

Research on Traffic Congestion Area Identification and Traffic Control Based on Multi-source Data Fusion

Han Bai

Shandong Jiaotong University

Haoyu Chen

Shandong Jiaotong University

Liang Zhao

Shandong Jiaotong University

Tianzheng Wei

`weitianzheng0220@163.com`

Shandong Jiaotong University

Jian Lu

Shandong Jiaotong University

Xiuguang Wang

Shandong Zhengqu Traffic Engineering Co., Ltd

Article

Keywords: Traffic congestion identification, Traffic control, Multi-source data, Deep learning

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10797349/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Research on Traffic Congestion Area Identification and Traffic Control Based on Multi-source Data Fusion

Han Bai¹, Haoyu Chen¹, Liang Zhao¹, Tianzheng Wei^{1,*}, Jian Lu¹, Xiuguang Wang²

¹ School of Transportation and Logistics Engineering, Shandong Jiaotong University, Jinan 250357, China;

² Shandong Zhengqu Traffic Engineering Co., Ltd, Jinan, China.

* Corresponding author: E-mail: weitianzheng0220@163.com

Abstract: With the acceleration of urbanization, traffic congestion has become a major issue affecting urban operational efficiency and residents' travel experience. In this context, accurately identifying recurrent traffic congestion areas, predicting their dynamic evolution trends, and formulating effective traffic flow optimization strategies have become critical to improving traffic management levels. This study combines geomagnetic traffic flow data, vehicle trajectory data, and mobile phone signaling data to propose a multi-source data-based approach for traffic congestion identification, prediction, and control. A grid-based partitioning method combined with the DBSCAN clustering algorithm is employed to refine the urban road network and accurately identify recurrent congestion areas. A Bayesian network is introduced to analyze congestion diffusion and quantify the propagation probability of traffic congestion across road segments. Furthermore, a hybrid CNN-BiLSTM model based on the self-attention mechanism (SA-CNN-BiLSTM) is proposed, which leverages the self-attention mechanism to dynamically weight key features, thereby improving prediction accuracy. The results show that the SA-CNN-BiLSTM model reduces the prediction error (MAE) by 23% compared with traditional prediction models and demonstrates stronger adaptability and robustness during peak hours, effectively predicting the evolution trends of traffic congestion. The proposed methods are validated using the Hongjialou area of Jinan as a case study, confirming their effectiveness. The findings provide data-driven decision support for urban traffic management authorities, facilitating the transformation of traffic management from experience-driven to intelligent decision-making and enhancing the operational efficiency of urban transportation.

Keywords: Traffic congestion identification; Traffic control; Multi-source data; Deep learning

1. Introduction

Against the backdrop of rapid urbanization, the population of urban residents and the total number of automobiles have been increasing dramatically. According to the Ministry of Public Security, the national motor vehicle ownership reached 453 million units in 2024, with the number of licensed drivers exceeding 542 million. The enormous traffic demand has placed tremendous pressure on urban road networks. Commuting congestion has evolved into a common governance challenge across metropolitan areas and has become a critical issue for improving urban travel satisfaction. In response to traffic congestion, concepts such as "smart cities" and "intelligent transportation" have attracted widespread attention in recent years. Multi-source data-based traffic congestion analysis provides more comprehensive and precise decision-making support for traffic management. For instance, trajectory data can be analyzed to identify traffic flow variation trends, providing support for traffic optimization and management [1]; real-time traffic data can be combined to adjust signal timing, thereby improving road capacity; and big data analytics can be employed to optimize urban road planning and travel routes, reducing road network load [2]. How to leverage multi-source data to accurately identify traffic congestion areas and optimize traffic management has become an important research topic.

This paper proposes a multi-source data fusion approach for traffic congestion identification, prediction, and control, constructing a traffic congestion management framework based on deep learning and optimization algorithms. In addition, a simulated annealing algorithm is employed to design a peak-hour traffic flow allocation strategy, achieving overall optimization of the road network. The main objectives of this study are as follows:

(1) To achieve congestion area identification through a road grid-based analysis method, accurately locating recurrent traffic congestion areas in cities;

(2) To combine the identification results of recurrent congestion areas with relevant event definitions, conduct in-depth analysis of the spatiotemporal evolution patterns of congestion, and design a short-term traffic congestion prediction model accordingly.

1.1 Paper Organization

The remainder of this paper is organized as follows. Section 2 summarizes the related work on traffic flow prediction and congestion identification. Section 3 describes the data processing and multi-source data fusion methods. Section 4 introduces the deep learning methods for traffic congestion identification, congestion propagation, and congestion prediction. Section 5 presents the experimental results and discussion of the traffic flow prediction model and traffic congestion area identification. Finally, Section 6 concludes the paper.

2. Literature Review

This section discusses the relevant literature on data processing, congestion identification, congestion propagation, and congestion prediction techniques in the context of traffic flow prediction.

2.1 Current Research on Traffic Data Mining

By integrating multi-source data such as onboard locators, geomagnetic sensors, and video surveillance, and employing machine learning and graph mining techniques for modeling and analysis, traffic data mining research aims to improve the accuracy and real-time performance of traffic flow prediction, congestion identification, and intelligent traffic control. Goshi Sato et al. [3] utilized data collected by onboard sensors to determine road operating conditions and transmitted the information in real time to vehicles and roadside servers within the area through V2X communication networks, enabling drivers to accurately assess the current driving environment and providing technical support for the development of intelligent transportation systems. Hu et al. [4] partitioned the road network into regular grids based on the location information of traffic data and mapped the number of traffic accidents in each grid to pixel values of an image, achieving effective integration of data with spatial maps and providing a new method for spatial distribution analysis of traffic accidents.

2.2 Current Research on Traffic Congestion Identification

To improve the accuracy and timeliness of traffic congestion identification, researchers have comprehensively employed advanced traffic flow theory, data mining techniques, and intelligent optimization algorithms to determine the evolution patterns of congestion areas and optimize traffic flow scheduling strategies. Chiabaut et al. [5] proposed a new method for real-time estimation of highway traffic conditions and travel time by applying Gaussian mixture models and k-means algorithms to cluster observation days in historical datasets. Gore et al. [6] characterized congestion levels using a probability-based hazard function approach and generated reliability-based highway level of service (LOS) thresholds, proposing a new measure called the Congestion Index (CI) to assist in quantifying the intensity and duration of congestion. Drliciac et al. [7] considered both horizontal and vertical characteristics of traffic flow for capacity assessment, concluding that when road capacity is exceeded or nearly reached, traffic flow "always" decreases significantly compared to when capacity is not exceeded. Liu, Bowie et al. [8] proposed a graph-based, low-cost traffic forecasting and congestion detection framework, along with a threshold-based method for detecting traffic congestion using traffic volume and image-based traffic occupancy.

2.3 Current Research on Traffic Congestion Prediction

To improve the accuracy of traffic congestion prediction and its adaptability to complex environments, researchers have comprehensively employed deep learning models (such as spatiotemporal graph convolutional networks and Transformers) alongside traditional methods to effectively capture the spatiotemporal dependencies of traffic flow. Deng et al. [9] utilized two-dimensional convolution operators to explore spatiotemporal relationships and proposed a uniformly designed improved random search sequential optimization deep convolutional network that achieves

satisfactory prediction even with incomplete data. Chen et al. [10] predicted traffic flow by integrating gradient boosting regression trees with the least absolute shrinkage operator, achieving good prediction accuracy even under abnormal traffic conditions. Li et al. [11] proposed an improved support vector regression model for short-term traffic flow prediction, optimizing SVM parameters to obtain the optimal linear combination and effectively improving prediction accuracy. Liu, Yutian et al. [12] introduced TC-Predictor, a novel neural network architecture integrating congestion-condition adaptive graph convolutional networks (GCN) to capture both global road network structure and local dynamic patterns, dynamically adjusting the adjacency matrix based on congestion levels at each prediction step. They also introduced a data-driven metric to emphasize high-flow scenarios and employed the Sparrow Search Algorithm (SSA) for hyperparameter optimization. Salamanis, Athanasios et al. [13] provided accurate predictions under both normal and abnormal conditions. Their model uses the density-based clustering algorithm DBSCAN to identify traffic patterns under normal and abnormal conditions, and applies different prediction models to each individual cluster representing a traffic pattern. Nearest neighbor and support vector regression algorithms from machine learning, as well as the ARIMA model from time series analysis, were trained and tested.

2.4 Current Research on Traffic Control

In the field of traffic control, researchers have comprehensively employed boundary control, dynamic partitioning, and signal optimization methods, analyzing network macroscopic fundamental diagrams (MFD) and vehicle queue lengths to improve system stability and control precision. The introduction of reinforcement learning has further enhanced the adaptability and real-time performance of control strategies. Yan et al. [14] proposed a dynamic traffic subarea partitioning algorithm to achieve dynamic updating of traffic subareas. Zhao et al. [15] utilized reinforcement learning strategies for signal control optimization, employing a phase-duration-aware reinforcement learning method in their signal timing scheme. Xu et al. [16] distinguished between public transit and private vehicles during model construction, proposing a bus signal priority optimization strategy for signal control. Budholiya et al. [17] proposed using the YOLOv5 algorithm, a Greedy-based Genetic Algorithm (GGA), Deep CNN, and LSTM for traffic monitoring and congestion control in VANETs. The YOLOv5 algorithm was used for vehicle detection with high speed and precision. GGA was employed for feature selection to handle numerous features in the data and improve computational speed. Deep CNN with LSTM was used for classification to process larger datasets and improve accuracy.

Current research on traffic flow prediction and congestion management exhibits a trend toward deep integration of multiple technologies. At the data level, real-time perception is achieved through onboard sensors and V2X communication, with grid-based mapping enhancing spatial analysis capabilities. In congestion identification and evaluation, quantitative methods based on clustering, hazard functions, and graph structures have been developed. The core advancement in prediction models lies in the use of deep learning (such as CNN, GCN, and LSTM) to more accurately capture spatiotemporal dynamic features, with clustering and parameter optimization improving adaptability to abnormal and complex scenarios. Control strategies are evolving toward dynamic and intelligent approaches, such as real-time optimization through reinforcement learning and bus priority signaling. Overall, research is transitioning from single data sources or models toward closed-loop intelligent systems integrating perception, analysis, prediction, and control. This paper focuses on urban recurrent congestion, constructing a high-quality dataset through data cleaning and parameter calibration, and employing grid-based and clustering methods to identify congestion areas. Based on traffic flow theory, the congestion evolution is decomposed into four stages: initiation, propagation, persistence, and dissipation, with its mechanisms analyzed from spatiotemporal dimensions. Subsequently, a prediction model integrating the self-attention mechanism with CNN-BiLSTM is constructed to achieve accurate short-term congestion early warning. Finally, a simulated annealing algorithm is used to optimize macroscopic traffic control strategies, forming a complete technical closed loop of "identification-analysis-prediction-regulation" to provide scientific decision support for urban traffic governance.

