Transportmetrica A: Transport Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ttra21

A novel hybrid deep learning model with ARIMA Conv-LSTM networks and shuffle attention layer for short-term traffic flow prediction

Ali Reza Sattarzadeh, Ronny J. Kutadinata, Pubudu N. Pathirana & Van Thanh Huynh

To cite this article: Ali Reza Sattarzadeh, Ronny J. Kutadinata, Pubudu N. Pathirana & Van Thanh Huynh (2025) 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, 21:1, 2236724, DOI: 10.1080/23249935.2023.2236724

To link to this article: https://doi.org/10.1080/23249935.2023.2236724

Published online: 26 Jul 2023.

TRANSPORTMETRICA A: TRANSPORT SCIENCE
2025, VOL. 21, NO. 1, 2236724
https://doi.org/10.1080/23249935.2023.2236724

Taylor & Francis
Taylor & Francis Group

A novel hybrid deep learning model with ARIMA Conv-LSTM networks and shuffle attention layer for short-term traffic flow prediction

Ali Reza Sattarzadeh[a], Ronny J. Kutadinata[b], Pubudu N. Pathirana[a] and Van Thanh Huynh [a]

[a]School of Engineering, Deakin University, Waurn Ponds, Australia; [b]ARRB – National Transport Research Organisation, Port Melbourne, Australia

ABSTRACT

Traffic flow prediction requires learning of nonlinear spatio-temporal dynamics which becomes challenging due to its inherent nonlinearity and stochasticity. Addressing this shortfall, we propose a new hybrid deep learning model based on an attention mechanism that uses multi-layered hybrid architectures to extract spatial–temporal, nonlinear characteristics. Firstly, by designing the autoregressive integral moving average (ARIMA) model, trends and linear regression are extracted; then, integration of convolutional neural network (CNN) and long short-term memory (LSTM) networks leads to better understanding of the model's correlations, serving for more accurate traffic prediction. Secondly, we develop a shuffle attention-based (SA) Conv-LSTM module to determine significance of flow sequences by allocating various weights. Thirdly, to effectively analyse short-term temporal dependencies, we utilise bidirectional LSTM (Bi-LSTM) components to capture periodic features. Experimental results illustrate that our Shuffle Attention ARIMA Conv-LSTM (SAACL) model provides better prediction than other comparable methods, particularly for short-term forecasting, using PeMS datasets.

ARTICLE HISTORY
Received 7 June 2022
Accepted 7 July 2023

KEYWORDS
Intelligent transportation systems; traffic flow prediction; deep learning ARIMA-LSTM; Conv-LSTM; shuffle attention layers

1. Introduction

Urban traffic congestion is a widespread phenomenon that significantly influences many aspects of daily life (Rostami-Shahrbabaki et al. 2020). Extending road infrastructure to increase traffic capacity may not be a sustainable solution to the congestion problem. Intelligent Transportation System (ITS) and its advanced traffic control and management strategies can provide alternate solutions to alleviate congestion, reduce emission, lower fuel consumption, and improve safety (Rostami-Shahrbabaki et al. 2020). A critical component of the modern traffic management system is its ability in forecasting short-term traffic flow. A reliable estimation algorithm can dynamically predict urban traffic flow with

CONTACT Van Thanh Huynh v.huynh@deakin.edu.au School of Engineering, Deakin University, Waurn Ponds, VIC 3216, Australia

acceptable accuracy in the short and long term based on historical traffic data. However, predicting short-term traffic flow is problematic because of its inherent chaotic and dynamic characteristics.

Many existing traffic forecasting methods considered temporal correlations, including the Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA works on time-series data by capturing the standard temporal structures in the data. Hamed, Al-Masaeid, and Said (1995) customised and applied the ARIMA model to predict traffic flow, which also yielded acceptable results based on experiments. The ARIMA model and its time-series-based variations, such as Kohonen ARIMA (KARIMA) (Voort, Dougherty, and Watson 1996), subset ARIMA (Lee and Fambro 1999), seasonal ARIMA (SARIMA) (Williams and Hoel 2003), were confirmed to produce an acceptable prediction on stochastic traffic flow. However, with all their power in linear analysis, ARIMA-like time series models cannot handle nonlinearities in traffic flow due to a lack sufficient encapsulation of the dependencies between historical data. Besides ARIMA, other temporal dependence techniques that have been utilised in traffic flow prediction are parametric, like the Kalman filters (Okutani and Stephanedes 1984) and non-parametric, such as the support vector regression machine (Wu, Ho, and Lee 2004; Yao, Shao, and Gao 2006), the k-nearest neighbour model (Zhang, He, and Lu 2009), the Bayesian model (Shiliang, Changshui, and Guoqiang 2006), and the partial neural network model (Huang et al. 2014). Nevertheless, the temporal dependence approaches, while considering dynamic changes of traffic conditions, ignore spatial dependence assuming that road networks do not affect such dynamic changes.

One of the most effective traffic flow prediction methods is artificial neural networks (ANNs) that capture traffic flow uncertainties and complex nonlinearities (Zheng et al. 2021). As opposed to conventional ANN algorithms, deep learning models can extract many features and correlations from raw data due to their multilayer structure. Recent developments in deep learning have led to a considerable evolution in the traffic area, both in control and prediction. Deep learning algorithms have gained popularity in recent years and have been effectively suggested to predict traffic flow due to their inherent characteristics in their structure (Huang et al. 2014; Lv et al. 2014; Wu et al. 2018). Because of advances in processing capacity, deep learning models can be developed to capture more intricate traffic patterns. Although deep learning models predict the future traffic trend accurately, the majority of them can only be used for traffic networks as simple as a series of road segments; however, urban network structures are typically far more complex. Furthermore, present models attempt to capture spatiotemporal correlations in traffic, although these patterns change over time. A deeper knowledge of the spatial–temporal correlations in traffic networks would benefit traffic control policy and management (Yan, Ma, and Pu 2021). While deep learning approaches are highly effective in capturing nonlinearities and patterns in complex data, relying solely on deep learning models for traffic flow prediction may result in certain limitations such as incomplete extraction of complex features of traffic flow, and handling the spatiotemporal and periodic features simultaneously to extract underlying traffic characteristics. To address the aforementioned issues, we present a hybrid deep learning algorithm with the ARIMA model and a shuffle attention mechanism, resulting in an accurate time series prediction. The proposed solution incorporates the advantages of both strategies and efficiently captures both linear and nonlinear characteristics of traffic flow data. The experimental findings illustrate that the presented strategy outperforms single models in predictive performance.

The following are the key contributions of this research: (1) A hybrid technique for predicting short-term traffic flow based on ARIMA model and the Conv-LSTM network; (2) The proposed architecture can extract both the linear characteristics and the spatial–temporal features of the traffic flow data; (3) Assigning different attention levels of traffic flow sequence at various intervals by customising a shuffle attention mechanism for the Conv-LSTM component. The suggested technique can automatically identify influential times, locations, and the relevance of each flow sequence to the total prediction performance without auxiliary data; (4) Using a real-world dataset, we conduct experiments to evaluate the efficacy of the suggested strategy. Experiments and implementations show that the proposed combined model of deep learning and shuffle attention mechanism outperforms previously designed algorithms in short-term forecasting of urban traffic flow with greater accuracy.

The rest of this paper is structured as follows. The related work is presented in Section 2. In Section 3, we describe the methodology, and we explain our suggested model for predicting traffic flow. In Section 4, we do experimental tests and discussion on a real-world dataset and evaluate predictive accuracy with various current approaches. The paper is summarised in section 5.

2. Related work

This section reviews the most recent studies on short-term traffic flow prediction. There are mostly three types of short-term traffic flow prediction: parametric, non-parametric, and hybrid. Typically, parametric approaches are based on the assumption of specified functions for some independent or dependent variables. The linear time-series models can extract the trend and periodicity pattern from the data. Parametric methods infer linear features and patterns and do not derive spatiotemporal correlations. The autoregressive moving average (ARMA) and the Kalman filtering model (KF) are one of the most important parametric methods. The autoregressive integrated moving average (ARIMA) model is a prominent model-based approach (Lu et al. 2021). Parametric models have an equational basis and are widely applicable to achieving steady traffic flow. However, when traffic circumstances change dramatically, the models may reveal evident weaknesses, which can easily cause data loss and overfitting. The ARIMA model is used to forecast traffic flow on highways and urban expressways (Hamed, Al-Masaeid, and Said 1995; Levin and Tsao 1980). Different forms of the ARIMA model were proposed that somehow improve the prediction's accuracy. For instance, a KARIMA model was proposed to solve the non-linearity of traffic data by using the Kohonen network (Voort, Dougherty, and Watson 1996). ARIMAX model was proposed to rectify forecasting performance (Williams and Hoel 2003). One of the branches of the ARIMA model in which the prediction performance has been improved is the ARIMAX model, in which the inclusion of other explanatory variables provides the extraction of turning points. A Bayesian seasonal ARIMA was presented for traffic flow prediction by employing the Bayesian to increase prediction precision (Ghosh, Basu, and O'Mahony 2007). From the literature review, traditional approaches encountered serious challenges. First, when faced with a large volume of data, they had no ability to handle them and extract patterns (Miglani and Kumar 2019). They were primarily intended to predict specific locations, they could not be utilised to estimate the traffic of whole metropolitan areas (Xie et al. 2022). Second, in real traffic conditions, they were unable to deal with the

unexpected and dynamic changes in different traffic conditions (Yan, Ma, and Pu 2021). The time-series static models that were mentioned might not be able to capture the nonlinear and stochastic nature of the traffic flow well, at the same time producing larger prediction errors.

