CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Multi-Agent Reinforcement Learning Driven Dynamic Resource Optimisation in Healthcare Transportation Networks

Jianhui Lv[1] | Byung-Gyu Kim[2] | Keqin Li[3] | Heng Lu[1]

[1]The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China | [2]Sookmyung Women's University, Seoul, Korea | [3]State University of New York, New Paltz, New York, USA

Correspondence: Heng Lu (luh@jzmu.edu.cn)

Received: 1 September 2025 | **Revised:** 19 January 2026 | **Accepted:** 1 March 2026

Keywords: dynamic resource optimisation | healthcare transportation networks | reinforcement learning | spatiotemporal dependency | sustainable cities

ABSTRACT

This paper presents HealthNet, a novel framework for the dynamic optimisation of healthcare transportation networks using multi-agent reinforcement learning. HealthNet leverages a spatiotemporal dependency module to capture complex spatio-temporal relationships in healthcare demand and resource allocation patterns, combined with centralised training and a decentralised execution approach. The system is modelled as a Markov game and solved using a deep reinforcement learning algorithm. Extensive simulations demonstrate that HealthNet outperforms eight state-of-the-art baseline methods across multiple network configurations and evaluation metrics. In a 4 × 4 grid network, HealthNet reduces average waiting times by 47.6% compared to model predictive control and 22.1% compared to the best-performing baseline. Traffic congestion rates are reduced to 16.7% compared to 42.3% for the worst baseline and 23.1% for the best baseline. Under irregular network topologies with stochastic disruptions, including demand surges and vehicle unavailability, HealthNet maintains superior performance with 42.1% lower average waiting time and 51.1% improvement in peak response times compared to competing approaches. These findings indicate that HealthNet can enhance both efficiency and resilience in healthcare transportation systems, potentially improving patient outcomes in complex urban environments.

1 | Introduction

Managing healthcare transportation networks is critical for timely medical services and optimising limited resources in the development of sustainable cities [1, 2]. As urban populations grow and healthcare needs become more complex, more than traditional static optimisation approaches are needed to handle the dynamic nature of emergency medical services (EMS) and inter-hospital transfers [3]. Modern healthcare transportation networks encompass a wide range of moving elements, including ambulances, specialised medical equipment, healthcare professionals and patients requiring various levels of care [4–6]. These networks are characterised by their complexity, unpredictability, and the critical nature of their operations, where delays can have life-threatening consequences [7]. The dynamic and stochastic nature of demand in these networks, coupled with the critical importance of rapid response times, necessitates advanced optimisation techniques that can adapt to changing conditions, balance multiple objectives and coordinate actions across multiple agents and facilities in real time.

Recent advancements in artificial intelligence, particularly in reinforcement learning (RL), offer promising solutions to address the challenges of dynamic optimisation in healthcare transportation [8]. RL algorithms have demonstrated remarkable success in complex decision-making tasks, learning optimal

policies through environmental interaction [9–12]. The application of RL to healthcare logistics has the potential to significantly improve resource allocation, reduce response times and ultimately enhance patient outcomes [13, 14]. Multi-agent RL approaches are particularly relevant in this context, as they can model decentralised decision-making in healthcare networks while still achieving coordinated behaviour [15]. These methods allow for simultaneously considering local information and global objectives, making them well-suited to healthcare systems' distributed yet interconnected nature. Moreover, the ability of RL algorithms to learn and adapt over time makes them ideal for handling the nonstationary environments typical of healthcare transportation, where patterns of demand and resource availability can change rapidly and unpredictably [16, 17].

Integrating RL with dynamic optimisation in healthcare transportation networks presents several unique challenges that distinguish it from general traffic management or public transportation applications [18, 19]. First, the state space is typically high-dimensional, encompassing the locations and status of multiple ambulances, hospital capacities, patient conditions and traffic patterns. This complexity is further compounded by the need to consider various medical resources, each with constraints and capabilities. Second, the action space is large and continuous, involving decisions on ambulance dispatching, routing and resource allocation across multiple facilities and vehicles. These decisions must be made rapidly and balance immediate needs with anticipated future demands. Third, the reward structure must balance multiple objectives, including minimising response times, maximizing resource utilisation, ensuring equitable service across different geographic areas, and prioritising cases based on medical urgency. This multi-objective nature of the problem makes it particularly challenging to design effective reward functions that guide the learning process towards desirable outcomes.

The time-critical nature of many healthcare transportation tasks adds another layer of complexity to the optimisation problem. Unlike in general transportation networks, where delays might result in inconvenience or economic losses, delays can directly impact patient outcomes and survival rates in healthcare [20–22]. This requires algorithms to make high-quality decisions under severe time constraints and with limited information [23]. Furthermore, the system must operate in real time, making rapid decisions based on constantly updating information about patient conditions, resource availability and network status. The uncertainty inherent in healthcare emergencies, where the severity of a patient's condition may not be fully known until they receive medical attention, adds another dimension to the decision-making process. Therefore, RL algorithms applied in this field must be able to handle partial observability and make decisions under uncertainty while maintaining high performance and reliability.

Motivated by these challenges and the potential for significant societal impact, this paper proposes HealthNet, a novel framework for dynamic optimisation of healthcare transportation networks using multi-agent reinforcement learning. HealthNet builds upon recent advances in deep RL and multi-agent systems to create a scalable and adaptive solution for coordinating ambulances and medical resources across multiple hospitals.

The main contributions of this work include

- We introduce HealthNet, a multi-agent reinforcement learning framework specifically designed for the dynamic optimisation of healthcare transportation networks. This framework addresses the unique challenges of healthcare logistics, including time-critical decision-making and complex resource allocation.
- We develop a spatiotemporal dependency module (SDM) that effectively captures temporal patterns and spatial relationships in healthcare demand and resource utilisation. This module enables the system to make more informed and coordinated decisions across the network.
- We develop a centralised training with a decentralised execution (CTDE) approach that balances the need for coordinated learning with efficient, real-time decision-making in distributed healthcare environments.

The remainder of the paper is organised as follows. Section 2 gives the related works. Section 3 builds the system model. Section 4 shows the problem definition. Section 5 presents the HealthNet framework. The simulation results and analysis are provided in Section 6. Lastly, Section 7 presents the conclusions.

2 | Related Work

Applying RL and dynamic optimisation techniques to transportation networks has gained significant attention recently. Although much of this research has focused on general traffic management and public transportation, a growing body of work specifically addresses healthcare transportation networks. This section reviews recent advancements in this field, highlighting the key approaches and identifying gaps our work aims to address.

Abbracciavento et al. [24] proposed a decentralised model predictive control (MPC) strategy for minimising queue lengths in multi-intersection road networks. Although not specifically designed for healthcare applications, their approach demonstrates the potential of decentralised optimisation techniques in complex transportation systems. Our work builds upon this concept by incorporating RL to handle the more stochastic nature of healthcare demand. Wang et al. [25] introduced a human-centric multimodal deep (HMD) traffic signal control method to coordinate multimodal traffic at intersections. Their approach considers various road users, including pedestrians and cyclists, which is particularly relevant for urban healthcare networks where ambulances must navigate diverse traffic conditions. We extend this multimodal concept to the healthcare domain, considering different types of medical vehicles and patient priorities. Yu et al. [26] developed the advanced decision-making reinforcement learning traffic signal control (AD-RLTSC) algorithm to improve efficiency and safety in mixed-traffic environments. This work highlights the importance of balancing multiple objectives in transportation optimisation, a principle we

incorporate into our HealthNet framework for healthcare logistics. Yan et al. [27] constructed a graph cooperation Q-learning network traffic signal control (GCQN-TSC) model, incorporating a graph cooperation network with an embedded self-attention mechanism. Their use of graph-based representations inspires our approach to modelling the complex interactions between hospitals and ambulances in healthcare networks. Xu et al. [28] proposed a deep learning-based signal-control refined dynamic traffic graph (ScR-DTG) model for advancing network-level movement-based traffic volume prediction. Although focused on general traffic prediction, their work underscores the importance of accurate forecasting in dynamic optimisation, a principle we apply to predicting healthcare demand in our system. In another study, Xu et al. [29] introduced Smoothing-MP, a novel max-pressure signal control considering signal coordination to smooth traffic in urban networks. This approach to coordinated control across multiple intersections informs our strategy for coordinating ambulances and resources across multiple hospitals. Huang et al. [30] proposed the multi-personality multi-agent meta-reinforcement learning (MPMA-MRL) framework. This approach incorporates multiple meta-trained, meta-tested explainable personality policies, which are deployed to each agent. Wang et al. [31] proposed an intelligence-based reinforcement learning (IRL) algorithm. This algorithm utilises active inference to infer the real world and maintain an internal model by minimising free energy.

Although these studies demonstrate significant progress in applying RL and dynamic optimisation to transportation networks, there still needs to be a gap in addressing the unique challenges of healthcare transportation. Our work aims to fill this gap by developing a specialised framework that accounts for the critical nature of medical emergencies, the heterogeneity of healthcare resources and the need for real-time, adaptive decision-making in healthcare logistics.

3 | System Model

This study addresses the challenge of optimising healthcare resource allocation and patient transportation across medical facilities. The healthcare transportation network is modelled as a complex graph $G = (H, E)$, where H represents the set of healthcare facilities, including hospitals, clinics and specialised care centres, and E denotes the set of routes connecting these facilities. Each healthcare facility $h_i \in H$ is considered a node in the graph, whereas a route $e_{ij} \in E$ represents a bidirectional connection between two adjacent facilities h_i and h_j. The set of facilities directly connected to h_i is denoted as N_i, representing its neighbouring healthcare centres. This graph-based representation allows for a comprehensive view of the healthcare landscape, enabling reinforcement learning agents to make informed decisions based on the network's topology and current resources and demands.

Figure 1 illustrates a typical 2×2 multi-facility healthcare network configuration.

