

Predicting Traffic Flow Using Enhanced Long Short-Term Memory

Hailong Dong

Lecturer, Department of Traffic Management Engineering, Henan Police College, Zhengzhou, 450046, China, E-mail: dong.dhl@outlook.com

Project Management

Received January 21, 2026; revised April 17, 2026; accepted September 5, 2026
Available online October 5, 2026

Abstract: In response to the urgent need for real-time, accurate traffic prediction in urban congestion, this study proposes an enhanced Long Short-Term Memory (LSTM)-based approach that leverages collaboration with cloud platforms. First, based on the LSTM model benchmark, a Convolutional Neural Network (CNN) is used to extract local spatial patterns, a double-layer LSTM is used to capture long-term temporal dependencies, and an attention mechanism is introduced to dynamically weight traffic characteristics over time. Second, the urban transportation network is modeled to better analyze vehicle management. Subsequently, local signal adaptive control is combined with regional signal collaborative control, and a hierarchical collaborative strategy for signal control that considers traffic flow balance across the road network is proposed. The results showed that the proposed method achieved a Root Mean Square Error (RMSE) of 15.8 vehicles per hour (veh/h) within 15 minutes and a Mean Absolute Percentage Error (MAPE) of 6.2%, both of which were significantly better than those of the comparison models. An ablation experiment confirmed that the CNN and attention mechanism reduced prediction errors by 18% and 12%, respectively. The proposed signal control strategy effectively reduced average delay time and reduced the number of overflows by 72%. The research method provides a scalable, high-precision, end-to-end solution for cloud-edge-integrated traffic prediction and collaborative management and control.

Keywords: Cloud platform; long short-term memory (LSTM); convolutional neural network (CNN); attention mechanism; traffic flow; signal control; prediction.

Copyright © Journal of Engineering, Project, and Production Management (EPPM-Journal).

DOI 10.32738/JEPPM-2026-50

1. Introduction

With the acceleration of global urbanization, traffic congestion has become a key bottleneck that hampers the sustainable development of cities. The increasingly severe traffic congestion has caused a series of social, economic, and environmental challenges, seriously affecting the quality of life and urban development (Lai et al., 2023). Enhancing the real-time accuracy of Traffic Flow Prediction (TFP) has become a core challenge in intelligent transportation systems. It can provide key data to support active management strategies, such as traffic guidance and light signal timing optimization (Al-Asaly et al., 2022; Bing et al., 2025). Currently, technologies such as fog computing, Internet of Vehicles (IoV), and big data have enabled the development of collaborative intelligent transportation systems characterized by mobile Internet and vehicle-road collaboration. Traditional cloud computing-based intelligent transportation systems offer clear advantages in deployment flexibility and scalability, but they still face challenges such as long end-to-end delays, traffic congestion, difficulty processing massive amounts of data, and high communication costs (Mallreddy and Vasa, 2023; Dogani et al., 2023). Accurate TFP is the basis for route planning, vehicle dispatching, signal control, and other applications. Currently, TFP methods in various countries mainly focus on three aspects: parametric, non-parametric, and hybrid models. For example, traditional recurrent neural networks often struggle to fully capture the complex spatiotemporal interweaving, leading to limited predictive accuracy (Ramamoorthi, 2023). The existing timing control system often ignores real-time traffic information, which can easily cause congestion. Single-intersection control lacks linkage, making it difficult to achieve overall coordination.

Therefore, this study aims not only to propose a high-precision technical algorithm but also to develop a comprehensive decision-support framework for urban operations and management. By integrating advanced deep learning predictive models with hierarchical collaborative signal control strategies, this study explores how data-driven intelligence can be translated into executable traffic engineering management practices within a cloud-edge collaborative architecture. Therefore, this study uses Long Short-Term Memory (LSTM) as the core and enhances it with a Convolutional Neural Network (CNN) and an Attention Mechanism (AM) to propose a CNN-Attention-LSTM model on a cloud computing platform. This model aims to deeply mine the inherent spatiotemporal laws in traffic data and achieve higher-precision

predictions. To address the challenges of traditional intelligent transportation systems, traffic signal flow feature capture and information control are implemented to better alleviate urban traffic congestion. The construction and evaluation of this framework aims to provide a new paradigm for the management and system optimization of transportation infrastructure projects in the context of smart cities, demonstrating how to improve the operational efficiency and reliability of public service systems through technological innovation, and providing a reference for system performance improvement and resource optimization allocation in the field of engineering management.

2. Literature Review

Faced with insufficient short-term TFP performance in smart cities and the difficulty of supporting intelligent urban traffic management, Zhou et al. (2023) introduced edge computing into the 5G IoV environment and integrated a deep convolutional Random Forest (RF) network to build a prediction model. The average delay cost was low, the average offloading utility was stable at around 70%, and recognition accuracy exceeded 95%. Smart city TFP ignores distributed data privacy. Yuan et al. (2023) developed a Federated Learning (FL) algorithm based on graph representation. They set up three major modules: long-term cycle capture, attention federation, and semantic capture within the local prediction model of the edge computing server. This algorithm could effectively implement TFP. In a smart city TFP, multi-party sensor data is stored and distributed in a zero-trust manner, and privacy issues hinder data sharing and integration. Wang et al. (2023) designed a privacy-preserving TFP method based on local sensitive hashing. This method achieved high predictive precision and efficiency while ensuring the privacy of sensor data. To address the fact that the mainstream TFP model does not fully utilize the node characteristics of the transportation network, thereby affecting prediction accuracy, Chen et al. (2023) first proposed a node connection strength index that uses nodes' historical transportation attributes. A Graph Convolutional Network (GCN) based on the exponential matrix was designed and paired with a new dynamics extractor to learn traffic flow features. The performance was shown to be better than that of the baseline method.

Intelligent traffic management systems need to accurately predict traffic congestion in real-time. Using data collected by the edge-layer cache server, Kishk et al. (2022) proposed a three-stage traffic congestion prediction strategy that includes congestion-direction detection and width-change decision-making. This strategy achieved 95% accuracy and only 5% error, outperforming other recent methods. In edge computing-enabled IoV, uneven spatiotemporal distribution of traffic flows can easily lead to a surge in load on edge servers, and the lack of reasonable computing offloading solutions between servers causes additional energy consumption and delays. Xu et al. (2023) proposed a GCN to mine relationships among road segments for accurate TFP estimation. Then, the deep deterministic policy gradient was used to drive prediction-based computational offloading methods, thereby obtaining the optimal solution. Comparative experiments demonstrated that this model had low prediction errors and superior performance in reducing total delay and energy consumption. Existing TFP research struggles to model the complex structural relationships of traffic networks due to insufficient dynamic correlations among spatial, temporal, and periodic features. Duan et al. (2022) proposed a fully dynamic, spatiotemporal GCN framework with self-attention based on Deep Learning (DL). Three components were designed: spatial and temporal graph attention and a fusion layer. Different periodic flow characteristics were designed. This framework had advantages in prediction accuracy, stability, and adaptability. Faced with the challenge that semantic correlation is ignored in TFP and centralized data training can easily lead to privacy leaks, Liu et al. (2023) proposed an intelligent TFP model based on multi-level FL, integrating a spatiotemporal graph DL model into the FL framework. This model could reduce communication overhead and improve prediction accuracy.

