acquire diverse experiences to make better decisions.

C. Traffic Light Controller used Q-learning

The aim of this research is to optimize the processing of vehicles at intersections by increasing throughput and reducing the standard deviation of queue lengths. Throughput is defined as the total of vehicles that can be handled at the intersections within a given time frame, while the standard deviation of queue lengths is used to ensure traffic balance across all road directions.

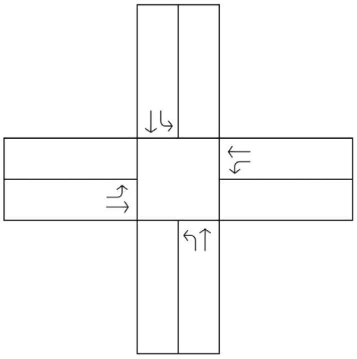


Fig. 2. 4-Lanes intersection

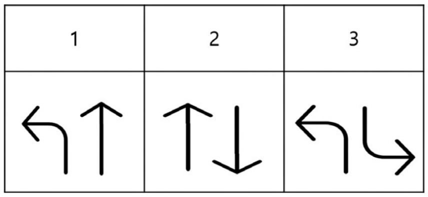


Fig. 3. Action set

Table 1
State transition table

Current state	Action	Next state
s1	a1 a3	{s3, s4, s5, s6, s7, s8} {s2, s3, s4, s6, s7, s8}
s2	a1 a2	{s3, s4, s5, s6, s7, s8} {s1, s3, s4, s5, s7, s8}
s3	a1 a3	{s1, s2, s5, s6, s7, s8} {s1, s2, s4, s5, s6, s8}
s4	a1 a2	{s1, s2, s5, s6, s7, s8} {s1, s2, s3, s5, s6, s7}
s5	a1 a3	{s1, s2, s3, s4, s7, s8} {s2, s3, s4, s6, s7, s8}
s6	a1 a2	{s1, s2, s3, s4, s7, s8} {s1, s3, s4, s5, s7, s8}
s7	a1 a3	{s1, s2, s3, s4, s5, s6} {s1, s2, s4, s5, s6, s8}
s8	a1 a2	{s1, s2, s3, s4, s5, s6} {s1, s2, s3, s5, s6, s7}

Table 2
Simulation parameters

Parameter	Value	Unit
The length of a road	5	km
The average speed of vehicles	10	km/h
The length of a vehicle	4.7	m
The distance between vehicles	1.3	m
The learning rate	0.1	–
The discount factor	0.9	–
ε	0.1	–
δ	0.5	–
Epoch	20	–

1) State and action

The number of lanes on the road determines the number of states at a given intersection. We presume that a right turn is also possible in the rightmost straight lane, if necessary, when there are two options on a road: left turn and straight. Therefore, an intersection with n lanes has 2*n states. The number of states is equivalent to the total number of lanes at an intersection. For example, a 4-lane intersection has 8 states, while a 5-lane intersection has 10 states. In Fig. 2, we can see that a 4-direction intersection has a total of 8 lanes, and the way with the highest total of number of vehicles determines the current state. From the set of actions that includes the direction of the current state, we choose the action that can result in the highest number of vehicles being able to move through the intersection.

Fig. 3 defines the three available action sets for a road. Each action set corresponds to a direction and selecting one will result in a green light only for that particular direction. Although the state may change in an n-direction intersection, The action set doesn't change.