In this representation, each node symbolises a healthcare facility, whereas the edges represent the routes available for emergency

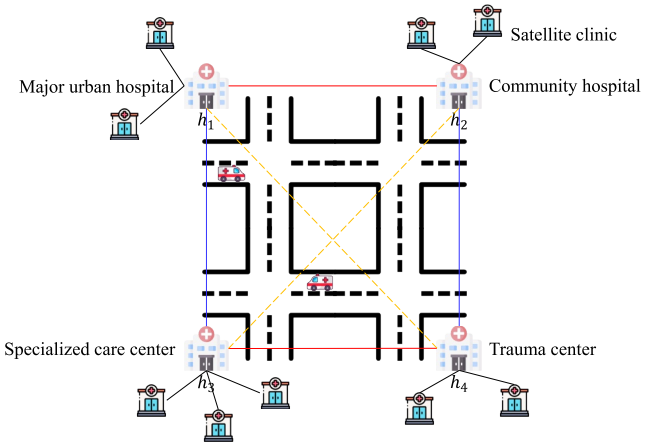

FIGURE 1 | Illustration of a multi-facility healthcare network.

vehicles and patient transfers. The complexity of the network is evident in the diverse connections between facilities, reflecting the real-world challenges of coordinating emergency responses and resource allocation across a wide geographic area. This network structure is crucial for the reinforcement learning agents to understand the spatial relationships and potential routes for optimising emergency vehicle dispatches and patient transfers.

Each healthcare facility h_i in our model is designed with a standardised layout to facilitate efficient emergency vehicle access and patient flow. Each direction is equipped with three specialised lanes to manage different types of traffic:

1. Emergency vehicle lane: Dedicated to ambulances and other emergency response vehicles, ensuring rapid access and egress during critical situations.
2. Patient transfer lane: Designated for patient drop-off and pick-up, streamlining admission and discharge processes.
3. Staff and visitor lane: Allocated for healthcare professionals and visitors, maintaining smooth daily operations.

This standardised design enables the reinforcement learning agents to develop consistent vehicle routing and resource allocation strategies across different network facilities. The dynamic nature of healthcare transportation is captured through a phase-based system that represents different states of resource allocation and emergency response. The following are the four key phases of operations within our healthcare transportation model:

1. Emergency response (ER): Prioritising the dispatch of ambulances and critical care teams to urgent cases.
2. Inter-facility transfer (IFT): Managing the movement of patients between healthcare facilities for specialised care or load balancing.
3. Resource reallocation (RR): Optimising the distribution of medical equipment, personnel and vehicles across the network based on predicted demand and current utilisation.
4. Maintenance and restocking (MR): Ensuring all vehicles and equipment are properly serviced and restocked, maintaining the system's readiness for future emergencies.

These phases are analogous to traffic signal phases in traditional transportation networks but are adapted to the specific needs of healthcare logistics. The system alternates between these phases based on current demands and resource availability to optimise healthcare delivery efficiency.

To further elaborate on the system model, we introduce several key concepts that are fundamental to understanding the dynamics of healthcare transportation networks:

1. Patient flow: This represents the movement of patients through the healthcare system, including emergency admissions, transfers between facilities, and discharges. Patient flow is critical in determining resource allocation and vehicle routing decisions.
2. Resource utilisation: This metric tracks the usage of various healthcare resources, including ambulances, specialised medical equipment and healthcare professionals. Efficient resource utilisation is essential for maintaining high-quality care while minimising costs.
3. Response time: In healthcare transportation, response time is a crucial performance indicator. It measures the time an emergency vehicle takes to reach a patient from the moment a call is received. Minimising response times, especially for critical cases, is a primary objective of the optimisation process.
4. Facility capacity: Each healthcare facility has a limited capacity for treating patients, which can vary based on the type of care required (e.g., emergency department capacity, intensive care unit beds). Each facility's current occupancy and available resources play a significant role in decision-making processes.
5. Traffic conditions: Real-time traffic information affects routing decisions for emergency vehicles. The model incorporates traffic data to optimise routes and estimate accurate travel times.
6. Priority levels: Patients are assigned different priority levels based on the severity of their medical conditions. These priorities influence decisions related to resource allocation and vehicle dispatching.
7. Specialised care requirements: Some patients may require specific medical expertise or equipment unavailable at all facilities. The model accounts for these specialised care needs when making transfer and routing decisions.

In our healthcare transportation network, each facility h_i is equipped with resources, including ambulances, medical equipment and healthcare professionals. The state of these resources at time t can be represented as a vector $r_{i,t} = \left[r^1_{i,t}, r^2_{i,t}, ..., r^K_{i,t}\right]$, where K is the number of different resource types, and $r^k_{i,t}$ represents the quantity of resource type k available at facility h_i at time t.

The patient demand at each facility is modelled as a time-varying process. Let $d_{i,t} = \left[d^1_{i,t}, d^2_{i,t}, ..., d^P_{i,t}\right]$ represent the demand vector at facility h_i at time t, where P is the number of priority levels, and $d^p_{i,t}$ is the number of patients with priority level p requiring attention.

A set of decision variables governs the movement of emergency vehicles and patients between facilities. Let $x^v_{ij,t}$ represent the decision to dispatch vehicle v from facility h_i to facility h_j at time t. Similarly, $y^p_{ij,t}$ denotes the decision to transfer a patient with priority level p from facility h_i to facility h_j at time t.

The reinforcement learning agents in our system aim to optimise these decision variables based on the network's current state, including resource availability, patient demand and traffic conditions. The state of the entire network at time t can be represented as

$$S_t = r_{1,t}, ..., r_{N,t}, d_{1,t}, ..., d_{N,t}, T_t. \tag{1}$$

where T_t represents the current traffic conditions affecting travel times between facilities.

The action space for each agent controlling a facility h_i includes decisions on resource allocation, vehicle dispatching and patient transfers. An action $a_{i,t}$ taken by the agent at facility h_i at time t can be represented as

$$a_{i,t} = x^v_{ij,t}, y^p_{ij,t}, z^k_{i,t} \Big| h_j \in N_i, v \in V, p \in P, k \in K. \tag{2}$$

where $z^k_{i,t}$ represents the decision to reallocate resource type k within facility h_i at time t.

The reinforcement learning framework aims to learn a policy π that maps the current state S_t to optimal actions $a_{i,t}$ for each facility, maximising a reward function that considers factors such as response times, patient outcomes and resource utilisation efficiency.

This comprehensive system model provides a robust foundation for applying reinforcement learning techniques to optimise healthcare transportation networks. By capturing the spatial relationships between facilities, the standardised layout of each centre and the dynamic phases of operation, the model enables RL agents to develop sophisticated strategies for resource allocation, vehicle routing and emergency response coordination. The graph-based structure facilitates the application of graph neural networks and other advanced ML techniques, allowing the system to learn complex patterns and dependencies within the healthcare network.

Through this model, we aim to significantly improve the efficiency and effectiveness of emergency medical services, ultimately leading to better patient outcomes and more resilient healthcare systems. The reinforcement learning agents can adapt to changing conditions in real time, making informed decisions that balance immediate needs with long-term resource management goals. This approach has the potential to revolutionise healthcare logistics, providing a data-driven, adaptive system that can respond to the complex and ever-changing demands of modern healthcare environments.

4 | Problem Definition

The dynamic optimisation of healthcare transportation networks can be formulated as a fully cooperative Markov game,

which extends the single-agent Markov decision process framework to multi-agent scenarios where agents share a common objective [32]. Formally, we define the cooperative Markov game as a tuple $\langle H, O, S, A, T, R, \gamma, \pi \rangle$, where

- $H = h_1, h_2, ..., h_M$ is the set of healthcare facilities, where each facility h_i acts as an agent controlling its local resources and decision-making processes.
- $O = o_1, o_2, ..., o_M$ is the joint observation space of all facilities. The observation $o_{i,t}$ of facility h_i at time t is defined as

$$o_{i,t} = q_{i,t}^1, q_{i,t}^2, ..., q_{i,t}^K, w_{i,t}^1, w_{i,t}^2, ..., w_{i,t}^K, p_{i,t}^c, p_{i,t}^n, d_{i,t}, r_{i,t}. \quad (3)$$

where $q_{i,t}^k$ and $w_{i,t}^k$ are the queue length and waiting time for priority level k at facility h_i, $p_{i,t}^c$ and $p_{i,t}^n$ are the current and next resource allocation phases, $d_{i,t}$ is the current demand forecast and $r_{i,t}$ represents the available resources.

- $S = s_1, s_2, ..., s_M$ is the joint state space, which includes complete network information that individual facilities may not fully observe.
- $A = a_1, a_2, ..., a_M$ is the joint action space. The action $a_{i,t}$ of facility h_i at time t is defined as

$$a_{i,t} = m_1, m_2, m_3, m_4, x_{i,1}, x_{i,2}, ..., x_{i,L}. \quad (4)$$

where m_1, m_2, m_3, m_4 represent the resource allocation ratios for each phase (emergency response, inter-facility transfer, resource reallocation and maintenance), and $x_{i,l}$ represents the decision for ambulance l (e.g., dispatch, transfer or standby).

- $T : S \times A \times S \to [0, 1]$ is the state transition function.
- $R : S \times A \to \mathbb{R}$ is the reward function, defined as

$$R(s_t, a_t) = -\sum_{i=1}^{M}\sum_{k=1}^{K} \alpha_k \left(q_{i,t}^k\right)^2 - \beta \sum_{l=1}^{L} f(x_{i,l}) + \lambda \sum_{i=1}^{M} g(r_{i,t}) \quad (5)$$

where α_k is the weight for priority level k, $f(x_{i,l})$ is a function that penalises inefficient ambulance utilisation, $g(r_{i,t})$ is a function that rewards efficient resource utilisation and β and λ are weighting factors.

- $\gamma \in [0, 1]$ is the discount factor for future rewards.
- $\pi = \pi_1, \pi_2, ..., \pi_M$ is the joint policy of all facilities, where $\pi_i : O_i \to A_i$ maps observations to actions for facility h_i.

