Short Term Prediction of Urban Traffic Flow based on Machine Learning Algorithms

Conghe Wang *

School of Automotive Studies, Tongji University, Shanghai, 201804, China

* Corresponding Author Email: 2053858@tongji.edu.cn

Abstract. With the increasing per capita car ownership worldwide, existing transportation facilities are unable to fully meet people's travel needs. The massive traffic flow can lead to traffic congestion, so predicting traffic flow can play a crucial role in urban traffic management. This study predicts the westbound traffic flow data of the Minnesota Department of Transportation (MN DoT) ATR 301 station on Interstate 94 in the United States for 28 consecutive days using a Long Short-Term Memory Network (LSTM) data prediction model. The Random Forest is utilized to analyze the association between attributes and the historical data has been separated into 1-hour intervals. The highway displays a bimodal pattern on weekdays and a unimodal pattern on holidays, according to the final forecast results. Furthermore, after making the necessary adjustments to the parameters, the simulated trend closely resembles the real value, proving that LSTM can be used to model short-term traffic flow. Accurately predicting urban traffic flow can make the management of urban traffic management systems more comprehensive and provide citizens with higher quality travel planning references.

Keywords: Traffic flow; LSTM; random forest; prediction.

1. Introduction

With the increase in the number of urban transportation vehicles in the world, urban transportation management urgently needs more refined and intelligent management methods and data processing methods to ensure the safety, efficiency, and order of urban traffic. As an important component of intelligent transportation systems (ITS), accurate and real-time traffic flow prediction is crucial for reducing traffic pressure [1]. It plays a crucial role in applications such as path planning and urban transportation management [2].

The prediction methods for short-term traffic flow can generally be divided into four categories: prediction methods based on statistical analysis theory, prediction methods based on micro simulation, prediction methods based on mixed models, and prediction methods based on artificial intelligence technology [3]. Among them, models based on statistical methods mainly include historical average models (HA), time series models (Time series), and autoregressive integrated moving average models (ARIMA), while ARIMA models and their improved seasonal autoregressive integrated moving average models (SARIMA). Most commonly used. The methods based on micro traffic simulation mainly include dynamic allocation theory [4].

In models based on intelligence theory, algorithms such as Support Vector Machine (SVM) and Random Forest are widely used. One useful technique for building an integrated classification model for traffic flow prediction is Random Forest, which integrates numerous decision trees for data prediction [5]. Due to the fact that most traffic data contain weather information, temperature information, etc., such high-dimensional traffic data can be handled by random forest, which can also manage missing numbers. Therefore, using random forest for feature analysis and data processing of traffic flow data can effectively improve the accuracy of predictions. With the improvement of big data technology and computing power, as well as the rapid development of artificial intelligence technology, deep learning technology is also widely used in traffic flow prediction [6].

In contrast to alternative prediction techniques, deep learning-based intelligent prediction techniques are capable of rapidly acquiring from large-scale traffic flow data complicated and time-varying non-line flow features in road networks. The accuracy of short-term traffic flow predictions can be significantly increased by deep learning technology, which also offers technical assistance for the creation of intelligent transportation systems and ongoing advancements in smart cities and transportation systems [7]. The use of deep learning technology for short-term prediction of urban traffic flow has become an inevitable trend. Among numerous deep neural networks, Multilayer Perceptron (MLP) was the first to be applied to short-term traffic flow prediction. As the number of traffic flow prediction scenarios increases, the complexity of the model continues to evolve. In the scenario of single point traffic flow prediction, Recurrent Neural Networks (RNN) and Temporal Convolutional Networks (TCN) are applied to predict time series.

The Long Short Term Memory Network (LSTM) model is a recursive neural network developed and optimized based on the RNN architecture. The introduction of gating mechanism in long short-term memory networks improves the long-range dependency problem of recurrent neural networks and effectively solves the problem of gradient explosion or disappearance in simple recurrent neural networks [8]. The Long Short-Term Memory (LSTM) model is a popular choice for a variety of short-term prediction scenarios since it outperforms other classic neural network models in handling long-distance sequence difficulties [9]. Meanwhile, due to its unique cyclic structure, LSTM can selectively learn and discard information, and its performance in short-term traffic prediction far exceeds traditional methods.

Many researchers have applied long short-term memory networks to traffic flow prediction, effectively characterizing the nonlinearity of traffic data and improving prediction accuracy by capturing long-term time related features in traffic flow data [10]. Doğan proposed a robust long short-term memory (R-LSTM) traffic flow prediction method based on robust empirical mode decomposition (REDM) algorithm and LSTM to address the randomness of traffic flow, which improves the stability of traffic flow prediction [11]. Zheng et al. used LSTM to predict traffic flow and analyzed the impact of various factors on the predictive performance of LSTM networks [12]. Shi et al. proposed a prediction method based on a combination of multiple linear regression and LSTM, which achieved the same prediction effect as non-missing data in the case of partial historical traffic loss in the target road section [13].