2) Reward

To reduce traffic congestion at an intersection, two parameters are utilized in configuring the reward function: the standard deviation of queue lengths in each direction and the throughput. A small standard deviation implies that all lanes have a similar queue length, leading to a balanced signal distribution and queue length. Intersections with reliable signals can accommodate a larger number of cars, making throughput a significant parameter. The value of τtp, which is an exponential function, decreases as the throughput value increases. In other words, higher throughput leads to lower τtp. The weighting factor α, which is a sigmoid function ranging from 0 to 1, adapts to the arrival rate of vehicles per hour. As more vehicles arrive, α approaches 1.

$$f(t) = \alpha \cdot (d_{\text{ql}}) + (1 - \alpha) \cdot (\tau^{\text{tp}}) \qquad (7)$$

$$r_{\text{t}} = log\delta\ (f(t)) \qquad (8)$$

As demonstrated in Eq. (7), the function is represented by the throughput and the standard deviation of queue lengths. Maximizing the reward (r) is achieved by minimizing the value of $f(t)$. The variable δ represents the base of the logarithmic function and its values range from 0 to 1.

The MDP diagram depicted in Figure 4 illustrates how an agent interacts with its environment, specifically at an intersection. At each time t, the environment transmits perceived information, which includes the queue lengths (q_l) and throughput (t_p) of all lanes within the intersection, to the agent. Subsequently, the agent computes the reward (r_t) obtained when transitioning from the previous state $(s_t\text{−}1)$ to the current state (s_t) and updates the Q-table accordingly. Following this, The current state (s_t) is then adjusted depending on the information perceived in the lanes with the lengthy queue. The action (at) is decided by the agent with the highest reward and relayed back to the environment. This action signifies the lanes (d_c) in which the green signal is activated at the current state. Ultimately, the green signal is enabled in the designated lanes at the street intersection. It is assumed that the environment operates in a deterministic manner. For detailed information, please refer to Table 1, which presents the state transition table.

4. Comparative Experimentation

A. Model of Simulation

The proposed method was evaluated at a 4-direction intersection, as depicted in Figure 2. The roadway had a length of 5 km, and the average speed of the vehicles was 10 km/h. Assuming that each vehicle was approximately 4.7 m long with a spacing of 1.3 m between them, one vehicle would occupy a queue space of up to 6 m.

To conduct the experiment, traffic data was gathered using a VISSIM simulator. Given the stochastic and dynamic nature of the traffic environment, the algorithm's learning rate was set to 0.1, and the exploration parameter (ε) was set to 0.1. Based on the initial experiments, the discount factor (δ) was established at 0.5. The discount factor of 0.9 was chosen because historical traffic data in Traffic Signal Control (TSC) applications holds less relevance than real-time data. For further details regarding the simulation parameters, please refer to Table 2.

To assess the effectiveness of the suggested algorithms and QL, simulation experiments were carried out. The assessment focused on three key metrics: queue length, standard deviation of queue lengths, and waiting time. Queue length represents the total number of vehicles waiting on the lane, while the standard deviation of queue lengths provides an indication of the balance between different directions of traffic flow. A lower standard deviation implies a better balance. Lastly, waiting time measures the duration for which a vehicle remains stationary before crossing the intersection.

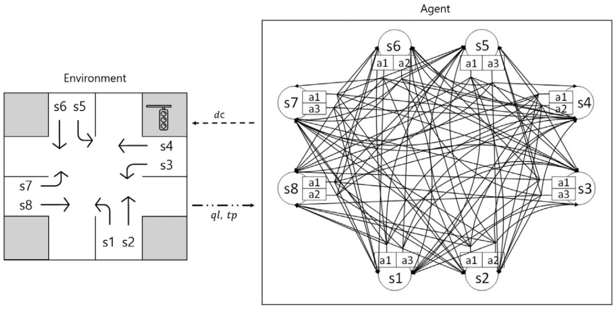
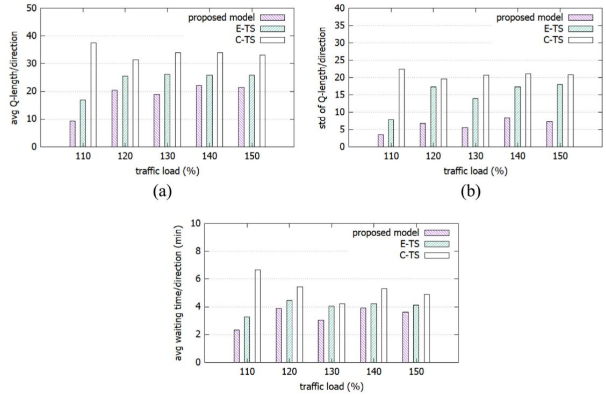