3. Data

3.1 Data Description

The data used in this study include vehicle trajectory data, geomagnetic data, and mobile phone signaling data. The online car-hailing trajectory data are collected in real time through the BeiDou Navigation Satellite System, containing fields such as administrative district code, latitude and longitude, and origin-destination points. After map matching, these data are used for traffic flow analysis and congestion prediction (see Table 1). The mobile phone signaling data are derived from communication interactions between mobile phones and base stations, which can reflect user movement trajectories and behavioral patterns, and are used for travel pattern and peak demand analysis (see Table 2). The geomagnetic data are collected through geomagnetic sensors, which can monitor vehicle distribution and driving status, and are used for evaluating road segment traffic flow and optimizing signal control (see Table 3). The three types of data, after processing and transformation, provide a multi-dimensional foundation for traffic research.

Table 1 Online car-hailing trajectory sample

Administrative Region	License Plate Number	Waiting Time	Longitude	Latitude	Arrival Area
370104	鲁 A****01	57s	117.0035	36.6985	Quanying Dream Factory Hotel
370103	鲁 A****83	28s	117.2140	36.7068	Longfeng Mountain Road Feiyue Avenue
370102	鲁 A****97	21s	117.0212	36.6568	Orange Hotel - Southwest Gate
370112	鲁 A****18	35s	117.0378	36.7053	Xian Hotel (Huangtai Branch)
370105	鲁 A****68	47s	117.1303	36.6557	Huachuang Guanli Center
...	...	...	...	...	...

Table 2 Mobile phone signaling data statistics

Mobile Positioning Time	Longitude	Latitude
20230702103313	117.0709	36.6631
20230702103313	117.0659	36.6719
20230702103312	117.0691	36.6662
20230702103308	117.0584	36.6793
20230702103308	117.0733	36.6807
20230702103306	117.0560	36.6749
20230702103300	117.0388	36.6790
...	...	...

Table 3 Geomagnetic data statistics

Date	Time	North Entrance	North Exit	East Entrance	East Exit	South Entrance	South Exit	West Entrance	West Exit
------	------	----------------	------------	---------------	-----------	----------------	------------	---------------	-----------

03-01	00:05	0	13	0	29	24	8	31	6
03-01	00:10	0	16	0	27	23	6	31	6
03-01	00:15	0	14	0	24	22	7	32	9
03-01	00:20	0	11	0	12	17	7	20	8
03-01	00:25	0	7	0	17	19	4	18	9
03-01	00:30	0	19	0	32	29	15	48	11
...	...	...	...	...	...	...	...	...	...

3.2 Data Preprocessing

To ensure data consistency, data alignment was first performed by converting the latitude and longitude coordinates from vehicle trajectories and mobile phone signaling into the unified WGS84 coordinate system. Subsequently, data cleaning was conducted with differentiated deduplication strategies for the three types of raw data (GPS-based spatiotemporal thresholds, window-based merging for geomagnetic data, and ID-grouped time aggregation for signaling data), outlier removal (standard deviation method for GPS, signal strength thresholds for signaling and geomagnetic data), and linear interpolation for missing value imputation, thereby improving data continuity and accuracy. Finally, trajectory smoothing and Savitzky-Golay filtering (combined with the moving average method) were applied to remove noise from the data, reducing fluctuations caused by geomagnetic signal errors or random jitter.

3.3 Multi-source Data Fusion

First, spatiotemporal alignment and data association were performed by standardizing all data timestamps to a uniform 15-minute interval and converting geographic coordinates to the WGS84 coordinate system. Meanwhile, the area enclosed by Lishan Road, Jiefang Road, Erhuan East Elevated Road, and Huayuan Road was selected as the study area (see Figure 1). In the multi-source feature parameter extraction, vehicle driving speeds were extracted from trajectory and geomagnetic data and aggregated as average speeds per unit time; traffic flows were extracted from geomagnetic and signaling data; and population flows were estimated from signaling data. These were combined with dynamic parameters such as speed and acceleration to comprehensively identify congestion areas.



Figure 1 Study Area

In the weight allocation-based data fusion, information entropy was first calculated, followed by the computation of difference coefficients and their normalization to obtain weights. Grey relational analysis was then combined to determine the final weights. Finally, the geomagnetic data, trajectory

data, and signaling data were weighted and fused to form a unified traffic congestion dataset.

Information entropy calculation

$$H_i = - \sum_{j=1}^n p_{ij} \log p_{ij} \quad (1)$$

Where p_{ij} is the probability of the j feature in the i sample.

Difference coefficient calculation and normalization

$$d_i = 1 - H_i \quad (2)$$

$$w_i = \frac{d_i}{\sum_{k=1}^m d_k} \quad (3)$$

where m is the total number of features.

(3) Grey relational analysis and weighted fusion

The weights are determined by calculating the degree of association between features and the reference sequence.

$$\rho_{ij} = \frac{\min \gamma_{ij}}{\max \gamma_{ij}} \quad (4)$$

where γ_{ij} is the absolute difference between the i feature and the reference sequence at the j sample.

$$X_{\text{fused}} = \sum_{i=1}^m w_i X_i \quad (5)$$

where X_i is the value of the i feature and w_i is its corresponding weight.

(4) Assume the features of geomagnetic data D_m , trajectory data D_g , and mobile signaling data D_s are $\{X_1, X_2, \dots, X_n\}$, with corresponding weights w_m , w_g , and w_s , respectively. Their feature values are combined through weight assignment to form a unified traffic congestion dataset.

$$D_f = w_m \cdot D_m + w_g \cdot D_g + w_s \cdot D_s \quad (6)$$

Among them, $w_m + w_g + w_s = 1$, ensuring that the sum of all feature weights is 1.

4. Method

This section presents the methodology organized around three aspects of urban traffic congestion: identification, propagation analysis and prediction, and flow optimization. First, the study area (Hongjialou area of Jinan) is spatially discretized using square grids, and a multi-source traffic dataset is constructed by integrating geomagnetic traffic flow data, vehicle trajectory data, and mobile phone signaling data. On this basis, a traffic congestion area identification method combining the DBSCAN clustering algorithm with the CNN-LSTM deep learning model is proposed, which extracts congestion feature vectors through clustering and then utilizes a convolutional neural network to extract spatial features and a long short-term memory network to capture temporal dependencies. Second, a congestion propagation analysis framework based on Bayesian networks is constructed to achieve quantitative reasoning of congestion diffusion probabilities between adjacent grids. Simultaneously, an SA-CNN-BiLSTM prediction model is designed, incorporating a self-attention mechanism to enhance the weight allocation of key spatiotemporal features and improve congestion prediction accuracy. Finally, a peak-hour traffic flow allocation optimization model is established with the objectives of minimizing total delay time and the congestion index, and an improved simulated annealing algorithm is employed for solution, providing theoretical and methodological support for traffic control strategy formulation.

4.1 Urban Grid Partitioning

The Hongjialou area of Jinan was selected as the study area. A square grid partitioning method was adopted, and the grids were classified into two types based on urban traffic network

characteristics: signalized intersection grids and unsignalized intersection grids. ArcGIS software was used to partition the study area into rectangular grid shapes (Figures 2 and 3), facilitating data management and analysis.

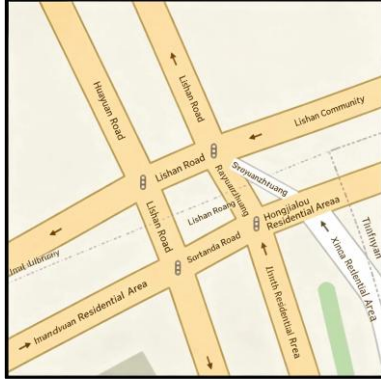


Figure 2 Signalized intersection grid



Figure 3 Unsignalized intersection grid

4.2 Congestion Identification Model Integrating Clustering Algorithm and CNN-LSTM

To construct a traffic congestion identification model integrating clustering algorithms with CNN-LSTM, DBSCAN was selected as the clustering algorithm for the congestion area identification stage, and a CNN-LSTM model was employed for the fusion identification of multi-source traffic data.

First, DBSCAN was applied to cluster the spatial grids. The clustering features of each region can be represented as a feature vector.

$$|N_{\epsilon}(p)| \geq MinPts \quad (7)$$

Here, $|N_{\epsilon}(p)|$ denotes the number of points contained in the ϵ -neighborhood of point p , and $MinPts$ is a preset threshold for the minimum number of points.

Through clustering analysis, a set of clusters $C = \{C_1, C_2, \dots, C_k\}$ is obtained, representing traffic flow, speed, crowd locations, vehicle positions, etc., where each cluster denotes a city grid area. For each cluster, the degree of congestion can be further assessed.

In the CNN layer, the feature vectors of each cluster are formed into an input matrix $X \in R^{n \times m}$, and local features are extracted through convolution operations. The specific operation formula of the CNN convolutional layer is shown in (8).

$$Y = X * W + b \quad (8)$$

Subsequently, the output Y of the CNN is used as the input to the LSTM layer. The LSTM's forget gate f_t , input gate i_t , output gate o_t , and cell state C_t are utilized to compute the final hidden state h_t at the current time step, thereby completing the recognition of traffic congestion. The basic formulas are shown in (9) to (13) [19].

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (9)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (10)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (11)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (12)$$

$$h_t = o_t * \tanh(C_t) \quad (13)$$

4.3 Bayesian Network-based Analysis of Urban Traffic Congestion Propagation

Based on urban grid division, the study area is divided into several grid cells. The traffic state of each grid (such as traffic flow X_i , vehicle speed Y_i , accident rate Z_i , and road condition W_i) is modeled using a Bayesian network [20].

