Traffic Flow Prediction Methods Based on Deep Learning

Haoxuan Yang[1,3,*], Yang Zhang[2]

[1]School of Intelligent Transportation Engineering, Beijing Jiaotong University, Beijing, China
[2]School of Computer Science, Nanchang University College of Science and Technology, Nanchang, China

[3]23251024@bjtu.edu.cn
*corresponding author

Abstract. In recent years, due to the rapid development of deep learning, it has been found that deep learning has shown good experimental results in traffic flow prediction, and is significantly better than traditional traffic flow prediction methods. This paper mainly classifies, sorts out, and summarizes the classic techniques and research status of traffic flow prediction based on deep learning. According to the length of traffic flow data, the traffic flow prediction method based on deep learning is further subdivided into long traffic flow prediction method and short traffic flow prediction method. The representative models of the two traffic flow prediction methods are analyzed and introduced in detail. Each representative model is expanded and extended, and the performance evaluation indicators of the long and short traffic flow prediction models are introduced. Finally, the possible future research directions and corresponding development trends in this field are summarized.

Keywords: Traffic flow prediction, deep learning, long traffic flow, short traffic flow.

1. Introduction

Traffic flow prediction refers to the utilization of various data analysis techniques and models to forecast road traffic conditions at a specific future time point or within a given period. It is pivotal in intelligent transportation systems, urban planning, traffic management, and public travel arrangements. Accurate traffic flow predictions empower traffic managers to conduct effective traffic guidance and engineering designs, thereby mitigating congestion and enhancing road utilization efficiency. Simultaneously, they assist individuals and enterprises in making more rational travel and logistics plans, saving time and costs. Based on the prediction object, traffic flow prediction can be categorized into short-term and long-term predictions.

Short-term traffic flow prediction typically involves forecasting traffic volume, speed, occupancy rate, and other parameters within the next few hours to a day. This type of prediction guides real-time or near-real-time traffic management and travelers, enabling them to make more informed decisions. Short-term prediction models need to swiftly respond to changes in traffic conditions, such as accidents, construction works, or special events causing congestion. Common methodologies include time series analysis, machine learning algorithms (e.g., Support Vector Machines, Random Forests, Neural

Networks), and deep learning approaches like Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs).

In contrast, long-term traffic flow prediction focuses on traffic trends and patterns over the next few days, weeks, or even months. This prediction is crucial for traffic planning, infrastructure investment decisions, large-scale event planning, and strategic traffic management. Long-term prediction faces challenges like capturing periodic patterns (e.g., seasonal variations, holiday effects) and trends. These models must consider additional external factors like seasonal changes, economic growth, and population migration. Traditional statistical models (e.g., ARIMA), machine learning methods, and deep learning models like Convolutional LSTM (ConvLSTM) and Spatio-Temporal Convolutional Neural Networks (STCNN) are commonly employed.

The current literature on the prediction of long and short traffic flows is not comprehensive. This paper summarizes deep learning-based traffic flow prediction methods, including representative models for both long-term and short-term predictions, and performance evaluation metrics, and explores future research directions and development trends.

2. Analysis of long traffic flow methods

For long-term traffic flow prediction, the primary issue lies in the accumulation of errors over time. Many predictive models rely on iterative methods, where each predicted value is fed back as new input. As the prediction window slides, the proportion of real data decreases while predicted data increases, leading to error amplification. To address this, the ACE-D model [1] has been proposed (Figure 1).