Non-parametric approaches do not have a fixed and formularised model. They have a strong learning ability in nonlinear and random data, and the trained model is dynamically able to cope with new conditions (Xing and Liu 2021). Researchers turn to nonlinear models when the amount of historical data increases thousands of times, and conventional methods lose their efficiency due to low accuracy and high computation load. Nonparametric models are capable of strong feature extraction and can modify mathematical models based on changes in the environment. The non-parametric model comprises the k-nearest neighbours algorithm (k-NN) which is a supervised learning method, and support vector regression (SVR) is the popular method in classification problems (Feng et al. 2019; Hong et al. 2011), and Artificial Neural Networks (ANNs) have provided significant results in recent years. Advanced non-parametric methods were proposed, which are based on machine learning algorithms (Xie et al. 2022). Compared to the parametric methods, they improved the prediction performance, but it was not enough. For instance, in densely populated urban areas, non-linear spatial–temporal and complex dependencies are not extracted by using these conventional non-parametric machine learning algorithms (Xie et al. 2022). One of the main reasons was the dramatic demand for computational powers resulted from the significant increase of data. Consequently, stronger algorithms were needed to be able to answer this demand and be able to handle new dynamic conditions automatically.

In the third category, hybrid approaches for traffic flow prediction have been presented by incorporating various strategies (Cetin and Comert 2006; Dimitriou, Tsekeris, and Stathopoulos 2008; Tan et al. 2009). In the above methods, there are shortcomings due to shallow structures and low perception of temporal and spatial characteristics (Lv et al. 2014). Zhang and Huang (2018) designed the traffic flow prediction algorithm based on Deep Belief Network (DBN) and genetic algorithm by applying different traffic flow patterns under various conditions. Yang, Dillon, and Chen (2017) developed a stack denoise autoencoder method to learn a hierarchical representation of urban traffic flow. Recently, deep learning has been successfully applied and studied in ITS to overcome urban traffic problems despite its complexities. The first deep learning model is a deep belief network utilising the multitask learning approach (Huang et al. 2014). A stacked autoencoder (SAE) model is a deep learning model, and its structure is made of deep layers to extract deep temporal and spatial data correlations (Lv et al. 2014). One well-known and practical deep learning method to extract the temporal features is the Long short-term memory (LSTM) method. The proposed LSTM network considers temporal correlations in traffic system via a two-dimensional network which is composed of many memory units (Zhao et al. 2017). Using a convolutional neural network (CNN), a deep learning architecture was presented to extract spatiotemporal traffic features for speed prediction in a large-scale traffic network (Ma et al. 2017). The CNN networks, due to the multi-layered structure, have been designed to learn the spatial characteristics and the recurrent neural network to discover the temporal characteristics of traffic flow data (Wu et al. 2018). The deep learning technique automatically extracts the intrinsic spatiotemporal correlations from raw data without preprocessing (Zheng et al. 2021). The Recurrent Neural Network-based (RNNs) algorithms like LSTM cannot extract spatial characteristics adaptively, and geographical information must

be separately encoded into the input. The CNN-based approaches do not directly represent temporal sequential interactions, which is considered a defect. Additionally, LSTMs, RNNs, and other algorithms lose sequence information when the input sequence is too long.

Several efforts have been made to integrate CNN and RNN structures to improve artificial intelligence powers. For instance, various systems have employed CNNs and RNNs to generate image/video descriptions. Vinyals et al. (2015) presented the visual attention mechanism for caption generation and the integrating of CNN and RNN layers have also had impressive results in visual activity recognition (Donahue et al. 2015), sentiment analysis (Wang et al. 2016), video classification (Wu et al. 2015), and 3D object classification (Socher et al. 2012). Traffic flow data has many characteristics in both time and space, like other machine learning applications. A deep learning framework, Spatiotemporal Graph Convolutional Networks (STGCN), was proposed to capture comprehensive spatiotemporal features of historical traffic data (Yu, Yin, and Zhu 2017). Li et al. (2017) solved the spatiotemporal forecasting problem by presenting a diffusion convolutional recurrent neural network (DCRNN) that extracts the spatiotemporal correlations. In this method, spatial features are captured with graph neural networks, while temporal dependencies are realised with the encoder-decoder architecture. Do et al. (2019) suggested a traffic flow prediction method that uses spatial and temporal attention layers to leverage spatial correlations between road sections and temporal correlations between time steps. An attention-based model was devised that automatically trains to determine the significance of previous traffic flow to extract temporal and spatial characteristics of historical data. To handle the nonlinearity characteristic of traffic flow, Zheng et al. (2021) was presented a hybrid deep learning Conv-LSTM model. It has been demonstrated that the suggested attention mechanism in this algorithm and Bi-LSTM module are also beneficial for the Conv-LSTM, which can improve prediction performance.

Although the latest deep learning prediction models offered some prediction advantages, there are still a number of outstanding issues that might hinder the efficacy of flow prediction under complex scenarios. First, tackling simultaneously the nonlinear, spatiotemporal characteristics as a whole as well as periodicity of lengthy data sequences is still a challenge. Second, information provided by traffic flow at different times or locations might not offer equal importance to the prediction performance. Prior works have already proposed hybrid models with different attention mechanisms to address the varying significance in different time segments of historic data towards learning and prediction. However, those proposed models, even with attention mechanism still unable to exploit at their best the various levels of importance and might not offer the best learning of the important part of data sequences. To address the gaps above, in this paper we offer a new structure that can tackle the linearity characteristics in the dataset based on the ARIMA model and dealing with the nonlinear spatiotemporal characteristics based on the combined CNN and LSTM layers to model spatial–temporal features and extract their correlations. The structure targets nonlinearity, spatio-temporal correlations, and periodicity, simultaneously. To address the second gap, in this paper we demonstrate that the proposed shuffle attention offers better and more accurate prediction. We carried out rigorous simulation with real world datasets and we demonstrate that the proposed architecture performs better than previous hybrid models as well as previous attention-based hybrid models.

In this paper, we propose a hybrid model to deal with nonlinearities, adverse spatiotemporal characteristics, stochasticity, via engaging a novel attention mechanism that

produces superior traffic flow prediction mainly due to significant improvements in feature extraction and generalisability.

3. Methodology

The traffic flow data containing correlations, and ARIMA is a classic and linear statistical approach for time series prediction, while the Conv-LSTM model extracts nonlinear correlations in a dataset. The proposed model combines the Conv-LSTM network and ARIMA model, each of which with its unique feature provides the ability to dynamically predict urban traffic flow. The ARIMA Conv-LSTM model inherits the nonlinear fitting superiority from the deep learning models and the high accuracy and power of the ARIMA model in the extraction of linearity and linear correlations. The proposed model in this research is realised for the short-term forecast of urban traffic flow. The training process is based on historical data aggregated by traffic sensors such as cameras and inductive loop detectors. We presented a novel hybrid algorithm that incorporates ARIMA, CNN layers, and bi-directional LSTM networks, capturing both spatiotemporal features and periodical features of traffic data with highly accurate prediction. This study also employs the Shuffle Attention (SA) layer to extract deep features. In the next section, we present more details on each component of our proposed traffic flow prediction model.

3.1. Proposed hybrid deep learning model for traffic flow prediction

Some shortcomings in different deep learning models lead to a weak network due to insufficient knowledge of correlations. For instance, deep learning single layer models cannot fully extract the complicated traffic flow features. For example, CNN usually extracts spatial features while LSTM is involved in extracting temporal characteristics. Previous hybrid deep learning methods work independently to extract traffic flow characteristics, including spatial, temporal, and periodic characteristics. Existing architectures, which are also complex, are not used in predicting traffic flow due to their independence. In this paper, the suggested model includes an ARIMA, Conv-LSTM component, and two separate bi-directional LSTM layers to capture daily, weekly, and periodic characteristics through which the model can learn and extract new traffic patterns. Figure 1 illustrates the overall architecture of the proposed model. The ARIMA model extracts the linear regression feature of traffic data because in a short time historical data, for example 5 min, the number of passing vehicles is not much different, and a linear model like ARIMA can predict with high accuracy linearity that is intrinsic in the data. As can be seen from the overall architecture, the traffic forecasting model is designed from several blocks. The blue block in Figure 1 includes deep learning algorithms and attention layers. These deep learning algorithms and attention layers are further detailed in Figure 2. A Conv-LSTM model combines convolutional layers (CNN) with LSTM networks, as described in detail in Section 3.3. First, the CNN part of the model processes the data and extracts the spatial features of the traffic network. After the initial processing with the CNN, the resulting one-dimensional output is then forwarded to an LSTM model to effectively capture the temporal correlations present in the traffic data. Another way to look at the Conv-LSTM is that it is an LSTM variation that includes a convolution operation within the LSTM cell. This model has a high prediction accuracy compared