Fig. 4. MDP diagram



(a)

(b)

(c)

Fig. 5. Performance comparison: (a) queue length, (b) queue length deviation, and (c) waiting time μ

B. Result and Analysis

The study compared the proposed model with two other QL models. The first model, referred to as "extension traffic signal" (E-TS), utilized the order of green lights on the road to make decisions on whether to extend or shorten the duration of the green light [16]. It can be considered as an upgraded version of fixed-time traffic lights. The second model, known as "cluster-traffic signal" (C-TS), employed a cluster-based QL technique to control traffic signals [17] . In the C-TS model, vehicles were grouped into clusters, and each cluster was allowed to cross the intersection during the green light phase. The reward for the C-TS model was calculated as the combined value of the queue length and waiting time, aiming to minimize congestion and delays.

To ensure precise analysis of the experiment's results, the measurement units were adjusted to align with the specific unit of direction, enhancing the accuracy and relevance of the analysis.

As depicted in Figure 5(a), the proposed algorithm exhibited superior performance compared to E-TS and C-TS models at 150% traffic load, with an average queue length approximately 20% shorter than E-TS and 65% shorter than C-TS. The flexibility and adaptability of the proposed method, which features an undetermined signal system, enabled it to outperform the other models as the total of arrivals increased. Figure 5(b) illustrates that the proposed algorithm achieved an average standard deviation value roughly 45% lower than E-TS and 70% lower than C-TS. This improved performance can be attributed to the fact that the proposed algorithm incorporates the standard deviation of queue lengths in the calculation of rewards. Additionally, Figure 5(c) contrasts the typical waiting time for each vehicle, revealing that the proposed algorithm boasted 13% less waiting time compared to E-TS and 35% less than C-TS, on average. Taken together, the results obtained from analyzing queue length, standard deviation of the queue length, and average waiting time confirm that the proposed algorithm effectively mitigates vehicle delays in a more balanced manner.

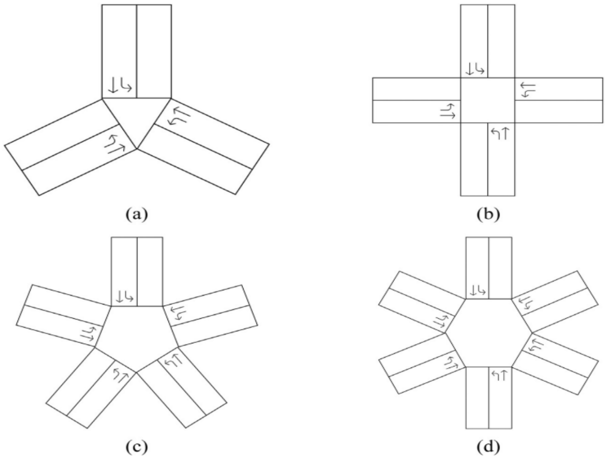


Fig. 6. n-lanes intersection: (a) 3-lanes, (b) 4-lanes, (c) 5-lanes, and (d) 6-lanes

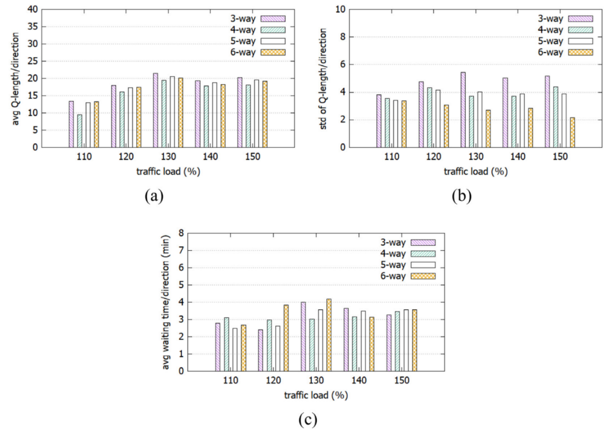


Fig. 7. Road initialization experiment per hour: (a) queue length, (b) standard deviation of queue lengths, and (c) waiting time