The deep learning method based on LSTM has been proven to be effective and accurate in short-term prediction of urban traffic flow. Therefore, this article considers applying this method to short-term prediction of urban traffic flow. This article first uses the random forest algorithm to perform feature analysis and selection on the westbound traffic volume data of MN DoT ATR 301 station on Interstate 94 in the United States from October 2012 to September 2018, then uses deep learning techniques such as long short-term memory neural networks for simulation research to demonstrate the effectiveness and advantages of this method compared to traditional methods. The goal of this article is to find an accurate algorithm prediction model, so that the analysis and prediction results can provide reference for travel, assist in planning travel routes, avoid risks such as traffic congestion, and take measures to alleviate impending road congestion, reducing environmental pollution and property losses caused by congestion [14].

2. Methods

2.1. Data Source

The UC Irvine Machine Learning Repository provides the data used for this study, which spans the years October 2, 2012, through September 30, 2018. This dataset includes the hourly traffic volume of Interstate 94 westbound at MN DoT ATR 301 station, totaling approximately 50000 pieces of data.

The highway is roughly located between Minneapolis and St. Paul, Minnesota. This dataset also provides hourly weather characteristics (such as temperature, rainfall, snowfall, and more detailed descriptions of weather), as well as holiday data. This research chose 28 days of data, spanning from April 29 to May 26, 2018, to conduct short-term trials on traffic flow prediction. Table 1 displays the data field types of the initial dataset.

Table 1. Variable description

Column names	Type	Explain	Role
traffic_volume	Int	Hourly I-94 ATR 301 reported westbound traffic volume	Target
date_time	Date	Hour of the data collected in local CST time	Feature
temp	Float	Average temp in kelvin	Feature
holiday	String	US National holidays plus regional holiday	Feature
rain_1h	Float	Rainfall measured in millimeters within one hour	Feature
snow_1h	Float	Snowfall measured in millimeters within one hour	Feature
clouds_all	Float	Cloud coverage percentage	Feature
weather_main	String	A concise description of the weather	Feature
weather_description	String	A detailed description of the weather	Feature

Traffic flow data is filtered according to holidays and extreme weather (such as rainstorm, thunderstorm, heavy snow, etc.), and some obviously wrong data are also removed. The data samples are shown in Table 2.

Table 2. The preprocessed data

holiday	date_time	temp	traffic_volume
None	2018/5/15 16:00:00	296.36	6888
None	2018/5/15 17:00:00	297.22	6548
None	2018/5/15 18:00:00	297.58	4953

In the Python data preprocessing stage, checks are made to find any empty cells or missing values. Based on the preliminary drawing results, it is found that although the dataset has been preprocessed and there are no blank or missing cells, there are still a small number of repeated and partially missing time periods, such as repeated or missing data at a certain moment. This type of situation can cause errors in the time series during the prediction process and lead to inaccurate prediction results. Therefore, duplicate data should be deleted first, and missing values should be avoided in selecting the required data. The initial data requires for the final algorithm is obtained. All features are first extracted for random forest voting during the first round of data processing, and the most pertinent characteristics are chosen for further examination. Data on traffic flow is gathered around-the-clock, and Figure 1 displays an initial graph of the data.

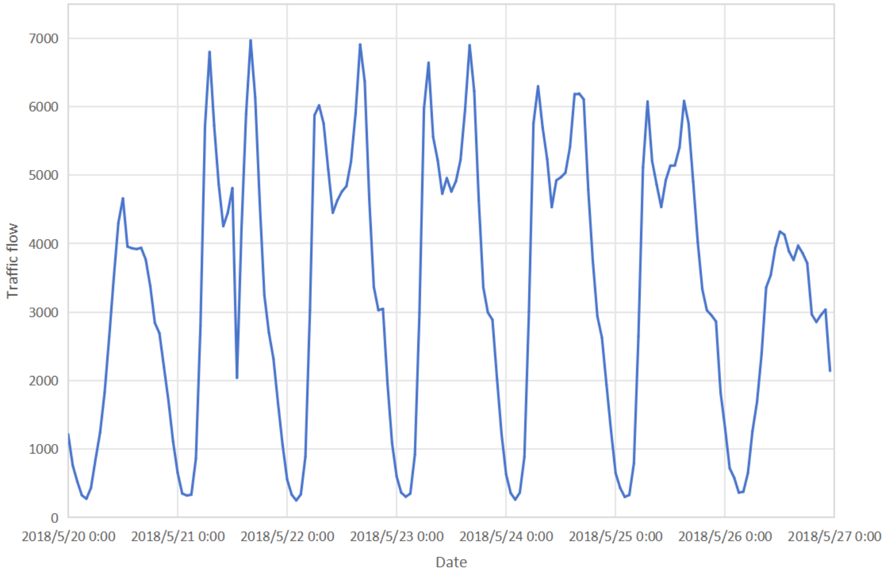

Fig. 1 Preliminary traffic flow

2.2. Random Forest

Multiple decision trees that are independent of one another make up a random forest. It is possible to identify the most representative features by training several subsets of features, voting collectively, or averaging their predictions. Figure 2 illustrates this procedure.