Past content generally relies on edge-cloud architecture. Improving prediction accuracy and privacy protection, node importance, semantic association, and dynamic spatiotemporal self-attention have become hot topics within the frameworks of 5G, FL, and graph convolution. However, spatially relevant road-segment screening and real-time signal feedback are not considered simultaneously, making it difficult for the model output to directly drive signal coordination. Although significant progress has been made in prediction accuracy, privacy protection, or dynamic spatiotemporal modeling in previous studies, there are generally two major limitations: First, most models focus on "prediction", and their outputs are difficult to directly and seamlessly transform into real-time signal cooperative control strategies that consider network topology. Secondly, although there is discussion of cloud-edge collaboration, few studies integrate spatial-correlation road screening, high-precision prediction, and hierarchical signal control into a unified, scalable, end-to-end framework for systematic design and verification. Therefore, to address the gap in previous research, a scalable, end-to-end cloud-edge collaboration paradigm is developed based on the LSTM model and cloud platforms.

3. Construction of the Cloud Computing TFP Model

3.1. TFP Model Construction Based on Enhanced LSTM

In the urban road network, the traffic conditions of a road section are significantly affected by its adjacent road sections. To quantify this spatial dependence, the study uses the Spearman correlation coefficient to measure the strength of correlation between traffic flow sequences across different road sections. The correlation coefficient R_{XY} between road sections X and road section Y is shown in Eq. (1) (Wang et al., 2024).

$$R_{XY} = 1 - \frac{6 \sum_{k=1}^N (R_{Xk} - R_{Yk})^2}{N(N^2 - 1)} \quad (1)$$

In Eq. (1) R_{Xk} and R_{Yk} are the ranks of X and Y at the k -th moment. N is the sample length. By calculating the correlation coefficient between the target predicted road section and all surrounding road sections, the road sections with the strongest correlation can be identified, forming a spatially related set of road sections (Xu et al., 2022). Afterward,

a spatiotemporal 2D matrix is constructed from this set of traffic flow data at consecutive time steps as input to the benchmark model. Eq. (2) is the input matrix M_t of the training sample.

$$M_t = \begin{bmatrix} F_t^{R1}, F_t^{R2}, \dots, F_t^{Rm} \\ f_t^{R1}, f_t^{R2}, \dots, f_t^{Rm} \\ f_{t-1}^{R1}, f_{t-1}^{R2}, \dots, f_{t-1}^{Rm} \\ \dots \dots \dots \dots \\ f_{t-n}^{R1}, f_{t-n}^{R2}, \dots, f_{t-n}^{Rm} \end{bmatrix} \quad (2)$$

In Eq. (2), f_{t-n}^{Rm} is the traffic flow of target road section n at time, with a total of m road sections. The rows of the M_t matrix represent the spatial dimension (different relevant road segments), and the columns represent the time dimension (historical time series). The matrix M_t is normalized to eliminate dimensionality effects and improve model training efficiency and stability. The LSTM model is used as the benchmark to implement TFP. To effectively extract spatial features from the spatiotemporal matrix M_t . This study treats the traffic value of the relevant road section as a 1D vector at each time step and slides it with a 1D CNN convolution kernel. Convolution operations capture local spatial correlation patterns between adjacent road sections, such as the flow-conduction effect of upstream sections on downstream sections (Alqarni, 2022). Afterward, after multi-layer convolution and activation function processing, the CNN module converts the original spatiotemporal matrix M_t into a feature sequence $Z = [z_1, z_2, z_3, \dots, z_p]$ at time step p . Each z_p is a feature vector that deeply condenses the spatial information at that moment. Then, the LSTM module is used to implement time-dynamic modeling and extract traffic-flow features. The input gate i_t , forget gate f_t , and output gate o_t at time t under the LSTM mechanism are shown in Eq. (3) (Xie et al., 2023).

$$\begin{cases} f_t = \sigma(M_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(M_i \cdot [h_{t-1}, x_t] + b_i) \\ o_t = \sigma(M_o \cdot [h_{t-1}, x_t] + b_o) \end{cases} \quad (3)$$

In Eq. (3) h_{t-1} means the output and structural information of the previous layer. h_t denotes the output, σ indicates the gate activation function, and x_t is the input value. M, b are the weight matrix and value of bias. After multiple processing steps, the LSTM module outputs a series of hidden states. This state can capture the evolution of traffic flow's spatial characteristics over time. Fig.1 is the structure diagram of the CNN-LSTM.

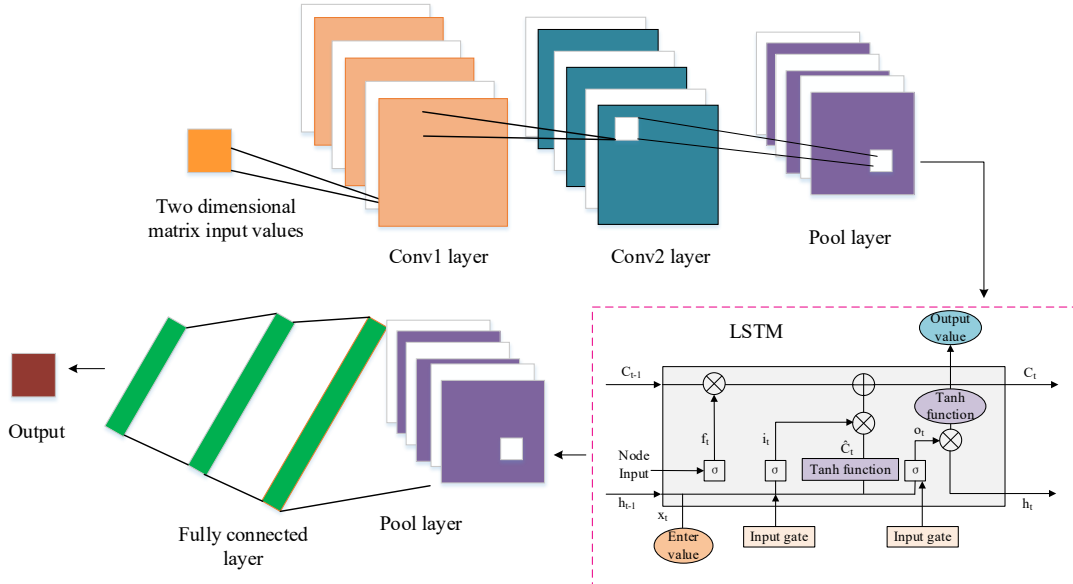


Fig. 1. Structure of the CNN-LSTM

In Fig.1, the input 2D matrix value is processed by CNN convolution, followed by two layers of convolution operations to obtain the local receptive field, and then passes through the fully connected layer. The results are fed into an LSTM for feature extraction, and the hidden state at each step is computed. In traffic prediction tasks, different historical moments do not have an equal impact on future traffic. Therefore, this study introduces AM into LSTM to dynamically allocate weights, allowing it to "pay attention" to the most important part of the sequence. Eq. (4) is the score expression of the weight α_i of AM (Paital, 2024).

$$\alpha_i = \frac{\exp(s_i)}{\sum_{i=1}^n \exp(s_i)} \quad (4)$$

In Eq. (4), s_i is the weight score in the traffic flow sequence. α_i is the time attention weight at the time step i . Fig.2 shows the CNN-LSTM framework combined with AM.

