Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/83/2025.19939

Urban Traffic Flow Prediction Algorithm Based on Attention Mechanism and LSTM

Penghao Zeng[1,a,*]

[1]*Central South University, Changsha, China*
a. 2995407356@qq.com
**corresponding author*

Abstract: Traffic flow prediction is one of the important techniques for urban intelligent transport construction. Due to the instability and complexity of the actual traffic flow changes in urban road networks, existing prediction models are difficult to extract the flow characteristics of traffic sections that are geographically on different grids at different times with limited training. Regarding this kind of problem, this paper proposes a prediction method based on transformer attention mechanism and multilayer LSTM model, firstly, the temporal characteristics of data are processed by four-layer LSTM network, then the attention mechanism of multiple heads is used to make predictions on spatial location characteristics, finally, the prediction results analysed by both of them are subjected to full-connectivity processing and constant adjustment of weights, and the extracted traffic flow characteristics with temporal and spatial characteristics are finally processed and outputted. spatial characteristics of the city is finally processed and the prediction results are output. After the related prediction experiments on the grid flow data and the related evaluation indexes of the comparative model experiments, the prediction effect of this model is better than the traditional prediction method, and it can make effective long- and short-term flow prediction of the traffic flow of the urban roads.

Keywords: urban traffic flow prediction, attentional mechanism, multilayer LSTM model, temporal and spatial properties.

1. Introduction

With the rapid development of economic globalisation in recent years, the scale of urban transport systems has continued to expand, and the traffic flow at traffic intersections in many cities has increased significantly. In order to effectively solve traffic congestion, facilitate people's travel, and enhance the sense of well-being of the city, real-time and accurate traffic congestion prediction is very important. In this context, the development of accurate traffic flow prediction models is crucial for managers, and they can provide reliable data support. These models can accurately assess road congestion and help to optimise the allocation of traffic resources to the greatest extent possible. In this way, not only can the overall efficiency of traffic management be effectively improved, but also the level of service can be significantly enhanced.

Domestic and international research in the field about traffic flow prediction is constantly deepening, seeking to find excellent ways to improve the prediction accuracy and speed of short-term traffic flow, and at present, the prediction methods are mainly divided into linear theoretical models

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/83/2025.19939

based on traditional statistics, nonlinear theoretical models, data-driven machine learning methods and deep learning based methods.

2. Background on relevant technological developments

Traditional statistical methods usually use qualitative and quantitative prediction, and typical time series models include Autoregressive Integrated Moving Average Model (ARIMA), Kalman filter model, such as Shen et al. proposed an adaptive Kalman filter model that can automatically adjust with the change of traffic flow level; Bayesian filter model, and Bayesian filter model, which can automatically adjust with the change of traffic flow level. Kalman filter model; Bayesian model, such as the improved Bayesian model proposed by Wang et al. to solve the defect of the long iteration step length of the weight calculation; nonlinear theoretical models based on the wavelet theoretical model, chaos theoretical model, nonparametric regression model, etc., based on the prediction method of machine learning, there are mainly neural network (NN) model, support vector machine (SVM), support vector machine (SVM), support vector machine (SVM), support vector machine (SVM), support vector machine (SVM), support vector machine (SVM), support vector machine (SVM), support vector machine (SVM), and so on. Support Vector Machine (SVM) model, such as Xu et al. proposed the ant colony optimisation based SVM to optimise the SVM parameters for short-term traffic flow prediction, and Li et al. proposed the genetic algorithm based optimisation based SVM to overcome the shortcomings of neural network model such as large training data demand and slow convergence speed.

Recently, academics have continued to search for more effective prediction techniques. With the increasing maturity of machine learning techniques, especially the excellent performance of neural networks in the field of prediction, it has attracted the attention of many researchers. They have used different types of neural networks to predict traffic flow. These neural network models are self-learning and are able to optimise the prediction performance by continuously adjusting their internal parameters. However, this learning process relies on a large sample size of data, which not only makes the convergence of the model potentially slow, but also carries the risk that the model may fall into a local optimum rather than a global optimal solution. Therefore, although neural networks show great potential in traffic flow prediction, how to improve the learning efficiency and optimise the algorithm is still a key issue to be addressed in current research. In order to be able to adaptively extract features with end-to-end learning capability, deep learning-based methods for urban traffic flow prediction have become a hot spot in current research.

In the field of deep learning, there are deep learning architectures for urban traffic flow prediction methods based on convolutional neural networks (CNNs), such as spatio-temporal multi-graph convolutional traffic flow prediction fused with knowledge graphs to achieve knowledge embedding and improve the anti-interference ability proposed by Li et al. and recurrent neural networks, GRU neural networks, and other deep learning architectures. For example, a model for solving traffic flow prediction using ConvGRU was proposed by Wang et al.

While dealing with longer time series data, traditional recurrent neural networks (RNNs) often face the problem of gradient vanishing, which limits their ability in capturing long-term dependencies in urban traffic flow. Since the recurrent structure of RNNs may cause the gradient value to decrease rapidly to near zero during backpropagation, this makes it difficult for the network to learn long-distance dependency information in time series. Therefore, although RNNs are theoretically capable of processing sequence data, in practice they have limitations in extracting long-term patterns in traffic flow.

Long Short-Term Memory (LSTM) network is an improvement of RNN network, which controls the flow of information in the network through a multi-layer gating mechanism of forgetting gate, input gate, and output gate, and captures the long-term dependencies in the sequential data while

DOI: 10.54254/2753-8818/83/2025.19939

reacting to the short-term changes through continuous training. And the design of LSTM hidden layer makes it perform well in dealing with complex sequence data with time dependency, which is widely used in the fields of natural language processing, speech recognition, and time series prediction, such as short-term traffic flow prediction based on stacked LSTM neural network proposed by Huang et al. which improves the ability of recurrent neural network to fit nonlinear and complex functions, and Duan et al. able to enhance temporal features for tunnel traffic flow prediction based on GWO-LSTM model. However, LSTM can only extract the correlation feature information between different historical time nodes, and cannot well obtain the spatial correlation feature information of the current area.