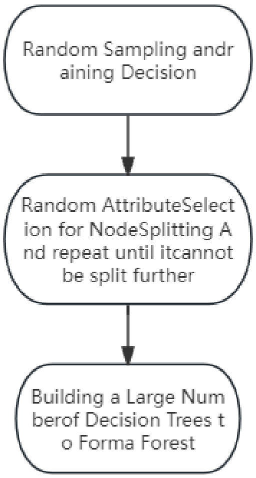

Fig. 2 Random Forest Flowchart

A random forest model can be constructed based on training data and its parameters can be adjusted according to validation data. The selection and processing methods of data can be adjusted based on the weight of each element's impact on the data in the random forest processing results.

2.3. LSTM Model

With the goal of improving the modeling capacity for long-term dependencies and resolving the issues of disappearing and exploding gradients in conventional RNN, LSTM is an upgraded recurrent neural network (RNN) architecture.

LSTM presents a memory cell with information storage and retrieval capabilities, as well as the ability to regulate information flow via gating mechanisms. Figure 3 illustrates the input, forget, and output gates—the three essential parts of an LSTM.

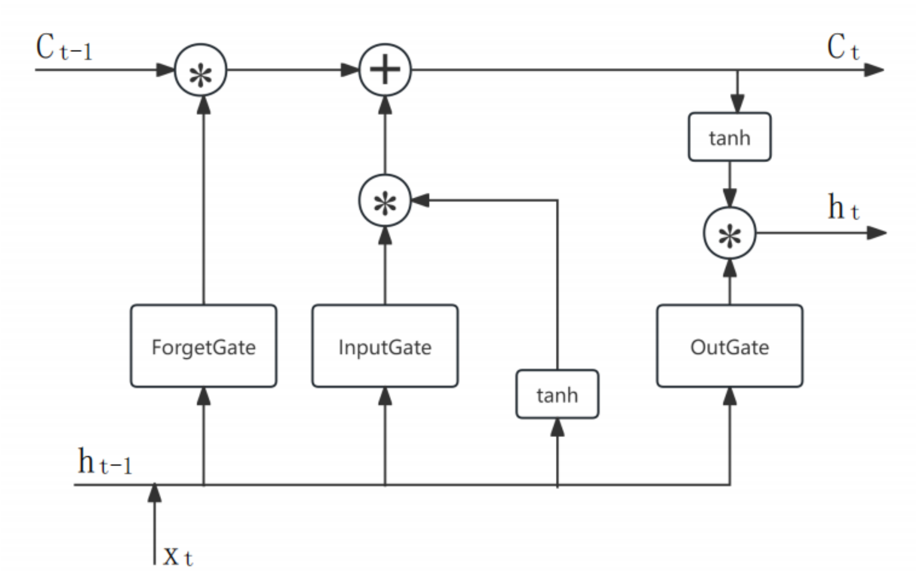

Fig. 3 LSTM Flowchart

Input gate, used to control the input coefficient when updating the memory state; Output gate, indicating how memory state values affect output; Forgetting gate, indicating how many past features are retained. The formula is as follows:

The formula for the forget gate:

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

The formula for the input gate:

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

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3}$$

The formula for the output gate:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{4}$$

$$h_t = o_t \tanh c_t \tag{5}$$

Where f_t is the forget gate unit, it is the input gate unit, it is the input gate unit, $O_t$is the output gate unit, and $h_{t-1}$is the hidden layer state. $W_f, W_i, W_c, W_o$are weight matrices, and b_f, b_i, b_c, b_o,are bias vectors. The sigmoid function and tanh function are used as process functions in the equation. These equations provide a detailed description of the various computational processes within the LSTM unit and enable it to capture and process long-term dependencies in sequential data, enhancing the data processing capability of the LSTM unit [15].

3. Results and Discussion

The analysis of the variation of traffic flow with dates in the dataset shows that in non-holiday situations, the average daily traffic flow on weekdays (i.e. Monday to Friday) is much higher than on Saturdays and Sundays, making it more prone to traffic congestion. The visualization effect is shown in Figure 4.

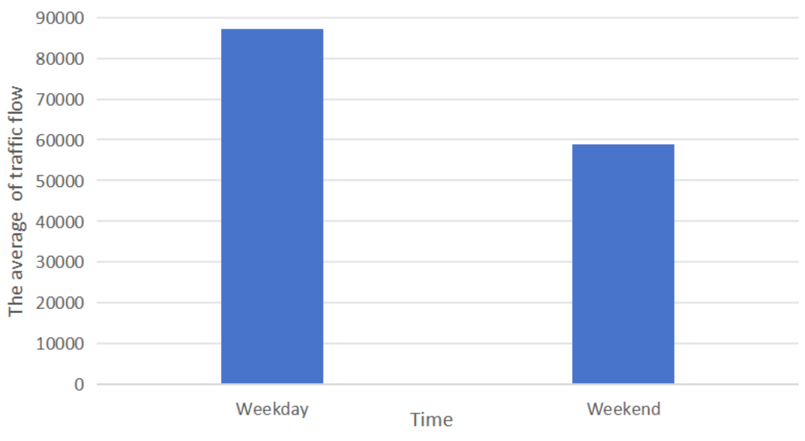

Fig. 4 The average daily traffic flow

3.1. Random Forest Results

A random forest is employed for feature correlation assessment because the source data contains eight distinct features and it is desired to study only the data that is most connected with the time feature. Like other characteristics, certain data must be transformed to integers in order to make the source data comparable because it is composed of string and floating-point types.

