ARCHIVES OF TRANSPORT *ISSN (print):* *0866-9546*
Volume 76, Issue 4, 2025 ***e-ISSN (online):*** *2300-8830*
DOI: *10.61089/aot2025.zf4t0674*

URBAN ROAD TRAFFIC CONGESTION INDEX PREDICTION BASED ON A HYBRID LIGHTGBM-LSTM MODEL

Sha LIU[1], Jian-Jie GAO[2*], Yi-Lin HONG[3], Jun-Chao ZHOU [4]

[1, 2, 3] Intelligent Policing Key Laboratory of Sichuan Province, Sichuan Police College, Luzhou, China
[1] Chengdu Traffic Management Bureau, Chengdu, China
[4] School of Mechanical Engineering, Sichuan University of Science & Engineering, Sichuan Zigong, China

***Abstract**:*

Accurate and timely prediction of the urban traffic congestion index (TCI) is crucial for implementing proactive traffic management and alleviating urban congestion. To address the limitations of single models in capturing both complex temporal dependencies and high-dimensional feature interactions, this paper proposes a novel hybrid prediction framework that synergistically integrates a Long Short-Term Memory (LSTM) network and a Light Gradient Boosting Machine (LightGBM). The model is designed to perform dual-stream learning: the LSTM module extracts medium- and long-term temporal patterns from historical TCI sequences, while the LightGBM module concurrently learns discriminative feature representations from the structured traffic data. A genetic algorithm (GA) is employed to optimize the fusion weights of the two components, constructing an adaptive and cohesive LightGBM-LSTM prediction model. The proposed framework was validated using real-world TCI data collected from three representative segments with varying congestion levels (mild, moderate, and severe) on Chengdu's Third Ring Road, covering a period from September to October 2024. The experimental results demonstrate that the hybrid model significantly outperforms both standalone LSTM and LightGBM baselines across all test scenarios. Specifically, it achieved accuracy improvements of 4.87% and 33.06% in mildly congested sections, 26.80% and 22.32% in moderately congested sections, and 47.87% and 10.47% in severely congested sections, respectively, measured by the Mean Absolute Percentage Error (MAPE). These findings confirm that the proposed GA-optimized LightGBM-LSTM hybrid model effectively enhances TCI prediction precision and robustness by leveraging complementary strengths of sequence learning and feature engineering. The study provides a reliable and efficient analytical tool for short-term traffic state forecasting, offering valuable support for the development of data-driven and refined urban traffic management strategies.

Keywords: *traffic prediction, hybrid model, LightGBM, LSTM, traffic congestion index*

To cite this article:

Liu, S., Gao J., Hong, Y., Zhou J., (2025). Urban road traffic congestion index prediction based on a hybrid LightGBM-LSTM model. Archives of Transport, 76(4), 175-190. https://doi.org/10.61089/aot2025.zf4t0674

Contact:
1) quick666s@163.com [https://orcid.org/0009-0003-8690-8748]; 2) gjj919323@163.com [https://orcid.org/0009-0000-2567-3373] – corresponding author; 3) yilin@scpolicec.edu.cn [https://orcid.org/0009-0002-8036-5204]; 4) zhou1987g@163.com [https://orcid.org/0000-0002-5747-3517]

Article is available in open access and licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0)

1. Introduction

The rapid pace of urbanization worldwide has placed unprecedented strain on urban transportation networks, leading to severe traffic congestion that undermines economic efficiency, environmental sustainability, and quality of life. In China, the challenge is particularly acute, with motor vehicle ownership exceeding 453 million and continuing to grow (Ministry of Public Security of the People's Republic of China, 2025,01,18). To manage this complexity, cities like Beijing and Shenzhen have adopted macro-level indicators such as the Traffic Congestion Index (TCI) to quantitatively evaluate and monitor network-wide performance. Accurate, short-term prediction of the TCI is therefore a critical task for enabling proactive traffic management and providing reliable traveler information.

In recent years, the rapid advancement of machine learning has provided powerful tools for processing large-scale, complex nonlinear datasets. Furthermore, deep learning, as an advanced extension of machine learning, offers enhanced capability in uncovering underlying patterns and logical relationships embedded in such data(Shi et al., 2020). While significant progress has been made in traffic forecasting—evolving from statistical models to advanced machine learning and deep learning techniques—achieving high accuracy for a complex, macro-level index like the TCI remains challenging. Standalone models, including the powerful Long Short-Term Memory (LSTM) network for temporal sequences and the efficient Light Gradient Boosting Machine (LightGBM) for feature-based learning, exhibit complementary strengths but also inherent limitations. LSTM models may overlook critical feature interactions, while LightGBM models often fail to fully capture complex temporal dependencies. Although hybrid models that combine different architectures have shown promise, most existing approaches have not deeply integrated the distinct, complementary advantages of tree-based ensembles and recurrent neural networks into a cohesive, optimally weighted framework.

To bridge this gap, this paper proposes a novel hybrid LightGBM-LSTM model for the precise prediction of the urban Traffic Congestion Index. The core innovation lies in the synergistic fusion of two components: the LightGBM module extracts high-level, discriminative features from the historical data, while the LSTM module learns the underlying medium- and long-term temporal dynamics. A genetic algorithm is employed to optimize the weight coefficients for integrating the outputs of these two components, creating a unified and robust prediction model. The primary contributions of this work are threefold:

1. We propose a novel hybrid forecasting architecture that deeply integrates LightGBM and LSTM to simultaneously leverage feature interactions and sequential patterns.
2. We introduce a genetic algorithm-based optimization strategy to determine the optimal fusion weights between the two model components, enhancing overall predictive performance.
3. We validate the proposed model using real-world TCI data from a major Chinese city, demonstrating its superiority over several standalone and benchmark hybrid models through comprehensive experiments.

The remainder of this paper is organized as follows: Section 2 provides a structured review of related work. Section 3 details the proposed LightGBM-LSTM methodology. Section 4 presents the experimental setup, results, and discussion. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Literature Review

Accurate traffic prediction is fundamental for intelligent transportation systems. Recent research has evolved from traditional statistical methods to sophisticated machine learning and hybrid approaches. This section reviews these methodological advancements with a focus on their application to traffic forecasting, culminating in the identification of the specific research gap addressed by this study.

2.1. Evolution of Traffic Prediction Methodologies

Early traffic prediction heavily relied on statistical time-series models, such as historical averages(Anitha et al., 2019), the Autoregressive Integrated Moving Average (ARIMA) (Dissanayake et al., 2021), Hidden Markov(Zhao et al., 2019). While computationally efficient, these parametric models often struggle with the non-linear and non-stationary nature of real-world traffic data, leading to limited

accuracy. The advent of machine learning introduced more flexible non-parametric models. However, shallow neural networks and support vector machines can be insufficient for capturing the complex, high-dimensional patterns in large-scale traffic datasets(Kuang et al., 2020).