The ACE-D model is a deep-learning architecture tailored for long-term traffic flow prediction. It employs an encoder-decoder structure built upon LSTM, where the encoder encodes the input traffic flow data sequence into semantic vectors, and the decoder translates these vectors into future traffic flow predictions. To enhance the model's ability to capture long-term dependencies, ACE-D incorporates a hard attention mechanism. This mechanism leverages a chain address search algorithm to swiftly identify patterns in historical data similar to the current prediction window. The search considers both periodic features and a forward search strategy to ensure useful information retrieval even when periodic features are obscure. Upon finding similar patterns, the model learns from their characteristics to assist predictions, strengthening its memory of long-term trends.
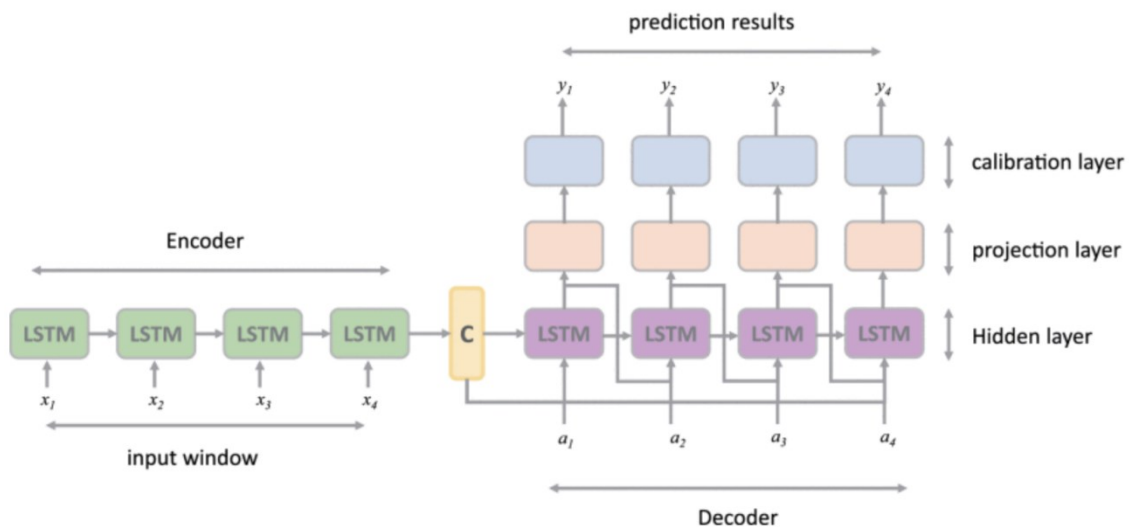
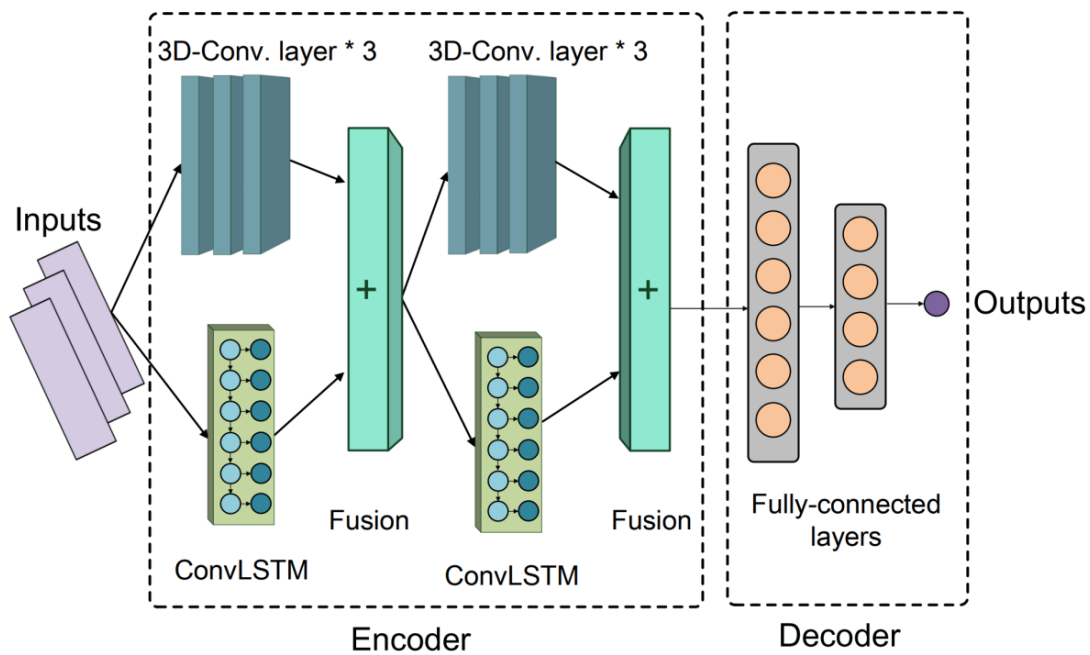**Figure 1.** Schematic diagram of the ACE-D model [1]

Furthermore, to refine and correct predictions, ACE-D adds a calibration layer to the decoder's output. This layer utilizes a control gate mechanism to categorize traffic flow data based on their magnitudes.

For low-magnitude (i.e., stable) data, the model directly outputs predictions, while for high-magnitude (i.e., volatile) data, the calibration layer adjusts them to enhance prediction detail accuracy.

By combining hard attention and a calibration layer, the ACE-D model effectively mitigates error accumulation in long-term predictions while maintaining sensitivity to traffic flow trends, demonstrating high accuracy and stability in long-term traffic flow prediction tasks.