The joint probability distribution of the system is expressed as the product of all nodes conditioned on their parent nodes, as shown in Equation (14), where parent nodes typically include spatial dependencies of adjacent grids and external factors.

$$P(X_1, Y_1, Z_1, W_1, \dots, X_n, Y_n, Z_n, W_n) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i)) \cdot P(Y_i | \text{Parents}(Y_i)) \cdot P(Z_i | \text{Parents}(Z_i)) \cdot P(W_i | \text{Parents}(W_i)) \quad (14)$$

The conditional probability table (CPT) specifically defines that the vehicle speed Y_i is jointly determined by the traffic flow X_i and the accident rate Z_i . Equation (15) presents the probability distribution of the speed being smooth, slow, or congested under different flow and accident rate conditions.

$$P(Y_i | X_i, Z_i) = \begin{pmatrix} P(Y_i = \text{smooth} | X_i = x_i, Z_i = z_i) \\ P(Y_i = \text{slow} | X_i = x_i, Z_i = z_i) \\ P(Y_i = \text{congested} | X_i = x_i, Z_i = z_i) \end{pmatrix} \quad (15)$$

In the propagation inference of traffic congestion, when data such as traffic flow and speed of grid i are known, Bayes' theorem can be used to infer the traffic state of the adjacent grid $i + 1$, and the propagation probability is calculated using equations (16)–(18).

$$P(a) = \frac{P(X_{i+1} | \text{Parents}(X_{i+1})) \cdot P(Y_{i+1} | \text{Parents}(Y_{i+1}))}{P(X_i, Y_i, Z_i, W_i)} \quad (16)$$

$$P(b) = \frac{P(Z_{i+1} | \text{Parents}(Z_{i+1})) \cdot P(W_{i+1} | \text{Parents}(W_{i+1}))}{P(X_i, Y_i, Z_i, W_i)} \quad (17)$$

$$P(B) = P(a) \times P(b) \quad (18)$$

Furthermore, under peak hours, special weather conditions, or emergency events, the Bayesian network supports dynamic prediction. For example, given that traffic flow increases and an accident occurs in a certain area, forward reasoning can be performed according to Equation (19) to update the vehicle speed, congestion level, and propagation probability of that area and its surrounding areas.

$$P(Y_{i+1} | X_i, Z_i) = \sum_{X_i, Z_i} P(Y_{i+1} | X_i, Z_i) \cdot P(X_i, Z_i) \quad (19)$$

Parameter learning and propagation probability are key steps in the traffic congestion prediction model. The former facilitates the extraction of key features from multi-source data, while the latter helps reflect the spatiotemporal dynamics of traffic flow. Since this study divides trajectory data into different spatiotemporal units [18] and identifies recurrent congestion grids based on this, the data volume is significantly reduced. Therefore, a Bayesian estimation method is adopted for parameter learning of the constructed Dynamic Bayesian Network (DBN) to estimate congestion propagation

probabilities.

For the visualization of urban traffic congestion propagation, the Hongjialou area was selected as the case study. The study scope mainly includes urban arterial roads such as Shanda Road and Jiefang Road, as well as large traffic attraction points including the Hongjialou Campus of Shandong University. These areas carry significant volumes of morning and evening peak commuter trips, as shown in Figure 4.

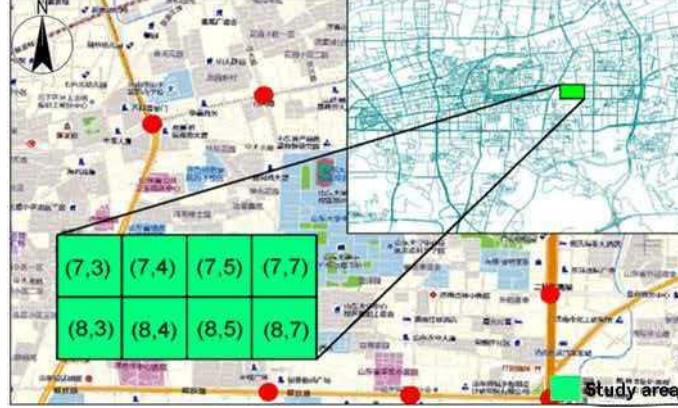


Figure 4 Study Area

Based on the multi-source dataset, the Bayesian network was trained to obtain the Conditional Probability Table (CPT), and the propagation process was visualized in the urban grid. As shown in Figure 5, grids (7,3) and (7,4) were the first to experience congestion, which subsequently propagated to grid (7,5). During the 07:40-07:55 period, grid congestion reached its maximum, covering 62.5% of the study area. In the following 10 minutes, the congestion situation decreased, and grid (8,5) transitioned to a free-flow state.

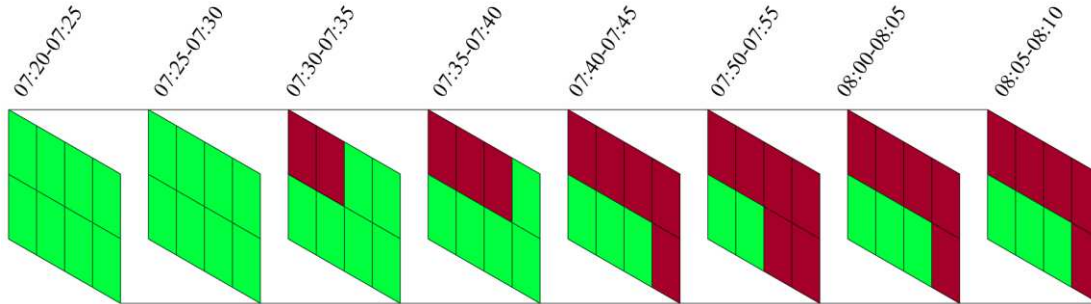


Figure 5 Congestion propagation subtree space-time cube graph

Investigating grids (7,3) and (7,4) where congestion first occurred, parameter learning and probability calculation of the Bayesian network revealed lateral propagation probabilities of 58% and 71%, respectively. During the morning peak period, the congestion grid occupancy rate reached 100%, with frequent propagation and relatively high propagation probabilities. These findings demonstrate significant research significance and reference value.

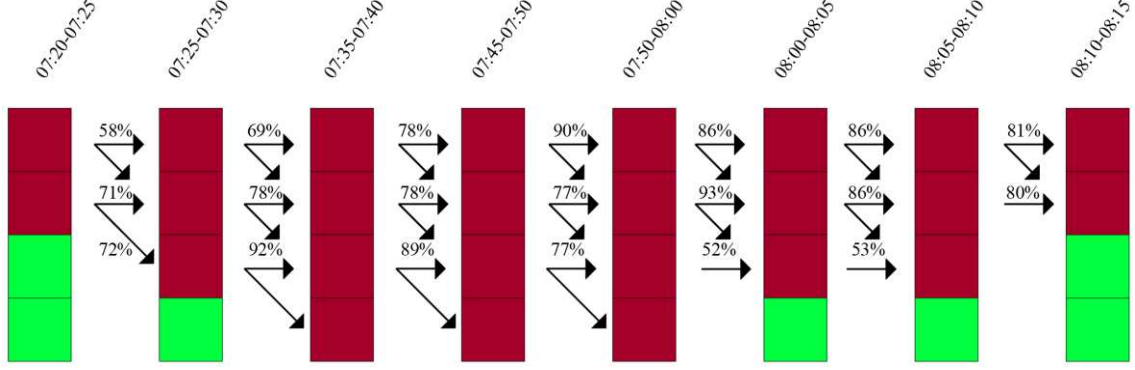


Figure 6 Congestion propagation subtree graph

4.4 SA-CNN-BiLSTM Traffic Congestion Prediction Model

For the traffic congestion prediction task, this study employs CNN to extract spatial features from traffic data, BiLSTM to capture temporal information, and introduces a self-attention mechanism to dynamically weight features based on their importance, thereby enhancing the model's attention to key areas and time periods and improving prediction accuracy and robustness.

To facilitate model construction, environment calibration is first performed: the study road network G is defined as a graph, where nodes represent intersections or key locations and edges represent connecting road segments; the fused dataset D is defined as a multi-dimensional feature matrix from multiple sources; and the traffic flow prediction parameters are defined as key variables for describing and predicting traffic flow dynamics.

In the feature parameter extraction stage, spatial feature extraction is primarily accomplished through convolution and pooling operations: the convolutional layer applies convolution kernels to the input data to extract local spatial features, while the pooling layer reduces the feature map dimensions and preserves the most important spatial features.

In the SA-CNN-BiLSTM congestion prediction model, the softmax function is used to convert the output layer into congestion probabilities. The softmax function is first defined: assuming the model output layer has C categories with raw outputs $z_1, z_2, \dots, z_C$, the softmax function converts each category score into a probability.

The input to the improved model is gridded traffic flow data. The first half of the model employs convolutional layers primarily for extracting spatial features from traffic flow data. To prevent overfitting and improve computational efficiency, pooling layers are added after the convolutional layers. The second half utilizes the BiLSTM model, designed to extract temporal features from traffic flow data. According to the topology, a Softmax layer is inserted between the convolutional network layers and the recurrent network layers [21], with the purpose of enhancing the spatial features of traffic flow data extracted by the convolutional network layers through the Softmax function.

4.5 Simulated Annealing-based Peak-hour Traffic Flow Allocation Model

4.5.1 Objective Function Definition

The objective function definition includes minimizing total delay time and minimizing the congestion index, representing traffic flow smoothness and traffic pressure on road segments, respectively.

(1) Minimizing total delay time

Delay time is closely related to vehicle speed and traffic flow. Assuming delay time is a function of traffic flow and vehicle speed, it can be expressed as Equation (20).

$$t_i(x_i, v_i) = \alpha_i \cdot \left(\frac{x_i}{c_i} - 1\right)^2 + \beta_i \cdot \left(\frac{v_i}{v_{max}}\right) \quad (20)$$

Where x_i is the traffic flow of road segment i , C_i is the capacity of that segment, v_i is the average vehicle speed of the segment, $v_{\max}$ is the maximum design speed of the segment, and α_i and β_i are weight coefficients.

This objective function can be expressed as the weighted sum of delay times across all road segments, as shown in Equation (21).

$$F(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^n t_i(x_i, v_i) \quad (21)$$

where $\mathbf{x}$ is the flow vector and $\mathbf{v}$ is the speed vector.

(2) Minimizing the congestion index

In this subsection, the congestion index is represented by the ratio of traffic flow to road capacity. Through the relationship between vehicle speed and traffic flow, the congestion index can be expressed as shown in Equation (22).

$$I_i(x_i, v_i) = \frac{x_i}{C_i} + \left(1 - \frac{v_i}{v_{\max}}\right) \quad (22)$$

The total congestion index is the sum of congestion indices across all road segments, as expressed in Equation (23).

$$F(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^n w_i \cdot I_i(x_i, v_i) \quad (23)$$

where w_i is the weight coefficient for road segment i , reflecting its relative importance.

4.5.2 Constraints

The model constraints include flow capacity constraints, total flow balance constraints, and vehicle speed constraints.

(1) Flow capacity constraint

The traffic flow on each road segment must be within its capacity range, as simply expressed in Equation (24).

$$0 \leq x_i \leq C_i, \quad \forall i = 1, 2, \dots, n \quad (24)$$

(2) Total flow balance constraint

Given that the peak-hour traffic demand is D , the total flow must equal the demand.

$$\sum_{i=1}^n x_i = D \quad (25)$$

(3) Vehicle speed constraint

The vehicle speed v_i should be within a reasonable range, i.e., the speed cannot exceed the maximum design speed $V_{\max}$ and cannot be lower than a certain minimum speed $V_{\min}$. In general, $V_{\min}$ is taken as 0.

$$V_{i_{\max \min}} \quad (26)$$

4.5.3 Simulated Annealing Algorithm Solution

The simulated annealing algorithm is employed to solve the aforementioned traffic flow allocation problem [26]. During algorithm execution, simulated annealing explores the solution space by randomly adjusting the traffic flow and vehicle speed on each road segment to minimize the objective function. The specific steps are as follows.