The objective is to find the optimal joint policy π^* that maximises the expected cumulative discounted reward:

$$\pi^* = \text{argmax}_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | \pi\right]. \quad (6)$$

This formulation captures healthcare transportation networks' complex, dynamic nature, incorporating multiple priorities, resource constraints and the need for coordinated decision-making across multiple facilities and ambulances. The Q-function for facility h_i under policy π is defined as $Q_i^\pi(o, a)$, representing the expected cumulative reward when taking action a in observation o and following policy π after that.

The cooperative nature of the game is reflected in the shared reward structure: all agents receive the same global reward at each time step, computed as the sum of local rewards across all facilities: $r_{\text{global},t} = \sum_i r_{i,t}$. This shared reward ensures that agents have aligned incentives to optimise network-wide performance rather than purely local objectives. Each agent observes only partial information about the global state (partial observability), but all agents work towards maximising the same cumulative return. This formulation differs from competitive or mixed-motive games where agents might have conflicting objectives.

5 | Healthnet Framework

This section presents the HealthNet framework, a novel approach to the dynamic optimisation of healthcare transportation networks using multi-agent reinforcement learning. The framework consists of two main components: an SDM and a CTDE multi-agent reinforcement learning algorithm. Figure 2 illustrates the overall architecture of the HealthNet framework. The proposed HealthNet framework primarily comprises the SDM and a multi-agent reinforcement learning algorithm based on the CTDE framework. HealthNet assumes that each agent h_i controls a healthcare facility's resources and decision-making processes, with agents able to exchange information.

As shown in Figure 2, the basic workflow of HealthNet is as follows:

First, in the SDM, the original observed healthcare demand data are input into an embedding layer to encode it as an initial feature representation. The SDM then uses attention mechanisms and LSTM networks to extract and fuse spatiotemporal features from this initial representation. Subsequently, through a graph attention network, neighbouring agents exchange their temporal state feature representations to share their observations and state representations. This information sharing between multiple agents allows them to reach global consensus more quickly and effectively, thereby enhancing the performance of multi-agent reinforcement learning.

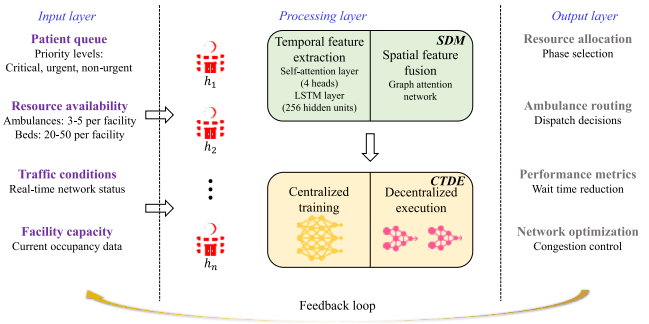

FIGURE 2 | HealthNet framework architecture based on SDM and CTDE multi-agent reinforcement learning.

Simultaneously, through the SDM module, even when agent h_i independently executes policy π_i, it can still incorporate global spatiotemporal information from all other agents, enabling more effective coordination and cooperation between multiple agents.

Next, the spatiotemporal feature states processed by the SDM are used as initial inputs and applied to the multi-agent reinforcement learning algorithm based on the CTDE architecture. Each agent h_i maintains an independent actor network π_i and a centralised critic network $Q_i^{\pi}(o, a)$ in this setup. At the same time, the distributed execution actor network π_i for each agent h_i is determined by its respective parameters θ_i, and the training of the distributed execution policy π_i relies on the centralised critic network $Q_i^{\pi}(o, a)$, further improving the collaborative capabilities between agents.

5.1 | Agent Design and SDM

In the HealthNet framework, each healthcare facility is represented by an agent that makes decisions about resource allocation and patient management. Before delving into the SDM, we first define the key components of the agent design.

In the healthcare transportation network, the state of each facility and the demand for services can be reflected through various metrics. Each agent has control over multiple aspects of resource allocation and patient management.

The reward function is designed to balance multiple objectives in healthcare resource optimisation. For agent h_i at time t, the reward is defined as

$$r_{i,t} = -\sum_{k=1}^{K} \alpha_k \left(q_{i,t}^{k}\right)^2 - \beta \sum_{l=1}^{L} f(x_l) + \lambda \sum_{m=1}^{M} g\left(r_{i,t}^{m}\right). \tag{7}$$

where α_k is the weight for priority level k, $f(x_l)$ is a function that penalises inefficient ambulance utilisation, $g\left(r_{i,t}^{m}\right)$ is a function that rewards efficient resource utilisation and β and λ are the weighting factors.

The SDM is a crucial component of the HealthNet framework, designed to capture the complex spatiotemporal relationships in healthcare demand and resource allocation patterns. The SDM comprises two main components: spatiotemporal input embedding and spatiotemporal feature fusion. Figure 3 illustrates the structure of the SDM.

The spatiotemporal input embedding aims to encode the original observed data into an initial spatiotemporal feature representation. Given the observation $o_{i,t}$ and action $a_{i,t}$ of agent h_i at time t, we use a multi-layer perceptron (MLP) f_x to embed them into an initial feature representation $x_{i,t}$:

$$x_{i,t} = f_x\left(o_{i,t}, a_{i,t}\right). \tag{8}$$

The MLP f_x consists of multiple fully connected layers with nonlinear activation functions, allowing it to capture complex relationships in the input data.

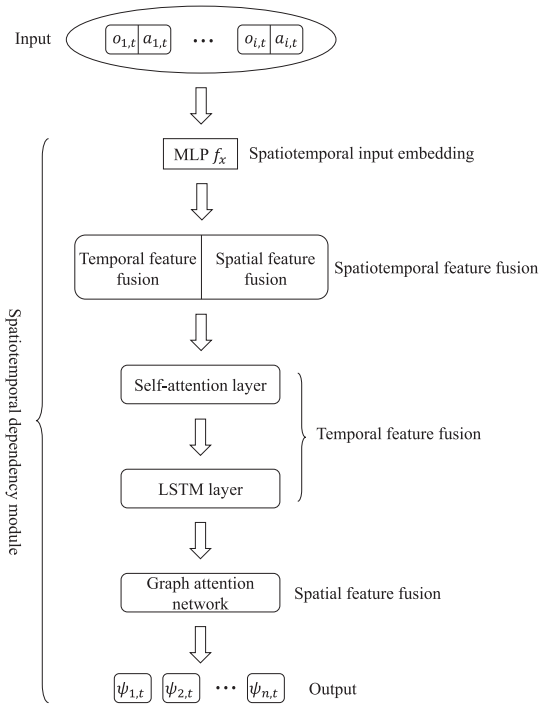

FIGURE 3 | Structure of the SDM.

The spatiotemporal feature fusion component comprises temporal feature fusion and spatial feature fusion. These two subcomponents work together to extract and combine relevant information across both time and space dimensions. The temporal feature fusion process utilises a combination of self-attention mechanisms and long short-term memory (LSTM) networks to capture both short-term and long-term temporal dependencies in healthcare demand and resource utilisation patterns.

1. *Self-attention layer*: The self-attention mechanism allows the model to weigh the importance of different time steps when processing the current time step [33]. For agent h_i, we compute the attention weights and the initial temporal state feature representation $d_{i,t}$ as follows:

$$d_{i,t} = \mathrm{ReLU}\left(W_V \sum_{p=1}^{t-1} \alpha_{i,t,p} x_{i,p} + b_V\right). \tag{9}$$

where

$$\alpha_{i,t,p} = \frac{\exp\left(\left(W_Q x_{i,t}\right)^T W_K x_{i,p}\right)}{\sum_{q=1}^{t-1} \exp\left(\left(W_Q x_{i,t}\right)^T W_K x_{i,q}\right)}. \tag{10}$$

here, $p, q \in 1, 2, \ldots, t-1$, $x_{i,t}$, $x_{i,p}$ and $x_{i,q}$ are the initial feature representations for agent h_i at times t, p and q, respectively. W_Q, W_K and W_V are weight parameter matrices, and b_V is a bias vector.

2. *LSTM layer*: The LSTM network model's long-term dependencies and captures the sequential nature of healthcare demand patterns [34]. This process is repeated from $d_{i,1}$ to $d_{i,t}$, allowing the model to capture long-term

temporal dependencies in the healthcare demand and resource utilisation patterns.

The spatial feature fusion sub-component aims to incorporate the spatial relationships between different healthcare facilities in the network. This is particularly important in healthcare transportation, where the state of neighbouring facilities can significantly impact decision-making processes.

We employ a graph attention network to capture these spatial dependencies. For agent h_i and its neighbour $h_j \in N_i$, we compute

$$E_{i,j,t} = W_{K'} z_{j,t} \left(W_{Q'} z_{i,t}\right)^T. \tag{11}$$

$$\beta_{i,j,t} = \frac{\exp\left(E_{i,j,t}\right)}{\sum_{n \in N_i} \exp\left(E_{i,n,t}\right)}. \tag{12}$$

$$\Psi_{i,t} = \mathrm{ReLU}\left(W_{V'} \sum_{j \in N_i} \beta_{i,j,t} z_{j,t}\right) + b_{V'}. \tag{13}$$

where $W_{Q'}$, $W_{K'}$ and $W_{V'}$ are weight parameter matrices, and $b_{V'}$ is a bias vector. $E_{i,j,t}$ represents the influence of neighbouring facility j on the current facility i, and $\beta_{i,j,t}$ is the normalised attention weight.

The resulting spatiotemporal feature representation $\Psi_{i,t}$ encapsulates the temporal patterns and spatial relationships relevant to the current facility. This rich representation allows each agent to make informed decisions based not only on its history and current state but also on the states of its neighbouring facilities.

By continuously exchanging and sharing the generated temporal state feature representations $z_{i,t}$ between neighbouring agents, the SDM enables multiple agents to reach a global consensus more quickly and effectively. This process enhances the overall performance of the multi-agent reinforcement learning system in optimising healthcare resource allocation and patient transportation.