However, it has potential shortcomings in terms of computational complexity and parameter sensitivity. The D-STN model [2], on the other hand, may provide improvements in these areas through its simplified architecture and training mechanism. (Figure 2)

The D-STN method is specifically designed to optimize the accuracy of long-term mobile traffic flow predictions by integrating two powerful deep learning architectures: ConvLSTM and 3D-ConvNet. It excels at capturing and analyzing complex spatio-temporal patterns in traffic data while demonstrating significant advantages in terms of computational complexity and parameter sensitivity.
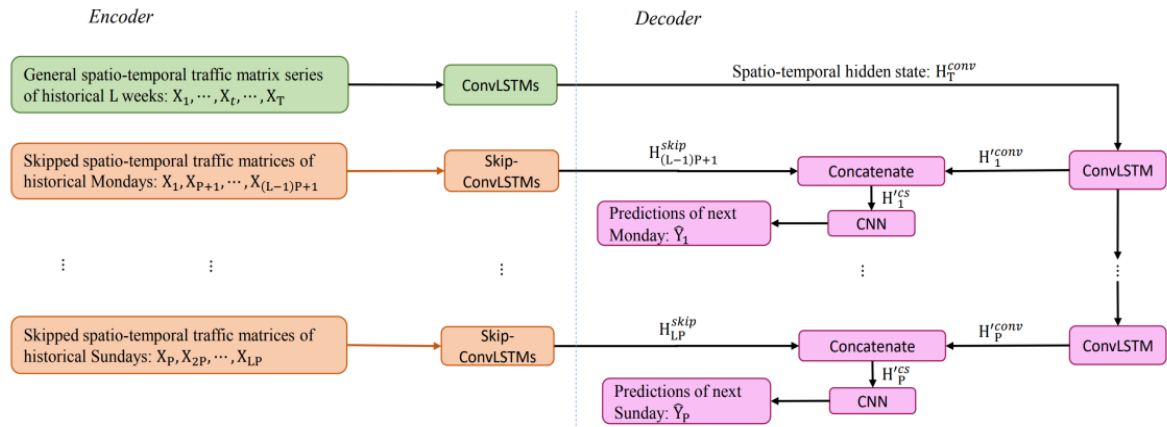

Figure 2. Schematic diagram of the D-STN model [2]

To address the common issue of cumulative errors in long-term predictions, D-STN introduces the Ouroboros Training Scheme (OTS), an innovative training mechanism that feeds the model's predictions back into the input during training, enabling continuous self-correction and enhancing its ability to predict long-term trends. Furthermore, D-STN employs a decay mechanism that gradually reduces the weight of model predictions and increases the weight of historical data averages, thereby mitigating prediction errors over time.

A notable feature of D-STN is its exceptional generalization ability, allowing the model to not only perform effectively on training data but also adapt to different geographical locations and traffic conditions, crucial for network management and resource allocation in practical applications. Overall, D-STN significantly improves the accuracy of long-term mobile traffic flow predictions through its advanced deep learning architecture and innovative training strategies, providing a powerful tool for traffic engineering and network optimization.

Although D-STN has advantages in terms of prediction accuracy and stability, its complex model structure can lead to reduced efficiency when dealing with complex traffic flow patterns. In contrast, the STCNN model [3] has stronger spatio-temporal capturing capabilities compared to D-STN (Figure 3).

Specifically, the STCNN method is an innovative spatio-temporal convolutional neural network architecture tailored for long-term traffic flow prediction. It adopts a classic encoder-decoder structure: During the encoder phase, STCNN first processes spatio-temporal flow data spanning consecutive days using ConvLSTM networks. This data encompasses traffic flow information from various locations within the network, capturing dynamic characteristics that evolve and transforming them into spatio-temporal hidden states, representing the general spatio-temporal dependencies of traffic flow. To further explore periodic traffic flow patterns, STCNN designs a Skip-ConvLSTM model that specifically processes skipped time series data (e.g. the same day of each week) to extract periodic flow patterns. These patterns are crucial for understanding and predicting long-term traffic flow changes.

Figure 3. Schematic diagram of the STCNN model [3]

During the decoder phase, STCNN employs another ConvLSTM network that decodes the spatiotemporal hidden states generated by the encoder, progressively generating future spatiotemporal hidden states for the upcoming week. These hidden states not only reflect the general spatio-temporal characteristics of traffic flow but also integrate the periodic flow patterns captured during the encoder phase. Finally, a Convolutional Neural Network (CNN) is utilized for the ultimate flow prediction. In this way, STCNN considers both daily spatio-temporal variations and periodic factors, enabling more precise and long-term traffic flow predictions. Experimental results demonstrate that the STCNN method outperforms other traditional and advanced prediction models, exhibiting superior prediction performance across multiple datasets.