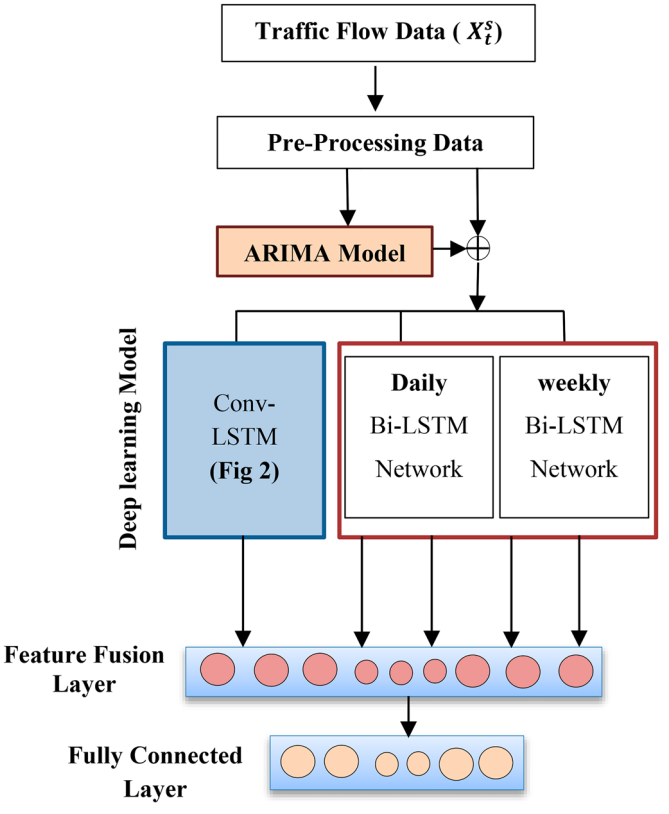

Figure 1. The overall architecture of the traffic flow prediction.

to others because Conv-LSTM performs matrix multiplication as a convolution operation. It extracts essential patterns by the convolution operations in multiple-dimensional data.

Figure 1 also depicts inclusion of Bi-LSTM in the overall proposed architecture. The reason for designing the Bi-LSTM component is that it can extract weekly and daily periodic correlations. Through Bi-Directional LSTM networks, regular patterns and new traffic flow features can be obtained. For example, traffic flow on holidays in a year is an essential feature through which the model can extract and consider them in forecasting. Traffic forecasting is performed by employing fully connected layers that receive information from the previous step, which is the feature fusion layer. This layer integrates the spatial and temporal information extracted by the Conv-LSTM network, enabling the model to capture the complex spatio-temporal correlations in the traffic data. The fully connected layers then produce the final prediction of the traffic flow. The proposed model also incorporates an attention mechanism, which automatically explores varying degrees of relevance of flow sequences at different periods and extracts more information from the input data. A detailed explanation of the attention mechanism will be provided in Section 3.6.

3.2. Autoregressive integrated moving average model, ARIMA

Previous experiments have demonstrated that the ARIMA model exhibits more acceptable performance for fixed and random times-series data, such as traffic data. Our proposed

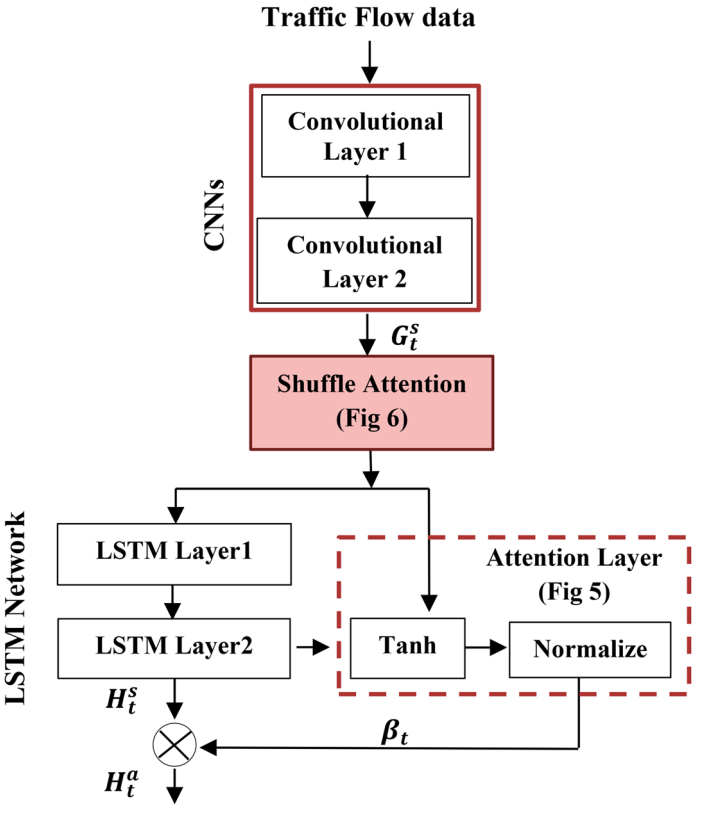

Figure 2. The Conv-LSTM module with an attention mechanism.

model also contains ARIMA to take advantage of its strengths in capturing trends in the traffic data. The ARIMA model is a popular time series model that combines the concepts of autoregression AR(p), integration, and moving average MA(q). These concepts are widely used in statistical modelling and analysis for predicting time series data. The ARIMA model is described by the equation below and is represented by the (p, D, q) parameters. The parameter D corresponds to the order of difference, p is the order of the autoregressive model, and q is the order of the moving-average model.

$$\left(1-\sum_{i=1}^{p}\emptyset_i L^i\right)(1-L)^D x_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{1}$$

The above equation describes an ARMA process, where x_t is the dependent variable, L is the lag operator, $\emptyset_i$ represents the AR parameter, θ_i denotes the moving average parameters, and ε_t is the error terms. Various methods have been developed to determine the p and q parameters. Box and Jenkins (Box et al. 2015) introduced the basic ARIMA algorithm, which consists of three main steps. The first step is to identify and select the type of model. The appropriate parameter D is employed to create a stationary time series at this stage. The second step involves estimating parameters (p, q), and finally, diagnostic statistics are used to evaluate the errors. If the variables and orders are appropriately selected, the ARIMA model effectively captures linear relationships in time series data.

3.3. Conv-LSTM component

The Conv-LSTM component is the central ingredient of the presented algorithm that strives to capture the nonlinear spatiotemporal features of the traffic flow. The convolutional neural networks and the LSTM are the main parts of this model, whose details and layers arrangement are depicted in Figure 2. The convolutional neural network comprises two separate one-dimensional components, whose output vector (G_t^s) enters the shuffle attention layer, which is presented in Figure 6 and will be stipulated later in Section 3.6.2. The LSTM network is also designed from two separate layers to be able to extract the temporal features between traffic flow data. The Conv-LSTM component receives traffic flow data as input which contains temporal and spatial information about the traffic network. To extract the spatial features, a one-dimensional convolution procedure is carried out on the traffic data and a sliding filter uses a convolution kernel filter to obtain a feature map of the prior layers.

The output of the convolutional layers, which processes the spatial correlations, is then fed into the LSTM layers to further capture the temporal correlations and characteristics of the traffic data. It is worth noting that a RNN also possesses a capability of interpreting hidden temporal characteristics in the sequential data (Funahashi and Nakamura 1993). Due to the vanishing gradient problem, RNNs may forget previous traffic conditions in the time series data, which can result in poor performs for long-term series. LSTM is therefore chosen as an appropriate method for learning long-term correlations, as it is able to memorise and store the long-term sequential data. Furthermore, the proposed model employs a stacked architecture, where multiple LSTM layers are sequentially arranged. Each layer receives the preceding layer's hidden state as its input, allowing the model to capture higher traffic flow characteristics. This approach is one of the prevalent practices used to improve the performance of deep neural networks.

3.3.1. The long-short term memory, LSTM

The LSTM model, proposed by Hochreiter and Schmidhuber (1997) is a special form of the RNN. The LSTM network can extract nonlinear correlations in historical data and retaining prior information over a long time due to its ability to handle large-dimension parameters and nonlinear activation functions in each layer. The LSTM network solves the issue of long-term learning correlations. It comprises recurrent network units that keep the values of short and long periods, stores information in memory cells, and is better at identifying and leveraging long-range features. As shown in Figure 3, the LSTM structure consists of input, forget, and output gates. Since the unnecessary information does not need to be stored, this information is removed by the first layer of the memory gate and can be defined as follows:

$$f_t = \sigma(W_f \times x_t + U_f \times h_{t-1} + b_f) \tag{2}$$

In the above equation, σ represents the activation function and W_f and U_f are the network weights, X_t is the value of the network input, h_{t-1} is the network output at time $t-1$, and b_f is the bias parameter.