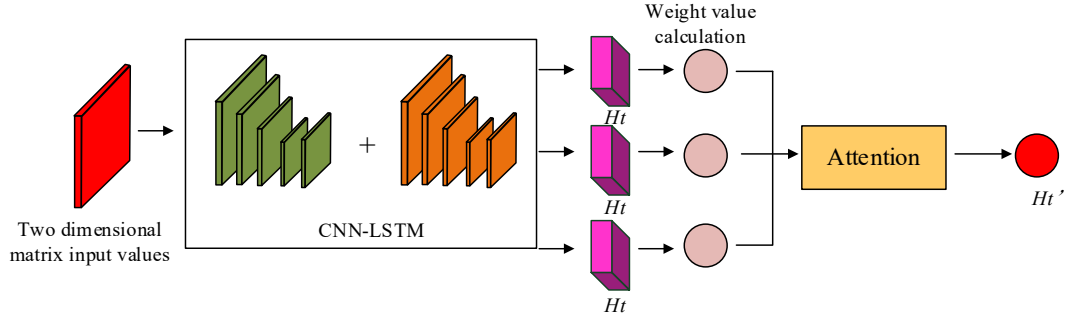


Fig. 2. Schematic diagram of NN-LSTM combined with AM

In Fig. 2, after preliminary processing of the traffic flow on the target road section, road sections with higher correlation are screened, and a 2D spatiotemporal traffic flow matrix is developed. The matrix is input into CNN-LSTM for convolution processing and hidden state calculation. The output sequence state uses the attention layer to automatically fit the weight distribution and calculate the spatiotemporal characteristics of different weights. After that, the spatiotemporal characteristics of the traffic flow are fused, and the prediction value at the next moment is finally output. Attention weight is not a hyperparameter that needs to be manually set. It is dynamically learned by the attention mechanism during model training, based on the input data. It enables the model to autonomously assign different importance to input features at different time steps, thereby capturing key information more intelligently.

3.2. Signal Control Based on Traffic Flow Balance in Road Network on Cloud Platform

Different management areas exhibit varying traffic flow conditions due to differences in social attributes and road network topology, which inevitably lead to higher accident rates. Therefore, based on TFP, this study designs an information control method under the cloud platform management architecture. This study first models the urban transportation network and sets it as a directed graph $G = (V, E)$. V is the intersection set, E means the road set, and the urban area set is $R = \{R_1, R_2, \dots, R_M\}$. Within the time interval Δt . The dynamic change of the number of vehicles is expressed in Eq. (5) (Yan et al., 2024).

$$Veh_q(t + \Delta t) = Veh_q(t) + \sum_p In_{q,p}(t) - \sum_g Out_{q,p}(t) \quad (5)$$

In Eq. (5), $In_{q,p}(t)$ and $Out_{q,p}(t)$ represent the number of vehicles entering and leaving R_q from area R_p at time t . $Veh_q(t)$ denotes the number of vehicles in R_q at time t . This study uses road pheromones to measure road congestion. Road pheromone represents the density of traffic flow on the road. The higher the value, the higher the density of traffic flow. The slower the vehicle travels, the more pheromones it leaves behind. If the mean vehicle velocity reaches the maximum limit, no pheromone will be left behind (Gao et al., 2023). When a car enters a certain area, the cloud server will record the vehicle information and the entry and exit times, and obtain its driving trajectory via communication. With the help of historical driving data, the traffic flow $M_t^{OD(q,p)}$ of the traffic flow starting point OD can be constructed. Its traffic flow matrix is expressed as Eq. (6) (Mei et al., 2023).

$$M_t^{OD(q,p)} = Num_{q \rightarrow p}(t) \quad (6)$$

In Eq. (6), by dynamically fusing the real-time matrix M_t^{OD} and the historical matrix M_t^{F-1} , the final traffic flow matrix can be obtained, as shown in Eq. (7).

$$M_t^F = \omega \cdot M_t^{OD} + (1 - \omega) \cdot M_t^{F-1} \quad (7)$$

In Eq. (7), ω is the weight. The M_t^F matrix reflects the long-term travel connections between regions and helps guide traffic flow more accurately under congestion conditions. Afterwards, this study uses dynamic correlation to measure the correlation between two divided areas, as shown in Eq. (8) (Hasanvand et al., 2023).

$$\Gamma_{(u,v),k}(t) = \sum_{j \in R} \alpha_{k,j}(t) \cdot \beta_{(u,v)} \quad (8)$$

In Eq. (8), the dynamic and static correlations between road (u,v) and area k are $\alpha_{k,j}(t)$ and $\beta_{(u,v)}$. Considering

differences in traffic light control across regions, this study designs a control strategy based on hierarchical collaboration from two perspectives: local and global. Fig. 3 shows the hierarchical collaborative control algorithm for light signals.

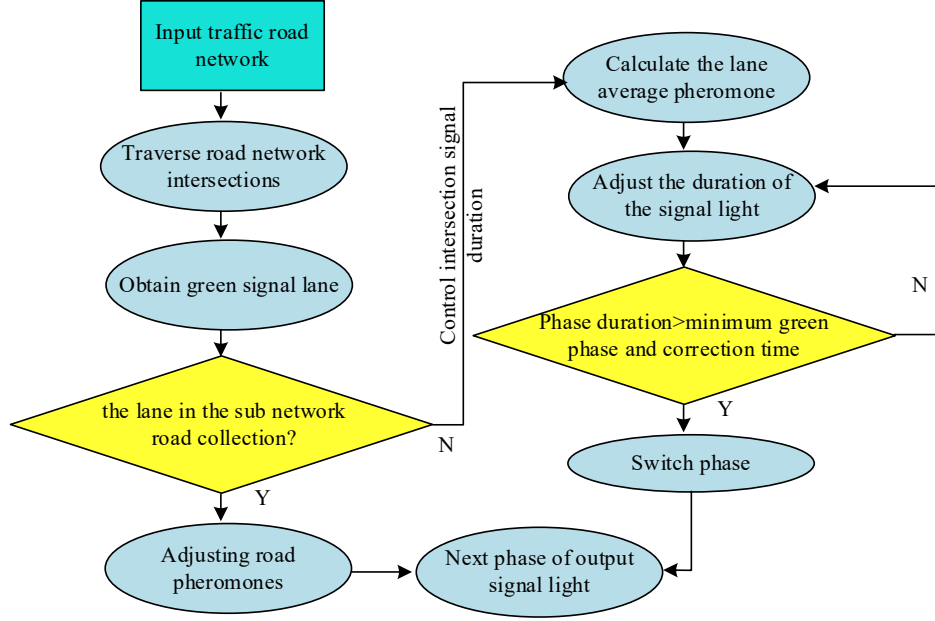


Fig. 3. Schematic diagram of light signal hierarchical collaborative control algorithm

In Fig. 3, road pheromones across different roads are adjusted to implement light signal control using a global approach, thereby reducing the duration of unnecessary light signals. If the acquired lane is not in the regional sub-network set, the relative duration of the intersection signal is adjusted using local signal control. Local signal control can use real-time lane pheromones to obtain dynamic adjustment information for adjusting signal phase durations. Specifically, at a single intersection, the green light duration is adjusted according to pheromones, as shown in Eq. (9).

$$T_{green}^{(p)}(t) = T_{base}^{(p)} + \alpha_c (\Phi_p(t) - \bar{\Phi}(t)) \quad (9)$$

In Eq. (9), $T_{base}^{(p)}$ is the baseline green light duration based on the Webster method. The α_c means the adjustment coefficient. $\bar{\Phi}(t)$ denotes the average pheromone of each entrance of the intersection. Local control is suitable for unsaturated traffic flow and quickly balances the pressure of each entrance of a single node. When it is detected that a traffic incident affects road a , the affected road set C_a is screened based on the shortest path, and its signal adjustment amplitude is calculated based on the correlation between the road and the incident road. Dynamically compressing the green light duration of high-correlation roads can prevent traffic flow from flooding into congested areas, and Eq. (10) is obtained.

$$\delta(l, t) = \tau(l, t) - \beta \frac{\tau(a, t) \cdot S_{Aa}(t)}{len(l, a)} \quad (10)$$

In Eq. (10), $\delta(l, t)$ is the adjusted pheromone of road l at time t . β indicates the weight parameter, τ is the fusion pheromone, and Aa is the area where the road is located. $S_{Aa}(t)$ is the correlation degree, and len is the shortest path length. Under the hierarchical collaborative combination strategy, in the normal state, each intersection performs locally adaptive signal control in response to local traffic. When a time trigger is detected, the cloud computing layer will perform collaborative control across regions to reduce the probability of traffic congestion. To better capture the spatiotemporal characteristics of traffic flow, this study relies on a cloud computing platform to implement the TFP model, from data preprocessing and feature engineering to LSTM model enhancement. Fig. 4 shows the TFP architecture on the cloud platform.