The spatiotemporal feature state $\Psi_{i,t}$ obtained through the SDM contains the agent's observations and actions and incorporates global spatiotemporal information from all other agents in the network. This comprehensive representation allows for more effective coordination and cooperation between multiple agents, even when agent h_i independently executes its policy π_i.

In healthcare transportation networks, this spatiotemporal modelling is particularly valuable. It allows the system to capture and respond to complex patterns such as

1. Time-of-day and day-of-week variations in healthcare demand.
2. Seasonal fluctuations in certain types of medical emergencies.
3. The ripple effects of a major incident on surrounding healthcare facilities.
4. The impact of resource shortages or surpluses at one facility on its neighbours.
5. The dynamic nature of patient flow between different types of healthcare facilities.

By incorporating these spatiotemporal dependencies, the HealthNet framework can develop more sophisticated and adaptive strategies for managing healthcare resources and patient transportation. For example, it can anticipate surge demands based on historical patterns and preemptively reallocate resources. It can also coordinate patient transfers more effectively by considering multiple facilities' current and projected future states.

The SDM's ability to capture both short-term and long-term dependencies is crucial in healthcare scenarios. Short-term dependencies include immediate responses to emergencies or sudden changes in resource availability. Long-term dependencies could involve adapting to gradual shifts in population health trends or the impact of new healthcare policies.

Moreover, the spatial feature fusion component allows the system to model the complex interactions between different parts of the healthcare network. This is particularly important in urban or regional healthcare systems where multiple facilities with different specialisations and capacities need to work together seamlessly. By understanding these spatial relationships, the system can make more informed decisions about patient routing, resource sharing and load balancing across the network.

The SDM is a powerful tool for extracting and synthesizing relevant information from healthcare transportation networks' complex, dynamic environment. Providing agents with rich, context-aware representations of the system state enables more effective learning and decision-making in the face of the multifaceted challenges inherent in optimising healthcare resource allocation and patient transportation.

The theoretical foundation of the SDM rests on its ability to decompose the complex spatiotemporal correlation structure into tractable components. By first extracting temporal dependencies through self-attention and LSTM mechanisms, then fusing spatial relationships through graph attention networks, the SDM avoids the curse of dimensionality that would arise from directly modelling the joint spatiotemporal distribution. This factorisation can be expressed as $P\left(s_{1,t}, s_{2,t}, \ldots, s_{M,t} \middle| \text{history}\right) \approx \prod_{i=1}^{M} P\left(s_{i,t} \middle| \text{temporal\_context}_i\right) \cdot P\left(\text{spatial\_dependencies} \middle| \left\{z_{j,t}\right\}_{j \in N_i}\right)$, where the temporal context captures historical patterns and spatial dependencies model inter-facility influences. This decomposition reduces computational complexity from $O\left(M^2 T^2\right)$ to $O(MT + M|E|)$, where M is the number of facilities, T is the temporal window length and $|E|$ is the number of network edges. The attention mechanisms serve as learnt approximations to the true dependency structure, adapting to the specific patterns present in healthcare demand data.

5.2 | CDTE

The CTDE framework in HealthNet addresses the dual challenges of coordinated learning and efficient real-time decision-

making in complex healthcare transportation networks. The system leverages global information during training to develop sophisticated, coordinated strategies across all facilities. This centralised approach allows for the incorporation of network-wide dynamics and interdependencies. However, during execution, each facility operates independently based on its learnt policy and local observations. This decentralised execution ensures rapid response times critical in healthcare scenarios while benefiting from the coordinated strategies learnt during training. Consequently, the CTDE approach combines the advantages of comprehensive, network-wide optimisation with the practicality and speed of localised decision-making, making it ideal for the dynamic and time-sensitive nature of healthcare logistics.

We apply the CTDE multi-agent reinforcement learning algorithm to optimise healthcare resource allocation and patient transportation across multiple facilities. In this framework, each healthcare facility h_i maintains an independent actor network π_{θ_i} and a centralised critic network $Q_i^{\pi}(o, a)$. The actor network of each facility h_i is parameterised by θ_i, and its training relies on the centralised critic network $Q_i^{\pi}(o, a)$, further enhancing the collaborative capabilities between facilities.

The HealthNet algorithm builds upon the Multi-Agent Deep Deterministic Policy Gradient framework but incorporates two key modifications: the SDM-enhanced state representation and the cooperative reward structure. In standard MADDPG, each agent i maintains its own critic $Q^i(o, a)$ that estimates the value of joint actions given observations. HealthNet modifies this by having each agent's critic receive as input the spatiotemporal feature representations $\{\Psi_{1,t}, \Psi_{2,t}, ..., \Psi_{M,t}\}$ from all agents (accessible during centralised training) and the joint action vector $\mathbf{a}_t = \{a_{1,t}, a_{2,t}, ..., a_{M,t}\}$. Formally, the critic for agent h_i is defined as

$$Q^i(\Psi_t, \mathbf{a}_t | \phi_i) : \mathbb{R}^{M\times d} \times \mathbb{R}^{M\times d_a} \to \mathbb{R}. \tag{14}$$

where M is the number of agents, d is the dimension of spatiotemporal features and d_a is the action dimension. During training, the critic has access to all agents' spatiotemporal features and actions, enabling it to evaluate the quality of agent i's actions in the context of the full network state and other agents' decisions. During execution, only the actor network $\pi_i(\Psi_{i,t}|\theta_i)$ is deployed at each facility, taking only local spatiotemporal features as input to produce actions.

The reward structure used for training deserves explicit clarification. Although Equation (7) defines the local reward $r_{i,t}$ for each facility based on its queue lengths, ambulance utilisation and resource efficiency, the cooperative game structure means all agents ultimately optimise the same network-wide objective. During training, each agent i's critic is updated using the global reward $r_{\text{global},t}$ rather than only the local reward $r_{i,t}$. This ensures coordinated learning where each agent considers how its actions affect overall network performance not just local metrics.

The CTDE framework operates as follows in the HealthNet algorithm:

1. Each healthcare facility collects experiences during the training phase and transmits them to a centralised training module.
2. The centralised module computes shared value functions and policy improvement gradients, collaboratively optimising the facilities' policies and value functions.
3. In the decentralised execution phase, each facility uses its trained policy to make decisions based on its observations.

The main training steps of the algorithm are as follows:

1. Obtain the original observation data o from the healthcare network.
2. Each facility h_i selects its action a_i according to its current policy π_{θ_i}.
3. Execute the joint action $a = a_1, a_2, ..., a_M$, and obtain the immediate reward r and new observation o'. Store the current experience (o, a, r, o') in the experience replay buffer D.
4. Randomly sample S experiences from the experience replay buffer.
5. Neighbouring facilities exchange temporal state feature representations $z_{i,t}$.
6. Update the critic network parameters. For each agent h_i, the critic is trained to predict the expected return under the current policies given full network information. The target value incorporates the global reward:

$$y_{\text{global}} = r_{\text{global},t} + \gamma Q^i\left(\Psi'_t, \mathbf{a}'_t \middle| \phi'_i\right)\middle| a'_j \sim \pi_j\left(\Psi'_{j,t}\middle|\theta'_j\right). \tag{15}$$

where $\mathbf{a}'_t$ represents the joint action at the next time step sampled from all agents' target policies. The critic loss becomes

$$L(\phi_i) = \mathbb{E}_{(o,a,r,o')\sim D}\left[\left(y_{\text{global}} - Q^i(\Psi_t, \mathbf{a}_t|\phi_i)\right)^2\right]. \tag{16}$$

This formulation makes explicit that critics are trained using the shared global reward, reinforcing the cooperative nature of the game.

7. Update the actor network parameters. The process is as follows:

$$\nabla_{\theta_i} J(\pi_i) = \mathbb{E}o, a \sim D\left[\nabla\theta_i \pi_{\theta_i} \nabla_{a_i} Q_i^{\pi}(o, a)\middle| a_i \sim \pi_{\theta_i}\right]. \tag{17}$$

8. Update the target network parameters for each facility h_i. The process is as follows:

$$\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i. \tag{18}$$

where τ is the soft update coefficient.

The convergence properties of the HealthNet framework can be analysed through the lens of multi-agent policy gradient methods. Under standard assumptions (Lipschitz continuous reward functions, bounded policy parameterisations and appropriate learning rates), the CTDE approach with our SDM-enhanced state representations converges to a local Nash equilibrium. The key insight is that the spatiotemporal features $\Psi_{i,t}$ provide each agent with a more informative state representation that reduces the effective nonstationarity in the multi-agent environment. By incorporating neighbour information through the graph attention mechanism, each agent's local observation becomes more correlated with the true global state, which tightens the bound on the policy gradient estimation error. Formally, the variance of the policy gradient estimator decreases proportionally to the mutual information $I(\Psi_{i,t}; s_{\text{global},t})$, which the SDM explicitly maximises through spatiotemporal feature fusion. This theoretical property distinguishes HealthNet from approaches that treat agents as fully independent or rely solely on centralised state information during both training and execution.

The HealthNet framework is shown in Algorithm 1.