The breakthrough came with deep learning, which excels at automatic feature extraction from raw data. In particular, Recurrent Neural Networks (RNNs) are naturally suited for sequential data. To overcomethe vanishing gradient problem in standard RNNs, the Long Short-Term Memory (LSTM) network was introduced (Hochreiter et al., 1997) and has since become a cornerstone for time-series forecasting. Its suitability for time-series traffic data has been demonstrated in various prediction applications. For example, Wang et al.(Ma et al., 2015)were among the first to introduce LSTM into traffic speed prediction, noting its advantages in handling nonlinear patterns and long-term dependencies. Subsequent studies by Zhong et al.(Zhong et al., 2018). Deep learning (DL) Neural Network (NN) approaches have been proven to extract temporal features from time series data, including Convolutional Neural Networks and Recurrent Neural Networks(Zhang et al., 2023). Approaches using DL algorithms have been focused on short-term traffic flow prediction(Gu et al., 2019).

LightGBM (Light Gradient Boosting Machine) is a training algorithm based on Gradient Boosting Decision Tree (GBDT), characterized by fast training speed, low memory consumption, and high prediction accuracy(Alafate et al., 2019).While GBDT struggles with maintaining accuracy and efficiency at large data scales(Kadiyala et al., 2018), LightGBM achieves significant improvements in training efficiency, memory usage, and scalability. It can directly process categorical features while effectively avoiding dimensionality explosion.

While deep learning models have improved traffic prediction accuracy, their growing complexity increases computational cost and tuning difficulty(Kumar et al., 2024). High-dimensional spatiotemporal data further exacerbates this optimization challenge. Although LightGBM has shown strong potential in time-series forecasting due to its efficiency, its deep integration for traffic congestion index prediction remains under-explored(F. Li et al., 2024).

2.2. Hybrid Models for Enhanced Traffic Forecasting

Platforms such as Baidu and Gaud provide congestion index to guide travel. However, the existence of a delay makes it unrealistic to avoid congestion in advance. Thus, it's really important to predict the congestion index in real-time.

The limitation of collection technology results in a lack of traffic characteristics, which leads to prediction errors. To solve this problem, it can be considered to enhance and generate features based on the original dataset. Yang Zhao et al.(Yang et al., 2021) reported a hybrid LSTM-GCN model with three-way temporal features (SCLN-TTF), leveraging Long Short-Term Memory Network(LSTM) to extract temporal features, and graph convolutional neural networks to extract spatial features; He et al.(He et al., 2021) introduced a technique for prediction based on trees and deep learning (DL). The essay described a means of feature generation, which is of great significance for feature extraction and generation. In a word, deep learning and tree-based theory can infer new features with limited features in the dataset, and have the ability to create features. In(L. Li et al., 2020), the LSTM-SPRVM model and the fuzzy comprehensive evaluation-based method were leveraged to predict and rank the congestion, and a traffic congestion prediction and visualization framework based on machine learning and a fuzzy comprehensive evaluation-MF-TCPV was proposed. Note that the framework can visually observe the results of congestion prediction, and the accuracy is also preferable.

With the increasing availability of large-scale datasets, research in traffic congestion prediction has progressively shifted towards deep learning methodologies(Cheng et al., 2022). Proposed hybrid models in the literature(Chu et al., 2020) demonstrate that these approaches yield substantial enhancements in prediction accuracy compared to conventional methods.

2.3. Identified Research Gap and Contribution

By synthesizing the literature reviewed above, a progressive trend is evident: research has evolved from standalone models to hybrid frameworks that attempt to combine algorithmic strengths. However, critical limitations persist: (1) Tree-based models (e.g., LightGBM) excel at feature learning but often neglect temporal dynamics when used in isolation in

traffic studies(N. Li et al., 2023); (2) Deep sequence models (e.g., LSTM) capture temporal patterns but may underutilize the powerful feature representation capabilities of tree ensembles(Cheng et al., 2022); (3) Most existing hybrids lack a principled and optimized mechanism to fuse these fundamentally different model families into a cohesive predictive engine(Cao et al., 2021).

This paper is motivated by the circumstance that the complementary strengths of LightGBM and LSTM have not been fully exploited within a unified framework for traffic prediction. Specifically, the question of how to optimally integrate feature-based learning with temporal sequence modeling to achieve superior prediction performance remains open.

To address these gaps and achieve more accurate and robust traffic congestion index (TCI) prediction, this study proposes a novel hybrid LightGBM-LSTM model. The model is designed to perform dual-stream learning: the LightGBM component extracts high-level discriminative features from historical and contextual data, while the LSTM component captures complex medium- and long-term temporal dependencies. Furthermore, a genetic algorithm is employed to optimize the fusion weights, ensuring an adaptive and effective integration of both components' outputs.

3. Methodology and Data

3.1. Definition of Traffic Congestion Index Prediction

The problem of traffic congestion prediction can be defined as forecasting future congestion conditions based on historical traffic data. To quantify the severity of congestion, the concept of the Traffic Congestion Index (TCI, also referred to as the Traffic Index or Traffic Performance Index) has been introduced. The TCI is a comprehensive metric whose values correspond to different levels of service, ranging from free-flow conditions to various degrees of congestion. A higher index value indicates more severe traffic congestion. This indicator provides an intuitive reflection and description of the operational status of urban roads or road networks (Specifications for urban traffic performance evaluation, 2016).

The Traffic Congestion Index reflects both the average travel speed on roads in a specific area and the public's perception of traffic congestion conditions. Compared to traditional traffic parameter indicators, it offers greater intuitiveness and clarity. When selecting parameters for constructing such an index, it is essential to ensure that the data is readily obtainable. In addition to conventional data collection techniques, leveraging open data sources—such as Amap, Baidu Maps, and related platforms—is highly recommended. This approach helps alleviate the challenges associated with limited data acquisition capabilities and facilitates more efficient and scalable traffic monitoring(Cheng et al., 2022). Therefore, this study adopts the Traffic Congestion Index provided by the Baidu Map Traffic Information Release and Analysis Platform as the foundational data metric for classifying traffic operational characteristics and predicting road traffic conditions.

The conversion formula between the Baidu Real-time Congestion Index and other traffic evaluation metrics is presented in Equation (1):

$$\mathrm{TPI} = F(\mathrm{TTR}) \tag{1}$$

Where TPI denotes the Traffic Performance Index, and TTR represents the Travel Time Ratio. The calculation formula for TTR is given in Equation (2).

$$\mathrm{TTR} = \frac{\overline{t_{kj}}}{t_j} \tag{2}$$

Where $\bar{t}_{kj}$ denotes the average time taken for vehicles to traverse road segment j during time interval k, and, $\bar{t}_{kj}$represents the travel time under free-flow conditions on the same segment. The classification of traffic states according to the Baidu Map Smart Transportation Platform is summarized in Table 1.