The input gate is a second gate, and at this stage, the model decides what information the input vector should contain and store in the cell. In input gate i_t updates the value and

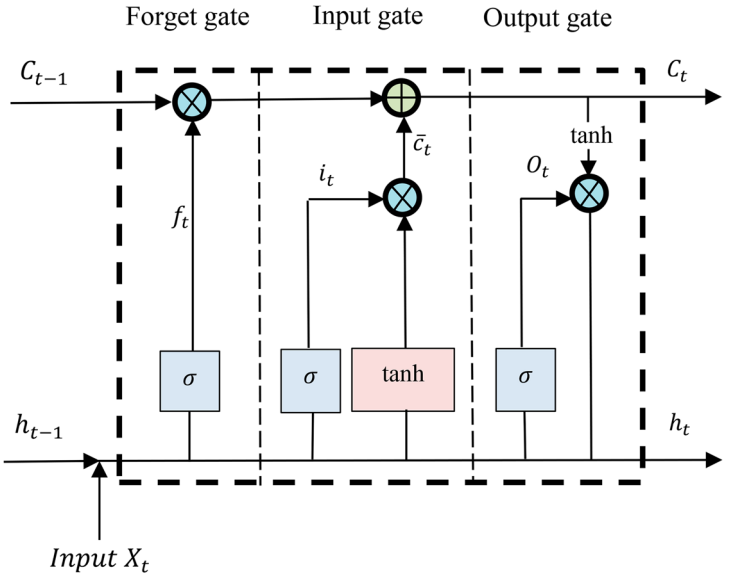

Figure 3. LSTM cell structure.

a tanh layer to generate a new state value C_t.

$$i_t = \sigma(W_i \times x_t + U_i \times h_{t-1} + b_i) \tag{3}$$

$$\bar{c}_t = \sigma(W_c \times x_t + U_c \times h_{t-1} + b_c) \tag{4}$$

where i_t defines the input threshold at time t, W_i, U_i, and U_c are the weights, b_c and $b_i$are bias parameters. In equation (5(, the cell states are updated at time t.

$$C_t = f_t \times C_{t-1} + i_t \times \bar{c}_t \tag{5}$$

Equation (6) shows the output gate formulation which generates the output information. W_o and U_i in equation (6) show the network weights and b_o is the expression of bias.

$$O_t = \sigma(W_o \times x_t + U_o \times h_{t-1} + b_o) \tag{6}$$

Equation (7) shows the current memory cell and the hidden unit at t time step. The last step provides valuable information as output after data passes through these three gates and forgets invalid information.

$$h_t = O_t \times \tanh(C_t) \tag{7}$$

3.3.2. Convolutional neural networks (CNN)

CNNs have been used to extract features in various applications, and this paper builds a CNN network to capture spatial data from a specific road network. A CNN network includes input, convolution, pooling, and fully connected layers. Finally, the output layer extract features and generates output. The main distinction between fully connected ANNs and CNNs is that CNN neurones are only linked to a smaller fraction of input, which minimises the overall parameters in the network. CNNs also employ the convolution function to extract

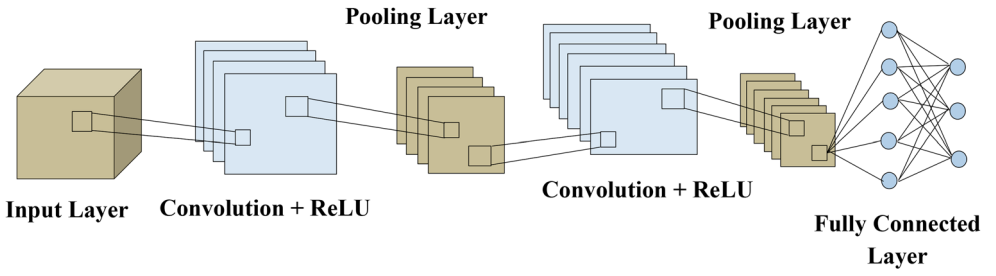

Figure 4. The CNN Structure.

crucial characteristics from matrices. It detects matrices patterns by evaluating the input features' spatial structure. To extract input features, CNNs do not require pre-processing. Furthermore, CNNs address the issue of overfitting that a standard NN suffers from.

Figure 4 illustrates the CNN structure, and the convolution layer is the core structural component of a convolution network, and its main objective is to extract and learn data characteristics.

The convolution layer provides a greater level of complexity to the CNN model. Each convolution layer employs several filters (kernels), each of them has its own set of weights. Convolution layers utilise filters with the same structure and form as the input matrix but are substantially smaller in size and dimensions. During the training process, the weights of filters are determined automatically based on the demand. Each convolution layer's filter is applied via the input layer, and the sum of the product of input and filter finally presents the feature map of each filter. Each feature map identifies a characteristic captured by a neural network (NN). In the CNN network training process with the backpropagation technique, there are two popular types of activation functions; Rectified Linear activation Unit (ReLU) and sigmoid function, in this study ReLU is adopted as the activation function. As a result, following the convolution layer, ReLU function is engaged to eliminate any negative values from the feature map. Following that, the pooling layer operates on the feature maps to reduce the spatial size of the matrix by minimising the representational dimensionality of each feature map, lowering the computational difficulty of the model. This layer speeds up the computations and eliminates the problem of overfitting. The resulting matrix is then flattened into a vector and sent into a fully connected layer, such as a NN.

3.4. Attention layers

The attention model is primarily focused on the inherent features patterns of data and indeed improves the efficiency of information processing (Zheng et al. 2021). The importance of historical traffic data in practice is not the same; therefore, more critical information need be considered for better forecasting performance. Furthermore, traffic conditions in specific locations at different times will have different effects on traffic flow forecasting. However, this importance is not realised merely by deep learning models alone like standard LSTM. The attention mechanism enables to Conv-LSTM component to determine the importance of data segments in the time series. This indeed improved the efficiency of the network and plays a significant role in accurately extracting deeper and essential features in the network. This paper introduces a shuffle attention (SA) unit to address this

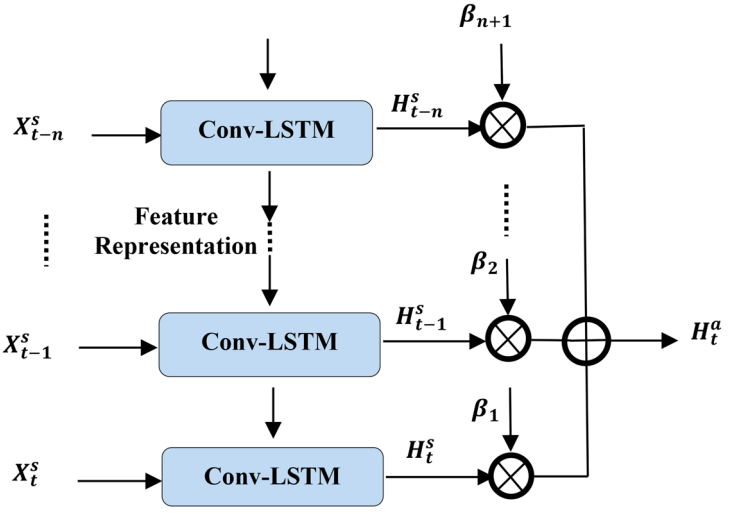

Figure 5. The attention Layers with Conv-LSTM networks (Zheng et al. 2021).

issue, employing shuffle units to effectively combine two types of attention mechanisms. There are two significant forms of attention mechanisms: channel attention and spatial attention. These approaches reinforce the critical characteristics by aggregating the same feature from all the positions using different aggregation techniques, transformations, and strengthening functions (Zhang and Yang 2021).

3.4.1. Attention layer

Figure 5 shows how the attention mechanism, and the Conv-LSTM network are integrated using the scoring method. As depicted in Figure 5, the output of Conv-LSTM network is computed as a weighted summation of the output of the LSTM network H^s_t as follows:

$$H^a_t = \sum_{k=1}^{n+1} \beta_k H^s_{t-(k-1),} \tag{8}$$

$$\beta_k = \frac{exp(s_k)}{\sum_{k=1}^{n+1} exp(s_k)} \tag{9}$$

In the above equation, $n+1$ indicates the flow sequence's length and β_k parameter is the value of temporal attention defined by the related formulation in equation (9). The traffic flow sequences, and related information has different importance and impact on traffic flow prediction, which is graded based on scores $S = [s_1, s_{2,}, \ldots, s_{n+1}]^T$, that shows the significance of each component in the sequence data. The scores $S = [s_1, s_{2,}, \ldots, s_{n+1}]^T$ demonstrate the importance of each component in the traffic flow data sequence, which can be presented as follows:

$$S_t = V_S^T \tanh(W_{hs} G^s_t + W_{ls} H^s_t). \tag{10}$$

In the above equation V_S, W_{Sx} and W_{hs} parameters are the learnable properties and H^s_t is the hidden output from the Conv-LSTM network. From equations (9) and (10), the attention

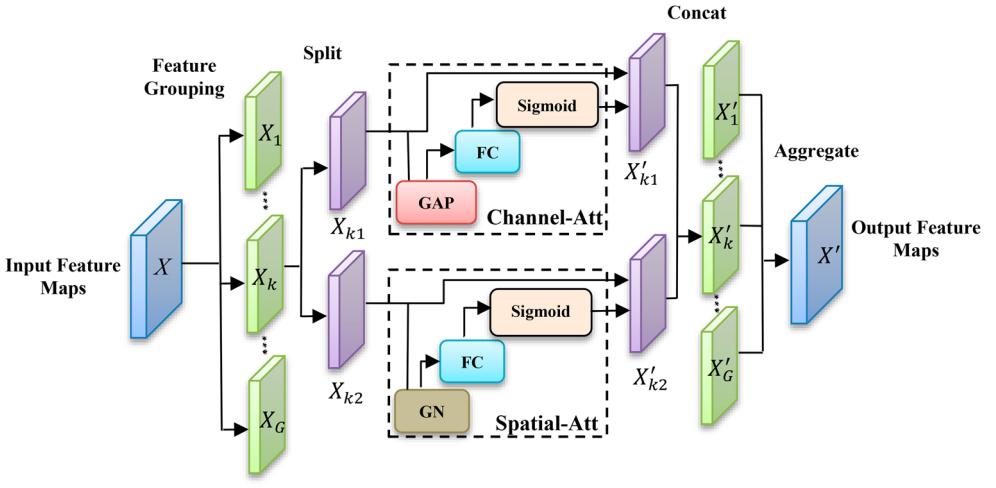

Figure 6. Shuffle Attention Mechanism.

value at time step t is affected by the CNNs layers (G_t^s) and the hidden variables H_t^s and its preceding n time steps.