For instance, the strings "cloud," "miss," and "clear" are used in the source data to describe the weather; in this study, those same strings are assigned comparable values of 1, 2, and 3 for comparison. Features including temperature, precipitation, snowfall, and weather details are compared using a random forest model that is time-indexed. The comparison results show a strong association with time and suggest that Kelvin temperature is rather significant. Time and Kelvin temperature are therefore chosen as prediction targets based on correlation analysis. Figure 5 displays the results of the comparison.

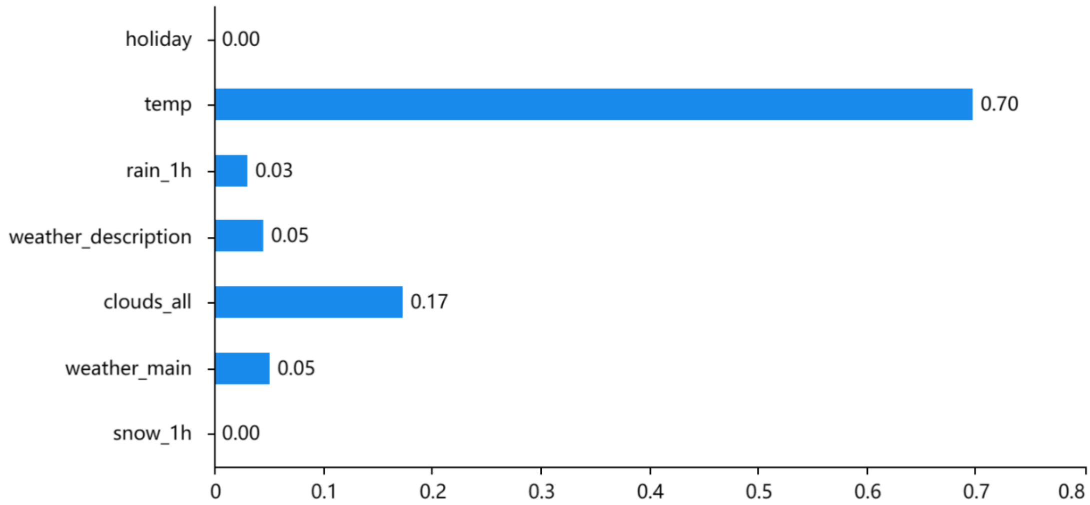

Fig. 5 Comparison chart of correlation with target features

3.2. Feature Selection

In the data preprocessing process, after deleting data with repeated time periods, it is noticed that there are incomplete data records or missing data at certain times every month from 2012 to 2016. The coherence of the time series is low, which would have a significant impact on the accuracy of the prediction results. Therefore, this part of the data is not used. Removing unnecessary data not only leads to a significant reduction in data volume, but also makes them closer to actual values, further improving the accuracy of subsequent predictions. There are many duplicate data in the 2017 data, and multiple sets of data records during some repeated time periods are not the same, which makes it difficult to determine the accuracy of this type of data. Therefore, these data cannot be effectively deleted and used. Compared with the 2018 data, the accuracy of this part of data is relatively low, and it is ultimately decided to use the 2018 traffic flow data.

Based on the conclusion of the above feature analysis, it can be concluded that Kelvin temperature has the highest correlation. Therefore, while ensuring that obvious erroneous data is avoided, based on the calculated annual temperature average, the traffic flow data for the 28 days period from April 29 (Sunday) to May 26 (Saturday) of that year is selected. The temperature values during this period are all within twice the standard deviation range of the temperature mean, which is relatively stable and accurate. This can ensure the accuracy and generalization of the prediction as much as possible and reduce the accidental impact of other features on traffic volume.

Based on the actual values of road traffic data in 2018, it can be observed that the majority of workdays display a bimodal mode and a very small portion display a unimodal mode, while non workdays display a unimodal mode. Normally, the bimodal values of workdays occur at 7am and 4pm respectively, which are the commuting time for workers before and after work. On weekends, the single peak value appears around noon, and the shift in peak time is usually within one hour.

However, the occurrence of non-bimodal values on a few working days is due to the influence of weather, temperature, or holidays on people's travel patterns. On May 28(Monday), 2018, extreme weather thunderstorms occurred, resulting in a significant decrease in traffic flow on that day (Figure 6). January 1, 2018, was New Year's Day and also a statutory holiday in the United States, which had caused a significant decrease in traffic flow (Figure 7).