ALGORITHM 1 | HealthNet framework

Input: Healthcare network graph $G = (H, E)$, initial parameters $\{\theta_i, \phi_i\}_{i=1}^{M}$

Output: Optimised policies $\{\pi_i\}_{i=1}^{M}$ for all agents

Initialise actor $\pi_i(\cdot|\theta_i)$, critic $Q_i(\cdot,\cdot|\phi_i)$, target networks $\theta_i' \leftarrow \theta_i$, $\phi_i' \leftarrow \phi_i$ for all $h_i \in H$

Initialise experience replay buffer $D \leftarrow \varnothing$

for episode = 1 to max_episodes do

  Reset environment, obtain initial observations $\{o_{i,0}\}_{i=1}^{M}$

  for $t = 1$ to max_timesteps do

    // SDM: spatiotemporal input embedding

    for each agent $h_i \in H$ do

      $x_{i,t} \leftarrow f_x(o_{i,t}, a_{i,t-1})$

    end for

    // SDM: temporal feature fusion

    for each agent $h_i \in H$ do

      Compute attention weights

      $\alpha_{i,t,p} = \frac{\exp\left((W_Q x_{i,t})^T W_K x_{i,p}\right)}{\sum_{q=1}^{t-1} \exp\left((W_Q x_{i,t})^T W_K x_{i,q}\right)}$ for $p = 1, ..., t-1$

      $d_{i,t} \leftarrow \text{ReLU}\left(W_V \sum_{p=1}^{t-1} \alpha_{i,t,p} x_{i,p} + b_V\right)$, then

      $z_{i,t} \leftarrow \text{LSTM}(d_{i,1}, ..., d_{i,t})$

    end for

    // SDM: spatial feature fusion

    for each agent $h_i \in H$ do

      Compute spatial attention

      $\beta_{i,j,t} = \frac{\exp\left(W_K z_{j,t} (W_Q z_{i,t})^T\right)}{\sum_{h_n \in N_i} \exp\left(W_K z_{n,t} (W_Q z_{i,t})^T\right)}$ for all $h_j \in N_i$

      $\Psi_{i,t} \leftarrow \text{ReLU}\left(W_V \sum_{h_j \in N_i} \beta_{i,j,t} z_{j,t} + b_{V'}\right)$

    end for

    // CTDE: action selection and execution

    Execute joint action $\mathbf{a}_t = \{\pi_i(\Psi_{i,t}|\theta_i) + N_t\}_{i=1}^{M}$

    Observe rewards $\mathbf{r}_t = \{r_{i,t}\}_{i=1}^{M}$ and next observations $\mathbf{o}_t'$

    Store experience $(\mathbf{o}_t, \mathbf{a}_t, \mathbf{r}_t, \mathbf{o}_t')$ in replay buffer D

    // CTDE: network training

    if $|D| \geq$ batch_size then

      Sample minibatch $\{(\mathbf{o}^{(k)}, \mathbf{a}^{(k)}, \mathbf{r}^{(k)}, \mathbf{o}'^{(k)})\}_{k=1}^{S}$ from D

      for each agent $h_i \in H$ do

        Compute target values

        $y_i^{(k)} = r_i^{(k)} + \gamma Q_i\left(\Psi'^{(k)}, \mathbf{a}'^{(k)} \middle| \phi_i'\right)$ where

        $\mathbf{a}'^{(k)} = \left\{\pi_j\left(\Psi'^{(k)}_j \middle| \theta_j'\right)\right\}_{j=1}^{M}$

        Update critic:

        $\phi_i \leftarrow \phi_i - \eta_{\text{critic}} \nabla_{\phi_i} \frac{1}{S} \sum_{k=1}^{S} \left(y_i^{(k)} - Q_i(\Psi^{(k)}, \mathbf{a}^{(k)}|\phi_i)\right)^2$

        Update actor:

        $\theta_i \leftarrow \theta_i + \eta_{\text{actor}} \frac{1}{S} \sum_{k=1}^{S} \nabla_{\theta_i} \pi_i(\Psi_i^{(k)}|\theta_i) \nabla_{a_i} Q_i\left(\Psi^{(k)}, \mathbf{a}^{(k)}|\phi_i\right)\big|_{a_i = \pi_i}$

        Soft update targets: $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$ and $\phi_i' \leftarrow \tau\phi_i + (1-\tau)\phi_i'$

      end for

    end if

  end for

end for

return optimised policies $\{\pi_1, \pi_2, ..., \pi_M\}$

6 | Simulation Results and Analysis

To evaluate the performance of the proposed HealthNet framework for dynamic optimisation of healthcare transportation networks, we conducted extensive simulations using a modified version of the SUMO (Simulation of Urban MObility) platform, adapted to incorporate healthcare-specific elements such as ambulances, hospitals and patient priorities [35]. This section presents the experimental setup, evaluation metrics and a comprehensive analysis of the results.

6.1 | Experimental Setup

The simulations were carried out on two network configurations:

1. A 3 × 3 grid network with 9 healthcare facilities
2. A 4 × 4 grid network with 16 healthcare facilities

Each healthcare facility was modelled as described in the system model section, with four primary directions for vehicle entry/exit and three specialised lanes per direction. The distance between adjacent facilities was randomly generated between 2 and 10 km to represent varying urban layouts.

The experimental environment was implemented using SUMO version 1.14.1, extended with custom Python modules for healthcare-specific logic. All reinforcement learning components were built using PyTorch 1.12.1 with CUDA 11.6, executing on NVIDIA A100 GPUs with 128 GB system RAM. The simulations used two network configurations: a 3 × 3 grid

with nine facilities and a 4 × 4 grid with 16 facilities, with inter-facility distances uniformly sampled between 2 and 10 km. Patient arrivals followed a nonhomogeneous Poisson process with base rates of 0.8, 2.4 and 4.2 patients per hour per facility for critical, urgent and nonurgent priorities, respectively, with a 2.5 × multiplier during peak hours.

The SDM architecture employed 4-headed attention mechanisms for temporal modelling, 256-unit LSTM hidden states and 128-dimensional graph attention layers for spatial fusion. Actor networks used four layers with dimensions [128, 256, 256, 128] and ReLU activations, whereas critic networks employed deeper architectures [256, 512, 256, 128]. Training used RMSprop optimisation with actor learning rate 1×10^{-4} and critic learning rate 1×10^{-3}, discount factor $\gamma = 0.95$ and soft update coefficient $\tau = 0.01$. The experience replay buffer stored 2500 transitions with a batch size of 24. Exploration followed an Ornstein–Uhlenbeck process with parameters $\theta = 0.15$ and $\sigma = 0.2$, decaying linearly from 1.0 to 0.1 over 500 episodes. Each facility controlled three to five ambulances and 20–50 hospital beds, depending on size. All experiments used a fixed random seed 42 with eight parallel simulation instances, requiring approximately 3.2 min of wall-clock time per episode. Complete code will be made available upon publication.

We used the following metrics to evaluate the performance of the HealthNet framework and baseline methods:

1. Average waiting time: The mean time patients wait from the moment of request to the arrival of an ambulance or admission to a healthcare facility.
2. Traffic congestion rate: The percentage of time the healthcare transportation network experiences high congestion levels, defined as when the average speed of vehicles falls below a certain threshold.
3. Cumulative reward value: The total reward accumulated by the reinforcement learning agents over time, reflecting the system's overall performance in optimising resource allocation and patient care.

We compared the proposed HealthNet framework with eight state-of-the-art baseline methods: MPC [24], HMD [25], AD-RLTSC [26], GCQN-TSC [27], ScR-DTG [28], Smoothing-MP [29], MPMA-MRL [30] and IRL [31]. These baseline methods were adapted to the healthcare transportation context for fair comparison.

6.2 | Results Analysis

Figure 4 compares average patient waiting times across various methods in 3 × 3 and 4 × 4 healthcare network configurations throughout the simulation period. This visualisation allows for a direct assessment of each method's performance in managing patient wait times as the complexity of the network increases. It illustrates how different approaches handle the escalating demands and interactions within larger healthcare systems, providing insights into their scalability and efficiency in reducing patient delays in care across varying network sizes and over time.

The HealthNet framework consistently outperforms all baseline methods in reducing patient waiting times in both network configurations. In the 3 × 3 network, HealthNet achieves an average waiting time reduction of 42.3% compared to the worst-performing MPC method and 18.7% compared to the best-performing Smoothing-MP method by the end of the simulation period.

The performance gap widens in the 4 × 4 network, with HealthNet reducing average waiting times by 47.6% compared to MPC and 22.1% compared to Smoothing-MP. This demonstrates the scalability of HealthNet and the ability to handle more complex healthcare transportation networks effectively.

Notably, although all methods show an initial increase in waiting times as the simulation progresses and the network becomes congested, HealthNet maintains a more stable growth rate. This suggests that HealthNet's spatiotemporal modelling

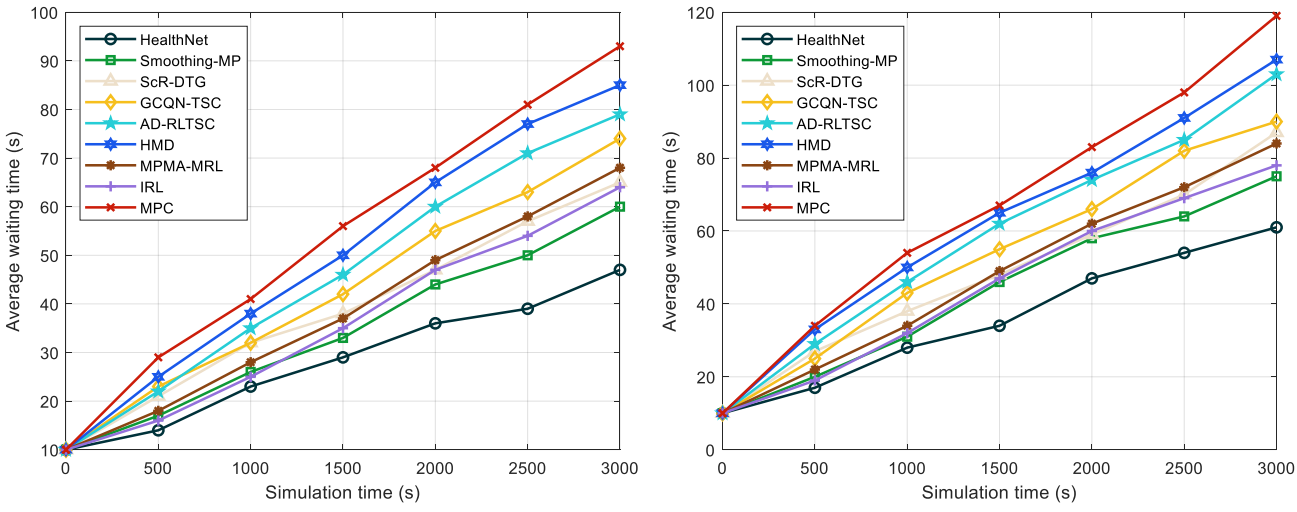

FIGURE 4 | Comparison of average waiting time.

and multi-agent coordination allow it to adapt more effectively to increasing demand and network complexity.