Table 1. Traffic State Classification Criteria on the Baidu Platform

Traffic State	Average Speed (km/h)	Traffic Performance Index (TPI)	Traffic State Level
Smooth	V≥44	TPI≤1.5	1
Congested	33≤V<44	1.5<TPI≤2	2
Heavy Congestion	16≤V<33	2<TPI≤4	3
Severe Congestion	V<16	TPI>4	4

3.2. The Proposed LightGBM-LSTM Hybrid Model

For traffic congestion index prediction, a hybrid LightGBM-LSTM prediction model is constructed, which consists of three core components: an LSTM module, a LightGBM module, and an integration module.

3.2.1. LSTM Module

The Long Short-Term Memory (LSTM) model is a special type of recurrent neural network (RNN) specifically designed to address the gradient vanishing and explosion problems commonly encountered in traditional RNNs when processing long sequence data. Its core innovation lies in the use of gating mechanisms and a cell state to achieve long-term memory retention and dynamic control of information flow. Building upon the traditional RNN structure, the LSTM introduces a cell state and three gating mechanisms: The Forget Gate determines which information from the previous time step should be discarded. The Input Gate controls whether new information from the current input should be stored. The Output Gate regulates the amount of information output from the current cell state.

Through these gating mechanisms, the LSTM selectively forgets and updates information, effectively mitigating the gradient-related issues of traditional RNNs and significantly enhancing modeling capability for time series data. The cell state acts as a continuous information pathway throughout all time steps, functioning like a "conveyor belt" that stores and transfers long-term memory. The internal structure of the LSTM cell is illustrated in Figure 1.

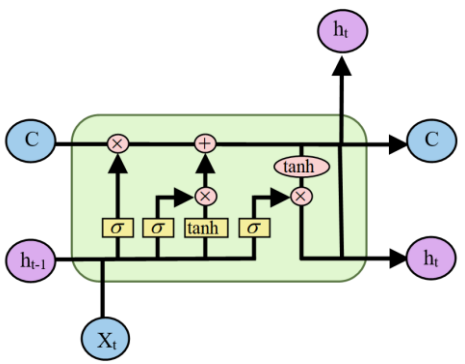


Fig. 1. Schematic Diagram of the Memory Cell Structure

The specific procedural steps of the LSTM algorithm are illustrated in Figure 2. In this process, the Epoch controls the overall number of training cycles, determining the depth of the model's learning from the data. The Iteration determines the frequency of parameter updates within each epoch, influencing both training stability and convergence speed. Through multiple epochs of training and continuous parameter optimization during each iteration, the model progressively reduces prediction errors and enhances overall performance.

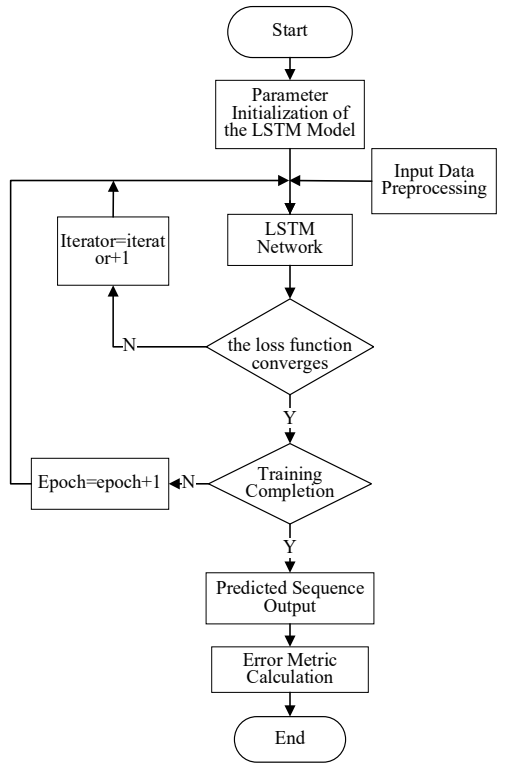


Fig. 2. Detailed Flowchart of the LSTM Algorithm

Owing to its unique memory structure, the LSTM model demonstrates significant advantages in the field of time series prediction, making it particularly suitable for scenarios such as traffic congestion index forecasting, which exhibits strong temporal dependencies and complex dynamic characteristics. Through the synergistic action of the cell state and gating mechanisms, the LSTM effectively captures

long-term patterns (e.g., periodic morning and evening peaks) and short-term fluctuations (e.g., impacts of unexpected incidents) in traffic data, while also mitigating the vanishing gradient problem commonly encountered in traditional recurrent neural networks.

3.2.2. LightGBM Module

The LightGBM algorithm is an efficient, optimized version of GBDT (Gradient Boosting Decision Tree). It incorporates several advanced techniques, including:

- Histogram-based Algorithm: Accelerates training by bucketing continuous features into discrete bins.
- Gradient-based One-Side Sampling (GOSS): Retains samples with large gradients and randomly discards those with small gradients to reduce data size.
- Exclusive Feature Bundling (EFB): Bundles mutually exclusive sparse features to reduce dimensionality.
- Leaf-wise Growth Strategy: Splits the leaf with maximum gain at each step, creating deeper trees for lower loss.

Additionally, LightGBM natively supports categorical features and feature parallelism, significantly enhancing efficiency and performance while reducing memory consumption. The overall workflow is illustrated in Figure 3.

3.3. Integration with Genetic Algorithm

By integrating multiple highly accurate prediction models, it is possible to leverage the strengths of each model while mitigating their individual limitations, thereby enhancing overall predictive performance. To address the shortcomings of single-model approaches, this study adopts an LSTM-LightGBM hybrid model that combines LSTM's capability to capture temporal trends with LightGBM's proficiency in processing historical features. This fusion leverages the advantages of both methods to improve the accuracy of short-term traffic congestion index forecasting.

Specifically, the hybrid model employs a linear weighting mechanism for integration, with optimal weight coefficients determined using a genetic algorithm. The mathematical formulation of the combined model is given by Equation (3):

$$f_{com} = w_1 f_{LightGBM} + w_2 f_{LSTM} \tag{3}$$

where:

w_1 denotes the weight coefficient for the LightGBM model,

w_2 denotes the weight coefficient for the LSTM model,

and the weights satisfy the constraint: $w_1+w_2=1$.