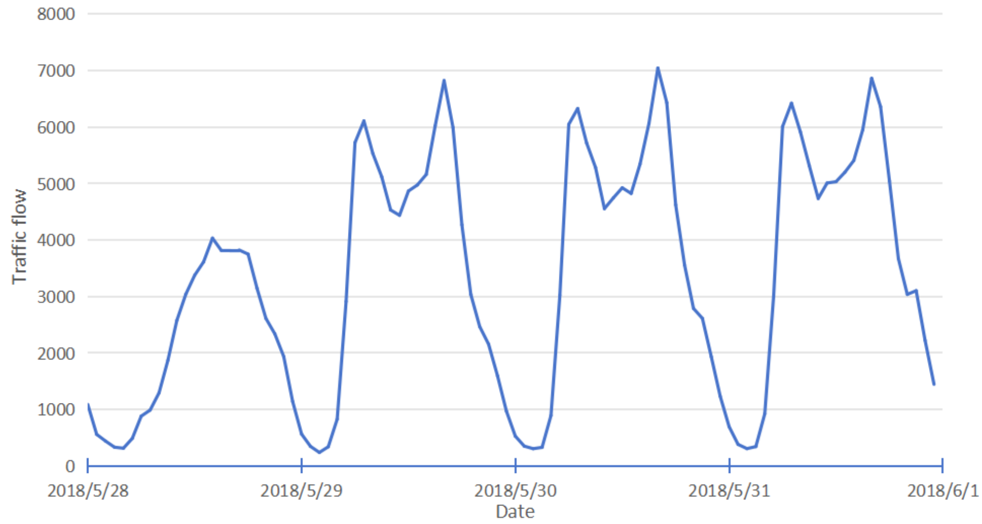

Fig. 6 Traffic flow in 2018.5.28

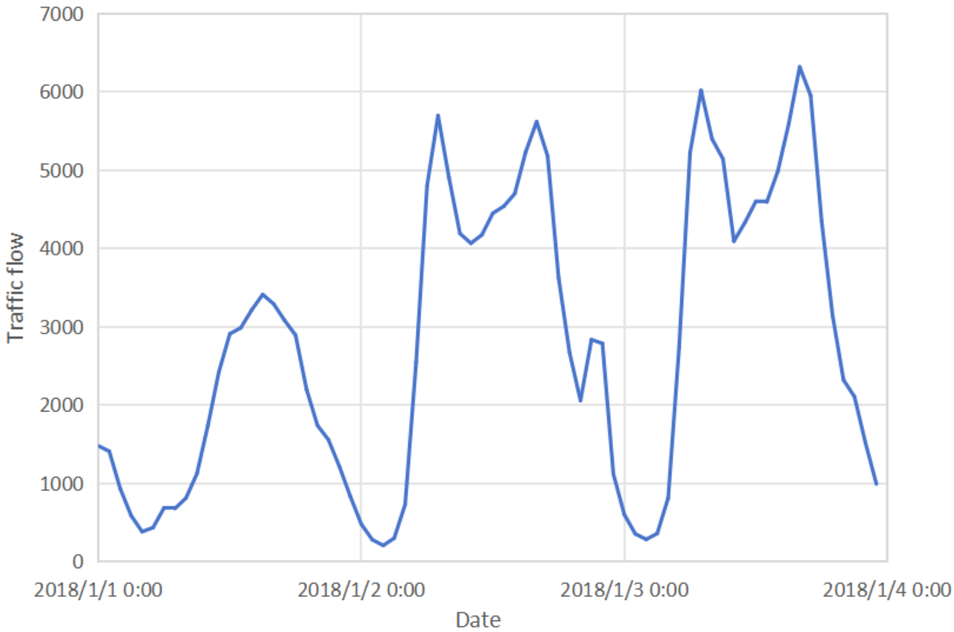

Fig. 7 Traffic flow in 2018.1.1

3.3. Traffic Flow Prediction

After understanding the general characteristics of the graph, this article selects the 28 days road traffic flow from April 29th, which is a total of four weeks, and uses LSTM for short-term traffic flow prediction. Table 3 shows the adjusted parameters after importing data into LSTM.

Table 3. Parameter Tuning

status	Hidden layer size	Batch_Size	Epochs	Figure
Traffic flow1	4	1	2000	8,9
Traffic flow2	4	1	1200	10,11
Traffic flow3	5	1	2000	12,13

Three predictions are made by changing the dimensions of the hidden layers and Epochs in the model. The comparison between the predicted results and the original data is shown in Figures 8, 10, and 12, and the difference between the predicted results and the original data is shown in Figures 9, 11, and 13.

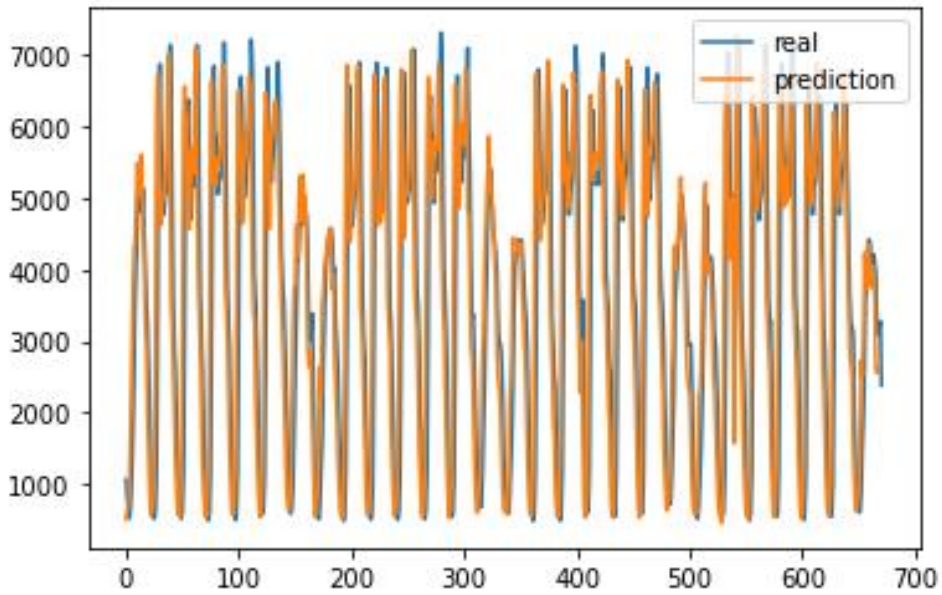

Fig. 8 Traffic flow prediction results (Traffic flow1)