3. Problem Definition and Description

3.1. Development of the Solution

Since the accuracy of urban traffic flow prediction is significantly affected by temporal dynamics and spatial interactions, current research often fails to fully integrate the composite features of time and space when dealing with urban traffic flow prediction, resulting in limited prediction accuracy. Therefore, by applying the multi-head attentional mechanism in the Transformer model, which provides the positional information of each element in the sequence through its positional encoding. Multiple attention heads process the information in parallel, and each head learns a different representation of the input, as a way to simulate various types of complex characteristics of traffic flow, such as Yang et al. proposed a short-term traffic flow prediction model for urban highways that combines a variant residual model and Transformer, which fuses short-term temporal and spatial characteristics of traffic flow in a specific dimension. In addition, the existing studies have failed to assign sufficient weights to specific time nodes that have a key influence in traffic flow dynamics, such as the transition point of peak hours.

3.2. Predictive Features

In order to improve the accuracy of urban traffic flow prediction, future research needs to develop prediction models that can integrate the temporal and spatial dimensions. These models should pay special attention to the key nodes by assigning higher weights to them to accurately reflect their central role in the overall traffic flow pattern. For example, the ship traffic flow prediction algorithm based on the attention mechanism and ConvLSTM proposed by Li et al. parses the extracted spatio-temporal feature information and outputs the prediction results, which is able to make effective long and short-term predictions for a certain area. Through this kind of method, it can provide more accurate data support for urban traffic planning and management, and then optimise the allocation of traffic resources to alleviate the problem of urban congestion.

4. Model Structure

4.1. Long and short-term memory networks

In the spatio-temporal prediction of urban traffic flow, the extraction of features related to different roads and terrain in the city is very critical for deep learning models, while the LSTM model is based on the traditional RNN architecture, which puts more emphasis on the extraction of the time series of data, and combines with its hidden layer to learn when to remember information and when to forget information, which makes the deep learning network capable of capturing the longterm dependencies, but it lacks the extraction of coordinate dimension and spatial features, and the spatial features have a significant impact on the prediction model.

DOI: 10.54254/2753-8818/83/2025.19939

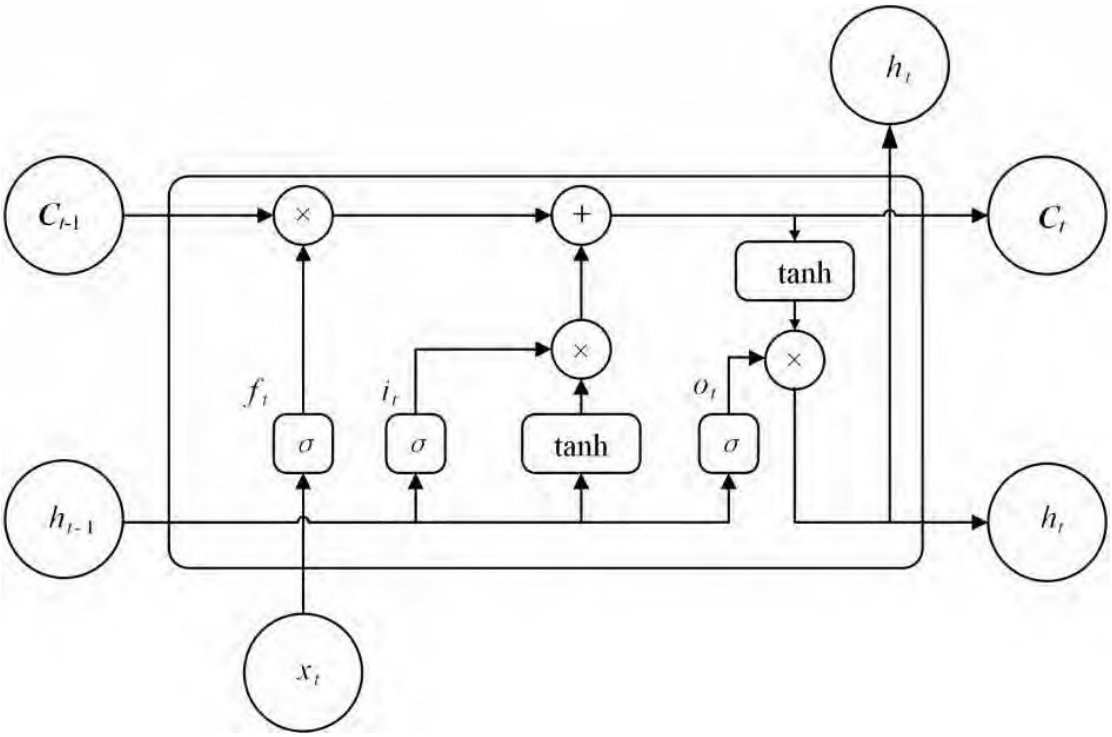

Figure 1: Schematic diagram of LSTM model structure

Therefore, after the LSTM algorithm is performed, the attention mechanism is also used to assist the prediction, and certain weights are assigned to integrate the results of the two, so as to achieve a comprehensive prediction. The computational procedure of applying LSTM on this model is as follows:

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

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i \quad (2)$$

$$\widetilde{C_t} = \tanh\,(W_c \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o \quad (5)$$

$$h_t = o_t * \tanh\,(C_t) \quad (6)$$

where σ is the sigmoid function, f_t is the activation value of the oblivion gate at time step t, h_{t-1} is the hidden state of the previous time step, i.e., the predicted value of the traffic flow of the previous cycle, x_t is the input of the traffic flow of the current time step, i_t is the activation value of the input gate, o_t is the activation value of the output gate , $\widetilde{C_t}$ is the candidate cell state, C_t is the cell state after the update, W_f is the weight of the forgetting gate, W_i is the weight of the input gate, W_c is the weight of the candidate cell state, and W_o is the weight of the output gate.