- (1) Initialization. Randomly select an initial traffic flow allocation scheme and speed scheme.
- (2) Generate neighborhood solutions. Based on the current flow and speed schemes, generate new schemes by fine-tuning the flow or speed to form neighborhood solutions.
- (3) Calculate the objective function. For each newly generated neighborhood solution, compute the objective function value.
- (4) Acceptance criterion. If the objective function value of the new solution is lower, accept it; otherwise, accept it with a certain probability determined by the temperature T and the difference in objective function values.
- (5) Cooling. Gradually reduce the temperature according to the temperature decay rule, decreasing the probability of accepting inferior solutions, until the temperature reaches its minimum or the stopping condition is met.

Through multiple iterations, the simulated annealing algorithm can find the optimal traffic flow allocation scheme, reducing traffic congestion and improving overall traffic efficiency.

5. Results

This section validates the effectiveness of the proposed methods from three dimensions: congestion identification, propagation prediction, and traffic flow optimization. In terms of congestion identification, the DBSCAN-CNN-LSTM model outperforms rule-based, clustering-based, and traditional CNN-LSTM models in recall, F1-score, and AUC-ROC, achieving an AUC-ROC of 0.96, demonstrating strong congestion level discrimination capability. In terms of congestion propagation and prediction, the Bayesian network successfully reveals the congestion diffusion patterns between grids, with the grid congestion occupancy rate reaching 100% during the morning peak and a lateral propagation probability of 71%. The SA-CNN-BiLSTM model achieves a prediction accuracy of approximately 94% during morning and evening peak hours, with MAE reductions of approximately 16% and 32% compared to BiLSTM and XGBoost models, respectively, and the residual distribution approaches a zero-mean normal distribution. In terms of traffic flow optimization, the optimization scheme based on the improved simulated annealing algorithm reduces average travel time by 12.5%, shortens average queue length by 22.4%, decreases delay time by 17.8%, and increases vehicle throughput by 7.6%. Comparative experiments with the genetic algorithm and particle swarm optimization further demonstrate that the improved simulated annealing algorithm achieves the best performance in average travel time, total network delay, and convergence time, exhibiting promising potential for practical application.

5.1 Numerical Experiments and Model Optimization Evaluation

5.1.1 Numerical Experiments

This subsection conducts numerical experiments using traffic data from the Hongjialou area of Jinan to validate the rationality and accuracy of the traffic congestion identification model based on the DBSCAN clustering algorithm and CNN-LSTM.

(1) Grid area

The grid area covers latitude and longitude ranges of [36.65987, 36.686022] and [117.038184, 117.073662], with a grid step size of $l = 20$ m. The dataset covers an area of 5.5 km². The study area includes both signalized and unsignalized grids, as detailed in Figure 7.

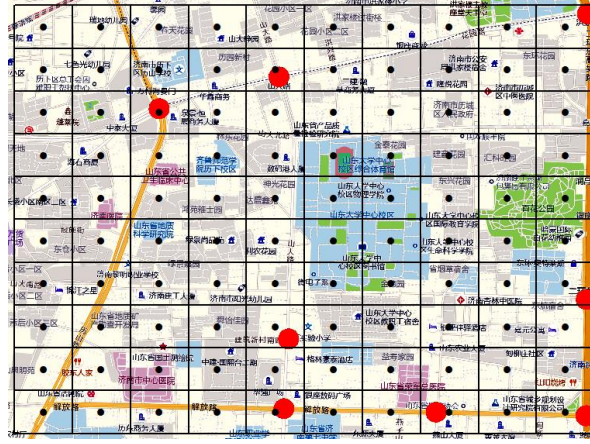


Figure 7 Grid division of the study area

(2) Multi-source dataset description

The dataset spans from March 11 to 17, 2024 (Monday to Sunday, with no holidays during this period), totaling 7 days. The dataset contains 181,963 records, including information on time, longitude, latitude, mileage, and speed.

(3) Grid traffic flow characteristic analysis

As observed from Figures 8-11, significant differences exist in traffic operational characteristics between signalized and unsignalized intersections. Both types of intersections exhibit distinct diurnal and weekly variation patterns. In comparison, signalized intersections show more pronounced speed reductions during peak hours with more significant vehicle queuing, while unsignalized intersections exhibit lower degrees of traffic flow control with relatively dispersed vehicle passage.

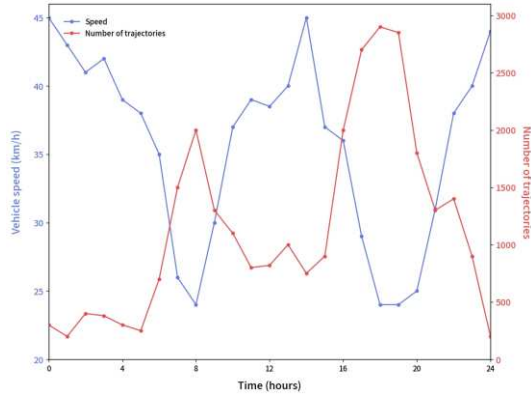


Figure 8 Daily changes in signal grid

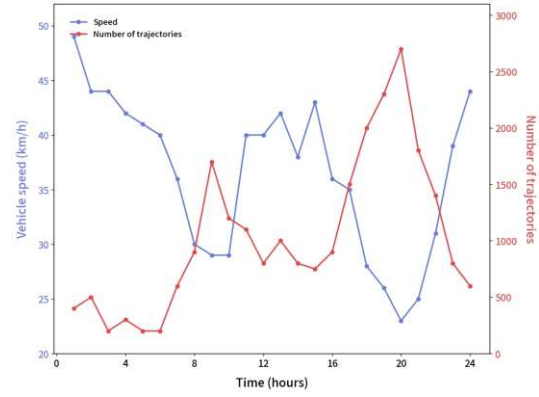


Figure 9 Daily changes in no-signal grid

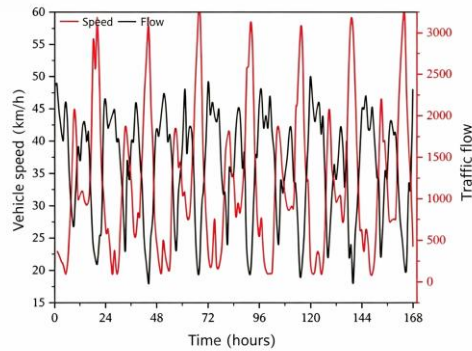


Figure 10 Signal grid cycle changes

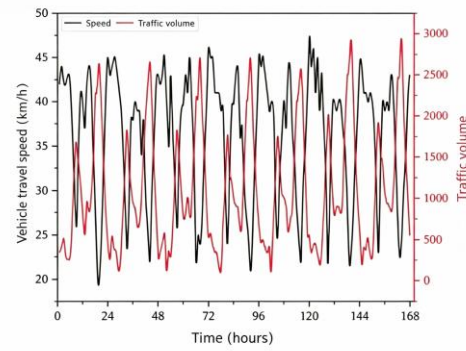


Figure 11 Weekly changes of no-signal grid

(4) Clustering effect analysis and visualization

Traffic areas are partitioned into multiple high-density regions classified into congestion levels 1 through 4. Colors gradually transition from yellow to red, indicating different levels from mild to severe congestion.

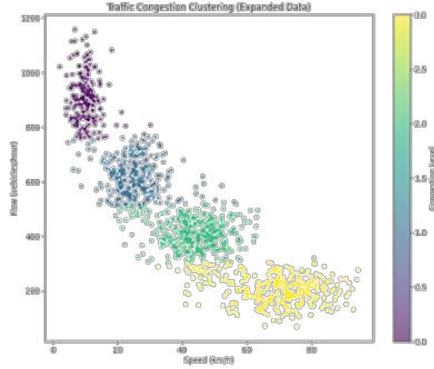


Figure 12 Congestion cluster analysis diagram

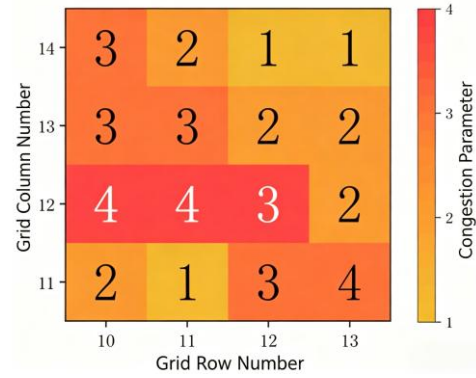


Figure 13 Congestion area identification map

5.1.2 Model Loss Function and Optimizer

Since the objective of congestion identification is a classification problem, the model's loss function can be optimized by combining the clustering and LSTM components. The loss function is cross-entropy loss, and the Adam optimizer is adopted.

Model parameters are optimized through iterative training, with the validation set loss monitored to prevent overfitting. An early stopping mechanism is employed: when the validation set loss shows no improvement for multiple consecutive epochs, training is terminated early. The identification results can be obtained through the above procedure.

5.1.3 Model Evaluation Analysis

To validate the effectiveness of the proposed traffic congestion identification model, the original dataset was first partitioned. To ensure verification reliability, 80% of the data was used as the training set and 20% as the test set. Heat maps were generated using ArcGIS software through coordinate projection, allowing intuitive observation of congestion conditions and point counts in the study area, further delineating the traffic congestion scope. As shown in Figures 14-17, the color within grids deepens from Level 1 to Level 4, with a notable increase in trajectory point counts.