3.4.2. Shuffle attention (SA)

One of the implied ideas of this paper, which is originally used in the field of ITS, is the use of shuffle attention (SA) integrated with CNN networks. The SA multiple channel features that boost the high-level CNN characteristics show the most promising performances on prediction tasks. The attention mechanism has been proven useful in a variety of computer applications. Zhang and Yang (2021) proposed a shuffle attention model, in which feature maps are collected, learned, and then shuffled in the channel dimension to interact with each other across channels.

We customised the basic SA module, which separates the input feature map into groups and employs the shuffle unit to combine spatiotemporal attention to extract more important features in less time. Following that, all secondary features are aggregated, and a 'channel shuffle' function is utilised to facilitate information transfer between secondary features. Finally, we illustrate the effect and validate the suggested SA's dependability. As shown in Figure 6, the entire structure of SA is illustrated and explained in four sections:

3.4.2.1. Feature classification. Let us consider when we are provided with a feature map as input $X \in R^{C\times W\times H}$, where H indicates height, W indicates width, and C indicates the feature map's channel number. Along the channel dimension, in SA unit X feature map divided into G groups, namely $X = [X_1, \ldots, X_G]$, $X_k \in R^{C/G\times W\times H}$, following the channel direction, each group is divided into two branches $X_{k1}, X_{k2} \in R^{C/2G*W*H}$. As illustrated in Figure 6 the first branch generates a channel or temporal attention maps, on the other hand, the second branch generates a spatial attention map by utilising the spatial correlation of characteristics.

3.4.2.2. Channel attention. Each layer of this channel focuses on what the input data means. To calculate channel attention, the spatial dimension of the input feature map is

compressed. Zhang and Yang (2021) presented a method to obtain the temporal features by a utilising the global average pooling (GAP), scaling, and activation to produce channel-wise statistics as $s \in R^{C/G\times 1\times 1}$, which can be formulated by shrinking X_{k1}through spatial dimension $W \times H$, and equation (12) represents the final output channel attention. One of the advantages of GAP in this mechanism is that the average of each feature map is extracted instead of adding fully connected layers at the top of the feature maps.

$$S = F_{GAP}(X_{k1}) = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H} X_{k1}(i,j), \tag{11}$$

$$X'_{k1} = \sigma(F_c(s)).X_{k1} = \sigma(W_1 s + b_1).X_{k1}, \tag{12}$$

where $W_1, b_1 \in R^{C/2G\times 1\times 1}$, are learnable parameters that constantly trained by the network. The essential part of neural networks is the sigmoid activation function (σ) and the fully connection (FC).

3.4.2.3. Spatial attention. This type of attention can be considered as an augmentation to channel attention. Normalisation layer that divides channels into groups and the Group Norm (GN) over X_{k2} to obtain space-wise statistics which is suggested by Wu and He (2018). Then, $F_c(\cdot)$ is used to improve the representation of X_{k2}. The following equation relates to spatial attention:

$$X'_{k2} = \sigma(W_2.GN(X_{k2}) + b_2).X_{k2}, \tag{13}$$

where $W_2, b_2 \in R^{C/2G\times 1\times 1}$ can be continually trained through the network.

3.4.2.4. Aggregation. Following the completion of the two types of attention learning characteristics, the two branches should be aggregated, $X'_k = [X'_{k1}, X'_{k2}] \in R^{C/G\times W\times H}$. After that, the sub-features are concatenated, and the channel shuffle procedure is completed.

4. Performance evaluation

4.1. Data description

To evaluate the performance of the proposed technique, we used a real-world dataset. The dataset was gathered from the California Department of Transportation's (Caltrans) Performance Measurement System (PeMS), commonly used for traffic prediction. The data for the experimental tests is from 7 observation locations for the six-month duration from 11 September 2017, to 4 March 2018. The data sampling interval was 5 min, resulting in 288 samples per day by seven sensors in different locations. The selected neighbourhood is on Street I980, District 4 in Oakland city. In this study, there were 44,336 training samples and 2016 test samples, which are used as input for the model. Figure 7(b) illustrates the sensors distribution of the I980 District 4 in the City of Oakland used for urban prediction. In this study, all sensors aggregate and store traffic data every 5 min ($\Delta = 5minutes$), the time window for all tests is set to $n = 15$ (time stamp), inferring 75-min ($t - n\Delta, \ldots, t - \Delta, t$) of historical data is used for training. Traffic flow with prediction horizon $h = 5$, 15, 30 and 60 min will be predicted.

As can be seen, Figure 7(a) plots the historical data of the traffic flow for five days from 18 September to 22 September 2017, at Street I980 District 4. Due to the fluctuations of

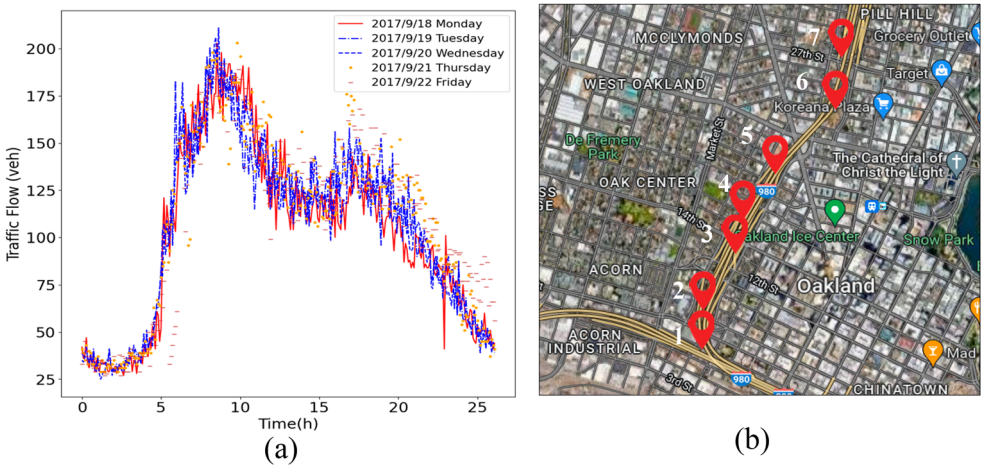

Figure 7. (a) The traffic flow from 18/9/2017 to 18/22/2017, (b) Sensor distribution of I-980 dataset.

historical data, it can be seen that the data in this spot follows a periodic pattern, which shows a traffic routine at different times of the day.

4.2. Measures of effectiveness

To evaluate the accuracy of forecasting models, we employed three measures of effectiveness, namely: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) which are conventionally defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_P - F_t|, \tag{14}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|F_P - F_t|}{v_i} \times 100\%, \tag{15}$$

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(F_P - F_t)^2}. \tag{16}$$

4.3. The model's prediction performance and discussion

4.3.1. Experimental design

We used the suggested model with several modules such as ARIMA, Conv-LSTM, Bi-LSTM, and the shuffle attention mechanism to predict the traffic flow. The TensorFlow framework is used to implement the model. In the analysis, the convolutional layer has ten filters with each having a size of 3. The sliding window's stride for the input flow data is set to 1. For training data, the batch size is set at 128. The activation function is the ReLU. The time window is set to 15 in all tests, indicating that 75-min historical data is utilised as a training unit.

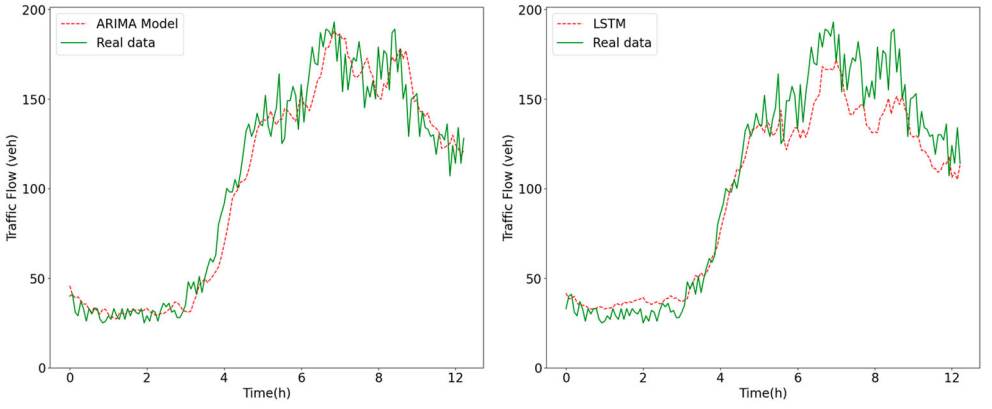

Figure 8. The predicted results of the ARIMA and LSTM model.

