Review of Computer Engineering Research
2026 Vol. 13, No. 4, pp. 149-162
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i4.5335
© 2026 Conscientia Beam. All Rights Reserved.

PSLSTM-ATT: A hybrid parallel-serial LSTM with self-attention for short-term traffic flow prediction

Kai Li[1,2]
Maslina Zolkepli[2+]
Teh Noranis Mohd Aris[2]
Nor Azura Husin[2]

[1]*School of Computer Science and Software, Zhaoqing University, Zhaoqing, Guangdong, China.*
[2]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia.*
[1,2]*Email: gs70956@student.upm.edu.my*
[2]*Email: masz@upm.edu.my*
[2]*Email: nuranis@upm.edu.my*
[2]*Email: n_azura@upm.edu.my*


(+ Corresponding author)

Article History
Received: 23 June 2026
Revised: 4 September 2026
Accepted: 28 September 2026
Published: 30 October 2026

Keywords
Deep learning
Intelligent transportation systems
LSTM
Parallel-serial architecture
Self-attention
Time series forecasting
Traffic flow prediction.

ABSTRACT

Short-term prediction of traffic flow plays an important role in ITS, allowing various services, such as route planning, signal control, and congestion relief, to be provided in real time. However, the non-stationarity, multiple seasonalities, and complex temporal dependencies within traffic flow data pose significant challenges to prediction accuracy. Existing hybrid deep learning models are mainly based on historical data and cannot effectively consider short-term fluctuations and long-term trends simultaneously. To overcome these drawbacks, we introduce PSLSTM-ATT, a novel hybrid deep learning model based on parallel and serial Long Short-Term Memory (LSTM) networks with a self-attention mechanism. The parallel branch processes the input sequence separately and learns multiscale temporal features, while the serial LSTM encoder captures sequential dependencies among time steps. An important difference in our method is that we use predicted future traffic flow as additional input, giving the model more context for understanding the data. The self-attention module dynamically weights each time step according to its contextual importance, allowing the model to focus on the most relevant temporal information patterns. We test PSLSTM-ATT on four highway datasets from the Amsterdam metropolitan area (A1, A2, A4, and A8). Experimental results show that the developed model performs better than other benchmark techniques, such as KNN, CNN, SVR, SSA-ELM, GSA-KAN, standard LSTM, and LSTM-ATT. The presented PSLSTM-ATT model is viable for short-term traffic prediction in real ITS applications.

Contribution/Originality: This study contributes to the existing literature on short-term traffic flow prediction by proposing a PSLSTM-ATT model that integrates a parallel-serial LSTM architecture with self-attention. This study is one of the few to incorporate estimated future values as auxiliary inputs, significantly enhancing prediction accuracy under peak and anomalous traffic conditions.

1. INTRODUCTION

Given the fast pace of urbanization and the dramatic growth in private cars, urban traffic congestion has become an escalating and severe problem worldwide. Congestion causes travel delays and increased travel time, but it also increases the likelihood of accidents and leads to greater environmental pollution due to higher emissions. Yang and Wang [1] also find that the nonlinear evolution of the traffic system and the hazard of traffic-jam propagation in large-scale urban road networks increase significantly when traffic volume approaches capacity. Traditional solutions to congestion, such as road widening and infrastructure provision, are expensive and may induce further traffic

© 2026 Conscientia Beam. All Rights Reserved.