4.2. Transformer-based Attention Mechanisms

The Attention Mechanism in the Transformer model is a powerful tool for enhancing a model's ability to make sense of sequential data. Attention Mechanism is a resource allocation strategy that allows a model to focus on the most important parts of the information it processes. In time series prediction, Attention Mechanism can be used to make predictions independently of traditional recurrent neural

DOI: 10.54254/2753-8818/83/2025.19939

nets. When applying the data in this paper for training, the Attention Mechanism with multiple heads is applied, which allows the model to capture information in different representation subspaces in parallel when reading the data line by line, and then synthesise the prediction results of the LSTM model to achieve a more accurate prediction, the specific structure is shown in the figure below:

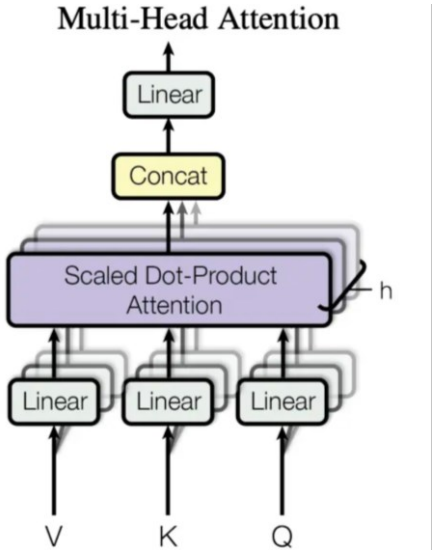

Figure 2: Model structure of the multi-attention mechanism

Where the attention mechanism of the four heads is computed as follows:

For the input sequence, it is first mapped to the representation space of queries (Q), keys (K) and values (V) respectively:

$$Q = W^Q \cdot x + b^Q \tag{7}$$

$$K = W^K \cdot x + b^K \tag{8}$$

$$V = W^V \cdot x + b^V \tag{9}$$

where W^Q, W^K, W^V are the learnable weight matrices, b^Q, b^K, b^V are the corresponding bias terms, and x is the input sequence.

Afterwards, for each head h, the dot product of query and key is computed, scaled dot product attention is performed, and softmax normalisation is performed to obtain the attention weights:

$$softmax\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) \tag{10}$$

Whered_k is the dimension of the key vector used to scale the dot product to prevent the gradient from vanishing or exploding.

In the multi-head attention mechanism, parallel computation and splicing of multiple heads is required to finally perform linear transformation and output:

$$\text{Multihead} = attention^1 + attention^2 + attention^3 + attention^4 \tag{11}$$

$$\text{Output} = \text{W}^\text{O} \cdot \text{Multihead} + \text{b}^\text{O} \tag{12}$$

4.3. Urban Traffic Flow Data Prediction Model Based on Multiple Attention Mechanism and LSTM

The model was fed with previously collected and preprocessed Beijing taxi traffic data into the coding network, which was first normalised by the data, created a data loader, and converted the dataset into

DOI: 10.54254/2753-8818/83/2025.19939

a PyTorch tensor, which was imported into the LSTM with Attention Mechanisms model, which firstly analysed the raw data in-depth by a multilayered long and short-term memory network in order to extract the key spatiotemporal features . Subsequently, these features are weighted and predicted with multiple heads of attention using the attention mechanism to ensure that the model can focus on those time points and spatial regions that have the greatest impact on the prediction results during the training process. Finally, the decoding network integrates the spatio-temporal features extracted by the encoding network to generate accurate predictions. After completing the prediction of both, the hidden layer is adjusted with a fully connected layer, the output layer fuses the two, and the prediction results are transformed into a 2D array and then back-normalised to obtain a traffic prediction that matches the real situation, which can be visualised and set up for model evaluation.

5. Experimental Results and Model Performance Evaluation

5.1. Experimental data collection and pre-processing

This article adopts the relevant data of Beijing taxi coordinates and time locations to experiment the effectiveness of the model, the data set has the longitude, latitude and time points of all the taxi vehicles appearing in Beijing within a week, through the relevant information of this data set can be processed to obtain the gridded flow data, so as to initially extract the spatial characteristics of geographic location, as shown in Fig. 1; it is also possible to combine the flow data of all days and then filter the data as a whole. The data can also be merged by combining all the days' flow data, and then filtered in general, taking one hour as the unit time interval to get the unit time gridded flow data, and further processed into a table type of data, dividing out each gridded flow change over time data table. This data can be imported into the LSTM model for model training (Figure 2).

In order to improve the data quality and remove the noise data, this paper refers to the gridded frame processing in the transbigdata function in python, which removes the irrational distribution data; the data that are at the edge of the grids or do not have spatial region characteristics, and a large number of data that are 0; the vehicle data that are stationary and classified as traffic data, and redundant data, so as to make the flow flow data between the grids sufficient and effective.

In the process of processing the main use of which GPS_to_grid, area_to_grid, area_to_params function, the collection of scattered GPS latitude and longitude data to generate a gridded rectangular grid with parameters.

The grid_params_optimize function is then used to optimise the grid parameters to give a large improvement in accuracy and to obtain the most reasonable unit size grid at the data distribution level.