The Genetic Algorithm (GA) is a global optimization method inspired by natural selection and genetic mechanisms, widely applied to solve complex problems—especially those that are nonlinear, multimodal, or combinatorial in nature. Drawing on principles from biological evolution, GA effectively explores complex search spaces by simulating genetic operations such as selection, crossover, and mutation.

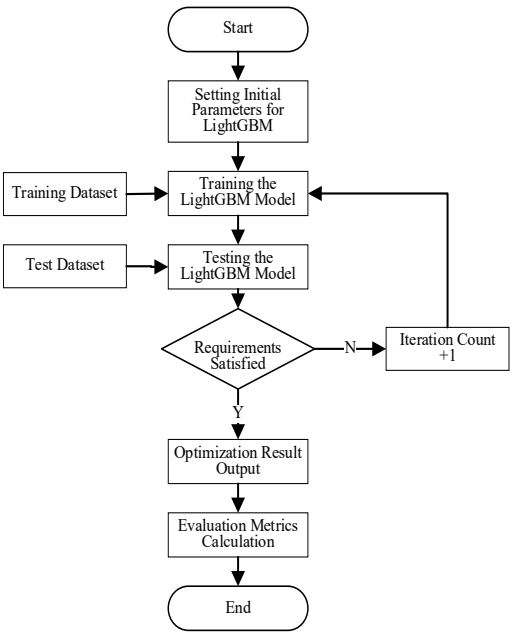


Fig. 3. Overall Workflow Diagram of the LightGBM Algorithm

Therefore, the optimal weight coefficients for the hybrid model can be determined using a genetic algorithm. The specific steps of the optimization process are summarized in Table 2.

Table 2. Steps for Calculating Optimal Weight Coefficients Using Genetic Algorithm

Step	Name	Description
1	Population Initialization	Randomly initialize multiple individuals, each representing a weight combination(w1,w2), satisfying w1+w2=1
2	Fitness Calculation	Compute the combined prediction result $f_{com} = w_1 f_{LightGBM} + w_2 f_{LSTM}$, and use the Mean Squared Error (MSE) as the fitness value.
3	Selection	Apply roulette wheel selection based on fitness values to choose superior individuals for the next generation.
4	Crossover	Perform simulated binary crossover on selected individuals to generate new weight combinations.
5	Mutation	Apply Gaussian mutation to a subset of individuals to increase diversity and avoid local optima.
6	Termination Condition	Stop evolution when the maximum number of generations (e.g., 100) is reached or the error converges (e.g., change<threshold for 10 consecutive generations).
7	Optimal Weight Extraction	Select the individual with the highest fitness as the optimal weight combination (w1*, w2*) for the final LSTM-LightGBM hybrid prediction.

3.4. Data Preparation and Experimental Setup

3.4.1. Data Preparation

Currently, numerous online map platforms leverage internet technology and big data analytics to deeply mine and process massive amounts of travel and location data, providing the public with real-time traffic information. For example, platforms such as Baidu Map Smart Transportation and Amap (Gaode Map) update every five minutes, displaying information about the top ten most congested road segments. The details include road names, congestion delay index, travel speed, travel time, and delay duration, among other metrics.(Hochreiter et al., 1997) This study utilizes traffic congestion index data from the Baidu Map Smart Transportation Analysis Platform for representative segments of Chengdu's Third Ring Road:

- Severely Congested Area: Western Third Ring Road, Section 3
- Moderately Congested Area: Southern Third Ring Road, Section 4
- Mildly Congested Area: Eastern Third Ring Road, Section 1

Based on this data, machine learning models are constructed to predict the traffic operational status. The experimental dataset consists of traffic congestion index values collected from September 1, 2024, to October 12, 2024(a sample is shown in Table 3). The dataset is divided as follows:

- Training set: The first 80% of the data.
- Validation set: The subsequent 20% of the data.
- Test set: Data from October 13, 2024(reserved for final evaluation).

The prediction task covers a full 24-hour period with a time granularity of 3 minutes.

Input Features

The model input includes the following features to capture spatiotemporal patterns:

1) **Road Segment Identifier**: Specific road segment (e.g., Western Third Ring Road Section 3).
2) **Congestion Level**: Severe, moderate, or mild.
3) **Historical Traffic Data**: Historical traffic congestion index values.
4) **Temporal Context**:

- Month and day of the week.
- Time period within the day (defined below).
- Binary indicator for weekends or public holidays.

Definition of Time Periods (Peak/Off-Peak Hours)

- Weekday Morning Peak:7:00–9:00
- Weekday Evening Peak:17:00–19:00
- Weekend/Holiday Morning Peak:10:00–12:00
- Weekend/Holiday Evening Peak:16:00–18:00
- Off-Peak Hours: All other time periods

This structured input enables the model to capture complex spatiotemporal patterns and contextual factors affecting traffic congestion. Subsequently, an empirical simulation is conducted to visually compare the model's predictions with actual observed values.

Table 3. Example of Traffic Congestion Index Data for Sections of Chengdu's Third Ring Road

Date	Time	Road Name	Road Length (km)	Congestion Index	Avg Speed (km/h)	Congested Length (km)
……	……	……	……	……	……	……
2024/9/29	8:24:00	Southern Third Ring Road, Section 4	5.678	1.832	45.43	0.052
2024/9/29	8:27:00		5.678	2.06	40.402	0.258
2024/9/29	8:30:00		5.678	2.105	39.538	0.412
2024/9/29	8:33:00		5.678	2.148	38.747	0.465
2024/9/29	8:36:00		5.678	2.06	40.402	0.315
2024/9/29	8:39:00		5.678	2.203	37.779	0.446
2024/9/29	8:42:00		5.678	2.247	37.04	0.494
2024/9/29	8:45:00		5.678	2.258	36.859	0.515
2024/9/29	8:48:00		5.678	2.352	35.386	0.663
2024/9/29	8:51:00		5.678	2.419	34.406	0.767
2024/9/29	8:54:00		5.678	2.43	34.25	0.776
……	……	……	……	……	……	……

3.4.2. Evaluation Metrics

This study employs the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE) as evaluation metrics. Among them, MAPE is an error metric commonly used in regression tasks, representing the proportion of prediction error relative to the true value. A smaller MAPE value indicates smaller model error and higher prediction accuracy. Its mathematical expression is shown in Equation (4):

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \tag{4}$$

where:
$\hat{y}_i$ denotes the predicted value of the i-th sample,
y_i denotes the actual value of the i-th sample,
n represents the total number of samples.
Meanwhile, RMSE serves as the fitness criterion, where a smaller value indicates better adaptability of an individual to the environment. Its mathematical expression is given by Equation (5):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(100y_i - 100y_{i'})^2} \tag{5}$$

where:
$y_{i'}$ represents the predicted value of the i-th sample,
y_i represents the actual value of the i-th sample,
N denotes the total number of samples.