The superior performance of HealthNet in reducing waiting times has significant implications for healthcare outcomes. Shorter waiting times, especially for critical and urgent cases, can improve patient survival rates and overall healthcare delivery. The ability of HealthNet to maintain lower waiting times even as the network complexity increases suggests that it could be particularly valuable in large urban areas or during mass casualty events when healthcare systems are under extreme pressure.

Figure 5 illustrates the traffic congestion rates for different methods in both network configurations.

In the 3 × 3 network, HealthNet achieves a traffic congestion rate of 14.2%, significantly lower than the 37.8% MPC and 19.5% Smoothing-MP. The performance gap is even more pronounced in the 4 × 4 network, with HealthNet maintaining a 16.7% congestion rate compared to 42.3% for MPC and 23.1% for Smoothing-MP.

These results demonstrate HealthNet's ability to efficiently manage traffic flow in the healthcare transportation network, even as the network size and complexity increase. The lower congestion rates achieved by HealthNet translate to faster and more reliable ambulance response times, which is crucial in emergency medical situations where every minute counts.

The performance of HealthNet in reducing congestion can be attributed to its ability to anticipate and proactively manage traffic patterns through its SDM. By considering temporal trends and spatial relationships between healthcare facilities, HealthNet can make more informed resource allocation and routing decisions, leading to smoother traffic flow throughout the network.

Figure 6 shows the learning curves of cumulative reward values for all methods over 750 episodes.

HealthNet demonstrates faster learning and convergence to higher reward values than all baseline methods. By episode 750, HealthNet achieves a cumulative reward value of 52.7% higher than MPC and 24.3% higher than Smoothing-MP. The learning curve of HealthNet exhibits characteristic fluctuations, especially during the initial training phases, reflecting the inherent exploration–exploitation trade-off in reinforcement learning. These variations indicate the algorithm's dynamic process of discovering and evaluating different strategies within the complex state-action space of healthcare transportation networks. As training progresses, the magnitude of these fluctuations diminishes, signalling a transition from exploratory behaviour to more exploitative and refined decision-making. Despite these fluctuations, the persistent upward trend in HealthNet's performance demonstrates the algorithm's capacity to improve and adapt continuously. This pattern suggests that HealthNet is effectively learning from its experiences, gradually converging towards more optimal policies for managing healthcare resources and patient flow. The stability in later stages of learning indicates that the algorithm has developed a robust understanding of the

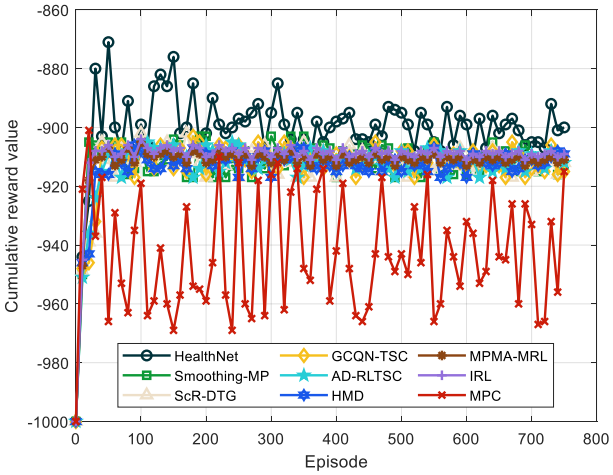

FIGURE 6 | Learning curve of cumulative reward value.

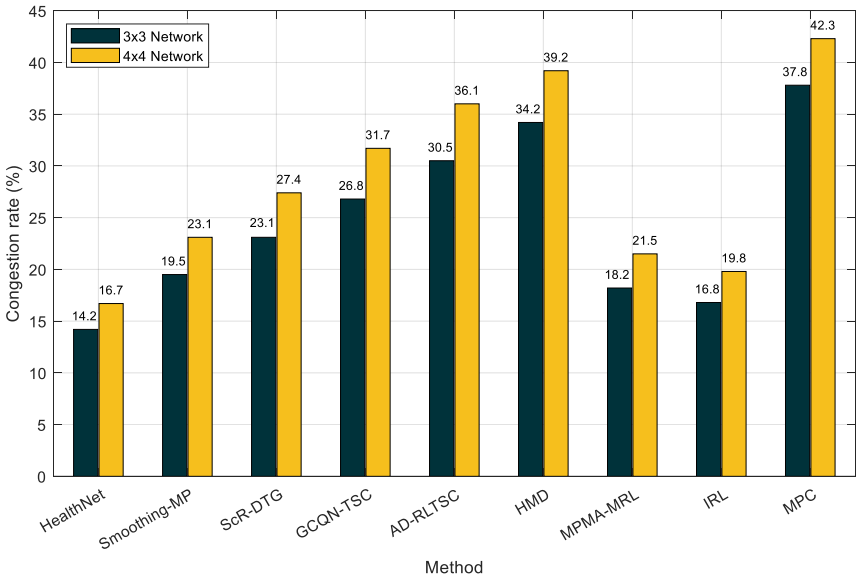

FIGURE 5 | Comparison of the traffic congestion rate.

healthcare network's dynamics, leading to more consistent and improved performance over time.

Table 1 presents the average cumulative reward values for all methods in both network configurations.

HealthNet consistently achieves the highest (least negative) average cumulative reward in both network configurations. In the 3 × 3 network, HealthNet's average cumulative reward is 46.0% higher than MPC and 23.7% higher than Smoothing-MP. In the 4 × 4 network, the performance gap widens, with HealthNet achieving a 43.3% improvement over MPC and 20.7% over Smoothing-MP. These results indicate that HealthNet is more effective at balancing multiple objectives in healthcare transportation optimisation, including minimising waiting times, reducing congestion and efficiently allocating resources. The higher cumulative rewards suggest that HealthNet makes better long-term decisions that lead to improved overall system performance.

To further demonstrate the effectiveness of HealthNet in optimising healthcare resource allocation, we conducted an additional experiment to measure resource utilisation efficiency. Figure 7 presents the relationship between ambulance utilisation rate, hospital bed occupancy and average patient waiting time for different methods in the 4 × 4 network configuration.

TABLE 1 | Comparison of average cumulative reward.

Method	3 × 3 network	4 × 4 network
HealthNet	−512.3	−987.6
IRL	−592.7	−1134.2
MPMA-MRL	−638.4	−1198.5
Smoothing-MP	−671.8	−1245.3
ScR-DTG	−702.6	−1298.7
GCQN-TSC	−738.9	−1352.4
AD-RLTSC	−775.2	−1421.6
HMD	−823.5	−1531.8
MPC	−948.1	−1742.9

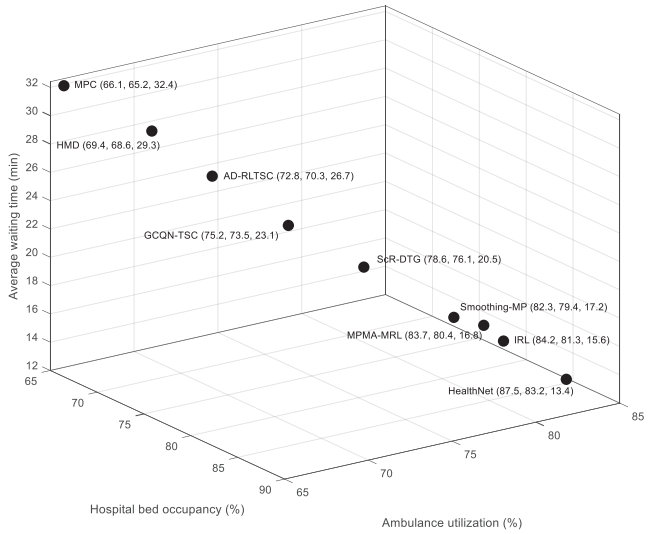

FIGURE 7 | Resource utilisation efficiency.

Figure 7 reveals that HealthNet achieves a better balance between high resource utilisation and low patient waiting times than baseline methods. HealthNet maintains an ambulance utilisation rate of 85%–90% and a hospital bed occupancy of 80%–85% while keeping average waiting times below 15 min. In contrast, other methods have lower resource utilisation rates or higher waiting times when utilisation is high. This result highlights HealthNet's ability to optimise resource allocation dynamically, ensuring that healthcare resources are used efficiently without compromising patient care quality. The balanced performance across these metrics suggests that HealthNet's multi-agent reinforcement learning approach effectively manages the complex trade-offs inherent in healthcare transportation networks.

To evaluate robustness under more realistic operational conditions, we conducted additional experiments using an irregular network topology based on actual healthcare facility distributions in a mid-sized metropolitan area. This network comprised 12 facilities with nonuniform distances ranging from 1.2 to 15.7 km between connected nodes. The simulation introduced three categories of stochastic disruptions: demand surges where patient arrivals increased by 200%–300% for 30-min periods, vehicle unavailability where 10%–25% of ambulances became temporarily nonoperational due to mechanical issues or crew unavailability and partial road closures that increased travel time on affected routes by 40%–80%.

Table 2 presents performance metrics across all nine methods under these irregular network and stochastic conditions. The results demonstrate HealthNet's ability to maintain performance advantages even when facing the unpredictability and complexity of real-world healthcare networks.

The irregular topology results reveal several insights into HealthNet's adaptive capabilities. First, the performance gap between HealthNet and baseline methods widens under stochastic conditions compared to the regular grid experiments. HealthNet reduces average waiting time by 42.1% compared to IRL and 51.1% compared to MPC, larger margins than observed in controlled grid networks.