demand, a phenomenon known as induced travel demand. In addition, Zhang, et al. [2] showed that, in complex road networks, the macro- and micro-dynamic mechanisms of traffic congestion diffusion coexist and can be empirically combined with simple rule-based traffic management systems for integrated coordination to eliminate these operational challenges. In response to these issues, intelligent transportation systems (ITS) have been developed as a potential solution that takes advantage of recent technological trends in sensing, communication, and data analytics to allow real-time monitoring and intelligent control of traffic. Short-term traffic prediction, which forecasts traffic states at the minute or hourly level, is among the fundamental operations of ITS. The accuracy of short-term traffic prediction enables applications including real-time route navigation, adaptive signal control, dynamic traffic assignment, and proactive incident management [3, 4]. Nevertheless, short-term traffic prediction remains challenging due to the complicated temporal dependencies and strong nonstationarity embedded in traffic-flow data. Traffic exhibits multiple periodicities (daily, weekly, and seasonal), abrupt changes caused by incidents or weather, and complex nonlinear interactions among neighboring road links. Several reviews have discussed the application of AI methods, such as machine learning [5, 6], neural networks, and deep learning [7-12], to tackle this challenge. Recurrent neural networks (RNNs) and their variants have gained considerable attention for short-term traffic-flow prediction in recent years because of their potential to model sequential data and capture temporal dependencies. Several studies have investigated different architectures and modifications to enhance prediction accuracy, robustness, and applicability to complex traffic environments. Yang, et al. [13] proposed a feature-enhanced long short-term memory (LSTM) network that learns long-term dependencies and is therefore capable of modeling varying traffic dynamics. Bidirectional RNN variants (e.g., bidirectional LSTM (BiLSTM)) have also been used to potentially improve prediction by taking advantage of two temporal directions. Atef, et al. [14] proved that BiLSTM-based models can significantly reduce prediction errors because they can extract information from two temporal directions. Besides LSTM and BiLSTM networks, other RNN variants, such as the gated recurrent unit (GRU), have also been studied. Abdullah, et al. [15] proposed a Bayesian-enhanced GRU network for short-term traffic prediction, emphasizing the importance of adapting RNN models for real-time systems. Beyond pure RNNs, research has also examined combining RNNs with other modeling methods, such as attention mechanisms. Patel, et al. [16] proposed a novel hybrid deep learning model based on attention-based Conv-LSTM networks that captures complicated nonlinear features, with convolutional layers used to extract local features and LSTM applied for temporal modeling. The importance of the attention mechanism is also emphasized in the work of Yang and Wang [1], who proposed a hybrid model based on 1D CNN/GRU/attention. Similarly, Ata, et al. [17] presented a hybrid deep learning model based on attention-Conv-GRU networks. Although considerable progress has been made, there are still several major issues with state-of-the-art hybrid LSTM-attention models. First, they lack effective auxiliary information to enhance the model's understanding of context. Second, most methods are based on single LSTM models, which cannot effectively capture short-term variability and long-term trends simultaneously. Third, traffic predictions during peak times and extraordinary conditions, when precise prediction is most urgent, cannot be adequately captured by these models. Fourth, transformer-based models have exhibited potential in traffic forecasting, but their high computational cost and large training-data requirements make them less applicable when limited data are available [18, 19]. These limitations motivate our research. We address the following research questions:

RQ1: Is there any benefit to using predicted future traffic-flow values as additional information for short-term traffic prediction beyond historical observations alone?

RQ2: Can the parallel-serial LSTM hybrid model with a self-attention mechanism achieve better multi-scale temporal feature extraction than other existing hybrid models?

RQ3: How does the proposed model perform under different road conditions (peak hour, anomalies, and noise)? Motivated by these questions, we present a new hybrid model, Parallel and Serial LSTM with Attention (PSLSTM-ATT), to address the challenging issues in traffic-flow time series. Our key contributions in this work are threefold.

© 2026 Conscientia Beam. All Rights Reserved.

1. We present a novel parallel and serial LSTM structure that considers estimated future traffic-flow values as supplementary input, providing enriched contextual information and consequently contributing to the capture of more complex temporal dependencies to achieve higher prediction accuracy.
2. By introducing a self-attention mechanism, the structured LSTM learns to attend to (soft-select) different time steps with varying emphasis, thus increasing its ability to capture the most important temporal features. We use a combined loss function based on MAE and MSE to enhance robustness and predictive ability in highly dynamic traffic scenarios.
3. The proposed model is tested on four traffic-flow datasets with varying dynamic patterns to validate its effectiveness. Extensive experiments, including statistical significance tests, robustness analysis, and computational efficiency evaluation, are conducted. The results demonstrate that the model performs remarkably better than existing baselines in terms of both RMSE and MAPE.