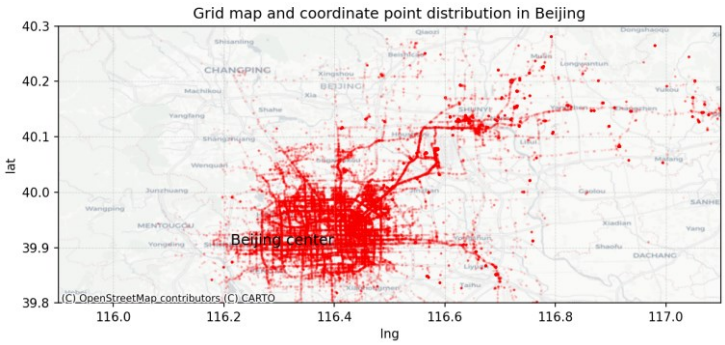

Figure 3: Spatial distribution of traffic datasets

DOI: 10.54254/2753-8818/83/2025.19939

grid_id	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(	2008-02-(
116.22_39	1125	2113	2534	2760	2676	2587	2679	2770	2331	2099	1832	1448	1191	659	603	606	616	590
116.22_40	61	101	78	89	74	80	91	81	95	155	156	126	121	113	122	140	136	148
116.22_40	203	414	482	466	524	500	571	469	495	557	532	544	463	271	246	212	211	188
116.22_40	20	46	35	69	43	47	49	55	88	150	119	104	93	92	103	116	104	103
116.38_39	27	44	44	33	72	89	80	72	110	114	132	128	174	154	160	134	159	133
116.38_39	27	19	27	28	21	19	21	21	42	50	57	28	36	39	50	45	43	44
116.38_39	573	1320	1272	1335	1301	1265	1171	1022	1038	1090	1103	1082	1141	550	720	853	918	875
116.38_39	5115	10213	10285	10463	9765	10496	11017	10577	10102	9506	8370	7458	6853	3450	3797	3641	3652	3606
116.38_39	5523	11064	10773	12011	11516	11491	13291	13241	11305	10096	8513	7140	6781	3624	4524	4295	4213	3647
116.38_40	62	97	95	119	111	91	152	173	190	182	197	186	212	93	97	170	161	141
116.38_40	381	843	641	670	784	959	1095	1041	985	1060	954	976	858	388	393	428	414	418
116.38_40	38	38	51	51	63	48	63	84	69	90	74	103	89	47	48	50	61	46
116.3_39.6	130	258	262	263	235	215	294	337	472	575	563	596	669	575	595	590	637	627
116.3_39.6	27	36	34	9	10	24	15	15	37	59	56	72	68	60	88	68	67	74
116.3_39.7	356	741	679	823	851	857	902	740	742	823	856	868	737	520	557	591	599	555
116.3_39.8	4952	10114	8829	10411	10253	10516	11171	9946	9133	8924	8024	7292	6268	3430	3241	2713	2746	2793
116.3_39.9	5504	12059	11417	12200	11832	11724	11264	11088	9399	8474	7361	5858	5053	2846	3126	2530	2305	2128
116.3_40.0	42	110	62	95	67	110	151	117	138	133	151	125	117	92	95	94	85	91
116.3_40.0	653	1464	1459	1332	1381	1595	1893	1880	1710	1860	1816	1761	1604	1066	1030	915	868	854
116.3_40.1	3	7	7	13	13	18	49	52	49	40	40	37	31	44	48	47	48	47
116.46_39	32	54	61	75	83	62	69	79	69	114	158	167	199	123	219	220	231	233
116.46_39	21	20	21	30	30	40	27	14	44	44	30	46	37	29	37	53	66	47
116.46_39	229	410	481	559	548	602	735	675	475	477	352	307	294	132	156	161	174	170
116.46_39	1340	3045	3383	3280	3404	3553	3725	3622	3385	3238	3202	2665	2541	1192	1278	1475	1402	1310
116.46_39	1825	3620	3577	4192	3865	3975	4695	4827	4177	3809	3642	3125	3087	1602	1918	2159	2188	2181
116.46_40	40	84	60	50	63	66	94	71	145	133	92	81	96	35	40	56	58	56
116.46_40	539	990	1028	1010	971	1260	1381	1036	1098	1019	995	952	1118	696	892	909	1055	1085
116.46_40	12	21	12	22	7	30	42	26	20	35	31	43	41	13	32	44	40	43
116.54_39	34	33	46	72	77	107	84	79	68	96	99	131	123	103	122	130	172	137
116.54_39	11	17	6	22	26	27	32	55	25	48	56	51	59	38	51	48	36	52
116.54_39	15	37	67	73	98	149	132	73	71	122	75	59	86	41	73	115	136	106
116.54_39	246	478	434	463	416	520	633	587	674	796	825	733	755	364	371	478	451	434
116.54_39	135	222	266	266	221	254	277	268	251	239	317	298	334	168	247	254	287	243

Figure 4: Data table after removing the noisy data and organising it

5.2. Forecasting area delineation

According to the latitude and longitude regions where the data are located, a rectangular boundary range is framed, and the highest and lowest points of latitude and longitude in the prediction region are where lng_{max} and lat_{max} denote the longitude and latitude of the highest point, respectively; lng_{min} and lat_{min} denote the longitude and latitude of the lowest point, respectively. Within the framed boundary, the prediction region is divided into m×n grids in this paper, and each grid represents a small prediction region, which is arranged according to the geographical location.

