JTAE

Journal of Technology in Aerospace Engineering

E- ISSN: 2676-4253

Vol. 10, No. 3, pp. 104-129, 2026

https://doi.org/10.22034/jtae.2026.10.3.8

Journal Homepage: https://www.jtae.ari.ac.ir



Original Research Paper

Evaluation and Comparison of Deterministic Search and Reinforcement Learning Algorithms for Routing in Network and Urban Environments

Iman Shafieenejad [1*] and Faezeh Shokri [2]

Aerospace Research Institute, Ministry of Science, Research and Technology, Tehran, Iran

ARTICLE INFO

Article History:

Received 23 February 2026
Revised 20 June 2026
Accepted 25 August 2026
Available Online 27August 2026

Keywords:

Optimal routing
A* Algorithm
Q-Learning
Double Q-learning
Statistical analysis

ABSTRACT

This study presents a comprehensive comparative evaluation of three routing algorithms in static network environments with complex obstacles. The algorithms include A* as a deterministic heuristic search method, Q-learning as a classical reinforcement learning approach, and Double Q-learning as an improved version designed specifically to reduce overestimation bias. Experiments were conducted in two distinct environments: a simple 25×25 grid network and a complex 40×40 urban environment. Algorithm performance was systematically assessed using multi-dimensional criteria including path length, execution time, success rate, path smoothness measured by number of turns, and learning stability across multiple episodes. To ensure statistical reliability of the results, each algorithm underwent 30 repeated independent runs with advanced statistical analyses including 95% confidence intervals, paired t-tests, and ANOVA to validate significance of observed differences. Findings revealed that A* consistently outperformed other algorithms across all quantitative metrics, generating optimal and smoothest paths with average length of 42 steps, minimal execution time of 0.0043 seconds, and only 7 turns. Although both reinforcement learning algorithms achieved 100% success rates in all experimental conditions, they required substantially longer execution times and produced lower-quality paths with considerably more turns (19 and 20 turns respectively). In the complex urban environment, performance disparities intensified significantly, with Double Q-learning optimality dropping from 91.3% to 90%. The results showed that reinforcement learning-based algorithms are better suited to unknown, dynamic, and uncertain environments due to their adaptive capabilities, whereas the A* algorithm achieves superior performance in static and well-defined environments.

* Corresponding Author's E-mail: shafieenejad@ari.ac.ir

How to Cite this Article:
I. Shafieenejhad and F. Shokri, "Evaluation and comparison of deterministic search and reinforcement learning algorithms for routing in network and urban environments," *Journal of Technology in Aerospace Engineering*, vol. 10, no. 3, pp. 104-129, 2026, (in Persian), https://doi.org/10.22034/jtae.2026.10.3.8.

COPYRIGHTS
© 2026 by the authors. Published by ARI. This article is an open access article distributed under the terms and conditions of The Creative Commons Attribution 4.0 International (CC BY 4.0)



فناوری در مهندسی هوافضا

سال ۱۴۰۵، دوره ۱۰، شماره ۳، صفحه ۱۰۴-۱۲۹

https://doi.org/10.22034/jtae.2026.10.3.8

Journal Homepage: https://www.jtae.ari.ac.ir

JTAE

شاپای الکترونیکی: ۲۶۷۶-۴۲۵۳

مقاله پژوهشی

ارزیابی و مقایسه الگوریتم‌های جستجوی قطعی و یادگیری تقویتی در مسیریابی محیط‌های شبکه‌ای و شهری

ایمان شفیعی‌نژاد[1]* و فائزه شکری[2]

۱- استادیار، پژوهشگاه هوافضا، وزارت علوم، تحقیقات و فناوری، تهران، ایران

۲- دانشجوی دکتری، پژوهشگاه هوافضا، وزارت علوم، تحقیقات و فناوری، تهران، ایران

اطلاعات مقاله

تاریخچه مقاله:

دریافت ۴ اسفندماه ۱۴۰۴

بازنگری ۳۰ خرداد ۱۴۰۵

پذیرش ۰۳ شهریور ۱۴۰۵

اولین انتشار ۰۵ شهریور ۱۴۰۵

واژه‌های کلیدی:

مسیریابی بهینه

آاستار

یادگیری کیو

یادگیری دوگانه کیو

تحلیل آماری

چکیده

این پژوهش به ارزیابی مقایسه‌ای جامع سه الگوریتم مسیریابی در محیط‌های شبکه‌ای ایستا با موانع پیچیده می‌پردازد. الگوریتم‌ها شامل آاستار به‌عنوان روش قطعی مبتنی بر جستجوی هیوریستیک، یادگیری کیو به‌عنوان رویکرد کلاسیک یادگیری تقویتی و یادگیری دوگانه کیو به‌عنوان نسخه بهبودیافته با هدف کاهش بیش‌برآورد هستند. آزمایش‌ها در دو محیط متفاوت شامل یک شبکه ساده ۲۵×۲۵ و یک محیط شهری پیچیده ۴۰×۴۰ انجام شد و عملکرد هر الگوریتم بر اساس معیارهای چندبعدی شامل طول مسیر، زمان اجرا، نرخ موفقیت، کیفیت مسیر (تعداد پیچش‌ها) و پایداری یادگیری مورد سنجش قرار گرفت. به‌منظور اطمینان از قابلیت اتکای نتایج، هر الگوریتم به‌صورت ۳۰ اجرای مکرر آزمایش شد و تحلیل‌های آماری پیشرفته شامل محاسبه فواصل اطمینان ۹۵ درصد، آزمون تی-تست جفتی و تحلیل واریانس بر روی داده‌ها اعمال گردید. یافته‌ها نشان داد که الگوریتم آاستار در تمامی معیارها عملکرد برتری داشته و با میانگین طول مسیر ۴۲ گام، زمان اجرای ۰/۰۰۴۳ ثانیه و ۷ پیچش، بهینه‌ترین و نرم‌ترین مسیر را تولید می‌کند. الگوریتم‌های یادگیری تقویتی اگرچه به نرخ موفقیت ۱۰۰ درصدی دست یافتند، اما به زمان اجرای به مراتب بیشتری نیاز داشتند و مسیرهای با کیفیت پایین‌تری (با ۱۹ و ۲۰ پیچش) ایجاد کردند. در محیط شهری، اختلاف عملکرد الگوریتم‌ها تشدید شد و افت بهینگی یادگیری دوگانه کیو کاهش یافت. یافته‌ها نشان دادند که الگوریتم‌های مبتنی بر یادگیری تقویتی در محیط‌های ناشناخته، پویا و دارای عدم قطعیت از قابلیت انطباق بیشتری برخوردارند، در حالی‌که الگوریتم آاستار در محیط‌های ایستا و شناخته‌شده عملکرد بهینه‌تری دارد.

پست الکترونیکی نویسنده مسئول: shafieenejad@ari.ac.ir

How to Cite this Article:

I. Shafieenejhad and F. Shokri, "Evaluation and comparison of deterministic search and reinforcement learning algorithms for routing in network and urban environments," *Journal of Technology in Aerospace Engineering*, vol. 10, no. 3, pp. 104-129, 2026, (in Persian), https://doi.org/10.22034/jtae.2026.10.3.8.

COPYRIGHTS

© 2026 by the authors. Published by ARI. This article is an open access article distributed under the terms and conditions of The Creative Commons Attribution 4.0 International (CC BY 4.0)

۱. مقدمه

مسیریابی بهینه در محیط‌های پیچیده و دارای موانع، از چالش‌های اساسی در حوزه‌های مختلفی همچون رباتیک خودمختار، سیستم‌های ناوبری و مدیریت لجستیک است. طراحی الگوریتم‌های کارآمد برای یافتن مسیر بهینه نه تنها کاربردهای عملی گسترده‌ای در صنعت و فناوری دارد، بلکه از منظر تئوریک نیز زمینه‌ای برای توسعه روش‌های محاسباتی و یادگیری ماشین فراهم می‌سازد.

در دهه‌های اخیر، دو رویکرد اصلی برای حل مسئله مسیریابی توسعه یافته‌اند: الگوریتم‌های قطعی مبتنی بر جستجو مانند آاستار و رویکردهای یادگیری تقویتی نظیر یادگیری کیو. آاستار به‌عنوان یک الگوریتم ابتکاری به‌دلیل تضمین بهینگی و کارایی در محیط‌های شناخته‌شده، به‌طور گسترده‌ای مورد استفاده قرار گرفته است. در مقابل، الگوریتم‌های یادگیری تقویتی با قابلیت یادگیری از تعامل با محیط، امکان مسیریابی در موقعیت‌های ناشناخته و پویا را فراهم می‌کنند. نسخه‌های بهبودیافته‌ای مانند یادگیری کیو دوگانه نیز برای کاهش مشکلاتی مانند بیش‌برآورد مقادیر عمل ارائه شده‌اند. واتکینز در مطالعه خود الگوریتم‌های یادگیری تقویتی کوانتومی مبتنی بر ارزش را معرفی کرد. این الگوریتم‌ها با استفاده از الگوریتم گروور معادلات را به‌روزرسانی می‌کنند. الگوریتم‌ها به دو دسته تقسیم می‌شوند: مبتنی بر تابع ارزش حالت و مبتنی بر تابع ارزش عمل. نتایج نشان می‌دهد که الگوریتم‌های مبتنی بر تابع ارزش عمل سریع‌تر همگرا می‌شوند، بازده بیشتری دارند و امکان استفاده از نرخ یادگیری بالاتر را فراهم می‌کنند [۱]. واتکینز و دایان الگوریتم یادگیری کیو را به‌عنوان یک روش ساده و افزایشی برای یادگیری بهینه تصمیم‌ها در محیط‌های مارکوفی کنترل‌شده معرفی کرده‌اند. این الگوریتم با بهبود پی‌درپی مقادیر کیو، کیفیت اعمال مختلف را ارزیابی می‌کند. نتایج نشان می‌دهد که یادگیری کیو با احتمال یک به مقادیر بهینه ارزش عمل‌ها همگرا می‌شود. همچنین به‌روزرسانی چند مقدار کیو به‌طور همزمان ارائه شده است [۲]. بورزآبادی و همکاران مجموعه‌ای از مسائل برنامه‌ریزی مسیر بهینه را با استفاده از رویکردی مبتنی بر نظریه اندازه حل کرده‌اند. در این روش ابتدا مسئله به‌صورت یک مسئله بهینه‌سازی فرموله شده و سپس با تعریف توابع کنترلی مصنوعی به مسئله کنترل بهینه تبدیل می‌شود. در ادامه با نگاشت مسئله به فضای اندازه‌های رادون مثبت، یک بازفرموله‌سازی انجام شده که منجر به تبدیل مسئله به یک مسئله برنامه‌ریزی خطی می‌شود. حل این مسئله امکان تولید مسیرهای تقریبی بهینه و بدون برخورد را فراهم کرده و کارایی روش در مثال‌های عددی مختلف نشان داده شده است [۳]. هاسلت و همکاران به بررسی مشکل بیش‌برآوردی مقادیر عمل در یادگیری کیو پرداخته‌اند. پژوهشگران نشان می‌دهند که ایده یادگیری کیو عمیق که ابتدا در محیط‌های جدولی معرفی شده بود، می‌تواند برای تقریب توابع در مقیاس بزرگ تعمیم یابد. با ارائه یک سازگارسازی خاص برای شبکه‌های کیو عمیق، الگوریتم حاصل نه تنها بیش‌برآوردی‌ها را کاهش می‌دهد بلکه باعث بهبود عملکرد می‌شود نگوین و همکاران الگوریتم‌های هم‌خط و ناهم‌خط یادگیری کیو را برای کنترل ربات‌های پاندول وارونه دوچرخ ارائه کردند. در این روش‌ها، داده‌ها بدون نیاز به مدل جمع‌آوری می‌شوند و سیستم به دو زیرسیستم تقسیم شده تا حجم داده‌ها کاهش یابد و سرعت محاسبات افزایش یابد. نتایج شبیه‌سازی نشان داد که تکنیک‌های غیر هم‌خط با کارایی داده‌ای بالاتر و دقت بیشتر، عملکرد بهتری در کنترل سیستم‌های پاندول وارونه دوچرخ دارند [۴]. وو و همکاران الگوریتم ترکیبی کلنی زنبور مصنوعی با یادگیری کیو را برای حل مسئله زمان‌بندی توزیع‌شده و انعطاف‌پذیر کارگاه با نگهداری پیشگیرانه ماشین‌ها ارائه کردند. این الگوریتم با مدل‌سازی خرابی ماشین‌ها تحت شرایط عملیاتی متغیر و استفاده از راهبرد نگهداری دوآستانه‌ای، زمان تکمیل کل را کاهش می‌دهد. عملگرهای جستجوی همسایگی به‌صورت تطبیقی توسط یادگیری کیو انتخاب شده و آزمایش‌ها نشان داد که روش پیشنهادی در مسائل با مقیاس‌های مختلف، عملکرد مؤثر و قابل‌توجهی دارد [۵]. گئو و همکاران الگوریتم کیو–ممتیک مبتنی بر یادگیری کیو و ممتیک را برای حل مسئله زمان‌بندی مشترک وسایل نقلیه در برداشت و حمل‌ونقل کشاورزی هوشمند ارائه کردند. الگوریتم با راهبرد مقداردهی اولیه ترکیبی، سه عملگر تقاطع تخصصی و جستجوی محلی تطبیقی هدایت‌شده توسط یادگیری کیو، جمعیت متنوعی تولید کرده و راه‌حل‌های با کیفیت بالا ارائه می‌دهد. نتایج آزمایش‌ها نشان داد که کیو–ممتیک عملکرد بهتری نسبت به الگوریتم‌های مرجع دارد و مؤلفه‌های الگوریتم به بهبود همگرایی و کیفیت راه‌حل کمک می‌کنند [۶]. سان و همکاران الگوریتم هیبریدی کیو و ای‌اودی‌وی[1] را برای مسیریابی چندمسیره در شبکه‌های پروازی معرفی کردند. این الگوریتم با استفاده از ارزیابی پایداری و پاداش دوگانه، همگرایی سریع‌تر و مسیریابی بهینه در مناطق خالی ارائه می‌دهد. نتایج نشان داد که این الگوریتم نسبت به پروتکل‌های ای‌اودی‌وی و کیوام‌آر[2]، تا ۴۰ درصد همگرایی سریع‌تر دارد و قابلیت اطمینان و عملکرد شبکه‌های پروازی را بهبود می‌بخشد [۷]. کیو و

1. Ad hoc On-Demand Distance Vector (AODV)

2. Q-learning based Multipath Routing (QMR)

همکاران الگوریتم کیوال‌اچ‌اچ[1] مبتنی بر یادگیری کیو را برای تخصیص وظایف اولویت‌بندی شده ارائه کردند. این الگوریتم با انتخاب خودکار ابتکارهای سطح پایین، هم‌زمان زمان تکمیل وظایف و زمان کل وظایف را به‌حداقل می‌رساند. نتایج شبیه‌سازی و مقایسه‌ها نشان داد که کیوال‌اچ‌اچ اثربخشی و کارایی بالایی در حل مسئله تخصیص وظایف دارد [۸]. دوان و همکاران الگوریتم چندهدفه کیوال‌سی‌پی‌اس‌او[2] مبتنی بر یادگیری کیو و بهینه‌سازی ازدحام ذرات چندهدفه را برای مسیریابی ربات‌ها در زمین‌های ناهموار ارائه کردند. این الگوریتم با سازگاری پویا پارامترها و عملگر تقاطع راهبردی، مسیرهای بهینه با کمترین طول و ناهمواری زمین تولید می‌کند. نتایج نشان داد که کیوال‌سی‌پی‌اس‌او نسبت به روش‌های پیشرفته، دقت، تنوع جمعیت و کارایی بهینه‌سازی بالاتری دارد و برای کاربردهای چندهدفه مسیریابی و تخصیص منابع مناسب است [۹]. گئو و همکاران الگوریتم کیو-ام‌اس‌اچ‌او[3] مبتنی بر یادگیری کیو و بهینه‌سازی چندهدفه را برای مسائل زمان‌بندی ارائه کردند. این الگوریتم با ساختارهای همسایگی ویژه و جستجوی متغیر مبتنی بر یادگیری کیو، کمینه‌سازی زمان تکمیل و میانگین زمان جریان را کاهش می‌دهد و در مقایسه با سایر روش‌ها عملکرد بهتری در حل مسئله دارد [۱۰]. ژانگ و همکاران الگوریتم کیوال‌ام‌اچ‌اچ‌ای[4] مبتنی بر یادگیری کیو و چندهدفه را برای زمان‌بندی ترکیبی توزیع‌شده با نگهداری پیشگیرانه و مصرف انرژی بهینه ارائه کردند. این الگوریتم با راهبردهای سطح‌بالا مبتنی بر یادگیری کیو زمان تکمیل و انتشار کربن را کاهش می‌دهد و در مقایسه با روش‌های پیشرفته چندهدفه، عملکرد بالایی در حل مسائل پیچیده زمان‌بندی انرژی- بهینه نشان می‌دهد [۱۱].