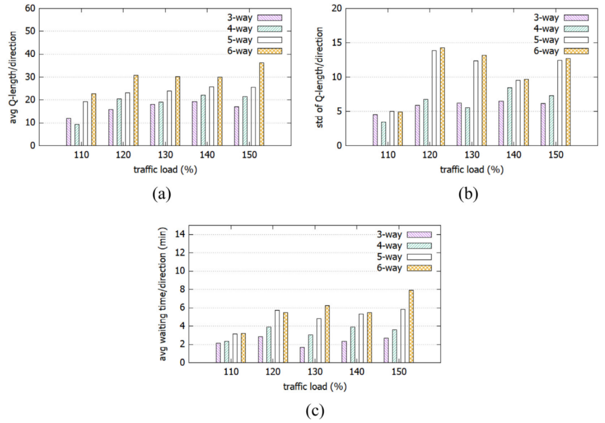


Fig. 8. Results of a 24-hour road initialization experiment, showcasing (a) the queue length, (b) the variability of queue lengths measured by the standard deviation, and (c) the waiting time

Figure 7 presents the results obtained from the experiment involving hourly road initialization. In Figure 7(a), it can be observed that the performance of the 3, 4, 5, and 6-lane intersections is quite similar. For a traffic load of 110%, the values are distributed between 09 and 14, while for a traffic load of 150%, the values are distributed between 19 and 20. Moving on to Figure 7(b), it is evident that the standard deviation of queue length is higher in the 3-way experiment compared to the other intersections. This can be attributed to the 3-way intersection having fewer roads, which imposes limitations on the possible action set. Specifically, the 3-way intersection lacks the option to execute a third action that enables both left turns from different roads. This restriction in action selection hinders the ability to control traffic in various combinations, occasionally resulting in the selection of unnecessary action combinations. Consequently, the standard deviation is higher in the 3-lane intersection. However, the absolute difference in values is not significantly notable. In terms of average waiting time, consistent results are observed across all intersections, indicating a fair distribution of signals for each direction.

Figure 8 illustrates the results of the road initialization experiment conducted every 24 hours. In Figure 8(a), it can be observed that the queue length of the 6-lane intersection surpasses that of the other intersections. This can be attributed to the 6-way intersection having the highest number of directions that require signal control within a limited time. Additionally, it is influenced by the accumulated delay resulting from the highest traffic load. Turning to Figure 8(b), it is evident that the standard deviation is higher for the 5-way and 6-lance intersections compared to the 3-way and 4-way intersections, particularly at 120% and 130% traffic loads. This indicates that the 5-way and 6-way intersections involve the calculation of more directions. Consequently, the standard deviation, which represents the balance of traffic on the roads, is less balanced in the 5-way and 6-way intersections compared to the 3-way and 4-lance intersections. The same, in Figure 8(c), an increase in waiting time is observed for the 5-lance and 6-lance intersections. Consequently, the proposed method demonstrates expandability to different intersection structures.

5. Conclusion

The present study introduced a traffic light control system utilizing QL, with a focus on incorporating standard deviation of queue lengths and throughput as key parameters. In comparison to previous QL-based research, the proposed method demonstrated favorable efficiency in terms of reducing both the standard deviation of queue lengths and the queue length itself, leading to shorter waiting times. This suggests that the traffic light control system effectively grasped the traffic flow dynamics and distributed lights accordingly. Additionally, this research explored a traffic light control approach that can be adapted to different intersection structures, emphasizing its expandability.

