Construction of Short-Term Traffic Flow Prediction Model Based on IoT and Deep Learning Algorithms

Xiaowei Sun[1], Huili Dou[2]*
Zhejiang Institute of Communications, Hangzhou 310012, China[1]
Institute of Rail Transit, Zhejiang Institute of Communications, Hangzhou 310012, China[2]

***Abstract*—On a global scale, traffic problems are an essential factor affecting urban operations, particularly challenging the frequent occurrence of traffic congestion and accidents. The solution to the problem requires real-time and accurate prediction of traffic flow. This article mainly explores the application of the Internet of Things and deep learning in traffic flow prediction, aiming to solve the problem where existing methods cannot meet the requirements of real-time and accuracy. IoT devices, such as road sensors and in-vehicle GPS devices, which provides rich information for traffic flow prediction. With the ability of deep learning, it can not only learn and abstract a large amount of complex traffic data but also handle traffic flow prediction tasks in various complex situations. During the model construction process, the complexity of the road network was fully considered, practical algorithms were designed to fuse multi-source data, and the structure of the model was optimized to meet the needs of real-time prediction. The experimental results show that the absolute error of the test results is generally less than 6km/h, which can better reflect the traffic speed of the road section in the future.**

***Keywords*—Internet of things; deep learning algorithm; short term traffic flow; prediction model**

I. INTRODUCTION

With increase in serious urban transportation problems, seriously affecting the functional operation of cities and the quality of life of citizens. Among them, the frequent occurrence of traffic congestion and accidents has become a common problem worldwide [1, 2]. Traffic flow prediction is crucial for urban traffic management [3, 4], as it can accurately and effectively predict traffic flow. Traffic management departments can arrange traffic police forces in advance, dispatch traffic lights reasonably, and effectively guide vehicles based on the prediction results, thereby alleviating traffic congestion and improving urban road capacity [5]. For example, when it is predicted that the traffic flow in a specific area will significantly increase, traffic control or evacuation work can be carried out in advance to avoid traffic congestion. For drivers and passengers, knowing the traffic flow situation in the future in advance can provide important references for their travel decisions, reduce waiting time, and improve travel efficiency [6]. Therefore, traffic flow prediction is also significant for the research and development. In areas such as autonomous driving, traffic signal optimization, and travel recommendation, accurate traffic flow prediction results are all needed [7].

In the information age of the 21st century, deep learning has seen significant growth in many fields. Traffic flow prediction, as a critical challenge, is gradually benefiting from these two technologies. The Internet of Things, also known as the extension of the Internet, connects various objects in the physical world, enabling them to collect and exchange data. The Internet of Things technology has played a considerable role in traffic flow prediction [8]. The use of IOT makes it possible to obtain large amounts of data, which greatly improves the accuracy and real-time nature of data sources. Deep learning, as an artificial intelligence algorithm, has also played an enormous role in traffic flow prediction. Deep learning can learn and understand a large amount of complex data and abstract valuable features [9, 10]. It has been successfully applied to traffic flow prediction. The application of deep learning enables traffic flow prediction models to understand and process complex traffic data, thereby improving the accuracy of predictions. Although the Internet of Things and deep learning have been successfully applied in traffic flow prediction, their potential still needs to be fully explored [11,12]. In terms of deep learning, how to design more effective models to handle more complex situations (such as traffic congestion, accidents, etc.) is also an important research direction. Therefore, constructing traffic flow prediction models based on the Internet of Things and deep learning is an essential direction in current traffic research.

However, although the Internet of Things and deep learning have been applied in traffic flow prediction, they still need to overcome many challenges. Data quality issues, such as sensor failures, network transmission issues, incomplete data, etc., can all affect the availability and accuracy of data [13, 14]. In terms of model design and optimization, how to design and optimize the model based on specific traffic flow prediction tasks and how to improve the interpretability of the model are all issues that need to be addressed. Real-time prediction problems require the model to have efficient data processing and computational capabilities to meet the needs of real-time prediction. The complexity of road networks and how to effectively integrate various factors into the model is a challenging issue. For the fusion of multi-source data, in addition to traffic flow data, weather data, social event data, social media data, etc., can also be utilized. How to effectively integrate these diverse data and improve prediction accuracy is a new challenge. Therefore, the significance of studying the short-term traffic flow prediction model based on the Internet of Things and deep learning algorithm is that it can significantly improve the efficiency of urban traffic management, reduce traffic congestion, and reduce the incidence of traffic accidents. Through real-time and accurate traffic flow prediction, traffic management departments can

allocate resources reasonably, optimize traffic signal scheduling, and guide vehicles to drive effectively. At the same time, this research has a significant impact on the development of intelligent transportation systems, especially in the fields of autonomous driving and traffic signal optimization.

II. Application of IoT and Deep Learning

A. Deep Learning Algorithm

The architecture of the Temporal Convolutional Network (TCN) is designed based on the characteristics of the latest convolutional system used for sequential data. It takes a time series as input and models the temporal correlation in each temporal data. Unlike the traditional Recurrent Neural Network (RNN) which recourses along the time axis of a sequence and introduces a large number of learning parameters, making the model difficult to optimize. TCN combines simplicity, autoregressive prediction, and very long memory without the recursive mechanism of RNN, which facilitates parallelization, as shown in Fig. 1. Therefore, TCN can effectively combine computational advantages with representation capabilities to achieve efficient and good predictive performance [15]. Due to the above advantages, TCN can be well applied to analyse data with strict order and is widely used for predicting various scenarios. Therefore, TCN is suitable for modelling and analysing time-series data sensors monitor. By capturing simple patterns in sensor time series and generating more complex patterns in higher-level layers, TCN can better extract temporal features [16].

Graph Convolutional Neural Networks (GCN) are feature extractors designed based on graph data [17]. The essence of GCN is to apply convolution to graph neural networks, which can flexibly extract structural information of graph data and reduce computational complexity. Due to the good complementary relationship between node attribute information and structural information in graph data, GCN can use the network layer to simultaneously learn the data structure and attribute information in the graph and use the two to represent the relationship between nodes [18, 19].

B. Application of Deep Learning in Traffic Flow Prediction

1) Reactive control of short-term traffic flow: The timing control method calculates the timing scheme based on historical traffic flow data and predetermined optimization objectives. The biggest drawback of this scheme is that it cannot adapt to the dynamic changes in traffic flow, resulting in limited control effectiveness. The reactive traffic signal control method adjusts the signal timing strategy based on existing traffic flow characteristics to improve control effectiveness without considering the impact of traffic flow prediction on control effectiveness [20, 21]. In recent years, green ratio optimization, phase difference optimization, mathematical programming, multi-objective optimization, dynamic programming and other methods have emerged in the field of reactive control [22]. Based on the analysis and research of the delay law of the vehicles at the intersection, an optimization model of phase difference adjustment of the wire control system is established. On the one hand, due to the lack of traffic flow forecasting mechanism, the control method does not consider the potential impact of the current control scheme on the future traffic conditions, so the control effect is limited. On the one hand, the control method lacks a traffic flow prediction mechanism. It needs to consider the potential impact of the current control plan on future traffic conditions, resulting in limited control effectiveness. On the other hand, most of these existing control methods adopt a single-machine computing environment, which cannot meet the real-time requirements of traffic optimization and control in the context of big data [23, 24].

