

دریافت مقاله: ۱۴۰۴/۰۶/۱۲

پذیرش مقاله: ۱۴۰۴/۰۷/۲۰

نوع مقاله: پژوهشی

کنترل تطبیقی چراغ راهنمایی با یادگیری تقویتی عمیق حساس به فاز و حافظه تجربه تفکیکی برای کاهش زمان انتظار و طول صف

صفورا اخلاقی*

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد اسلامی واحد قزوین، قزوین، ایران

پست الکترونیکی: Safoura.Akhlaghi@gmail.com

چکیده

مسئله کنترل تطبیقی چراغ راهنمایی در شبکه‌های شهری به دلیل پویایی تقاضا، ناپایداری جریان، و برهم‌کنش فازها چالش‌برانگیز است؛ روش‌های زمان‌بندی کلاسیک در مواجهه با عدم تعادل ورودی-خروجی و رخدادهای نادر، صف‌های طولانی و زمان‌های انتظار زیاد ایجاد می‌کنند. رویکرد پیشنهادی ما یک چارچوب سه‌مرحله‌ای یادگیری تقویتی عمیق است. (۱) مدل‌سازی وضعیت غنی شامل طول صف و تعداد خودرو هر خط، زمان انتظار تجمعی، فشار ترافیکی نرمال‌شده هر حرکت و حالت فعلی چراغ (امکان افزودن بازنمایی مکانی). (۲) تصمیم‌گیری حساس به فاز: برای هر فاز چراغ مسیر محاسباتی جداگانه در شبکه ارزش در نظر گرفته می‌شود تا تداخل گرادیان میان فازها حذف و پایداری افزایش یابد؛ سیاست با ϵ -greedy اجرا می‌شود. (۳) یادگیری با حافظه تفکیکی: تجربه‌ها برحسب جفت فاز-اقدام در میانگیرهای جدا ذخیره و با نمونه‌گیری متوازن آموزش داده می‌شوند. تابع پاداش ترکیبی از فشار، طول صف، زمان انتظار و تعداد عبوری‌هاست. ارزیابی روی دو مجموعه داده‌ها عمومی موناکو و هاربین نشان می‌دهد روش ما سریع‌تر

همگرا شده و پاداش بیشتری کسب می‌کند. در داده‌های شهر موناکو میانگین زمان انتظار نسبت به LSTM-PPO و CNN-DQN به ترتیب ۶۳ درصد و ۵۰ درصد در حالت خلوت و ۴۱ درصد و ۲۹ درصد در حالت شلوغ کاهش یافت؛ طول صف نیز ۴۴ درصد و ۲۹ درصد در حالت خلوت و ۴۲ درصد و ۳۰ درصد در حالت شلوغ کمتر شد. در داده‌های شهر هاربین کاهش‌ها برای زمان انتظار ۵۳ درصد و ۴۳ درصد در حالت خلوت و ۳۹ درصد و ۳۰ درصد در حالت شلوغ و برای طول صف ۴۳ درصد و ۲۷ درصد در حالت خلوت و ۴۴ درصد و ۳۱ درصد در حالت شلوغ به دست آمد.

واژه‌های کلیدی: کنترل تطبیقی چراغ راهنمایی، یادگیری تقویتی عمیق، حافظه تفکیکی، کاهش زمان انتظار، کاهش طول صف

۱. مقدمه

ازدحام ترافیکی در شهرهای بزرگ نه تنها باعث اتلاف زمان و افزایش مصرف سوخت می‌شود، بلکه اثرات اقتصادی و زیست‌محیطی قابل‌توجهی نیز ایجاد می‌کند. راهبردهای کنترل سیگنال سنتی اعم از برنامه‌های

* نویسنده مسئول

ثابت‌زمان یا برنامه‌های قابل‌تغییر محدود به دلیل نالیستایی تقاضا، ناهمگنی تقاطع‌ها و پویایی‌های چندمقیاسی جریان ترافیک، در پاسخ‌گویی بی‌درنگ دچار کاستی‌اند [۱،۲].

مرورهای جامع اخیر بر سامانه‌های مدیریت هوشمند ترافیک و کنترل تطبیقی چراغ‌ها نشان می‌دهد که حرکت از تنظیمات از پیش‌تعیین‌شده به کنترل مبتنی بر داده و هوش مصنوعی یک ضرورت است، اما چالش‌های مدل‌سازی حالت، قابلیت تعمیم و قیود پیاده‌سازی میدانی همچنان پابرجاست [۳-۱]. طی دو دهه گذشته، طیفی از رویکردها به‌کار گرفته شده است. در یک سوی طیف، بهینه‌سازی چرخه‌های ثابت با روش‌های فراابتکاری مانند کلونی زنبور عسل و روش‌های تکاملی، بهبودهایی در شرایط اشباع نشده/ اشباع شده گزارش کرده‌اند، ولی ذاتاً به سناریوهای ایستا و کالیبراسیون‌های پرهزینه وابسته‌اند [۴،۵]. در سوی دیگر، استفاده از حسگرها و اینترنت اشیاء برای پیش‌بینی تقاضا و تعدیل سیگنال‌ها، افق تازه‌ای گشوده اما با محدودیت پوشش حسگر، کیفیت داده و پایداری ارتباطی مواجه است [۶]. در محیط‌های متصل مبتنی بر ارتباطات خودرو- زیرساخت نیز، مرورهای نظام‌مند نشان می‌دهند که فرصت بهره‌گیری از اطلاعات سهموی و تجمعی خودروها برای کاهش وقفه‌ها وجود دارد، هر چند نفوذپذیری فناوری و استانداردهای ارتباطی، مانع مقیاس‌پذیری سریع است [۷]. افزون بر این، روش‌های مبتنی بر معیارهای اولویت/ فشار و الگوریتم‌های توزیع بار در تقاطع‌ها، با اتکا بر اطلاعات لحظه‌ای، تلاش می‌کنند تعارضات حرکتی را با حداقل‌سازی تأخیر حل کنند، اما اغلب به تنظیمات دستی حساس‌اند و در شبکه‌های بزرگ به سادگی تعمیم نمی‌یابند [۸،۹]. در بافت کشورهای درحال توسعه، داده‌های جمع‌سپار چشم‌اندازی کم‌هزینه‌تر فراهم کرده‌اند، ولی نالیمنی آماری و نوفه بالا، قابلیت اتکای کنترل را تحت‌تأثیر قرار می‌دهد [۱۰].

هم‌راستا با این مسیر، یادگیری تقویتی و به‌ویژه گونه‌های عمیق آن، به‌عنوان ابزارهایی برای یادگیری

سیاست‌های کنترل از داده‌های تعاملی مطرح شده‌اند. مرورهای نظام‌مند نشان می‌دهد یادگیری تقویتی می‌تواند معیارهایی مانند میانگین زمان انتظار، طول صف یا تعداد توقف را به‌صورت تطبیقی بهبود دهد؛ با این حال، حساسیت به بازنمایی حالت، کارایی نمونه، پایداری آموزش، و هماهنگی میان تقاطع‌ها همچنان چالش‌برانگیز است [۱-۳]. قیود بی‌درنگ (زمان تصمیم‌گیری در حد ثانیه)، عدم قطعیت حسگر، پاداش‌های تأخیردار و تغییرناپذیری سیاست در برابر تغییر الگوهای تقاضای ساعتی/ فصلی، از مهم‌ترین موانع انتقال این روش‌ها از شبیه‌سازی به میدان است [۲،۳]. از سوی دیگر، چارچوب‌های کنترل فعال و موازی نیز وعده داده‌اند با ساختاردهی بهتر حلقه کنترل و هم‌نهادی تصمیم‌ها، پایداری و پاسخ سریع‌تری ایجاد کنند، هرچند نیازمند طراحی دقیق صورت‌بندی حالت/ اقدام و تعامل ایمن با سیستم‌های موجود هستند [۱۱]. بر این مبنا، مسئله‌ای که این مقاله دنبال می‌کند توسعه و ارزیابی یک طرح کنترل تطبیقی چراغ مبتنی بر هوش مصنوعی است که: (الف) از بازنمایی حالت غنی شامل طول صف، زمان انتظار و شاخص‌های فشار/ اولویت استخراج‌شده از داده‌های میدانی استفاده کند؛ (ب) سیاستی حساس به فاز چراغ اتخاذ کند تا تصمیم‌ها با قیود فیزیکی و ایمنی سازگار باشند و (ج) با داده‌های ناهمگون پایدار بماند. نوآوری‌های اصلی این مقاله به شرح زیر است:

۱. معماری تصمیم‌گیری حساس به فاز با بازنمایی حالت غنی: یک بردار وضعیت مبتنی بر فیزیک ترافیک پیشنهاد می‌کنیم که طول صف، تعداد خودرو، زمان انتظار تجمعی و فشار نرمال‌شده هر حرکت را همراه با حالت فعلی چراغ وارد شبکه می‌کند. تصمیم‌گیری با زیرشبکه‌های تفکیک‌شده برای هر فاز چراغ انجام می‌شود؛ بدین ترتیب تداخل گرایان بین فازها حذف و سیاست‌ها برای هر فاز به‌طور اختصاصی آموخته می‌شوند.