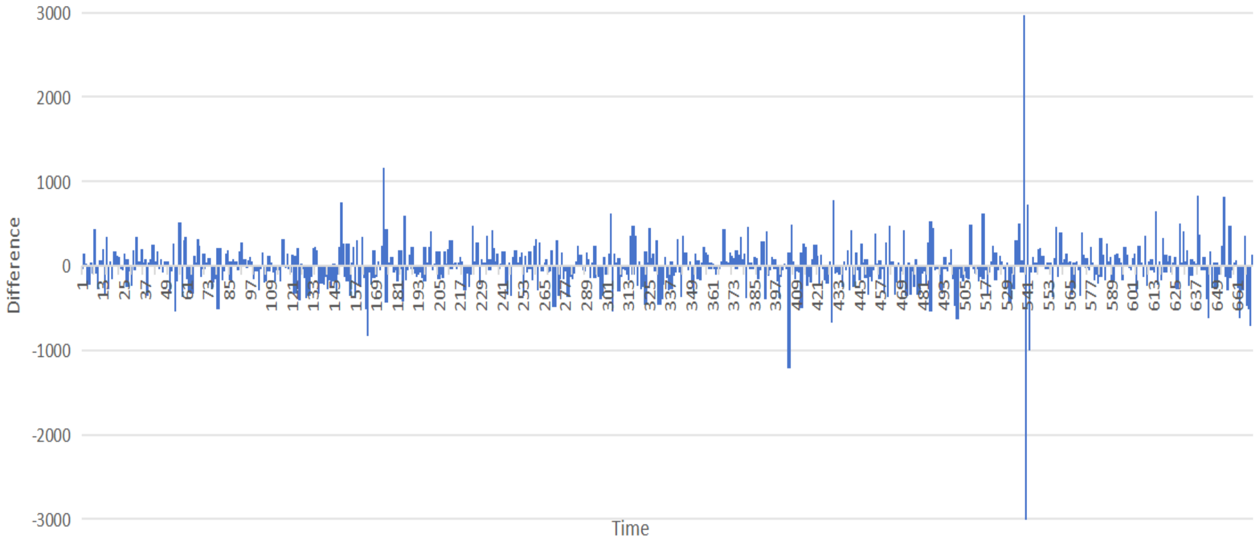

Fig. 9 Difference of Traffic flow prediction (Traffic flow1)

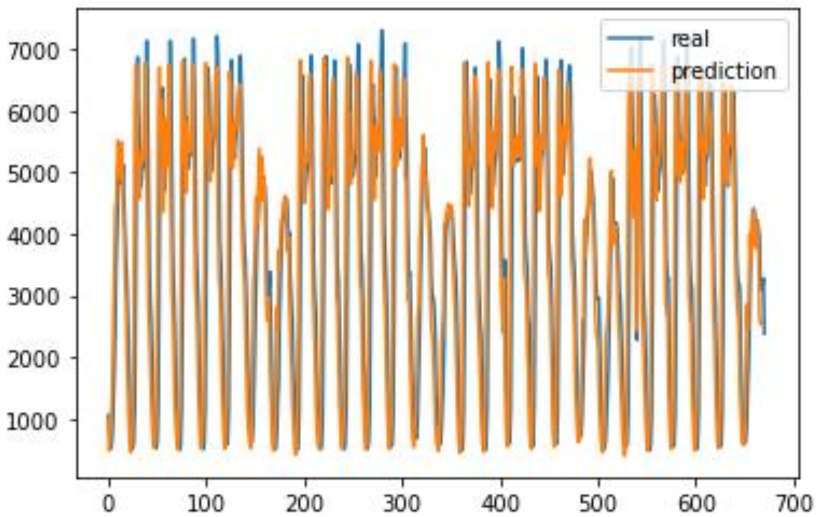

Fig. 10 Traffic flow prediction results (Traffic flow2)