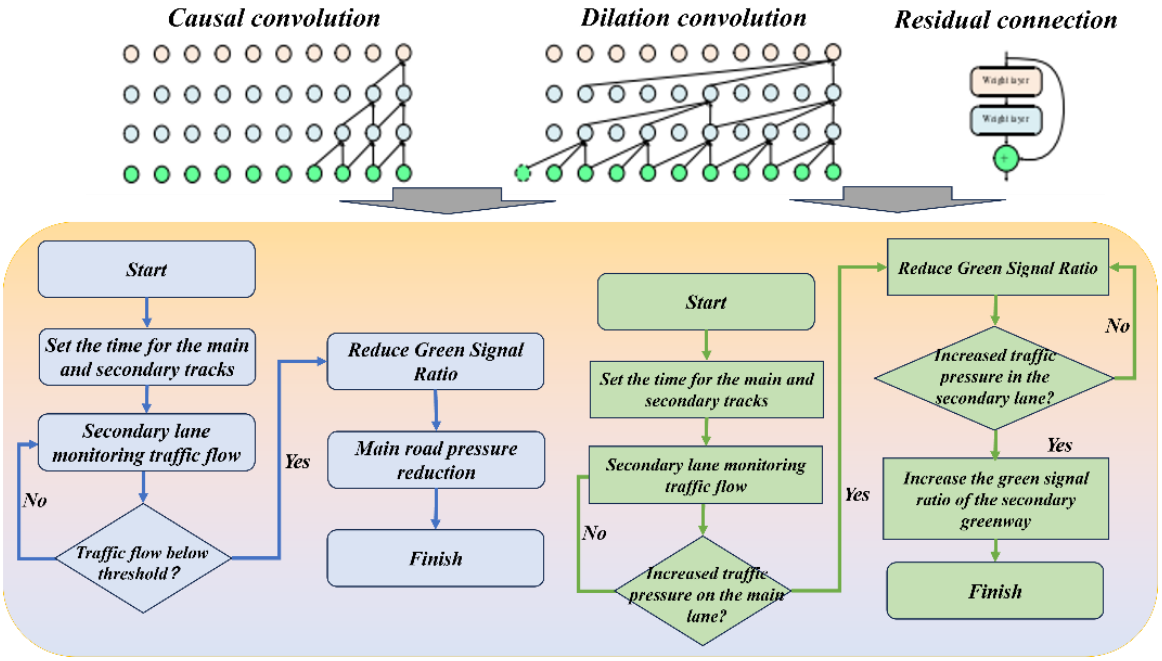

Fig. 1. Deep learning applied to short-term traffic models.

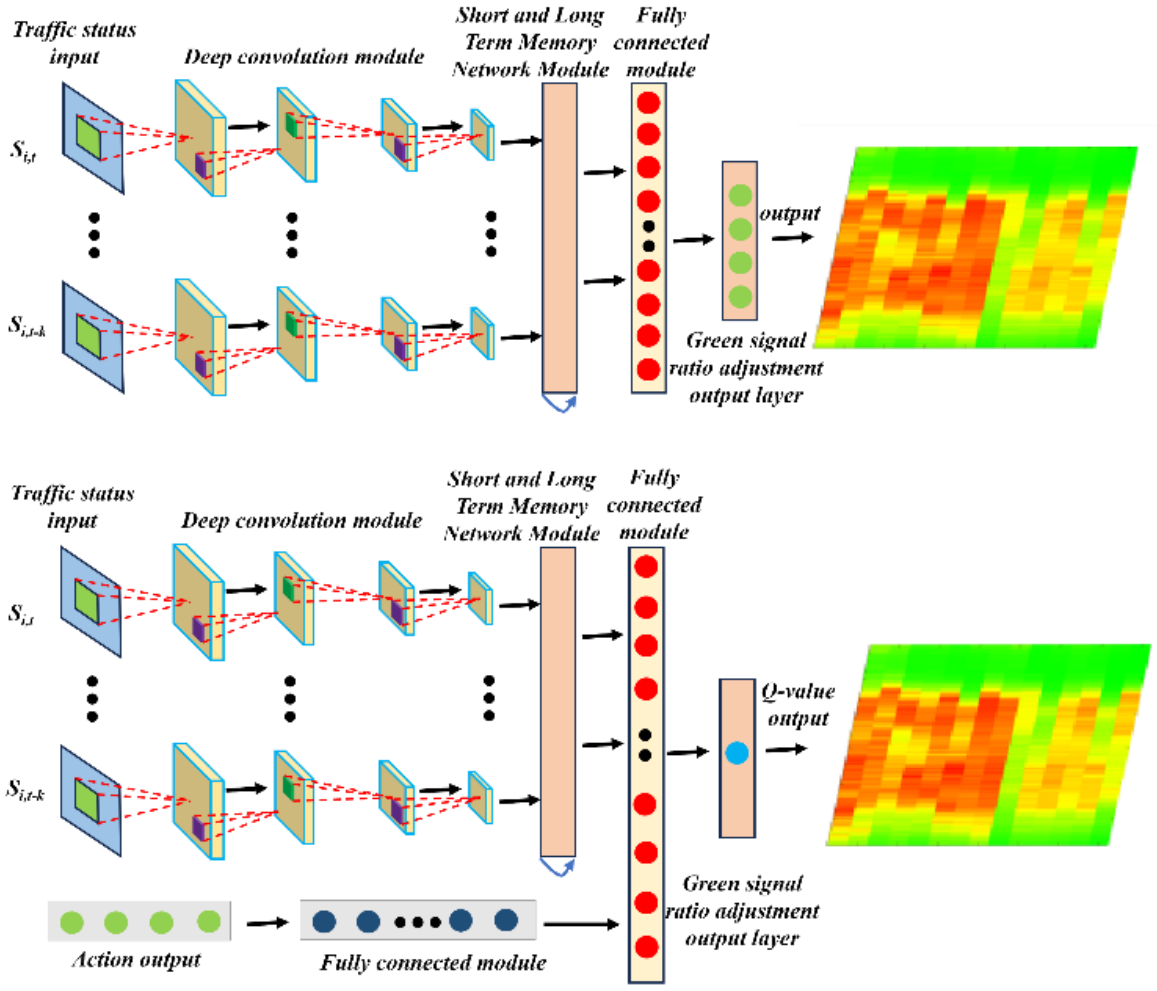

Fig. 2. Construction process of short-term traffic network image samples.

Similar to facial muscle movements forming different expressions, the spatiotemporal evolution of short-term traffic flow in the road network constitutes different forms of traffic. From a spatial perspective, short-term traffic network flows are interdependent and interrelated. Congestion on a road section may affect nearby or even further road sections. At the same time, in terms of time dimension, some similar but randomly fluctuating traffic flow characteristics will repeatedly appear on a road segment. Therefore, short-term traffic network flow feature learning needs to comprehensively consider the temporal and spatial characteristics, as well as periodic repeatability characteristics. As shown in Fig. 2, the construction process of short-term traffic network image samples is presented.