In Fig.4, the TFP includes a terminal layer, a cloud layer and a fog layer. The terminal layer consists of road-end geomagnetic sensors and vehicle-mounted units, which collect traffic parameters in real time. The fog layer deploys fog servers and roadside units in key traffic areas to handle data processing and local prediction. The cloud layer centrally stores global traffic history data and simultaneously performs traffic prediction. The cloud platform collects raw traffic flow data from sensors deployed across various road sections and filters out noise. Then, spatial correlation analysis is used to construct a 2D matrix that represents the spatiotemporal characteristics of traffic flow, which serves as input to the improved LSTM.

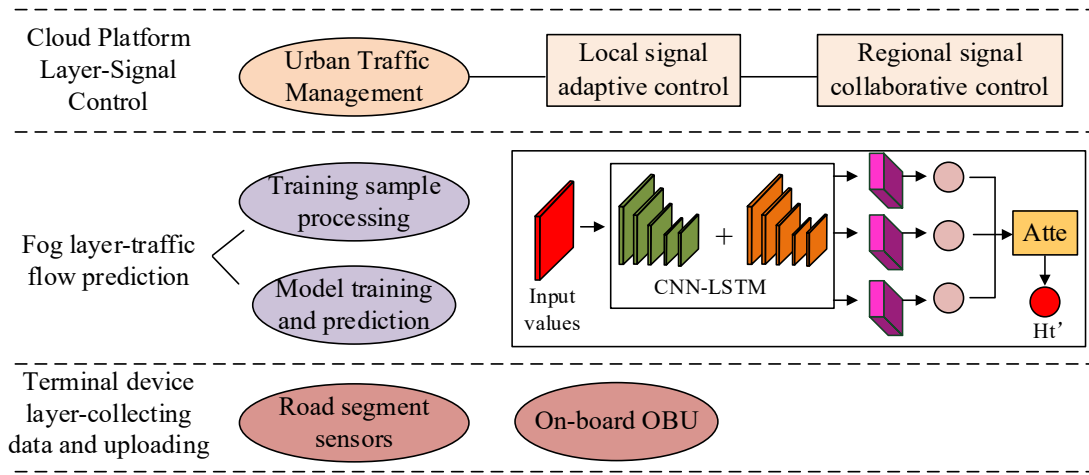


Fig. 4. TFP architecture based on cloud platform

4. Cloud Computing TFP Result Analysis

4.1. TFP Performance Evaluation of Enhanced LSTM Algorithm

To better evaluate the CNN-Attention-LSTM, predictive analysis is performed on a real dataset. This dataset is obtained through the Performance Measurement System sponsored by the California Department of Transportation. The dataset covers 17 weeks, and the traffic data obtained is used as the dataset. This paper uses indicators such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and training time to evaluate the algorithm's performance. Table 1 shows the configurations.

Table 1. Experimental environment and parameter settings

Server	Cloud server
Cloud nodes	16 vCPU Intel 8358 @ 2.6 GHz, 64 GB RAM, Ubuntu 20.04
Fog Node	8 vCPU, 16 GB RAM, deployed at roadside MEC (5G micro base station)
Batch size	1,024
Initial learning rate	1×10^{-3}
Dropout	0.15
Loss function	Huber- $\delta=1.0$
Optimizer	AdamW
Development language	Python3.7
DL Framework	Pytorch1.9.1

The filtering threshold for spatially related road segments directly affects the dimensionality and quality of the input feature matrix. A low threshold may introduce excessive noise, while a high threshold may result in the loss of useful secondary spatial information. The study compared the effects of different thresholds on the predictive performance (measured by RMSE) and computational complexity. The weight parameter α determines the compression strength of the duration of high-correlation green lights during congestion events. Sensitivity analysis was conducted in the SUMO simulation environment to determine its value. The correlation threshold and sensitivity results are shown in Table 2.

The results in Table 2 show that selecting 0.6 as the threshold achieves the best balance, ensuring low prediction error while moderately controlling the computational burden. When the pheromone control parameter is set to 100, the system performs best at reducing the average delay and overflow frequency in bottleneck areas. A value that is too small cannot effectively alleviate congestion, while a value that is too large may cause unnecessary congestion in the upstream road network. The proposed enhanced LSTM prediction model is compared with the Genetic Algorithm - Backpropagation Network (GA-BP), GCN - Transformer (GCN-Trans), Asynchronous Dilated Graph Convolutional Network (ADGCN), Particle Swarm Optimization - RF (PSO-RF), and PSO-LSTM models. Table 2 shows the traffic prediction results at different time periods.

In Table 3, at 15 minutes, CNN-Attention-LSTM performs best, with RMSE of 15.8 vehicles per hour (veh/h) and MAPE of 6.2%. As the prediction time interval grows, the errors of each model show an overall upward trend. The RMSE of CNN-Attention-LSTM at 30 minutes increases to 23.7 veh/h, and the MAPE increases to 8.5%. At 60 minutes, the RMSE further increases to 35.2 veh/h, and the MAPE increases to 10.2%, but the prediction error remains smaller than that of other methods. The prediction errors of ADGCN and GCN-Trans at each time interval are relatively low; their

maximum MAPE does not exceed 15%, and their performance is relatively stable. The prediction errors of PSO-LSTM, PSO-RF, and GA-BP are relatively high. Their MAPE values at 60 minutes reach 15.6%, 17.9%, and 20.3%, and the overall error increase is more obvious. The traffic prediction performance of the CNN-Attention-LSTM is better than that of other models, and the improvement becomes more pronounced as the time interval increases. Two road sections are selected to compare the performance of the above algorithm in extracting spatiotemporal traffic-flow features, as shown in Fig. 5.

Table 2. Correlation threshold and sensitivity results

Correlation threshold						
Threshold	Enter the number of road segments		RMSE (veh/h, 15min)		MAPE (% , 15min)	
0.4	15	16.5	6.5	0.4	15	16.5
0.6	9	15.8	6.2	0.6	9	15.8
0.8	4	17.1	6.9	0.8	4	17.1
Sensitivity results						
Pheromone control parameter values— β	Average delay (s/km)		Average queue length (m)		Overflow frequency (times/h)	
50	35.1		89		3.8	
100	32.3		84		3.1	
150	34.5		92		3.5	
200	37.8		98		4.2	

Table 3. Traffic predictions of different algorithms at different time periods

Model	15 min		30 min		60 min	
Indicator	RMSE (veh/h)	MAPE (%)	RMSE (veh/h)	MAPE (%)	RMSE (veh/h)	MAPE (%)
CNN-Attention-LSTM	15.8	6.2	23.7	8.5	35.2	10.2
ADGCN	16.5	6.8	24.8	9.2	36.9	12.8
GCN-Trans	17.2	7.1	26.1	9.8	38.4	13.5
PSO-LSTM	18.9	8.3	28.5	10.9	42.7	15.6
PSO-RF	22.4	9.7	31.2	12.4	47.3	17.9
GA-BP	25.1	11.5	35.8	15.1	53.6	20.3