در ادامه مطالعات پیشین، ژو و همکاران الگوریتم کیو-ام‌پی‌سی‌ای-ال‌اس[5] مبتنی بر یادگیری کیو و تکامل چندجمعیتی با جستجوی محلی را برای زمان‌بندی ماشین‌های موازی با منابع تولید شبکه‌ای مشترک معرفی کردند. الگوریتم پیشنهادی با بهره‌گیری از جستجوی محلی دوگانه و تنظیم پویای پارامترها از طریق یادگیری کیو، تأخیر وزنی کل را کاهش می‌دهد. این رویکرد در مقایسه با الگوریتم‌های فراابتکاری پیشرفته، عملکرد قابل‌توجهی در بهبود برنامه‌ریزی انعطاف‌پذیر شبکه‌ای از خود نشان می‌دهد [۱۲]. زو و همکاران الگوریتم ژنتیک دو جمعیتی هدایت‌شده با کیوجی‌جی‌ای[6] را برای حل مسئله زمان‌بندی جریان تولید توزیع‌شده با ماشین‌هایی با کارایی پردازش فازی ارائه می‌کنند. نتایج تجربی بیانگر آن است که کیوجی‌جی‌ای نسبت به الگوریتم‌های استاندارد ان‌اس‌جی‌ای-۲[7]، پی‌پی‌او[8] و ام‌اوای‌ای/ دی[9]، بهبود قابل توجهی در معیارهای ابرحجم و فاصله نسلی دارد و تعادل بهتری بین پایداری راه‌حل و کارایی محاسباتی ایجاد می‌کند [۱۳]. اوزسویدان الگوریتم جستجوی حریصانه بهبودیافته با یادگیری کیو را برای حل مسئله زمان‌بندی جریان تولید هیبریدی انعطاف‌پذیر با زمان‌های آماده‌سازی وابسته به توالی معرفی می‌کند. پژوهشگر نشان می‌دهد که استفاده از روش ارزیابی وضعیت تجمعی و سازوکار یادگیری کیو، امکان تنظیم خودکار پارامترهای الگوریتم و رویه‌های جستجوی محلی را فراهم می‌کند. نتایج تجربی بیانگر آن است که روش پیشنهادی نسبت به الگوریتم‌های متداول و حل‌کننده سی‌پلکس[10] عملکرد برتری دارد و کارایی و قابلیت اطمینان بالاتری در مسائل زمان‌بندی واقعی و مصنوعی ارائه می‌دهد [۱۴]. دینچر و همکاران الگوریتم یادگیری کیو و بهینه‌سازی دسته‌ای ذرات چندهدفه فازی مولکولی را برای بهینه‌سازی راهبردهای سرمایه‌گذاری در فناوری‌های انرژی تجدیدپذیر ارائه می‌کنند. پژوهشگران نشان می‌دهند که راه‌حل‌های نوآورانه، بازدهی چرخه عمر و مدل‌های چرخشی انرژی غیرمتمرکز مهم‌ترین عوامل در بهبود عملکرد این فناوری‌ها هستند [۱۵]. ژانگ و همکاران الگوریتم خوشه‌بندی نیمه‌ثابت اس‌اف‌سی-کیوال-آی‌ای‌سی‌او[11] را برای حفظ تعادل انرژی در شبکه‌های حسگر بی‌سیم ارائه می‌کنند. پژوهشگران نشان می‌دهند که ترکیب یادگیری کیو و بهینه‌سازی کلنی مورچه، توزیع بار و طول عمر شبکه را بهبود می‌بخشد و اس‌اف‌سی-کیوال-آی‌ای‌سی‌او نسبت به الگوریتم‌های موجود عملکرد برتری دارد [۱۶]. لی و همکاران الگوریتم تکراری حریصانه‌ای مبتنی بر یادگیری کیو را برای زمان‌بندی گروهی چندسناریویی پیشنهاد کردند. در این روش، یادگیری کیو برای بهبود انتخاب جستجوی محلی و افزایش تنوع راه‌حل‌ها به‌کار رفت. نتایج نشان داد الگوریتم پیشنهادی از نظر کارایی و پایداری از روش‌های موجود برتر است [۱۷]. ژوان و همکاران برای زمان‌بندی جریان انعطاف‌پذیر توزیع‌شده با محدودیت بافر ترکیبی، الگوریتم

1. Q-Learning based Hyper-Heuristic (QLHH)
2. Multi-objective Q-Learning based Competitive Particle Swarm Optimization (MOQLCPSO)
3. Q-learning based Multi-objective Spotted Hyena Optimizer (Q-MSHO)
4. Q-Learning based Multi-objective Hyper-Heuristic Algorithm (QLMHHA)
5. Q-learning based Multi-Population Co-Evolutionary Algorithm with Local Search (Q-MPCEA-LS)
6. Q-learning Guided Genetic Algorithm (QGGA)
7. Non-dominated Sorting Genetic Algorithm II (NSGA-II)
8. Proximal Policy Optimization (PPO)
9. Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D)
10. IBM ILOG CPLEX Optimization Studio (CPLEX)
11. Stable Fuzzy Clustering with Q-learning based Improved Ant Colony Optimization (SFC-QL-IACO)

کیوال‌ایی‌دی‌ای[1] مبتنی بر یادگیری کیو را ارائه کردند. این روش با ترکیب یادگیری کیو و نگاشت ، تعادل میان جستجوی محلی و سراسری را برقرار کرده و در مقایسه با روش‌های موجود، نتایج دقیق‌تر و کارآمدتری در کاهش هزینه‌های تأخیر ارائه می‌دهد [۱۸]. وانگ و همکاران یک الگوریتم ژنتیک چندهدفه تطبیقی مبتنی بر یادگیری کیو برای بهینه‌سازی توازن خطوط جداسازی پسماندهای الکترونیکی توسعه دادند. این روش با تنظیم پویای راهبرد جستجو از طریق یادگیری کیو، مصرف انرژی و عدم‌تعادل ناشی از شکست وظایف را کاهش می‌دهد. نتایج آزمایش‌ها نشان داد که مدل پیشنهادی ضمن بهبود کارایی خط، درآمد را تا حدود ۱۱/۹۸ درصد افزایش می‌دهد [۱۹]. شبیه‌سازی واقع‌گرایانه محیط، فاصله میان نتایج شبیه‌سازی و رفتار واقعی خودرو را کاهش می‌دهد. این سکو از آاستار برای مسیر سراسری، نوار کشسان زمان‌بندی‌شده برای مسیر محلی و منحنی‌های بزیه برای بهینه‌سازی مسیر استفاده می‌کند. ویژگی‌های فیزیکی خودرو در فرایند بهینه‌سازی لحاظ شده و نتایج نشان می‌دهد مسیرهای بهینه‌شده در محیط واقعی با شبیه‌سازی تطابق بالایی دارند. این رویکرد ایمنی، تطبیق‌پذیری و قابلیت عملیاتی الگوریتم‌های مسیریابی خودران را افزایش می‌دهد [۲۰]. زنگ و همکاران یک الگوریتم خوشه‌بندی نوین مبتنی بر سازوکار مسیریابی کلونی مورچگان ارائه کردند که نقاط داده را با استفاده از مورچگان خوشه‌بندی می‌کند. این روش با استخراج بهتر ویژگی‌های داده و افزایش خطا، دقت خوشه‌بندی را نسبت به الگوریتم‌های سنتی بهبود می‌دهد. الگوریتم به‌صورت یک سامانه پویا با مدل‌های تولید کلونی، مسیریابی و خوشه‌بندی رنگ‌آمیزی طراحی شده و آزمایش‌ها روی داده‌های شبیه‌سازی و واقعی نشان داد عملکرد آن در بیشتر شاخص‌های ارزیابی برتر از الگوریتم‌های مقایسه‌ای است [۲۱].

جیوفرید و همکاران یک مطالعه تجربی بر روی الگوریتم‌های یادگیری چندعاملی برای مسیریابی هم‌زمان انجام دادند که مسیرهای بدون برخورد برای چندین عامل را محاسبه می‌کند. این پژوهش الگوریتم‌های مبتنی بر یادگیری را با روش‌های سنتی جست‌وجو و برنامه‌ریزی مقایسه کرده و کاربردپذیری آن‌ها در محیط‌های واقعی را ارزیابی می‌کند. نتایج نشان داد برخی الگوریتم‌های یادگیری چندعاملی عملکردی نزدیک به روش‌های سنتی با کیفیت مسیر مشابه و تلاش محاسباتی کمتر دارند که نشان‌دهنده پتانسیل بالای آن‌ها برای کاربردهای عملی است [۲۲]. ژو و همکاران یک نسخه بهینه‌شده از الگوریتم یادگیری کیو برای مسیریابی محلی ربات‌های متحرک ارائه کردند که با بهبود مقداردهی اولیه جدول کیو، سیاست انتخاب عمل و تابع پاداش، سرعت همگرایی و کارایی مسیریابی را افزایش می‌دهد. این روش از تکنیک آرام‌اس‌پراپ[2] برای تنظیم پویای نرخ یادگیری استفاده کرده و عملکرد ربات در محیط‌های پیچیده آزمایش شد. نتایج شبیه‌سازی نشان داد الگوریتم پیشنهادی در معیارهای گام‌ها، پاداش اکتشافی، تغییر نرخ یادگیری و زمان اجرا نسبت به الگوریتم‌های موجود بهبود قابل توجهی دارد [۲۳]. گائو و همکاران مسئله مسیریابی چندعاملی با رعایت محدودیت‌های زمانی را بررسی کردند و دو الگوریتم جستجوی مبتنی بر بهینه‌سازی و تقریب را ارائه دادند: سی‌بی‌اس[3] مبتنی بر محدودیت و اس‌بی‌اس[4]. با هدف بهبود مقیاس‌پذیری در محیط‌های پرتراکم، آن‌ها الگوریتم مبتنی بر یادگیری تقویتی پی‌دی‌سی‌ال[5] را معرفی کردند که از سازوکارهای پیشرفته بهره می‌برد. نتایج تجربی نشان داد الگوریتم اس‌بی‌اس بهینه عمل کرده و در محیط‌های پر از موانع موفقیت بالاتری دارد، در حالی که پی‌دی‌سی‌ال راه‌حل‌های قابل قبولی را با سرعت بیشتر ارائه می‌دهد [۲۴]. ونو و گوروسامی یک مرور جامع بر الگوریتم‌های مسیریابی برای ناوبری خودکار ارائه دادند که روش‌های کلاسیک، فرابتکاری و مبتنی بر هوش مصنوعی را بررسی می‌کند. این مطالعه چالش‌های محیط‌های پویا، محدودیت‌های غیرهولونومیک و سطوح متفاوت دانش محیطی را تحلیل کرده و نقاط قوت و محدودیت هر دسته الگوریتم را برجسته می‌کند. همچنین روندهای نوظهور مانند ادغام یادگیری ماشین و یادگیری تقویتی و مسیرهای تحقیقاتی آینده برای بهبود تطبیق‌پذیری و عملکرد سیستم‌های مسیریابی در محیط‌های پیچیده مورد بحث قرار گرفته است [۲۵]. آسلان و دمیرجی یک الگوریتم پلاسمای ایمنی با یادگیری کیو برای مسیریابی پهپادها ارائه کردند. این الگوریتم با بهبود روند درمان مدل‌شده و حذف پارامترهای کنترل اختصاصی، کارایی جستجو را افزایش می‌دهد. آزمایش‌ها در دوازده سناریوی مختلف نشان داد مسیرهای یافت‌شده توسط کیو–لیپا[6] عملکرد بهتری نسبت به سایر تکنیک‌های بهینه‌سازی هوشمند دارند و مدیریت مبتنی بر یادگیری تقویتی تأثیر مثبت بر عملکرد کلی الگوریتم دارد [۲۶]. وانگ و همکاران الگوریتم آرجی‌او– آرآرتی استار[7] را برای مسیریابی پهپاد در محیط‌های پیچیده ارائه کردند که با نمونه‌برداری احتمالی منطقه‌ای، تنظیم گام تطبیقی، میدان پتانسیل مصنوعی بهینه‌شده، کارایی و همگرایی مسیر را بهبود می‌بخشد. این الگوریتم با استفاده از توسعه دوطرفه درخت، بهینه‌سازی حریصانه و محدودیت‌های قابلیت اجرا، کیفیت مسیر و

1. Q-Learning based Estimation of Distribution Algorithm (QLEDA)
2. Root Mean Square Propagation (RMSprop)
3. Conflict-Based Search (CBS)
4. Simulation-Based Search (SBS)
5. Priority-based Deep Coordination Learning (PDCL)
6. Q-LIPA
7. Regionally Greedy Optimization - Rapidly-exploring Random Tree Star (RGO-RRT*)

سرعت جستجو را افزایش می‌دهد. شبیه‌سازی‌ها نشان داد آرجی‌او- آرآرتی استار نسبت به الگوریتم‌های معیار، مسیرهای کوتاه‌تر و زمان برنامه‌ریزی کمتری ارائه می‌دهد و برای ناوبری پهپاد در محیط‌های پیچیده بسیار مناسب است [۲۷]. نی و همکاران یک الگوریتم کلنی زنبور مصنوعی یکپارچه چند راهبرد مبتنی بر یادگیری کیو برای مسیریابی وسایل نقلیه ارائه کردند که با بهره‌گیری از راهبردهای جستجوی متنوع و انتخاب هوشمندانه آن‌ها از طریق یادگیری کیو، عملکرد جستجو را بهبود می‌بخشد. این الگوریتم با پیکربندی‌های مؤثرتر حالت و عمل، نسبت به روش‌های قبلی توانایی بهینه‌سازی بالاتری دارد. ارزیابی‌ها روی توابع معیار و کاربردهای مسیریابی نشان داد الگوریتم کلنی زنبور مصنوعی یکپارچه چندراهبرد مبتنی بر یادگیری کیو در حل مسائل عددی و عملی، عملکرد برتری نسبت به سایر الگوریتم‌های فراابتکاری دارد [۲۸].

در[۲۹] یک الگوریتم مسیریابی مبتنی بر یادگیری کیو دوگانه چندمرحله‌ای ارائه شده است که با استفاده از پاداش پویا بر اساس تعداد پرش و مرکزیت گره، انتخاب گره بعدی را بهینه می‌کند و موجب افزایش نرخ تحویل و کاهش تأخیر و سربار می‌شود. همچنین در [۳۰] یک عامل هوشمند مبتنی بر یادگیری کیو و فرآیند تصمیم مارکوف معرفی شده است که با یادگیری تطبیقی و تصمیم‌گیری خودکار، عملکرد مسیریابی را بهبود داده و نرخ تحویل را افزایش می‌دهد. در [۳۱] یک چارچوب مسیریابی مبتنی بر یادگیری تقویتی عمیق دوگانه همراه با یادگیری فدرال برای شبکه‌های حسگر بی‌سیم ارائه شده است که با هدف کاهش مصرف انرژی و بهبود کارایی مسیرها طراحی شده و در معیارهای تأخیر و مقیاس‌پذیری عملکرد بهتری نشان می‌دهد. همچنین در [۳۲] یک پروتکل مسیریابی مبتنی بر یادگیری کیو چندهدفه برای شبکه‌های پهپادی ارائه شده است که با تنظیم تطبیقی پارامترها و بهبود سازوکار اکتشاف و بهره‌برداری، موجب افزایش نرخ تحویل و کاهش تأخیر و مصرف انرژی می‌شود.

غنی‌فر و همکاران یک چارچوب مبتنی بر شبکه عصبی از فیزیک تناسبی- مشتقی برای بازتولید بلادرنگ مسیر پهپادهای چندروتوره ارائه کرده‌اند. در این روش مسیر به‌صورت تطبیقی و با استفاده از بازخورد حسگرها و مدل فیزیکی سیستم به‌روزرسانی می‌شود. تابع هزینه شبکه شامل ترم‌های تناسبی و مشتقی بوده و معادلات دینامیکی سیستم در فرآیند یادگیری لحاظ شده است که موجب سازگاری فیزیکی و حفظ امکان‌پذیری مسیر می‌شود. نتایج شبیه‌سازی نشان می‌دهد این روش خطای ردیابی را به‌طور قابل توجهی کاهش داده و عملکرد بهتری نسبت به روش‌های کلاسیک در شرایط نویز و اغتشاش دارد [۳۳]. همچنین غنی‌فر و همکاران در پژوهشی دیگر، یک روش جدید برای برنامه‌ریزی مسیر پهپادهای چندروتوره با عنوان میدان پتانسیل تطبیقی مبتنی بر منطق فازی الهام‌گرفته از کنترل PID (FPID-APFM)[1] ارائه کرده‌اند. در این روش محدودیت‌های روش‌های کلاسیک میدان پتانسیل مانند گیر افتادن در نقاط بهینه محلی و نوسان نزدیک موانع با استفاده از یک سیستم استنتاج فازی برطرف می‌شود. این سیستم به‌صورت بلادرنگ ضرایب نیروهای جاذبه و دافعه را تنظیم می‌کند و با الهام از ساختارPID ، پایداری و نرمی مسیر را بهبود می‌دهد. نتایج شبیه‌سازی در محیط‌های ایستا و پویا نشان می‌دهد که روش پیشنهادی از نظر ایمنی، سازگاری و کیفیت مسیر عملکرد بهتری نسبت به روش‌های مرسوم دارد [۳۴]. گاندی و نارماوالا به بررسی پروتکل‌های مسیریابی مبتنی بر یادگیری تقویتی در شبکه‌های فرصت‌طلب پرداخته و دو روش RLON[2] و DQON[3] را بر پایه یادگیری کیو و یادگیری کیو دوگانه ارائه کرده‌اند که با هدف انتخاب بهینه گره بعدی و افزایش سازگاری با تغییرات شبکه طراحی شده‌اند. این روش‌ها در بهبود نرخ تحویل، کاهش تأخیر و کاهش سربار نسبت به روش‌های مرجع عملکرد بهتری نشان می‌دهند[۳۵]. سینگ و همکاران یک پروتکل مسیریابی مبتنی بر یادگیری کیو دوگانه برای شبکه‌های فرصت‌طلب با عنوان ORAL[4] ارائه کرده‌اند که با استفاده از تخمین‌گر وزن‌دار کیو و مکانیزم پاداش احتمالاتی، مشکل بیش‌برآوردی و کم‌برآوردی را کاهش می‌دهد [۳۶]. ژو و همکاران یک راهبرد مسیریابی تطبیقی مبتنی بر یادگیری کیو دوگانه بهبودیافته برای اینترنت اشیای ماهواره‌ای پیشنهاد کرده‌اند که با مدل‌سازی شبکه به‌عنوان محیط یادگیری تقویتی و بهینه‌سازی تابع پاداش و پارامترها، موجب بهبود عملکرد در شبکه‌های پویا شده است [۳۷]. پاتیل و همکاران نیز رویکردی مبتنی بر یادگیری تقویتی برای بهینه‌سازی مسیریابی ترافیک شهری ارائه کرده‌اند که از یادگیری کیو دوگانه بهبودیافته و بازپخش تجربه اولویت‌دار استفاده می‌کند. نتایج نشان می‌دهد این روش نسبت به الگوریتم‌های پایه همگرایی سریع‌تر، کارایی بهتر و عملکرد مناسب‌تری در شبکه‌های بزرگ و پیچیده دارد [۳۸].

شفیعی‌نژاد در پژوهش‌های خود به توسعه سامانه‌های هوشمند هدایت و کنترل برای ربات‌های هوایی پرداخته است. وی در یک پژوهش، با الهام از رفتار اجتماعی ملخ‌ها و بهره‌گیری از منطق فازی، روشی برای هدایت گروهی ربات‌های کروی هوایی و

1. Fuzzy PID-inspired Adaptive Potential Field Method
2. Reinforcement Learning based Opportunistic Networks
3. Double Q-Learning based Opportunistic Networks
4. Off-Policy Reinforcement-based Adaptive Learning

بهینه‌سازی برخط مسیر ارائه کرده که موجب بهبود هماهنگی و دقت حرکت جمعی شده است [۳۹]. همچنین در پژوهشی دیگر، یک ربات هوایی تک‌بال الهام‌گرفته از دانه افرا را همراه با سامانه هدایت بهینه مقاوم و کنترل پیش‌بین طراحی کرده که نتایج آن نشان‌دهنده افزایش پایداری، دقت و مقاومت سامانه در برابر اغتشاشات محیطی است [۴۰].

اگرچه مطالعات متعددی به بررسی جداگانه این الگوریتم‌ها پرداخته‌اند، اما تحقیقات محدودی به مقایسه نظام‌مند عملکرد آن‌ها در محیط‌های با پیچیدگی‌های مختلف، همراه با تحلیل آماری جامع و ارزیابی چندمعیاره پرداخته‌اند. به‌ویژه، مقایسه‌ای که هم‌زمان معیارهای کمی (طول مسیر، زمان اجرا) و کیفی (نرم‌بودن مسیر، پایداری یادگیری) را در محیط‌های ایستا و پویا بررسی کند، در ادبیات موجود کمرنگ است. هدف اصلی این پژوهش، ارزیابی مقایسه‌ای جامع سه الگوریتم آاستار، یادگیری کیو و یادگیری دوگانه کیو در مسئله مسیریابی است. اهداف اختصاصی این مطالعه عبارتند از: پیاده‌سازی و آزمایش الگوریتم‌ها در محیط‌های شبکه‌ای با موانع، ارزیابی عملکرد بر اساس معیارهای چندبعدی شامل بهینگی، کارایی و کیفیت مسیر، تحلیل آماری با اجراهای مکرر و آزمون‌های فرضیه و بررسی تأثیر افزایش پیچیدگی محیط بر عملکرد الگوریتم‌ها. این پژوهش با هدف بررسی و تحلیل سه مسئله کلیدی در حوزه مسیریابی محیط‌های ایستا طراحی شده است. نخست، مزایا و محدودیت‌های الگوریتم آاستار در مقایسه با روش‌های یادگیری تقویتی مورد ارزیابی قرار می‌گیرد. سپس، میزان اثربخشی یادگیری دوگانه کیو در بهبود پایداری یادگیری و کاهش مشکل بیش‌برآورد نسبت به نسخه استاندارد یادگیری کیو بررسی می‌شود. در نهایت، تأثیر افزایش پیچیدگی محیط بر عملکرد نسبی این الگوریتم‌ها تحلیل خواهد شد.