Table 1. Prediction performance with various proposed components on urban-I980.

	Time Step											
	5 min			15 min			30 min			60 min		
Model	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
ARIMA	15.24	11.3%	23.12	16.80	14.50%	23.42	17.63	18.6%	27.84	22.73	17.6%	29.12
CNN	15.14	11.3%	21.93	16.02	12.96%	22.14	17.12	14.6%	23.45	20.67	15.8%	27.04
Conv-LSTM#	14.70	11.5%	20.69	15.44	12.30%	21.73	16.93	13.2%	22.37	18.87	14.3%	25.44
Conv-LSTM∗	13.87	10.1%	19.16	15.81	11.20%	20.93	16.14	12.1%	21.48	17.61	13.1%	24.12
SAACL	**13.36**	**9.6%**	**17.86**	**14.74**	**10.6%**	**19.84**	**15.26**	**11.9%**	**20.82**	**16.38**	**12.7%**	**22.31**

Note: # with Bi-LSTM, ∗ with Attention.

4.3.2. Prediction performance result of the proposed model

The linear ARIMA model and nonlinear LSTM model are the two models that were evaluated in this study, and their performance is discussed. As demonstrated in Figure 8, the linear ARIMA and nonlinear LSTM models have not achieved satisfactory performance in capturing significant correlations in the traffic flow prediction task. Although the ARIMA model is linear, it can identify the ascending and descending slope and the linear traffic flow trends. However, it does not have enough generalisation power. One of the influential features of this model is that the ARIMA model extracts an overall representation of traffic patterns. Furthermore, this determination of data behaviour can be a great help for the proposed model, which we will discuss in more detail below. The LSTM model, as shown in Figure 8 (right), has unstable performance despite having structure of neural networks. This is due to the inability to extract spatial features (Fu, Zhang, and Li 2016).

As previously stated, traffic flow data was fed into the Conv-LSTM component to capture spatiotemporal characteristics. To demonstrate the efficacy of the suggested method, we evaluated the prediction effectiveness of the Conv-LSTM model with the CNN-LSTM while the attention mechanism or Bi-LSTM component is not included to evaluate the performance of this model and the results have been presented in Table 1. The Conv-LSTM component captures spatial characteristics from historical data by feeding it independently into the convolution layers, while the temporal features are extracted from data separately into the LSTM network.

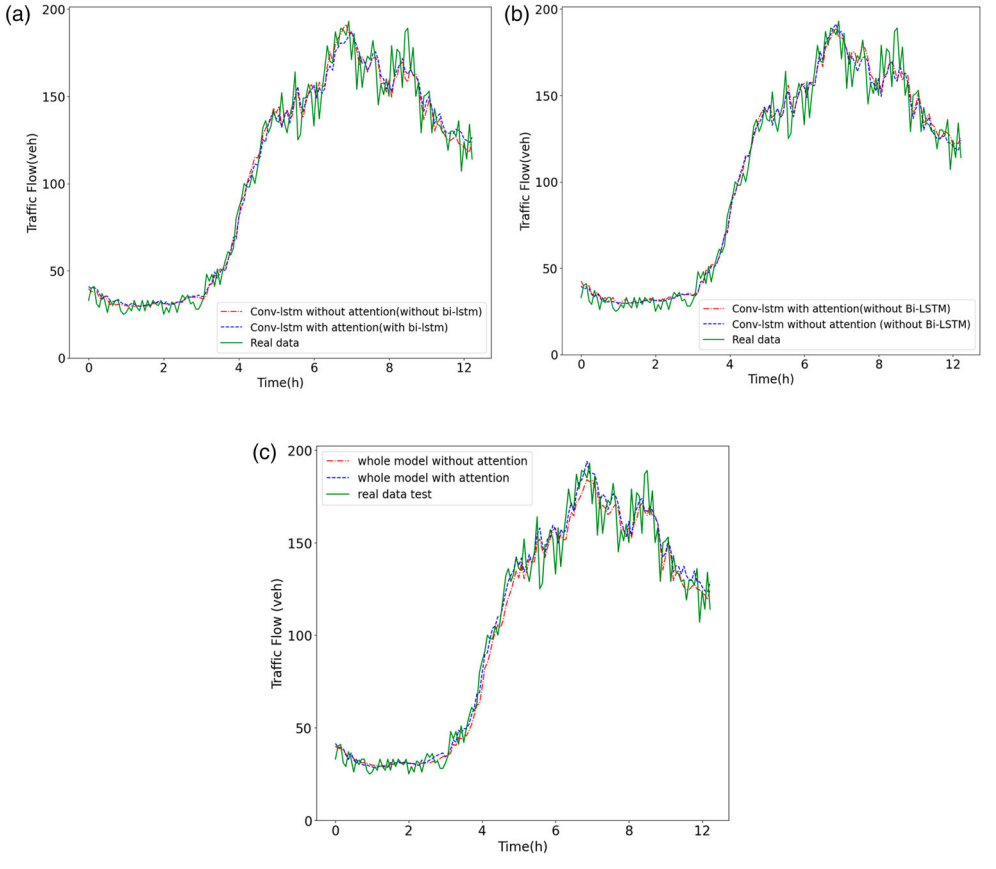

Figure 9. An illustration of the prediction performance with various proposed modules. (a) AT Conv-LSTM and Bi-LSTM and CNN-LSTM, (b) Conv-LSTM W/O Attention (Without Bi-LSTM), (c) Conv-LSTM W/O Bi-LSTM (W/O Without Attention).

Consequently, the spatial and temporal characteristics are concatenated and captured more efficiently in a dense or fully connected layer. Figure 9(a) visualises the performance evaluation of the designed forecast model over a 5-min forecast horizon from 0:00 to 12:00 AM. The Attention Conv-LSTM (AT-Conv-LSTM) and Bi-LSTM modules demonstrated superior performance and is closer to the ground truth than the Conv-LSTM module without attention layer and Bi-LSTM module especially during periods when there is a large fluctuation in traffic volume. In addition, we evaluated the prediction performance of the attention mechanisms. As discussed, traffic temporal and spatial characteristics of the traffic flow are captured by feature grouping and feature aggregation. Since the importance of data is not the same at different times. Attention mechanism automatically adjusts the weights accordingly when predicting traffic flow. It gives the Conv-LSTM module more generalisation power to extract essential parts of data and dependencies with greater accuracy and improved quality. Figure 9(b) visualises the model overall performance concerning the existence of the attention mechanism. By comparing the blue and red figures, it can be inferred that the forecasting model is closer to the ground truth and has followed the traffic fluctuations by integrating with the attention mechanism. We used a shuffle attention

layer after the CNNs neural network, which significantly impacted traffic forecasting error. Indeed, when the traffic volume has large fluctuations, the shuffle attention mechanism predicts better than when there are no attention mechanisms.

The suggested Bi-LSTM component can capture the temporal dependencies, such as daily patterns and weekly periodicities in historical traffic data. Consequently, the Bi-LSTM module by forward and backward passes can capture more periodic characteristics, improving prediction performance. Table 1 also depicts the MAE, MAPE, and RMSE prediction results and the effect of different modules by improving the generalisation of representations and prediction accuracy. According to the criteria and measurement matrices, the proposed model has a lower prediction error than the other models in all prediction horizons. From Figure 9 and Table 1, which show the prediction performance in terms of MAE, MAPE, and RMSE, it can be seen that the Bi-LSTM module significantly impacts the overall performance of the model, and by adding this module the model can achieve smaller prediction error. The difference between Figure 9(a) and (c) is in the implementation of the Bi-LSTM module. This also demonstrates that extracting periodic features enhances the forecasting performance of the traffic volume. According to the prediction results in Table 1, the traffic volume predicted by the Conv-Bi-LSTM module is more accurate than the CNN without Bi-LSTM. The attention layer's, in comparison to the Bi-LSTM module, significantly improves the presented predictive performance of the model. This improvement is due to the engagement of attention mechanism in interpreting deep learning models by giving attention weight to each data sequence element based on its significance (Do et al. 2019).