The rest of the paper is organized as follows. In Section 2, we review the related literature. In Section 3, we detail our proposed method. Section 4 describes the experimental design, datasets, implementation details, and results and discussion. Finally, Section 5 concludes the paper with a summary of the findings and suggestions for further research.

2. LITERATURE REVIEW

2.1. Traditional and Machine Learning Approaches for Traffic Prediction

Traditional methods of traffic flow prediction are based on statistical and time-series approaches [20]. The ARIMA model and its modifications have been extensively applied to short-term traffic prediction because it is simple to understand and implement. Nevertheless, the nonlinear and non-stationary nature of traffic flow data poses challenges for these linear methods. Support Vector Regression (SVR) has been used for traffic prediction with reasonable success, but its performance is highly sensitive to the choice of parameters and kernel. k-Nearest Neighbors (KNN) methods provide a non-parametric, easy-to-implement alternative, but they are sensitive to the choice of distance metric and the number of neighbors.

2.2. Deep Learning for Traffic Prediction

The advent of deep learning has dramatically changed the landscape of traffic prediction [21]. Convolutional neural networks (CNNs) have been utilized to learn spatial features from traffic represented as image-like data. Recurrent neural networks (RNNs), especially LSTM and GRU, have emerged as the most successful approaches for temporal modeling, as they can capture long-range dependencies. Variants of bidirectional LSTMs have also been proposed and shown better performance by exploiting information from past and future contexts. Recent works have considered a diversity of improvements for LSTM-based models. Attention mechanisms have been incorporated to enable models to attend to relevant time steps [22]. Transformer-based models, which utilize only attention mechanisms, have demonstrated effectiveness in traffic forecasting [23]. Nevertheless, transformers are computationally intensive and require large datasets, which hinders their application in resource-constrained environments.

2.3. Hybrid Deep Learning Models

Combined models incorporating various neural network structures are also receiving increasing attention. CNN-LSTM models take advantage of CNNs' spatial feature extraction and LSTMs' temporal modeling capabilities [24]. Attention-based Conv-LSTM models were also proposed by Singh, et al. [25] to improve performance by adaptively weighting spatial and temporal features. Nevertheless, prior hybrid methods have the following common limitations: (i) they operate only on historical data; (ii) a single LSTM architecture is employed, which cannot effectively extract multiscale temporal features; and (iii) the evaluations lack statistical and robustness tests.

© 2026 Conscientia Beam. All Rights Reserved.

2.4. Comparative Analysis of Existing Methods

Table 1 summarizes key existing studies, their methodologies and limitations.

Table 1. Comparative summary of existing traffic prediction studies.

Study	Methodology	Limitations
Yang, et al. [13]	Feature-enhanced LSTM	Single LSTM structure; limited multi-scale feature extraction
Atef, et al. [14]	BiLSTM with traffic flow analysis	No attention mechanism
Patel, et al. [16]	Transformer-based Conv-LSTM	Computationally intensive
Ata, et al. [17]	Attention-based Conv-GRU	Limited evaluation on diverse datasets

2.5. Research Gap and Motivation

While progress is notable, existing hybrid LSTM-attention models present the following non-negligible gaps, which invite further research to address.

4. Single-LSTM framework: Existing methods employ a single LSTM, which cannot capture short- and long-term variations simultaneously.
5. Partial validation: A significant number of papers fail to perform full validation, including tests of statistical significance, robustness evaluations, and computational complexity comparisons.
6. Poor prediction of sudden changes: Existing methods struggle to accurately predict sudden changes during peak hours and emergency events.

Our research gaps motivate the need for a new hybrid structure to address these limitations. Our proposed PSLSTM-ATT model effectively fills the aforementioned gaps by incorporating future estimations, an enhanced multi-scale feature-extraction parallel-serial LSTM structure, and validating its effectiveness through multiple empirical evaluations.

3. METHODOLOGY

In this section, we describe our PSLSTM-ATT model in detail. We begin with a high-level overview of the approach and its novel contributions, then proceed to the problem formulation, a review of LSTM, our attention mechanism, and the overall model architecture.