2) Short-term traffic flow model predictive control: Based on the model predictive control (MPC) framework, the predictive model predicts the future traffic dynamics, and the potential control performance of the candidate scheme is calculated [25, 26]. In the current control cycle, the first element of the optimization sequence is applied to the traffic system model and restarts the next round of the rolling optimization process based on the feedback traffic status and prediction model. By comprehensively considering the cumulative impact trends, visionary control decisions are generated. Therefore, in urban expressway traffic control, MPC can synergistically adjust the traffic flow at the ramp entrance and exit. In highway traffic control, MPC can coordinate and solve problems such as speed restrictions, lane allocation, and release time of vehicle queues on on-ramps. The rolling time domain method of MPC can further plan the path selection problem of multiple travellers. Abstract boundary control and path guidance as economic MPC problems is to improve the mobility of urban networks. Furthermore, in the MPC framework, macro traffic flow and exhaust emission models are introduced to reduce the probability of traffic congestion and reduce pollution emissions. However, to apply the MPC control strategy to more complex traffic network systems, the contradiction between the time required optimizing the objective function in the prediction time domain and the real-time performance of online control still needs to be solved urgently [27].

In order to reduce the solving time of the MPC objective function, the entire road network is decomposed into several regional subnets to accelerate the calculation process. In addition, by parameterizing the macro traffic prediction model,

it reduces the time of online computation in the rolling time domain. Based on the improved macro traffic flow model, and quadratic programming provide a solution for improving the real-time performance of MPC in traffic flow control. However, they reduce the calculation time for solving the objective function in the prediction time domain. They can only roughly describe macroscopic traffic flow phenomena such as traffic density, traffic flow, occupancy rate, etc., which, to some extent, reduces the control effect.

3) Deep reinforcement learning control for short-term traffic flow: Deep reinforcement learning is a feedback-based iterative learning method based on a deep learning evaluation mechanism. Deep learning involves constructing a hierarchical neural network that simulates human brain thinking. The development of deep learning to this day mainly includes CNN, deep belief networks DBN, DSAEs, and LSTM, each with its advantages and applicability. However, these deep learning methods focus on the learning of traffic flow characteristics at the segment level. When traffic control rises to the regional road network, the best control timing may be missed due to the inability to obtain the interconnectivity between segment traffic flows [28, 29].

The training of deep learning in the context of big data takes several days or even weeks, so reducing the time cost of model training while ensuring training accuracy has become a hot topic in academic and industrial research [30]. Distributed deep learning is a powerful tool to accelerate deep neural network training. By introducing predictive hierarchical caching strategy in distributed training, it can improve the cache hit rate of data, shorten synchronization time and network blocking times. Secondly, through the sparse gradient compression mechanism of entropy, the propagation gradient threshold can be determined dynamically and the data volume of the propagation gradient can be compressed to reduce the communication load. By quantifying the performance differences of each node and dynamically allocating the training batches of each node, the time of each iteration between nodes is approximately consistent, thereby improving the impact of gradient obsolescence on convergence in asynchronous parallel optimization [31].

4) Distributed parallel processing of traffic flow big data: The rapid development of the Internet of Things and artificial intelligence has provided strong support for the interconnection of vehicles, pedestrians, traffic lights, roadside equipment, and traffic management centers. It is necessary to establish new theories and methods for traffic network flow adaptive control and achieve the next generation of data-driven intelligent transportation systems (ITS). In the context of the Internet of Vehicles (IoV), the traffic flow data collected by multi-source heterogeneous sensors is rapidly increasing, and the era of big data in transportation has arrived. Cloud computing uses a universal computing model to deploy computing tasks to a computing resource pool, allowing users to transparently access computing resources, storage space, and information services according to their needs. It is one of the most effective methods for processing big data. It has the self-maintenance and management function of virtual computing resources and can dynamically acquire or release computing resources to adapt to dynamic application workloads.

In a traffic control system, if all raw data is sent to a remote traffic control center for processing and analysis using cloud computing, it requires extremely high network bandwidth. In addition, when optimization decisions are returned from cloud computing centers to traffic signal controllers, local traffic dynamics may have undergone significant changes. This poses hidden dangers to the safety and real-time performance of traffic control. Edge computing is an expansion of cloud computing architecture, pushing some computing intelligence, data processing, storage, and services from the cloud to the network's edge. It enables analysis and processing to occur on the side of the data source, avoiding response delays or data security risks caused by long-distance, high-capacity data communication as much as possible.

III. Construction of Short-Term Traffic Flow Prediction Model Based on IoT and Deep Learning Algorithms

A. Overall Architecture

As shown in Fig. 3, the model predictive control architecture is deployed on a cloud computing platform to collaboratively control the signal timing strategies of various intersections from a global perspective in order to improve the traffic capacity of vehicles in the road network and alleviate traffic congestion. Establish information channels between transportation networks and cloud computing through communication technologies like the Internet and 5G. The location, speed, and intersection status of vehicles in the transportation network are collected through multi-source sensors and then uploaded to the cloud control center. The nonanalytical prediction model of the cloud control center predicts the trend of traffic flow changes in the future based on the traffic status collected at the current time, pre-set control requirements, and control the sequences generated by optimization algorithms and provides an evaluation of the cumulative control performance of the control sequence in the future. Using the distributed computing of cloud computing, multiple computing nodes participate in the calculation to accelerate the optimal control sequence solution. In the current control cycle, the first strategy in the optimal control sequence is selected and applied to the traffic network flow system. The rolling time domain method is used to continuously implement this process and effectively control the traffic flow of the road network.