اهمیت این پژوهش از دو جنبه نظری و عملی قابل تبیین است. از منظر نظری، این مطالعه به درک بهتر مزایا و محدودیت‌های رویکردهای قطعی در مقابل روش‌های یادگیری‌محور کمک می‌کند. از جنبه عملی، یافته‌های این تحقیق می‌تواند مبنایی برای انتخاب الگوریتم مناسب با توجه به الزامات خاص هر کاربرد (از جمله اولویت سرعت، بهینگی یا قابلیت انطباق) فراهم آورد. همچنین، ارائه تحلیل آماری قوی بر اساس ۳۰ اجرای مکرر، قابلیت اطمینان نتایج را افزایش می‌دهد. ساختار این مقاله به‌صورت زیر سازماندهی شده است: در بخش دوم، روش تحقیق و جزئیات پیاده‌سازی الگوریتم‌ها ارائه می‌شود. بخش سوم به تجزیه و تحلیل نتایج حاصل می‌پردازد. در بخش چهارم، درباره یافته‌ها در مقایسه با مطالعات پیشین بحث می‌شود و در نهایت، در بخش پنجم نتیجه‌گیری مطرح می‌گردد.

۱- ۱ توضیح الگوریتم‌ها

۱- ۱- ۱ الگوریتم آاستار[1]

الگوریتم آاستار یک الگوریتم جستجوی گرافی است که برای یافتن کوتاه‌ترین مسیر از یک گره شروع به یک گره هدف به‌کار می‌رود. این الگوریتم که در سال ۱۹۶۸ توسعه داده شد، ترکیبی از دو الگوریتم دیگر است: جستجوی حریصانه بر اساس بهترین–اولین که برای رسیدن سریع به هدف طراحی شده اما بهینه بودن مسیر را تضمین نمی‌کند و الگوریتم دایجسترا که کوتاه‌ترین مسیر را پیدا می‌کند اما ممکن است این کار به کندی انجام شود. الگوریتم آاستار با ترکیب هوشمندانه این دو روش، هم بهینه بودن مسیر را تضمین می‌کند و هم با استفاده از تابع هیورستیک مناسب، عملکرد بسیار کارآمدی دارد. بخش اصلی الگوریتم آاستار تابع ارزیابی f(n) است که برای هر گره n محاسبه می‌شود. این تابع، هزینه تخمین‌زده‌شده برای ارزان‌ترین مسیر از شروع به هدف که از گره n می‌گذرد را نشان می‌دهد. تابع f(n) از مجموع دو بخش تشکیل شده است:

$$f(n) = g(n) + h(n) \qquad (۱)$$

در الگوریتم آاستا، g(n) نشان‌دهنده هزینه واقعی مسیر از گره شروع تا گره فعلی n است و این مقدار دقیق و قطعی می‌باشد. از طرف دیگر، h(n) یک تابع هیورستیک است که به‌صورت تخمینی هزینه باقیمانده از گره n تا گره هدف را برآورد می‌کند و هوشمندی و کارایی الگوریتم آاستار به کیفیت این تابع هیورستیک بستگی دارد. برای اینکه الگوریتم آاستار بتواند کوتاه‌ترین مسیر را تضمین کند، تابع هیورستیک h(n) باید دو شرط مهم داشته باشد. شرط اول پذیرندگی است: تابع هیورستیک هرگز نباید هزینه واقعی مسیر باقیمانده را بیش‌برآورد کند. به‌عبارت دیگر، برای هر گره n باید داشته باشد:

$$h(n) \leq h*(n) \qquad (۲)$$

که در آن h*(n) هزینه واقعی کوتاه‌ترین مسیر از n به هدف است. شرط دوم سازگاری است که به‌ویژه در گراف‌هایی با بررسی پویا اهمیت دارد. یک تابع هیورستیک سازگار است اگر برای هر گره n و هر جانشین آن m که از طریق عمل a به‌دست می‌آید، داشته باشد:

$$h(n) \leq c(n,a,m) + h(m) \qquad (۳)$$

که در آن c(n, a, m) هزینه حرکت از n به m است. این نابرابری، معادل شرط مثلث در هندسه است و تضمین می‌کند که f(n) در طول کاهش پیدا نمی‌کند. یک هیورستیک سازگار، لزوماً پذیرفتنی نیز هست [۴۱].

1. A Star

۱- ۱- ۲ یادگیری کیو[1]

یادگیری تقویتی[2](RL) یک چارچوب ریاضیاتی است که برای توسعه عامل‌های کامپیوتری طراحی شده و امکان یادگیری رفتار بهینه را با ارتباط دادن پاداش‌های گذشته و جاری با اعمال انجام‌شده فراهم می‌کند. این چارچوب در حوزه‌هایی مانند هوش تجاری، مالی و بازی‌ها کاربردهای موفق فراوانی داشته و برای تصمیم‌گیری‌های پی‌درپی در محیط‌های ناشناخته و با داده‌های حجیم بسیار مناسب است. در این چارچوب، یادگیری کیو به‌عنوان یک الگوریتم یادگیری تقویتی بدون مدل مطرح است که با استفاده از تابع کیو، پاداش‌های احتمالی آینده اعمال در حالت‌های مشخص را تخمین می‌زند و از طریق به‌روزرسانی‌های تکراری، یادگیری سیاست بهینه را ممکن می‌سازد [۵].

الگوریتم یادگیری کیو بر اساس معادله بلمن[3] کار می‌کند:

$$Q(s,a) = E[r + \gamma * max\, Q(s',a')] \qquad (۴)$$

در رابطه (۴)، s نشان‌دهنده حالت فعلی است و a عمل انتخاب‌شده در این حالت می‌باشد. پس از اجرای عمل a، پاداش دریافتی با نماد r مشخص می‌شود و حالت بعدی که به آن منتقل می‌شود، با 's نمایش داده می‌شود. ضریب تنزیلγ تعادل بین پاداش‌های فوری و پاداش‌های آینده را تعیین می‌کند و max Q(s', a') بهترین ارزش قابل دستیابی از حالت بعدی 's را نشان می‌دهد که بیشترین پاداش بلندمدت ممکن را منعکس می‌کند. اجزای اصلی یادگیری کیو شامل چند بخش مهم هستند. اولین بخش، مقادیر کیو یا مقادیر عمل است که نشان‌دهنده پاداش‌های مورد انتظار برای انجام یک عمل در یک حالت مشخص می‌باشند. این مقادیر به مرور زمان و بر اساس قانون به‌روزرسانی اختلاف زمانی[4] اصلاح می‌شوند. دوم، پاداش‌ها و اپیزودها هستند. در این بخش عامل با انجام اعمال مختلف از حالتی به‌حالت دیگر حرکت کرده و در این مسیر پاداش دریافت می‌کند. سوم، اختلاف زمانی است که عامل برای به‌روزرسانی مقادیر کیو از فرمول اختصاصی به‌روزرسانی اختلاف زمانی استفاده می‌کند، فرمولی که ترکیبی از پاداش فوری و تخمین پاداش‌های آینده را در نظر می‌گیرد.

$$Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A)) \qquad (۵)$$

در این تعریف‌ها، S نشان‌دهنده وضعیت فعلی عامل است و A عملی است که عامل در حالت S انجام می‌دهد. پس از انجام عمل، عامل به حالت بعدی S′ منتقل می‌شود و A′ بهترین عملی است که در حالت S′ می‌تواند انتخاب شود R .پاداشی است که عامل در نتیجه انجام عمل A در حالت S دریافت می‌کند و α (آلفا) نرخ یادگیری است که میزان تأثیر اطلاعات جدید بر مقادیر کیو قدیمی را تعیین می‌کند. در الگوریتم‌های یادگیری تقویتی، سیاست (ε) برای ایجاد تعادل بین دانش فعلی و جست‌وجوی گزینه‌های جدید استفاده می‌شود. در این سیاست، عامل میان بهره‌برداری از دانش فعلی و اکتشاف گزینه‌های جدید تعادل برقرار می‌کند. عامل با احتمال 1-ε عملی را برمی‌گزیند که بیشترین مقدار کیو را دارد و به این ترتیب پاداش مورد انتظار را حداکثر می‌کند. با این‌حال، با احتمال ε اقدام به انتخاب تصادفی می‌کند تا مسیرهای ناشناخته را بررسی کرده و امکان دستیابی به راهبردهای بهتر را افزایش دهد.

۱- ۱- ۳ یادگیری کیو دوگانه[5]

یادگیری کیو دوگانه یک روش بهبود یافته از الگوریتم استاندارد یادگیری کیو است که توسط هسلت در سال ۲۰۱۰ معرفی شد و برای حل مشکل بیش‌برآوردی[6] در مقادیر کیو طراحی شده است. در یادگیری کیو استاندارد، از همان مقادیر کیو هم برای انتخاب عمل بهینه و هم برای ارزیابی ارزش آن عمل استفاده می‌شود که این مسئله می‌تواند منجر به بیش‌برآوردی سیستماتیک در مقادیر کیو گردد و در نتیجه یادگیری سیاست‌های زیربهینه و کاهش پایداری الگوریتم را به دنبال داشته باشد. در فرمول به‌روزرسانی یادگیری کیو یعنی معادله (۵) مقدار $max\ Q(s',a')$ هم عمل بهینه را انتخاب و هم ارزش آن را ارزیابی می‌کند. از آنجاکه مقادیر کیو برآوردی هستند و شامل نویز می‌باشند، ماکزیمم گرفتن از این مقادیر منجر به بیش‌برآوردی می‌شود، این بیش‌برآوردی به‌خصوص در محیط‌های پیچیده و با نویز می‌تواند مشکل‌ساز شود. ایده اصلی یادگیری کیو دوگانه استفاده از دو تابع ارزش مستقل (QA و QB) است. این دو تابع به‌صورت متناوب برای انتخاب عمل بهینه و ارزیابی ارزش آن استفاده می‌شوند. فرمول‌های به روزرسانی یادگیری کیو دوگانه به این‌صورت می‌باشند:

$$E[max\ Q(s',a')] \geq max\ E[Q(s',a')] \qquad (۶)$$

$$a* = argmax\ QA(s',a') \qquad (۷)$$

$$QA(s,a) \leftarrow QA(s,a) + \alpha * [r + \gamma * QB(s',a*) - QA(s,a)] \qquad (۸)$$

$$a* = argmax\ QB(s',a') \qquad (۹)$$

$$QB(s,a) \leftarrow QB(s,a) + \alpha * [r + \gamma * QA(s',a*) - QB(s,a)] \qquad (۱۰)$$

1. Q- Learning
2. Reinforcement learning
3. Bellman Equation
4. Temporal Difference
5. Double Q-Learning
6. Overestimation

معادله (۸ و۱۰) به‌روزرسانی در الگوریتم یادگیری کیو دوگانه، مقدار کیو برای عمل a در حالت s را اصلاح می‌کند. پاداش دریافتی r و ارزش بهترین عمل در حالت بعدی QB با ضریب تنزیل γ ترکیب شده و تحت تأثیر نرخ یادگیری α به‌مقدار کیو فعلی افزوده می‌شوند. استفاده از دو جدول کیو به‌صورت متقابل، بیش‌برآوردی را کاهش داده و یادگیری پایدار را تضمین می‌کند.

به‌منظور درک بهتر فرآیند پیاده‌سازی و ارزیابی الگوریتم‌های مسیریابی، روند اجرای کلی این الگوریتم‌ها در شکل (۱) ارائه شده است. این نمودار مراحل اصلی از آماده‌سازی محیط تا ذخیره نتایج را به‌صورت منظم نشان می‌دهد. همچنین شکل (۲)، روندنمای مقایسه عملکرد سه الگوریتم آاستار، یادگیری کیو و یادگیری دوگانه کیو را نشان می‌دهد.



شکل ۱– روند نمای اجرای الگوریتم‌های مسیریابی.

Fig. 1. Flowchart of routing algorithm execution.



شکل ۲– روند نمای مقایسه عملکرد الگوریتم‌های مسیریابی آاستار، یادگیری کیو و یادگیری دوگانه کیو.

Fig. 2. Flowchart for Performance Comparison of A*, Q-Learning, and Double Q-Learning Routing Algorithms.

این پژوهش با ارائه یک شبه‌کد ساختاریافته، چارچوب نظام‌مندی برای مقایسه سه الگوریتم مسیریابی ارائه می‌دهد. شبه‌کد پیشنهادی با تعریف ورودی‌های استاندارد شامل ابعاد شبکه، موانع، نقطه شروع

و هدف، یکپارچگی آزمایش‌ها را تضمین می‌کند. در مرحله اول، محیط آزمایشی با ساخت شبکه و تعریف موانع آماده‌سازی می‌شود که پایه‌ای مشترک برای اجرای تمام الگوریتم‌ها فراهم می‌آورد. اجرای الگوریتم آاستار مبتنی بر جستجوی هیوریستیک بوده و مسیر بهینه را بازمی‌گرداند. در مقابل، الگوریتم‌های یادگیری تقویتی شامل یادگیری کیو و یادگیری دوگانه کیو، هر یک با مکانیزم‌های آموزش خاص خود (به‌ترتیب با استفاده از یک جدول و دو جدول ارزش) پیاده‌سازی شده‌اند. این ساختار سلسله مراتبی امکان مقایسه‌ای بین رویکردهای قطعی و یادگیری‌محور را فراهم می‌سازد.

Pseude code

Algorithm: Comparative Path Planning
Input: grid_size, obstacles, start, goal, num_episodes=1000, num_runs=30
Output: paths, performance_metrics

1. Environment Setup:
  Create grid with obstacles, actions=Up,Down,Left,Right
Reward: +1000 (goal), -50 (obstacle), -1 (step with revisit penalty)

2. A* Algorithm:
  Initialize openList with start (g=0, h=Manhattan), closedList=[]
  While openList not empty:
    current = node with min f in openList
    If current == goal: reconstruct path via parent pointers
    For each neighbor: if valid and not in closedList:
      g_new = current.g+1, h_new = Manhattan(neighbor,goal)
      Update or add neighbor to openList

3. Q-Learning Algorithm:
  Initialize Q = rand×0.01, set α=0.3 (decays), γ=0.95, ε:1.0→0.01
  For episode=1:num_episodes:
    state = start, visited = zeros
    For step=1:max_steps:
      visited[x,y]++
      action = random if rand<ε else argmax_a Q(state,a)
      next_state, reward = transition(state,action)
      Q(state,action) = Q(state,action) + α×[reward + γ×max_a Q(next_state,a) - Q(state,action)]
      state = next_state, if state==goal break
    Decay ε, and α every 100 episodes
  Extract path: start→goal by argmax_a Q(state,a)

4. Double Q-Learning Algorithm:
  Initialize Q1=rand×0.01, Q2=rand×0.01, set α=0.3 (decays), γ=0.95, ε:1.0→0.01
  For episode=1:num_episodes:
    state = start, visited = zeros
    For step=1:max_steps:
      visited[x,y]++
      Q_avg = (Q1+Q2)/2
      action = random if rand<ε else argmax_a Q_avg(state,a)
      next_state, reward = transition(state,action)
      If rand<0.5:
        best = argmax_a Q1(next_state,a)
        Q1 = Q1 + α×[reward + γ×Q2(next_state,best) - Q1]
      Else:
        best = argmax_a Q2(next_state,a)
        Q2 = Q2 + α×[reward + γ×Q1(next_state,best) - Q2]
      state = next_state, if state==goal break
    Decay ε, and α every 100 episodes
  Extract path: Q_avg=(Q1+Q2)/2, start→goal by argmax_a Q_avg(state,a)

5. Statistical Analysis (30 runs):
  For each run: store path_length, time, success, rewards, convergence
  Compute: mean, std, median, min, max, 95% CI
  Tests: t-test, ANOVA, Cohen's d
  Quality: turns (smoothness), optimality = (optimal_length/found_length)×100

6. Visualization:
  Plot paths (A*:green, Q:blue, DQ:red), learning curves, comparative tables
  Return: paths, metrics, results

شبه‌کد ارائه شده در این بخش، رویکردی جامع برای ارزیابی کمی الگوریتم‌ها اتخاذ کرده است. محاسبه معیارهای چندبعدی شامل طول مسیر، زمان اجرا و کیفیت مسیر، امکان ارزیابی همه‌جانبه عملکرد را میسر می‌سازد. طراحی آزمایش با ۳۰ اجرای مکرر هر الگوریتم، پایایی نتایج را از طریق تحلیل‌های آماری پیشرفته شامل محاسبه شاخص‌های مرکزی، پراکندگی، فواصل اطمینان و آزمون‌های فرضیه تضمین می‌کند. بخش نمایش نتایج نیز با ترسیم مسیرها بر روی شبکه، تولید نمودارهای یادگیری و ایجاد جداول مقایسه‌ای، هم ارزیابی بصری و هم تحلیل عددی را پوشش می‌دهد. این ساختار دووجهی – تحلیل کمی و کیفی– چارچوبی علمی برای مقایسه الگوریتم‌های مسیریابی در محیط‌های پیچیده ارائه می‌دهد که قابلیت تعمیم به کاربردهای واقعی را دارد.

۲. تحلیل و ارزیابی عملکرد الگوریتم‌ها

تحلیل‌های آماری صورت‌گرفته در سه جدول مجزا، ابعاد مختلف عملکرد این الگوریتم‌ها را به تفصیل بررسی می‌کند.

جدول ۱- مقایسه عملکرد کلی الگوریتم‌ها در محیط شبکه‌ای.

Table 1. Comparison of overall algorithm performance in the network environment.

Performance metric	A*	Q-Learning	Double Q-learning
Path length (steps)	42	42	46
Execution time (seconds)	0.0169	0.3595	0.3915
Success rate (%)	100.0	100.0	91.3
Number of turns	7	19	20
Smoothness index	0.125	0.050	0.048
Turn-to-path length ratio	0.167	0.452	0.435

همان‌طور که در جدول (۱) مشاهده می‌شود، الگوریتم آاستار در تمام معیارهای کمی عملکرد برتری را نسبت به دو الگوریتم یادگیری تقویتی نشان می‌دهد. این الگوریتم توانست با طول مسیر دقیقاً ۴۲ گام (بدون هیچ پراکندگی) و در زمان اجرای بسیار کوتاه ۰/۰۰۴۳ ثانیه، مسیری بهینه و با کیفیت بالا (تنها ۷ پیچش) ارائه دهد. در مقابل، الگوریتم‌های یادگیری کیو و یادگیری دوگانه کیو اگرچه به نرخ موفقیت ۱۰۰ درصدی دست یافتند، اما زمان اجرای آن‌ها به‌ترتیب ۶۴ و ۷۴ برابر آاستار بود. همچنین، مسیرهای تولید شده توسط این الگوریتم‌ها کیفیت پایین‌تری داشتند که با شاخص نرم‌بودن ۰/۰۵ و ۰/۰۴۸ (در مقابل ۰/۱۲۵آاستار) و نسبت پیچش به طول ۰/۴۵۲ و ۰/۴۳۵ (در مقابل ۰/۱۶۷ آاستار) قابل مشاهده است.