To ensure the reproducibility of the experiments and to provide context for the computational efficiency discussed later, the hardware and software configurations used for model training and evaluation in this study are detailed in Table 4.

Table 4. Experimental Environment Configuration

Item	OS	CPU	RAM	IDE	Storage
Specification	Windows10	AMD Ryzen5	8GB	VSCode	128GB

4. Result Analysis and Performance Comparison

4.1. Analysis of Single Model Results

To preliminarily and intuitively assess the predictive capability of our proposed hybrid model, we first present a visual comparison between the predicted and the actual observed traffic congestion index (TCI) over a continuous time series. Figure 4-6 illustrates this comparison for a selected test period, allowing for an initial observation of how closely the model's output follows the real traffic fluctuations.
(1) Eastern Third Ring Road, Section 1
As shown in the prediction visualization in Figure 4, the predicted values from the LSTM model (red dashed line) align more closely with the true values (blue curve). Particularly during morning and evening peak hours, the LSTM demonstrates a stronger capability to capture abrupt changes in the traffic index. In contrast, the predictions from LightGBM (green dotted curve) exhibit an overall smoother trend. Although LightGBM shows some deviations at peak points, it still maintains a reasonable fit to the overall trend of the ground truth data.

(2) Southern Third Ring Road, Section 4 (Moderately Congested Area)

As observed in the prediction visualization of Figure 5, the predicted values from the LSTM model (red dashed line) demonstrate a strong alignment with the ground truth (blue curve). The model effectively captures the rapid upward trends in traffic congestion during both morning and evening peak hours, despite slight overestimation during certain intervals, such as multiple spikes in the evening peak. Overall, the prediction trend closely matches the actual data.

In contrast, the predictions generated by the LightGBM model (green dotted line) exhibit noticeable lag during peak periods, leading to underestimation of congestion levels in some high-traffic intervals. Additionally, its predictions during off-peak hours (e.g., 0:00–6:00) display minor fluctuations and occasional overestimation in stable periods, indicating inconsistent performance across varying traffic conditions.

(3) Western Third Ring Road, Section 3 (Severely Congested Area)

As can be seen from Figure 6, the predicted values of the LSTM model (red dashed line) accurately capture multiple traffic peaks, demonstrating particularly strong performance during morning and evening rush hours. The overall prediction trend remains stable throughout the day without significant abnormal fluctuations, although slight overshooting (overestimation) occurs in certain localized segments.

In contrast, the predictions of the LightGBM model (green dotted line) exhibit noticeable lag during peak periods and generally underestimate peak values. This indicates LightGBM's limited ability to model abrupt changes under severe congestion conditions. Additionally, anomalous jumps appear during some off-peak intervals, primarily due to its difficulty in capturing complex nonlinear temporal dependencies.

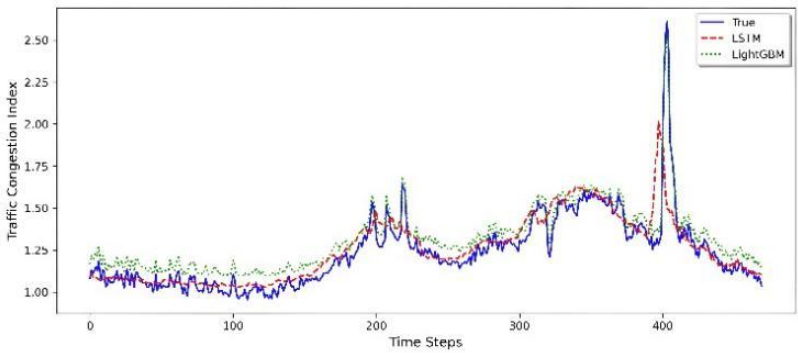


Fig. 4. Visualization of LightGBM and LSTM Predictions in Mildly Congested Sections

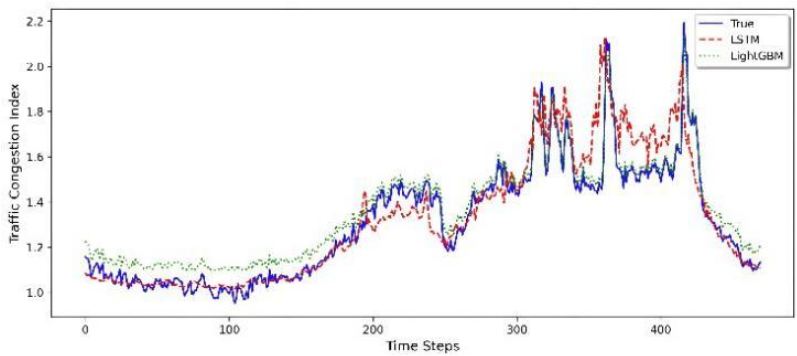


Fig. 5. Visualization of LightGBM and LSTM Predictions in Moderately Congested Sections

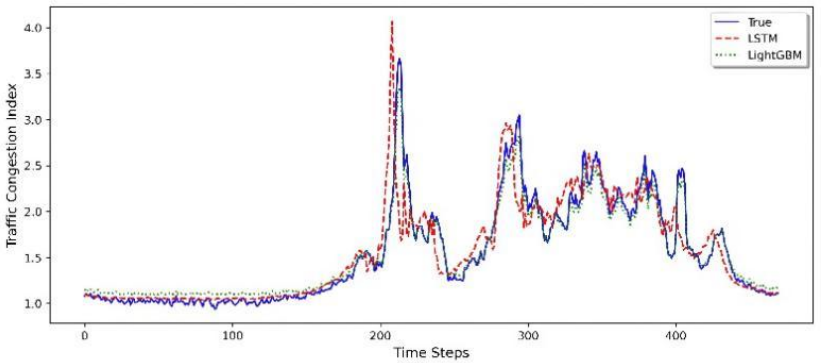


Fig. 6. Visualization of LightGBM and LSTM Predictions in Severely Congested Sections

Table 5. Mean Absolute Percentage Error (MAPE) Across Different Congestion Levels (%)

Congestion Type	LSTM	LightGBM
Eastern Ring Rd Sec1 (Mild)	4.80	6.82
Southern Ring Rd Sec4 (Moderate)	4.87	4.59
Western Ring Rd Sec3 (Severe)	8.09	4.71

Using MAPE as the evaluation metric, the predictive performance of the LightGBM and LSTM models for traffic congestion index across different types of road segments is compared in Table 5. The following observations can be made:

In the Eastern Third Ring Road, Section 1 (mild congestion), the error of LSTM is 4.80%, significantly lower than LightGBM's 6.82%. This indicates that LSTM is more effective at capturing overall trends and slight fluctuations, particularly excelling in accurately predicting minor variations during peak and off-peak hours. LightGBM, in contrast, shows higher error in this scenario, likely due to its limited ability to model subtle nonlinear variations in mild congestion conditions.