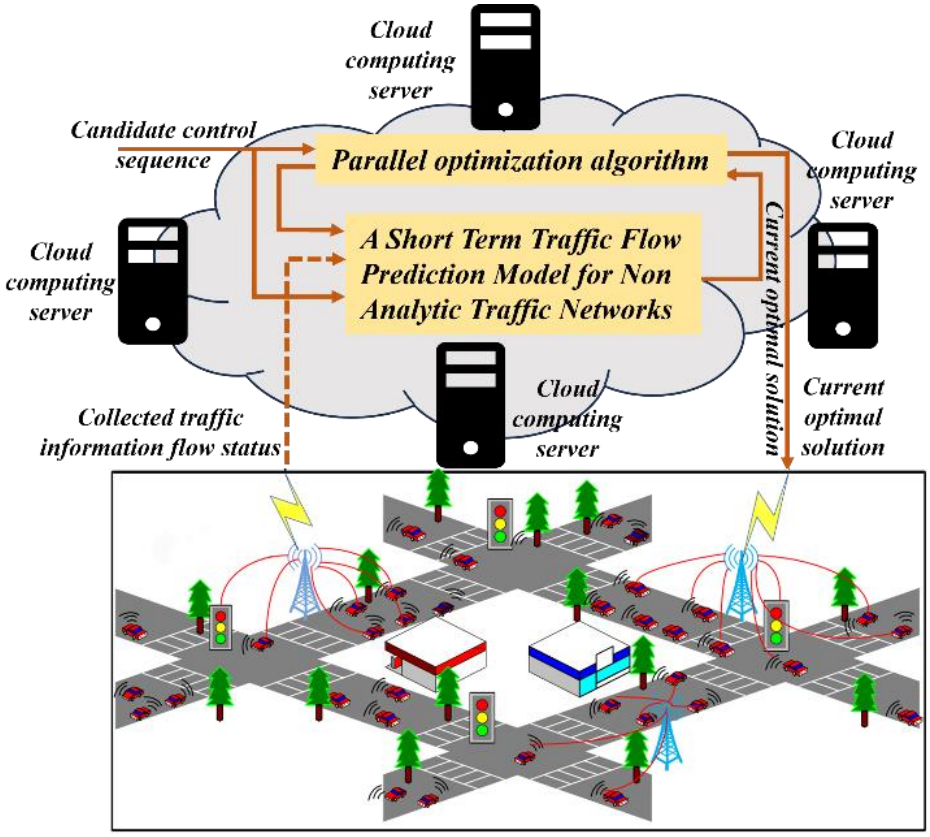

Fig. 3. Overall architecture diagram.

B. Short Term Traffic Flow Simulation Modelling

In urban road networks, vehicles face complex road conditions on their driving routes, as shown in Fig. 4.

1) Maximum speed limit: The maximum allowable driving speed depends on the road infrastructure and equipment. When the road conditions and driving equipment performance are good, the maximum driving speed can increase correspondingly. Otherwise, it is necessary to reduce the maximum driving speed to meet actual needs.

2) District-specific speed limits: The urban road network has many special areas, such as hospitals, schools, military administration areas, and signalized intersections, where vehicles need to slow down appropriately.

3) Temporary speed limit: When encountering sudden situations such as roadbed maintenance, abnormal weather, and traffic accidents, the relevant traffic management department will issue a temporary speed limit notice, and vehicles should slow down in advance and pass slowly when driving to the section. $V_{lim}(x)$ dynamically divides the road into a series of sections with different speed restrictions.

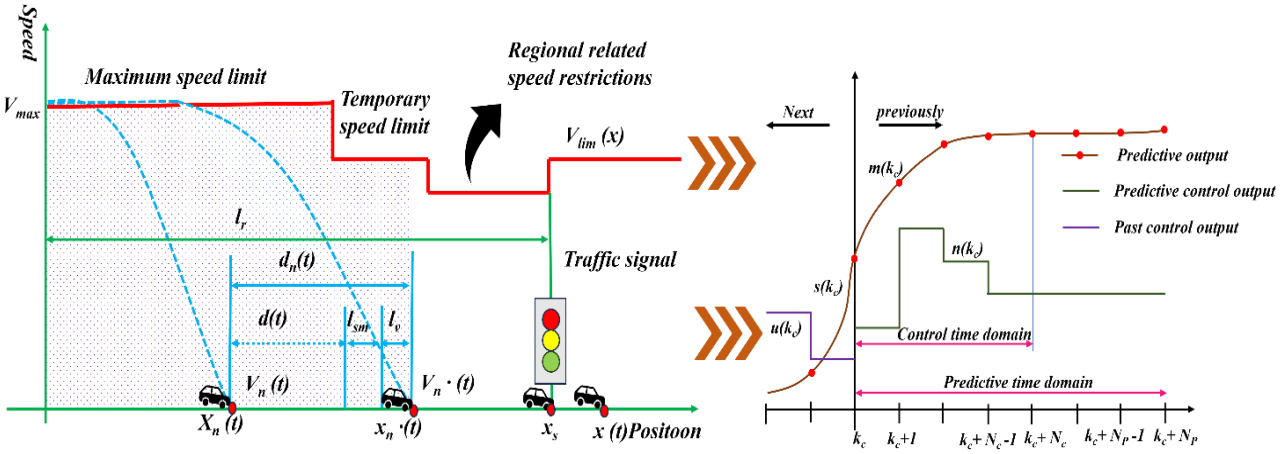

Fig. 4. Spatiotemporal constraints on vehicle motion.

C. Establishment of Short-term Traffic Flow Model Prediction Model

Decompose the complete dataset D into R parts, with each part represented by Dr. (r=1, 2,.. R). Therefore, dataset D can be represented as a set of subsets of data, Dr., as shown in Formula (1):

$$D = U_{\mathrm{r}=1}^{R} D^{\mathrm{r}} = U_{\mathrm{r}=1}^{R} U_{\mathrm{n}=1}^{N} D_{\mathrm{n}}^{\mathrm{r}} \tag{1}$$

Among them, Nr represents the size of the r-th data subset, and Dr n represents the nth data on the r-th data subset.

The objective function of parallel training of CNN-LSTM model is shown in Formula (2):

$$J = \min \frac{1}{N} \sum_{\mathrm{r}=1}^{R} j^{r} = \min \frac{1}{N} \sum_{\mathrm{r}=1}^{R} \sum_{n=1}^{N} j_{n}^{r} \tag{2}$$

order $j_n^r = J_n^r = \left\| y_n^r - \hat{y}_n^r \right\|^2 / 2M$, Here y_n^r and $\hat{y}_n^r$ are the observation and prediction vectors for the n-th sample in the r-th dataset, respectively, The training of the CNN-LSTM model based on the complete dataset D is to minimize the objective function described by the formula, thereby obtaining ideal weights and biases, known as global learning parameters. Furthermore, The weights and biases trained by minimizing local objective function Jr on data subset Dr are called local learning parameters. For parallel feature forward learning processes, the output values of different types of network layers are synchronously calculated in parallel based on corresponding data subsets. At time t, denoted by $a_{n,j,c}^{r,l}(t)$, the CNN layer extracts local feature values based on the data subset Dr. The calculation Formula (3) is as follows:

$$a_{n,j,c}^{r,l}(t) = \sigma\left(\sum_{i=1}^{N_c^{l-1}} a_{n,i,c}^{r,l-1}(t) * \omega_{j,i,c}^{r,l} + b_{j,c}^{r,l}\right) \tag{3}$$

Among them, c represents the convolutional layer of the CNN-LSTM model:

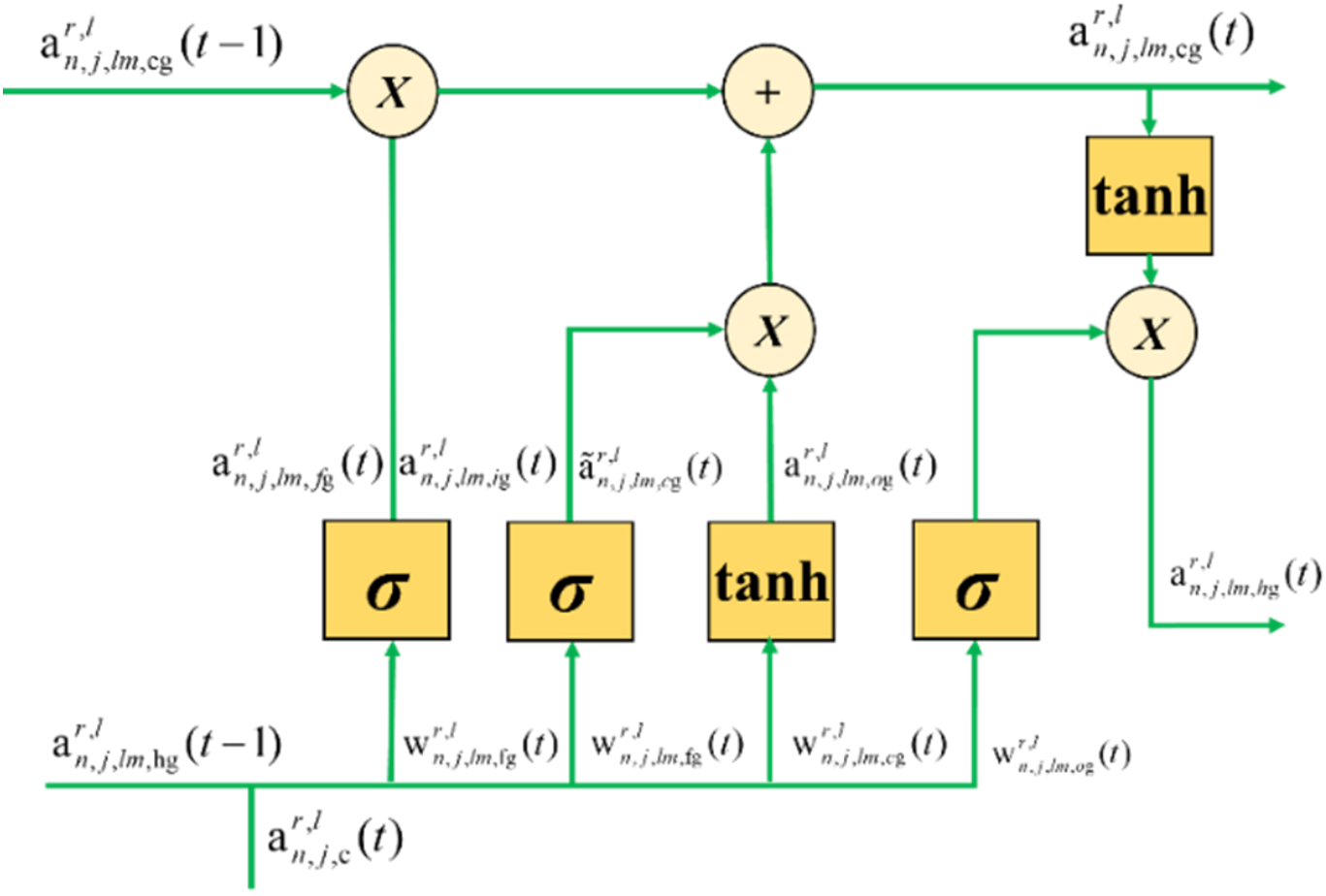

Fig. 5. Structural diagram of LSTM.

As shown in Fig. 5, in the LSTM module, the output values of forgetting gates (such as formulas), input gates (such as formulas), cellular states (such as formulas), output gates (such as formulas), and implicit states (such as formulas) at time t are represented by $a^{r,l}_{n,j,lm,\mathrm{fg}}(t)$, $a^{r,l}_{n,j,lm,ig}(t)$, $a^{r,l}_{n,j,lm,cg}(t)$, $a^{r,l}_{n,j,lm,og}(t)$, and, respectively. The calculation formulas for these locally activated feature values are shown in (4) - (9):

$$a^{r,l}_{n,j,lm,fg}(t)=\sigma(\sum_{i=1}^{N^l_{hg}} w^{r,l}_{j,i,lm,fg}(t-1)+\sum_{i=N^l_{hg}+1}^{N^l_{hg}+N^{l-1}_c} w^{r,l}_{j,i,lm,fg}a^{r,l-1}_{n,i,c}(t)+b^{r,l}_{j,lm,fg}) \tag{4}$$

$$a^{r,l}_{n,j,lm,ig}(t)=\sigma(\sum_{i=1}^{N^l_{hg}} w^{r,l}_{j,i,lm,hg}a^{r,l}_{n,i,lm,hg}(t-1)+\sum_{i=N^l_{hg}+1}^{N^l_{hg}+N^{l-1}_c} w^{r,l}_{j,i,lm,ig}a^{r,l-1}_{n,i,c}(t)+b^{r,l}_{j,lm,ig}) \tag{5}$$

$$a^{r,l}_{n,j,lm,cg}(t)=a^{r,l}_{j,i,lm,fg}(t)a^{r,l}_{n,i,lm,cg}(t-1)+a^{r,l}_{n,i,lm,ig}(t)\tilde{a}^{r,l}_{n,i,lm,cg}(t) \tag{6}$$

$$\tilde{a}^{r,l}_{n,j,lm,cg}(t)=\tanh(\sum_{i=1}^{N^l_{hg}} w^{r,l}_{j,i,lm,cg}a^{r,l}_{n,i,lm,hg}(t-1)+\sum_{i=N^l_{hg}+1}^{N^l_{hg}+N^{l-1}_c} w^{r,l}_{j,i,lm,cg}a^{r,l-1}_{n,i,c}(t)+b^{r,l}_{j,lm,cg}) \tag{7}$$

$$a^{r,l}_{n,j,lm,og}(t)=\sigma(\sum_{i=1}^{N^l_{hg}} w^{r,l}_{j,i,lm,og}a^{r,l}_{n,i,lm,hg}(t-1)+\sum_{i=N^l_{hg}+1}^{N^l_{hg}+N^{l-1}_c} w^{r,l}_{j,i,lm,og}a^{r,l-1}_{n,i,c}(t)+b^{r,l}_{j,lm,og}) \tag{8}$$

$$a^{r,l}_{n,j,lm,hg}(t)=a^{r,l}_{n,i,lm,og}(t)\tanh(a^{r,l}_{n,i,lm,cg}(t)) \tag{9}$$

Let $a^{r,l}_{n,j,f}(t)$ represent the network output of the fully connected layer at time t, and the calculation Formula is shown in (10):

$$a^{r,l}_{n,j,f}(t)=\sigma(\sum_{i=1}^{N^{l-1}_f} w^{r,l-1}_{j,i,f}(t)+b^{r,l}_{j,f}) \tag{10}$$

Among them, f represents the fully connected layer. When layer l is a fully connected layer, i represents the i-th input neuron, j represents the j-th output neuron. $w^{r,l-1}_{j,i,f}(t)$ and $b^{r,l}_{j,f}$ represent the weight and bias of layer l (which is a fully connected layer) on the r-th data subset, respectively, and N^{l-1}_f represents the number of neurons in the previous layer; σ(·) represents the RELU activation function.