۲. آموزش پایدار با حافظه تجربه تفکیکی و پاداش چندمعیاره: تجربه‌ها برحسب جفت فاز- اقدام در

در [۱۴] محققان یک شبکه عصبی زیست‌الهام شده معرفی کردند که هر فاز چراغ را با نورون برانگیزاننده مدل می‌کند. این سیستم با داده‌های صف خودرو تغذیه می‌شود و قابلیت تغییر پویا یا حتی حذف یک فاز چراغ را در شرایط نامتوازن ترافیک دارد. آزمایش‌ها در SUMO اجرا شدند و نتایج نشان دادند این مدل نسبت به کنترل چرخه‌ای سنتی واکنش سریع‌تر و تطبیق بهتری با تغییرات ناگهانی تقاضای ترافیک دارد. در [۱۵] نویسندگان رویکردی را برای طراحی یک سامانه چراغ راهنمای هوشمند در هنگ‌کنگ ارائه کرده‌اند که بر پایه یادگیری ماشین و تشخیص شیء عمل می‌کند. داده‌های ورودی از طریق ویدئوهای ضبط‌شده در تقاطع‌ها به دست آمده و سیستم قادر است نوع و تعداد وسایل نقلیه و عابران پیاده را شناسایی کند. سپس مدل‌های ترافیکی ساخته و در محیط شبیه‌سازی SUMO مقایسه‌ای میان سامانه‌های موجود و سامانه پیشنهادی انجام شده است. نتایج نشان داد که روش پیشنهاد شده در مدیریت جریان ترافیک عملکرد بهتری دارد و توانسته زمان‌بندی‌ها را به صورت خودکار و تطبیقی تنظیم کند.

در [۱۶] نویسندگان سامانه DeepSIGNAL-ITS را معرفی کردند که ترکیبی از یادگیری عمیق (YOLO برای تشخیص خودروها) و یادگیری تقویتی (Proximal Policy Optimization) است. داده‌ها در لبه پردازش شده و فقط اطلاعات ساختاریافته به فضای ابری ارسال می‌شود. شبیه‌سازی‌ها در SUMO نشان دادند که زمان انتظار وسایل نقلیه به طور متوسط بیش از ۳۰ درصد کاهش یافته و در برخی مناطق حتی تا ۵۰ درصد نیز بهبود یافته است. روش جدید در [۱۷] یک چارچوب کنترل چراغ‌های راهنمایی در مقیاس شبکه‌ای ارائه می‌دهد که بر شناسایی گره‌های بحرانی در شبکه ترافیکی و بهینه‌سازی آنها با یادگیری تقویتی عمیق متمرکز است. ابتدا با مدل‌سازی سه‌لایه‌ای و داده‌های سفر، گره‌هایی که بیشترین تأثیر را بر کارایی شبکه دارند شناسایی می‌شوند. سپس الگوریتم

میانگیرهای جدا ذخیره و با نمونه‌گیری متوازن آموزش داده می‌شوند تا سوگیری داده‌ها نسبت به رخداد‌های پرتکرار رفع شود.

۳. اعتبارسنجی جامع و بهبودهای کمی معنادار نسبت به مراجع: چارچوب روی دو مجموعه داده‌ها عمومی موناکو و هاربین در دو سناریوی شلوغ و خلوت در برابر دو روش مرجع [۳۰] LSTM-PPO و [۳۱] CNN-DQN ارزیابی شد. روش پیشنهادی زمان انتظار را حداکثر ۶۳ درصد و طول صف را تا ۴۴ درصد کاهش می‌دهد.

بخش‌های بعدی این مقاله به شرح زیر سازماندهی شده‌اند: بخش دوم به بررسی کارهای مرتبط می‌پردازد. در بخش سوم روش پیشنهادی مقاله ارائه می‌شود. بخش چهارم نتایج تجربی را ارائه می‌دهد. در بخش پنجم مقاله کارهای آینده و نتیجه‌گیری ارائه می‌شود.

۲. کارهای پیشین

در [۱۲] نویسندگان یک چارچوب مبتنی بر هوش مصنوعی برای بهینه‌سازی چراغ‌های راهنمایی شهری ارائه دادند. روش آنها شامل ترکیب الگوریتم‌های عصبی، ژنتیک و یک الگوریتم ترکیبی بود تا نقاط ضعف رویکردهای منفرد را جبران کند. داده‌های حسگرها و دوربین‌ها وارد این مدل شده و پس از پیش‌پردازش به الگوریتم داده شدند. نتایج نشان داد مدل ترکیبی به سرعت همگرا می‌شود و کاهش حدود ۲ درصد در نرخ اشغال مسیرها به دست می‌آورد. روش جدید در [۱۳] با تمرکز بر بسترهای کشور‌های در حال توسعه طراحی شد. این مقاله از بینایی ماشینی (برای شمارش خودروها توسط Raspberry Pi) و الگوریتم Q-learning بهره می‌برد. حالت‌ها شامل تعداد خودرو و مدت زمان سبز فعلی بودند و پاداش بر اساس طول صف تعریف شد. شبیه‌سازی در SUMO و آزمایش میدانی در یک تقاطع واقعی نشان داد که صف خودروها حدود ۲۹ درصد و زمان انتظار و زمان تلف‌شده حدود ۵۰ درصد کاهش یافتند.

یادگیری تقویتی عمیق روی این گره‌ها اعمال می‌شود. نتایج شبیه‌سازی نشان دادند که این روش تأخیر متوسط و زمان سفر را نسبت به روش‌های پایه به شکل محسوسی کاهش می‌دهد.

در [۱۸] محققان بر کنترل هوشمند چراغ راهنمایی مبتنی بر یادگیری تقویتی عمیق و بینایی ماشین متمرکز شده‌اند. آن‌ها از الگوریتم‌های تشخیص وسیله نقلیه برای استخراج داده‌های ترافیکی استفاده کرده و سپس با مدل‌های یادگیری تقویتی چراغ‌ها را به صورت تطبیقی کنترل کرده‌اند. نتایج شبیه‌سازی نشان داده‌اند که این روش توانسته است زمان سفر و ازدحام را کاهش دهد و در عین حال بهره‌وری سیستم را افزایش دهد. در [۱۹] نویسندگان یک مدل ترکیبی مبتنی بر CNN-LSTM و الگوریتم Proximal Policy Optimization برای کنترل چراغ‌های راهنمایی در محیط چندعاملی ارائه کرده‌اند. ایده اصلی این است که ابتدا وضعیت تقاطع‌ها (شامل فاز چراغ، سطح و تغییرات تراکم و سرعت خودروها) به صورت بردار حالت مدل‌سازی شده و سپس توسط CNN-LSTM برای پیش‌بینی الگوهای مکانی-زمانی ترافیک پردازش شود. محیط آزمایش شامل دو تقاطع مجاور در شبیه‌ساز SUMO بوده و نتایج نشان داد که روش پیشنهادی توانست تا ۹۲ درصد کاهش در زمان انتظار متوسط در سناریوهای ترافیک متوسط ایجاد کند.

در [۲۰] نویسندگان یک روش یادگیری تقویتی چندعاملی مبتنی بر بازی تعاونی را برای کنترل هماهنگ چراغ‌های راهنمایی توزیع‌شده پیشنهاد کرده‌اند. در این مدل هر تقاطع به عنوان یک عامل در نظر گرفته می‌شود که علاوه بر یادگیری راهبرد محلی خود، با عامل‌های مجاور نیز همکاری و هماهنگی دارد. با ترکیب توابع Q محلی در یک شبکه ترکیبی، عملکرد کل شبکه ترافیکی بهینه‌سازی می‌شود. نتایج شبیه‌سازی نشان دادند که این رویکرد نسبت به روش‌های سنتی میانگین سرعت وسایل نقلیه را افزایش داده و زمان تأخیر و انتشار آلاینده‌ها را کاهش

می‌دهد. در [۲۱] روشی با عنوان NSGA-II ارائه شده است. هدف این مدل بهینه‌سازی چندهدفه (کاهش زمان تأخیر متوسط و کاهش تعداد توقف‌ها در هر چرخه) در شبکه تقاطع‌ها است. برای اعتبارسنجی، داده‌های واقعی از شهر بلو هوریزونته (برزیل) استفاده شد و مدل در شبیه‌ساز AIMSUN پیاده‌سازی گردید. نتایج نشان داد که این روش زمان تأخیر خودروها را تقریباً به نصف کاهش داده است. در [۲۲] نویسندگان ترکیب منطق فازی و یادگیری Q عمیق را برای کنترل چراغ‌های راهنمایی معرفی کرده‌اند. در این رویکرد، توالی فاز چراغ‌ها توسط یادگیری Q عمیق تعیین می‌شود و زمان سبز چراغ با استفاده از کنترل‌کننده فازی تنظیم می‌گردد. مزیت اصلی این روش توانایی انطباق هم‌زمان با تغییرات ترافیکی و حفظ پایداری سیستم است. در [۲۳] یک الگوریتم یادگیری Q عمیق مبتنی بر ارتباطات خودرو به خودرو برای کنترل سیگنال‌های ترافیکی در محیط چندعاملی پیشنهاد شده است. در این روش، وضعیت لحظه‌ای ترافیک از طریق ارتباطات خودرو به خودرو دریافت شده و به عامل‌های یادگیری تقویتی ارسال می‌شود. این مدل با در نظر گرفتن اطلاعات شبکه‌ای و تعامل میان تقاطع‌ها، توانسته است در شبیه‌سازی SUMO نسبت به روش‌های سنتی و الگوریتم‌های یادگیری تقویتی پایه، زمان انتظار و طول صف‌ها را به طور محسوسی کاهش دهد.

در [۲۴] نویسندگان یک چارچوب یادگیری تقویتی عمیق چندعاملی بر پایه روش بازیگر-منتقد ارائه کردند. برای حل مشکل ابعاد زیاد و مشاهده ناقص محیط، دو راهکار پیشنهاد شد: اضافه کردن مشاهدات تقاطع‌های همسایه و استفاده از یک ضریب تنزیل فضایی. آزمایش روی شبکه مصنوعی و شبکه واقعی شهر موناکو نشان داد که این روش نسبت به الگوریتم‌های پایه پایداری و کارایی بالاتری دارد. در [۲۵] روشی با عنوان NACRL ارائه شده است که تمرکز آن بر سه هدف اصلی است: بهبود کارایی (کاهش صف و زمان انتظار)، ارتقای ایمنی (محدود کردن تغییرات

کردند. این روش ترکیبی از الگوریتم یادگیری تقویتی عمیق و شبکه LSTM است تا پویایی حرکت خودروها در طول زمان بهتر درک شود. برخلاف روش‌های رایج که فقط به یک عکس لحظه‌ای از تقاطع تکیه می‌کنند، این الگوریتم توالی حرکت خودروها را به‌عنوان اطلاعات ورودی در نظر می‌گیرد. آزمایش‌ها در محیط شبیه‌ساز SUMO انجام شدند و نتایج نشان دادند که DRQN-TLC توانست زمان سفر متوسط خودروها را ۲۳ درصد و زمان انتظار کلی را ۱۰ درصد نسبت به الگوریتم یادگیری تقویتی معمولی کاهش دهد.