جدول ۲- مقایسه ویژگی‌های یادگیری.

Table 2. Comparison of learning characteristics.

Learning characteristics	Q-learning	Double Q-learning
Average final Rewarr	957.38	956.52
Standard deviation of reward	0.86	1.41
Convergence episodes	78	74
Variance reduction	-	-767.2%
Effect size (Cohen's d)	-	0.455

نتایج ارائه‌شده در جدول (۲) نشان می‌دهد که بین دو الگوریتم یادگیری تقویتی تفاوت‌های معنی‌داری در معیارهای یادگیری وجود دارد. یادگیری کیو با میانگین پاداش نهایی ۹۵۷/۳۸ و انحراف معیار ۰/۸۶، عملکرد پایدارتری نسبت به یادگیری دوگانه کیو (میانگین پاداش ۹۵۶/۵۲ با انحراف معیار ۱/۴۱) داشت. نکته جالب توجه، سرعت همگرایی یادگیری کیو دوگانه است که با ۷۴ اپیزود به‌طور معنی‌داری سریع‌تر از یادگیری کیو (۷۸ اپیزود) همگرا شد. با این وجود، اندازه اثر کوهن (Cohen's d) معادل ۰/۴۵۵ نشان می‌دهد که تفاوت طول مسیر بین این دو الگوریتم در حد متوسط است. کاهش واریانس۷۶۷/۲- درصدی در یادگیری دوگانه کیو نشان‌دهنده افزایش ناپایداری در این الگوریتم است.

جدول ۳- مقایسه عملکرد در محیط‌های مختلف.

Table 3. Performance comparison across different environments.

Environment	Algorithm	Path length (steps)	Execution time (s)	Optimality (%)
Network environment	A*	42	0.0169	100%
	Q-learning	42	0.3595	100%
	Double Q-learning	46	0.3915	91.3%
Urban environment	A*	72	0.0244	100%
	Q-learning	74	0.6607	97.3%
	Double Q-learning	80	0.7652	90%

جدول (۳) عملکرد الگوریتم‌ها را در دو محیط متفاوت مقایسه کرده است. در محیط شبکه‌ای ساده، آاستار و یادگیری کیو هر دو به بهینگی ۱۰۰ درصدی دست یافتند، اما یادگیری دوگانه کیو با ۹۱/۳ درصد بهینگی عملکرد ضعیف‌تری داشت. با افزایش پیچیدگی محیط به‌حالت شهری (همان‌طور که در جدول ۳ نشان داده شده است)، اختلاف عملکرد الگوریتم‌ها تشدید شد. طول مسیر آاستار از ۴۲ به ۷۲ گام افزایش یافت (۷۱ درصد رشد)، در حالی‌که زمان اجرای آن تنها ۴۴ درصد افزایش پیدا کرد. الگوریتم‌های یادگیری تقویتی افت بهینگی قابل توجهی در محیط شهری تجربه کردند: یادگیری کیو از ۱۰۰ درصد به ۹۷/۳ درصد و یادگیری دوگانه کیو از ۹۱/۳ درصد به ۹۰ درصد کاهش یافت. این نتایج نشان می‌دهد که با افزایش پیچیدگی محیط، مزیت نسبی الگوریتم‌های یادگیری کاهش می‌یابد.

اختلاف بیشتر عملکرد الگوریتم‌ها در محیط شهری را می‌توان به افزایش پیچیدگی مسئله مسیریابی نسبت داد. در این محیط، با بزرگ‌تر شدن فضای حالت، افزایش تعداد موانع و ایجاد مسیرهای باریک و متعدد، فرآیند تصمیم‌گیری دشوارتر می‌شود و الگوریتم‌های مبتنی بر یادگیری ناچارند تعداد بسیار بیشتری از حالت‌ها و مسیرهای ممکن را ارزیابی کنند. علاوه بر این، طولانی‌تر شدن مسیرها موجب افزایش تعداد تصمیم‌گیری‌های متوالی شده و احتمال انحراف از مسیر بهینه را افزایش می‌دهد. در مقابل، الگوریتم آاستار با بهره‌گیری از تابع هیوریستیک، از دانش ساختاری محیط برای هدایت جستجو استفاده می‌کند و بنابراین با افزایش پیچیدگی محیط، افت عملکرد کمتری را تجربه می‌کند. به‌همین دلیل، تفاوت عملکرد میان الگوریتم‌های یادگیری و آاستار در محیط شهری بیش از محیط شبکه‌ای ساده نمایان شده است.

جدول ۴- پارامترهای کلیدی شبیه‌سازی الگوریتم‌های مسیریابی.

Table 4. Key parameters for simulating routing algorithms.

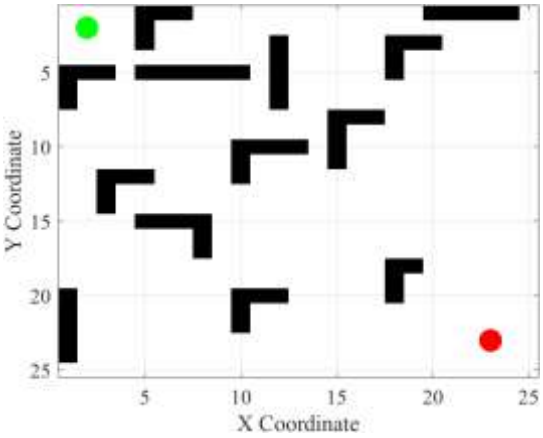
Parameter	A*	Q-learning	Double Q-learning
Environment	25×25 (Grid) / 40×40 (Urban)	25×25 (Grid) / 40×40 (Urban)	25×25 (Grid) / 40×40 (Urban)
Number of episodes	—	1000	1000
Number of independent runs	30	30	30
Learning rate (α)	—	0.3 (decaying)	0.3 (decaying)
Discount factor (γ)	—	0.95	0.95
Exploration policy (ε)	—	1.0 → 0.01	1.0 → 0.01

جدول (۴) پارامترهای اصلی مورد استفاده در پیاده‌سازی هر سه الگوریتم را به‌صورت مجزا برای دو محیط شبکه‌ای و شهری ارائه می‌دهد.

نمودارهای مقایسه هر سه الگوریتم آاستار، یادگیری کیو و یادگیری کیو دوگانه در شکل‌های (۳) تا (۳۱) ارائه شده است.

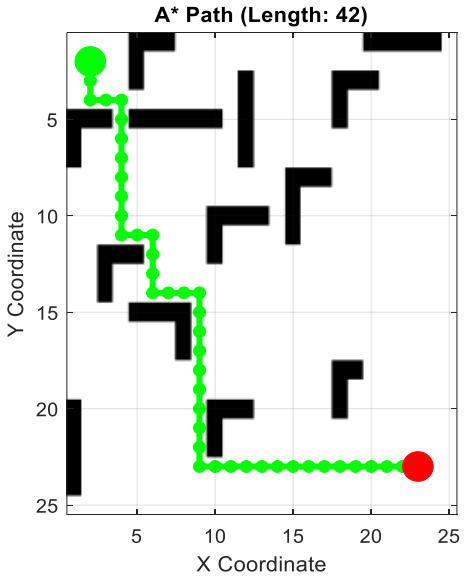
شکل (۳) محیط شبکه‌ای اصلی مورد استفاده در این پژوهش را نمایش می‌دهد. این محیط شامل یک شبکه‌ی دوبعدی با ابعاد 25×25 خانه است که تعدادی موانع، ثابت در نظر گرفته شده‌اند تا شرایط واقعی و چالش‌برانگیزتری برای مسیریابی ایجاد شود. وجود این موانع سبب می‌شود عامل ناچار به انتخاب مسیرهای غیرمستقیم و تصمیم‌گیری‌های پیچیده‌تری در مواجهه با موانع شود. در این محیط، نقطه سبز نشان‌دهنده موقعیت شروع و نقطه‌ی قرمز بیانگر مقصد نهایی است. هدف اصلی، یافتن مسیری ایمن، بهینه و بدون برخورد با موانع میان این دو نقطه می‌باشد. در ادامه، عملکرد سه الگوریتم آاستار، کیوو یادگیری کیو دوگانه در این محیط مورد مقایسه قرار گرفته است تا دقت، کارایی و پایداری آن‌ها در شرایط یکسان ارزیابی شود.

شکل (۴) مسیر به‌دست‌آمده توسط الگوریتم آاستار را در محیط شبکه‌ای نشان می‌دهد. همان‌گونه که مشاهده می‌شود، این الگوریتم با استفاده از ترکیب تابع هزینه‌ی واقعی و تابع برآوردی[1] توانسته است مسیری را میان نقطه شروع و نقطه‌ی هدف بیابد. در این شبیه‌سازی، عامل با طی ۴۲ حرکت توانسته از موقعیت اولیه (نقطه سبز) به مقصد نهایی (نقطه قرمز) برسد، بدون آنکه با هیچ‌یک از موانع برخوردی داشته باشد.



شکل ۳- نقشه شبکه‌ای با موانع پیچیده.

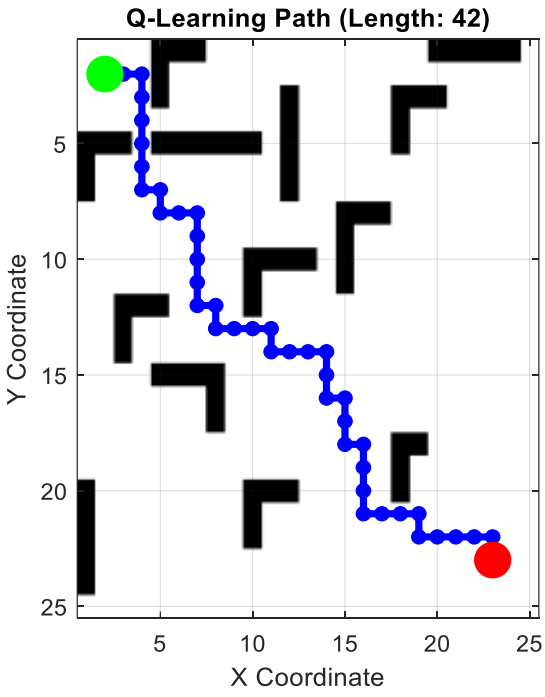
Fig. 3. Network map with complex obstacles.



شکل ۴- مسیر آاستار.

Fig. 4. A* path.

1. Heuristic Function



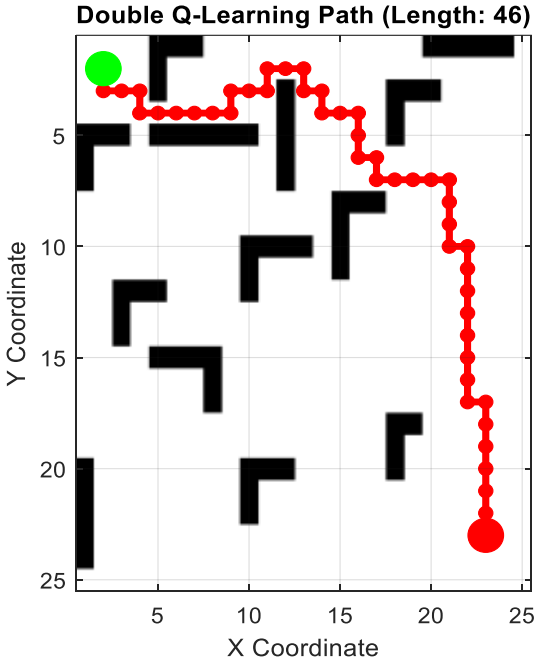
شکل ۵– مسیر یادگیری کیو.

Fig. 5. Q-Learning path

شکل (۵) مسیر حاصل از اجرای الگوریتم یادگیری کیو را در محیط شبکه‌ای نشان می‌دهد. در این شبیه‌سازی، عامل بدون داشتن آگاهی اولیه از محیط و تنها بر پایه یادگیری تجربی توانسته است مسیر مناسبی از نقطه شروع به نقطه‌ی هدف بیابد. همان‌طور که دیده می‌شود، عامل پس از طی ۴۲ گام موفق شده است به مقصد برسد. فرآیند یادگیری از طریق به‌روزرسانی مقادیر کیو در طول تعامل‌های متوالی با محیط انجام گرفته است، به‌طوری‌که عامل با گذشت زمان توانسته سیاست بهینه‌ای را برای حرکت در مسیر بدون برخورد با موانع بیاموزد. نتیجه نهایی نشان می‌دهد که الگوریتم یادگیری کیو قادر است بدون نیاز به مدل محیط، مسیر قابل‌قبولی مشابه الگوریتم‌های کلاسیک مبتنی بر جست‌وجو مانند A* پیدا کند.

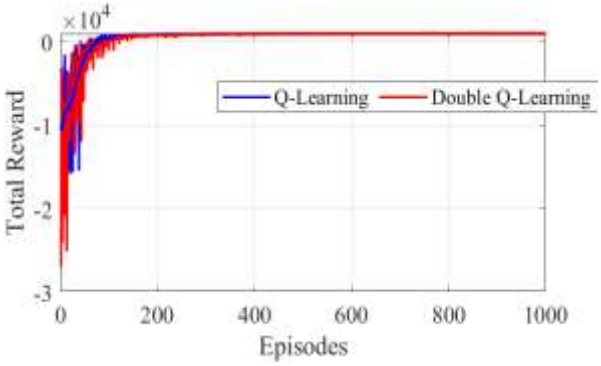
شکل (۶) مسیر به‌دست آمده توسط الگوریتم یادگیری کیو دوگانه را نشان می‌دهد. در این روش، فرآیند یادگیری با استفاده از دو مجموعه‌ی مجزای کیو جدول انجام می‌شود تا مشکل بیش‌برآوردی مقادیر کیو که در الگوریتم یادگیری کیو وجود دارد، کاهش یابد. عامل توانسته است پس از طی ۴۲ حرکت متوالی از نقطه شروع به نقطه هدف برسد و مسیری ایمن و بدون برخورد با موانع را بیابد.

شکل (۷) روند تغییر پاداش کل در طول فرایند یادگیری را برای دو الگوریتم یادگیری کیو و یادگیری کیو دوگانه نشان می‌دهد. همان‌گونه که مشاهده می‌شود، هر دو الگوریتم در مراحل ابتدایی نوسانات شدیدی دارند، اما با افزایش تعداد اپیزودها، مقادیر پاداش به‌تدریج افزایش یافته و به مقدار ثابت نزدیک می‌شوند که نشان‌دهنده همگرایی سیاست یادگیری است. الگوریتم یادگیری کیو دوگانه (منحنی قرمز) نسبت به یادگیری کیو (منحنی آبی) دارای نوسانات کمتر و همگرایی سریع‌تر است. این موضوع بیانگر پایداری بیشتر در فرآیند یادگیری و کاهش خطای بیش‌برآوردی مقادیر کیو در الگوریتم یادگیری کیو دوگانه است. منحنی‌های صاف‌شده نیز این تفاوت را به‌خوبی نشان می‌دهند، به‌طوری‌که یادگیری کیو دوگانه زودتر از یادگیری کیو به پاداش پایدار می‌رسد. در مجموع، نتایج این نمودار تأیید می‌کند که استفاده از رویکرد دوگانه در به‌روزرسانی مقادیر کیو، باعث افزایش بهبود کارایی یادگیری در مسیریابی می‌شود.



شکل ۶– مسیر یادگیری کیو دوگانه.

Fig. 6. Double Q-learning path.

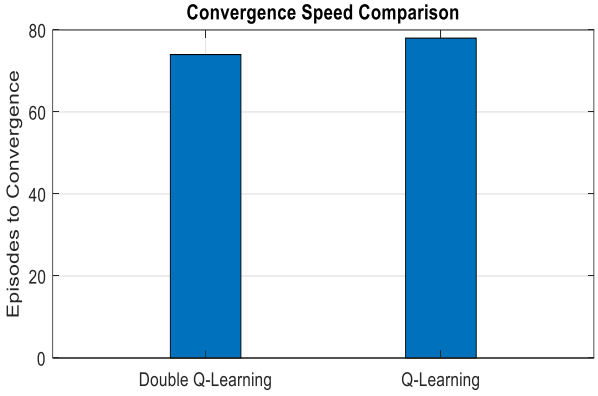


شکل ۷– مقایسه منحنی‌های یادگیری.

Fig. 7. Comparison of learning curves.

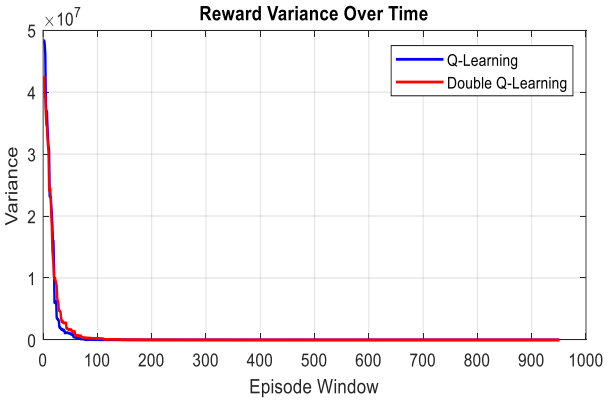
شکل (۸) روند همگرایی دو الگوریتم کیو و یادگیری کیو دوگانه را نشان می‌دهد. همان‌طور که مشاهده می‌شود، الگوریتم کیو برای رسیدن به نقطه همگرایی به تقریباً ۷۸ اپیزود نیاز دارد، در حالی‌که الگوریتم یادگیری کیو دوگانه سریع‌تر عمل کرده و پس از حدود ۷۴ اپیزود به همگرایی می‌رسد. این مقایسه نشان می‌دهد که یادگیری کیو دوگانه می‌تواند در محیط‌های مشابه با سرعت بیشتری همگرا شود و فرآیند تصمیم‌گیری سریع‌تری داشته باشد.

شکل (۹) واریانس پاداش در طول زمان دو الگوریتم کیو و یادگیری کیو دوگانه را نشان می‌دهد. محور افقی بازه‌های اپیزودهای آموزش و محور عمودی میزان واریانس را نمایش می‌دهد. هر دو الگوریک در مراحل اولیه (اپیزودهای ۰–۲۰۰) از واریانس بالایی برخوردارند که نشان‌دهنده عدم ثبات در یادگیری اولیه است. با پیشرفت فرآیند آموزش، واریانس در هر دو الگوریتم به‌طور کلی روند نزولی دارد.