3. Analysis of short traffic flow methods

3.1. Short-distance traffic flow prediction based on CNN

These methods typically transform traffic flow data into a time series input, extract local and global features through convolutional layers, and use these features to predict future traffic flow. The model architecture includes an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer. During training, the mean squared error is used as the loss function, and parameters are adjusted via an optimizer.

Zhang, Zhu, and others [4] proposed a spatio-temporal feature selection algorithm (STFSA) based on CNN. The STFSA algorithm selects the optimal spatio-temporal data as input for the CNN model by analyzing the spatio-temporal correlations in traffic data. It first converts the selected spatio-temporal data into a two-dimensional matrix and feeds it into the CNN model. The CNN then extracts and learns features from the input data to complete the traffic flow prediction. However, the prediction performance of the CNN model is highly dependent on parameter selection, and the algorithm does not account for other influential factors. It focuses solely on spatio-temporal characteristics, ignoring key factors such

as weather and road conditions. Consequently, the trained model is very sensitive to abnormal data and easily affected by noise.

Later, Sun, Wu, and others [5] introduced a multi-branch model called Traffic Flow Forecasting Network (TFFNet) (as shown in Figure 4), which combines deep convolutional neural networks (CNNs) and residual networks (ResNets). TFFNet also incorporates a novel data preprocessing method to calculate city-wide and suburban traffic flow based on taxi GPS trajectories. Additionally, TFFNet considers both temporal and spatial dependencies and integrates external influencing factors, thereby addressing the issue of the trained model being susceptible to noisy data and improving prediction accuracy.
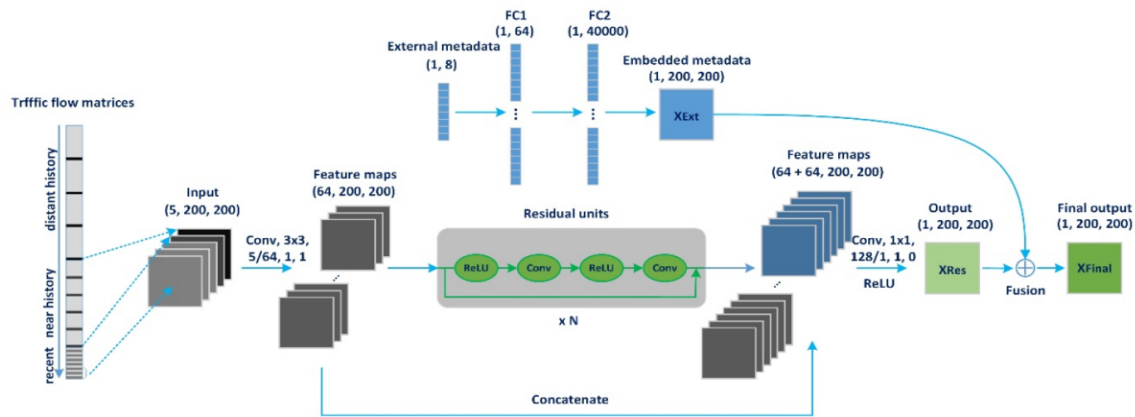
Figure 4. TFFNet architecture. Conv: convolution layer; FC: fully connected layer [5].

Chen, Yang, and others [6] proposed a prediction model with high prediction accuracy and low computational complexity (Spatio-temporal Decoupled 3D DenseNet with Attention ResNet, abbreviated as ST-D3DDARN). The ST-D3DDARN model established four branches, namely proximity, cycle, trend, and external factors, and adopted a decoupled 3D convolutional neural network (3D CNN), which solved the problems of long training time and slow convergence speed of traditional 3D CNN models. Finally, to deal with the complex spatial correlation of urban traffic flow, the model designed a residual network that combines spatial self-attention and coordinate attention mechanisms.

3.2. Short-term traffic flow prediction based on RNN

LSTM is a special type of RNN. The model controls the transmission and forgetting of information by introducing "memory units" and "gating mechanisms" (including input gates, forget gates, and output gates), and can effectively capture long-term dependencies. First, the historical traffic data is converted into a time series and input into the model. Then, the long-term and short-term dependencies in the data are captured through the LSTM layer, and the prediction results are output using the fully connected layer. During the training process, the mean square error is used as the loss function, and the model parameters are adjusted through the optimizer. Finally, the model is evaluated and optimized on the test data.