بررسی‌های این بخش دو شکاف اصلی را آشکار می‌کند. نخست، بخش بزرگی از کارهای یادگیری تقویتی برای کنترل چراغ یا به یک تقاطع منفرد و شبیه‌ساز محدود شده‌اند یا سیاست‌های آموخته‌شده را بدون توجه صریح به دینامیک فازها تفسیر می‌کنند؛ این پاداش محوری صرف می‌تواند به سیاست‌هایی هم‌ارزش از نظر عدد پاداش ولی نامطلوب در عمل بینجامد. در این مقاله، با تصمیم‌گیری حساس به فاز (حفظ/تغییر) و توازن‌دهی هدفمند به تجربه‌ها در فرآیند یادگیری، این نقص برطرف شده و از سوگیری به نفع اعمال پرتکرار جلوگیری می‌شود. دوم، در سطح شبکه، بسیاری از روش‌ها به طراحی شهودی حالت و پاداش متکی‌اند که پایداری یادگیری و هماهنگی چندتقاطع را تضعیف می‌کند. این مقاله با غنی‌سازی حالت بر پایه شاخص‌های فشاری و عدم تعادل ورودی-خروجی و تعریف پاداشی که به کاهش تأخیر و تخلیه گلوگاه‌ها جهت می‌دهد، صورتی عملی و مقیاس‌پذیر ارائه می‌کند که همسو با رفتار مطلوب شبکه است و در سناریوهای مختلف بهبود معنادار نسبت به خطوط پایه نشان می‌دهد. جدول ۱ خلاصه کارهای گذشته را نشان می‌دهد.

۳. روش پیشنهادی

برای مقابله با پیچیدگی‌های کنترل چراغ‌های راهنمایی در تقاطع‌های شهری، روشی نوآورانه و چندفازی پیشنهاد

مکرر فاز چراغ‌ها) و تقویت هماهنگی در مقیاس شبکه. این الگوریتم بر پایه چارچوب آموزش متمرکز و اجرای غیرمتمرکز توسعه یافته و از سازوکار توجه برای درک بهتر وابستگی‌های مکانی-زمانی میان تقاطع‌ها استفاده کرده است. آزمایش‌ها در یک شبکه مصنوعی ۵×۵ انجام شدند. نتایج نشان داد که NACRL نسبت به الگوریتم‌های مرجع عملکرد بهتری در کاهش تراکم، بهبود هماهنگی و افزایش پایداری الگوریتم دارد.

در [۲۶] نویسندگان یک چارچوب یادگیری تقویتی عمیق چندعاملی بر پایه روش بازیگر-منتقد معرفی کرده‌اند. هدف اصلی، کنترل هماهنگ چراغ‌های راهنمایی در شبکه‌های ترافیکی پیچیده است. برای حل مشکل ابعاد زیاد و مشاهده ناقص محیط، دو راهکار به‌کار رفته: (۱) استفاده از اطلاعات تقاطع‌های همسایه و (۲) اعمال ضریب تنزیل فضایی برای مدل‌سازی تأثیر متقابل تقاطع‌ها. این روش با ساختار آموزش متمرکز و اجرای غیرمتمرکز پیاده‌سازی شد و در محیط شبیه‌ساز SUMO آزمایش گردید. نتایج نشان دادند که روش پیشنهادی در مقایسه با الگوریتم‌های پایه، پایداری و همگرایی بهتری داشته و توانسته است زمان انتظار و تراکم خودروها را کاهش دهد. در [۲۷] روشی با عنوان NACRL ارائه شده است که تمرکز آن بر سه هدف اصلی است: بهبود کارایی (کاهش صف و تأخیر)، ارتقای ایمنی (کاهش تغییرات مکرر فاز چراغ‌ها) و افزایش هماهنگی در سطح شبکه. این الگوریتم با ترکیب سازوکار توجه برای درک بهتر ارتباطات مکانی-زمانی و شبکه LSTM برای مدل‌سازی پویایی زمانی طراحی شده است. NACRL با ساختار آموزش متمرکز و اجرای غیرمتمرکز توسعه یافته و روی دو بستر ارزیابی شد: یک شبکه مصنوعی ۵×۵ و یک شبکه واقعی از شهر موناکو. نتایج نشان داد که این روش نسبت به الگوریتم‌های مرجع توانسته است تأخیر، تراکم و نرخ تصادف را کاهش دهد و هماهنگی میان تقاطع‌ها را بهبود بخشد.

در [۲۸] نویسندگان الگوریتمی به نام DRQN-TLC معرفی

جدول ۱: مزایا و معایب روش‌های مختلف

معایب	مزایا	مرجع
ارزیابی فقط روی یک مقطع ترافیکی	کاهش نرخ اشغال حدود ۲ درصد با مدل ترکیبی	[۱۲]
نبود آزمایش میدانی و اتکا به شبیه‌سازی	کاهش ۵۰ درصد زمان انتظار و زمان تلف‌شده	[۱۳]
آزمون فقط روی یک تقاطع و مفروضات ساده‌شده	واکنش سریع به تغییرات ناگهانی تقاضا	[۱۴]
محدودیت شبیه‌ساز و هزینه بالای حسگرها	بهبود عملکرد ITLS نسبت به FCTL در شبیه‌سازی	[۱۵]
نیاز به زیرساخت پرهزینه	کاهش بیش از ۳۰ درصد زمان انتظار و کاهش آلاینده‌ها	[۱۶]
محدودیت در تعمیم‌پذیری به کل شبکه‌های بسیار بزرگ	شناسایی و بهینه‌سازی گره‌های بحرانی	[۱۷]
اتکا به دقت الگوریتم‌های بینایی در شرایط آب‌وهوایی	کاهش زمان سفر و افزایش بهره‌وری	[۱۸]
ارزیابی محدود در دو تقاطع	کاهش چشمگیر زمان انتظار	[۱۹]
پیچیدگی بالا در هماهنگی عامل‌ها در مقیاس بزرگ	بهبود سرعت، کاهش تأخیر و آلاینده‌ها	[۲۰]
وابستگی به داده و شرایط خاص شهر بلو هوریزونته	کاهش ۵۰ درصد زمان تأخیر در داده واقعی	[۲۱]
نیاز به تنظیم دقیق پارامترهای فازی و شبکه عصبی	پایداری و کاهش آلودگی	[۲۲]
نیازمند زیرساخت ارتباطی پیشرفته	کاهش زمان انتظار و صف	[۲۳]
نیازمند ارتباط بین عامل‌ها و محاسبات سنگین	بهبود پایداری و کاهش زمان انتظار در شبکه‌های بزرگ	[۲۴]
پیچیدگی محاسباتی بالا و وابستگی به داده‌های دقیق	بهبود هم‌زمان کارایی، ایمنی و هماهنگی در شبکه تقاطع‌ها	[۲۵]
نیاز به ارتباط بین عامل‌ها و محاسبات سنگین	بهبود پایداری و کاهش زمان انتظار در شبکه‌های بزرگ	[۲۶]
پیچیدگی محاسباتی بالا به دلیل استفاده هم‌زمان از مکانیزم توجه و LSTM	کاهش تراکم و زمان انتظار در شبکه	[۲۷]
وابستگی شدید به کیفیت داده تصویری	کاهش ۲۳ درصد زمان سفر و ۱۰ درصد زمان انتظار	[۲۸]

اساس ترکیب‌های مختلف حالت چراغ- اقدام ذخیره شده و در زمان یادگیری، نمونه‌گیری متوازن از این حافظه‌ها انجام می‌شود. این سازوکار از تمرکز بیش از حد روی برخی حالت‌های پرتکرار جلوگیری کرده و موجب افزایش پایداری یادگیری می‌شود. روندنمای روش پیشنهادی در شکل ۱ نشان داده شده است.

۳-۱. فاز اول: مدل‌سازی وضعیت

در مرحله نخست از روش پیشنهادی، مسئله کلیدی تعریف و بازنمایی وضعیت محیط ترافیکی است. هدف این فاز آن است که تمامی اطلاعات مؤثر بر تصمیم‌گیری عامل یادگیرنده به صورت کمی و دقیق به برداری از ویژگی‌ها تبدیل شود تا هم جنبه‌های فیزیکی جریان ترافیک پوشش داده شود و هم امکان استخراج الگوهای پیچیده توسط شبکه یادگیری فراهم گردد. وضعیت ترافیک در یک تقاطع

می‌کنیم که با ترکیب دیدگاه‌های نظری حمل‌ونقل و یادگیری تقویتی عمیق طراحی شده است. این روش با هدف کاهش زمان سفر، بهبود توزیع ترافیک و جلوگیری از تراکم در شبکه شهری توسعه یافته و از سه فاز اصلی تشکیل شده است. در فاز اول، وضعیت محیط ترافیکی با استفاده از شاخص‌های کلیدی نظیر طول صف، تعداد وسایل نقلیه، زمان انتظار و فشار ترافیکی مدل‌سازی می‌شود؛ به‌طوری‌که بیان ریاضی فشار هر حرکت به صورت اختلاف نرمال شده تعداد خودروهای ورودی و خروجی تعریف می‌گردد. در فاز دوم، تصمیم‌گیری عامل هوشمند به‌گونه‌ای انجام می‌شود که نه تنها وابسته به وضعیت محیط، بلکه حساس به حالت فعلی چراغ راهنمایی نیز باشد. بدین ترتیب، برای هر حالت چراغ مسیر تصمیم‌گیری مجزا طراحی شده و عامل اقدام بهینه (تغییر یا عدم تغییر حالت چراغ) را انتخاب می‌کند. در فاز سوم، فرآیند آموزش تقویت می‌شود؛ به این صورت که تجربیات عامل در حافظه‌های تفکیکی بر

آن، تعداد کل خودروهای حاضر در هر خط، با نماد V_i به عنوان متغیر مکمل در نظر گرفته می شود. این متغیر تراکم واقعی خط را نشان داده و مبنای محاسبات ظرفیت و اشباع مسیر است. علاوه بر این، عامل باید به زمان انتظار خودروها نیز دسترسی داشته باشد.