In the Southern Third Ring Road, Section 4 (moderate congestion), the prediction errors of both models are relatively close, with LSTM at 4.87% and LightGBM at 4.59%. Although LightGBM exhibits a slightly lower error, LSTM demonstrates stronger capabilities in temporal modeling, effectively capturing overall trends even in moderately fluctuating segments. This suggests complementary strengths between the two models for such scenarios, allowing flexible model selection based on real-time requirements or specific fluctuation characteristics.

In the Western Third Ring Road, Section 3 (severe congestion), the error of LSTM increases significantly to 8.09%, substantially higher than LightGBM's 4.71%. This reveals that under conditions of intense congestion with frequent and sharp fluctuations, LSTM tends to exhibit prediction overshooting. While LightGBM achieves a lower overall error, it lacks responsiveness to sudden changes, potentially underestimating the severity of congestion during peak periods. Thus, both models show certain limitations when applied individually in severely congested areas, highlighting the need for a combined approach to enhance prediction accuracy and robustness.

To improve the accuracy of traffic congestion index forecasting, this study proposes a hybrid model integrating LSTM and LightGBM to leverage their complementary advantages. The specific complementary benefits of this combined approach are summarized in Table 6.

4.2. Construction of Hybrid Prediction Model and Forecasting Results

The strategy for constructing the hybrid model involves using a genetic algorithm to determine the optimal weight coefficients w_1 and w_2 for the linear weighted combination of the LightGBM and LSTM models. The Root Mean Square Error (RMSE) is employed as the fitness criterion, where a smaller value indicates better adaptability of an individual solution to the environment. The iterative process of optimizing the weights is illustrated in Figure 7.

Table 6. Comparison of Advantages and Disadvantages among LSTM, LightGBM, and the Hybrid Model

Model	Advantages	Disadvantages	Applicable Scenarios
LSTM	Suitable for long-term trend prediction; accurately captures peak congestion	High computational cost; may overshoot in some peak predictions	All-day traffic prediction, especially for long-term trend analysis
LightGBM	Fast computation speed; suitable for short-term prediction	Lag in peak predictions; tends to underestimate sudden congestion	Short-term rapid prediction, such as immediate congestion warning
LightGBM-LSTM Hybrid	Combines long-term learning ability and fast computation; balances errors	Requires careful weight adjustment between models	All-day traffic prediction + short-term rapid correction

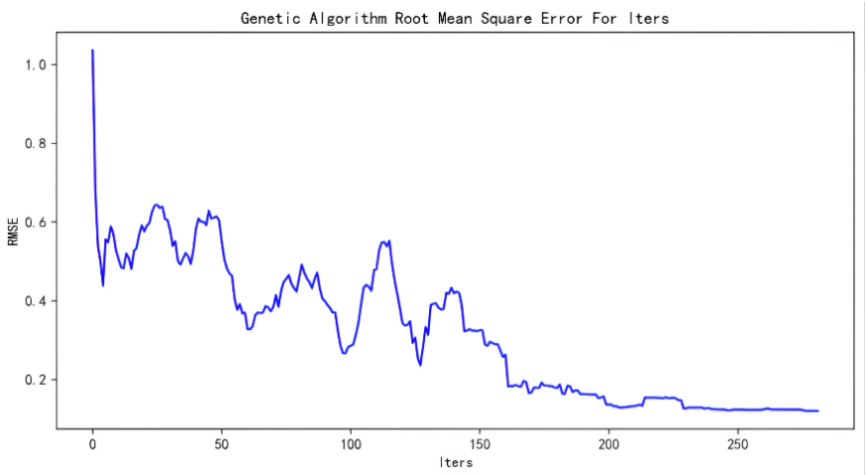


Fig. 7. Iterative Optimization of Weight Coefficients Based on Root Mean Square Error (RMSE) Evaluation

As can be seen from the figure, when the population iteration reaches 200 generations, the RMSE decrease stabilizes, indicating error convergence. The optimal weight coefficients are obtained as w=[0.6875，0.3125], thereby achieving the hybrid prediction model as shown in Equation (6):

$$f_{com} = 0.6875 f_{LightGBM} + 0.3125 f_{LSTM} \quad (6)$$

Through simulation experiments, the LSTM-LightGBM hybrid model was applied to predict traffic congestion indices for the three road segments. The predicted values were compared visually with the actual values, and the experimental results are presented below.

(1) Eastern Third Ring Road, Section 1 (Mild Congestion Area)

As shown in Figure 8, the prediction curve of the LSTM-LightGBM hybrid model closely follows the overall trend of the ground truth values. The predictions of the hybrid model (red dashed line) align highly with the actual values (blue curve) across most time periods, indicating a low overall prediction error. During several critical intervals, particularly the morning peak (7:30–9:00) and evening peak (17:30–19:00), the model accurately captures the rapid increases and decreases in the congestion index, significantly outperforming the predictions of standalone LSTM or LightGBM models.

Furthermore, during off-peak hours, such as from 0:00–6:00 and after 21:00, the hybrid model demonstrates stable performance. The prediction curve shows no significant jumps or abnormal fluctuations and remains closely aligned with the true values, reflecting its strong fitting capability. This indicates that the LSTM-LightGBM hybrid model not only enhances responsiveness to sudden changes during peak hours but also maintains high stability and robustness under low-congestion conditions, making it suitable for all-day traffic prediction tasks in mildly congested areas.

(1) Southern Third Ring Road, Section 4 (Moderate Congestion Area)

As shown in Figure 9, the prediction curve of the LSTM-LightGBM hybrid model closely aligns with the overall trend of the actual values. The predictions of the hybrid model (red dashed line) remain highly consistent with the true values (blue curve) during most time periods, demonstrating its strong predictive capability in moderately congested scenarios.

During peak hours, specifically the morning peak (approximately 7:00–9:30) and evening peak (approximately 17:30–20:00), the hybrid model accurately tracks changes in the actual traffic index, effectively capturing multiple consecutive peaks and exhibiting excellent adaptability to rapid traffic fluctuations. Although minor prediction deviations occur in certain intervals, such as 15:00–16:30, the overall error is significantly reduced compared to standalone LSTM or LightGBM models.

Furthermore, during off-peak hours, including the early morning (0:00–6:00) and late night (after 21:00), the hybrid model maintains stable and smooth predictions without significant jumps or abnormal fluctuations. The predicted curve adheres closely to the actual values, highlighting its enhanced robustness and generalization ability.