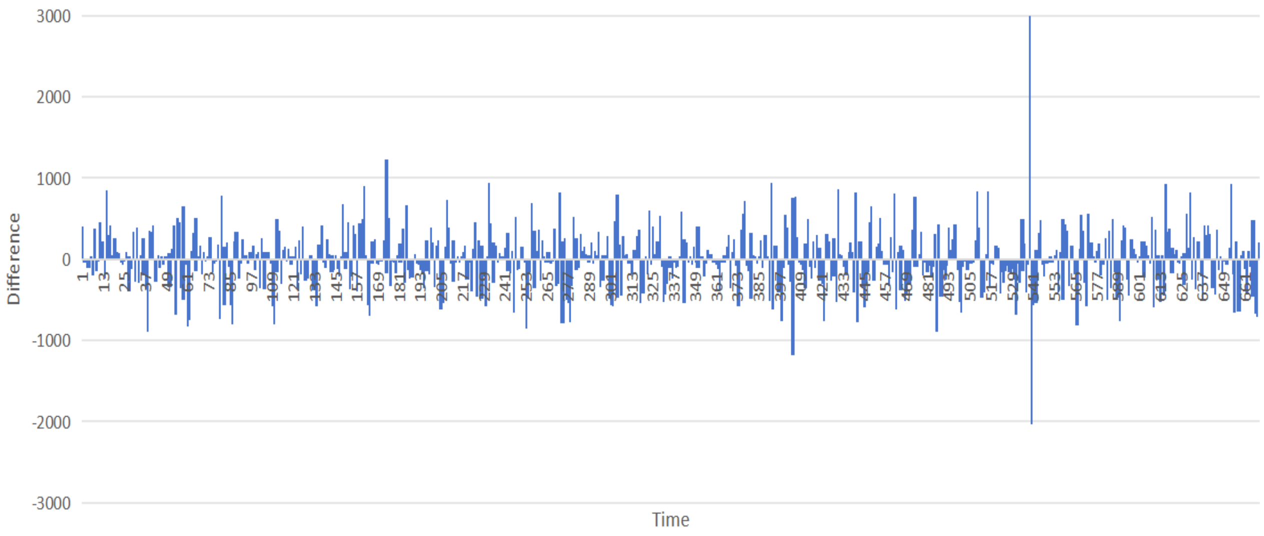

Fig. 11 Difference of Traffic flow prediction results (Traffic flow2)

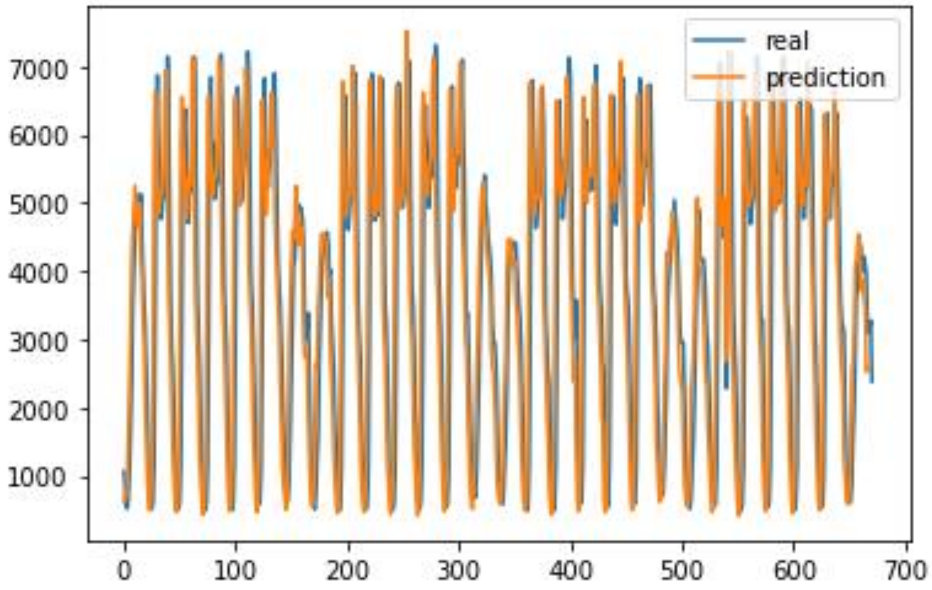

Fig. 12 Traffic flow prediction results (Traffic flow3)

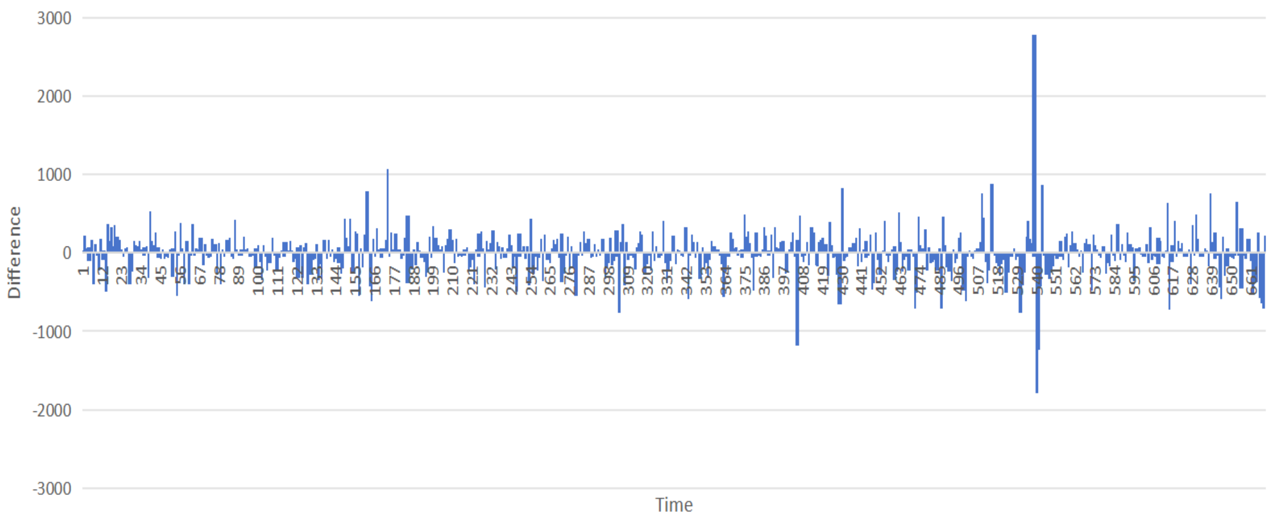

Fig. 13 Difference of Traffic flow prediction results (Traffic flow3)