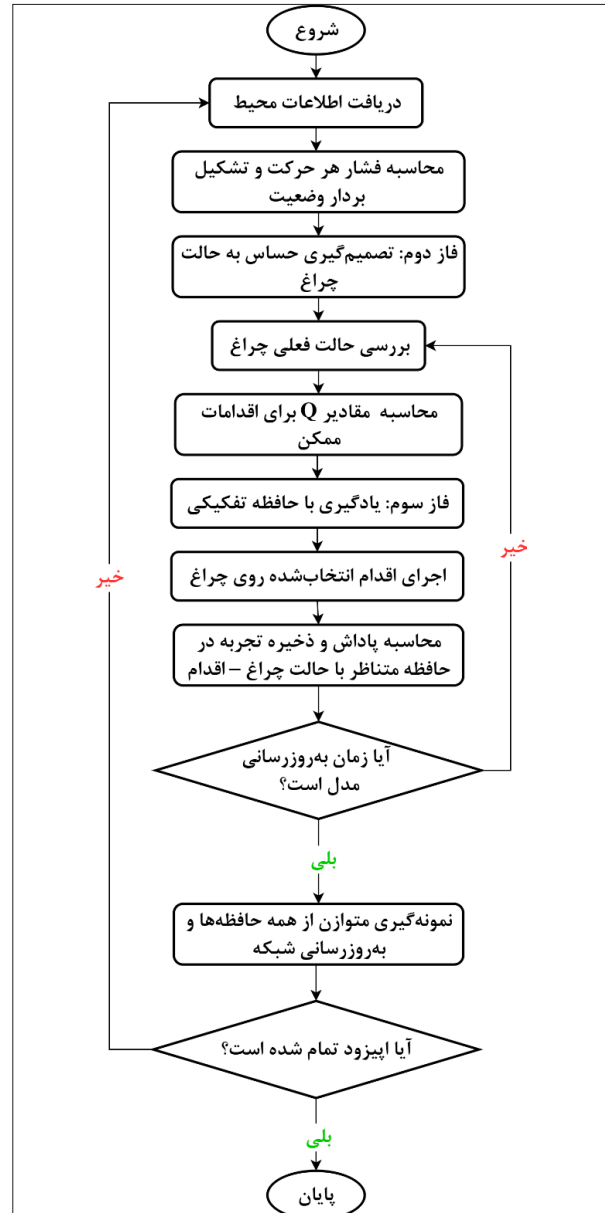
برای هر خط ورودی i ، زمان انتظار تجمعی W_i تعریف می شود که برابر است با مجموع زمان هایی که خودروهای حاضر در آن خط تا لحظه فعلی منتظر مانده اند. برای خودرو j در زمان t ، زمان انتظار به صورت بازگشتی تعریف می شود:

$$W_j(t) = \begin{cases} W_j(t-1) + 1 & \text{speed} < 0.1 \text{ m/s} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

این فرمول نشان می دهد که هر بار که خودرو شروع به حرکت کند، زمان انتظارش صفر می شود و فقط در شرایط توقف افزایش می یابد. بنابراین، مجموع زمان های انتظار همه خودروهای یک خط، شاخصی روشن از میزان ازدحام و ناکارایی مدیریت فعلی چراغ راهنمایی به دست می دهد. یکی دیگر از مفاهیم مهم که در این فاز به وضعیت اضافه می شود، فشار ترافیکی است. فشار برای هر حرکت (l, m) به صورت اختلاف بین تراکم خودروهای خط ورودی l و خط خروجی متناظر m تعریف می شود. تراکم یک خط به صورت نسبت تعداد خودروهای حاضر به ظرفیت بیشینه آن خط بیان می شود. اگر $x(l)$ تعداد خودروهای حاضر در خط l و $x_{\max}(l)$ ظرفیت بیشینه آن خط باشد، فشار حرکت از رابطه زیر به دست می آید:

$$w(l, m) = \frac{x(l)}{x_{\max}(l)} - \frac{x(m)}{x_{\max}(m)} \quad (2)$$

این تعریف نشان می دهد که فشار مثبت زمانی رخ می دهد که تعداد خودروهای ورودی نسبت به خروجی بیشتر باشد و نیاز به تخلیه آن مسیر وجود دارد، در حالی



شکل ۱. روندنمای روش پیشنهادی.

شهری را می توان مجموعه ای از پارامترها دانست که نشان دهنده تراکم وسایل نقلیه، توزیع آنها در مسیرها و شرایط سیگنال های کنترلی است. یکی از مؤلفه های اصلی در این بازنمایی، طول صف وسایل نقلیه در هر خط ورودی است. اگر l_1 نشان دهنده طول صف در خط ورودی l باشد، این متغیر بیان می کند که چه تعداد خودرو در انتظار عبور از تقاطع هستند. طول صف نه تنها شاخصی از شدت تقاضای ترافیکی در یک مسیر است، بلکه در محاسبه تأخیر و فشار ترافیکی نیز نقشی بنیادی دارد. همراه با

۳-۲. فاز دوم: تصمیم‌گیری حساس به حالت چراغ

پس از مدل‌سازی دقیق وضعیت محیط، در فاز دوم عامل یادگیرنده باید تصمیم بگیرد که آیا چراغ راهنمایی در حالت فعلی باقی بماند یا به حالت بعدی تغییر کند. ویژگی اصلی این فاز، طراحی سازوکاری است که تصمیم‌گیری عامل را وابسته به حالت فعلی چراغ کند. در واقع، یک تصمیم یکسان (مثلاً تغییر چراغ) ممکن است در دو حالت مختلف چراغ نتایج کاملاً متفاوتی ایجاد کند؛ بنابراین، نیاز است که مسیر تصمیم‌گیری برای هر حالت چراغ مجزا در نظر گرفته شود. برای این منظور، مدل یادگیری تقویتی یک شبکه ارزش تعریف می‌کند که ورودی آن وضعیت محیط s_t و خروجی آن ارزش اقدام‌های ممکن $Q(s_t, a)$ است.

اقدام‌ها شامل دو انتخاب هستند: $a \in \{0, 1\}$ که در آن $a = 0$ به معنای حفظ حالت فعلی چراغ (H_c) و $a = 1$ به معنای تغییر به حالت بعدی (H_{c+1}) است. برای هر حالت چراغ، یک دروازه تصمیم‌گیری طراحی می‌شود تا مسیر محاسباتی شبکه به صورت تفکیک‌شده عمل کند. به عبارت دیگر، اگر چراغ در حالت $H_c = 0$ باشد، تنها زیرشبکه متناظر فعال می‌شود و مقدار $Q(s, a)$ برای اقدامات ممکن در همین حالت محاسبه می‌شود و اگر چراغ در حالت $H_c = 1$ باشد، مسیر دیگری از شبکه فعال می‌گردد. این طراحی از تداخل تصمیم‌گیری میان حالت‌های متفاوت چراغ جلوگیری کرده و به عامل امکان می‌دهد تصمیم‌های تخصصی‌تر و دقیق‌تری برای هر حالت اتخاذ کند. مقدار Q برای هر جفت وضعیت-اقدام بر اساس معادله بلمن به‌روزرسانی می‌شود:

$$Q(s_t, a_t) \leftarrow r_t + \gamma a' \max_{a'} Q(s_{t+1}, a') \quad (5)$$

که در آن r_t پاداش دریافت شده در زمان t ، $\gamma \in [0, 1]$ ضریب تنزیل برای آینده، s_{t+1} وضعیت جدید پس از اجرای اقدام a_t و $\max_{a'} (Q(s_{t+1}, a'))$ بیشترین ارزش اقدام ممکن در وضعیت آینده است. انتخاب اقدام در هر گام بر اساس

که فشار منفی نشان‌دهنده اشباع خروجی و احتمال ایجاد انسداد است. فشار کل یک تقاطع به‌عنوان مجموع قدرمطلق فشارهای همه حرکات تعریف می‌شود:

$$P_i = \left| \sum_{(l,m) \in i} W(l, m) \right| \quad (3)$$

مقدار P_i شاخصی از عدم تعادل توزیع ترافیک میان خطوط ورودی و خروجی است و هر چه بزرگ‌تر باشد، نشان‌دهنده وضعیت ناپایدارتر و نیاز فوری‌تر به مداخله کنترلی است- علاوه بر این متغیرها، حالت فعلی چراغ راهنمایی نیز بخشی از وضعیت محسوب می‌شود. حالت چراغ، که آن را با H_c نمایش می‌دهیم، بیان می‌کند که کدام حرکت‌ها در لحظه فعلی مجاز (سبز) یا ممنوع (قرمز) هستند. برای هر حرکت (l, m) ، اگر چراغ سبز باشد، مقدار سیگنال $a(l, m) = 1$ و در غیر این صورت $a(l, m) = 0$ خواهد بود. این متغیر به عامل یادگیرنده امکان می‌دهد تصمیم‌های خود را بر اساس ترکیب حالت چراغ و شرایط ترافیکی اتخاذ کند. در کنار داده‌های عددی، می‌توان از بازنمایی تصویری نیز برای توصیف موقعیت مکانی خودروها در تقاطع بهره گرفت. اگر ماتریسی با نام M تعریف شود که موقعیت خودروها را در شبکه سلولی خطوط ورودی و خروجی ثبت می‌کند، این ماتریس به‌عنوان ورودی شبکه‌های عصبی هم‌آمیختگی استفاده شده و به عامل کمک می‌کند الگوهای مکانی ازدحام را شناسایی کند. در نهایت، وضعیت کلی در زمان t به صورت‌برداری از همه متغیرهای معرفی‌شده تعریف می‌شود:

$$s_t = \{L_t, V_t, W_t, w(l, m), H_c, M\} \quad (4)$$

این بازنمایی جامع باعث می‌شود که عامل هوشمند، در هر گام زمانی، هم از شاخص‌های کلاسیک مهندسی ترافیک مانند فشار و تراکم استفاده کند و هم از داده‌های رفتاری مانند زمان انتظار تجمعی و موقعیت مکانی خودروها بهره گیرد.