(2) Western Third Ring Road, Section 3 (Severe Congestion Area)

As shown in Figure 10, the predicted values of the hybrid model (red dashed line) generally align well with the actual values (blue curve), indicating that the model possesses strong trend-fitting capability even under severe congestion scenarios and maintains stable prediction performance throughout the day. Particularly during peak hours (e.g., 7:30–9:00 and 17:30–20:00) and off-peak periods (e.g., 0:00–6:00), the prediction error is significantly reduced compared to that of individual models.

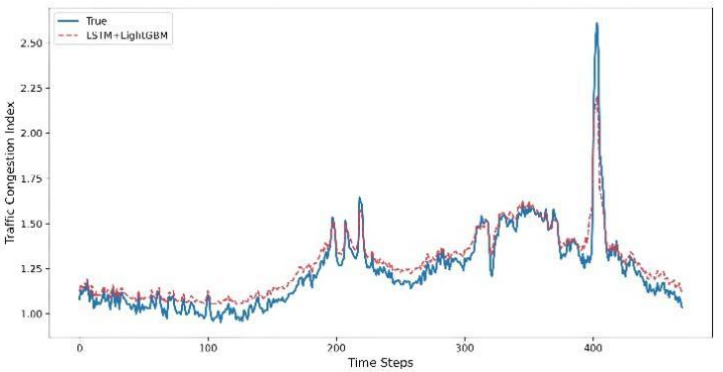


Fig. 8. Visualization of LSTM-LightGBM Hybrid Model Predictions in Mildly Congested Sections

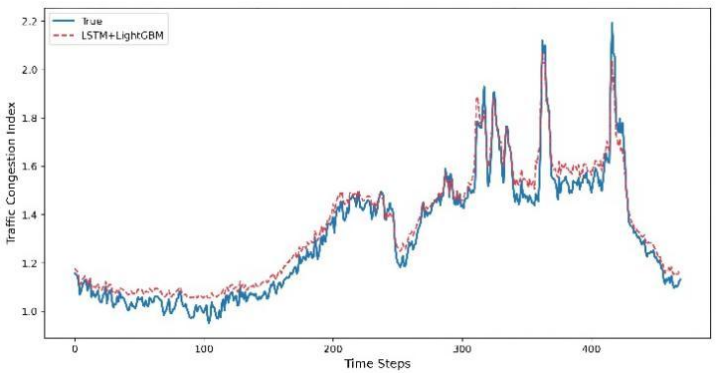


Fig. 9. Visualization of LSTM-LightGBM Hybrid Model Predictions in Moderately Congested Sections

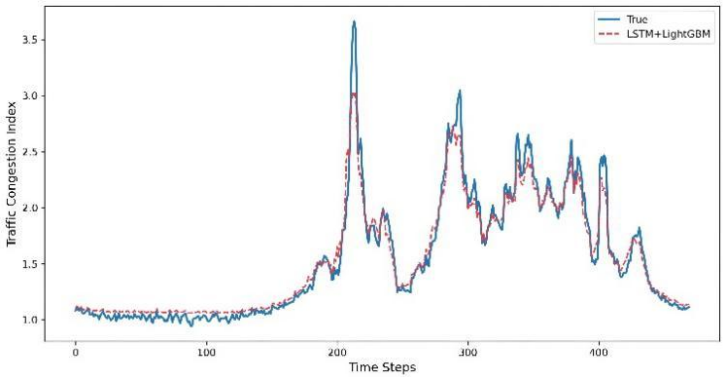


Fig. 10. Visualization of LSTM-LightGBM Hybrid Model Predictions in Severely Congested Sections

During peak intervals, the hybrid model effectively predicts the rising process of the traffic congestion index and captures abrupt changes in traffic conditions in a timely manner, thereby mitigating the lagging issue observed in the standalone LightGBM model. Meanwhile, compared to the potential overshooting problem of the LSTM model, the hybrid model provides smoother and better-fitting predictions at multiple peaks, demonstrating enhanced adaptability to drastic fluctuations.

Furthermore, during off-peak hours, such as late night after 22:00, the prediction curve shows no abrupt jumps and remains consistent with the actual data, highlighting the model's robustness and resistance to interference. Although the predictions slightly underestimate a few extreme peak values around 11:00 or 18:30, the error is substantially smaller compared to individual models, confirming that the hybrid model exhibits stronger responsiveness and generalization ability in handling sudden congestion.

Using the Mean Absolute Percentage Error (MAPE) as the evaluation metric for prediction performance, the results of the LightGBM model, LSTM model, and LightGBM-LSTM hybrid model across different congestion levels are summarized in Table 7.

The experimental results demonstrate that the LightGBM-LSTM hybrid model significantly improves prediction accuracy across all congestion regions. Compared to the individual LSTM and LightGBM models, the hybrid model achieves accuracy improvements of 4.87% and 33.06% in mildly congested areas, 26.80% and 22.32% in moderately congested areas, and most notably, 47.87% and 10.47% in severely congested areas. These results fully illustrate that the hybrid model effectively integrates the temporal modeling capabilities of LSTM with the feature representation strengths of LightGBM, exhibiting enhanced predictive power and robustness when dealing with complex and dynamic traffic environments.

Table 7. Prediction MAPE (%) of Different Models in Various Congestion Regions

Congestion Region	LSTM	LightGBM	LightGBM-LSTM
Eastern Ring Rd Sec1 (Mild)	4.80	6.82	4.5657
Southern Ring Rd Sec4 (Moderate)	4.87	4.59	3.5647
Western Ring Rd Sec3 (Severe)	8.09	4.71	4.2171

It can be observed that the LightGBM-LSTM hybrid model achieves significant improvements in accuracy compared to the individual models. The calculated accuracy enhancement of the hybrid model across various congestion regions is illustrated in Table 8.

Table 8. Prediction Accuracy Improvement of LightGBM-LSTM Hybrid Model in Different Congestion Regions (%)

Congestion Region	Improvement vs LSTM	Improvement vs LightGBM
Eastern Ring Rd Sec1 (Mild)	4.87	33.06
Southern Ring Rd Sec4 (Moderate)	26.80	22.32
Western Ring Rd Sec3 (Severe)	47.87	10.47

5. Conclusion

To address the limitations of low prediction accuracy when using standalone LSTM or LightGBM models for traffic congestion index forecasting, this study proposes a hybrid model that integrates the strengths of both approaches. The main conclusions are as follows:

As a deep learning model, LSTM excels at capturing long-term dependencies and effectively learns overall traffic flow trends. However, it may exhibit overshooting during peak traffic periods. In contrast, LightGBM shows advantages in short-term feature extraction and rapid response to anomalous data but often lags in responding to sudden congestion events due to its limited ability to model temporal dependencies.