When dividing the region, considering that the data failure rate should be controlled within a certain range and the quality of the selected data should be guaranteed, the size of individual grids is adjusted, and the latitude and longitude of each grid line spacing is adjusted to 0.08 in the pre-processing stage, and the grids with less data volume, insignificant spatial features, and most of the time points of 0 are merged. Finally, 112 grids are set to be divided in the prediction area, and a total of 10097262 sample data of taxis in Beijing from 2 February 2008 to 8 February 2008 are selected. The ratio of the training data to the test data is set at 8:2. The sample data from 13:00 on 2 February 2008 to 3:00 on 7 February 2008 are selected as the training set to train the model, and the sample data from 4:00 on 7 February 2008 to 13:00 on 8 February 2008 are selected as the test data to evaluate the performance of the model.

5.3. Experimental environment setting

In this paper, Python language and Pytorch-based deep learning framework are used to achieve the prediction of urban traffic flow. The epoch of the training times is set to 100, and the loss function adopts the Mean Square Error Loss (MSE), which is more sensitive to outliers and can better capture the stability of the model prediction, calculated as MSE = (1/n) Σ(yi - ŷi)^2, and the optimiser adopts the Adam optimiser in thetorch. Meanwhile, in the training environment, in order to avoid model overfitting, the Dropout function is used, which increases the sparsity of the network by randomly discarding some neurons in the network during the training process, resulting in a different topology of the network in each training iteration, and the coefficient of Dropout is set to 0.1.

DOI: 10.54254/2753-8818/83/2025.19939

Table 1: Parameters of the experimental environment

Parameter name	Parameter setting
Number of training sessions	*100*
optimiser	*Adam*
loss function	*MSE*
Dropout rate	*0.1*
time step	*12*
learning rate	*0.001*
sample size	*112*

5.4. Model structure parameterisation

After establishing the prediction model according to the preliminary ideas, this paper optimises the model configuration by constantly adjusting the structural parameters. In the network setup, the coding format of four-layer LSTM combined with four-head attentional mechanism is used, and the size of the hidden layers are all set to 128, and the output of the last time step of the training set is taken as the input data of the test set. After analysing the longitudinal temporal characteristics of the dataset and continuously debugging the code for prediction, it was decided to set the step size to 12, i.e., the data from the last 12 time points was used to predict the traffic at the next time point. The detailed settings are shown in the table below:

Table 2: Predictive model structural parameters

Unit level	**Parameter setting**
Four-layer LSTM cell	**footsteps:12** **hidden layer:128**
Module on the four-headed attention mechanism	**footsteps:12** **hidden layer:128**

5.5. Predictive model evaluation indicators

After completing the construction and training of the model with LSTM and attention mechanism set up, it is necessary to judge the prediction performance of the model with relevant evaluation indexes. According to the characteristics of urban traffic flow, it was decided to choose the mean absolute error (MAE), root mean square error (RMSE), and normalised root mean square error (NRMSE) to evaluate the prediction results of the model. The specific formulas are as follows:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \tag{13}$$

n is the sample size, i.e. the total number of observations in the dataset. y_i is the ith observation representing the actual measured value. y_i' represents the value predicted by the ith model.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \tag{14}$$

The RMSE is the square root of the MAE, which has the same units as the observations.

$$\frac{RMSE}{\max(y)-\min(y)} \tag{15}$$

Where max(y) and min(y) are the maximum and minimum values of the observations, respectively.

DOI: 10.54254/2753-8818/83/2025.19939

5.6. Evaluation of experimental results

After reading the data of all the grids and going through the training cycle, this paper selects one of the grids and visualises its training process and prediction process, showing its pre-training image as well as the comparison image of the prediction image with the actual value.

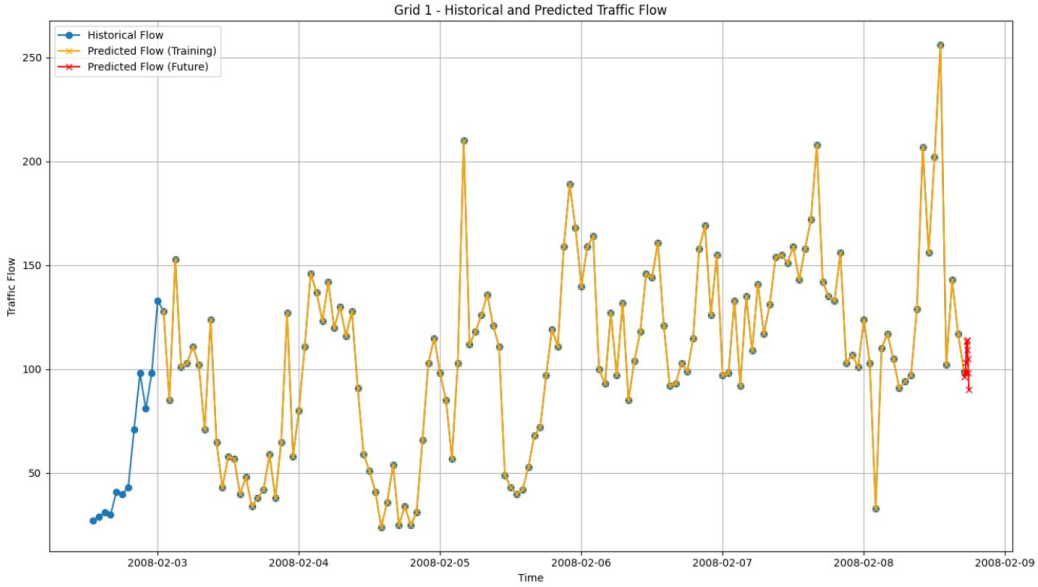

Figure 5: Pre-training test

During the training process, as shown in Fig. 6, the overall trend of the loss function MSE is decreasing after multiple training cycles, reflecting the steady improvement of the training effect, and finally achieving a good fit.