Based on the classical gradient descent criterion, the relationship between global learning parameters and local learning parameters in the parallel error backpropagation process is derived layer by layer. The calculation formulas for updating the global learning parameters $w^l_{j,i,f}(t)$ and $b^l_{j,f}(t)$ in the fully connected layer at time step t are shown in Formulas (11) - (12):

$$w^l_{j,i,f}(t)=w^l_{j,i,f}(t-1)-\eta\frac{\partial J}{\partial w^l_{j,i,f}}=\frac{1}{R}\sum_{r=1}^{R} w^l_{j,i,f}(t) \tag{11}$$

$$b^l_{j,f}(t)=b^l_{j,f}(t-1)-\eta\frac{\partial J}{\partial b^l_{j,f}}=\frac{1}{R}\sum_{r=1}^{R} b^l_{j,f}(t) \tag{12}$$

Among them, the local weights and biases in the fully connected layer are calculated as shown in Formulas (13) - (14):

$$w^{r,l}_{j,i,f}(t)=w^l_{j,i,f}(t-1)-\frac{\eta}{N}\sum_{n=1}^{N^r} R\delta^{r,l}_{n,j,f}a^{r,l-1}_{n,i,f} \tag{13}$$

$$b^{r,l}_{j,f}(t)=b^l_{j,f}(t-1)-\frac{\eta}{N}\sum_{n=1}^{N^r} R\delta^{r,l}_{n,j,f} \tag{14}$$

The global adaptive learning rate of parallel training of the CNN-LSTM model can be obtained by calculating the local gradient sum, as shown in Formulas (15) - (16):

$$\eta(t)=\frac{l_r}{\mu+\sqrt{G(t)}} \tag{15}$$

$$G(t)=\rho G(t-1)+(1-\rho)g(t) \tag{16}$$

Among them, G(t) represents the sum of squares of gradients with attenuation factors; ρ Similar to the attenuation factor in the momentum gradient descent method, it represents the impact of past gradients on current parameter updates, typically taking a value of 0.9; Lr represents the basic learning rate; μ It is a minimal constant that prevents the denominator from being zero. The adaptive learning rate has advantages over the traditional fixed learning rate, because it can adjust the learning rate according to the gradient of the parameter itself, so as to achieve better convergence effect. Regardless of whether the training data set D is evenly divided or inevenly divided, the adaptive learning rate ensures that the convergence results are almost identical to the serial training method. Therefore, the parallel training theory of the CNN-LSTM model ensures that the global learning features of the large dataset can be obtained from the parallel learning of each decomposed data subset.

IV. Model Experiment and Result Analysis

Fig. 6 shows the trend of CP curves for MAE indicators using different prediction methods. For the prediction tasks of traffic network flow in 5min, 15min, 30min, and 60min, the CNN-LSTM prediction method has 100%, 85.71%, 85.71%, and 71.4% of MAE errors controlled within 20 for expressways, respectively. Compared to other prediction methods, the CP curve of the CNN-LSTM prediction method is always located at the top left of the graph in different traffic network flow prediction tasks, indicating that the CNN-LSTM method has more advantages in improving the accuracy of traffic network flow prediction. The universal ability measures the adaptability of the CNN LSTM model to prediction tasks in

different traffic scenarios. As can be seen from the figure, MAE prediction error of CNN-LSTM method in different traffic network flow prediction tasks is mainly kept between 12.31 and 19.05. This shows that CNN-LSTM model has competitive adaptability in various prediction scenarios under the same prediction time domain. However, the MAE indices of DTR and SVR methods fluctuate greatly under different prediction tasks. These two prediction models are susceptible to differences in the prediction time domain, so their predictive universal ability could be better in different traffic scenarios. CNN-LSTM is more stable than other methods in terms of universal prediction ability. Through comparison, it was further found that the prediction accuracy of the CNN-LSTM method in larger prediction time domain tasks is lower than that in smaller prediction time domain tasks. This is because under the same training cycle, as the prediction time domain increases, the difficulty of predicting the future multi-step traffic dynamic evolution trend increases. In the future, the deep learning model structure will be improved to enhance the feature extraction ability of traffic network flow in larger prediction time domains.

MPC online optimization is carried out rolling, and the end of the current control cycle starts the subsequent predictive time domain optimization. Therefore, short-term traffic network flow predictive control based on the rolling time domain includes multiple predictive time domain optimization processes. Fig. 7 shows the computational efficiency of predicting time domain optimization for each control cycle. The MPC control scheme based on Spark cloud parallel optimization takes much less time at each control time step than the single machine serial optimization MPC control scheme. Especially for single-machine computing environments, all chromosomes sequentially call the traffic network flow prediction model circularly to obtain evaluation values for control effectiveness. This calculation method requires a high computational time cost for non-analytical micro prediction models.