را شکل می‌دهد و خروجی آن یک اقدام مشخص است که مستقیماً به چراغ اعمال می‌شود. این اقدام یا حفظ حالت فعلی است یا تغییر آن و بلافاصله اثر آن بر وضعیت جدید محیط و پاداش حاصل در فاز بعدی محاسبه خواهد شد.

۳-۳. فاز سوم: یادگیری با حافظه تفکیکی

پس از آن که عامل در فاز دوم اقدام مناسب را انتخاب و اجرا کرد و پاداش مربوطه محاسبه شد، فرآیند یادگیری آغاز می‌شود. هدف این فاز به‌روزرسانی شبکه تصمیم‌گیری است تا عامل بتواند در گام‌های آینده تصمیم‌های بهینه‌تری بگیرد. یکی از چالش‌های اصلی در یادگیری تقویتی، عدم توازن در داده‌های آموزشی است؛ زیرا برخی ترکیب‌های حالت چراغ-اقدام بسیار بیشتر از سایر ترکیب‌ها رخ می‌دهند. برای نمونه، باقی ماندن در یک حالت چراغ (مثلاً $a = 0, H_c = 0$) ممکن است بسیار پرتکرار باشد، در حالی که تغییر چراغ در حالت خاصی (مثلاً $a = 1, H_c = 1$) کمتر رخ دهد. اگر همه این داده‌ها در یک حافظه مشترک ذخیره شوند، عامل بیشتر روی داده‌های پرتکرار آموزش دیده و در وضعیت‌های کم‌رخداد عملکرد ضعیفی خواهد داشت. برای رفع این مشکل، در روش پیشنهادی یک حافظه تفکیکی طراحی می‌شود. به این صورت که حافظه تجربه به چند زیرحافظه تقسیم می‌شود و هر کدام مختص یک ترکیب حالت چراغ-اقدام است. به‌طور رسمی، اگر $H_c \in \{0, 1\}$ حالت چراغ و $a \in \{0, 1\}$ اقدام انتخاب شده باشد، مجموعه حافظه‌ها به صورت زیر تعریف می‌شود:

$$D = \{D(H_c, a) | H_c \in \{0, 1\}, a \in \{0, 1\}\} \quad (8)$$

هر تجربه به‌صورت چهارگانه زیر ذخیره می‌شود:

$$e_t = (s_t, a_t, R_t, s_{t+1}) \quad (9)$$

تجربه e_t در حافظه مربوط به جفت (H_c, a) ذخیره

سیاست ϵ -greedy انجام می‌شود. در این سیاست، عامل با احتمال ϵ یک اقدام تصادفی (برای کاوش فضای حالت‌ها) و با احتمال $1 - \epsilon$ اقدام متناظر با بیشترین مقدار Q را انتخاب می‌کند:

$$a_t = \begin{cases} \text{random action} & \text{with probability } \delta \\ \arg \max_a Q(s_t, a) & \text{with probability } 1 - \delta \end{cases} \quad (6)$$

این رویکرد باعث می‌شود عامل همواره بین کاوش برای یافتن تصمیم‌های جدید و بهره‌برداری از تصمیم‌های شناخته‌شده بهینه تعادل برقرار کند. پس از انتخاب و اجرای اقدام، کیفیت آن باید سنجیده شود. به این منظور، یک تابع پاداش تعریف می‌شود که میزان بهبود یا وخامت وضعیت ترافیک را نشان می‌دهد. در روش پیشنهادی، تابع پاداش بر پایه ترکیبی از فشار ترافیکی، طول صف‌ها، زمان انتظار و تعداد خودروهای عبور کرده تعریف شده است:

$$R_t = - \sum_{(l,m)} W(l,m) - \alpha_1 \sum_i L_i - \alpha_2 \sum_i W_i + \alpha_3 N_t \quad (7)$$

در رابطه (۷)، W_i زمان انتظار تجمعی خودروها در خط i ، N_t تعداد خودروهایی که در گام زمانی t موفق به عبور از تقاطع شده‌اند و ضرایب $\alpha_1, \alpha_2, \alpha_3$ پارامترهایی هستند که اهمیت نسبی هر معیار را مشخص می‌کنند. ویژگی مهم این فاز آن است که تصمیم‌گیری عامل تنها تابعی از پارامترهای ترافیکی (مانند طول صف یا فشار) نیست، بلکه به حالت فعلی چراغ نیز حساس است. برای مثال، اگر چراغ در حالت $H_c = 0$ باشد (فرضاً سبز برای شرق-غرب)، عامل ممکن است تصمیم بگیرد به دلیل فشار زیاد در شمال-جنوب به تغییر چراغ اقدام کند؛ در حالی که اگر همان شرایط صف و فشار در حالت $H_c = 1$ رخ دهد، تصمیم متفاوتی به دست می‌آید. این حساسیت به حالت چراغ، الگوریتم را از بسیاری از مدل‌های ساده گذشته متمایز می‌کند. در نهایت، فاز دوم فرآیند تصمیم‌گیری عامل

جدول ۲: پارامترهای شبیه‌سازی.

پارامتر	مقدار
تابع فعال‌سازی	ReLU
نرخ یادگیری	۰/۰۰۱
مقدار اولیه ϵ	۱
مقدار نهایی ϵ	۰/۰۱
نرخ کاهش ϵ	۰/۹
تابع خطا	MSE
بهینه‌ساز	Adam

مناسبتی داشته باشد.

۴. آزمایش‌ها

برای ارزیابی کارایی روش پیشنهادی، از دو مجموعه داده عمومی استفاده شده است که در مرجع [۲۹] معرفی شده‌اند. این مجموعه داده‌ها به ترتیب مربوط به شهر موناکو در کشور فرانسه و شهر هاربین در کشور چین هستند. انتخاب این داده‌ها به دلیل پوشش شرایط ترافیکی متنوع (از جمله تراکم بالا در مناطق پرتردد و تراکم متوسط در شرایط عادی) صورت گرفته است تا عملکرد الگوریتم پیشنهادی در محیط‌های واقعی و با الگوهای رفتاری متفاوت وسایل نقلیه مورد بررسی قرار گیرد.

داده‌های شهر موناکو یک سناریوی شهری باز برای SUMO است که کل ناحیه شهری و پیرامونی موناکو (حدود ۷۰-۷۴ کیلومترمربع) را مدل می‌کند؛ شبکه چندلایه شامل خیابان‌ها، ۷۴ تونل، ۹۶ پل، حدود ۱۵ هزار ساختمان (همراه با ارتفاع)، تقسیم‌بندی ناحیه‌ای و لایه حمل‌ونقل عمومی با تقریباً ۱۷۶ ایستگاه اتوبوس در ۲۴ مسیر و ۱۲ ایستگاه قطار است. داده‌های شهر هاربین بر بستر SUMO یک شبکه شهری بزرگ‌مقیاس را با ابعاد تقریبی ۶۰۰×۷۰۰ مترمربع و ۳۴۱ تقاطع سیگنال‌دار مدل می‌کند. پیکربندی آزمایش شامل گام‌های زمانی ثابت (حلقه‌های ۳۶۰۰ قدم)، استقرار چندعامل (۴ عامل با پوشش شعاعی حدود ۵۰۰ متر) و ثبت شاخص‌های ارزیابی مانند زمان انتظار است. پیاده‌سازی روش پیشنهادی با استفاده

می‌شود. در هنگام یادگیری، از هر حافظه به تعداد مساوی نمونه‌گیری می‌شود تا تعادل برقرار گردد. فرض کنیم ظرفیت Min batch برابر B باشد و تعداد حافظه‌ها برابر K. در این صورت از هر حافظه تعداد $\frac{B}{K}$ نمونه انتخاب می‌شود. به این ترتیب، عامل تجربه کافی از تمام حالت‌ها و اقدامات خواهد داشت و تعادل در یادگیری حفظ می‌شود. به‌روزرسانی شبکه بر اساس کمینه‌سازی خطای میانگین مربعات میان خروجی شبکه و مقدار هدف صورت می‌گیرد. تابع خطا به صورت زیر تعریف می‌شود:

$$L(\theta) = E_{(s,a,R,s') \sim U(D)} [(y - Q(s,a;\theta))^2] \quad (10)$$

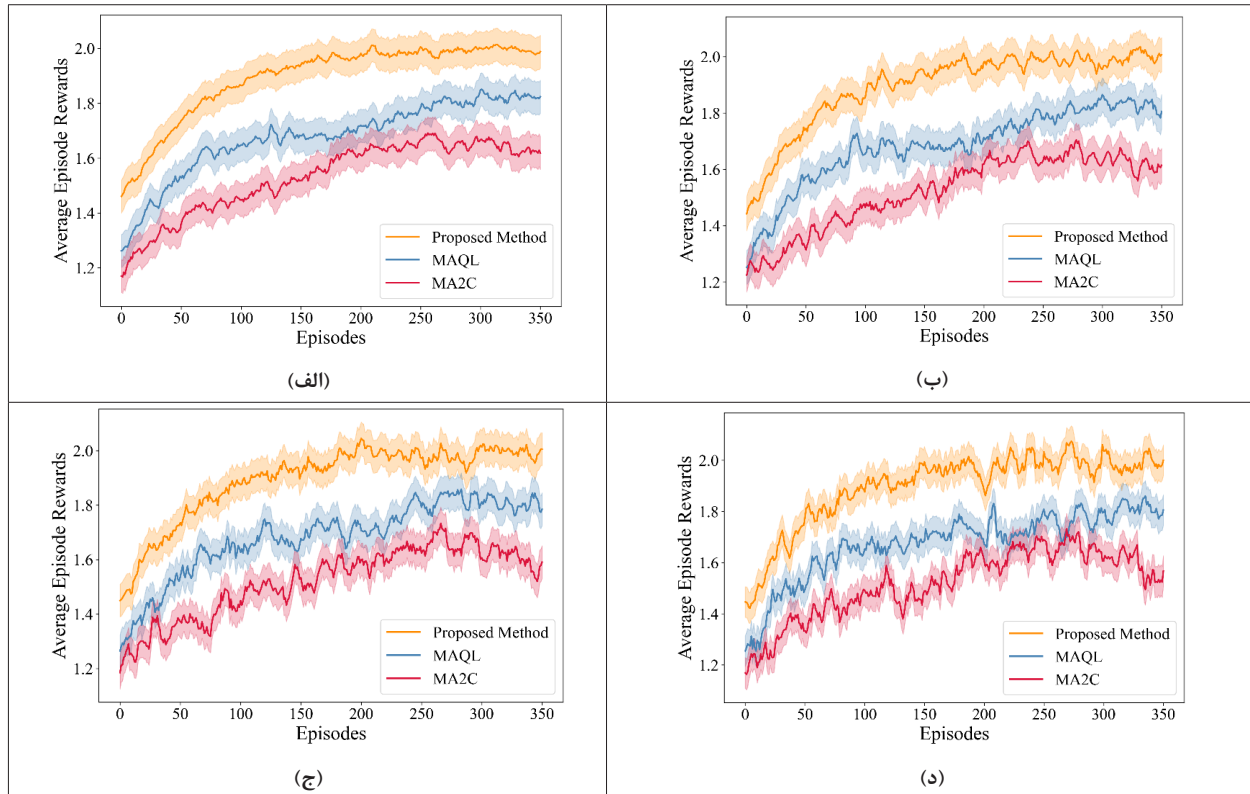
که در آن θ پارامترهای شبکه $Q, U(D)$ عمل انتخاب تصادفی از نمونه‌های موجود در حافظه‌ها و y مقدار هدف است که از معادله بلمن به دست می‌آید:

$$y = R + \gamma a' \max Q(s', a'; \theta^-) \quad (11)$$

که در آن θ^- پارامترهای شبکه هدف است که به‌طور دوره‌ای به‌روزرسانی می‌شوند تا از نوسانات یادگیری جلوگیری شود. به این ترتیب، به‌روزرسانی پارامترهای شبکه Q بر اساس گرادینان کاهش انجام می‌شود:

$$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta) \quad (12)$$

که در آن η نشانگر نرخ یادگیری است. ویژگی مهم استفاده از حافظه تفکیکی آن است که از سوگیری داده‌ها جلوگیری می‌کند. به عبارت دیگر، حتی اگر در طول شبیه‌سازی داده‌های مربوط به اقدام $\mathbf{a} = \mathbf{0}$ (حفظ حالت چراغ) بیشتر تولید شود، در فرآیند آموزش نمونه‌های مربوط به اقدام $\mathbf{a} = \mathbf{1}$ (تغییر حالت چراغ) نیز به میزان مساوی در نظر گرفته می‌شود. این طراحی موجب می‌شود عامل در مواجهه با شرایط نادر اما بحرانی نیز عملکرد



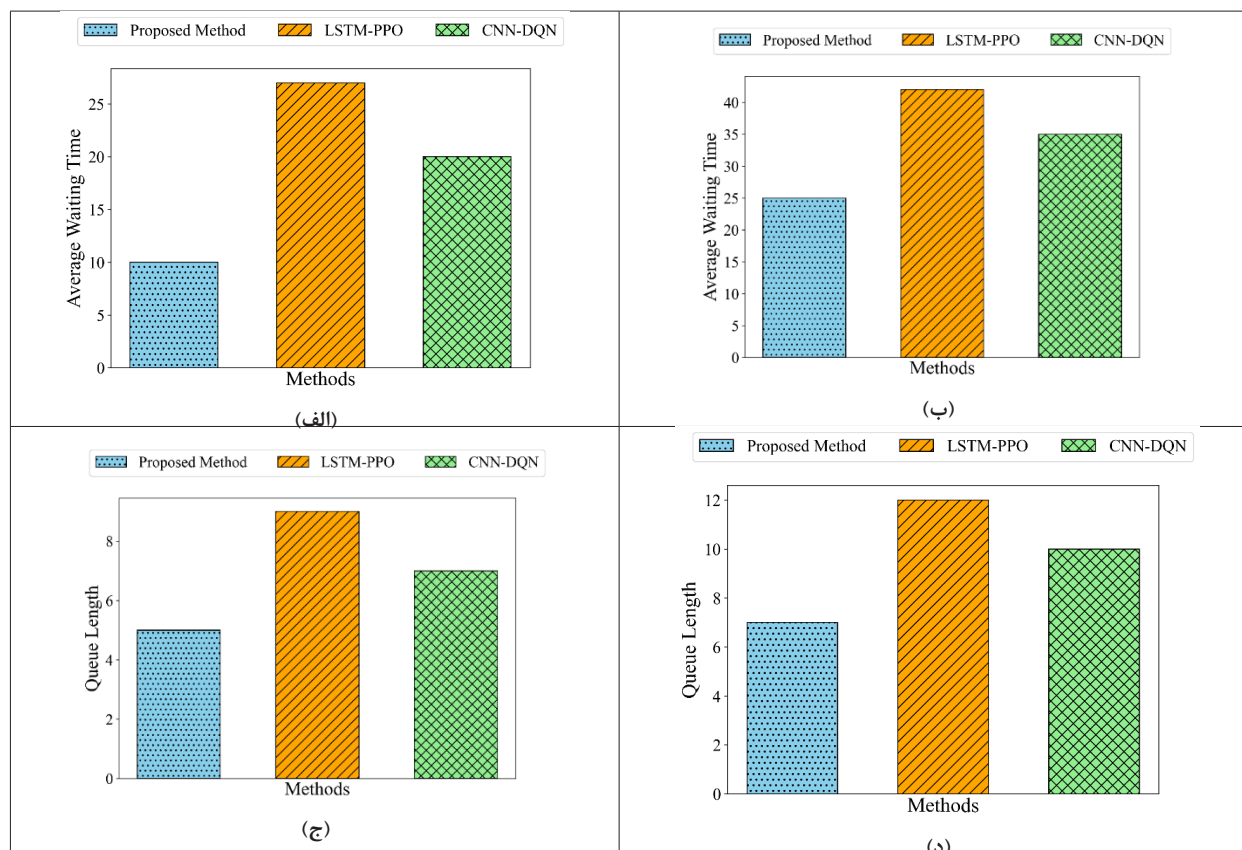
شکل ۲. مقایسه پاداش میانگین اپیزودها برای روش پیشنهادی در برابر دو روش مرجع در چهار نرخ جست و جو (الف) $\epsilon=0.9$ ، (ب) $\epsilon=0.7$ ، (ج) $\epsilon=0.5$ و (د) $\epsilon=0.3$.

این، تصمیم‌گیری حساس به حالت چراغ موجب می‌شود به جای یک مسیر مشترک، زیرشبکه متناظر با هر فاز چراغ فعال و به روزرسانی شود؛ این تفکیک تداخل گرایان‌ها را کاهش می‌دهد و نوسان یادگیری را کم می‌کند، چیزی که در باریک‌تر بودن ناحیه اطمینان شکل‌ها نیز دیده می‌شود. سوم، تابع پاداش چندمعیاره با ترکیب فشار، طول صف، زمان انتظار و تعداد عبوری‌ها سیگنال یادگیری معنادارتری تولید می‌کند و عامل را به سمت کاهش ازدحام واقعی هدایت می‌نماید. نهایتاً، حافظه تجربه تفکیکی برای هر جفت حالت چراغ-اقدام از سوگیری به سمت رخدادهای پرتکرار جلوگیری می‌کند و با نمونه‌برداری متوازن پوشش یکنواختی از حالات نادر اما بحرانی فراهم می‌سازد؛ این سازوکار به همگرایی باثبات‌تر و میانگین پاداش بزرگ‌تر نسبت به دو روش مرجع [۲۶] MA2C و [۲۹] MAQL می‌انجامد.

شکل ۳ نتایج شهر موناکو را در دو معیار میانگین

از زبان برنامه‌نویسی پایتون (با استفاده از کتابخانه‌های عددی و یادگیری عمیق) انجام شد و تمامی آزمایش‌ها در شبیه‌ساز SUMO از طریق رابط TraCI اجرا گردید. مقادیر پارامترهای شبیه‌سازی الگوریتم در جدول ۲ نشان داده شده‌اند.