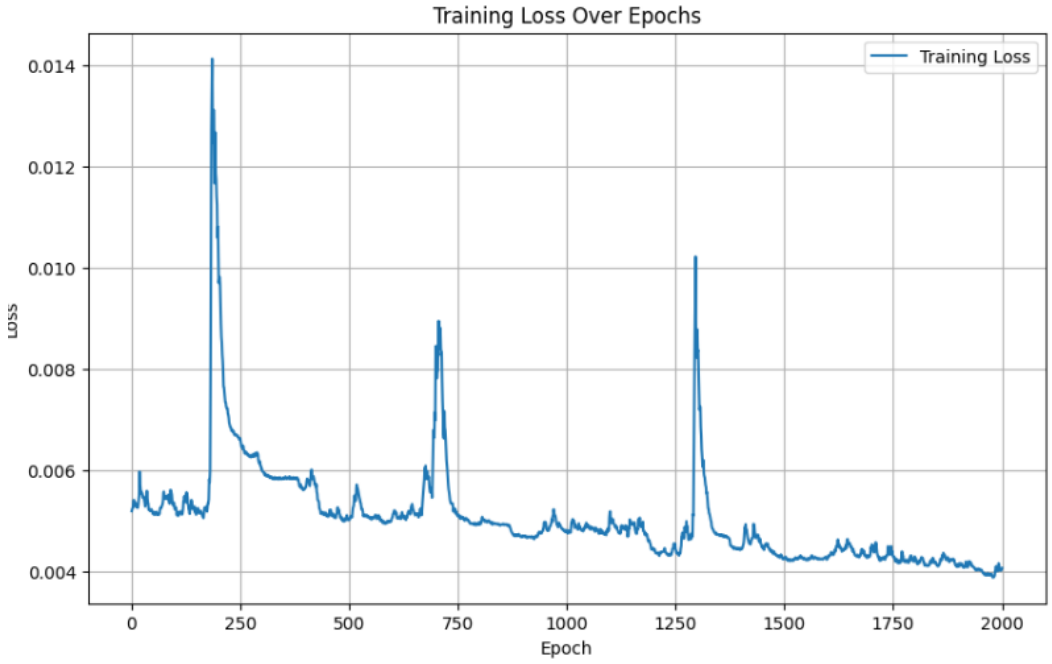

Figure 6: Loss function variation with training period for LSTM and comb attention mechanism based prediction model

DOI: 10.54254/2753-8818/83/2025.19939

In the field of predictive modelling of urban flows, there is a proliferation of methods aimed at improving prediction accuracy and generalisation. From classical statistical models to modern machine learning techniques, each model has its unique advantages and limitations. The following is a comparative analysis of the prediction model in this paper with other different models.

In comparing the model performance, this paper has chosen three prediction models for comparison: gated recurrent unit (GRU), LSTM prediction model alone, and prediction model based on the attention mechanism of transformer alone, and the data comparison of the test indexes obtained is shown in the following table:

Table 3: Performance comparison of different prediction models

predictive model		RMSE	MSE	MAE
GRU	**train**	**29.1459**	**849.4849**	**23.3091**
	test	**30.4729**	**928.5961**	**24.054**
Transformer	**train**	**34.9142**	**1219.0039**	**27.6692**
	test	**28.3584**	**804.2008**	**21.487**
LSTM	**train**	**29.8907**	**893.4564**	**24.2402**
	test	**29.664**	**879.9549**	**21.8926**
LSTM + Attention	**train**	**29.0867**	**846.0361**	**22.2759**
	test	**27.2616**	**743.1937**	**22.5173**

From table 3, it can be seen that both the training cycle and the testing cycle, the model of LSTM combined with the attention mechanism used in this paper is more accurate, with less error, and the performance is more superior compared to the traditional prediction model, and the prediction ability is significantly improved.