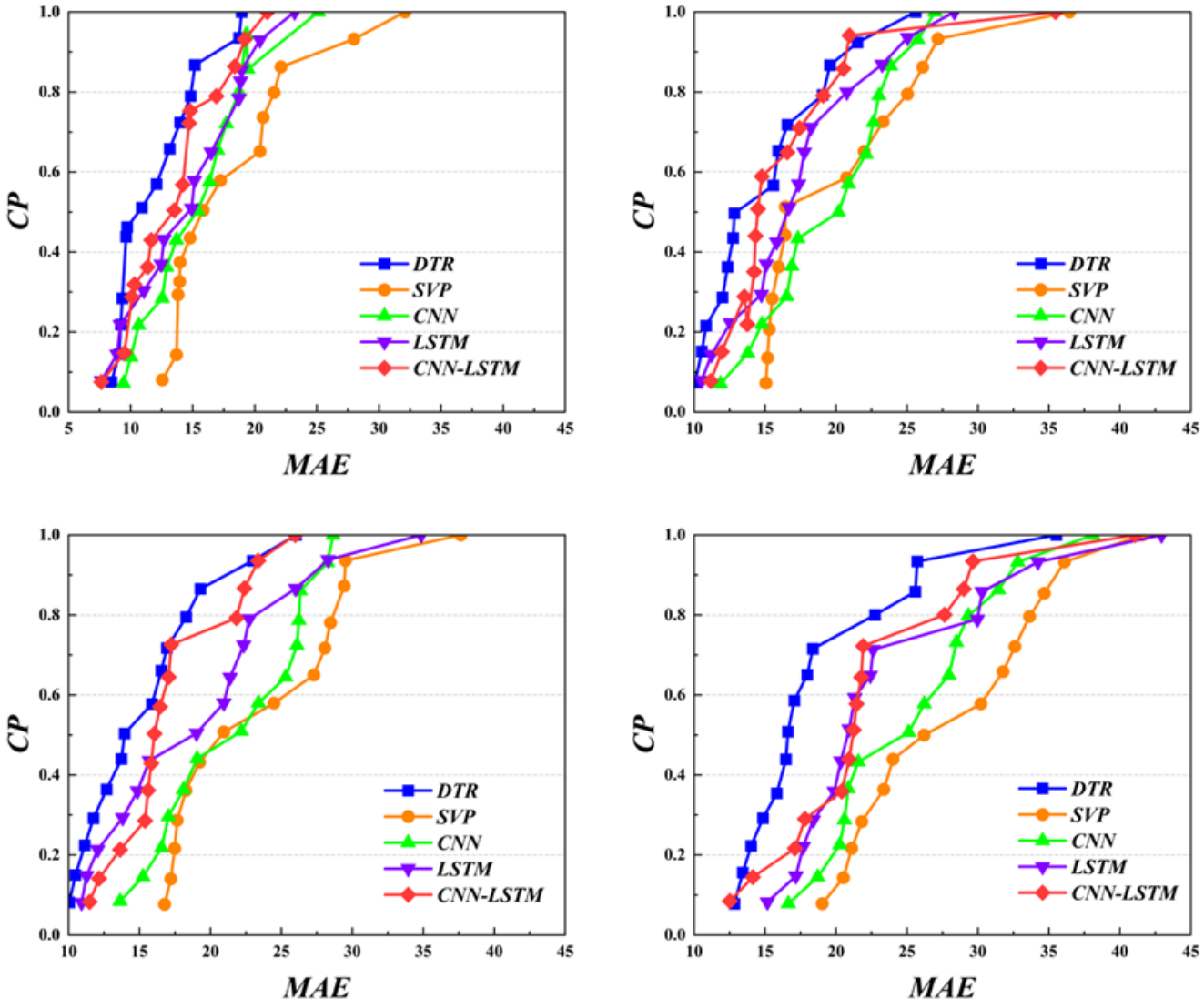

Fig. 6. Cumulative distribution traffic flow prediction.

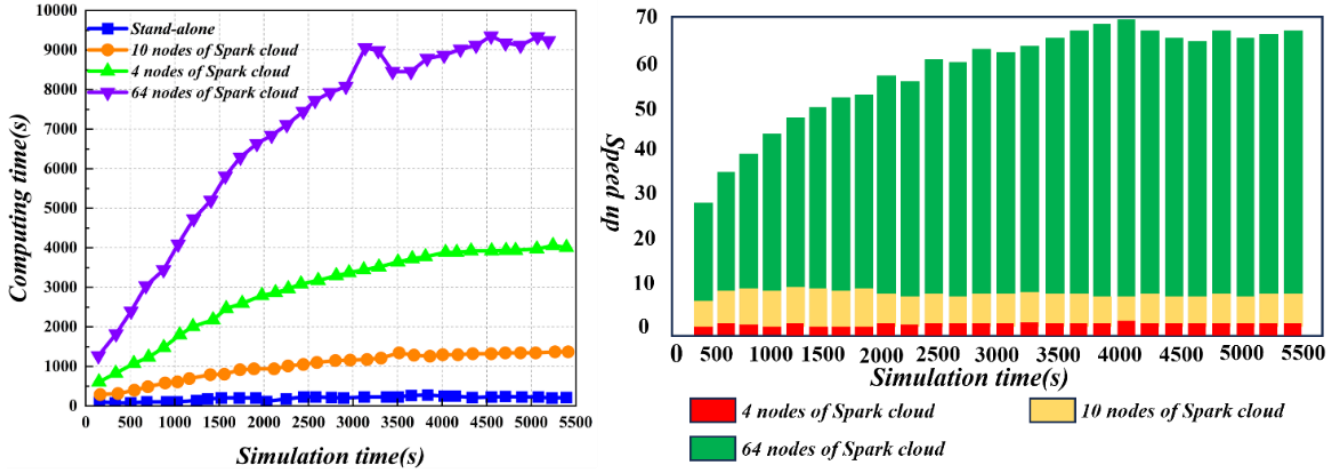

Fig. 7. Computational efficiency of predicting time-domain optimization for each control cycle.

On the contrary, on the Spark cloud, multiple chromosomes are divided into several subpopulations and distributed to various worker nodes. The more nodes there are, the fewer chromosomes are assigned to each worker node, resulting in a more minor computational task. The chromosomes on all worker nodes call the traffic network flow prediction model to obtain the control effect evaluation values of chromosomes in a parallel manner, thereby reducing optimization time. As shown in Fig. 7, at the beginning of the simulation, the acceleration ratio of parallel optimization based on Spark cloud is lower than that of single-machine serial optimization. With the deepening of the simulation, the efficiency of the late parallel computation is improved significantly and remains stable. This is because Spark initially takes some time to load the data. However, after data is cached to the memory, Spark can directly obtain data from the memory when it is invoked again, which improves computing efficiency.

From Fig. 8, it can be observed that the prediction results of the nonparametric regression algorithm are better than those of the BP neural network method. However, the improved algorithm in this paper has a particular improvement in prediction accuracy compared to the basic nonparametric regression algorithm, and the absolute error of the prediction results is generally below 6km/h, which can better reflect the future traffic speed situation of the road segment. In summary, it can be seen that the improved prediction method in this paper shows good prediction ability in the overall traffic speed fitting and specific prediction results. Compared with the general algorithm, this method has obvious improvement in prediction accuracy.

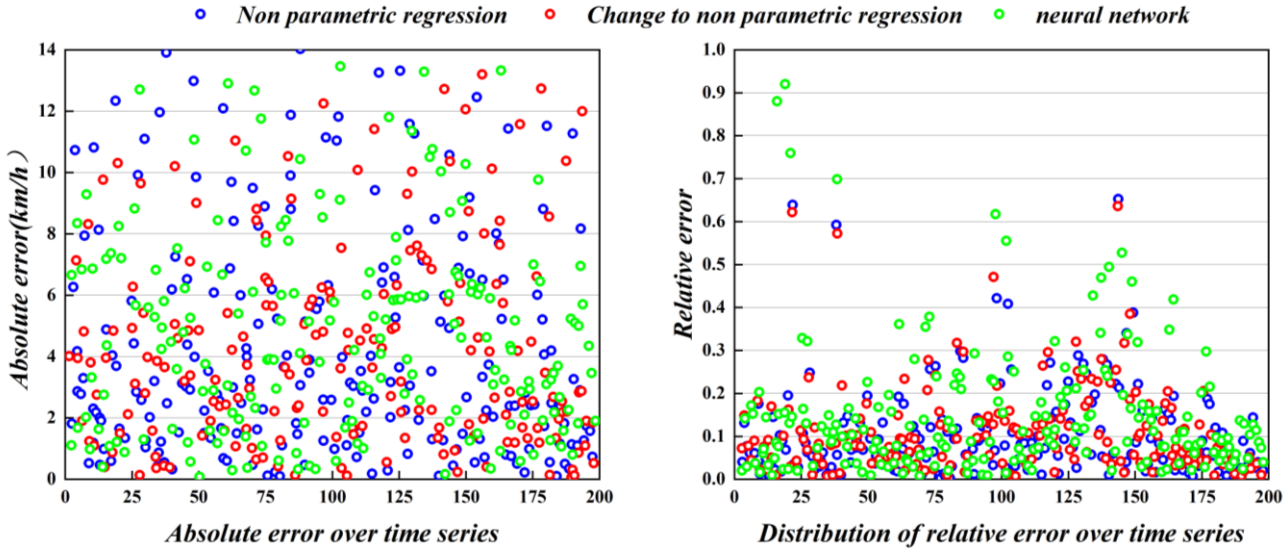

Fig. 8. Error distribution over time series.