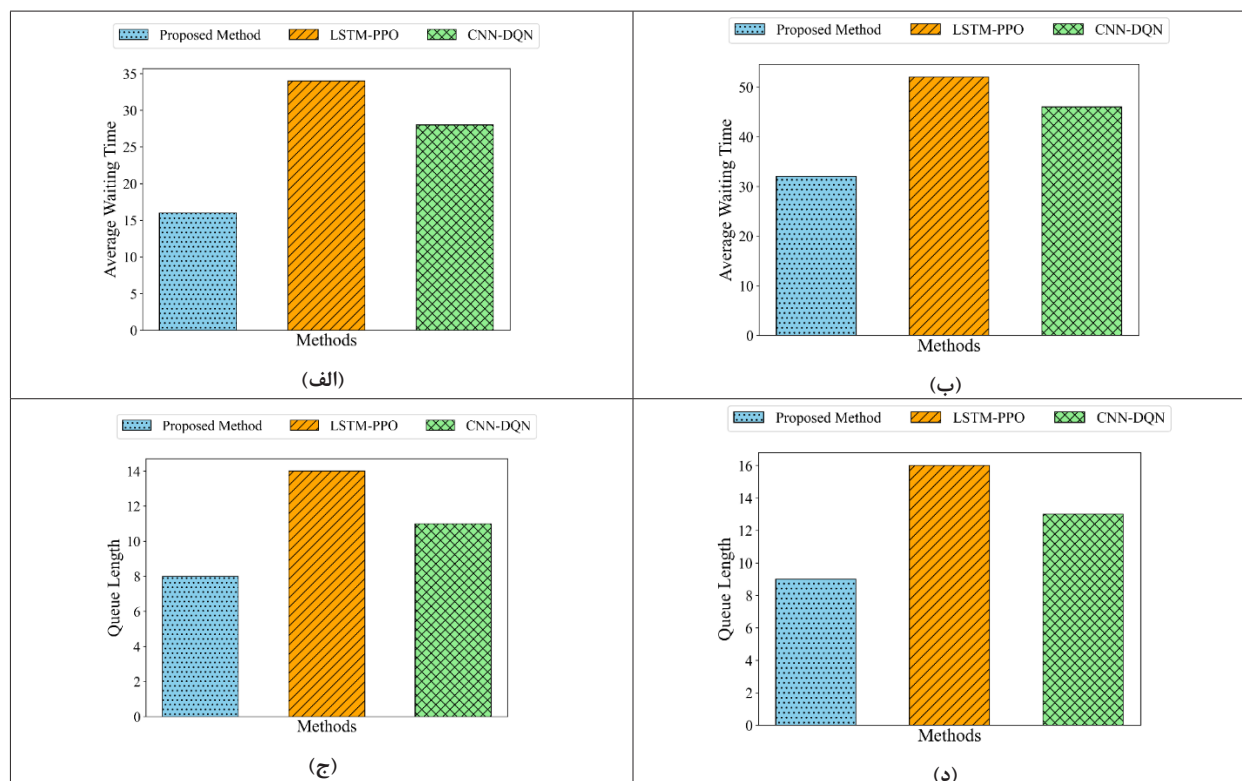
شکل ۲ (الف-د) نشان می‌دهد روش پیشنهادی در تمام نرخ‌های جست و جو (به‌ویژه در $\epsilon=0.9$) سریع‌تر به پاداش‌های بالاتر و رفتار پایدارتر می‌رسد و از دو روش مرجع [۲۶] MA2C و [۲۹] MAQL پیشی می‌گیرد. دلیل این برتری، نخست بازنمایی وضعیت غنی و منطبق با واقعیت ترافیک است؛ علاوه بر طول صف و تعداد خودرو، زمان انتظار تجمعی و فشار ترافیکی نرمال‌شده‌ی ورودی-خروجی نیز در وضعیت وارد می‌شود و در کنار آن، حالت فعلی چراغ صراحتاً مدل می‌گردد؛ در نتیجه عامل به شاخص‌های علی ازدحام و عدم تعادل شبکه دسترسی مستقیم دارد و سیاستی عمومی‌تر می‌آموزد. علاوه بر



شکل ۳. مقایسه عملکرد روش پیشنهادی با دو مرجع دیگر روی مجموعه داده‌ها شهر موناکو در دو سناریوی ترافیکی. (الف) میانگین زمان انتظار خلوت، (ب) میانگین زمان انتظار شلوغ، (ج) طول صف خلوت و (د) طول صف شلوغ.

تراکم، جریان عبوری یکنواخت‌تری تولید می‌کند و توانایی بیشتری در جلوگیری از تجمع خودروها دارد. همچنین مشاهده می‌شود بهبود نسبی در غیرپیک بیشتر از پیک است که بیانگر آن است که روش پیشنهادی در شرایط با تغییرپذیری بالاتر جریان ورودی، سیاست‌های کارتری اتخاذ می‌کند و از اتلاف ظرفیت سبز می‌کاهد؛ در پیک نیز با وجود فشار ترافیکی زیاد، همچنان زمان انتظار و صف را کمتر از دو مرجع نگه می‌دارد. در مجموع، در هر دو معیار و هر دو سناریو، روش پیشنهادی به‌طور پایدار از LSTM-PPO و CNN-DQN بهتر عمل کرده و کاهش معنی‌داری در تأخیر و طول صف ایجاد می‌کند؛ بنابراین برای کنترل سیگنال در تقاطع‌های پرتردد موناکو گزینه مؤثرتری است. شکل ۴ نتایج شهر موناکو را در دو معیار میانگین زمان انتظار و طول صف برای دو سناریوی پیک (شلوغ) و غیرپیک (خلوت) و سه روش، روش پیشنهادی، [۳۰]

زمان انتظار و طول صف برای دو سناریوی پیک (شلوغ) و غیرپیک (خلوت) و سه روش، روش پیشنهادی، [۳۰] و LSTM-PPO و CNN-DQN [۳۱] نشان می‌دهد. در سناریوی پیک، روش پیشنهادی زمان انتظار را به ۲۵ واحد کاهش می‌دهد، در حالی که LSTM-PPO و CNN-DQN به ترتیب ۴۲ و ۳۵ واحد هستند؛ یعنی کاهش ۴۰ درصد نسبت به [۳۰] و ۲۹ درصد نسبت به [۳۱]. در سناریوی غیرپیک، شکاف پررنگ‌تر است: روش پیشنهادی ۱۰ واحد در برابر ۲۷ (LSTM-PPO) و ۲۰ (CNN-DQN) ثبت می‌کند که معادل کاهش ۶۳ درصد و ۵۰ درصد است. الگوی مشابهی در طول صف دیده می‌شود: در پیک، طول صف روش پیشنهادی ۷ در برابر ۱۲ و ۱۰ است (کاهش ۴۲ درصد و ۳۰ درصد نسبت به [۳۰] و [۳۱]) و در غیرپیک، ۵ در برابر ۹ و ۷ (کاهش ۴۴ درصد و ۲۹ درصد). این اختلاف‌ها نشان می‌دهد روش پیشنهادی در هر دو سطح



شکل ۴. مقایسه عملکرد روش پیشنهادی با دو مرجع دیگر روی مجموعه داده‌ها شهر هاربین در دو سناریوی ترافیکی. (الف) میانگین زمان انتظار خلوت، (ب) میانگین زمان انتظار شلوغ، (ج) طول صف خلوت و (د) طول صف شلوغ.

۱۴ و ۱۱ گزارش می‌شود که معادل کاهش ۴۳ درصد و ۲۷ درصد است. برتری بیشتر در غیرپیک بیانگر آن است که روش پیشنهادی در شرایط با تنوع جریان و ظرفیت سبز استفاده نشده، توزیع بهتری از زمان بندی فازها ارائه می‌کند و اتلاف ظرفیت را به حداقل می‌رساند.

در [۳۲] یک مطالعه بر روی ترافیک شهر همدان انجام شده است. این داده‌ها از تقاطع‌های منطقه ۱ شهر همدان در ساعاتی با ترافیک زیاد، متوسط و کم جمع‌آوری شده است. تعداد خودروهای پشت چراغ قرمز، مسیرهای موجود در تقاطع و محدوده ترافیکی از مشخصات این مجموعه داده است. دسترسی به این مجموعه داده براساس درخواست از نویسندگان است. جدول ۳ عملکرد روش پیشنهادی را در سناریوهای پیک و غیرپیک برای شهر همدان نشان می‌دهد و بیانگر عمومیت‌پذیری چارچوب در بستری متفاوت از نمونه‌های بین‌المللی است. همان‌گونه که دیده می‌شود، روش پیشنهادی در هر دو سناریو کمترین

LSTM-PPO و CNN-DQN [۳۱] نشان می‌دهد. در سناریوی پیک، میانگین زمان انتظار روش پیشنهادی ۳۲ واحد است؛ این مقدار برای LSTM-PPO و CNN-DQN به ترتیب ۵۲ و ۴۶ واحد ثبت شده است. بنابراین روش پیشنهادی نسبت به [۳۰] حدود ۳۸ درصد و نسبت به [۳۱] حدود ۳۰ درصد کاهش انتظار ایجاد می‌کند. الگوی مشابهی در طول صف دیده می‌شود: مقدار روش پیشنهادی ۹ در برابر ۱۶ (LSTM-PPO) و ۱۳ (CNN-DQN) است که به معنای کاهش ۴۳ درصد و ۳۱ درصد می‌باشد. این اعداد نشان می‌دهند حتی تحت فشار ترافیکی بالا، سیاست یاد گرفته شده توسط روش پیشنهادی توانسته است نرخ تخلیه صف را افزایش دهد و از رشد صف‌های پشت چراغ جلوگیری کند. در سناریوی غیرپیک شکاف عملکردی پررنگ‌تر است. میانگین زمان انتظار روش پیشنهادی ۱۶ واحد در برابر ۳۴ و ۲۸ است؛ یعنی به ترتیب ۵۳ درصد و ۴۳ درصد بهبود نسبت به [۳۰ و ۳۱]. برای طول صف نیز مقادیر ۸ در برابر

پاداش نهایی و نوسانات را فراهم می‌کند. در شرایط پیک و غیرپیک، کاهش تأخیر و صف با روش پیشنهادی (به ترتیب ۶۶ و ۴۴ درصد) نسبت به مراجع به‌دست آمده و این مزیت در هر دو شهر پایدار است. مجموعه این شواهد نشان می‌دهد چارچوب ارائه شده هم کارا و هم قابل تعمیم است و می‌تواند مبنایی برای پیاده‌سازی‌های میدانی باشد. در آینده، گسترش به شبکه‌های چندتقاطع بزرگ‌مقیاس، بهره‌گیری از داده‌های چندوجهی (حسگرهای دیداری و سامانه‌های متصل) و اعمال قیود ایمنی و عدالت عبوری می‌تواند دامنه کاربرد و اطمینان‌پذیری سیستم را بیش از پیش افزایش دهد.