In Fig. 5(a), during the morning peak (7 o'clock - 10 o'clock), when the recall is 0.2, the precision of CNN-Attention-LSTM is 0.95, GCN-Trans is 0.92, PSO-LSTM is 0.88, ADGCN is 0.85, PSO-RF is 0.82, and GA-BP is 0.8. When the recall reaches 0.8, the precision of each algorithm drops to approximately 0.60, 0.55, 0.50, 0.48, 0.45, and 0.42. In Fig.5(b), the trend of the evening peak (17 o'clock - 19 o'clock) is similar to that of the morning peak, and the precision of each algorithm at the same recall is slightly lower. The recall value of CNN-Attention-LSTM (0.87) is higher than that of ADGCN, GCN-Trans (0.84), PSO-LSTM (0.83), PSO-RF (0.82), and GA-BP (0.81). In Fig. 5(c), when the recall is 0.2, the precision of CNN-Attention-LSTM is nearly 1. When the recall is 0.8, it is about 0.85, which is better than that of other comparison algorithms. Morning and evening peak traffic flow is complex, and the CNN-Attention-LSTM and GCN-Trans can better capture spatiotemporal characteristics and perform better. GA-BP and PSO-RF are insufficient for processing complex scene features and yield poor predictive performance. The CNN-Attention-LSTM is used for ablation experiments, as shown in Fig. 6.

In Fig. 6(a), at different time intervals (5, 10, 15, and 20 min), the MAE value of CNN-Attention-LSTM is the lowest, indicating that its prediction error is the smallest. LSTM has the relatively highest MAE value. At 20 minutes, the MAE of LSTM is 22 veh/h, CNN-LSTM is 18 veh/h, and CNN-Attention-LSTM is 16 veh/h. In Fig. 6(b), at 10 minutes, the RMSE of LSTM is about 20 veh/h, CNN-LSTM is 17 veh/h, and CNN-Attention-LSTM is about 16 veh/h, indicating that adding CNN and AM can improve model performance. In Fig. 6(c), CNN-Attention-LSTM has the relatively lowest prediction error at each time interval, which is smaller than that of LSTM and CNN-LSTM. In TFP, CNN can capture spatial dependencies between adjacent road segments or regions, identify traffic change patterns in local areas, and provide the model with richer spatial information. AM can automatically learn the importance of weights of different time steps or features. It allows the model to focus on key time points or features while ignoring irrelevant information, thereby more

accurately capturing changes in flow. CNN supplements spatial features, AM optimizes attention to key information, and together they synergize with LSTM's ability to process time series to comprehensively capture the spatiotemporal characteristics of traffic flow, effectively reducing prediction errors. The TFP results of the CNN-Attention-LSTM are compared with the real values, as shown in Fig. 7.

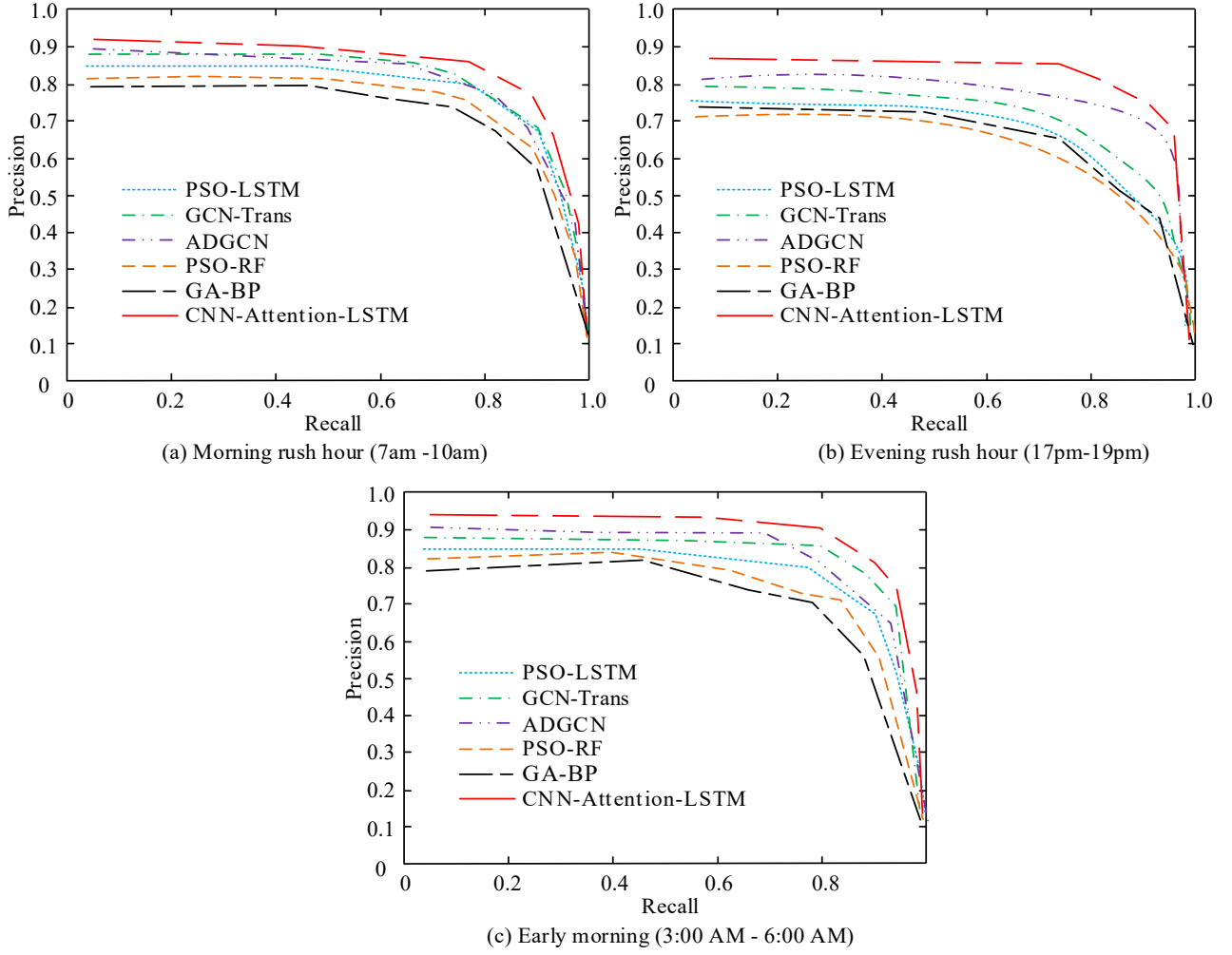


Fig. 5. Performance of extracting spatiotemporal features of traffic flow

In Fig. 7, the comparison between the TFP results of the CNN-Attention-LSTM and the real value at five-minute and ten-minute intervals shows that the prediction curve (dashed line) better fits the changing trend of the real value curve (solid line). During peak and through traffic periods, the predicted values can be addressed in a timely manner. Although there is some deviation, the overall deviation does not exceed 5%, which is competitive for the application.

4.2. Cloud Platform TFP Result Inspection

Based on the SUMO platform, 25,000 vehicle flows are randomly generated. The number of roads is set to 300, the length of a single road is 200 m, the number of traffic lights is 100, the time step is 3s, and the weight is 100. The proposed light signal system control scheme is analyzed, as shown in Fig. 8.

In Fig. 8(a), in the first 600 s in the central area, the number of cars under the Webster scheme is higher than under the single-control and collaborative-control schemes, and this trend is clear. When the time exceeds 600 s, the traffic flow under the collaborative control scheme is significantly reduced, indicating that the scheme effectively alleviates road congestion. Fig. 8(b) and Fig. 8(c) show that the number of cars in the collaborative control scheme is higher than in the other two schemes, thereby delaying the increase in vehicle congestion and reducing the number of vehicles in the later period. The performance of the three schemes is analyzed, as listed in Table 4.