References

[1] Lei Chai, Guojiang Shen and Wei Ye, "The Traffic Flow Model for Single Intersection and its Traffic Light Intelligent Control Strategy," 2006 6th World Congress on Intelligent Control and Automation, Dalian, 2006, pp. 8558-8562.

[2] A. A. Zaidi, B. Kulcsár and H. Wymeersch, "Back-Pressure Traffic Signal Control with Fixed and Adaptive Routing for Urban Vehicular Networks," in IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 8, pp. 2134-2143, Aug. 2016.

[3] Mannion, P., Duggan, J., Howley, E. (2016). An Experimental Review of Reinforcement Learning Algorithms for Adaptive Traffic Signal Control. In: McCluskey, T., Kotsialos, A., Müller, J., Klügl, F., Rana, O., Schumann, R. (eds) Autonomic Road Transport Support Systems. Autonomic Systems. Birkhäuser, Cham.

[4] X. Cheng and Z. Yang, "Intelligent Traffic Signal Control Approach Based on Fuzzy-Genetic Algorithm," 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Jinan, China, 2008, pp. 221-225.

[5] Junxia Gao, Jiangeng Li, Xiaohua Zhao and Yangzhou Chen, "Two-stage fuzzy control of urban isolated intersection signal for complex traffic conditions," Fifth World Congress on Intelligent Control and Automation, Hangzhou, China, 2004, pp. 5287-5291, vol. 6.

[6] W. Yi-Fei and G. Zheng, "Research on Polling Based Traffic signal Control Strategy with Fuzzy Control," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 2018, pp. 500-504.

[7] L. Zhi-Yong and M. Feng-wei, "On-line Reinforcement Learning Control for Urban Traffic Signals," 2007 Chinese Control Conference, Zhangjiajie, China, 2007, pp. 34-37.

[8] Q. Yang et al., "Adaptive Multimodal Continuous Ant Colony Optimization," in IEEE Transactions on Evolutionary Computation, vol. 21, no. 2, pp. 191-205, April 2017.

[9] Askerzade I. N., and Mahmood M., "Control the extension time of traffic light in single junction by using fuzzy logic," *International Journal of Electrical & Computer Sciences IJECS–IJENS*, 10, 48–55, 2010.

[10] Y. Liao and X. Cheng, "Study on Traffic Signal Control Based on Q-Learning," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, 2009, pp. 581-585.

[11] L. Li, Y. Lv and F. -Y. Wang, "Traffic signal timing via deep reinforcement learning," in IEEE/CAA Journal of Automatica Sinica, vol. 3, no. 3, pp. 247-254, 10 July 2016.

[12] Mousavi, S.S., Schukat, M. and Howley, E. (2017), Traffic light control using deep policy-gradient and value-function-based reinforcement learning. IET Intell. Transp. Syst., 11: 417-423.

[13] Anon B.C. Silva, D. Oliveira, A.L.C. Bazzan and E.W. Basso, "Adaptive Traffic Control with Reinforcement Learning," Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS06), ACM Press, 2006, pp. 80–86.

[14] G. Li, R. Gomez, K. Nakamura and B. He, "Human-Centered Reinforcement Learning: A Survey," in IEEE Transactions on Human-Machine Systems, vol. 49, no. 4, pp. 337-349, Aug. 2019.

[15] D. Pandey and P. Pandey, "Approximate Q-Learning: An Introduction," 2010 Second International Conference on Machine Learning and Computing, Bangalore, India, 2010, pp. 317-320.

[16] Chin Y. K., Bolong N., Kiring A., Yang S. S. and Teo K. T. K., "Q-learning based traffic optimization in management of signal timing plan," in *International Journal of Simulation, Systems, Science & Technology* 12 29–35, 2011.

[17] A. Boukerche, D. Zhong and P. Sun, "A Novel Reinforcement Learning-Based Cooperative Traffic Signal System Through Max-Pressure Control," in IEEE Transactions on Vehicular Technology, vol. 71, no. 2, pp. 1187-1198, Feb. 2022.