Li, Zhang, and others [7] compared RNN with three different RNN structural units - long short-term memory network (LSTM), gated recurrent unit (GRU), and autoencoders (SAEs). First, the data was preprocessed and normalized. Then the data was put into different models for model construction and training. The results showed that RNN had a good effect on traffic flow prediction for specific roads and could achieve short-term prediction. Among them, LSTM had the smallest prediction error and performed best, so it can be concluded that LSTM is the best model among RNN variants.

Subsequently, Zhang, Cheng, and others [8] proposed an improved genetic algorithm-optimized LSTM model (IGA-LSTM) (as shown in Figure 5), which improves the optimization effect by improving the standard genetic algorithm and dynamically adjusting the mutation rate and crossover

rate. The improved genetic algorithm (IGA) is used to optimize multiple parameters in the LSTM model (such as the number of hidden units, the number of training rounds, etc.), making the IGA-LSTM model outperform the traditional LSTM model in terms of convergence speed and prediction accuracy.
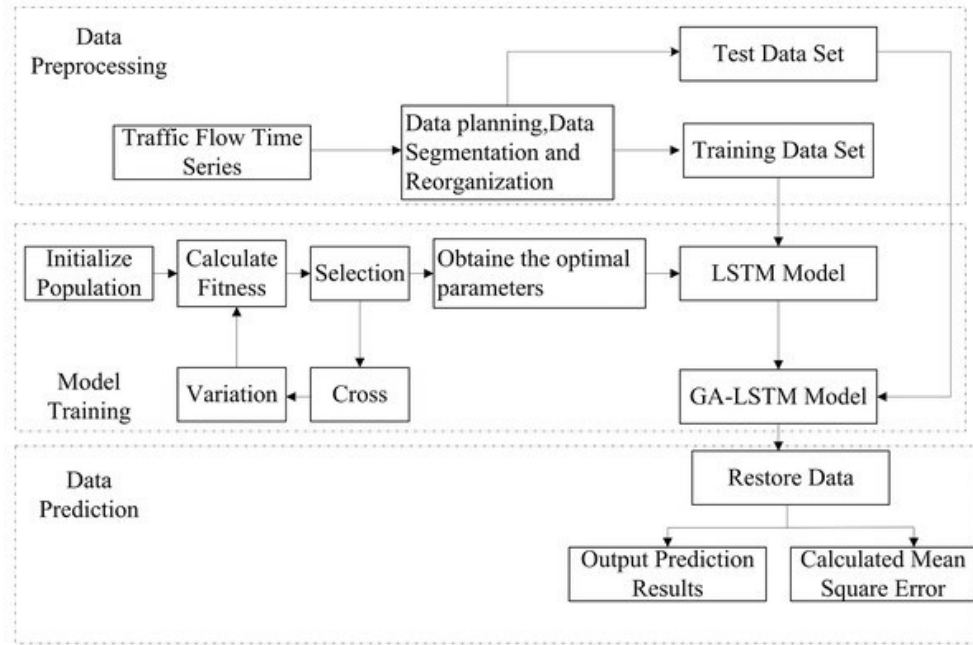
Figure 5. Framework diagram of traffic flow forecast by the IGA-LSTM model [8].

Zhang, Jiang, and others [9] proposed a graph convolution long short-term memory network (AGC-LSTM) model enhanced by the attention mechanism, which captures the spatial dependence between flight segments by using a graph convolution network (GCN), and then uses an LSTM layer to capture the dependencies between times. The attention mechanism is introduced into the GCN to help the model focus on some key nodes in the route network in each layer. Moreover, the model also performs multi-step predictions to ensure the robustness of the model in different prediction periods. This model significantly improves the accuracy of short-term traffic flow prediction by combining the spatial feature extraction capabilities of GCN, the time series processing capabilities of LSTM, and the key feature focusing of the attention mechanism.

3.3. Hybrid traffic flow prediction based on CNN and LSTM
The hybrid model effectively combines the advantages of CNN and LSTM to achieve traffic flow prediction. Initially, the model inputs historical traffic data—such as traffic flow, speed, and density—into a CNN layer to extract the spatial relationships among various traffic parameters. The spatial features extracted are then passed to the LSTM network to capture temporal dependencies across different time slices, allowing for a comprehensive analysis of traffic flow.