Second, peak waiting times during demand surges reveal how different methods handle stress conditions. HealthNet maintains peak times 12.3% lower than IRL and 47.2% lower than MPC. The spatiotemporal feature fusion allows HealthNet to detect early indicators of developing congestion patterns and proactively reallocate resources before facilities become overwhelmed. During vehicle unavailability events, HealthNet's multi-agent coordination enables facilities with surplus capacity to compensate for constrained neighbours, maintaining network-level performance despite local disruptions.

Third, the resource utilisation metric under irregular conditions demonstrates HealthNet's ability to balance efficiency with reliability. While maintaining the highest utilisation rate at 81.3%, HealthNet avoids the brittleness that high utilisation

sometimes introduces. The CTDE framework allows each facility to maintain awareness of network-wide resource states, preventing situations where local optimisation decisions inadvertently create bottlenecks elsewhere in the system. This coordinated resource management becomes particularly valuable during partial road closures, where naive greedy approaches might route multiple ambulances to the same facility despite degraded access, whereas HealthNet distributes load accounting for dynamic accessibility constraints.

The irregular network experiments validate that HealthNet's architecture translates to environments more representative of actual healthcare logistics challenges. The combination of attention-based temporal modelling and graph-based spatial fusion proves robust to topology variations and stochastic perturbations that characterise real-world operations.

The reward function in Equation (7) balances multiple objectives through weighted terms for queue length penalties, ambulance utilisation and resource efficiency. To understand how different reward formulations influence learning and performance, we conducted ablation experiments with five reward variants applied to the 3×3 network configuration. These variants systematically removed or modified specific components to isolate their contributions.

Table 3 presents the performance of HealthNet under different reward function configurations, demonstrating how reward design choices propagate through the learning process to final system behaviour.

The full reward formulation in Equation (7) achieves the best balance across metrics, validating the design choices. When the ambulance utilisation term is removed by setting λ to zero, the average waiting time increases by 10.6% and the convergence slows by 18.3%. Without utilisation incentives, agents adopt conservative policies that keep ambulances idle near facilities rather than proactively positioning them to serve anticipated demand. This conservative behaviour reduces system responsiveness but paradoxically appears safer to the learning algorithm in early training phases, requiring more episodes to discover that active resource positioning improves outcomes.

Removing the resource efficiency term by setting δ to zero degrades performance less severely than eliminating utilisation incentives but introduces greater variance in resource distribution across facilities. The resource balance variance metric quantifies how unevenly resources are allocated, calculated as the standard deviation of resource utilisation rates across all facilities. Without efficiency incentives, facilities sometimes accumulate surplus capacity while neighbours face shortages, as agents lack explicit motivation to maintain balanced network-level resource states. This imbalance manifests primarily during high-demand periods when coordinated resource sharing becomes most valuable.

Applying uniform priority weights rather than differentiated α_k values for different patient priority levels removes the system's ability to appropriately triage responses. Average waiting time increases by 14.9%, disproportionately affecting critical patients who no longer receive preferential resource allocation. The

TABLE 2 | Performance under an irregular network with stochastic disruptions.

Method	Avg waiting time (s)	Peak waiting time (s)	Congestion rate (%)	Resource utilisation (%)
HealthNet	42.3	127.4	21.2	81.3
IRL	49.7	145.2	24.8	78.6
MPMA-MRL	53.1	152.8	26.3	76.4
Smoothing-MP	58.4	168.5	28.9	73.2
ScR-DTG	64.2	182.1	31.5	70.8
GCQN-TSC	69.8	195.3	34.2	68.1
AD-RLTSC	73.5	208.7	36.8	65.7
HMD	78.9	223.6	39.4	62.5
MPC	86.4	241.2	43.7	58.9

TABLE 3 | Ablation study on reward function components.

Reward configuration	Avg waiting time (s)	Convergence episodes	Cumulative reward	Resource balance variance
Full reward	47.3	410	−512.4	0.1
Without utilisation term	52.1	485	−628.7	0.2
Without resource efficiency	49.6	445	−567.3	0.3
Uniform priority weights	54.2	520	−691.5	0.1
Quadratic queue penalty	45.7	390	−485.6	0.1
Linear queue penalty	51.4	560	−634.2	0.2

convergence penalty reflects the additional learning time required when the reward signal fails to distinguish between delay impacts across priority levels, forcing the algorithm to rediscover priority-based scheduling through trial and error.

The quadratic queue penalty experiment modifies the queue length term from squared to cubed, increasing the nonlinearity of congestion penalties. This formulation achieves slightly better average waiting time and faster convergence, as the steeper penalty gradient more aggressively punishes long queues. However, the quadratic formulation in Equation (7) was retained for the main experiments because the cubic penalty occasionally induced oscillatory behaviour where agents over-reacted to temporary queue buildups, creating inefficient resource movements. The quadratic form provides sufficient incentive for queue management while maintaining smoother learning dynamics.

Conversely, linearising the queue penalty to first-order removes the escalating urgency as queues grow. Convergence slows by 36.6% as the learning signal becomes weaker, and final performance degrades substantially. Linear penalties fail to capture the nonlinear relationship between queue length and patient outcomes, where the marginal harm of each additional waiting patient increases as total wait time grows.

These ablation results demonstrate that each component of the reward function contributes meaningfully to overall system performance. The multiplicative interaction between terms creates an objective that guides learning towards policies balancing responsiveness, efficiency and appropriate priority handling. Alternative reward formulations could achieve similar or better performance through different component combinations, but these experiments establish that the chosen formulation represents a well-reasoned design that successfully navigates the multi-objective optimisation challenge inherent in healthcare transportation networks.

6.3 | Practical Deployment Considerations

Although simulation results demonstrate promising performance, transitioning HealthNet from experimental settings to operational healthcare environments introduces practical considerations that warrant attention.

Deploying HealthNet requires establishing unified data pipelines that integrate hospital capacity databases, ambulance GPS tracking, traffic monitoring networks and patient dispatch records. Many healthcare systems currently maintain these data sources in isolated silos with incompatible formats. Implementation would necessitate standardised interfaces, potentially through healthcare information exchange protocols such as HL7 FHIR, with edge computing infrastructure at each facility to meet latency requirements for real-time decision-making.

Rather than replacing existing emergency medical services dispatch platforms, HealthNet would function most effectively as a decision support layer providing routing recommendations to human dispatchers who retain ultimate authority. This hybrid approach addresses liability concerns while allowing gradual trust-building as the system demonstrates reliability over time.

Regulatory requirements vary across jurisdictions. In the United States, systems influencing emergency medical response may require FDA validation studies. European Union deployment must address GDPR requirements and the proposed AI Act's provisions for high-risk healthcare applications. Documentation of decision-making processes becomes necessary for regulatory compliance and post-incident review, with reward function design directly impacting acceptance based on how the system balances efficiency metrics against equity considerations.

Privacy protections must be embedded throughout the architecture. Individual patient health information can be abstracted to priority levels and resource requirements without exposing protected details. Federated learning approaches, where facilities train local model components without sharing raw data, offer pathways to enhance privacy while maintaining system performance.

System reliability for life-critical applications requires graceful degradation, where partial network outages result in reduced optimisation rather than complete failure. Regular retraining cycles must account for distribution shift as healthcare demand patterns evolve, with automated monitoring detecting performance degradation below acceptable thresholds.

7 | Conclusions

This study presented HealthNet, a novel multi-agent reinforcement learning framework for the dynamic optimisation of healthcare transportation networks. The framework addressed the critical challenges of managing complex, time-sensitive healthcare logistics in urban environments. By integrating an SDM with a CTDE approach, HealthNet significantly improved key performance metrics such as patient waiting times, resource utilisation and network congestion management. Simulation results showed that HealthNet consistently outperformed state-of-the-art baseline methods across various scenarios. In the 4×4 network configuration, HealthNet reduced average waiting times by 47.6% compared to the worst-performing baseline and 22.1% compared to the best-performing baseline. The framework's ability to reduce waiting times and congestion while maintaining high resource utilisation efficiency translated to tangible benefits in healthcare delivery, potentially leading to improved patient outcomes and more resilient healthcare systems. The simulations, while extensive, were conducted in controlled environments and may not fully capture the unpredictability of real-world healthcare emergencies. Furthermore, exploring the potential for transfer learning to adapt the model to new geographic areas or healthcare systems with minimal retraining could enhance its scalability.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under No. 62202247.

Funding

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data available on request from the authors.

References

1. J. Lv, A. Slowik, and K. Li, "Secure Output-Feedback Control of Transportation Cyber-Physical Systems for Emergency Medical Services Under Stealthy Attacks," *IEEE Transactions on Intelligent Transportation Systems* 26, no. 9 (September 2025): 14179–14191, https://doi.org/10.1109/TITS.2024.3516937.

2. O. Sabuncu and B. Bilgehan, "Revolutionizing Healthcare 5.0: Blockchain-Driven Optimization of Drone-to-Everything Communication Using 5G Network for Enhanced Medical Services," *Technology in Society* 77 (June 2024): 102552, https://doi.org/10.1016/j.techsoc.2024.102552.

3. J. Lv, A. Slowik, K. Li, and H. Jiang, "Distributed Edge Intelligence for Rapid In-Vehicle Medical Emergency Response in Internet-of-Vehicles," *IEEE Internet of Things Journal* 12, no. 5 (March 2025): 4750–4760, https://doi.org/10.1109/JIOT.2024.3516947.

4. J. Lv, S. Rani, and K. Li, "Intelligent Multi-level Network Optimization for Medical Logistics in Underground Transportation Systems: A Computational Intelligence Approach," *Computers & Industrial Engineering* 209 (November 2025): 1–13, https://doi.org/10.1016/j.cie.2025.111451.