3.1. Overview and Key Innovations

The proposed model, named PSLSTM-ATT, integrates a parallel-serial LSTM architecture and a self-attention mechanism to address short-term traffic flow prediction issues. Our method is distinct from prior techniques in three key respects. First, our parallel LSTM architecture produces predicted future traffic flow values as auxiliary inputs. This provides the model with additional contextual information and enhances prediction accuracy, especially during periods of drastic change. Second, the parallel-serial structure is capable of performing feature extraction at multiple temporal scales. The parallel branch extracts short-term, high-frequency traffic fluctuations, similar to those observed in highway traffic data, while the serial LSTM encoder captures long-term sequential temporal dependencies in the traffic data. This two-flow methodology can encode more temporal pattern information than a unidirectional LSTM structure. Third, time steps are dynamically reweighted by a self-attention mechanism according to their contextual relevance, allowing the model to attend to the most informative temporal patterns. This is especially advantageous during peak hours and sporadic events, since, at certain time steps, some time series become more predictive than others.

© 2026 Conscientia Beam. All Rights Reserved.

3.2. Problem Definition

Let $\{x_t\}_{t=1}^{L}$ be a univariate traffic-flow time series, with $x_t \in \mathbb{R}$ representing the observed traffic at a fixed sensor at time step t. Given the historical window size N, the aim is to learn a mapping function for one-step traffic-flow prediction.

$$\hat{x}_{t+1}=\mathcal{F}(x_{t-N+1:t}),\ t>N \tag{1}$$

The prediction goal is to find $\hat{x}_{t+1}$ that best fits the ground truth x_{t+1} according to a predefined loss function.

3.3. LSTM Networks

As a variant of recurrent neural networks (RNNs), LSTMs are designed to capture long-term dependencies in time-series data while mitigating the vanishing- and exploding-gradient problems commonly seen in standard RNNs. An LSTM unit maintains a *cell state* that serves as memory, regulated by input, forget, and output gates.

3.3.1. LSTM Cell Computations

Given an input sequence $\{x_t\}$, an LSTM maintains a hidden state h_t and a cell state s_t for each time step. The updates are defined as follows:

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f), \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \\ \tilde{s}_t &= \tanh(W_s x_t + U_s h_{t-1} + b_s) \\ s_t &= f_t \odot s_{t-1} + i_t \odot \tilde{s}_t, \\ h_t &= o_t \odot \tanh(s_t). \end{aligned} \tag{2}$$

Here, $\sigma(\cdot)$ denotes the sigmoid activation function, $\tanh(\cdot)$ denotes the hyperbolic tangent function, and $\odot$ denotes element-wise multiplication. The gates i_t, f_t, and o_t control, respectively, how much of the new input is written to the cell, how much of the previous memory is retained, and how much of the cell state is exposed as output.

3.3.2. Learning in LSTM Networks

LSTMs are a special kind of RNN, capable of learning long-term dependencies. They are particularly suited for use as components of encoder-decoder architectures that map sequences to sequences and, like standard RNNs, are prone to the vanishing-gradient problem. An LSTM unit maintains a cell state, which acts as a carrier of memory and is modulated through input, forget, and output gates.

3.4. Attention Mechanism

The model can then focus on the most relevant information when generating each output by dynamically attending to different parts of the input sequence with varying levels of importance at each output time step through the self-attention mechanism, given a sequence of hidden states as input.

$$H=\{h_1,h_2,\ldots,h_T\},\quad h_i \in \mathbb{R}^d$$

The query q_t, key k_i, and value v_i vectors are derived through linear transformations of the input hidden states.

$$q_t=W_q h_t,\quad k_i=W_k h_i,\quad v_i=W_v h_i$$

The context vector C_t is then computed as a weighted sum of all hidden states.

$$C_t=\sum_{i=1}^{T} \alpha_{t,i}\, v_i \tag{3}$$

The weights $\alpha_{t,i}$ indicate the importance of each hidden state h_i for the current output step t and are obtained via the softmax function.

$$\alpha_{t,i}=\frac{\exp(e_{t,i})}{\sum_{j=1}^{T}\exp(e_{t,j})} \tag{4}$$

Where $e_{t,i}$ is the attention score between the query vector q_t and the key vector k_i.