In this regard, Pervez Khan and colleagues [10] proposed a more complex hybrid architecture that integrates CNN, LSTM, and Transposed CNN. This model starts with a traffic congestion map sourced from the Seoul Transportation Operation and Information Service (TOPIS). Compared to traditional models, it extracts spatial features through CNN first, followed by using LSTM to analyze historical traffic flow relationships. Finally, the processed features are fed into the Transposed CNN, which transforms low-dimensional time series features into high-resolution traffic congestion images for prediction. This multi-layered processing enables a more precise prediction of traffic conditions. However, due to the integration of multiple networks, this model has a high computational complexity and often deals with a significant amount of irrelevant information during the learning process, leading to increased consumption of computational resources.

To further optimize this process, Liu, Ma, and colleagues [11] introduced the CCSANet model(as shown in Figure 6), which combines the Correlation Convolutional Long Short-Term Memory Network (Corr-ConvLSTM) with a self-attention mechanism. First, the Corr-ConvLSTM model calculates the correlations between continuous traffic features, enhancing the ability of ConvLSTM to capture continuous spatial characteristics. The ConvLSTM then simultaneously extracts spatial and temporal features, better reflecting the spatiotemporal relationships. Finally, by incorporating the self-attention mechanism, the model can take into account external factors influencing network traffic. This series of improvements enables CCSANet to simultaneously capture complex spatiotemporal features, ultimately enhancing the accuracy of network traffic prediction.
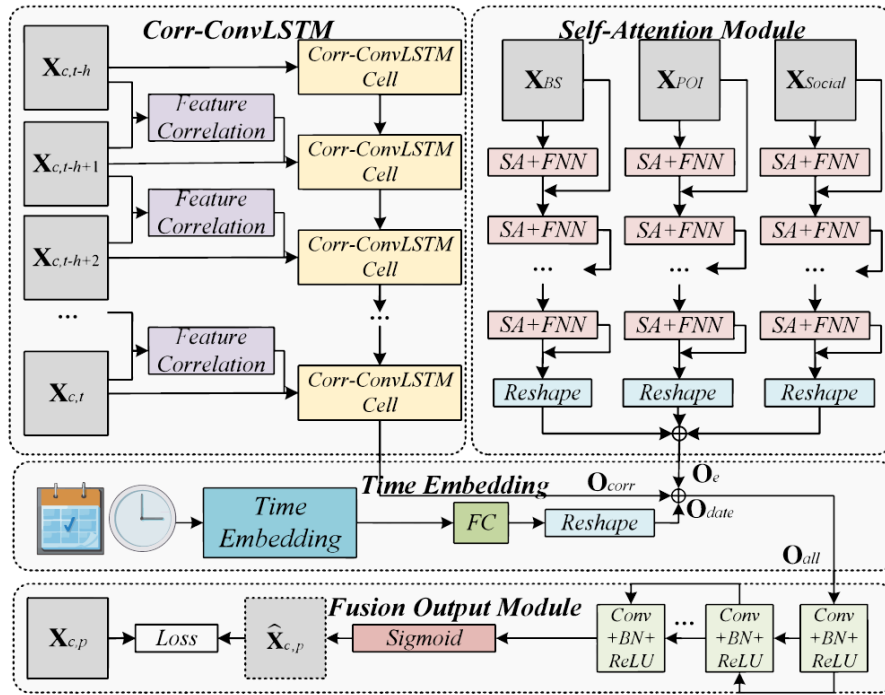
Figure 6. Overview of the proposed CCSANet [11].

Ali Reza Sattarzadeh and others [12] proposed a new hybrid deep learning model (SAACL) that combines the ARIMA model, convolutional long short-term memory network (Conv-LSTM), and shuffle attention mechanism (Shuffle Attention Layer). The paper introduced a new attention mechanism, Shuffle Attention, which assigns different weights to different traffic flow sequences, so that the model can automatically identify the importance of each period and location to the overall prediction, thereby improving the prediction performance. The ARIMA model and the Bi-LSTM model were also introduced. The ARIMA model is used to extract linear features in traffic flow data, and the Bi-LSTM model is used to capture periodic features in traffic flow. By combining these models, this new hybrid deep learning model can better capture the complex features in traffic flow data and improve the accuracy of prediction.