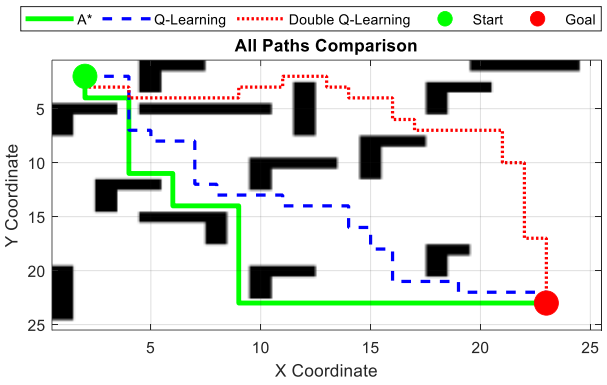
شکل ۸– مقایسه سرعت همگرایی.

Fig. 8. Comparison of convergence speed.



شکل ۹– واریانس پاداش در طول زمان.

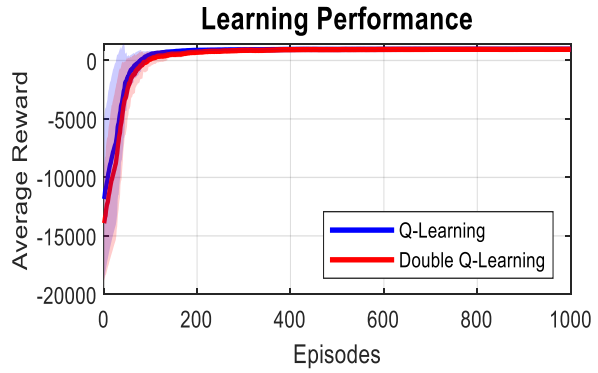
Fig. 9. Reward Variance over time.



شکل ۱۰– مقایسه همه مسیرها.

Fig. 10. Comparison of all paths.

شکل (۱۰) مسیرهای به‌دست آمده توسط سه الگوریتم آاستار، کیو و یادگیری کیو دوگانه را در محیطی شامل موانع مقایسه می‌کند. همان‌طور که دیده می‌شود، مسیر سبز رنگ مربوط به الگوریتم آاستار به‌عنوان مسیر مرجع و بهینه در نظر گرفته شده است. مسیرهای آبی و قرمز که به‌ترتیب متعلق به کیو و یادگیری کیو دوگانه هستند، نشان می‌دهند هر دو الگوریتم یادگیری تقویتی توانسته‌اند مسیرهایی نزدیک به مسیر بهینه تولید کنند. الگوریتم یادگیری کیو دوگانه در مقایسه با کیو مسیر کوتاه‌تر و هموارتری است و نوسانات کمتری در حرکت میان موانع دارد، در حالی که کیو مسیر طولانی‌تری با انحراف بیشتری نسبت به مسیر بهینه طی کرده است. به‌طور کلی، هر سه الگوریتم توانسته‌اند از نقاط شروع (دایره سبز) تا مقصد (دایره قرمز) برسند، اما الگوریتم آاستار مسیر بهینه‌تری را پیدا کرده و یادگیری کیو دوگانه از میان روش‌های یادگیری تقویتی عملکرد نزدیک‌تری به آن داشته است.

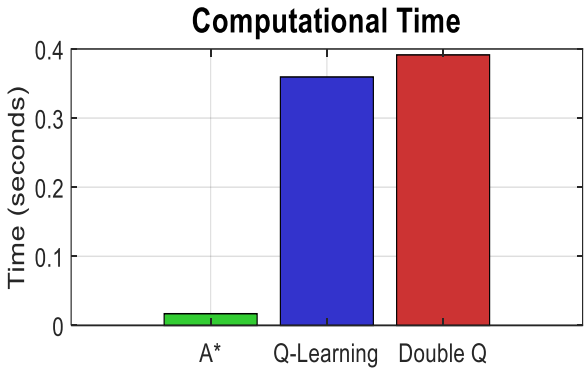


شکل ۱۱– عملکرد یادگیری.

Fig.11. Learning performance.

شکل (۱۱) عملکرد یادگیری دو الگوریتم یادگیری کیو (آبی) و یادگیری کیو دوگانه (قرمز) را بر حسب میانگین پاداش در طول اپیزودها نشان می‌دهد. همان‌طور که دیده می‌شود، هر دو الگوریتم در مراحل

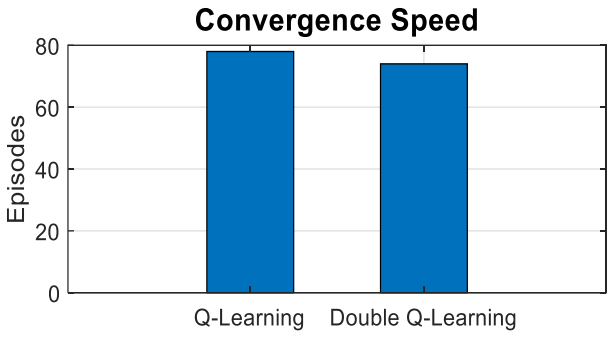
ابتدایی یادگیری (حدود ۰ تا ۱۰۰ اپیزود) با نوسانات قابل توجهی روبه‌رو هستند، اما به‌تدریج مقدار پاداش افزایش یافته و هر دو به ناحیه پایداری (نزدیک صفر) می‌رسند. الگوریتم یادگیری کیو دوگانه نسبت به یادگیری کیو روند همگرایی پایدارتری دارد و نوسانات کمتری را در طول فرآیند یادگیری تجربه می‌کند. این موضوع نشان می‌دهد که تفکیک به‌روزرسانی‌های کیو در دو مجموعه جداگانه (ویژگی کلیدی یادگیری کیو دوگانه) باعث کاهش بیش‌برآورد پاداش‌ها و در نتیجه بهبود فرایند یادگیری می‌شود. در نهایت، هر دو الگوریتم به عملکرد نهایی مشابهی دست می‌یابند، اما یادگیری کیو دوگانه با سرعت بیشتر و نوسان کمتری به همگرایی می‌رسد.



شکل ۱۲– زمان محاسباتی.

Fig. 12. Computational time.

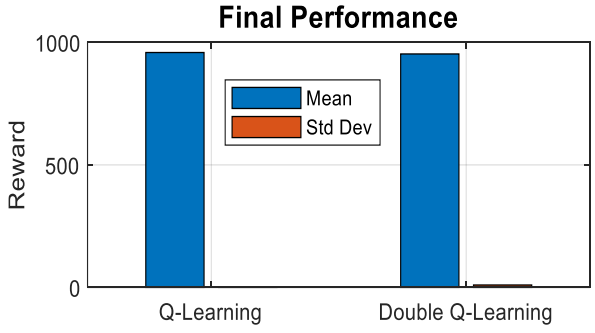
در شکل (۱۲)، زمان محاسباتی سه الگوریتم مورد مقایسه قرار گرفته است. مطابق نتایج، الگوریتم آاستار با میانگین زمان اجرای حدود ۰/۰۱۶۹ ثانیه، به‌طور چشمگیری سریع‌تر از دو الگوریتم یادگیری تقویتی عمل می‌کند. در مقابل، زمان محاسباتی الگوریتم‌های یادگیری کیو و یادگیری کیو دوگانه تقریباً یکسان و در حدود ۰/۳۵۹۵ و ۰/۳۹۱۵ ثانیه است. این تفاوت نشان دهنده پیچیدگی محاسباتی بالاتر این روش‌ها در مقایسه با الگوریتم جست‌وجوی کلاسیک آاستار می‌باشد. در الگوریتم‌های یادگیری کیو و یادگیری کیو دوگانه، عامل برای دستیابی به سیاست بهینه ناگزیر است طی اپیزودهای متوالی، محیط را کاوش کرده و مقادیر کیو را به‌صورت تدریجی به‌روزرسانی کند. این در حالی‌است که الگوریتم آاستار یک روش قطعی و غیرتکراری است که تنها در یک مرحله مسیر بهینه را محاسبه می‌کند .علاوه بر این، تفاوت اندک بین زمان محاسباتی یادگیری کیو و یادگیری کیو دوگانه نشان می‌دهد که افزودن ساختار دوم کیو در الگوریتم کیو دوگانه، اگرچه دقت برآورد مقادیر کیو را افزایش می‌دهد، اما تأثیر زیادی بر هزینه زمانی ندارد.



شکل ۱۳– سرعت همگرایی.

Fig. 13. Convergence speed.

در شکل (۱۳)، سرعت همگرایی دو الگوریتم یادگیری کیو و یادگیری کیو دوگانه بر اساس تعداد اپیزودهای لازم تا رسیدن به پایداری مقادیر کیو مورد مقایسه قرار گرفته است. همان‌گونه که مشاهده می‌شود، الگوریتم کیو دوگانه نسبت به کیو با تعداد اپیزودهای کمتری به همگرایی می‌رسد؛ به‌طور مشخص، یادگیری کیو دوگانه حدود ۵ درصد سریع‌تر از کیو همگرا شده است. این نتیجه بیانگر آن است که افزودن مجموعه‌ی دوم مقادیر کیو در الگوریتم یادگیری کیو دوگانه باعث کاهش سوگیری بیش‌براوردی در برآورد مقادیر کیو می‌شود؛ در نتیجه تخمین پاداش‌ها دقیق‌تر و روند یادگیری پایدارتر می‌گردد. در الگوریتم کیو، استفاده از یک تابع کیو واحد باعث می‌شود در هنگام انتخاب عمل‌هایی با پاداش بالا، مقادیر کیو بیش از حد واقعی برآورد شوند، که این امر منجر به نوسان و کندی در همگرایی می‌گردد. در مقابل، در یادگیری کیو دوگانه فرآیند انتخاب و به‌روزرسانی مقادیر کیو، این مشکل را کاهش داده و باعث افزایش پایداری و تسریع در همگرایی می‌شود. به‌طور کلی، نتایج این مقایسه نشان می‌دهد که هرچند هر دو الگوریتم قادر به یافتن سیاست بهینه هستند، اما یادگیری کیو دوگانه از نظر سرعت یادگیری و همگرایی عملکرد بهتری نسبت به الگوریتم کیو دارد.

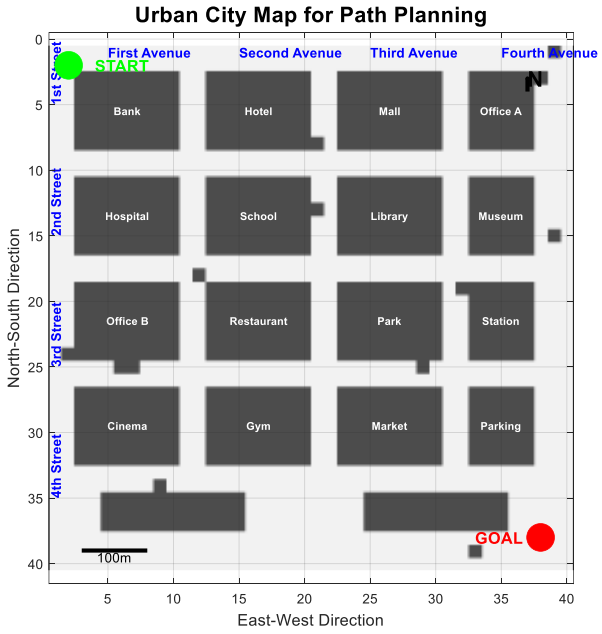


شکل ۱۴– عملکرد نهایی.

Fig. 14. Final performance.

در شکل (۱۴)، عملکرد نهایی دو الگوریتم کیو و یادگیری کیو دوگانه از نظر میانگین پاداش و انحراف معیار مقایسه شده است. نتایج نشان می‌دهد که هر دو الگوریتم به میانگین پاداش تقریباً یکسان و نزدیک به مقدار حداکثری (حدود ۱۰۰۰) دست یافته‌اند. همچنین، انحراف معیار پاداش‌ها برای هر دو روش برابر است که بیانگر پایداری مشابه در عملکرد نهایی آن‌ها است.

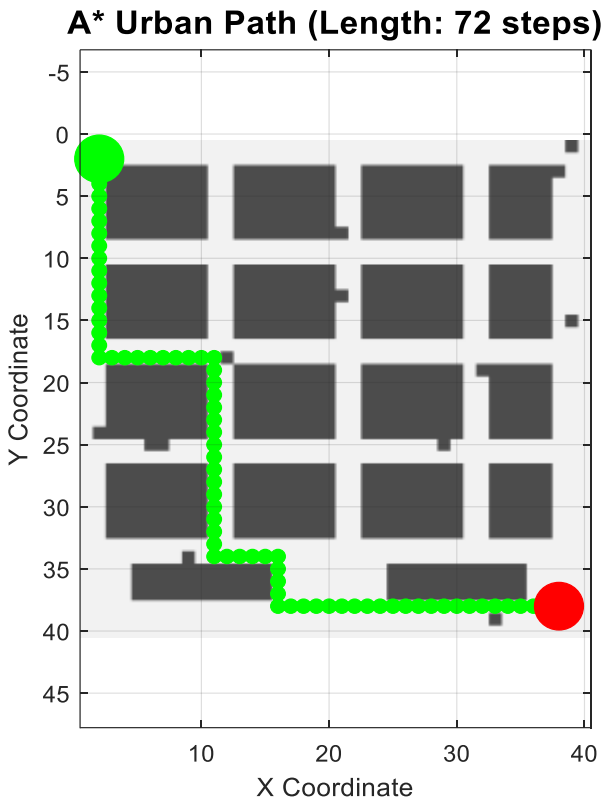
این نتایج نشان می‌دهد که در محیط آزمایشی حاضر، هر دو الگوریتم توانسته‌اند به سیاست بهینه و سطح عملکرد مشابهی دست یابند. برابری انحراف معیار نیز حاکی از آن است که یادگیری کیو دوگانه در این مسئله خاص، مزیت قابل توجهی از نظر کاهش نوسانات نسبت به کیو نداشته است. به‌عبارت دیگر، اگرچه یادگیری کیو دوگانه به طور نظری برای کاهش بیش‌براوردی طراحی شده است، اما در این محیط، تفاوت چشمگیری وجود ندارد و در نتیجه هر دو الگوریتم به عملکرد نهایی تقریباً یکسانی دست یافته‌اند.



شکل ۱۵– نقشه شهری برای برنامه‌ریزی مسیر.

Fig. 15. Urban map for path planning.

شکل (۱۵) نقشه‌ای شهری را برای مسئله برنامه‌ریزی مسیر در یک محیط شهری شبیه‌سازی‌شده نشان می‌دهد. در این نقشه، ساختمان‌ها و موانع اصلی مانند بانک، بیمارستان، مدرسه، رستوران، و پارک به‌صورت بلوک‌های تیره نمایش داده شده‌اند، در حالی که مسیرهای بین آن‌ها نشان‌دهنده خیابان‌ها و فضاهای باز قابل‌حرکت برای عامل هوشمند هستند. نقطه سبز در گوشه بالا–چپ نقشه به‌عنوان مبدأ و نقطه قرمز در پایین–راست به‌عنوان مقصد مشخص شده است. این محیط با ساختار شبکه‌ای منظم، شرایطی واقعی برای ارزیابی عملکرد الگوریتم‌های مسیریابی مانند آاستار، یادگیری کیو و یادگیری کیو دوگانه فراهم می‌کند. در چنین محیطی، عامل باید با در نظر گرفتن موانع متعدد، کوتاه‌ترین و ایمن‌ترین مسیر را از مبدأ تا مقصد بیابد. این نقشه برای تحلیل توانایی الگوریتم‌ها در یادگیری مسیر بهینه در محیط‌های پیچیده شهری طراحی شده است.

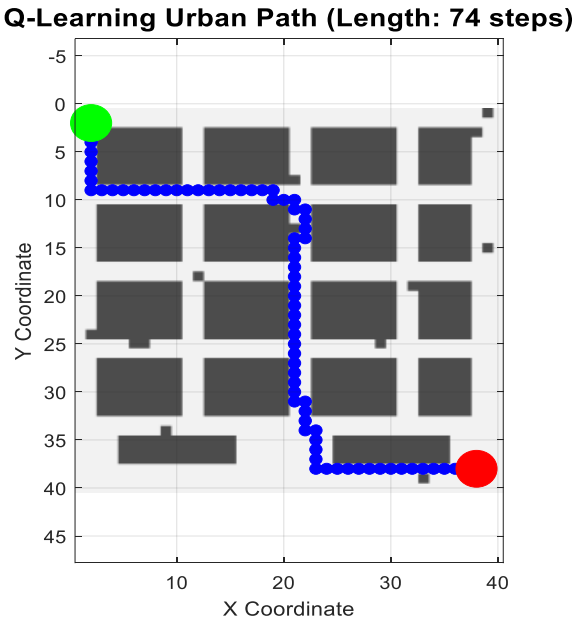


شکل ۱۶– مسیر شهری آاستار.

Fig. 16. A* urban path.

بر اساس شکل (۱۶)، مسیر بهینه توسط الگوریتم آاستار در یک محیط شهری با طول ۷۲ گام نمایش داده شده است. مسیر شناسایی شده از موانع ساختمانی (نقاط تیره) عبور کرده و توانسته است مسیری بهینه را بیابد. این مسیر از نواحی با تراکم متغیر موانع گذر کرده و انعطاف‌پذیری الگوریتم آاستاردر تطبیق با محیط‌های شهری پیچیده را به‌خوبی نشان می‌دهد.

کارایی این الگوریتم در سناریوی حاضر به‌واسطه‌ی به‌کارگیری یک تابع هیورستیک ترکیبی تضمین شده است که به طور همزمان فاصله اقلیدسی تا هدف و میزان مجاورت با موانع شهری را در فرآیند تصمیم‌گیری لحاظ می‌کند. این ویژگی باعث شده تا الگوریتم آاستار برای کاربردهای رباتیک شهری، ناوبری خودران و سیستم‌های هوشمند حمل‌ونقل گزینه‌ای قابل اعتمادباشد؛ زیرا می‌تواند در محیط‌های متغیر و پرتراکم، مسیرهایی ایمن و بهینه را تولید کند.

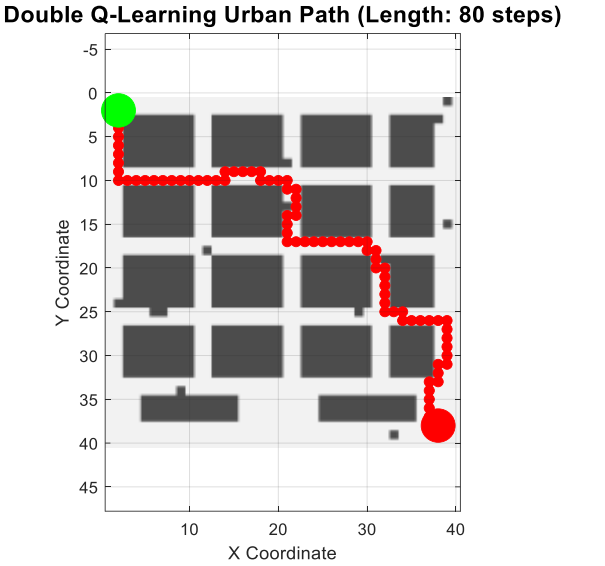


شکل ۱۷– مسیر شهری یادگیری کیو۳.

Fig. 17. Q-learning urban path.