4.3.3. Discussion on performance comparison with different prediction algorithms

We conducted experimental tests to investigate the prediction performance of the proposed model. We compared the performance of the proposed model with other existing methods mentioned in the literature in order fairly and comprehensively validate the effectiveness of the model. Also in our comparison, we investigated the performance of other hybrid models, such as ARIMA + LSTM and ARIMA + CNN, in order to understand the positive impact of the hybridisation and addition of deep learning components, such as attention layers, convolutional layers, etc., on our proposed model. The comparison in Figure 10(a) and Table 2 shows the prediction performance of the hybrid deep learning models (combined ARIMA, LSTM, and CNN) which infers that hybridisation reduces the prediction error compared to the stand-alone ARIMA. The ARIMA model captures the linear features of time-series data, and then the CNN and LSTM empower the ARIMA to extract the spatiotemporal features. This is because ARIMA + LSTM, as a result of the neural network architecture extracting temporal features of data, performs better than ARIMA model. The inability to extract other inherent features, such as spatial and periodic features, leads to poor prediction performance compared to the proposed model. As the ARIMA + CNN model able to extract special dependencies (Jin et al. 2021), Figure 10(a) concur its superior performance while Table 2 outlines comparative performances with other alternative approaches. However, as the prediction horizons increase, this improvement becomes less prominent due to longer time horizons. Figure 10(b) illustrates the proposed model's output, which combines the ARIMA model and Conv-LSTM by incorporating the shuffle attention layer. As seen in Figure 10(b), as a result of the nonlinear shuffle-AT-Conv-LSTM, the ARIMA model is able to predict severe traffic data fluctuations. The impact of this hybrid

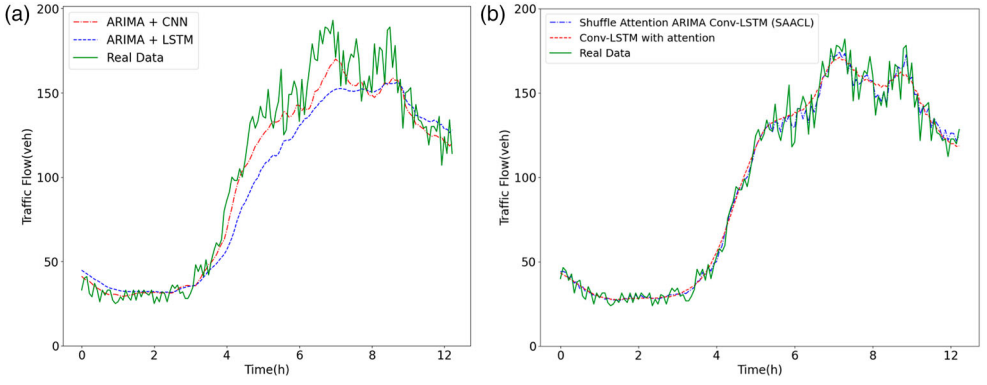

Figure 10. An illustration of the prediction performance with the aggregation time length is 5 min. (a) Hybrid model prediction without Attention layer, (b) Hybrid model prediction with attention layer (AT Conv-LSTM and Shuffle-AT-ARIMA Conv-LSTM).

Table 2. Performance comparison of different algorithms for traffic flow prediction on urban-I980.

	Time Step											
	5 min			15 min			30 min			60 min		
Model	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE
SVR	16.39	16.7%	20.82	18.03	18.0%	23.14	20.03	19.3%	25.92	24.00	21.9%	30.90
LSTM	14.77	11.1%	20.05	16.50	12.4%	22.57	19.00	14.7%	25.59	23.60	18.3%	30.75
ARIMA + LSTM	14.65	10.8%	19.87	16.51	12.3%	22.44	18.27	14.1%	25.04	22.36	17.1%	28.64
ARIMA + CNN	14.41	10.7%	19.74	16.32	12.3%	22.31	18.06	13.9%	24.91	22.19	16.5%	28.52
SAE	14.32	11.1%	19.61	16.25	12.1%	22.12	17.86	13.4%	24.54	21.31	16.2%	28.57
DNN-BTF	14.05	10.9%	19.32	15.55	11.5%	21.37	16.97	12.8%	23.06	19.12	14.8%	25.88
DCRNN	13.79	10.7%	18.88	14.79	11.5%	20.43	16.05	12.4%	22.18	18.43	14.2%	25.74
AT-Conv-LSTM	13.49	10.1%	18.56	14.34	10.8%	20.08	15.48	11.4%	21.26	16.65	12.3%	23.26
SAACL	**13.36**	**9.6%**	**17.86**	**14.74**	**10.6%**	**19.84**	**15.26**	**11.9%**	**20.82**	**16.38**	**12.7%**	**22.31**

structure with attention layer is evident from Figure 10(b), with the red graph denoting traffic flow data with greater accuracy.

In addition, we also evaluated several other prediction models to further investigate their performance. The results are presented in Table 2. Among the evaluated models, the SVR method produces the highest prediction error. This is likely due to its inability to sufficiently extract the spatial and periodic features of data, owing to its use of kernel function (Wu, Ho, and Lee 2004). Although LSTM model demonstrate relatively better performance due to the use of NN, the weakness in extracting special features prevents it from producing an acceptable prediction accuracy. The SAE model provides better prediction performance than the previous models, as this method inherently considers the spatial and temporal correlations (Lv et al. 2014). Due to the use of a multi-layered architecture, the DNN-based prediction model (DNN-BTF) (Wu et al. 2018) can extract inherent spatial and temporal features, and due to a recurrent neural network layers, it has significantly improved the prediction performance. The DCRNN model, by diffusion graph convolution, can dynamically extract spatial dependencies and temporal features with a recurrent neural network producing much more acceptable performance than the previous models. AT-Conv-LSTM model (Zheng

et al. 2021) has improved accuracy compared to other models by effectively extracting spatial–temporal and periodic features, indeed owing to the attention mechanism. The reason for this strong performance is that the traffic flow's spatial and temporal characteristics are intertwined, and the attention mechanisms and Bi-LSTM modules naturally extract these features effectively.

The proposed Shuffle Attention-ARIMA Conv-LSTM, is different from other methods, and presents a new attention mechanism called shuffle attention and a novel architecture for deep learning models that employs the linear ARIMA model to extract linear features and improve the predictive performance. According to the proposed architecture, traffic flow data is entered into the ARIMA model to represent the traffic data pattern and linearity features. The ARIMA model can significantly affect time series data due to its autoregressive (AR) and moving average (MA) models. The superiority of the hybrid strategy is predominantly as a result of incorporation the shuffle AT-Conv-LSTM network and the ARIMA model.

5. Conclusion

In this paper, we presented a new deep learning architecture to predict traffic flow. The ARIMA model, convolutional neural network, shuffle attention mechanism, and LSTM network comprised in a unique design to handle complicated nonlinear dynamic characteristics intrinsic to traffic flow data. The experiments proved that when spatiotemporal characteristics are treated simultaneously in the Conv-LSTM module, more relevant features are extracted from the data, which is vital in the dynamic prediction of urban traffic. Moreover, the suggested attention strategy and shuffle attention benefit the Conv-LSTM module to improve prediction performance. Furthermore, the presented Bi-LSTM component can extract daily and weekly patterns and periodic traffic information, enhancing predictive performance. According to experiments performed with real traffic data, the outcomes reveal that the presented model has superior forecasting capabilities over existing approaches.

There are two layers in the LSTM network. The two convolution neural networks extract the spatial features while the two Bi-LSTM networks capture the daily and weekly temporal features with potential periodic dependences. The performance of the forecasting model was evaluated using the actual flow dataset located at I980 Oakland city. The results show that the LSTM network and the ARIMA linear model cannot independently predict traffic flow to an acceptable accuracy particularly due to its inability to extract and exploit spatial features in complex data. The shuffle attention mechanism implemented in the model by scoring the features extracted by deep learning layers automatically detects the most important elements of each flow sequence at different times, and by using this strategy, the predictive performance is significantly improved. The proposed architecture, named Shuffle Attention-ARIMA Conv-LSTM (SAACL) can extract spatiotemporal and periodic features more effectively than ARIMA, LSTM, and conventional deep learning networks.

In the future, we will consider other mechanisms that can facilitate learning dynamics and hierarchical characteristics in sequential data by tackling several issues, such as long-term correlations in sequential datasets and multidimensional dynamic dependencies.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Van Thanh Huynh http://orcid.org/0000-0001-8668-3145

References

Box, George E. P. , Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. Wiley.

Cetin, M., and G. Comert. 2006. "Short-Term Traffic Flow Prediction with Regime Switching Models." *Transportation Research Record* 1965 (1): 23–31. https://doi.org/10.1177/0361198106196500103.

Dimitriou, L., T. Tsekeris, and A. Stathopoulos. 2008. "Adaptive Hybrid Fuzzy Rule-Based System Approach for Modeling and Predicting Urban Traffic Flow." *Transportation Research Part C: Emerging Technologies* 16 (5): 554–573. https://doi.org/10.1016/j.trc.2007.11.003.

Do, L. N., H. L. Vu, B. Q. Vo, Z. Liu, and D. Phung. 2019. "An Effective Spatial-Temporal Attention Based Neural Network for Traffic Flow Prediction." *Transportation Research Part C: Emerging Technologies* 108: 12–28. https://doi.org/10.1016/j.trc.2019.09.008.

Donahue, J., L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. 2015. *Long-Term Recurrent Convolutional Networks for Visual Recognition and Description*. https://doi.org/10.1109/CVPR.2015.7298878.

Feng, X., X. Ling, H. Zheng, Z. Chen, and Y. Xu. 2019. "Adaptive Multi-Kernel SVM With Spatial–Temporal Correlation for Short-Term Traffic Flow Prediction." *IEEE Transactions on Intelligent Transportation Systems* 20 (6): 2001–2013. https://doi.org/10.1109/TITS.2018.2854913.

Fu, R., Z. Zhang, and L. Li. 2016. "Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction." In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, November 11–13, 2016. https://doi.org/10.1109/YAC.2016.7804912.