4. Method performance comparison

4.1. Analysis of methods based on long-distance traffic flow prediction

Table 1. Advantages and disadvantages of long-distance traffic flow prediction methods

Model	Advantages	Disadvantages
D-STN	High accuracy and stability in long-term forecasting Combines STN predictions with historical statistic Suitable for different regions and conditions	Complex model structure May require substantial computational resources
STCNN	Strong spatiotemporal capturing ability Specifically designed for traffic flow data Suitable for real-time or near-real-time forecasting	May require adjustment to adapt to new traffic pattern Might be slightly inferior to D-STN in long-term forecasting
ACE-D	Accuracy and stability Uses attention mechanism to enhance forecasting accuracy Suitable for long-term sequence forecasting	May be sensitive to parameter adjustments May be less efficient when dealing with very long sequences

D-STN excels at long-term traffic forecasting with its complex model that combines predictions with historical data, although it may be computationally intensive (Table 1). STCNN is superior for real-time traffic analysis by capturing spatiotemporal relationships but might need tuning for new traffic patterns and may not match D-STN for the longest forecasts. ACE-D balances accuracy and stability using an attention mechanism to focus on historical traffic similarities, yet it could be inefficient for very long sequences and sensitive to parameter adjustments. Collectively, these models can form a robust predictive system, with D-STN offering long-term views, STCNN providing real-time insights, and ACE-D ensuring stability across different prediction horizons.

4.2. Analysis of methods based on short-distance traffic flow prediction

The experimental comparison of short traffic flow prediction methods is shown in Table 2. The main comparison factors include experimental data sets and evaluation indicators.

Table 2. Experimental comparison of short traffic flow prediction on Caltrans PeMS

Model	Time step:5 min			Time step:15 min			Time step:30 min			Time step:60 min		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
CNN	15.14	11.30%	21.93	16.02	12.96%	22.14	17.12	14.60%	23.45	20.67	15.80%	27.04
LSTM	14.77	11.10%	20.05	16.50	12.40%	22.57	19.00	14.70%	25.59	23.60	18.30%	30.75
DNN-BTF	14.05	10.90%	19.32	15.55	11.50%	21.37	16.97	12.80%	23.06	19.12	14.80%	25.88
DCRNN	13.79	10.70%	18.88	14.79	11.50%	20.43	16.05	12.40%	22.18	18.43	14.20%	25.74
SAACL	13.36	9.60%	17.86	14.74	10.60%	19.84	15.26	11.90%	20.82	16.38	12.70%	

As can be seen from Table 2, the prediction results of the five models at different times are compared, and the mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE) of each model in each period are shown. The experimental data comes from the California Department of Transportation's (Caltrans) Performance Measurement System (PeMS). Among the methods for short-term traffic flow prediction, the prediction results of the CNN model are not as good as those of other models in each period, but because it was proposed earlier and many models are improved based on CNN, it is widely used in traffic flow prediction models. Although the DCRNN model performs well in short-term predictions, it slightly decreases in long-term predictions. The LSTM model is less effective in dealing with long-term time dependencies. The SAACL model shows the best

prediction performance at all time steps, with the highest prediction accuracy and stability, but because it integrates multiple models, the model structure is complex, the demand for computing resources is high, and the training time of the model is long.

5. Future development trends

Nowadays, deep learning technology has been widely used in the field of traffic flow prediction. However, there are still some problems with traffic flow prediction, such as the complexity of spatiotemporal feature modeling, the difficulty of multimodal data fusion, adversarial training robustness enhancement, etc.

5.1. Long-Term traffic flow prediction future development trends

Enhanced Model Robustness Improving model resilience against noise and missing data is crucial. This can be achieved through techniques like data augmentation, anomaly detection, attention mechanisms, and adversarial training to ensure accurate predictions even with imperfect data.

Lightweight and Real-Time Models With the advancement of 5G and 6G networks, there is a need for models that are not only accurate but also lightweight and capable of real-time performance. Research may focus on developing new network architectures, applying knowledge distillation, or employing model pruning to reduce computational complexity and resource requirements.

Integration of Multimodal Data Future models may integrate diverse data sources such as social media, weather data, and economic indicators to enhance prediction accuracy by capturing a broader range of influencing factors.

5.2. Short-Term traffic flow prediction future development trends

Enhance the robustness of the model In the real world, traffic flow data is often affected by noise or missing data. Improving the robustness of the model is an important development direction. In the future, we can study the use of data enhancement, anomaly detection, attention mechanism, adversarial training, and other technical methods to improve the prediction accuracy of the model when facing abnormal data.

Lightweight and real-time model With the rapid development of 5G and 6G networks, network traffic prediction models face the dual challenges of real-time performance and computational efficiency in applications. Therefore, how to design lightweight, computationally efficient, and highly accurate models is an important development direction in the future. In the future, new network architectures, knowledge distillation, or model migration methods can be used to reduce the computational complexity and storage requirements of the model.