5. Q. Wu, J. Fang, J. Zeng, J. Wen, and F. Luo, "Monte Carlo Simulation-Based Robust Workflow Scheduling for Spot Instances in Cloud Environments," *Tsinghua Science and Technology* 29, no. 1 (February 2024): 112–126, https://doi.org/10.26599/tst.2022.9010065.

6. J. H. Lv, B. G. Kim, B. D. Parameshachari, A. Slowik, and K. Q. Li, "Large Model-Driven Hyperscale Healthcare Data Fusion Analysis in Complex Multi-Sensors," *Information Fusion* 115 (March 2025): 102780, https://doi.org/10.1016/j.inffus.2024.102780.

7. M. Filipovska and H. S. Mahmassani, "Spatio-Temporal Characterization of Stochastic Dynamic Transportation Networks," *IEEE Transactions on Intelligent Transportation Systems* 24, no. 9 (September 2023): 9929–9939, https://doi.org/10.1109/tits.2023.3276190.

8. Y. Matsuo, Y. LeCun, M. Sahani, et al., "Deep Learning, Reinforcement Learning, and World Models," *Neural Networks* 152 (August 2022): 267–275, https://doi.org/10.1016/j.neunet.2022.03.037.

9. Y. Liu, M. Y. Yang, and Z. G. Guo, "Reinforcement Learning Based Optimal Decision Making Towards Product Lifecycle Sustainability," *International Journal of Computer Integrated Manufacturing* 35, no. 10–11 (November 2022): 1269–1296, https://doi.org/10.1080/0951192x.2022.2025623.

10. H. C. Liu, Z. Y. Huang, X. Y. Mo, and C. Lv, "Augmenting Reinforcement Learning With Transformer-Based Scene Representation Learning for Decision-Making of Autonomous Driving," *IEEE Transactions on Intelligent Vehicles* 9, no. 3 (March 2024): 4405–4421, https://doi.org/10.1109/tiv.2024.3372625.

11. Y. Jiang, K. Di, R. Qian, et al., "Optimizing Risk-Aware Task Migration Algorithm Among Multiplex UAV Groups Through Hybrid Attention Multi-Agent Reinforcement Learning," *Tsinghua Science and Technology* 30, no. 1 (April 2024): 318–330, https://doi.org/10.26599/tst.2024.9010013.

12. T. M. Hu, B. Luo, C. H. Yang, and T. W. Huang, "MO-MIX: Multi-Objective Multi-Agent Cooperative Decision-Making With Deep Reinforcement Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 10 (October 2023): 12098–12112, https://doi.org/10.1109/tpami.2023.3283537.

13. C. Yu, J. M. Liu, S. M. Nemati, and G. S. Yin, "Reinforcement Learning in Healthcare: A Survey," *ACM Computing Surveys* 55, no. 1 (January 2023): 1–36, https://doi.org/10.1145/3477600.

14. D. J. Jagannath, R. J. Dolly, G. S. Let, and J. D. Peter, "An IoT Enabled Smart Healthcare System Using Deep Reinforcement Learning," *Concurrency and Computation: Practice and Experience* 34, no. 28 (December 2022): e7403, https://doi.org/10.1002/cpe.7403.

15. A. Z. Al-Marridi, A. Mohamed, and A. Erbad, "Optimized Blockchain-Based Healthcare Framework Empowered by Mixed Multi-Agent Reinforcement Learning," *Journal of Network and Computer Applications* 224 (April 2024): 103834, https://doi.org/10.1016/j.jnca.2024.103834.

16. S. Maheshwari, P. K. Jain, and K. Kotecha, "Route Optimization of Mobile Medical Unit With Reinforcement Learning," *Sustainability* 15, no. 5 (March 2023): 3937, https://doi.org/10.3390/su15053937.

17. A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, and M. Guizani, "Route Optimization of Mobile Medical Unit With Reinforcement Learning," *IEEE Internet of Things Journal* 10, no. 24 (December 2023): 21982–22007, https://doi.org/10.1109/jiot.2023.3288050.

18. E. Akin, "Deep Reinforcement Learning-Based Multirestricted Dynamic-Request Transportation Framework," *IEEE Transactions on Neural Networks and Learning Systems, Early Access* (December 2023).

19. H. R. Su, Y. D. Zhong, J. Y. J. Chow, B. Dey, and L. Jin, "EMVLight: A Multi-Agent Reinforcement Learning Framework for an Emergency Vehicle Decentralized Routing and Traffic Signal Control System," *Transportation Research Part C: Emerging Technologies* 146 (January 2023): 103955, https://doi.org/10.1016/j.trc.2022.103955.

20. X. Fu, Q. F. Nie, X. B. Li, J. Liu, S. Nambisan, and S. Jones, "The Role of the Built Environment in Emergency Medical Services Delays in Responding to Traffic Crashes," *Journal of Transportation Engineering A* 148, no. 10 (October 2022): 04022085, https://doi.org/10.1061/jtepbs.0000726.

21. N. Bagheri, S. Yousefi, and G. Ferrari, "Software-Defined Traffic Light Preemption for Faster Emergency Medical Service Response in Smart Cities," *Accident Analysis & Prevention* 196 (March 2024): 107425, https://doi.org/10.1016/j.aap.2023.107425.

22. D. Noonan, M. Ryan, D. Whelan, and D. O'Neill, "Medical Fitness to Drive, Emergency Service Vehicles and Crash Risk," *Irish Journal of Medical Science* 192, no. 5 (October 2023): 2487–2493, https://doi.org/10.1007/s11845-023-03301-0.

23. T. Y. Zhao, Y. C. Tang, Q. M. Li, and J. Q. Wang, "Resilience-Oriented Network Reconfiguration Strategies for Community Emergency Medical Services," *Reliability Engineering & System Safety* 231 (March 2023): 109029, https://doi.org/10.1016/j.ress.2022.109029.

24. F. Abbracciavento, F. Zinnari, S. Formentin, A. G. Bianchessi, and S. M. Savaresi, "Multi-Intersection Traffic Signal Control: A Decentralized MPC-Based Approach," *IFAC Journal of Systems & Control* 23 (March 2023): 100214, https://doi.org/10.1016/j.ifacsc.2022.100214.

25. L. Z. Wang, Z. L. Ma, C. Y. Dong, and H. Wang, "Human-Centric Multimodal Deep (HMD) Traffic Signal Control," *IET Intelligent Transport Systems* 17, no. 4 (April 2023): 744–753, https://doi.org/10.1049/itr2.12300.

26. D. Yu, S. G. Wei, and L. G. Chai, "Traffic Signal Control in Mixed Traffic Environment Based on Advance Decision and Reinforcement Learning," *Transportation Safety and Environment* 4, no. 4 (November 2022): tdac027, https://doi.org/10.1093/tse/tdac027.

27. L. P. Yan, L. L. Zhu, K. Song, et al., "Graph Cooperation Deep Reinforcement Learning for Ecological Urban Traffic Signal Control,"

Applied Intelligence 53, no. 6 (March 2023): 6248–6265, https://doi.org/10.1007/s10489-022-03208-w.

28. M. Y. Xu, T. Z. Qiu, J. Fang, H. Y. He, and H. T. Chen, "Signal-Control Refined Dynamic Traffic Graph Model for Movement-Based Arterial Network Traffic Volume Prediction," *Expert Systems with Applications* 228 (October 2023): 120393, https://doi.org/10.1016/j.eswa.2023.120393.

29. T. Xu, S. Barman, and M. W. Levin, "Smoothing-MP: A Novel Max-Pressure Signal Control Considering Signal Coordination to Smooth Traffic in Urban Networks," *Transportation Research Part C: Emerging* 166 (September 2024): 104760, https://doi.org/10.1016/j.trc.2024.104760.

30. S. J. Huang, C. N. Sun, R. Q. Wang, and D. Pompili, "Toward Adaptive and Coordinated Transportation Systems: A Multi-Personality Multi-Agent Meta-Reinforcement Learning Framework," *IEEE Transactions on Intelligent Transportation Systems* 26, no. 8 (August 2025): 12148–12161, https://doi.org/10.1109/tits.2025.3560227.

31. Y. H. Wang, Y. He, F. R. Yu, K. S. Wu, and S. Z. Chen, "Intelligence-Based Reinforcement Learning for Dynamic Resource Optimization in Edge Computing-Enabled Vehicular Networks," *IEEE Transactions on Mobile Computing* 24, no. 3 (March 2025): 2394–2406, https://doi.org/10.1109/tmc.2024.3506161.

32. W. X. Zhang, T. Zhao, Z. K. Zhao, Y. J. Wang, and F. R. Liu, "An Intelligent Strategy Decision Method for Collaborative Jamming Based on Hierarchical Multi-Agent Reinforcement Learning," *IEEE Transactions on Cognitive Communications and Networking* 10, no. 4 (August 2024): 1467–1480, https://doi.org/10.1109/tccn.2024.3373640.

33. Y. Ma, J. Q. Li, Y. F. Hu, and H. Chen, "A Battery Prognostics and Health Management Technique Based on Knee Critical Interval and Linear Complexity Self-Attention Transformer in Electric Vehicles," *IEEE Transactions on Intelligent Transportation Systems* 25, no. 8 (August 2024): 10216–10230, https://doi.org/10.1109/tits.2024.3355436.

34. A. Ala, V. Simic, D. Pamucar, and N. Bacanin, "Enhancing Patient Information Performance in Internet of Things-Based Smart Healthcare System: Hybrid Artificial Intelligence and Optimization Approaches," *Engineering Applications of Artificial Intelligence* 131 (May 2024): 107889, https://doi.org/10.1016/j.engappai.2024.107889.

35. X. Li, W. H. Xu, T. Q. Wang, and Y. Yuan, "Optimizing Integrated Eco-Driving Control and Holding Strategy for Real-Time Bus Bunching Mitigation," *IEEE Transactions on Intelligent Transportation Systems* 25, no. 7 (July 2024): 7568–7582, https://doi.org/10.1109/tits.2024.3362345.