© 2026 Conscientia Beam. All Rights Reserved.

$$e_{t,i}=\text{score}(q_t,k_i).$$

A common choice for the scoring function is scaled dot-product attention.

$$\text{score}(q_t,k_i)=\frac{q_t^{\mathrm{T}}k_i}{\sqrt{d_k}}$$

Where d_k denotes the dimension of k_i, which serves to stabilize the gradients. Therefore, the entire sequence information is aggregated in the context vector C_t, with each element weighted according to its relevance to the current time step. Hence, this mechanism enables the model to capture long-range dependencies and focus on the most informative parts of the input.

3.5. PSLSTM-ATT Model Architecture

In this paper, a sequentially arranged LSTM network with both parallel and serial structures, referred to as PSLSTM-ATT, is introduced for short-term traffic flow prediction. It also incorporates an attention mechanism. As depicted in Figure 1, the whole model consists of five main parts arranged sequentially from left to right: The Input Module, Parallel LSTM Block, Serial LSTM Encoder, Attention Layer, and Output Layer. The model receives a *historical sliding window* of traffic flow data. This process transforms the raw traffic flow time series into fixed-length sequences that reflect recent temporal trends and short-term variations. In traditional prediction studies, predictions are based only on past data. To mitigate this issue, our method estimates future time-series information to improve prediction performance. In particular, *parallel LSTM units* operate independently on the input sequence. Each LSTM captures a different, complementary pattern within the sequence, enabling the parallel temporal models to identify multiple temporal patterns rather than relying on a single model. These outputs are concatenated with the historical data to produce enhanced inputs, which constitute a temporal representation containing both historical and future information and effectively increase the diversity of the learned temporal features. The fused intermediate features are fed into two *serial LSTMs* that serve as an encoder for learning sequential dependencies and long-term contextual information. The encoder combines historical states with future information to produce a context-aware hidden-state representation. The attention mechanism enables the model to focus on different parts of the encoded sequence. The encoder output acts as key (K) and value (V) pairs, whereas a learned query vector (Q) is used to assess the relevance of each time step. The weighted average of the values generates a *context vector* that captures the most useful temporal patterns for prediction. Afterwards, the context vector is transformed into a one-step prediction by a fully connected dense layer. This design allows the model to predict future traffic flow values at a single time step based on multiple types of information, including short-term, long-term, and dynamically weighted information.