By combining these two models, the proposed LightGBM-LSTM hybrid model effectively integrates LSTM's capacity for temporal pattern capture with LightGBM's strengths in short-term pattern recognition. The hybrid model significantly improves prediction accuracy across various congestion levels, demonstrating enhanced stability, timeliness, precision, and stronger generalization capability.

6. Limitations and Future Work

Despite its promising performance, this study has several limitations that should be acknowledged. Firstly, the model's development relied exclusively on data from specific segments of Chengdu's Third Ring Road over a limited timeframe. While effective for the studied context, this constrained scope necessitates caution regarding the model's generalizability to other urban road networks with distinct traffic patterns, infrastructure characteristics, or regional driving behaviors.

Secondly, and equally important, the current model does not incorporate various external factors known to significantly impact traffic conditions. Weather conditions (e.g., rain, snow, fog), traffic incidents, road construction, large-scale events, and temporary traffic control measures can cause substantial, often sudden, changes in congestion that are not fully captured by historical TCI data alone. The absence of these variables represents a limitation in modeling non-recurring congestion events.

Future research will therefore focus on two primary directions:

- Expanding Data Scope and Assessing Generalizability:Applying and validating the proposed hybrid framework across multiple cities and diverse road network types to thoroughly evaluate its transferability and robustness.
- Integrating Multi-Source Data: Developing mechanisms to effectively incorporate real-time and forecasted external data (e.g., weather, incident reports) into the model architecture. Exploring feature engineering and fusion techniques for these factors is expected to further enhance prediction accuracy, particularly for anomaly-driven congestion, leading to a more comprehensive and robust traffic forecasting system.

Acknowledgment

The work is supported by Sichuan Science and Technology Program under Grant No. 2024NSFSC2029. Supported by Intelligent Policing Key Laboratory of Sichuan Province, No. ZNJW2026ZZZD002.

References

1. Alafate, J., & Freund, Y. S. (2019). Faster boosting with smaller memory. Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper_files/paper/2019/hash/3ffebb08d23c609875d7177ee769a3e9-Abstract.html
2. Anitha, E. B., Aravinth, R., et al. (2019). Prediction of road traffic using naive bayes algorithm. Int. J. Eng. Res. Technol, 7(1), 1-4. https://doi.org/10.17577/IJERTCONV7IS01042
3. Cao, P., Dai, F., et al. (2021). A survey of traffic prediction based on deep neural network: Data, methods and challenges. Paper presented at the International conference on cloud computing. https://doi.org/10.1007/978-3-030-99191-3_2
4. Cheng, W., Li, J.-l., et al. (2022). Combination predicting model of traffic congestion index in weekdays based on LightGBM-GRU. Scientific reports, 12(1), 2912. https://doi.org/10.1038/s41598-022-06975-1

5. Chu, Z., Yu, J., et al. (2020). LPG-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis, and deep gated recurrent unit network. Information Sciences, 535, 107-129. https://doi.org/10.1016/j.ins.2020.05.042
6. Dissanayake, B., Hemachandra, O., et al. (2021). A comparison of ARIMAX, VAR and LSTM on multivariate short-term traffic volume forecasting. Paper presented at the Conference of open innovations association, FRUCT. https://doi.org/10.5281/zenodo.4514955
7. Gu, Y., Lu, W., et al. (2019). An improved Bayesian combination model for short-term traffic prediction with deep learning. IEEE Transactions on Intelligent Transportation Systems, 21(3), 1332-1342. https://doi.org/10.1109/TITS.2019.2939290
8. He, H., & Fan, Y. (2021). A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. Expert Systems with Applications, 176, 114899. https://doi.org/10.1016/j.eswa.2021.114899
9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
10. Kadiyala, A., Kumar, A., et al. (2018). Applications of python to evaluate the performance of decision tree-based boosting algorithms. Environmental Progress, 37(2), 618-623. https://doi.org/10.1002/ep.12888
11. Kuang, L., Hua, C., et al. (2020). Traffic volume prediction based on multi-sources GPS trajectory data by temporal convolutional network. Mobile Networks Applications, 25(4), 1405-1417. https://doi.org/10.1007/s11036-019-01458-6
12. Kumar, N., Martin, H., et al. (2024). Enhancing Deep Learning-Based City-Wide Traffic Prediction Pipelines Through Complexity Analysis. Data Science for Transportation, 6(3), 24. https://doi.org/10.1007/s42421-024-00109-x
13. Li, F., Nie, W., et al. (2024). Network traffic prediction based on PSO-LightGBM-TM. Computer Networks, 254, 110810. https://doi.org/10.1016/j.comnet.2024.110810
14. Li, L., Lin, H., et al. (2020). MF-TCPV: A machine learning and fuzzy comprehensive evaluation-based framework for traffic congestion prediction and visualization. Ieee Access, 8, 227113-227125. https://doi.org/10.1109/ACCESS.2020.3043582
15. Li, N., Zou, F., et al. (2023). Vehicle traveling speed prediction based on LightGBM algorithm. Paper presented at the International Conference on Genetic and Evolutionary Computing. https://doi.org/10.1007/978-981-99-9412-0_1
16. Ma, X., Tao, Z., et al. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, 187-197. https://doi.org/10.1016/j.trc.2015.03.014
17. Ministry of Public Security of the People's Republic of China. (2025,01,18). https://www.mps.gov.cn/n2254314/6409334/c9939035/content.html
18. Shi, X., Qi, H., et al. (2020). A spatial–temporal attention approach for traffic prediction. IEEE Transactions on Intelligent Transportation Systems, 22(8), 4909-4918. https://doi.org/10.1109/TITS.2020.2983651
19. Specifications for urban traffic performance evaluation. (2016). (Vol. 32). Beijing: Standards Press of China. GB/T33171-2016.
20. Yang, Z., Tang, R., et al. (2021). Short-term prediction of airway congestion index using machine learning methods. Transportation Research Part C: Emerging Technologies, 125, 103040. https://doi.org/10.1016/j.trc.2021.103040
21. Zhang, X., Huang, K., et al. (2023). Urban short-term traffic flow prediction algorithm based on cnn-lstm model. Paper presented at the 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE). https://doi.org/10.1109/ICCECE58074.2023.10135384
22. Zhao, S.-x., Wu, H.-w., et al. (2019). Traffic flow prediction based on optimized hidden Markov model. Paper presented at the Journal of Physics: Conference Series. https://doi.org/10.1088/1742-6596/1168/5/052001

23. Zhong, Y., Xie, X., et al. (2018). A new method for short-term traffic congestion forecasting based on LSTM. Paper presented at the IOP Conference Series: Materials Science and Engineering. https://doi.org/10.1088/1757-899X/383/1/012043