By comparing the data, it is found that increasing the number of hidden layer sizes in the prediction model can reduce the error between the predicted and actual values, while reducing the number of complete traversals of the dataset (Epochs) can significantly increase the error.

Consequently, it is evident from the highway traffic flow forecast chart that weekday traffic flow displays a bimodal pattern, but weekend traffic flow displays a unimodal pattern. Furthermore, the graph shows that daily traffic is at its peak between 7 a.m. and 4 p.m., which is in line with earlier estimations of the real traffic flow on the route. The LSTM model's simulated data trends exhibit peaks at nearly identical sites, suggesting that the model has captured the general features and patterns of traffic flow and demonstrating the model's initial accuracy in predicting traffic flow.

4. Conclusion

Through research, this article finds that the predicted results of the highway on weekdays show a bimodal pattern, while on holidays, the predicted results all show a unimodal pattern. The bimodal hours between work days are 7am and 4pm respectively, while the single peak hours on weekends and holidays are around 12pm. This study can provide more comprehensive and accurate data analysis for urban traffic management systems, and provide better planning references for residents' travel, which can make urban transportation more orderly and convenient.

In some cases, the consistency between traffic volume and the above patterns and trends is relatively low. This is because traffic volume is affected by a few sporadic factors, such as extreme weather conditions that prevent travel, and travel surges or declines caused by statutory holidays. This type of data accounts for a very low proportion in the overall tens of thousands of data, so a few predicted results have high errors.

For such more complex problems, in the future, it may be possible to optimize the prediction model through joint analysis of weather characteristics and statutory holidays, making the prediction results more comprehensive and accurate.

References

[1] Zhang Kunpeng, Zheng Liang, Liu Zijian, et al. A deep leaning based multitask model for network-wide traffic speed prediction. Neurocomputing, 2020, 396: 438-450.

[2] Zhou Feng, Wang Shipu, Zhang Kunpeng. Short-term Traffic Flow Prediction Model of Urban Road Network Based on Graph Transformer Network. Science Technology and Engineering, 2024, 24(10): 4307-4316.

[3] He Yiyang, Yin Feng, Yuan Ping. Survey of short-term traffic flow prediction models in intelligent transportation system. Modern Computer, 2020(16): 8-11.

[4] Long Jiancheng, Guo Jiaqi. Review and prospect of dynamic traffic assignment research. Journal of Transportation Systems Engineering and Information Technology,2021,21(5): 125-138.

[5] Lv Hongyan, Feng Qian. A Review of Random Forest Algorithm Research. Journal of Hebei Academy of Sciences, 2019, 36(03): 37-41.

[6] Zhu Shiwei, Ye Baolin, Wu Weimin. A review and simulation study on short-term traffic flow prediction methods based on deep learning. Software Guide, 23(02), 182-193.

[7] Cao Yu, Wang Cheng, Wang Xin et al. Urban Road Short-term Traffic Flow Prediction based on Spatiotemporal Node Selection and Deep Learning. Journal of Computer Applications, 2020, 40(5): 1488-1493.

[8] Wang Xin. Research on Short-term Passenger Flow Prediction of Urban Rail Transit Based on Deep Learning and Multi-feature Fusion. Lanzhou Jiaotong University, 2024.

[9] Feng Biyu. Research on Short-term Passenger Flow Prediction of Urban Rail Transit Based on CNN-LSTM Combined Model[D]. East China Jiaotong University, 2021.

[10] Shao Hongxin, Soong Boon-Hee. Traffic Flow Prediction with Long Short-Term Memory Networks (LSTMs). Proceedings of 2016 IEEE Region 10 Conference Singapore: IEEE, 2016: 2986-2989.

[11] Doğan Erdem. Robust-LSTM: A Novel Approach to Short-traffic Flow Prediction based on Signal Decomposition. Soft Computing, 2022.26: 5227-5239.

[12] Zheng Haifeng, Lin Feng, Feng Xinxin et al. A Hybrid Deep Learning Model with Attention-based Conv-LSTM Networks for Short-Term Traffic Flow Prediction. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(11): 6910-6920.

[13] Shi Ruize, Du Lijing. Multi-Section Traffic Flow Prediction based on MLR-LSTM Neural Network. Sensors, 2022, 22(19): 7517.

[14] Kong Siqin. Short-Term Traffic Flow Prediction based on VMD-HSSA-BP Neural Network Model. Electronic Design Engineering, 2024, 32(10): 1-7.

[15] Liu Weiyuan et al. Research on Holiday Passenger Flow Prediction in Suzhou Rail Transit. Urban Rapid Rail Transit, 2021, 34(05): 66-73.