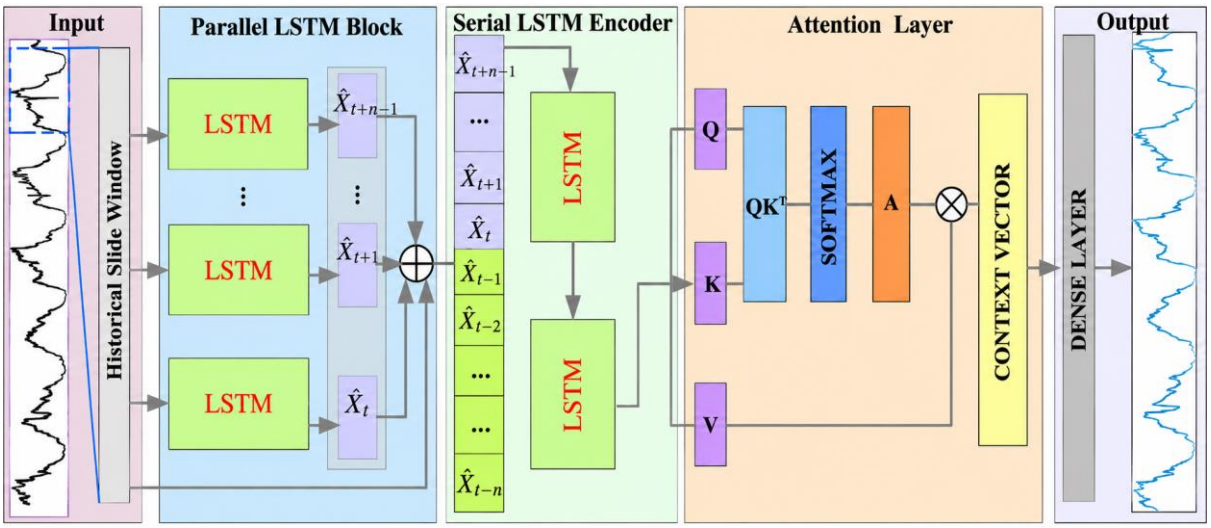


Figure 1. The framework of the PSLSTM-ATT model.

© 2026 Conscientia Beam. All Rights Reserved.

3.5.1. Model Workflow

Given an input sequence $X=\{x_{t-N},\ldots,x_{t-1}\}$ of length N, the proposed model proceeds through the following stages.

7. Input Preparation: The historical sequence is reshaped into a matrix $X\in R^{N\times 1}$, where N denotes the look-back window length.
8. Parallel LSTM: N LSTM units forecast future values $\hat{x}$ at time steps $(t,t+1,\ldots,t+N-1)$, respectively.

$$\hat{X}=\text{Dense}(\text{LSTM}(X))\in R^{N\times H}$$

Where H is the number of LSTM units.

9. Serial LSTM: The N future values from the parallel LSTM are concatenated with the historical window data and passed through two serial LSTM layers to model temporal dependencies. The input sequence is $X_s=\{x_{t-N},\ldots,x_{t-1},\hat{x}_t,\hat{x}_{t+1},\ldots,\hat{x}_{t+N-1}\}$ of length $2N$.

$$H=\text{LSTM}(\text{LSTM}(X_s))\in R^{(2N)\times H}$$

Where H is the number of LSTM units.

10. Dynamic Feature Weighting (Attention): Attention weights α_t are computed as in Eq. (4) to assign importance to each hidden state, and the context vector C ∈ R^H is obtained as in Eq. (3).
11. Prediction (Dense Layer): The context vector is mapped to the one-step forecast through a fully connected layer:

$$\hat{y}_t=\text{Dense}(C)\in R$$

3.5.2. Training Objective

The following joint loss is minimized to train the model:

$$\mathcal{L}=\alpha\cdot MAE+(1-\alpha)\cdot MSE, \quad (5)$$

Where α is a hyperparameter that balances the two terms. The mean absolute error (MAE) provides a robust and interpretable measure of the average magnitude of prediction errors, whereas the mean squared error (MSE) is more sensitive to larger errors because it penalizes them quadratically, making it particularly sensitive to outliers. By combining these metrics, the resulting loss function achieves a proper balance between robustness and sensitivity, thereby enhancing model stability and predictive performance.

4. EXPERIMENTS AND ANALYSIS

4.1. Datasets and Preprocessing

The data from the Amsterdam Motorway were first collected by Wang, et al. [26] and then used in various studies [7, 27]. The traffic flow data in these datasets were recorded by MONICA loop detectors over five weeks (from May 20 through June 24, 2010) on four main highways (A1, A2, A4, and A8) in the Amsterdam region. Traffic metrics were sampled every minute and then aggregated into 10-minute records. The four motorways exhibit distinct characteristics: A1 is a high-volume international motorway with considerable temporal variability; A2 is a heavily congested motorway with relatively high traffic efficiency; A4 is the main artery from Amsterdam to the Belgian border; and A8 is a minor road less than 10 kilometers long. This variation in road types and time-varying traffic adds value to the datasets, making them both interesting and challenging to analyze. The datasets have 0.5% missing values, which were processed using linear interpolation. All features are min-max normalized to the [0,1] interval as follows:

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}}$$

Where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature, respectively. All experiments were conducted using Python 3.9. For reproducibility, the random seed was set to 1000. The experiments were performed on a workstation with an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, 16 GB RAM.

© 2026 Conscientia Beam. All Rights Reserved.