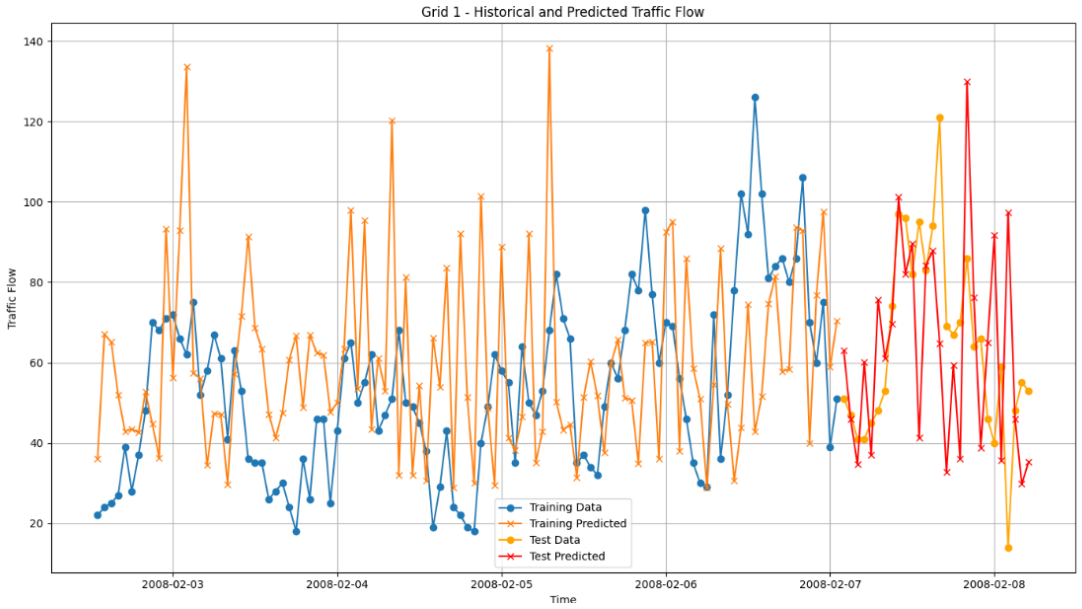

Figure 7: Prediction results based on LSTM model

DOI: 10.54254/2753-8818/83/2025.19939

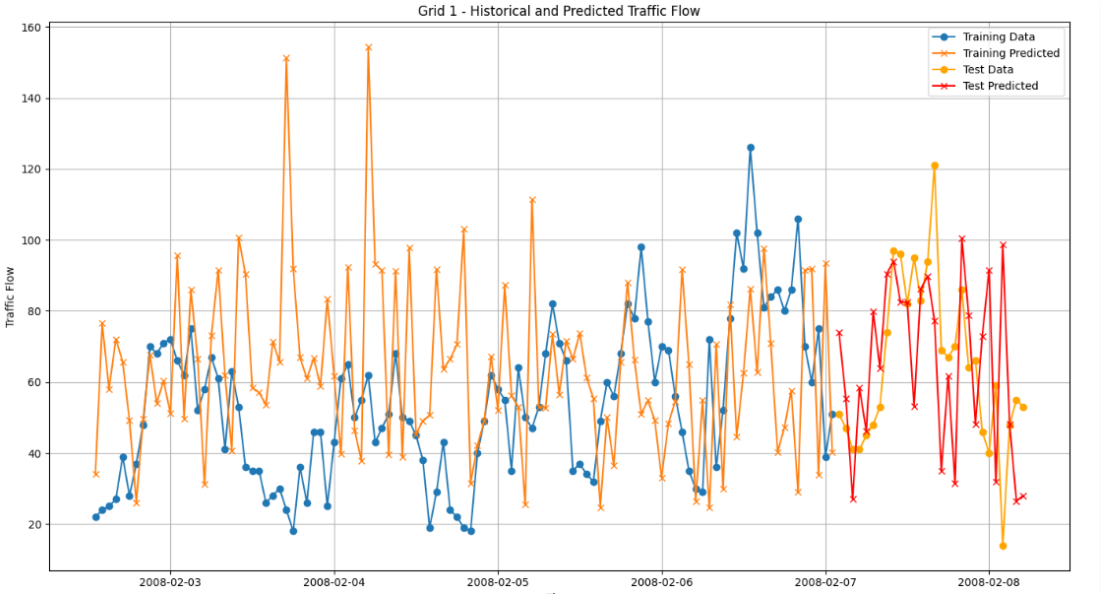

Figure 8: Model prediction results based on transformer attention mechanism

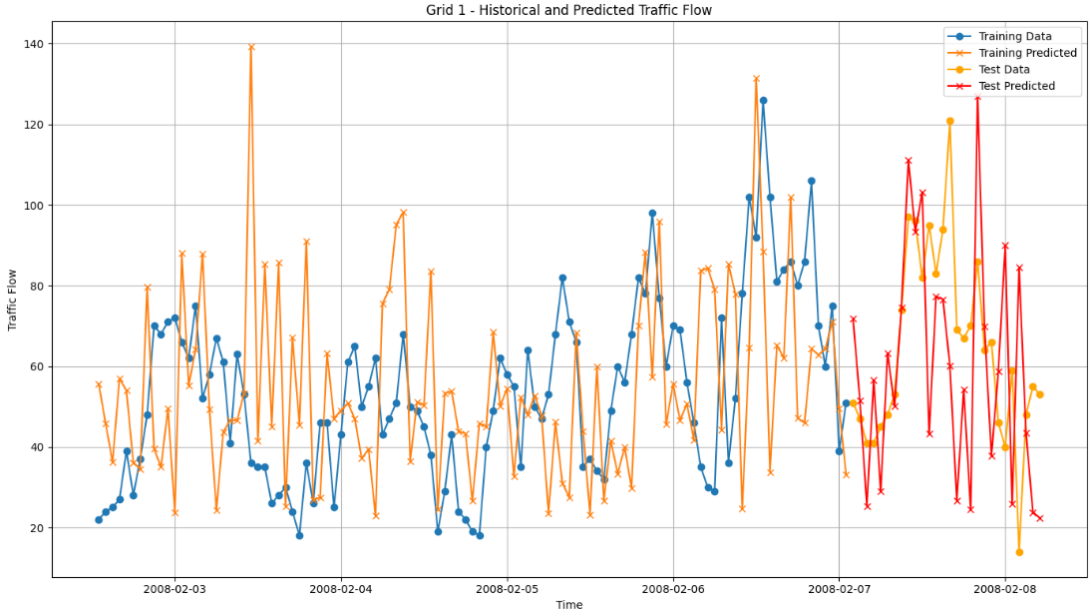

Figure 9: Prediction results based on GRU network

DOI: 10.54254/2753-8818/83/2025.19939

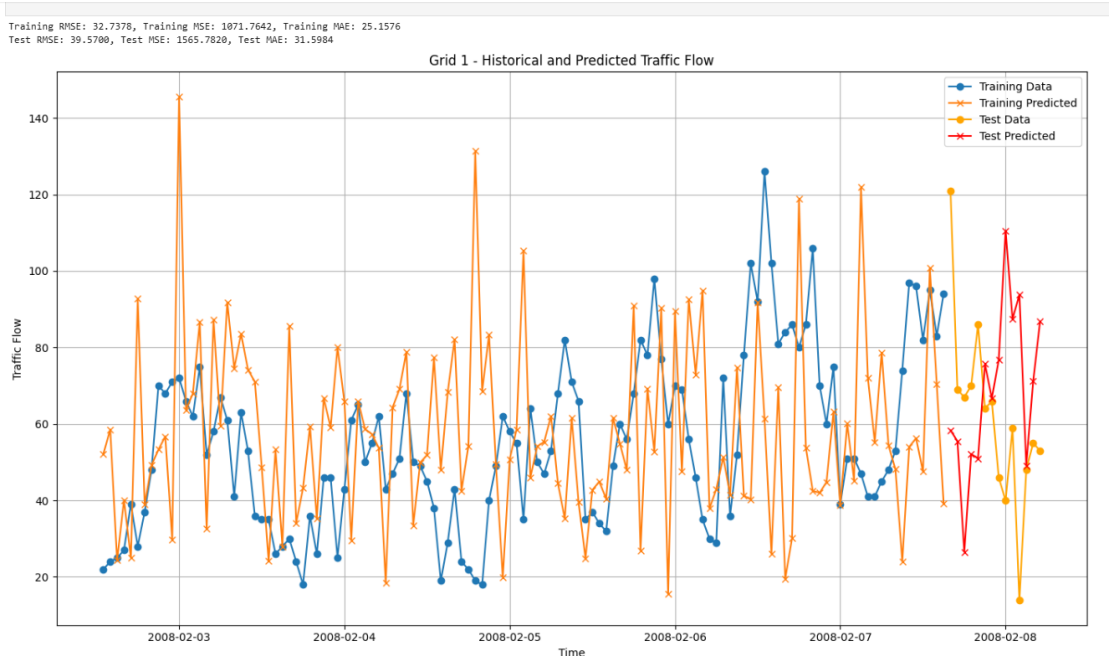

Figure 10: Prediction results based on LSTM with attention mechanism mode

By observing the comparison images between the final predicted traffic flow and the real traffic flow under the same training cycle of several prediction models, it can be seen that the LSTM combined with the attention mechanism algorithm selected in this paper has a significant improvement in the prediction performance compared with some traditional prediction models; in the evaluation criteria data display in Table x, it can be observed that the performance evaluation index comparison of these models for traffic flow prediction In Table x, it can be observed that the performance of these models for traffic flow prediction has obvious advantages, which can better fit the temporal, spatial and other characteristics of urban traffic flow, and has a good learning function.

6. Conclusion And Future Work

The rapid advancement of information technology has significantly enhanced the role of machine learning and artificial intelligence in urban traffic flow forecasting, optimizing the dynamic management and balance of supply and demand in urban traffic, thereby ensuring the stable operation of urban transportation systems. This study proposes a traffic flow forecasting method based on the Transformer attention mechanism and multi-layer LSTM models, integrating complex features of temporal sequences and spatial positions to provide a new solution for accurate prediction of urban traffic flow.

In comparative experiments, the model outperformed traditional forecasting models across several evaluation metrics, demonstrating a significant reduction in Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Normalized Root Mean Square Error (NRMSE), reflecting substantial improvements in training efficiency and predictive accuracy. However, the model still exhibits certain biases in forecasting during peak hours, especially when traffic flow surges, indicating that there is room for improvement in handling extreme situations.