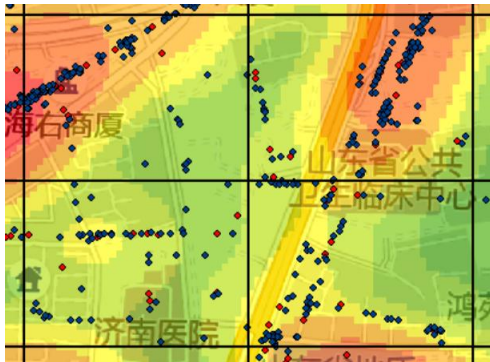


Figure 14 Level 1 Congestion Grid

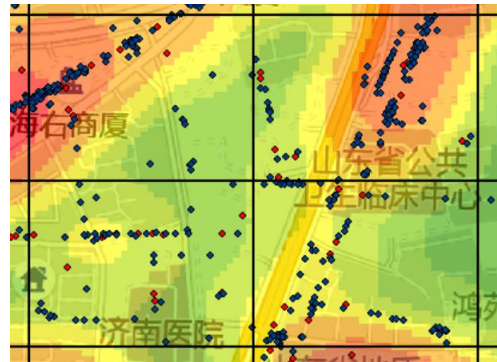


Figure 15 Level 2 Congestion Grid

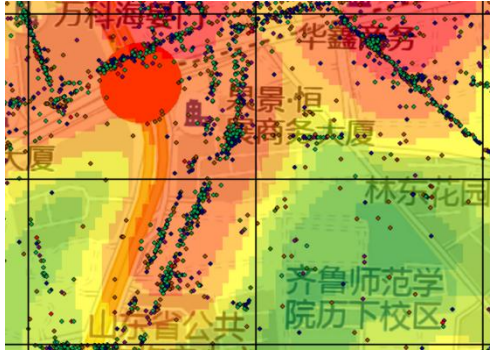


Figure 16 Level 3 Congestion Grid

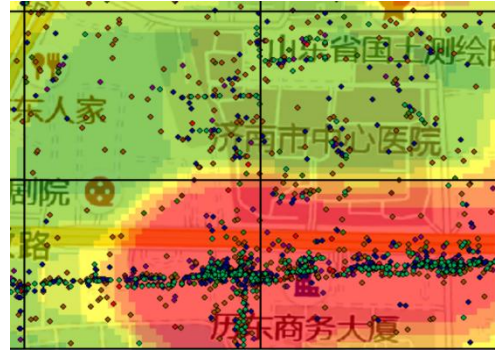


Figure 17 Level 4 Congestion Grid

Furthermore, to evaluate the performance of the proposed traffic congestion identification model, four common evaluation metrics were employed: recall, F1-score, and AUC-ROC curve (Area Under the Curve). Three alternative congestion identification models were used for comparative validation: a rule-based model (Model 1), a clustering-based model (Model 2), and a traditional CNN-LSTM model (Model 3). The comparison results are presented in Table 4.

Table 4 Model parameter comparison table

Model	recall rate	F1 score	AUC-ROC
My model	0.94	0.92	0.96
Model 1	0.68	0.70	0.80
Model 2	0.82	0.81	0.88
Model 3	0.90	0.87	0.92

From the comparison of accuracy, precision, and AUC-ROC values, it can be observed that the proposed model outperforms other models in identifying various types of congestion areas, demonstrating strong identification capability and the ability to accurately distinguish congestion levels at different thresholds.

5.2 Model Experimental Results and Analysis

5.2.1 Experimental Environment

The experiments were conducted in the PyCharm development environment using Python libraries including Keras, NumPy, and Pandas for model data training and data processing, and Scikit-learn for model performance evaluation. The experimental environment consists of Windows 11 Home Chinese Edition, with a 12th Gen Intel Core i7-12700H CPU, NVIDIA GeForce RTX 3060 Laptop GPU, 16 GB RAM, and Python 3.7.

5.2.2 Model Evaluation Results Analysis

In traffic congestion prediction, traffic flow characteristics vary significantly across different time periods. The morning and evening peak hours (7:00-9:00 and 17:00-19:00) pose greater challenges for prediction models due to high traffic density and complex traffic environments. Therefore, to further evaluate the adaptability and stability of the models under different congestion states, this study compared the MAE and MSE of SA-CNN-BiLSTM (Model 1), BiLSTM (Model 2), and XGBoost (Model 3) during morning and evening peak hours. Figures 18 and 19 illustrate the error performance of each model during peak hours. Through this comparison, the prediction advantages of SA-CNN-BiLSTM in complex traffic environments can be intuitively analyzed, and its stability and robustness under congested conditions can be verified.

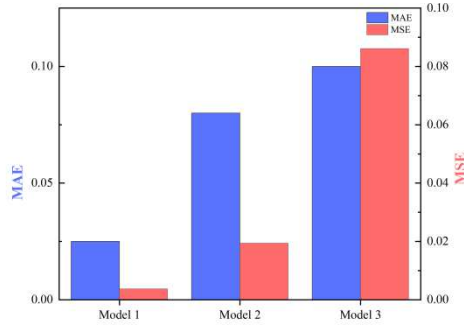


Figure 18 Comparison of MAE and MSE of the model during the morning peak period

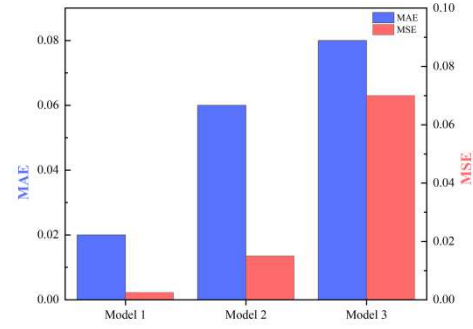


Figure 19 Comparison of MAE and MSE for the evening peak model

As shown in Figures 18 and 19, the proposed model exhibits lower MAE and MSE than both comparison models, indicating higher prediction accuracy and better data fitting capability, particularly in handling deviations and outliers.

To further evaluate the performance of the three models across different peak periods, an error band analysis method was adopted to compare prediction errors during morning and evening peak hours. Figures 20 and 21 present the error band analysis comparisons of the three models during morning and evening peak periods. It can be observed that the prediction errors of the three models exhibit different fluctuation amplitudes and stability across different time periods.

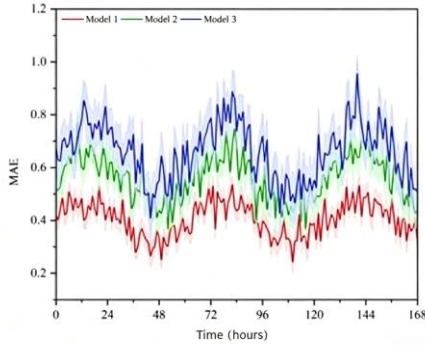


Figure 20 Analysis of model morning peak error band

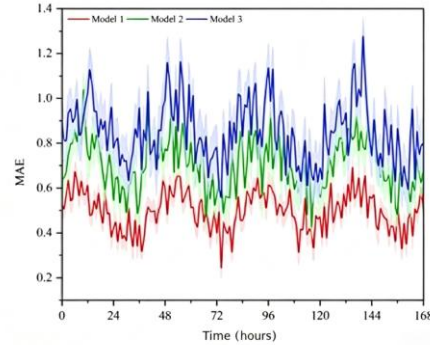


Figure 21 Model evening peak error band analysis

During the morning peak period, Model 1 exhibits a relatively narrow error band, with the mean absolute error (MAE) consistently maintained at a low level and small standard deviation (STD) fluctuations. The MAE of Model 1 is approximately 0.53 with a standard deviation of approximately 0.07, and the error band is relatively concentrated, indicating consistent prediction results with low variability. In comparison, Model 2 exhibits a slightly wider error band with an MAE of approximately 0.63 and a standard deviation of 0.09, indicating lower prediction accuracy and higher variability during the morning peak. Model 3 performs the worst, with an MAE of approximately 0.78, a standard deviation of 0.12, and a wide error band with large fluctuation amplitudes, reflecting poor adaptability during high-flow periods.

During the evening peak period, the error bands of the three models show more pronounced fluctuation changes. Model 1 maintains a narrow error band with an MAE of approximately 0.57 and a standard deviation of 0.08. Although the error slightly increases, it still maintains good accuracy stability. The error band of Model 2 widens, with an MAE of approximately 0.72 and a standard deviation of 0.10, indicating relatively unstable prediction results during the evening peak. Model 3

continues to exhibit a wide error band with an MAE of approximately 0.85 and a standard deviation of 0.13, showing poor prediction accuracy and large fluctuations.

Overall, Model 1 demonstrates superior performance during both morning and evening peak periods, particularly in terms of error fluctuation amplitude and stability. The results further confirm this conclusion: Model 1 achieves lower MAE than the other two models in both peak periods, and its error band variability is significantly smaller, demonstrating good prediction stability.

5.2.3 Model Comparison Analysis

This study compares actual values with the prediction results of three different models to explore the adaptability and fitting accuracy of each model in handling traffic flow at signalized and unsignalized intersections. The following four figures present the congestion index variation trends under the four scenarios, providing detailed data support for subsequent model evaluation.

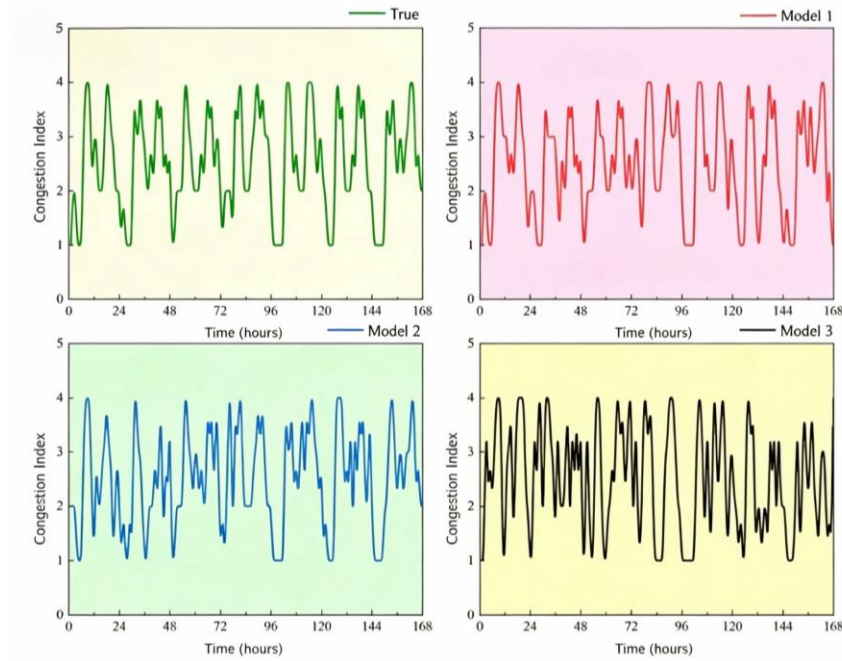


Figure 22 Comparison of predictions for unsignalized intersections during the morning peak period