In Table 4, regional collaboration dynamically compresses green lights through "pheromone-correlation", effectively preventing upstream traffic from continuing to pour into the bottleneck. Its average delay time does not exceed 40 s/km, and it effectively reduces queuing time and overflow occurrences. The Webster scheme and single control show obvious delays, with average values of 68.4 s/km and 52.7 s/km, respectively. Table 5 shows the application performance.

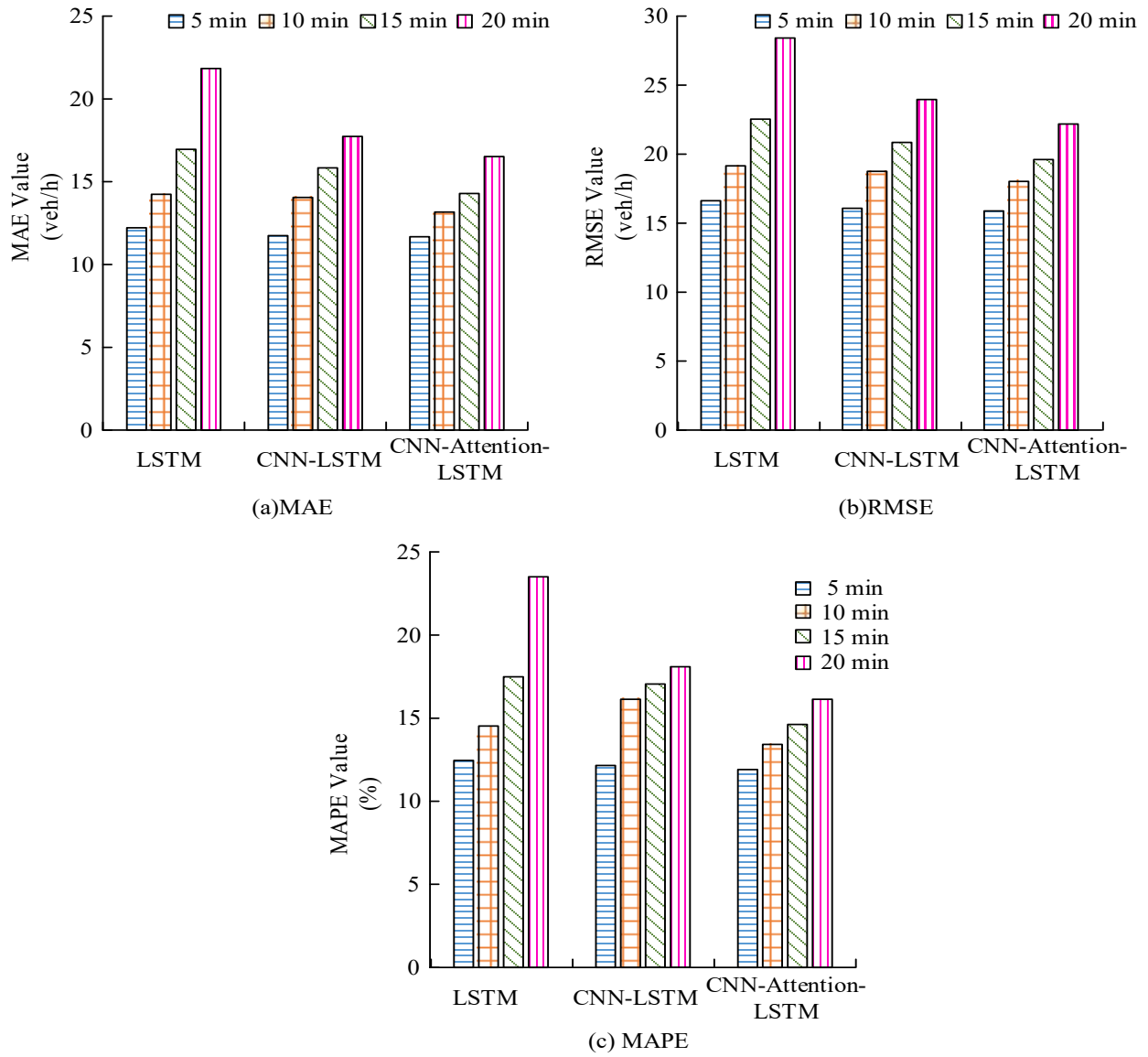


Fig. 6. Enhanced LSTM ablation results

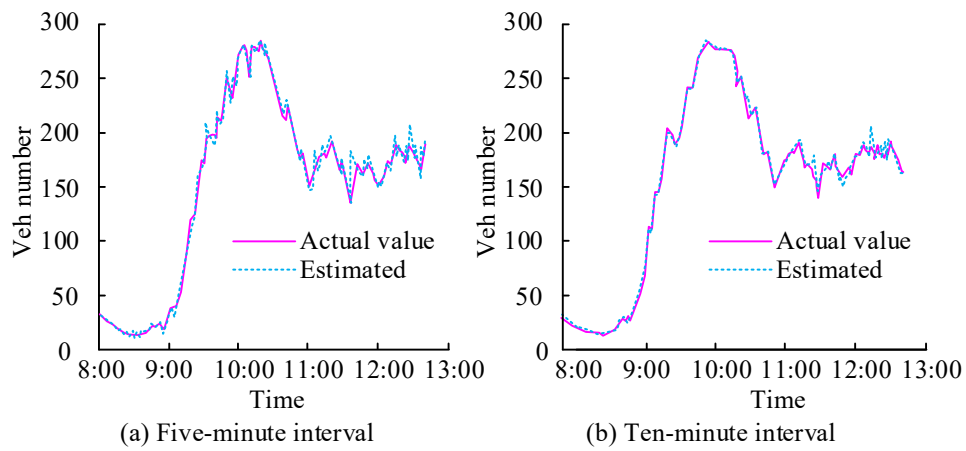
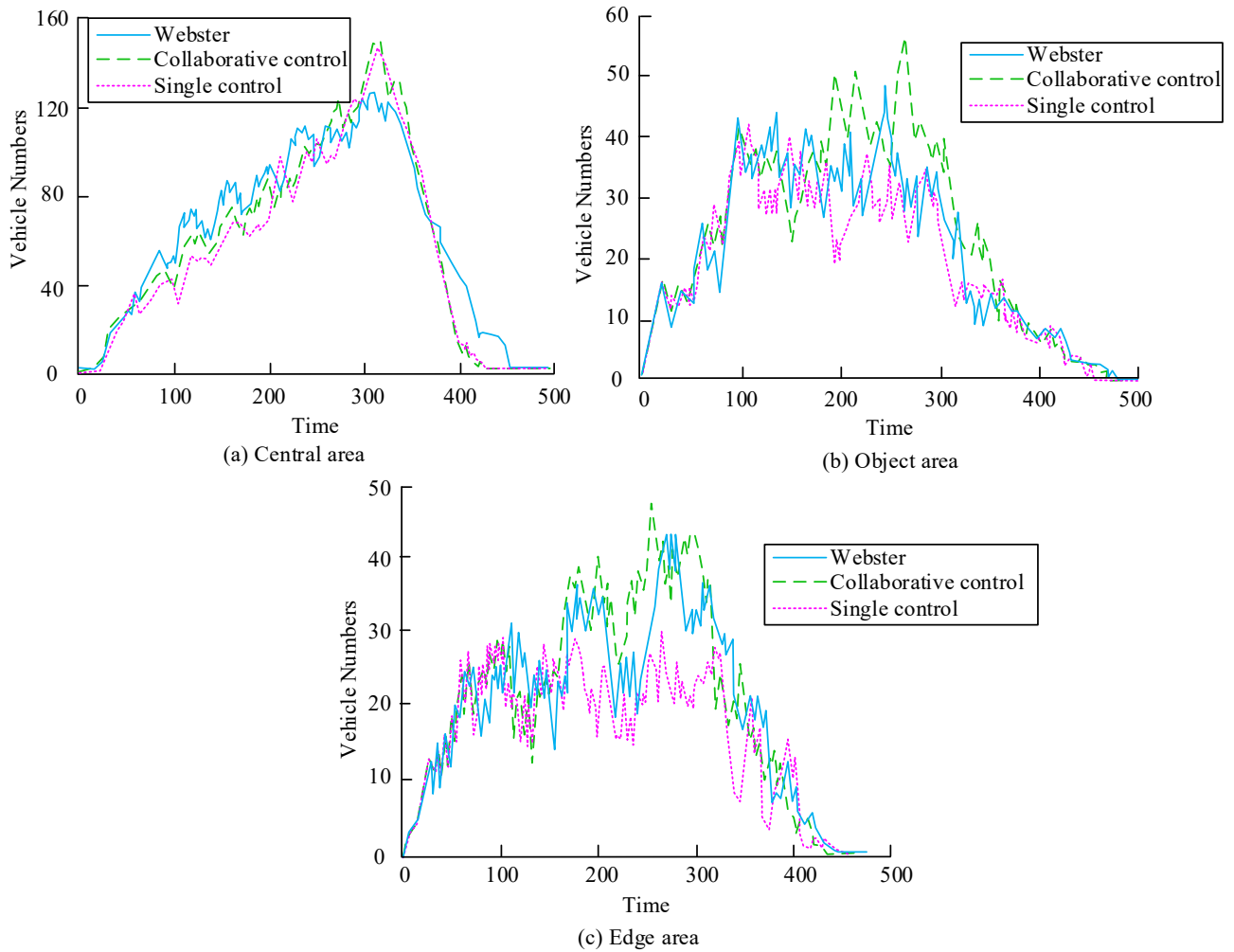