Funahashi, K-i, and Y. Nakamura. 1993. "Approximation of Dynamical Systems by Continuous Time Recurrent Neural Networks." *Neural Networks* 6 (6): 801–806. https://doi.org/10.1016/S0893-6080(05)80125-X.

Ghosh, B., B. Basu, and M. O'Mahony. 2007. "Bayesian Time-Series Model for Short-Term Traffic Flow Forecasting." *Journal of Transportation Engineering* 133 (3): 180–189. https://doi.org/10.1061/(ASCE)0733-947X(2007)133:3(180).

Hamed, M. M., H. R. Al-Masaeid, and Z. M. B. Said. 1995. "Short-term Prediction of Traffic Volume in Urban Arterials." *Journal of Transportation Engineering* 121 (3): 249–254. https://doi.org/10.1061/(ASCE)0733-947X(1995)121:3(249).

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, W.-C., Y. Dong, F. Zheng, and S.-Y. Wei. 2011. "Hybrid Evolutionary Algorithms in a SVR Traffic Flow Forecasting Model." *Applied Mathematics and Computation* 217 (15): 6733–6747. https://doi.org/10.1016/j.amc.2011.01.073.

Huang, W., G. Song, H. Hong, and K. Xie. 2014. "Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning." *IEEE Transactions on Intelligent Transportation Systems* 15 (5): 2191–2201. https://doi.org/10.1109/TITS.2014.2311123.

Jin, K., J. Wi, E. Lee, S. Kang, S. Kim, and Y. Kim. 2021. "TrafficBERT: Pre-Trained Model with Large-Scale Data for Long-Range Traffic Flow Forecasting." *Expert Systems with Applications* 186: 115738. https://doi.org/10.1016/j.eswa.2021.115738.

Lee, S., and D. B. Fambro. 1999. "Application of Subset Autoregressive Integrated Moving Average Model for Short-Term Freeway Traffic Volume Forecasting." *Transportation Research Record* 1678 (1): 179–188. https://doi.org/10.3141/1678-22.

Levin, M., and Y.-D. Tsao. 1980. "On Forecasting Freeway Occupancies and Volumes (Abridgment)." *Transportation Research Record*, 773.

Li, Y., R. Yu, C. Shahabi, and Y. Liu. 2017. "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting." Preprint, arXiv:1707.01926.

Lu, S., Q. Zhang, G. Chen, and D. Seng. 2021. "A Combined Method for Short-Term Traffic Flow Prediction Based on Recurrent Neural Network." *Alexandria Engineering Journal* 60 (1): 87–94. https://doi.org/10.1016/j.aej.2020.06.008.

Lv, Y., Y. Duan, W. Kang, Z. Li, and F.-Y. Wang. 2014. "Traffic Flow Prediction with Big Data: A Deep Learning Approach." *IEEE Transactions on Intelligent Transportation Systems* 16 (2): 865–873. https://doi.org/10.1109/TITS.2014.2345663.

Ma, X., Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang. 2017. "Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction." *Sensors* 17 (4): 818. https://doi.org/10.3390/s17040818. https://www.mdpi.com/1424-8220/17/4/818.

Miglani, A., and N. Kumar. 2019. "Deep Learning Models for Traffic Flow Prediction in Autonomous Vehicles: A Review, Solutions, and Challenges." *Vehicular Communications* 20: 100184. https://doi.org/10.1016/j.vehcom.2019.100184.

Okutani, I., and Y. J. Stephanedes. 1984. "Dynamic Prediction of Traffic Volume Through Kalman Filtering Theory." *Transportation Research Part B: Methodological* 18 (1): 1–11. https://doi.org/10.1016/0191-2615(84)90002-X.

Rostami-Shahrbabaki, M., A. A. Safavi, M. Papageorgiou, P. Setoodeh, and I. Papamichail. 2020. "State Estimation in Urban Traffic Networks: A two-Layer Approach." *Transportation Research Part C: Emerging Technologies* 115: 102616. https://doi.org/10.1016/j.trc.2020.102616.

Shiliang, S., Z. Changshui, and Y. Guoqiang. 2006. "A Bayesian Network Approach to Traffic Flow Forecasting." *IEEE Transactions on Intelligent Transportation Systems* 7 (1): 124–132. https://doi.org/10.1109/TITS.2006.869623.

Socher, R., B. Huval, B. Bath, C. D. Manning, and A. Ng. 2012. "Convolutional-Recursive Deep Learning for 3d Object Classification." *Advances in Neural Information Processing Systems* 25: 656–664.

Tan, M., S. C. Wong, J. Xu, Z. Guan, and P. Zhang. 2009. "An Aggregation Approach to Short-Term Traffic Flow Prediction." *IEEE Transactions on Intelligent Transportation Systems* 10 (1): 60–69. https://doi.org/10.1109/TITS.2008.2011693.

Vinyals, O., A. Toshev, S. Bengio, and D. Erhan. 2015. "Show and Tell: A Neural Image Caption Generator." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Voort, Mvd, M. S. Dougherty, and S. M. Watson. 1996. "Combining Kohonen Maps with Arima Time Series Models to Forecast Traffic Flow." *Transportation Research Part C: Emerging Technologies* 4 (5): 307–318. https://doi.org/10.1016/S0968-090X(97)82903-8.

Wang, J., L.-C. Yu, Lai KR, and X. Zhang. 2016. "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Williams, B. M., and L. A. Hoel. 2003. "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results." *Journal of Transportation Engineering* 129 (6): 664–672. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664).

Wu, Y., and K. He. 2018. "Group Normalization." In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wu, C.-H., J.-M. Ho, and D.-T. Lee. 2004. "Travel-time Prediction with Support Vector Regression." *IEEE Transactions on Intelligent Transportation Systems* 5 (4): 276–281. https://doi.org/10.1109/TITS.2004.837813.

Wu, Y., H. Tan, L. Qin, B. Ran, and Z. Jiang. 2018. "A Hybrid Deep Learning Based Traffic Flow Prediction Method and Its Understanding." *Transportation Research Part C: Emerging Technologies* 90: 166–180. https://doi.org/10.1016/j.trc.2018.03.001.

Wu, Z., X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. 2015. "Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification." [conference presentation] *Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia*. https://doi.org/10.1145/2733373.2806222.

Xie, Y., J. Niu, Y. Zhang, and F. Ren. 2022. "Multisize Patched Spatial-Temporal Transformer Network for Short- and Long-Term Crowd Flow Prediction." *IEEE Transactions on Intelligent Transportation Systems* 23 (11): 21548–21568. https://doi.org/10.1109/TITS.2022.3186707.

Xing, L., and W. Liu. 2021. "A Data Fusion Powered Bi-Directional Long Short Term Memory Model for Predicting Multi-Lane Short Term Traffic Flow." *IEEE Transactions on Intelligent Transportation Systems* 23 (9): 16810–16819. https://doi.org/10.1109/TITS.2021.3095095.

Yan, H., X. Ma, and Z. Pu. 2021. "Learning Dynamic and Hierarchical Traffic Spatiotemporal Features With Transformer." *IEEE Transactions on Intelligent Transportation Systems* 23 (11): 22386–22399. https://doi.org/10.1109/TITS.2021.3102983.

Yang, H. F., T. S. Dillon, and Y. P. Chen. 2017. "Optimized Structure of the Traffic Flow Forecasting Model With a Deep Learning Approach." *IEEE Transactions on Neural Networks and Learning Systems* 28 (10): 2371–2381. https://doi.org/10.1109/TNNLS.2016.2574840.

Yao, Z.-s., C.-f. Shao, and Y.-l. Gao. 2006. "Research on Methods of Short-Term Traffic Forecasting Based on Support Vector Regression." *Journal of Beijing Jiaotong University* 30 (3): 19–22.

Yu, B., H. Yin, and Z. Zhu. 2017. "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting." Preprint, arXiv:1709.04875.

Zhang, X.-l., G.-g. He, and H.-p. Lu. 2009. "Short-term Traffic Flow Forecasting Based on K-Nearest Neighbors non-Parametric Regression." *Journal of Systems Engineering* 24 (2): 178–183.

Zhang, Y., and G. Huang. 2018. "traffic Flow Prediction Model Based on Deep Belief Network and Genetic Algorithm." *IET Intelligent Transport Systems* 12 (6): 533–541. https://doi.org/10.1049/iet-its.2017.0199.

Zhang, Q.-L., and Y.-B. Yang. 2021. "Sa-net: Shuffle Attention for Deep Convolutional Neural Networks." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Zhao, Z., W. Chen, X. Wu, P. C. Y. Chen, and J. Liu. 2017. "LSTM Network: A Deep Learning Approach for Short-Term Traffic Forecast." *IET Intelligent Transport Systems* 11 (2): 68–75. https://doi.org/10.1049/iet-its.2016.0208.

Zheng, H., F. Lin, X. Feng, and Y. Chen. 2021. "A Hybrid Deep Learning Model With Attention-Based Conv-LSTM Networks for Short-Term Traffic Flow Prediction." *IEEE Transactions on Intelligent Transportation Systems* 22 (11): 6910–6920. https://doi.org/10.1109/TITS.2020.2997352.