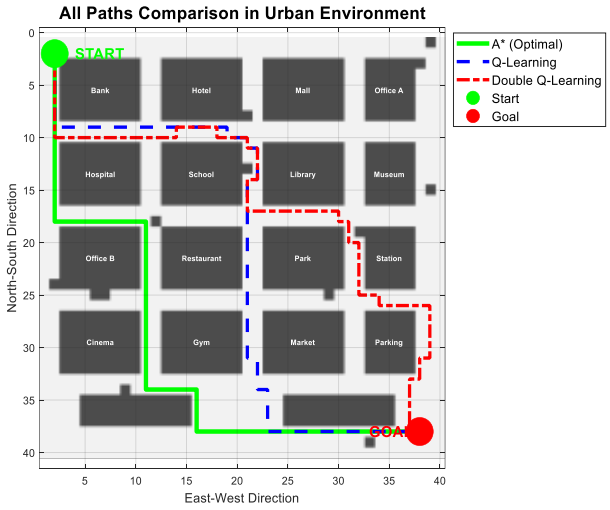
شکل (۱۷)، مسیر نهایی تولیدشده توسط الگوریتم یادگیری تقویتی یادگیری کیو را در یک محیط شبیه‌سازی‌شده شهری به تصویر می‌کشد. این محیط دوبعدی، شامل موانع ثابت (بلوک‌های خاکستری) است که نمایانگر ساختمان‌ها و مناطق غیرقابل عبور هستند. عامل از نقطه شروع (دایره سبز) در مختصات تقریبی (۳,۱) حرکت خود را آغاز کرده و با موفقیت به نقطه هدف (دایره قرمز) در مختصات تقریبی (۳۸,۳۸) رسیده است. مسیر یافت‌شده که با نقاط آبی مشخص شده، یک مسیر کاملاً ایمن و بدون برخورد با موانع است که طول آن معادل ۷۴ گام است. این نتیجه، کارایی بالای الگوریتم یادگیری کیو را در حل مسائل مسیریابی در محیط‌های ایستا و دارای محدودیت به اثبات می‌رساند. این الگوریتم با موفقیت سیاستی را فرا گرفته است که عامل را قادر می‌سازد تا با حرکات صحیح خود مجموع پاداش خود را برای رسیدن به هدف، بیشینه کند.

شکل (۱۸)، مسیر تولید شده توسط الگوریتم یادگیری کیو دوگانه در یک محیط شهری شبیه‌سازی‌شده را نمایش می‌دهد. این مسیر با موفقیت از نقطه شروع به نقطه هدف می‌رسد و نشان می‌دهد که الگوریتم توانسته است یک سیاست معتبر برای مسیریابی بدون برخورد با موانع را فرا بگیرد. با این‌حال، تحلیل دقیق‌تر مسیر یادگیری کیو دوگانه نشان می‌دهد که اگرچه این یک راه‌حل موفقیت‌آمیز است، اما از نظر طول، یک مسیر زیربهینه محسوب می‌شود. این مسیر به وضوح طولانی‌تر و دارای نوسان‌های بیشتری نسبت به مسیر بهینه (آاستار) است. این پدیده، ماهیت اکتشافی و فرآیند یادگیری مبتنی بر آزمون و خطای الگوریتم‌های یادگیری تقویتی را نشان می‌دهد.



شکل ۱۸– مسیر شهری یادگیری کیو دوگانه.

Fig. 18. Double Q-learning urban path.

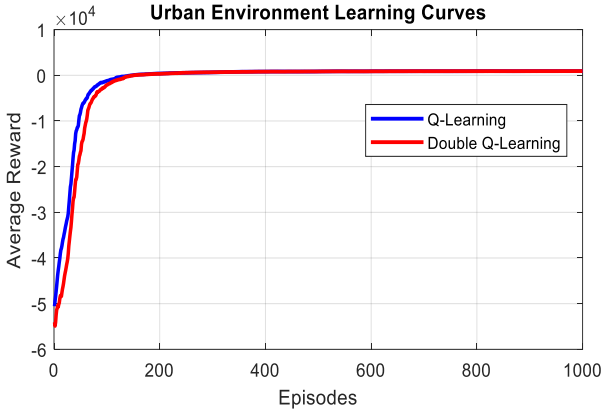


شکل ۱۹– مقایسه همه مسیرها در محیط شهری.

Fig. 19. Comparison of all paths in the urban environment.

شکل (۱۹) یک تحلیل مقایسه‌ای بصری از مسیرهای تولید شده توسط سه الگوریتم مسیریابی مجزا را در یک محیط شهری شبیه‌سازی‌شده ارائه می‌دهد: الگوریتم جستجوی آاستار (خط سبز)، یادگیری کیو (خط آبی) و یادگیری کیو دوگانه (خط قرمز). هر سه الگوریتم باید یک مسیر بهینه از نقطه شروع به نقطه هدف پیدا کنند. همان‌طور که مشاهده می‌شود، الگوریتم آاستار که به‌عنوان معیار بهینه

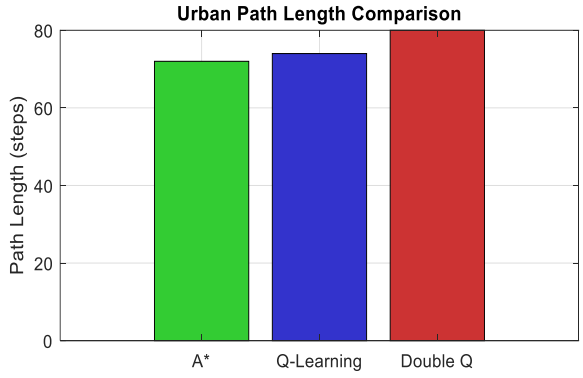
در نظر گرفته شده، مسیری کوتاه و مستقیم را پیدا کرده است. در مقابل، دو الگوریتم مبتنی بر یادگیری تقویتی مسیرهای متفاوتی را انتخاب کرده‌اند. این نتایج به وضوح برتری الگوریتم جستجوی آاستار را در یافتن کوتاه‌ترین مسیر در یک محیط ایستا و با نقشه مشخص نشان می‌دهد. از آنجاکه آاستار از اطلاعات کامل محیط برای یافتن مسیر بهینه استفاده می‌کند، مسیر آن به‌عنوان یک خط پایه برای ارزیابی کارایی الگوریتم‌های دیگر عمل می‌کند. در مقابل، الگوریتم‌های کیو و یادگیری کیو دوگانه که از طریق آزمون و خطا و بدون دانش قبلی از نقشه، سیاست حرکتی خود را یاد می‌گیرند، مسیرهای دیگری را یافته‌اند. نکته قابل توجه دیگر، مقایسه بین دو الگوریتم یادگیری تقویتی است؛ در این مسئله خاص، مسیر حاصل از الگوریتم کیو، کوتاهتر از مسیر یادگیری کیو دوگانه است. این امر نشان می‌دهد که بهبودهای بیش‌براوردی در الگوریتم یادگیری کیو دوگانه لزوماً در هر آزمایش منجر به یافتن مسیر بهتر یا کوتاه‌تر نمی‌شود. در مجموع، نمودار تفاوت ذاتی بین الگوریتم‌های مبتنی بر جستجوی قطعی و روش‌های یادگیری تقویتی را به‌خوبی به تصویر می‌کشد.



شکل ۲۰– تحلیل همگرایی پاداش متوسط در الگوریتم‌های یادگیری کیو و یادگیری دوگانه کیو.

Fig. 20. Analysis of average reward convergence in q-learning and double q-learning algorithms.

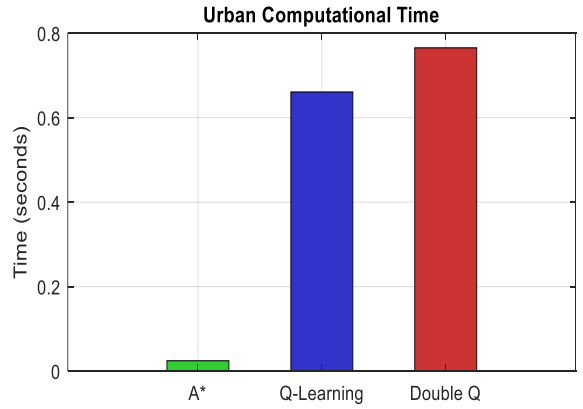
درشکل (۲۰) روند همگرایی پاداش متوسط دو الگوریتم یادگیری کیو و یادگیری دوگانه کیو در طول ۱۰۰۰ اپیزود نمایش داده شده است. همان‌طور که مشاهده می‌شود، هر دو الگوریتم از پاداش متوسط منفی در ابتدای فرآیند یادگیری شروع کرده و به تدریج به سمت صفر همگرا می‌شوند. در مراحل اولیه، یادگیری دوگانه کیو پاداش متوسط کمتری (منفی‌تر) نسبت به یادگیری کیو نشان می‌دهد که می‌تواند ناشی از محافظه‌کاری بیشتر این الگوریتم در برآورد ارزش اقدامات باشد. با این‌حال، پس از حدود ۳۰۰ اپیزود، عملکرد هر دو الگوریتم تقریباً یکسان شده و هر دو به پاداش متوسط صفر می‌رسند که نشان‌دهنده دستیابی به یک سیاست بهینه است.



شکل ۲۱– مقایسه طول مسیر شهری.

Fig. 21. Urban path length comparison.

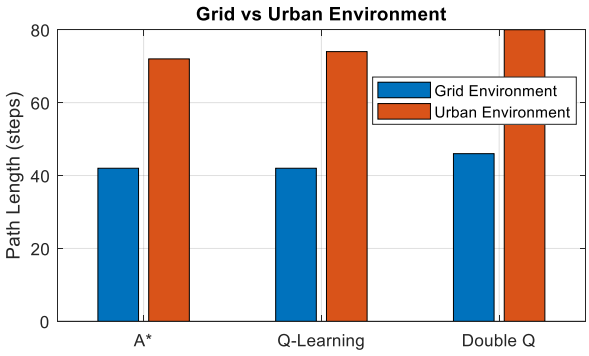
شکل (۲۱) نشان می‌دهد که الگوریتم آاستار با طول مسیر ۷۲ گام کوتاه‌ترین و در نتیجه بهینه‌ترین مسیر را یافته است. الگوریتم یادگیری کیو با طول ۷۴ گام عملکردی نزدیک به آاستار داشته و توانسته مسیر تقریباً بهینه‌ای را در محیطی شامل موانع شهری پیچیده شناسایی کند. در مقابل، الگوریتم یادگیری کیو دوگانه با طول مسیر ۸۰ گام مسیر بلندتری طی کرده است که نشان‌دهنده دقت کمتر یا همگرایی کندتر آن در این محیط خاص است. به‌طور کلی، این نتایج بیانگر آن است که الگوریتم آاستار همچنان در یافتن مسیرهای دقیق‌تر و کوتاه‌تر نسبت به روش‌های یادگیری تقویتی برتری دارد، اما کیو نیز توانسته با اختلاف اندکی عملکردی قابل‌قبول و نزدیک به مسیر بهینه ارائه دهد. تفاوت اندک میان طول مسیرها حاکی از آن است که روش‌های یادگیری تقویتی، با تنظیمات مناسب، می‌توانند در آینده جایگزینی هوشمند و مستقل برای الگوریتم‌های کلاسیکی مانند آاستار در محیط‌های واقعی باشند.



شکل ۲۲– زمان محاسباتی در محیط شهری

Fig. 22. Computational time in urban environment.

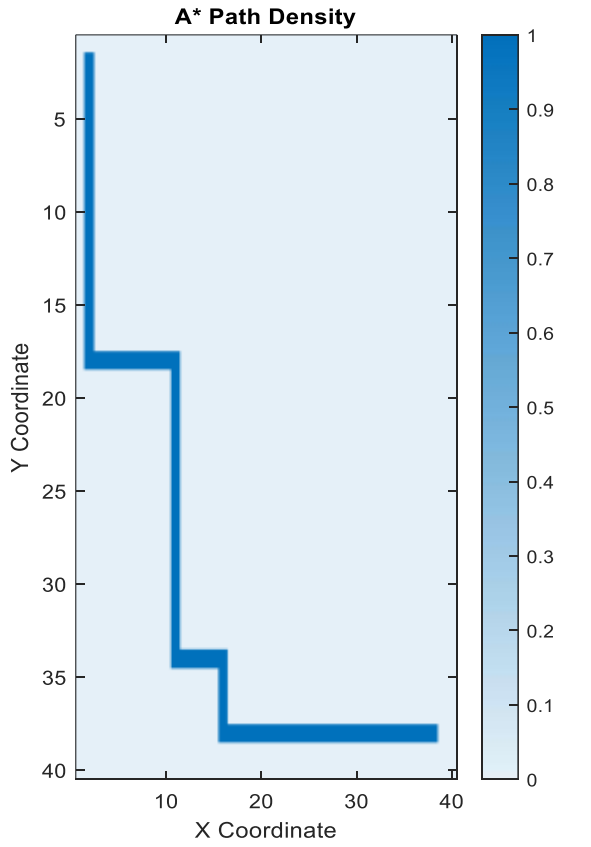
شکل (۲۲) زمان محاسباتی الگوریتم‌ها را در محیط شهری نشان می‌دهد. الگوریتم A* با میانگین زمان ۰/۰۲۶ ثانیه سریع‌ترین عملکرد را داشته است. این برتری ناشی از ماهیت قطعی و ساختار جستجوی هدایت‌شده الگوریتم A* است که بدون نیاز به فرایندهای یادگیری تکراری، مسیر بهینه را مستقیماً محاسبه می‌کند. در مقابل، الگوریتم‌های کیو و یادگیری کیو دوگانه به‌ترتیب با زمان‌های ۰/۶۶۲ و ۰/۷۶۶ ثانیه، زمان محاسباتی بیشتری نیاز دارند که ناشی از فرآیند یادگیری مبتنی بر تعامل و به‌روزرسانی تدریجی مقادیر کیو است. به‌طور کلی، این مقایسه نشان می‌دهد که در محیط‌های ایستا و با دسترسی کامل به نقشه، A* از نظر سرعت اجرای لحظه‌ای برتری محسوسی دارد؛ اما الگوریتم‌های یادگیری تقویتی در محیط‌های پویا و ناشناخته قابلیت انطباق و تصمیم‌گیری خودکار را فراهم می‌کنند. این ویژگی برای سامانه‌های ناوبری خودران و ربات‌های هوشمند اهمیت راهبردی دارد.



شکل ۲۳– مقایسه محیط شبکه‌ای و محیط شهری.

Fig. 23. Comparison of network environment and urban environment.

شکل (۲۳) مقایسه عملکرد الگوریتم‌ها در دو محیط شبکه‌ای و شهری را نشان می‌دهد. هر سه الگوریتم A*، کیو و یادگیری کیو دوگانه در محیط شبکه‌ای مسیرهایی با طول یکسان، یعنی ۴۲ گام، تولید کرده‌اند که بیانگر ساختار ساده محیط است. در این شرایط، همه الگوریتم‌ها بدون چالش قابل‌توجهی قادر به یافتن مسیر بهینه بوده‌اند. با این وجود در محیط شهری که شامل موانع نامنظم و مسیرهای پیچیده‌تر است، تفاوت عملکردها آشکار می‌شود. طول مسیر بهینه‌ی الگوریتم A* در این محیط ۷۰ گام بوده است، در حالی‌که الگوریتم کیو مسیر را با حدود ۷۳ گام و یادگیری کیو دوگانه با ۸۰ گام طی کرده‌اند. این افزایش طول مسیر نشان‌دهنده تأثیر مستقیم پیچیدگی محیط و تراکم موانع بر دقت و همگرایی الگوریتم‌های یادگیری تقویتی است. در نتیجه، می‌توان گفت که در محیط‌های ساده، همه روش‌ها عملکرد مشابهی دارند، اما با افزایش پیچیدگی، عملکرد الگوریتم‌ها متفاوت است.

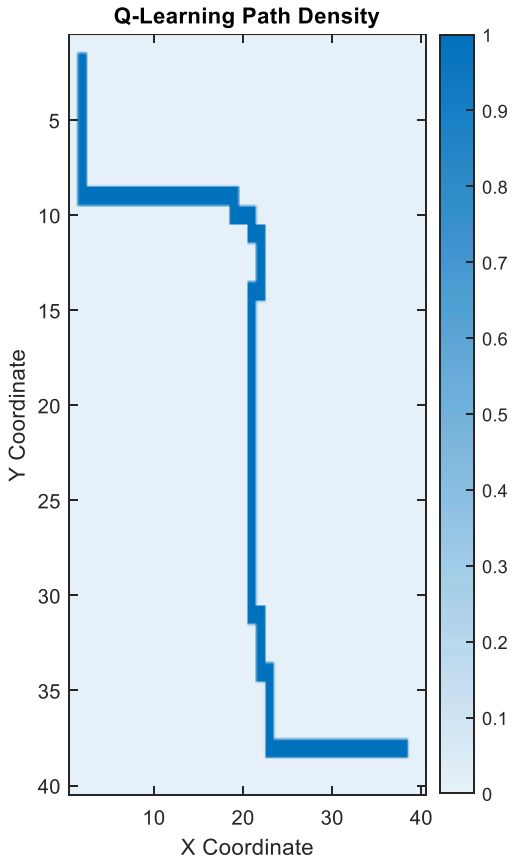


شکل ۲۴– چگالی مسیر A*.

Fig. 24. A* path density.

شکل (۲۴)، یک نقشه حرارتی از چگالی مسیر تولید شده توسط الگوریتم A* را در فضای مسئله نشان می‌دهد. محورهای افقی و عمودی مختصات محیط را نشان می‌دهند و رنگ هر پیکسل بیانگر میزان تکرار یا احتمال عبور عامل از آن نقطه است، که در آن رنگ سفید نشان‌دهنده حداکثر چگالی (مقدار ۱) و رنگ سیاه نشان‌دهنده چگالی صفر است. همانطور که دیده می‌شود، تنها یک مسیر واحد، کاملاً مشخص و با رنگ سفید وجود دارد و تمام نقاط دیگر محیط به رنگ سیاه هستند. این مشاهدهٔ کلیدی نشان می‌دهد که الگوریتم در تمام اجراها، دقیقاً یک مسیر یکسان را برای رسیدن از نقطه شروع به هدف انتخاب کرده است. این نتیجه به‌طور مستقیم ماهیت قطعی و بهینه الگوریتم A* را منعکس می‌کند. الگوریتم A*، به‌عنوان یک الگوریتم جستجوی آگاهانه که از نقشه کامل محیط استفاده می‌کند، با داشتن یک تابع هیوریستیک مشخص، همواره یک مسیر بهینه را پیدا می‌کند. برخلاف الگوریتم‌های مبتنی بر یادگیری یا تصادفی که ممکن است در مراحل اکتشاف مسیرهای متعددی را امتحان کنند و در نتیجه نقشه چگالی پراکنده‌تری ایجاد کنند. A* هیچ‌گونه عدم قطعیتی در انتخاب گام بعدی خود ندارد.