Fig. 7. Comparison between predicted and actual traffic flow values

**Fig. 8.** Control effect of light signal system**Table 4.** Application performance of three solutions

Indicator	Webster	Single control	Proposed
Average delay (s/km)	68.4±2.1	52.7±1.6	32.3±1.3
Average queue (m)	137±5	108±4	84±3
Overflow frequency (times/h)	11.2±0.7	6.3±0.5	3.1±0.4
Fuel emissions (g/km)	284±9	241±7	198±6

Table 5. Application efficiency of signal control strategies in traffic flow

Indicator	Traditional timing control	Inductive control	Proposed
Regional average travel time (s)	360	320	290
Average delay time at intersections (s)	125	110	95
Average speed on the road network (km/h)	22.5	25.0	28.0
Queue length at key intersections (m)	280	240	200
Traffic volume (vehicles/hour)	1,800	1,900	2,050
Number of parking times (times/vehicle·kilometer)	2.8	2.3	1.7
Fuel consumption reduction rate (%)	-	-	12.5
Duration of congestion (minutes/event)	45.0	38.0	25.0
Control strategy response delay (s)	N/A	15-30	<5
System reliability (% of normal operating time)	98.5	97.0	99.5

In Table 5, the control strategy based on cloud platform design reduces the regional average travel time by 19.4% compared with traditional control. Active prediction and coordination eliminate "waiting for green lights" and "queue overflow," significantly improving traffic efficiency. The average delay at the intersection decreases by 24.0% compared with traditional control, and traffic volume increases by 13.9%. The proposed method effectively reduces response delay (<5 s) and improves system reliability (uptime ratio is 99.5%). This method significantly shortens the recovery time for sudden congestion events and relies on the cloud platform architecture to enhance the anti-interference capability and reliability of the entire control system. To test the generalization ability and scalability of the proposed framework in different urban environments, three typical regional networks with different traffic characteristics were constructed in the SUMO simulation platform: 1) Central Business District (CBD): with obvious tidal traffic characteristics and a dense road network. 2) Residential area: The peak hours in the morning and evening are prominent, but the flow during the off-peak period is relatively low. 3) Mixed zone: including commercial, residential, and main roads, with complex transportation modes. The collaborative control strategy was tested on these networks separately. The results are shown in Table 6.

Table 6. Performance of the model in different simulated urban scenarios

Scenario	Average delay reduction rate (%)	Average queue length reduction rate (%)	Regional traffic capacity improvement rate (%)
Raw dataset scenario	52.8%	38.7%	13.9%
CBD	48.5%	35.1%	11.2%
Residential area	55.2%	41.5%	15.3%
Mixed zone	46.1%	33.6%	10.5%

From Table 6, although the strategy's performance (e.g., delay reduction rate) slightly decreases in more complex CBD and mixed-zone scenarios, significant optimization effects are still achieved (with reduction rates exceeding 46%). In residential areas with relatively simple transportation modes, their performance is even better than the original scene. This indicates that the proposed CNN-Attention-LSTM prediction model and hierarchical collaborative control strategy exhibit satisfactory generalization and robustness and can adapt to urban areas with varying topological structures and traffic demands.

5. Conclusion

From the perspective of engineering systems management, a cloud-based CNN-Attention LSTM traffic prediction and control framework has been successfully developed and validated for controlling traffic flow and light signal. This provides a closed-loop management solution for urban traffic operation managers, from data collection, intelligent prediction, and real-time control. This end-to-end systematic approach transforms isolated traffic prediction tasks into a dynamically intervened and optimized operational process, providing new management tools and ideas for solving complex urban congestion challenges. In the experiment, the predicted RMSE of CNN-Attention-LSTM dropped to 15.8 veh/h, which was 8.1% lower than GCN-Trans and 16.4% lower than PSO-LSTM. The predicted RMSE at 60 minutes was 35.2 veh/h, which was 4.6% lower than ADGCN's. Compared with the Webster timing scheme, the average delay decreased from 68.4 s/km to 32.3 s/km, a reduction of more than 40%, and the average queue length decreased, with the overflow frequency decreasing by 72.3%. The regional traffic flow increased from 1,800 veh/h to 2,050 veh/h, a 13.9% increase, thereby reducing the system response delay. However, this research still has shortcomings: the LSTM hyperparameters must be manually tuned, potentially affecting performance on cross-provincial road networks. In addition, when predicting traffic flow, unstructured information such as weather, accident texts, and related extreme emergency scenarios is not integrated. Based on the research framework, managers can use high-precision short-term forecasting results to identify potential congestion bottlenecks in advance, implement preventive signal-timing adjustments or issue warnings, and shift from passive response to active intervention. The predicted results can dynamically deploy public resources such as traffic police and evacuation equipment, prioritize their allocation to areas where congestion is imminent, maximize resource utilization efficiency, assist managers in formulating the optimal traffic evacuation path, and shorten the impact time of events. The multiple indicators provided by the research framework can serve as a quantitative basis for evaluating the effectiveness of traffic management strategies in different regions and time periods. Through long-term analysis of data generated by this system, policymakers can accurately identify normalized congestion nodes and bottleneck sections in the urban transportation network, providing robust data support for investment decisions in infrastructure projects such as road expansion and intersection improvements. The construction and operation of cloud platforms require significant investment. The government can draw on research frameworks, design public-private partnership models, encourage technology companies to participate in the construction and operation of smart transportation infrastructure, and improve the transportation data governance framework. Future research will introduce multi-modal fusion neural networks that integrate unstructured information such as weather, traffic accident texts, social media emotions, etc., and combine them with event knowledge graphs to enhance the model's emergency adaptability to extreme emergencies through zero or few sample learning. Simultaneously, a distributed self-learning and federated transfer learning mechanism is designed to enable the model to perform lightweight self adjustment and hyperparameter optimization on edge nodes, while safely drawing on model knowledge from other similar cities, thereby providing more flexible and scalable large-scale traffic prediction and control services for smart cities. The study presents a clear path for traffic managers to utilize existing and future technologies to transform their decision-making models from passive responses to historical experience and lagging data to active management based on real-time prediction and system optimization. Managers can deploy similar systems

to achieve early warning of traffic congestion, intelligent, collaborative control of traffic lights, and rapid emergency response, thereby enabling more effective, refined decisions in daily operations and ultimately improving the operational efficiency and service level of the entire urban transportation network.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors confirm that no AI tools were used in the preparation of this manuscript.