مراجع

1. El Mokhi, C., Erguig, H., Hmina, N., & Hachimi, H. (2025, April). Intelligent traffic management systems: A literature review on AI-Based traffic light control. In International Conference on Advanced Sustainability Engineering and Technology (pp. 154–171). Cham: Springer Nature Switzerland.
2. Agrahari, A., Dhabu, M. M., Deshpande, P. S., Tiwari, A., Baig, M. A., & Sawarkar, A. D. (2024). Artificial intelligence-based adaptive traffic signal control system: A comprehensive review. *Electronics*, 13(19), 3875.
3. Noaen, M.; Naik, A.; Goodman, L.; Crebo, J.; Abrar, T.; Abad, Z.S.H.; Bazzan, A.L.C.; Far, B. (2022). Reinforcement Learning in Urban Network Traffic Signal Control: A Systematic Literature Review. *Expert Syst. Appl.*, 99, 116830.
4. Jovanović, A.; Teodorović, D. (2017). Pre-Timed Control for an under-Saturated and over-Saturated Isolated Intersection: A Bee Colony Optimization Approach. *Transp. Plan. Technol.*, 40, 556–576.
5. Ahmed, E.K.E.; Khalifa, A.M.A.; Kheiri, A. (2018). Evolutionary Computation for Static Traffic Light Cycle Optimisation. *Proceedings of ICCCEEE, Khartoum, Sudan*, 12–14 August 2018; pp. 1–6.
6. Neelakandan, S.; Berlin, M.A.; Tripathi, S.; Devi, V.B.; Bhardwaj, I.; Arulkumar, N. (2021). IoT-Based Traffic Prediction and Traffic Signal Control System for Smart City. *Soft Comput.*, 25, 12241–12248.
7. Jing, P.; Huang, H.; Chen, L. (2017). An Adaptive Traffic Signal Control in a Connected Vehicle Environment: A Systematic Review. *Information*, 8, 101.
8. Kim, M.; Schrader, M.; Yoon, H.-S.; Bittle, J.A. (2023). Optimal Traffic Signal Control Using Priority Metric Based on Real-Time Measured Traffic Information. *Sustainability*, 15, 7637.

متوسط زمان انتظار را نسبت به دو روش یادگیری عمیق LSTM-PPO و CNN-DQN و همچنین روش چرخه ثابت ثبت کرده است. این برتری، همسو با نتایج ارائه شده در شکل‌های ۲ تا ۴، نشان می‌دهد طراحی بازنمایی وضعیت غنی و تصمیم‌گیری حساس به فاز چراغ به کاهش تأخیر در شرایط واقعی شبکه‌های شهری کشور نیز می‌انجامد. تفاوت معنادار با چرخه ثابت حاکی از آن است که سیاست‌های تطبیقی آموخته شده، از ظرفیت سبز بهتر استفاده کرده و از انباشت تأخیر در ساعات اوج جلوگیری می‌کنند؛ در غیرپیک نیز با توزیع کارآمد زمان‌بندی فازها از اتلاف ظرفیت جلوگیری می‌شود.

جدول ۳: مقایسه زمان انتظار در داده‌های شهر همدان

روش	پیک - متوسط زمان انتظار (ثانیه)	غیرپیک - متوسط زمان انتظار (ثانیه)
روش پیشنهادی	۳۳	۱۴
LSTM-PPO	۷۰	۴۶
CNN-DQN	۵۷	۳۴
چرخه ثابت	۱۵۲	۸۷

۵. نتیجه‌گیری و کارهای آینده

در این پژوهش، برای مسئله کنترل تطبیقی چراغ راهنمایی در شبکه‌های شهری، یک چارچوب یادگیری تقویتی عمیق معرفی شد که سه مؤلفه کلیدی دارد: بازنمایی وضعیت غنی و منطبق با فیزیک ترافیک (طول صف، تعداد خودرو، زمان انتظار تجمعی، فشار نرمال شده و حالت فعلی چراغ)، تصمیم‌گیری حساس به فاز چراغ با مسیرهای محاسباتی مجزا، و آموزش پایدار مبتنی بر حافظه تجربه تفکیکی و تابع پاداش چندمعیاره. ارزیابی روی دو مجموعه داده‌ها عمومی شهرهای موناکو و هاربین نشان داد که روش پیشنهادی در همه سناریوها سریع‌تر همگرا می‌شود، پایداری بیشتری دارد و در معیارهای میانگین زمان انتظار و طول صف بهبودهای معنی‌داری ایجاد می‌کند. تحلیل حساسیت نرخ جست‌وجو نیز نشان داد $\epsilon=0.9$ بهترین توازن میان سرعت یادگیری، سطح

- reinforcement learning for traffic signal control. *Journal of Big Data Analytics in Transportation*, 2(3), 263-274.
24. Wu, P., Wei, W., Zheng, L., Hu, Z., & Essa, M. (2023). Cycle-level traffic conflict prediction at signalized intersections with LiDAR data and Bayesian deep learning. *Accident Analysis & Prevention*, 192, 107268.
25. Fang, J., You, Y., Xu, M., Wang, J., & Cai, S. (2023). Multi-objective traffic signal control using network-wide agent coordinated reinforcement learning. *Expert Systems with Applications*, 229, 120535.
26. Chu, T., Wang, J., Codecà, L., & Li, Z. (2020). Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE transactions on intelligent transportation systems*, 21(3), 1086-1095.
27. Calvo, J. A., & Dusparic, I. (2018, December). Heterogeneous Multi-Agent Deep Reinforcement Learning for Traffic Lights Control. In *AICS* (pp. 2-13).
28. Choe, C. J., Baek, S., Woon, B., & Kong, S. H. (2018, November). Deep Q learning with LSTM for traffic light control. In *2018 24th Asia-Pacific Conference on Communications (APCC)* (pp. 331-336). IEEE.
29. Wang, T., Cao, J., & Hussain, A. (2021). Adaptive Traffic Signal Control for large-scale scenario with Cooperative Group-based Multi-agent reinforcement learning. *Transportation research part C: emerging technologies*, 125, 103046.
30. Huang, L., & Qu, X. (2023). Improving traffic signal control operations using proximal policy optimization. *IET Intelligent Transport Systems*, 17(3), 592-605.
31. Faqir, N., Loqman, C., & Boumhidi, J. (2022). Deep Q-learning approach based on CNN and XGBoost for traffic signal control. *Int. J. Adv. Comput. Sci. Appl*, 13(9).
32. Yasinian, H., & Esmailpour, M. (2022). Distributed learning automata based approach to inferring urban structure via traffic flow. *Applied Intelligence*, 52(2), 1338-1350.
9. Zaghal, R.; Thabatah, K.; Salah, S. (2017). Towards a Smart Intersection Using Traffic Load Balancing Algorithm. *Proceedings of the 2017 Computing Conference*, London, UK, 18–20 July 2017; pp. 485–491.
10. Mishra, S.; Singh, V.; Gupta, A.; Bhattacharya, D.; Mudgal, A. (2023). Adaptive Traffic Signal Control for Developing Countries Using Fused Parameters Derived from Crowd-Source Data. *Transp. Lett.*, 15, 296–307.
11. Tian, Y.; Liu, S.; Yan, X.; Zhu, T.; Zhang, Y. (2024). Active Control Method of Traffic Signal Based on Parallel Control Theory. *IEEE J. Radio Freq. Identif.*, 8, 334–340.
12. Zheng, Y. (2024). Intelligent Signal Optimization Algorithm Based on Artificial Intelligence in Intelligent Traffic Control Systems. *Procedia Computer Science*, 247, 445-452.
13. Hurtado-Gómez, J., Romo, J. D., Salazar-Cabrera, R., Pachon de la Cruz, A., & Madrid Molina, J. M. (2021). Traffic signal control system based on intelligent transportation system and reinforcement learning. *Electronics*, 10(19), 2363.
14. Castro, G. B., Hirakawa, A. R., & Martini, J. S. (2017). Adaptive traffic signal control based on bio-neural network. *Procedia Computer Science*, 109, 1182-1187.
15. Ng, S. C., & Kwok, C. P. (2020). An intelligent traffic light system using object detection and evolutionary algorithm for alleviating traffic congestion in Hong Kong. *International journal of computational intelligence systems*, 13(1), 802-809.
16. Medvei, M. M., Bordei, A. V., Niță, Ș. L., & Țăpuș, N. (2025). DeepSIGNAL-ITS—Deep Learning Signal Intelligence for Adaptive Traffic Signal Control in Intelligent Transportation Systems. *Applied Sciences*, 15(17), 9396.
17. Xu, M., Wu, J., Huang, L., Zhou, R., Wang, T., & Hu, D. (2020). Network-wide traffic signal control based on the discovery of critical nodes and deep reinforcement learning. *Journal of Intelligent Transportation Systems*, 24(1), 1-10.
18. Zhao, Z., Wang, K., Wang, Y., & Liang, X. (2024). Enhancing traffic signal control with composite deep intelligence. *Expert Systems with Applications*, 244, 123020.
19. Faqir, N., Ennaji, Y., Chakir, L., & Boumhidi, J. (2025). Hybrid CNN-LSTM and Proximal Policy Optimization Model for Traffic Light Control in a Multi-Agent Environment. *IEEE Access*.
20. Zhang, Z., Qian, J., Fang, C., Liu, G., & Su, Q. (2021). Coordinated control of distributed traffic signal based on multiagent cooperative game. *Wireless Communications and Mobile Computing*, 2021(1), 6693636.
21. Leal, S. S., & de Almeida, P. E. M. (2023). Traffic light optimization using non-dominated sorting genetic algorithm (NSGA2). *Scientific reports*, 13(1), 15550.
22. Tunc, I., & Soylemez, M. T. (2023). Fuzzy logic and deep Q learning based control for traffic lights. *Alexandria Engineering Journal*, 67, 343-359.
23. Tan, K. L., Sharma, A., & Sarkar, S. (2020). Robust deep