V. CONCLUSIONS

This article conducts in-depth research on short-term traffic flow prediction and constructs a prediction model based on the Internet of Things and deep learning algorithms. Through analysis of actual traffic data and model validation, we have drawn the following conclusions:

Many real-time traffic data, including vehicle sensors, traffic cameras, and traffic lights, has been obtained using

Internet of Things technology. These data provide comprehensive and accurate traffic status information, providing an essential foundation for short-term traffic flow prediction. By comparing the experimental results, it was found that deep learning algorithms have better performance and accuracy in traffic flow prediction tasks compared to traditional machine learning algorithms.

The prediction model proposed by this research institute based on the Internet of Things and deep learning algorithms provides strong support for actual traffic management and decision-making. This model can help traffic management departments better plan road resources, optimize traffic signal timing, and provide real-time traffic congestion information to drivers and traffic participants to improve traffic efficiency and reduce traffic congestion.

ACKNOWLEDGMENT

This study combines the Internet of Things technology and deep learning algorithm to build a short-term traffic flow prediction model, which effectively solves the problem of traffic congestion and frequent accidents. However, future research needs to further explore the combination of 5G communication technology, advanced edge computing technology and existing models to achieve more efficient data processing and transmission, improve the model's faster response to real-time data, improve the accuracy and real-time forecasting, and provide strong support for more intelligent and efficient traffic management and urban operation.

FUNDING

This work was supported by Science and Technology Plan Project of Zhejiang Provincial Department of Transportation: Research and System Development with Road Asset Digital Technology and System Development based on highly accurate map (No. 202203).

REFERENCES

[1] Xia M. Traffic congestion index calculation based on BP neural network[J]. Advances in Computer, Signals and Systems, 2021, 5(1).

[2] Hassija V, Gupta V ,Garg S , et al. Traffic Jam Probability Estimation Based on Blockchain and Deep Neural Networks. IEEE Transactions on Intelligent Transportation Systems, 2020, PP(99).

[3] Yaqin Y, Yue X, Yuxuan Z, et al. Dynamic multi-graph neural network for traffic flow prediction incorporating traffic accidents. Expert Systems With Applications, 2023, 234.

[4] Xijun Z ,Jiwen L . Traffic flow prediction based on GRU-BP combined neural network. Journal of Physics: Conference Series, 2021, 1873(1).

[5] Guo Y, Lu L. Application of a Traffic Flow Prediction Model Based on Neural Network in Intelligent Vehicle Management. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(3).

[6] Yi L, Mingsheng L,Yunchi X, et al. Traffic flow prediction model based on gated graph convolution with attention. Journal of Physics: Conference Series, 2023, 2493(1).

[7] Ameya A K, Shravan R, Ananya D , et al. Traffic flow prediction models – A review of deep learning techniques. Cogent Engineering, 2022, 9(1).

[8] Oreja M J, Gozalvez J. A Comprehensive Evaluation of Deep Learning-Based Techniques for Traffic Prediction. IEEE Access, 2020, 8.

[9] Yang L, Yaolun S,Yan Z, et al. WT-2DCNN: A convolutional neural network traffic flow prediction model based on wavelet reconstruction. Physica A: Statistical Mechanics and its Applications, 2022, 603.

[10] Ismaeel G A, Janardhanan K ,Sankar M , et al. Traffic Pattern Classification in Smart Cities Using Deep Recurrent Neural Network. Sustainability, 2023, 15(19).

[11] Qianqian Z, Nan C, Siwei L. FASTNN: A Deep Learning Approach for Traffic Flow Prediction Considering Spatiotemporal Features. Sensors, 2022, 22(18).

[12] Yuanmeng Z, Jie C, Hong Z, et al. A deep learning traffic flow prediction framework based on multi-channel graph convolution. Transportation Planning and Technology, 2021, 44(8).

[13] Zhao Z, Hao Y, Xianfeng Y. A Transfer Learning–Based LSTM for Traffic Flow Prediction with Missing Data. Journal of Transportation Engineering, Part A: Systems, 2023, 149(10).

[14] Bernardo G, José C, Helena A. A survey on traffic flow prediction and classification. Intelligent Systems with Applications, 2023, 20.

[15] Chen C ,Ziye L ,Shaohua W , et al. Traffic Flow Prediction Based on Deep Learning in Internet of Vehicles. IEEE Transactions On Intelligent Transportation Systems, 2021, 22(6).

[16] Jiang L, Luofeng J. Traffic Flow Prediction Method Based on Deep Learning. Journal of Physics: Conference Series, 2020, 1646 (1).

[17] Hong Z, Sunan K ,XiJun Z , et al. Dynamic Spatial–Temporal Convolutional Networks for Traffic Flow Forecasting. Transportation Research Record, 2023, 2677(9).

[18] Emerging Technologies; Research Conducted at Beijing Institute of Technology Has Updated Our Knowledge about Emerging Technologies (A hybrid deep learning based traffic flow prediction method and its understanding). Computers, Networks & Communications, 2018.

[19] Wu Y, Tan H, Qin L, et al. A hybrid deep learning based traffic flow prediction method and its understanding. Transportation Research Part C, 2018, 90.

[20] Yingya G, Yufei P, Run H, et al. Capturing spatial–temporal correlations with Attention based Graph Convolutional Network for network traffic prediction. Journal of Network and Computer Applications, 2023, 220.

[21] Siyuan F, Shuqing W, Junbo Z, et al. A macro–micro spatio-temporal neural network for traffic prediction. Transportation Research Part C, 2023, 156.

[22] Bo W, L. H V, Inhi K, et al. Distributional prediction of short-term traffic using neural networks. Engineering Applications of Artificial Intelligence, 2023, 126(PC).

[23] Rui H, Cuijuan Z, Yunpeng X, et al. Deep spatio-temporal 3D dilated dense neural network for traffic flow prediction. Expert Systems with Applications, 2024, 237(PA).

[24] Xian Y, Yin-Xin B, Quan S. STHSGCN: Spatial-temporal heterogeneous and synchronous graph convolution network for traffic flow prediction. Heliyon, 2023, 9(9).

[25] Robert J ,Young T K ,Emily G , et al. Tailoring Mission Effectiveness and Efficiency of a Ground Vehicle Using Exergy-Based Model Predictive Control (MPC). Energies, 2021, 14(19).

[26] Artificial Neural Network; Findings on Artificial Neural Network Discussed by Investigators at Ryerson University. Internet Networks & Communications, 2017.

[27] Sunday S O. The Application of Model Predictive Control (MPC) to Fast Systems such as Autonomous Ground Vehicles (AGV). IOSR Journal of Computer Engineering, 2014, 16(3).

[28] Yaqin Y, Yue X, Yuxuan Z, et al. Dynamic multi-graph neural network for traffic flow prediction incorporating traffic accidents. Expert Systems with Applications,2023,234.

[29] Yi X, Liangzhe H, Tongyu Z, et al. Generic Dynamic Graph Convolutional Network for traffic flow forecasting. Information Fusion, 2023, 100.

[30] Dongran Z, Jun L. Multi-view fusion neural network for traffic demand prediction. Information Sciences,2023,646.

[31] Xiaoxiao S, Xinfeng W, Boyi H, et al. Multidirectional short-term traffic volume prediction based on spatiotemporal networks. Applied Intelligence, 2023, 53(20).