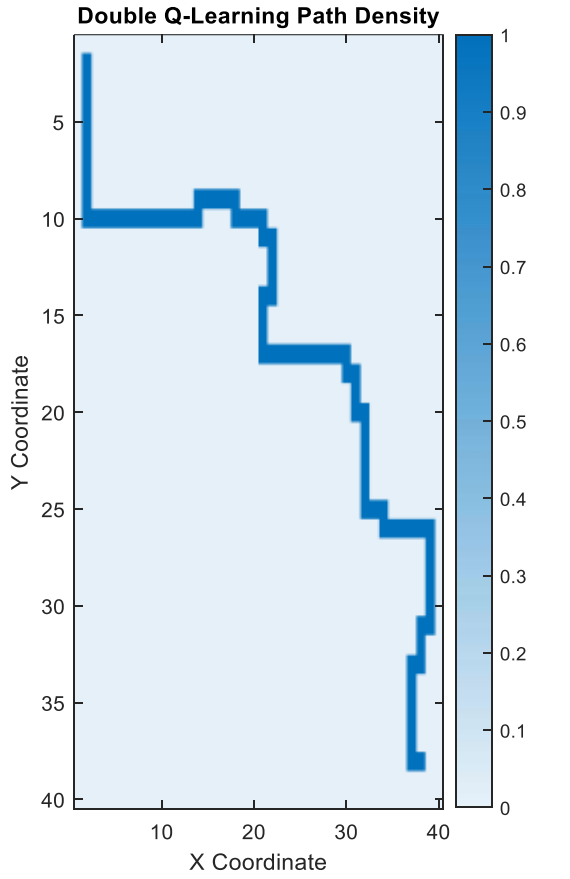
بنابراین، چگالی مسیر بر روی این خط ۱۰۰ درصد و در سایر نقاط صفر است، زیرا الگوریتم هیچ‌گاه از مسیر بهینه خود منحرف نمی‌شود. این نمودار به‌عنوان یک خط پایه ایده‌آل عمل می‌کند و نشان می‌دهد که یک راه‌حل قطعی و بهینه در این محیط چه شکلی دارد و می‌تواند معیاری برای سنجش میزان پراکندگی و کارایی الگوریتم‌های دیگر باشد.



شکل ۲۵– چگالی مسیر یادگیری کیو.

Fig. 25. Q-learning path density.

شکل (۲۵)، نقشه چگالی مسیر حاصل از عملکرد الگوریتم کیو را پس از اتمام فرآیند یادگیری نمایش می‌دهد. این نقشه حرارتی، یک مسیر کاملاً مشخص و با چگالی بالا (رنگ سفید) را در پس‌زمینه‌ای با چگالی صفر (رنگ سیاه) به تصویر می‌کشد. این تمرکز بالای چگالی بر روی یک خط سیر واحد، نشان‌دهنده همگرایی موفقیت‌آمیز الگوریتم و یادگیری یک سیاست قطعی و پایدار است؛ به این معنا که پس از آموزش کافی، عامل در اجراهای متوالی به‌صورت مداوم یک مسیر مشابه را برای رسیدن به هدف انتخاب می‌کند. نکته قابل توجه در این نمودار، شکل مسیر در نقاط تغییر جهت، به‌ویژه در مختصات (۲۰،۱۰) و (۲۲،۳۵)، است که دارای ساختاری پلکانی و غیرمستقیم می‌باشد. این پدیده، محدودیت‌های فضای عمل گسسته عامل را منعکس می‌کند؛ جایی‌که برای ایجاد یک حرکت مورب یا یک پیچ بهینه، عامل ناچار به انجام یک سری حرکات متناوب عمودی و افقی است. بنابراین، این نمودار نه تنها پایداری سیاست نهایی فراگرفته‌شده را تأیید می‌کند، بلکه جزئیات دقیقی از نحوه تعامل عامل با محدودیت‌های محیط و بهینه‌سازی محلی مسیر خود را نیز آشکار می‌سازد.

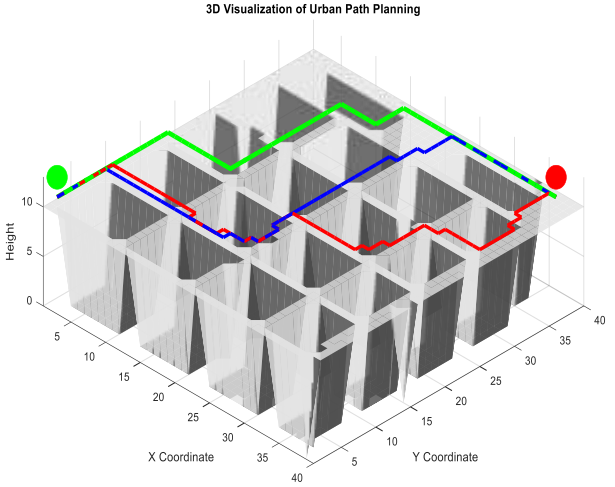


شکل ۲۶– چگالی مسیر یادگیری کیو دوگانه.

Fig. 26. Double Q-learning path density.

شکل (۲۶)، نقشه چگالی مسیر را برای الگوریتم یادگیری کیو دوگانه به نمایش می‌گذارد و به‌وضوح نشان‌دهنده همگرایی الگوریتم به یک سیاست کاملاً پایدار و قطعی است. وجود یک مسیر واحد با چگالی بالا (رنگ سفید) در مقابل پس‌زمینه با چگالی صفر (رنگ سیاه) مؤید آن است که عامل پس از اتمام فرآیند یادگیری، به‌طور مداوم یک استراتژی حرکتی یکسان را برای رسیدن به هدف انتخاب می‌کند. نکته تحلیلی مهم در این نمودار، شکل مسیر است که در مقایسه با مسیر بهینه، پیچیده‌تر و طولانی‌تر به‌نظر می‌رسد و دارای ساختاری پلکانی و پر از تغییر جهت‌های کوچک است. این ویژگی، نتیجه مستقیم فرآیند یادگیری اکتشافی است؛ جایی‌که

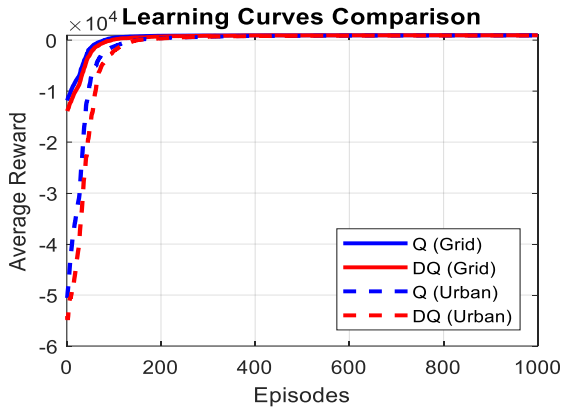
الگوریتم از طریق آزمون و خطا، یک مسیر ایمن و موفق را کشف کرده و آن را به‌عنوان سیاست نهایی خود تقویت نموده است، حتی اگر این مسیر کوتاه‌ترین راه ممکن نباشد. بنابراین، این نمودار به‌خوبی نشان می‌دهد که یادگیری کیو دوگانه چگونه یک راه‌حل کارآمد را پیدا می‌کند



شکل ۲۷– نمایش سه‌بعدی برنامه‌ریزی مسیر شهری.

Fig. 27. 3D Visualization of urban path planning.

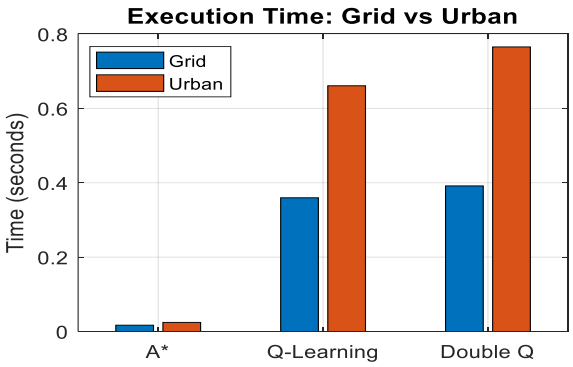
شکل (۲۷)، یک تصویر سه‌بعدی جامع از نتایج برنامه‌ریزی مسیر در یک محیط شبیه‌سازی‌شده شهری را ارائه می‌دهد و به مقایسه مسیرهای نهایی سه الگوریتم مجزا می‌پردازد. ویژگی برجسته این نمودار، تفاوت در استراتژی‌های مسیریابی است؛ الگوریتم A* استار یک مسیر بهینه و نسبتاً مستقیم را در امتداد لبه شمالی محیط انتخاب کرده است، در حالی‌که الگوریتم‌های یادگیری تقویتی کیو و یادگیری کیو دوگانه مسیرهای کاملاً متفاوتی انتخاب کرده‌اند. این مقایسه، تفاوت‌های بنیادین در رویکرد الگوریتم‌ها را به‌خوبی آشکار می‌سازد. مسیر بهینه A* استار، که به‌عنوان یک معیار عملکرد عمل می‌کند، نتیجه مستقیم یک الگوریتم جستجوی قطعی با دسترسی کامل به نقشه است و کوتاه‌ترین مسیر ممکن را انتخاب می‌کند. در مقابل، مسیرهای زیربهینه حاصل از کیو و یادگیری کیو دوگانه، ماهیت فرآیند یادگیری مبتنی بر آزمون و خطا را نشان می‌دهد. این الگوریتم‌ها بدون دانش قبلی، یک سیاست حرکتی معتبر را فرا گرفته‌اند که اگرچه لزوماً کوتاه‌ترین نیست، اما عامل را با موفقیت به‌هدف می‌رساند. تفاوت بین مسیرهای آبی و قرمز نیز نشان می‌دهد که چگونه تغییرات جزئی در مکانیزم یادگیری می‌تواند منجر به همگرایی به سیاست‌های نهایی متفاوتی شود. در مجموع، این نمودار به‌طور مؤثری مصالحه بین الگوریتم‌های با اطلاعات کامل و توانایی یادگیری تقویتی در محیط‌های ناشناخته را به تصویر می‌کشد.



شکل ۲۸– مقایسه منحنی‌های یادگیری.

Fig. 28. Comparison of learning curves.

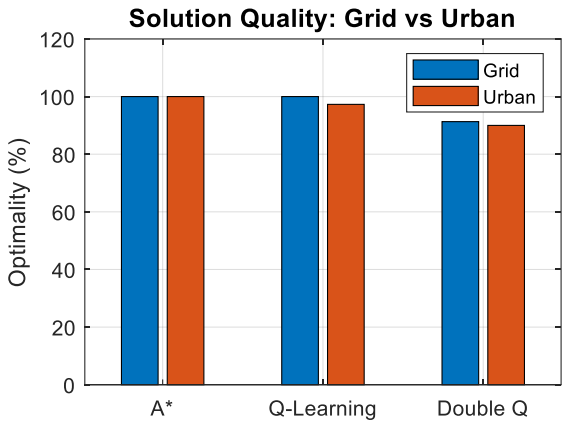
شکل (۲۸)، یک تحلیل مقایسه‌ای جامع از منحنی‌های یادگیری چهار سناریوی مختلف را ارائه می‌دهد که عملکرد الگوریتم‌های کیو و یادگیری کیو دوگانه را در دو محیط مجزا، یعنی یک محیط ساده شبکه‌ای و یک محیط پیچیده‌تر شهری، به تصویر می‌کشد. واضح‌ترین نتیجه‌ای که از این نمودار استنباط می‌شود، تأثیر مستقیم پیچیدگی محیط بر سرعت و کارایی یادگیری است. منحنی‌های مربوط به محیط شبکه‌ای (خطوط ممتد) به‌طور قابل توجهی سریع‌تر از منحنی‌های محیط شهری(خطوط منقطع) به سطح پاداش بهینه نزدیک به صفر همگرا می‌شوند. این امر نشان می‌دهد حل مسئله در محیط ساده‌تر نیازمند اپیزودهای آموزشی کمتری است. در مقایسه بین خود الگوریتم‌ها، مشاهده می‌شود که در هر دو محیط، الگوریتم استاندارد کیو (خطوط آبی) در مراحل اولیه یادگیری، نرخ بهبود سریع‌تری نسبت به یادگیری کیو دوگانه (خطوط قرمز) دارد. با این‌حال، نکته کلیدی این است که علی‌رغم این تفاوت‌های جزئی در دینامیک یادگیری و تأثیر غالب پیچیدگی محیط، تمام چهار رویکرد در نهایت به یک سطح عملکردی بسیار بالا و تقریباً یکسان همگرا می‌شوند که این موضوع کارایی و استحکام هر دو الگوریتم را در یافتن یک سیاست موفق برای حل مسئله مسیریابی در هر دو محیط تأیید می‌کند.



شکل ۲۹– زمان اجرای محیط شبکه‌ای و محیط شهری.

Fig. 29. Execution time in network vs. urban environment.

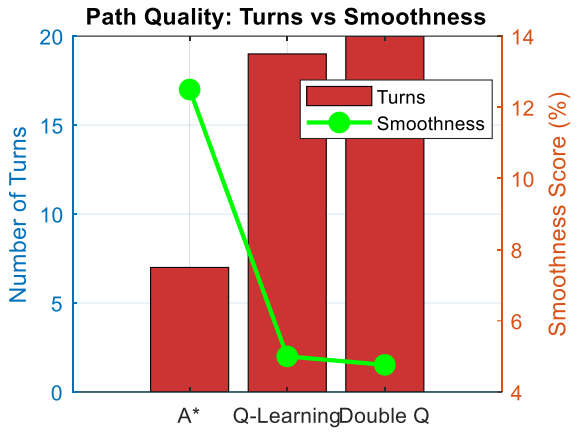
شکل (۲۹)، یک مقایسه کمی از زمان اجرا سه الگوریتم مسیریابی (آاستار ,کیو ,یادگیری کیو دوگانه) در دو محیط مجزا، یعنی شبکه‌ای و شهری، ارائه می‌دهد. یافته‌های این نمودار بیانگر تفاوت چشمگیر در زمان محاسباتی میان الگوریتم آاستار و الگوریتم‌های یادگیری تقویتی کیو و یادگیری کیو دوگانه است. الگوریتم آاستار با زمان اجرایی بسیار پایین، کارایی محاسباتی بالایی را نشان می‌دهد که ناشی از ماهیت قطعی و جستجوی مستقیم آن است. در مقابل، الگوریتم‌های کیو و یادگیری کیو دوگانه به‌دلیل فرآیند یادگیری تکرارشونده و به‌روزرسانی مداوم مقادیر کیو در طول اپیزودها، به‌طور قابل‌توجهی زمان بیشتری نیاز دارند. همچنین، نتایج نشان می‌دهد که پیچیدگی محیط عامل مؤثری در افزایش زمان اجرا است؛ به طوری‌که در محیط شهری، زمان محاسباتی هر سه الگوریتم به‌صورت معناداری بیشتر از محیط شبکه‌ای است. در نهایت، مقایسه بین دو الگوریتم یادگیری تقویتی نشان می‌دهد که زمان اجرای آن‌ها بسیار نزدیک است.



شکل ۳– کیفیت راه‌حل محیط شبکه‌ای و محیط شهری.

Fig. 30. Solution quality in network vs. urban environment.

شکل (۳۰) مقایسه‌ای کمی از کیفیت راه‌حل سه الگوریتم آاستار، کیو و یادگیری کیو دوگانه را در دو محیط شبکه‌ای و شهری نشان می‌دهد، که در آن کیفیت بر اساس درصد بهینگی ارزیابی شده است. برجسته‌ترین یافته، عملکرد الگوریتم آاستار است که در هر دو محیط با دستیابی مداوم به ۱۰۰ درصد بهینگی، نقش خود را به‌عنوان الگوریتم معیار تأیید می‌کند. در مقابل، عملکرد الگوریتم‌های یادگیری تقویتی وابستگی بیشتری به پیچیدگی محیط دارد. در محیط ساده شبکه‌ای، هر دو الگوریتم کیو و یادگیری کیو دوگانه قادر به دستیابی کامل به راه‌حل بهینه (۱۰۰ درصد) هستند، اما در محیط پیچیده‌تر شهری، افت محسوسی در کیفیت مشاهده می‌شود؛ به‌طوری‌که کیو همچنان عملکردی نزدیک به بهینه (حدود ۹۸ درصد) دارد، در حالی‌که کیفیت راه‌حل در یادگیری کیو دوگانه به حدود ۹۰ درصد کاهش می‌یابد. این یافته نشان می‌دهد که هرچند هر دو الگوریتم یادگیری توانایی یافتن راه‌حل‌های معتبر را دارند، اما افزایش پیچیدگی محیط و احتمال همگرایی به بهینه‌های محلی می‌تواند سبب انتخاب مسیرهای زیربهینه شود، پدیده‌ای که در این مورد برای یادگیری کیو دوگانه بارزتر است.



شکل ۳۱– مقایسه عملکرد الگوریتم‌های برنامه‌ریزی مسیر بر اساس معیارهای تعداددور و نرمی مسیر.

Fig. 31. Comparison of path planning algorithms based on turn count and path smoothness criteria.

در شکل (۳۱) عملکرد الگوریتم‌های مختلف برنامه‌ریزی مسیر از نظر دو معیار تعداد دورها و نرمی مسیر مقایسه شده است. همان‌طور که مشاهده می‌شود، الگوریتم آاستار با کمترین تعداد دور (۷/۵) و بیشترین میزان نرمی (۱۳/۰ درصد) کارایی بهتری را در تولید مسیرهای مستقیم و هموار نشان می‌دهد. در مقابل، الگوریتم یادگیری کیوو یادگیری کیو دوگانه اگرچه در کاهش تعداد دورها عملکرد متوسطی دارند (۱۹/۰ و ۴/۵ به‌ترتیب)، اما در معیار نرمی مسیر به‌طور محسوسی ضعیف‌تر عمل کرده‌اند. این نتایج حاکی از آن است که الگوریتم‌های مبتنی بر جستجوی اکتشافی مانند آاستار در محیط‌های ایستا می‌توانند مسیرهای بهینه‌تر و روان‌تری تولید کنند، در حالی که روش‌های یادگیری تقویتی ممکن است در پویایی‌های محیطی انعطاف بیشتری داشته باشند ولی نیاز به تنظیمات بیشتری برای بهبود کیفیت هندسی مسیر نهایی دارند.

در بررسی نتایج به‌دست‌آمده، مشاهده شد که در محیط ساده شبکه‌ای، الگوریتم یادگیری کیو دوگانه نسبت به یادگیری کیو عملکرد ضعیف‌تری از نظر درصد بهینگی مسیر نشان می‌دهد که این موضوع در نگاه اولیه ممکن است با هدف اصلی این الگوریتم در کاهش بیش‌برآوردی ناسازگار به‌نظر برسد. از این‌رو، لازم است این رفتار به‌صورت دقیق‌تری در چارچوب ویژگی‌های محیط و ماهیت الگوریتم تحلیل شود.

یادگیری کیو دوگانه اساساً با هدف کاهش بیش‌برآوردی در محیط‌های تصادفی و دارای نویز طراحی شده است، نه در محیط‌های ساده، قطعی و با فضای حالت محدود. در محیط شبکه‌ای ساده مورد استفاده در این پژوهش، به‌دلیل ثابت بودن موانع و قطعی بودن گذارها، میزان بیش‌برآوردی در یادگیری کیو استاندارد ناچیز است؛ بنابراین مزیت اصلی الگوریتم یادگیری کیو دوگانه در چنین شرایطی به‌خوبی ظاهر نمی‌شود. در مقابل، استفاده از دو جدول ارزش مستقل موجب محافظه‌کارانه‌تر شدن برآورد مقادیر ارزش شده و می‌تواند سرعت دستیابی به سیاست بهینه را در محیط‌های ساده کاهش دهد و در برخی موارد به همگرایی کندتر یا زیر بهینه منجر شود.