Future research directions should focus on enhancing the model's predictive accuracy and robustness by integrating multi-source data, such as weather conditions, special events, and holidays. This approach is expected not only to further improve the prediction of traffic flow but also to provide more comprehensive and detailed decision support for urban traffic management. Moreover, the potential applications of this model extend beyond urban road traffic to other modes of transportation,

DOI: 10.54254/2753-8818/83/2025.19939

such as public transit and non-motorized traffic, which is of significant importance for reducing urban traffic congestion and promoting green travel. Overall, this study provides strong technical support for the construction of smart urban transportation systems and contributes new ideas to the sustainable development of the sharing economy and environmental protection sectors.

References

[1] *Shen Leixiao, Lu Yuhang, Guo Jianhua, Kalman filter short-time traffic flow prediction on common national and provincial highways Adaptability study 10. 3963/j. jssn. 1674-4861. 2021. 05. 015*

[2] *Wang J. Deng W. Zhao J. B. Short-time traffic flow prediction based on improved Bayesian combinatorial model 1001 - 0505(2012) 01-0162-06*

[3] *P. Xu P. Jiang F. R. Short-duration traffic flow prediction based on ant colony optimisation support vector machine 10.3969 /j.issn.1000-386x.2013.03.066*

[4] *Li Xiaobin Traffic flow prediction based on genetic algorithm optimised support vector machine 000-7180 (2010) 10-0186-03*

[5] *Liu Qidong Research on traffic flow prediction based on spatio-temporal multi-scale convolutional network 2021*

[6] *WANG Yusen, JING Zhiyong, WEI Lin, GAO Yufei, SHI Lei, WANG Qingxian, TAO Yongcai, WANG Xiangjie. A method to solve the traffic flow prediction problem using ConvGRU[J/OL]. Small Microcomputer Systems. https://link.cnki.net/urlid/21.1106.TP.20230915.1043.004*

[7] *Kai Li, Shaowei Huang Short-term traffic flow prediction based on stacked LSTM neural network 1009-8445(2022)05-0079-06*

[8] *Duan ZX, Du WX. Research on Tunnel Traffic Flow Prediction and Lighting Regulation Based on GWO-LSTM Model[J]. Modern Tunnelling Technology, 2024, 61(3): 157-165. DUAN Zhongxin, DU Wanxin. Study on Tunnel Traffic Flow Prediction and Lighting Regulation Based on GWO-LSTM Model[J]. Modern Tunnelling Technology, 2024, 61(3): 157-165*

[9] *G. Li, W. Song, Z. Chen Ship traffic flow prediction algorithm based on attention mechanism and ConvLSTM 10. 118967/jsjksx.230800067 2023*