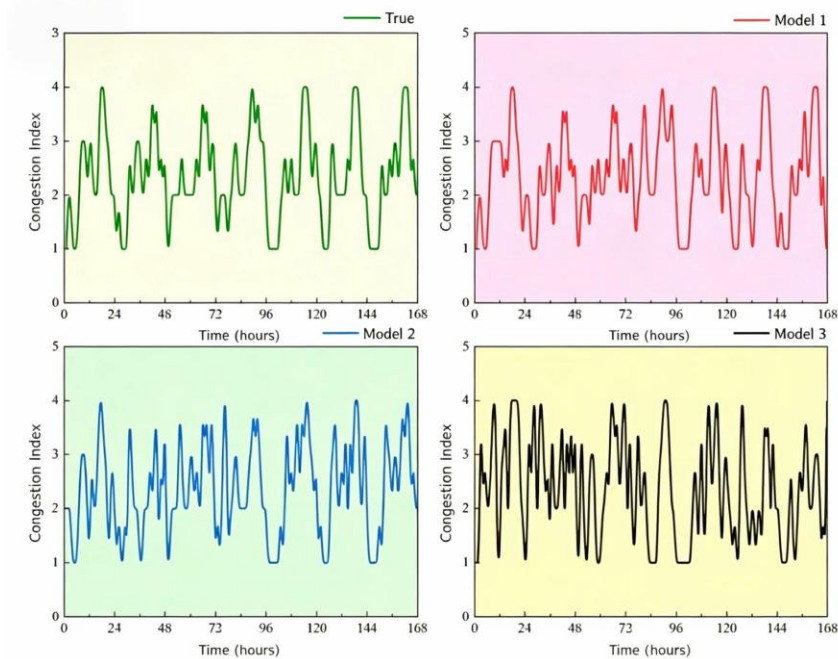


Figure 23 Comparison of predictions for unsignalized intersections during the evening peak period

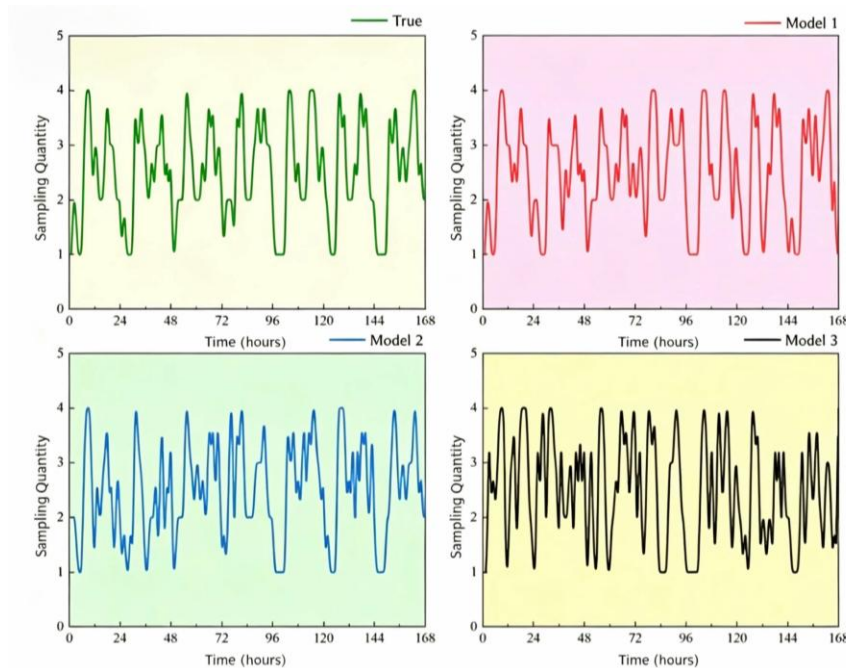


Figure 24 Comparison of predictions for signalized intersections during the morning peak period

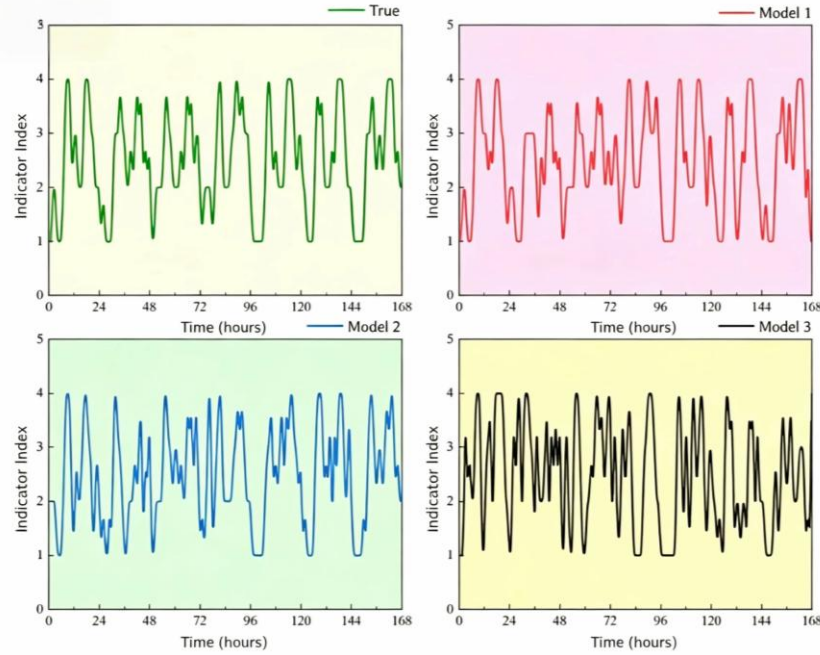


Figure 25 Comparison of predictions for signalized intersections during the evening peak period

As shown in Figures 22-25, SA-CNN-BiLSTM (Model 1) demonstrates significant advantages over the other two models in evening peak traffic congestion prediction, particularly in prediction accuracy. Specifically, the prediction accuracy of Model 1 during high-congestion periods improved by approximately 23% compared to BiLSTM (Model 2). XGBoost (Model 3) exhibits lower prediction accuracy during the evening peak with a wider error range; during high-congestion periods, its prediction error typically falls between 15% and 20%, failing to effectively reflect actual traffic conditions.

Overall, Model 1 significantly outperforms Models 2 and 3 in evening peak prediction accuracy. During peak periods, the predicted values of Model 1 are highly consistent with the actual values, achieving an accuracy of approximately 94%, demonstrating strong adaptability and prediction stability in complex traffic environments. In contrast, Models 2 and 3 exhibit substantial deviations in their prediction results, failing to accurately capture the actual variations in traffic flow. Considering both prediction accuracy and stability, Model 1 demonstrates significant advantages in the evening peak traffic congestion prediction task, capable of more accurately reflecting the dynamic changes in traffic conditions.

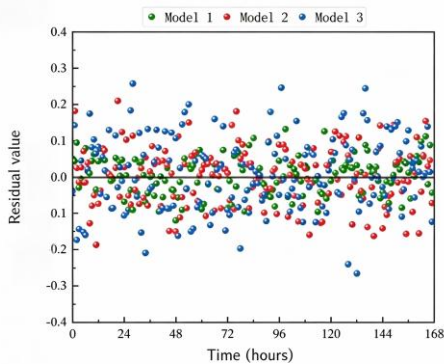


Figure 26 Residual value comparison

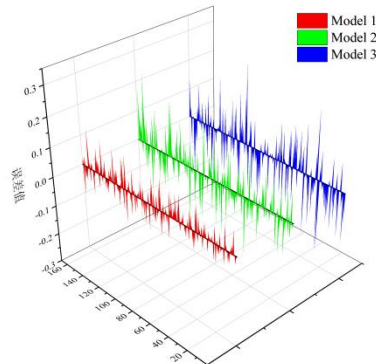


Figure 27 Error distribution trend diagram

Through analysis of the residual values in Figure 26 and the error trend distribution in Figure 27 for the three models, it is evident that during the morning peak, the prediction accuracy of Model 1 is significantly superior to Models 2 and 3. Specifically, the residual fluctuations of Model 1 are relatively small, and its residual distribution approximates a normal distribution with a mean close to zero, indicating small prediction errors and high stability. In contrast, Models 2 and 3 exhibit larger residual fluctuations and more dispersed residual distributions. Particularly during high-congestion periods, the errors of Models 2 and 3 increase significantly, indicating lower prediction accuracy during these periods.

In the quantitative analysis of prediction errors, Model 1 achieves errors close to zero in the vast majority of time periods, with prediction results very close to actual values and an accuracy of approximately 94%. Model 2 achieves an accuracy of approximately 85%, and Model 3 approximately 78%. These data indicate that Model 1 improves accuracy by approximately 23% and 20% compared to Models 2 and 3, respectively, with particularly outstanding performance during peak periods. Model 1 not only outperforms the other models in overall accuracy but also predicts peak-hour traffic congestion more accurately, thereby demonstrating higher reliability in practical applications.

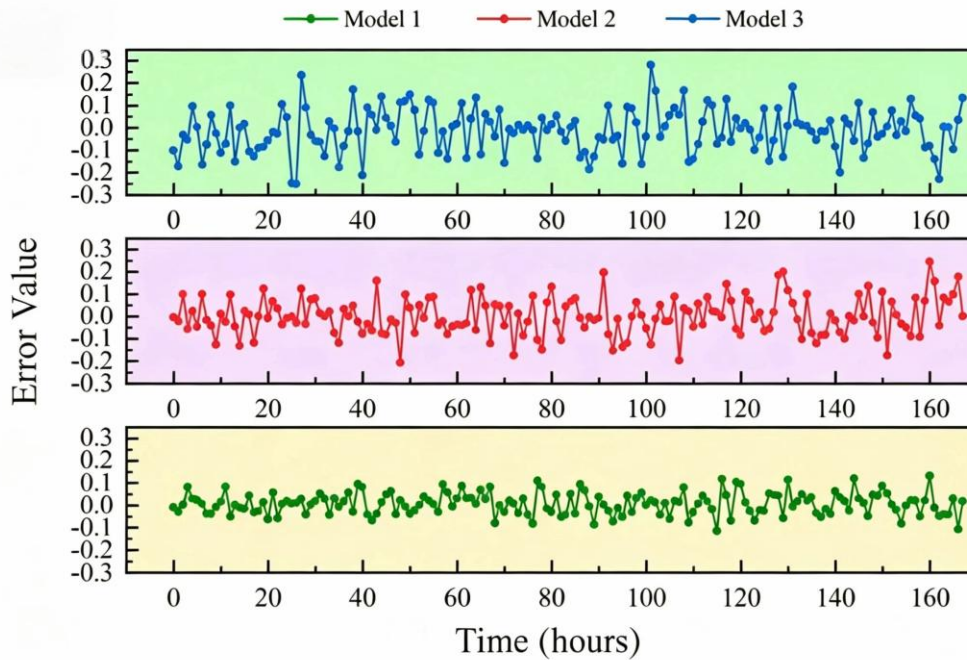


Figure 28 Error distribution hourly chart

According to the analysis of Figure 28, the proposed model demonstrates significant prediction advantages in performance, with notably lower error values and higher accuracy compared to traditional prediction models. From the error distribution perspective, the model not only effectively reduces prediction errors but also exhibits stronger robustness in stability. In contrast, traditional prediction models show larger error fluctuations and weaker prediction stability. Particularly in complex traffic scenarios or under abnormal data conditions, the error values of traditional models exhibit significant upward trends. Furthermore, the mean absolute errors of traditional models remain at high levels, making it difficult to accurately capture potential spatiotemporal features in diverse traffic data environments. The proposed model, through the introduction of deeper feature extraction and optimization mechanisms, not only significantly reduces prediction bias but also more accurately reflects the actual traffic flow variation characteristics, fully demonstrating the superiority and applicability of the model.