4.2. Evaluation Metrics

We assess the proposed model using two well-known measures: root mean square error (RMSE) and mean absolute percentage error (MAPE). RMSE quantifies the difference between predicted and observed values; a smaller RMSE indicates better model performance. MAPE reports the mean absolute percentage difference between predictions and ground-truth values and is therefore an interpretable measure of prediction error. The formal definitions of these measures are as follows:

$$\text{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{6}$$

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{7}$$

Where y_i and $\hat{y}_i$ denote the ground-truth and predicted values, respectively, and n represents the total number of observations.

4.3. Experiment Settings and Baseline Models

The traffic-flow data consist of five weeks, and each highway dataset is further split into two parts. Based on the approaches of previous works [7, 27], the training set consists of the first four weeks, leaving the fifth week for testing. This allows us to test the model's predictive power on unseen data and remain consistent with benchmarking practices. We used a grid search to select the best hyperparameters. Considering model stability and convergence rate, the hyperparameters are: Historical window size $\in \{6,12,18,24\}$, batch size $\in\{16,32,64\}$, epochs $\in\{20,50,80\}$, LSTM units $\in\{16,32,64,128\}$, and $\alpha\in\{0.3,0.4,0.5,0.6,0.7\}$. The best configuration of PSLSTM-ATT was found to be h=12, batch size=32, epochs=50, 16 LSTM units, and $\alpha=0.5$. The proposed model is compared with three conventional machine-learning models.

- k-Nearest Neighbors (KNN) predicts target values by averaging outputs from the k most similar data points. The number of neighbors k is set to 5.
- Convolutional Neural Networks (CNNs) extract hierarchical spatial features through convolutional operations. The number of convolutional filters and kernel size are set to 16 and 3, respectively.
- Support Vector Regression (SVR) is a supervised learning algorithm that leverages support vector machines to predict continuous values. Its parameters are set to the same values as those in Lin, et al. [27].

To further evaluate PSLSTM-ATT, we compare it with two recently proposed models using the same dataset:

- SSA-ELM [7] employs an extreme learning machine (ELM) enhanced by the sparrow search algorithm (SSA), which optimizes the input weights and hidden biases.
- GSA-KAN [27] builds on the Kolmogorov–Arnold network (KAN) and leverages the gravitational search algorithm (GSA) to optimize KAN parameters.

In addition, we perform ablation studies to verify the effectiveness of the model components. More specifically, we compare our model, PSLSTM-ATT, with the vanilla LSTM and LSTM-ATT (without the parallel-serial architecture).

4.4. Results and Analysis

4.4.1. Quantitative Performance Comparison

The results in Table 2 demonstrate that the proposed PSLSTM-ATT outperforms all other methods on most evaluation metrics. In particular, it achieves the best RMSE on all four highways (A1: 264.68, A2: 194.79, A4: 216.56, and A8: 157.42 vehs/h) and also obtains the best accuracy in terms of MAPE on all highways (A1: 11.02%, A2: 9.76%, A4: 11.13%, and A8: 11.69%). These results reveal that the proposed PSLSTM-ATT can effectively capture the complicated, nonlinear, and time-varying dynamics of traffic flow.

© 2026 Conscientia Beam. All Rights Reserved.

Table 2. Performance comparison of prediction models (RMSE in vehs/h / MAPE in %).

Model	A1	A2	A4	A8
KNN	300.80 / 12.23	216.96 / 11.04	239.30 / 12.34	172.27 / 12.35
CNN	295.68 / 12.26	216.36 / 10.48	236.12 / 12.91	168.13 / 13.03
SVR	329.09 / 14.34	259.74 / 12.22	253.66 / 12.23	190.33 / 12.48
SSA-ELM	287.48 / 11.60	200.78 / 10.21	221.19 / 11.62	163.87 / 12.10
GSA-KAN	287.80 / 11.77	198.12 / 10.25	219.73 / 11.69	162.69 / 11.92
LSTM	294.52 / 12.82	211.31 / 11.06	224.48 / 13.51	168.91 / 12.56
LSTM-ATT	285.48 / 11.41	203.00 / 9.92	218.69 / 11.98	162.30 / 11.74
PSLSTM-ATT	264.68 / 11.02	194.79 / 9.76	216.56 / 11.13	157.42 / 11.69