6. Conclusion

This paper reviews deep learning-based traffic flow prediction methods, providing an in-depth analysis of classic models for both long-term and short-term traffic flow predictions, along with their respective advantages and disadvantages. In long-term forecasting, models such as D-STN, STCNN, and ACE-D stand out, each with its strengths: D-STN is accurate and stable for long-term predictions, STCNN is adept at real-time analysis and capturing spatio-temporal relationships, and ACE-D balances accuracy and stability through an attention mechanism. For short-term predictions, models like LSTM and CNN show varying performance at different time steps, with the SAACL model demonstrating the highest prediction accuracy and stability across all time steps. However, these models still face challenges such as data noise and computational efficiency in practical applications.

In the future, enhancing model robustness and developing lightweight, real-time models will be key directions for development. Technical approaches like data augmentation, anomaly detection, and attention mechanisms can improve model prediction accuracy in the face of abnormal data. Moreover, with the rapid development of 5G and 6G networks, designing models that are computationally efficient and capable of real-time performance will become increasingly important.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1]   Wang, Z., Sun, X., & Ding, Z. (2021). Long-Term Traffic Flow Forecasting Based on LSTM Encoder-Decoder Architecture. IEEE Transactions on Intelligent Transportation Systems, 22(10), 6561-6570. https://doi.org/10.1109/TITS.2020.2995546

[2]   Zhang, C., & Patras, P. (2018). Long-Term Mobile Traffic Forecasting Using Deep Spatio-Temporal Neural Networks. In Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc) (pp. 1-12). Los Angeles, CA, USA. https://doi.org/10.1145/3209582.3209606

[3]   He, H., & Zhang, J. (2019). STCNN: A Spatio-Temporal Convolutional Neural Network for Long-Term Traffic Prediction. In Proceedings of the 20th IEEE International Conference on Mobile Data Management (MDM) (pp. 226-235). Hong Kong. https://doi.org/10.1109/MDM.2019.0053

[4]   Zhang, W., Yu, Y., Qi, Y., Shu, F., & Wang, Y. (2019). Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning. Transportmetrica A: Transport Science, 15(2), 1688–1711. https://doi.org/10.1080/23249935.2019.1637966

[5]   Sun, S., Wu, H., & Xiang, L. (2020). City-Wide Traffic Flow Forecasting Using a Deep Convolutional Neural Network. Sensors, 20(2), 421. https://doi.org/10.3390/s20020421

[6]   Chen, J., Yang, G., Zhang, Z., et al. (2024). ST-D3DDARN: Urban traffic flow prediction based on spatio-temporal decoupled 3D DenseNet with attention ResNet. PloS one, 19(6), e0305424. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0305424

[7]   Li, Z., Li, C., Cui, X., & Zhang, Z. (2021). Short-term Traffic Flow Prediction Based on Recurrent Neural Network. 2021 International Conference on Computer Communication and Artificial Intelligence (CCAI) (pp. 81-85). Guangzhou, China. https://doi.org/ 10.1109/CCAI50917.2021.9447466

[8]   Zhang, J., Qu, S., Zhang, Z., et al. (2022). Improved genetic algorithm optimized LSTM model and its application in short-term traffic flow prediction. PeerJ Computer Science, 8, e1048. https://peerj.com/articles/cs-1048/

[9]   Zhang, Y., Xu, S., Zhang, L., et al. (2024). Short-term multi-step-ahead sector-based traffic flow prediction based on the attention-enhanced graph convolutional LSTM network (AGC-LSTM). Neural Computing and Applications. https://doi.org/10.1007/s00521-024-09827-3

[10]   Ranjan, N., Bhandari, S., Zhao, H. P., Kim, H., & Khan, P. (2020). City-Wide Traffic Congestion Prediction Based on CNN, LSTM, and Transpose CNN. IEEE Access, 8, 81606-81620. https://doi.org/10.1109/ACCESS.2020.2991462

[11]   Ma, X., Zheng, B., Jiang, G., & Liu, L. (2023). Cellular Network Traffic Prediction Based on Correlation ConvLSTM and Self-Attention Network. IEEE Communications Letters, 27(7), 1909-1912. https://doi.org/10.1109/LCOMM.2023.3275327

[12]   Sattarzadeh, A. R., Kutadinata, R. J., Pathirana, P. N., et al. (2023). A novel hybrid deep learning model with ARIMA Conv-LSTM networks and shuffle attention layer for short-term traffic flow prediction. Transportmetrica A: Transport Science, 1-23. https://www.tandfonline.com/doi/full/10.1080/23249935.2023.2236724