5.3 Simulation Experiments and Results Analysis

5.3.1 Simulation Platform and Experimental Configuration

The experiments in this subsection employ the VISSIM traffic simulation platform, with a typical urban congestion area as the simulation object. The road network includes two major intersections, three arterial roads, and multiple secondary roads, as shown in Figure 29. Traffic flow inputs are constructed based on the fused dataset (geomagnetic flow, GPS trajectories, and signaling data), with the simulation time set to the weekday morning peak period (07:00-09:00). All vehicle types, route choices, and signal timing parameters are constructed based on actual data to ensure the realism of the simulation scenario.



Figure 29 road network diagram



Figure 30 Road network simulation diagram

To validate the effectiveness of the optimization strategy, two sets of simulation experimental schemes were designed.

(1) Scheme A (Baseline scheme). Simulation is conducted according to current actual road traffic signal timing and vehicle driving behavior without any optimization intervention, reflecting the original state of traffic operations during peak hours. This scheme reproduces typical problems such as high traffic flow density, low vehicle passage efficiency, and long average delays during peak hours, providing a baseline for subsequent optimization effect evaluation.

(2) Scheme B (Optimized scheme). Signal timing for key road segments is optimized based on the improved simulated annealing algorithm, and the optimization results are imported into the VISSIM platform for simulation. This scheme aims to validate the effectiveness of the proposed method in improving peak-hour traffic congestion and enhancing vehicle passage capacity.

5.3.2 Simulation Indicators and Evaluation System

To evaluate the application effectiveness of the improved simulated annealing algorithm in urban peak-hour traffic control within the context of multi-source data fusion, a real road network model was constructed on the VISSIM platform. Data were extracted using its built-in modules and interfaces, combined with Python for data processing and visualization. A simulation evaluation system comprising four main indicators was established: average travel speed, reflecting the average driving speed of vehicles across the entire simulation network to assess network operational efficiency; average delay time, measuring vehicle waiting time caused by queuing and signal control, which is a key indicator for evaluating traffic congestion levels; queue length, monitoring queuing conditions at major intersections during peak hours to extract maximum and average queue lengths and evaluate the effectiveness of signal timing adjustments in alleviating traffic pressure at bottleneck sections; and vehicle throughput, primarily used for comparative analysis of the improvement in passage efficiency at arterial roads or key intersections under different traffic control schemes during peak hours.

5.3.3 Simulation Results Comparative Analysis

To more intuitively demonstrate the changes in traffic conditions before and after optimization, heat maps were employed as an analysis tool. Heat maps can clearly reflect the spatial distribution of road congestion levels and their temporal variation trends, intuitively presenting the performance of different road segments under different traffic states.

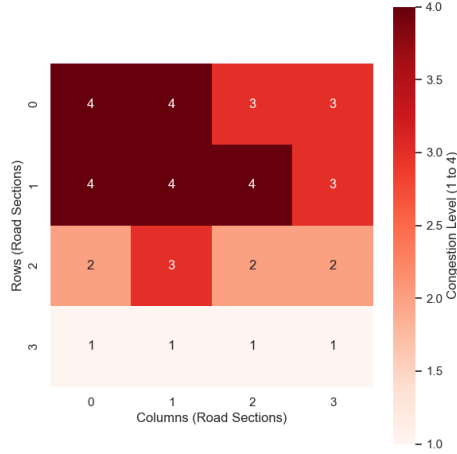


Figure 31 Traffic congestion heat map of Scheme A

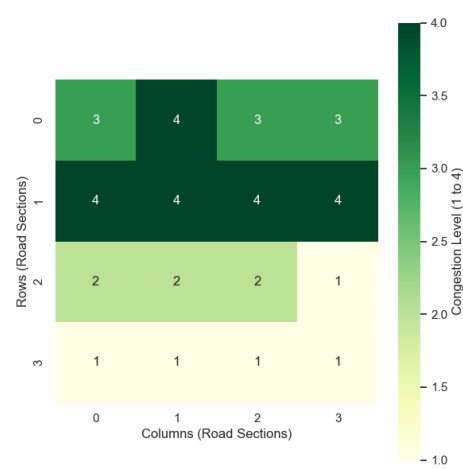


Figure 32 Traffic congestion heat map of Scheme B

As shown in Figures 31 and 32, after implementing traffic flow optimization measures, congestion levels on multiple road segments were significantly improved. Areas that were originally at Level 4 congestion before optimization were reduced to Level 3 or lower congestion levels after optimization. Overall, the congestion distribution became more uniform, and many road segments that previously experienced high-intensity congestion showed significantly enhanced traffic mobility, reflecting the effectiveness of the optimization scheme in alleviating traffic pressure and improving road capacity.

Table 5 Comparison of main traffic operation indicators before and after optimization

Indicator Name	Scheme A	Scheme B	Optimization Rate
average travel time (s)	534.6	468.78	12.5%
average speed (km/h)	23.3	27.4	17.6%
average queue length (m)	101.7	78.9	22.4%
delay time (s)	29306	24091	17.8%
vehicle throughput (pcu)	2837	3051	7.6%

According to Table 5, the optimized average travel time is significantly reduced, indicating that the optimization scheme effectively decreases vehicle travel time on roads and improves passage efficiency. The average vehicle speed has increased, further reflecting the positive effects of the optimization scheme in alleviating traffic congestion and enhancing road capacity. The average queue length is notably shortened, indicating enhanced traffic flow fluidity and improved vehicle passage capacity. The reduction in delay time further validates the effectiveness of the traffic flow optimization scheme in reducing traffic congestion and delays. The number of vehicles passing per unit time has increased, with traffic flow and road segment capacity improving by approximately 7.6%.

5.4 Model Validation and Comparative Experimental Analysis

After completing the preliminary simulation experiments based on the improved simulated

annealing algorithm and validating its optimization effectiveness in specific traffic scenarios, this study further conducts model validation and comparative experimental analysis from the perspectives of generalization capability, adaptability, and comparative performance, aiming to more comprehensively evaluate the proposed method under different traffic conditions and algorithm backgrounds.

5.4.1 Data Partition and Validation Method

To ensure the practical adaptability of the model under non-training conditions, the original traffic dataset was partitioned into training and validation sets at a 7:3 ratio. Considering the temporal characteristics and periodicity of traffic data, a stratified sampling strategy by time period and category was adopted to avoid sample distribution bias misleading model performance, ensuring that the training and validation sets are representative in terms of flow levels, weather conditions, and time period types (weekday/holiday, peak/off-peak). Simultaneously, a unified indicator system consistent with the comparison models was introduced, selecting three representative indicators for evaluation: average travel time (ATT), total network delay (TDE), and convergence time (CT).

5.4.2 Comparative Algorithm Configuration

The genetic algorithm (GA) and particle swarm optimization (PSO) were selected for comparison. To ensure fairness of comparison, all algorithms were run under identical data, network structures, and simulation platforms. The core parameter settings of each algorithm are shown in Table 6.

Table 6 Algorithm core parameter settings

Algorithm Name	Core Parameters	Parameter Values
Genetic Algorithm	Population size, maximum number of iterations, crossover probability, mutation probability, selection method	50, 200, 0.8, 0.1, tournament selection
Particle Swarm Optimization (PSO)	Number of particles, maximum number of iterations, inertia weight, cognitive coefficient, social coefficient	30, 200, 0.7, 1.5, 1.5
Improved Simulated Annealing Algorithm	Initial temperature, final temperature, cooling coefficient, neighborhood perturbation method, maximum number of iterations	100, 0.1, 0.95, grid weight perturbation, 200

The computation process was conducted on the MATLAB platform with an Intel i7 processor, 16 GB RAM, and Windows 11 operating system.

5.4.3 Comparative Experimental Analysis

Based on the aforementioned algorithm configuration, the improved simulated annealing algorithm (ISA), genetic algorithm (GA), and particle swarm optimization (PSO) were respectively executed on the selected traffic network with actual traffic data. Performance evaluation was conducted from three key indicators: average travel time (ATT), total network delay (TDE), and convergence time (CT). The experimental results are presented in Table 7, where the comparison units for all three algorithms are in seconds.

Table 7 Performance comparison of different algorithms

Indicator name	simulated annealing algorithm	genetic algorithm	Particle Swarm Optimization (PSO) algorithm
ATT	541s	572s	599s

TDE	308s	342s	396s
CT	115s	222s	163s

Table 7 summarizes the performance of the improved simulated annealing algorithm (ISA), genetic algorithm (GA), and particle swarm optimization (PSO) on three key indicators: average travel time, total network delay, and convergence time. The improved simulated annealing algorithm demonstrates strong comprehensive capability in optimizing urban traffic flow allocation. Compared to the genetic algorithm and particle swarm optimization, the improved simulated annealing algorithm exhibits superior performance in both optimization effectiveness and computational efficiency, demonstrating promising potential for practical application.

6. Discussion

This study constructed a comprehensive framework for urban traffic congestion identification, propagation analysis, prediction, and control based on multi-source data fusion, and systematically validated it using the Hongjialou area of Jinan as the study area. In terms of congestion identification, the DBSCAN-CNN-LSTM model achieved an AUC-ROC of 0.96, with recall and F1-score both outperforming rule-based, clustering-based, and traditional CNN-LSTM models. These results demonstrate that combining the spatial density clustering capability of DBSCAN with the spatiotemporal feature extraction capability of CNN-LSTM can effectively improve the accuracy of congestion area identification. DBSCAN requires no preset number of clusters and can identify congestion areas of arbitrary shapes, overcoming the limitations of traditional methods such as k-means, which are sensitive to initial centers and struggle with non-convex spatial distributions. Meanwhile, the local spatial features extracted by CNN convolutional layers and the temporal dependencies captured by LSTM form a complementary relationship, enabling the model to achieve strong discrimination across different congestion levels. In terms of congestion propagation analysis, the Dynamic Bayesian Network-based analysis revealed the congestion diffusion patterns during morning peak hours in the Hongjialou area. The study found that grids (7,3) and (7,4) were the first to experience congestion, with lateral propagation probabilities of 58% and 71%, respectively. During the 07:40-07:55 period, the congestion grid coverage reached 62.5% of the study area, and the congestion grid occupancy rate peaked at 100% during the morning peak. These results indicate that urban traffic congestion does not occur randomly but propagates in a chain-like manner along specific spatial pathways, underscoring the importance of congestion source identification for targeted traffic interventions. In terms of congestion prediction, the SA-CNN-BiLSTM model achieved MAEs of 0.53 and 0.57 during morning and evening peak hours, with standard deviations of 0.07 and 0.08, respectively, and a prediction accuracy approaching 94%. This represents improvements of approximately 23% over BiLSTM and 20% over XGBoost. The incorporation of the self-attention mechanism enables the model to dynamically weight features based on data importance, outperforming comparison models in both error band fluctuation amplitude and stability, effectively reducing prediction uncertainty during peak hours. In terms of traffic flow optimization, the improved simulated annealing algorithm-based optimization scheme reduced congestion levels from Level 4 to Level 3 or lower on multiple road segments in VISSIM simulation, significantly decreasing average travel time and queue length. It demonstrated comprehensive advantages over genetic and particle swarm optimization algorithms across average travel time, total network delay, and convergence time.