همچنین در سازوکار یادگیری کیو دوگانه، مقدار ارزش با استفاده از یک جدول برای انتخاب عمل و جدول دیگر برای ارزیابی آن محاسبه می‌شود که این امر معمولاً منجر به برآوردهای کمتر نسبت به حداکثرگیری مستقیم در یادگیری کیو استاندارد می‌گردد. در محیط‌های ساده که مقادیر ارزش با سرعت بیشتری به حالت پایدار نزدیک می‌شوند، این رفتار محافظه‌کارانه می‌تواند مانع از رسیدن سریع به بهینه سراسری شود.

با این‌حال، مزیت اصلی یادگیری کیو دوگانه در محیط‌های پیچیده‌تر و دارای عدم قطعیت بیشتر آشکار می‌شود. نتایج حاصل از محیط شهری پیچیده‌تر نشان می‌دهد که کاهش کارایی این الگوریتم در مقایسه با محیط ساده کمتر بوده و پایداری بیشتری نسبت به یادگیری کیو استاندارد دارد، به‌طوری‌که افت عملکرد آن در محیط پیچیده محدودتر است. این موضوع بیانگر آن است که یادگیری کیو دوگانه در شرایط پیچیده‌تر و غیرقطعی، رفتار پایدارتر و قابل‌اعتمادتری از خود نشان می‌دهد.

کاربردهای عملی الگوریتم‌های مسیریابی صرفاً به ربات‌های زمینی و سیستم‌های ناوبری شهری محدود نمی‌شود، بلکه این روش‌ها در طراحی مسیر پهپادها نیز کاربرد گسترده‌ای یافته‌اند. پهپادها به‌دلیل ماهیت پروازی و نیاز به پرهیز از موانع هوایی و زمینی، به الگوریتم‌های مسیریابی کارآمد و قابل اعتمادی نیاز دارند. روش‌های قطعی مانند A*Star به‌دلیل سرعت بالا و تضمین بهینگی، گزینه‌های مناسبی برای مسیریابی لحظه‌ای پهپادها در محیط‌های شناخته‌شده هستند. در مقابل، الگوریتم‌های یادگیری تقویتی مانند یادگیری کیو و یادگیری کیو دوگانه به‌دلیل قابلیت یادگیری از محیط و انطباق با شرایط پویا، برای مأموریت‌های پهپادی در محیط‌های ناشناخته و متغیر بسیار مناسب هستند. با توجه به نتایج این پژوهش، انتخاب الگوریتم مناسب برای مسیریابی پهپادها باید بر اساس عواملی مانند سطح دانش از محیط، نیاز به سرعت محاسباتی، اهمیت بهینگی مسیر و الزامات پایداری انجام شود [۲۶].

۳. نتیجه‌گیری

این پژوهش با انجام ارزیابی مقایسه‌ای جامع بین سه الگوریتم A*Star، یادگیری کیو و یادگیری دوگانه کیو در محیط‌های مسیریابی ایستا با موانع پیچیده، به نتایج قابل توجهی دست یافت. یافته‌ها به‌وضوح نشان داد که الگوریتم A*Star در محیط‌های شناخته‌شده و ایستا از برتری مطلق برخوردار است. این الگوریتم نه تنها کوتاه‌ترین مسیر (۴۲ گام) را با تضمین بهینگی کامل ارائه داد، بلکه با کمترین زمان اجرا (۰/۰۰۴۳ ثانیه) و بهترین کیفیت مسیر (۷ پیچش، شاخص نرم‌بودن ۰/۱۲۵) عملکردی بی‌نظیر از خود نشان داد. ثبات نتایج A*Star در ۳۰ اجرای مکرر (با انحراف معیار صفر در طول مسیر) نیز بر قابلیت اطمینان بالای این روش تأکید دارد. در مقابل، الگوریتم‌های یادگیری تقویتی اگرچه به موفقیت ۱۰۰ درصدی در یافتن مسیر دست یافتند، اما با چالش‌های متعددی روبرو بودند. یادگیری کیو توانست در محیط ساده به بهینگی کامل دست یابد، اما مسیرهای تولید شده از کیفیت پایین‌تری (۱۹ پیچش) برخوردار بودند و زمان اجرای آن ۶۴ برابر A*Star بود. یادگیری دوگانه کیو نیز اگرچه با هدف بهبود پایداری طراحی شده بود، اما نه تنها نتوانست برتری معنی‌داری در کاهش بیش‌برآورد نشان دهد، بلکه با افت بهینگی به ۹۱/۳ درصد و افزایش زمان اجرا به ۷۴ برابر A*Star مواجه شد. افزایش پراکندگی نتایج (انحراف معیار ۱/۴۱ در پاداش) نیز حاکی از ناپایداری بیشتر این الگوریتم نسبت به انتظارات نظری بود. تأثیر پیچیدگی محیط به وضوح در نتایج مشهود است. در محیط شهری پیچیده‌تر، عملکرد الگوریتم‌های یادگیری تقویتی با افت محسوس‌تری همراه بود. به‌طور خاص، یادگیری دوگانه کیو در محیط شهری با کاهش ۱۰ درصدی بهینگی نسبت به A*Star مواجه شد، در حالی که زمان اجرای آن همچنان به‌طور قابل توجهی بیشتر بود. این یافته نشان می‌دهد که با افزایش پیچیدگی محیط، مزیت نسبی الگوریتم‌های یادگیری‌محور کاهش می‌یابد. بر اساس یافته‌های این پژوهش، توصیه‌های کاربردی زیر ارائه می‌شود:

۱) در محیط‌های ایستا و شناخته‌شده، الگوریتم A*Star به‌عنوان گزینه اول و ترجیحی پیشنهاد می‌شود.

۲) در محیط‌های ناشناخته یا پویا که نیاز به قابلیت انطباق وجود دارد، یادگیری کیو می‌تواند گزینه مناسبی باشد، هرچند باید هزینه محاسباتی بالاتر و کیفیت مسیر پایین‌تر را پذیرفت.

۳) استفاده از یادگیری دوگانه کیو تنها در شرایطی توصیه می‌شود که پایداری یادگیری اولویت اصلی باشد و کاهش جزئی در بهینگی قابل قبول باشد.

محدودیت‌های پژوهش حاضر شامل تمرکز بر محیط‌های ایستا، استفاده از شبکه‌های گسسته و در نظر نگرفتن عوامل پویا مانند موانع متحرک است. برای پژوهش‌های آتی، بررسی عملکرد این الگوریتم‌ها در محیط‌های پویا و پیوسته، ادغام روش‌های هیوریستیک با یادگیری عمیق و توسعه الگوریتم‌های ترکیبی که مزایای هر دو رویکرد قطعی و یادگیری‌محور را در بر گیرند، پیشنهاد می‌شود. همچنین، بررسی تأثیر انواع مختلف توابع پاداش و استراتژی‌های اکتشاف می‌تواند به بهینه‌سازی عملکرد الگوریتم‌های یادگیری تقویتی در مسیریابی بینجامد.

تعارض منافع

هیچگونه تعارض منافع توسط نویسندگان بیان نشده است

مراجع

[1] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[2] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992, https://doi.org/10.1007/BF00992698.

[3] A. H. Borzabadi, A. V. Kamyad, M. H. Farahi, and H. H. Mehne, "Solving some optimal path planning problems using an approach based on measure theory," *Applied Mathematics and Computation*, vol. 170, no. 2, pp. 1418–1435, 2005, https://doi.org/10.1016/j.amc.2005.01.035.

[4] B. Q. A. Nguyen, N. T. Dang, T. T. Le, and P. N. Dao, "On-policy and Off-policy Q-learning algorithms with policy iteration for two-wheeled inverted pendulum systems," *Robotics and Autonomous Systems*, 2025, Art. No. 105111, https://doi.org/10.1016/j.robot.2025.105111.

[5] R. Wu, E. Luo, X. Li, H. Tang, and Y. Li, "Hybrid artificial bee colony algorithm with Q-learning for distributed heterogeneous flexible job shop scheduling problem considering machine preventive maintenance," *Swarm and Evolutionary Computation*, vol. 98, 2025, Art. no. 102134, https://doi.org/10.1016/j.swevo.2025.102134.

[6] X. Guo, Q.-K. Pan, W. Zhang, Z.-H. Miao, X.-L. Jing, and H.-Y. Sang, "A Q-learning-assisted memetic algorithm for joint vehicle scheduling problem for harvesting and transportation in smart agriculture," *Swarm and Evolutionary Computation*, vol. 96, 2025, Art. no. 102007, https://doi.org/10.1016/j.swevo.2025.102007.

[7] C. Sun, L. Hou, S. Yu, and J. Shu, "HQA: Hybrid Q-learning and AODV multi-path routing algorithm for flying Ad-Hoc networks," *Vehicular Communications*, 2025, Art. no. 100947, https://doi.org/10.1016/j.vehcom.2025.100947.

[8] X. H. Qiu, S. J. Tian, A.-F. Liu, Y.-H. Wei, H. Sekiya, and Y.-J. Choi, "Q-learning-based hyper-heuristic algorithm for priority and precedence dual-driven task assignment in spatial crowdsourcing," *Expert Systems with Applications*, vol. 297, 2026, Art. no. 129327, https://doi.org/10.1016/j.eswa.2025.129327.

[9] Z. Duan, Y. Zhang, Z. Shao, Z. Xu, and Z. Xiang, "Advanced robot path planning on rough terrain: A Q-learning-based multi-objective PSO algorithm," *Applied Soft Computing*, 2025, Art. no. 113798, https://doi.org/10.1016/j.asoc.2025.113798.

[10] J. Guo, B. Peng, B. Du, K. Wang, and Y. Li, "Q-learning-based multi-objective spotted hyena algorithm for flexible open shop scheduling problem with consideration of preventive maintenance and travel/setup times," *Journal of Manufacturing Systems*, vol. 82, pp. 224–238, 2025, https://doi.org/10.1016/j.jmsy.2025.06.001.

[11] Z. Q. Zhang, X. Y. Wu, B .Qian, R. Hu, and J. B. Yang, "A Q-learning-based multi-objective hyper-heuristic algorithm for energy-efficient integrated distributed hybrid flow-shop scheduling with preventive maintenance," *Computers & Operations Research*, 2025, Art. no. 107267, https://doi.org/10.1016/j.cor.2025.107267.

[12] L. Xu ,T. Zhou, K. Li, J. Chen, and H. Zhang, "Q-learning-driven multi-population cooperative evolutionary algorithm with local search for scheduling of network-shared manufacturing resources," *Computers & Operations Research*, vol. 180, 2025, Art. no. 107076, https://doi.org/10.1016/j.cor.2025.107076.

[13] G .Zuo, Z. Jia, Z. Wu, J. Shi, and G. Wang, "A Q-learning guided dual population genetic algorithm for distributed permutation flow shop scheduling problem with machine having fuzzy processing efficiency," *Expert Systems with Applications*, 2025, Art. no. 127882, https://doi.org/10.1016/j.eswa.2025.127882.

[14] F. B. Ozsoydan, "A trajectory-based algorithm enhanced by Q-learning and cloud integration for hybrid flexible flowshop scheduling problem with sequence-dependent setup times: A case study," *Computers & Operations Research*, vol. 181, 2025, Art. no. 107079, https://doi.org/10.1016/j.cor.2025.107079.

[15] H. Dinçer *et al.*, "Q-learning algorithm and molecular fuzzy multi-objective particle swarm optimization-based decision-making approach to circular economy-oriented investment alternatives for renewable energy technologies," *Information Sciences*, vol. 718, 2025, Art. no. 122378,

https://doi.org/10.1016/j.ins.2025.122378.

[16] Z. Zhaohui, Z. Jiaqi, and L. Jing, "Q-learning-based semi-fixed clustering routing algorithm in WSNs," *Ad Hoc Networks*, vol. 174, 2025, Art. no. 103837, , https://doi.org/10.1016/j.adhoc.2025.103837.

[17] Z. Li, Y. Wang, Y. Han, K. Gao, and J. Li, "Q-Learning-Driven Accelerated Iterated Greedy Algorithm for Multi-Scenario Group Scheduling in Distributed Blocking Flowshops," *Knowledge-Based Systems*, vol. 317, 2025, Art. no. 113424, https://doi.org/10.1016/j.knosys.2025.113424.

[18] H. Xuan, Q.-Q. Zheng, L. Lv, and B. Li, "Q-learning based estimation of distribution algorithm for scheduling distributed heterogeneous flexible flow-shop with mixed buffering limitation," *Engineering Applications of Artificial Intelligence*, vol. 149, 2025, Art. no. 110537, https://doi.org/10.1016/j.engappai.2025.110537.

[19] K. Wang, X. Ma, Y. Li, Y. Luo, Y. Li, and L. Gao, "An adaptive genetic algorithm based on Q-learning for energy-efficient e-waste disassembly line balancing and rebalancing considering task failures," *Journal of Manufacturing Systems*, vol. 80, pp. 1–19, 2025, https://doi.org/10.1016/j.jmsy.2025.02.009.

[20] G. Wang, Z. Zhang, and A. Nazarova, "A digital twin-based platform for testing and optimization of path planning algorithms," *Simulation Modelling Practice and Theory*, 2025, Art. no. 103206, 2025, https://doi.org/10.1016/j.simpat.2025.103206.

[21] S. Zeng, Y. Wang, X. Yu, H. Song, X. Guo, and Z. Li, "A dyeing clustering algorithm based on ant colony path-finding mechanism," *Engineering Applications of Artificial Intelligence*, vol. 136, 2024, Art. no. 108941, https://doi.org/10.1016/j.engappai.2024.108941.

[22] A. Giuffrida, N. Basilico, and F. Amigoni, "An empirical evaluation of learning-based multi-agent path finding algorithms in warehouse environments," *Robotics and Autonomous Systems*, vol. 194, 2025, Art. no. 105149, https://doi.org/10.1016/j.robot.2025.105149.

[23] Q. Zhou, Y. Lian, J. Wu, M. Zhu, H. Wang, and J. Cao, "An optimized Q-Learning algorithm for mobile robot local path planning," *Knowledge-Based Systems*, vol. 286, 2024, Art. no. 111400, https://doi.org/10.1016/j.knosys.2024.111400.

[24] J. Gao, Y. Li, X. Shi, and K. Yan, "Customer satisfaction optimization in due-aware multi-agent path finding," *Information Sciences*, 2025, Art. no. 122782, https://doi.org/10.1016/j.ins.2025.122782.

[25] S. Venu and M. Gurusamy, "A Comprehensive Review of Path Planning Algorithms for Autonomous Navigation," *Results in Engineering*, 2025, Art. no. 107750, https://doi.org/10.1016/j.rineng.2025.107750.

[26] S .Aslan and S. Demirci, "An immune plasma algorithm with Q-learning based pandemic management for path planning of unmanned aerial vehicles," *Egyptian Informatics Journal*, vol. 26, 2024, Art. no. 100468, https://doi.org/10.1016/j.eij.2024.10046.

[27] Y. Wang and G. Wang, "An adaptive and efficient path planning algorithm for UAV navigation in complex environments," *Computers & Operations Research*, vol. 185, 2026, Art. no. 107296, https://doi.org/10.1016/j.cor.2025.107296.

[28] X. Ni, W. Hu, Q. Fan, Y. Cui, and C. Qi, "A Q-learning based multi-strategy integrated artificial bee colony algorithm with application in unmanned vehicle path planning," *Expert Systems with Applications*, vol. 236, 2024, Art. no.121303, https://doi.org/10.1016/j.eswa.2023.121303.

[29] J. Wu, H. Jin, S. Cai, and L. Liu, "A delay tolerant network routing algorithm based on multi-step double Q-learning," *IET Communications*, vol. 17, no. 11, pp. 1321–1333, 2023, https://doi.org/10.1049/cmu2.12626.

[30] M. S. El, "Intelligent routing agent based on Q-learning and Markov decision processes for routing optimization in DTN networks," *International Journal of Intelligent Networks*, vol. 6, pp. 97–112, 2025, https://doi.org/10.1016/j.ijin.2025.07.001.

[31] N. Manogaran *et al.*, "Developing a Novel Adaptive Double Deep Q-Learning-Based Routing Strategy for IoT-Based Wireless Sensor Network with Federated Learning," *Sensors*, vol. 25, no. 10, 2025, Art. no. 3084, https://doi.org/10.3390/s25103084.

[32] J. Liu *et al.*, "QMR:Q-learning based Multi-objective optimization Routing protocol for Flying Ad Hoc Networks," *Computer Communications*, vol. 150, pp. 304–316, 2020, https://doi.org/10.1016/j.comcom.2019.11.011.

[33] M. Ghanifar, A. Nikkhah, M. Kamzan, M. Teshnehlab, and M. Tayefi, "Online trajectory regeneration for multirotors via a proportional-derivative physics-informed neural network," *IEEE Access*, vol. 13, pp. 145168 - 145190, 2025, https://doi.org/10.1109/ACCESS.2025.3598862.

[34] M. Ghanifar, M. Omidali, A. Nikkhah, M. Teshnehlab, and M. Tayefi, "Fuzzy PID-inspired Potential Field Path Planning for Quadrotors in Environmental Monitoring," *Journal of Technology in Aerospace Engineering*, vol. 37, Special Issue, pp. 1–12, 2025, https://doi.org/10.22034/jtae.2026.1400.

[35] J. Gandhi and Z. Narmawala, "Q-Learning and Double Q-Learning Based Routing in Opportunistic Networks," *International Journal of Networked and Distributed Computing*, vol. 14, 2026, Art. no. 13, https://doi.org/10.1007/s44227-026-00090-7.

[36] J. Singh, S. K. Dhurandher, I. Woungang, and L. Barolli, "Double Q-learning based routing protocol for opportunistic networks," *Journal of High Speed Networks*, vol. 29, no. 1, pp. 1–14, 2023, https://doi.org/10.3233/JHS-22201.

[37] J. Zhou, X. Gong, L. Sun, Y. Xie, and X. Yan, "Adaptive Routing Strategy Based on Improved Double Q-Learning for Satellite Internet of Things," *Security and Communication Networks*, vol. 2021, no. 1, 2021, Art. no. 5530023, https://doi.org/10.1155/2021/5530023.

[38] M. Patil, P. Tambolkar, and S. Midlam-Mohler, "Optimizing Traffic Routes With Enhanced Double Q-Learning," *IET Intelligent Transport Systems*, vol. 19, no. 1, 2025, Art. no. e70002, https://doi.org/10.1049/itr2.70002.

[39] I. Shafieenejad, "Design of swarm guidance for aerial spherical robots with social intelligence inspired by grasshopper' optimal motion and fuzzy logic," *Aerospace Knowledge and Technology*, vol. 14, no. 1 ,pp. 193–216, 2025, (in Persian), 20.1001.1.23221070.1404.14.1.11.4.

[40] I. Shafieenejad, "Designing optimal Robust Guidance and Model Predictive controller for Mono-Wing Aerial Robot by the Maple Seed Tree Inspired," *Journal of Aeronautical Engineering*, vol. 27, no. 1, pp. 56–74, 2025, (in Persian), https:doi.org/10.22034/joae.2025.481623.1251.

[41] S. J. Russell, S. Russell, and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2020.