References

- Al-Asaly, M. A., Bencherif, M. A., Alsanad, A., and Hassan, M. M. (2022). A deep learning-based resource usage prediction model for resource provisioning in an autonomic cloud computing environment. *Neural Computing and Applications*, 34(13), 10211–10228. doi: 10.1007/s00521-021-06665-5.
- Alqarni, A. A. (2022). Majority vote-based ensemble approach for distributed denial of service attack detection in cloud computing. *Journal of Cyber Security and Mobility*, 11(2), 265–278. doi: 10.13052/jcsm2245-1439.1126.
- Bing, L., Jian, P. G., Ling, X., Hong, H. W., and Xiao, H. Z. (2025). Digital twin traffic flow prediction based on graph federated learning in intelligent IoT. *China Communications*, 22(7), 290–305. doi: 10.23919/JCC.fa.2024-0014.202507.
- Chen, J., Wang, W., Yu, K., Hu, X., Cai, M., and Guizani, M. (2023). Node connection strength matrix-based graph convolution network for traffic flow prediction. *IEEE Transactions on Vehicular Technology*, 72(9), 12063–12074. doi: 10.1109/TVT.2023.3265300.
- Dogani, J., Khunjush, F., Mahmoudi, M. R., and Seydali, M. (2023). Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism. *The Journal of Supercomputing*, 79(3), 3437–3470. doi: 10.1007/s11227-022-04782-z.
- Duan, Y., Chen, N., Shen, S., Zhang, P., Qu, Y., and Yu, S. (2022). FDSA-STG: Fully dynamic self-attention spatio-temporal graph networks for intelligent traffic flow prediction. *IEEE Transactions on Vehicular Technology*, 71(9), 9250–9260. doi: 10.1109/TVT.2022.3178094.
- Gao, B., Chen, Q., Liu, Y., Wan, K., and Li, K. (2023). Predictive cruise control under cloud control system for urban bus considering queue dissipation time. *IEEE Transactions on Intelligent Vehicles*, 8(4), 2639–2649. doi: 10.1109/TIV.2023.3235352.
- Hasanvand, M., Nooshyar, M., Moharamkhani, E., and Selyari, A. (2023). Machine learning methodology for identifying vehicles using image processing. *Artificial Intelligence and Applications*, 1(3), 170–178. doi: 10.47852/bonviewAIA3202833.
- Kishk, A. M., Badawy, M., Ali, H. A., and Saleh, A. I. (2022). A new traffic congestion prediction strategy (TCPS) based on edge computing. *Cluster Computing*, 25(1), 49–75. doi: 10.1007/s10586-021-03377-2.
- Lai, Q., Tian, J., Wang, W., and Hu, X. (2023). Spatial-temporal attention graph convolution network on edge cloud for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(4), 4565–4576. doi: 10.1109/TITS.2022.3185503.
- Liu, L., Tian, Y., Chakraborty, C., Feng, J., Pei, Q., Zhen, L., and Yu, K. (2023). Multilevel federated learning-based intelligent traffic flow forecasting for transportation network management. *IEEE Transactions on Network and Service Management*, 20(2), 1446–1458. doi: 10.1109/TNSM.2023.3280515.
- Mallreddy, S. R. and Vasa, Y. (2023). Predictive maintenance in cloud computing and DevOps: ML models for anticipating and preventing system failures. *NVEO-Natural Volatiles & Essential Oils Journal*, 10(1), 213–219. doi: 10.53555/nveo.v10i1.5751.
- Mei, P., Karimi, H. R., Huang, C., Chen, F., and Yang, S. (2023). Remaining driving range prediction for electric vehicles: Key challenges and outlook. *IET Control Theory & Applications*, 17(14), 1875–1893. doi: 10.1049/cth2.12486.
- Paital, S. (2024). Cloud IoT platforms for smart city traffic optimization. *Optimality*, 1(2), 300–308. doi: 10.22105/opt.v1i2.61.
- Ramamoorthi, V. (2023). Applications of AI in cloud computing: Transforming industries and future opportunities. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(4), 472–483. doi: 10.32628/CSEIT2390910.
- Wang, B., He, Y., Shui, Z., Jin, Q., and Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. *Applied and Computational Engineering*, 64(1), 89–94. doi: 10.54254/2755-2721/64/20241350.
- Wang, F., Li, G., Wang, Y., Rafique, W., Khosravi, M. R., Liu, G., and Qi, L. (2023). Privacy-aware traffic flow prediction based on multi-party sensor data with zero trust in smart city. *ACM Transactions on Internet Technology*, 23(3), 1–19. doi: 10.1145/3511904.
- Xie, P., Ma, M., Li, T., Ji, S., Du, S., Yu, Z., and Zhang, J. (2023). Spatio-temporal dynamic graph relation learning for urban metro flow prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(10), 9973–9984. doi: 10.1109/TKDE.2023.3269771.
- Xu, X., Jiang, Q., Zhang, P., Cao, X., Khosravi, M. R., Alex, L. T., and Dou, W. (2022). Game theory for distributed IoT task offloading with fuzzy neural network in edge computing. *IEEE Transactions on Fuzzy Systems*, 30(11), 4593–4604. doi: 10.1109/TFUZZ.2022.3158000.

- Xu, X., Yang, C., Bilal, M., Li, W., and Wang, H. (2023). Computation offloading for energy and delay trade-offs with traffic flow prediction in edge computing-enabled IoV. *IEEE Transactions on Intelligent Transportation Systems*, 24(12), 15613–15623. doi: 10.1109/TITS.2022.3221975.
- Yan, H., Xu, X., Bilal, M., Xia, X., Dou, W., and Wang, H. (2024). Customer centric service caching for intelligent cyber-physical transportation systems with cloud-edge computing leveraging digital twins. *IEEE Transactions on Consumer Electronics*, 70(1), 1787–1797. doi: 10.1109/TCE.2023.3326969.
- Yuan, X., Chen, J., Yang, J., Zhang, N., Yang, T., Han, T., and Taherkordi, A. (2023). FedSTN: Graph representation driven federated learning for edge computing enabled urban traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(8), 8738–8748. doi: 10.1109/TITS.2022.3157056.
- Zhou, S., Wei, C., Song, C., Pan, X., Chang, W., and Yang, L. (2023). Short-term traffic flow prediction of the smart city using 5G internet of vehicles based on edge computing. *IEEE Transactions on Intelligent Transportation Systems*, 24(2), 2229–2238. doi: 10.1109/TITS.2022.3147845.



Hailong Dong was born in 1987. He is a native of Xuchang, Henan Province. He received his bachelor's degree in Transportation Engineering from Lanzhou Jiaotong University, where he studied from 2006 to 2010, and his master's degree in Transportation Planning and Management from the same institution, completed in 2013. He has served as a lecturer in the Department of Traffic Management Engineering at Henan Police College since April 2014.