However, this study has several limitations. First, the data time span is only 7 days (March 11-17, 2024), with a relatively small sample size. The generalization capability of the model under different seasons, holidays, special weather conditions, and emergency events requires further validation. Second, the study was validated only in the Hongjialou area of Jinan, where the road network structure is predominantly grid-based. The applicability of the model to other urban road network configurations such as ring roads and expressway networks remains to be tested. Third, the data fusion employs the entropy weight method and grey relational analysis for static weight allocation without considering dynamic changes in the reliability of each data source under different conditions such as peak versus off-peak hours or clear versus rainy weather. Fourth, the "black-box" nature of deep learning models makes congestion prediction results lack intuitive physical interpretation, which may

affect managers' trust and willingness to adopt the predictions. Fifth, the VISSIM simulation only designed two comparison groups (baseline and optimized), without covering emergency scenarios such as traffic accidents, road construction, or severe weather, leaving the robustness of the optimization strategy under abnormal conditions insufficiently verified.

Based on the aforementioned limitations, future research can be deepened in the following directions. First, dynamic data fusion mechanisms should be introduced, applying attention mechanisms or federated learning to the multi-source data fusion process to adaptively adjust the fusion weights of each data source according to different time periods and scenarios, thereby improving fusion accuracy and robustness. Second, the data collection period should be extended to 3-6 months, covering diverse traffic scenarios including weekdays, holidays, and severe weather. Cross-regional validation should be conducted in multiple cities or different types of road networks, combined with transfer learning methods to reduce data requirements and training costs in new scenarios. Third, the prediction model should be deployed on edge computing devices, combined with stream data processing frameworks to develop online real-time prediction systems, achieving minute-level congestion early warning and dynamic signal control in a closed loop. Fourth, interpretable analysis methods such as SHAP or LIME should be introduced to quantify the contribution of each input feature to congestion prediction results, enhancing the transparency of model decisions and the trust of managers. Fifth, traffic control optimization should be expanded from single peak-hour flow allocation to a comprehensive optimization framework encompassing multiple scenarios including emergency accident response, severe weather management, and large-scale event support, constructing a more complete urban traffic intelligent management and control system. Sixth, hierarchical coordination between macroscopic flow optimization and microscopic signal control should be explored, using the optimization results of the simulated annealing algorithm as upper-level guidance for reinforcement learning-based signal control, achieving a macro-micro integrated traffic management and control strategy.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Shandong Provincial Natural Science Foundation project (ZR2025QC741), the Inovative Team Project of Ji'nan Government (202333036), Shandong Jiaotong University Talent Introduction Research Start-up Fund (BS2023035).

Ethics declaration

This study does not involve experiments on human subjects or the use of human tissue samples. All data used in this research—including geomagnetic traffic flow data, vehicle trajectory data, and mobile phone signaling data—were collected from publicly available or licensed sources and were anonymized prior to analysis. No personal identifiable information was accessed or used in this study. Therefore, institutional ethics committee approval and informed consent were not required.

References:

- [1] Zhao S F. Research and Implementation of Traffic Flow Prediction Method Integrating Spatio-Temporal Characteristics [D]. Sichuan: University of Electronic Science and Technology of China, 2024.
- [2] Zhang W. Visual Analysis of Urban Traffic and TravelBased on Historical Trajectory Data [D]. Sichuan: Southwest University of Science and Technology, 2024.
- [3] Sato G, Sakuraba A, Uchida N, et al. A new road state information platform based on crowd sensing on challenged network environments[J]. Internet of Things, 2022, 18: 100214.

- [4] Hu Z, Zhou J, Zhang E. Improving traffic safety through traffic accident risk assessment[J]. Sustainability, 2023, 15(4): 3748.
- [5] Chiabaut N, Faitout R. Traffic congestion and travel time prediction based on historical congestion maps and identification of consensual days[J]. Transportation Research Part C: Emerging Technologies, 2021, 124: 102920.
- [6] Gore N, Pulugurtha S S, Arkatkar S, et al. Congestion index and reliability-based freeway level of service[J]. Journal of Transportation Engineering, Part A: Systems, 2021, 147(6): 04021027.
- [7] Drliciac M, et al. Research on vehicle congestion group identification for evaluation of traffic flow parameters[J]. Sustainability, 2024, 16(5): 1861.
- [8] Liu B, et al. A graph-based framework for traffic forecasting and congestion detection using online images from multiple cameras[J]. IEEE Access, 2024, 12: 3756-3767.
- [9] Deng S, Jia S, Chen J. Exploring spatial-temporal relations via deep convolutional neural networks for traffic flow prediction with incomplete data[J]. Applied Soft Computing, 2019, 78: 712-721.
- [10] Chen X, Zhang S, Li L. Multi-model ensemble for short-term traffic flow prediction under normal and abnormal conditions[J]. IET Intelligent Transport Systems, 2019, 13(2): 260-268.
- [11] Li C, Xu P. Application on traffic flow prediction of machine learning in intelligent transportation[J]. Neural Computing and Applications, 2021, 33(2): 613-624.
- [12] Liu Y, et al. Traffic congestion predictor: A spatiotemporal graph neural network with congestion-conditional adaptive mechanism and optimization algorithm[J]. Expert Systems with Applications, 2025, 277: 127200.
- [13] Salamanis A, et al. Identifying patterns under both normal and abnormal traffic conditions for short-term traffic prediction[J]. Transportation Research Procedia, 2017, 22: 665-674.
- [14] Yan F, Zhang M, Shi Z. Dynamic partitioning of urban traffic network sub-regions with spatiotemporal evolution of traffic flow[J]. Nonlinear Dynamics, 2021, 105(1): 911-929.
- [15] Zhao W, Ye Y, Ding J, et al. Ipdalight: Intensity-and phase duration-aware traffic signal control based on reinforcement learning[J]. Journal of Systems Architecture, 2022, 123: 102374.
- [16] Xu T, Barman S, Levin M W, et al. Integrating public transit signal priority into max-pressure signal control: Methodology and simulation study on a downtown network[J]. Transportation Research Part C: Emerging Technologies, 2022, 138: 103614.
- [17] Budholiya A, Manwar A B. Efficient traffic monitoring and congestion control with GGA and deep CNN-LSTM using VANET[J]. Multimedia Tools and Applications, 2024, 83(28): 70937-70960.
- [18] Lyu X Y, Shi Q. Intersection traffic flow forecasting based on the smooth long short-term memory model [J]. Journal of Nantong University (Natural Science Edition), 2022, 21(01): 37-43.
- [19] Qu X P. Analysis of Recurrent Congestion Propagation Pattern sand Congestion Prediction Based on Taxi GPS Data [D]. Chang'an University, Xi'an, China, 2023.
- [20] Cheng X Y, Qu X P, Zhang X Y, et al. Analysis of recurrent congestion propagation mechanism based on dynamic Bayesian network[J]. Journal of South China University of Technology (Natural Science Edition), 2022, 50(11): 25-34.
- [21] Chen Yue, Yang Liu, Li Shuai, et al. Traffic congestion prediction algorithm based on CS-Bi LSTM framework [J]. Science Technology and Engineering, 2022, 22(29) : 12917-12926.
- [22] Chen X, Zhang W, Bai H, et al. A sigmoid-based car-following model to improve acceleration stability in traffic oscillation and following failure in free flow[J]. IEEE Transactions on

Intelligent Transportation Systems, 2024.

- [23] Alamsyah N, Yoga T P, Budiman B. Improving traffic density prediction using LSTM with parametric ReLU (PReLU) activation[J]. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 2024, 9(2): 154-160.
- [24] Li Y B, Li R C, Shi B, et al. Optimization of High Energy Efficiency for Self-Consistent Energy System in Highway Service Area via Simulated Annealing Algorithm-Genetic Algorithm [J]. Journal of Xi'an Jiaotong University, 2024, 58(01): 197-207+216.
- [25] Mustafa A S, Yussof S, Radzi N A M. Multi-objective simulated annealing for efficient task allocation in UAV-assisted edge computing for smart city traffic management[J]. IEEE Access, 2025.
- [26] Wijanarko H, Hisan V K, Fajri W, et al. Hybrid simulated annealing and random forest for traffic density prediction in VANETs[C]//2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT). IEEE, 2024: 377-383.
- [27] Yang H Q. Research of Urban Recurrent Congestion Evolution Based on Taxi GPS Data [D]. Heilongjiang: Harbin Institute of Technology, 2018.
- [28] Chen G, Chen L, Tan K, et al. Multimodal fusion intelligent driving fatigue monitoring and warning in police traffic scenarios[C]//2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM). IEEE, 2024: 574-579.
- [29] Ruan J T. Spatial-temporal dynamic model and load prediction of highway traffic flow based on clustering algorithm [D]. Hunan: Xiangtan University, 2023.
- [30] Fan B, Li P, Jin S, et al. Anomaly Detection Algorithm Based on Deep Learning and Gaussian Mixture Model [J]. Computer & Digital Engineering, 2024, 52(04): 989-994.
- [31] Li Shuai, Yang Liu, Zhao Xin hui. Prediction algorithm of short-term traffic congestion in urban area based on deep learning [J]. Science Technology and Engineering, 2023, 23(25) : 10866-10878.
- [32] Wang Q R, Gao H Y, Zhu C F. Research on highway traffic congestion prediction based on SA-GFSTCN[J]. Computer Engineering, 2024: 1-13.
- [33] Feng M, Zheng J, Ren J, et al. Towards big data analytics and mining for UK traffic accident analysis, visualization & prediction[C]//ICMLC 2020: 2020 12th International Conference on Machine Learning and Computing. New York: ACM, 2018: 225-229.
- [34] Pu Q K. Research on Freeway Traffic Congestion Detection Method Based on Multi-Source Data Fusion [D]. Chongqing: Chongqing University, 2021.
- [35] Ming X L. Research on Average Speed Prediction and Congestion Identification of Bus Lanes on Urban Trunk Road [D]. Chang'an University, Xi'an, China, 2023.