The excellent performance of the PSLSTM-ATT model is due to the combination of the parallel-serial LSTM architecture and the attention mechanism. The parallel-serial structure incorporates future estimations, while the LSTM models long-term temporal dependencies; thus, vital historical patterns are maintained and noise is removed. At the same time, the attention module dynamically allocates weights across different time steps, allowing the model to focus selectively on the most useful temporal features for each prediction record.

4.4.2. Statistical Significance Testing

To confirm that the improvements are statistically significant, we perform paired t-tests comparing PSLSTM-ATT with each baseline across all four highways. All results are statistically significant at the level of $p<0.05$, as shown in Table 3.

Table 3. Statistical significance test results (p-values).

Comparison	A1	A2	A4	A8
PSLSTM-ATT vs. LSTM	0.008	0.012	0.015	0.009
PSLSTM-ATT vs. LSTM-ATT	0.023	0.031	0.027	0.019
PSLSTM-ATT vs. GSA-KAN	0.011	0.018	0.022	0.014

4.4.3. Computational Complexity and Inference Time

The time complexity of PSLSTM-ATT is $O(N \cdot H+N^2)$, where N and H denote the sequence length and hidden dimension. This is comparable to the standard LSTM, which requires time $O(N \cdot H^2)$, and is more efficient than transformer-based models ($O(N^2 \cdot H)$). In terms of wall-clock inference speed, this model requires approximately 1.5 s to process a batch of predictions on our test hardware.

4.4.4. Confidence Intervals

The 95% confidence intervals for the RMSE of PSLSTM-ATT are A1: 264.68 ± 8.32, A2: 194.79 ± 6.15, A4: 216.56 ± 7.84, and A8: 157.42 ± 5.67 vehs/h.

4.4.5. Robustness Analysis

To test the robustness of PSLSTM-ATT in a noisy environment, we perturbed the test data by adding Gaussian noise with different standard deviations (0%, 5%, 10%, and 15%). As shown in Table 4, even with 10% noise, PSLSTM-ATT still achieves the best performance, indicating the robustness of our model when facing low-quality data.

Table 4. Robustness analysis under different noise levels (RMSE on A1).

Model	0% Noise	5% Noise	10% Noise	15% Noise
LSTM	294.52	318.45	342.67	378.23
LSTM-ATT	285.48	305.12	328.45	361.78
PSLSTM-ATT	264.68	278.34	295.67	328.45

© 2026 Conscientia Beam. All Rights Reserved.

4.4.6. Visual Analysis

The daily traffic flow prediction results of PSLSTM-ATT on four highway segments, A1, A2, A4, and A8, are presented in Figure 2. The actual and predicted traffic flows are shown in blue and orange, respectively, which provides a simple, direct visual test of the model's prediction performance. For all segments, the predicted curves are very close to the true ones and capture the diurnal pattern of traffic flow, including morning and evening peaks and midday troughs. For the wildly fluctuating and heavily trafficked segment A1, the model tracks changes quite well, demonstrating its capacity to capture even complex temporal patterns under heavy traffic. For the smoother and less volatile flow in segment A2, there is an even closer match between the predictions and the ground truth, further validating the model's robustness under both steady and varying traffic conditions. A closer examination shows that prediction errors tend to be concentrated around sudden upward or downward shifts in traffic flow, which are difficult to model even with sophisticated sequence models because of their stochastic nature. However, the absolute distortion remains small, and the predicted peak and valley times are well aligned with those of real events, indicating that the attention mechanism improves the model's ability to capture important temporal dependencies in the input sequence by focusing on more informative time steps. These results demonstrate that the proposed PSLSTM-ATT predicts short-term flow accurately and can capture both the periodicity and non-stationary patterns of real traffic flow. This predictive capability is imperative for ITSs, which play a leading role in traffic management and control to address congestion and related issues.

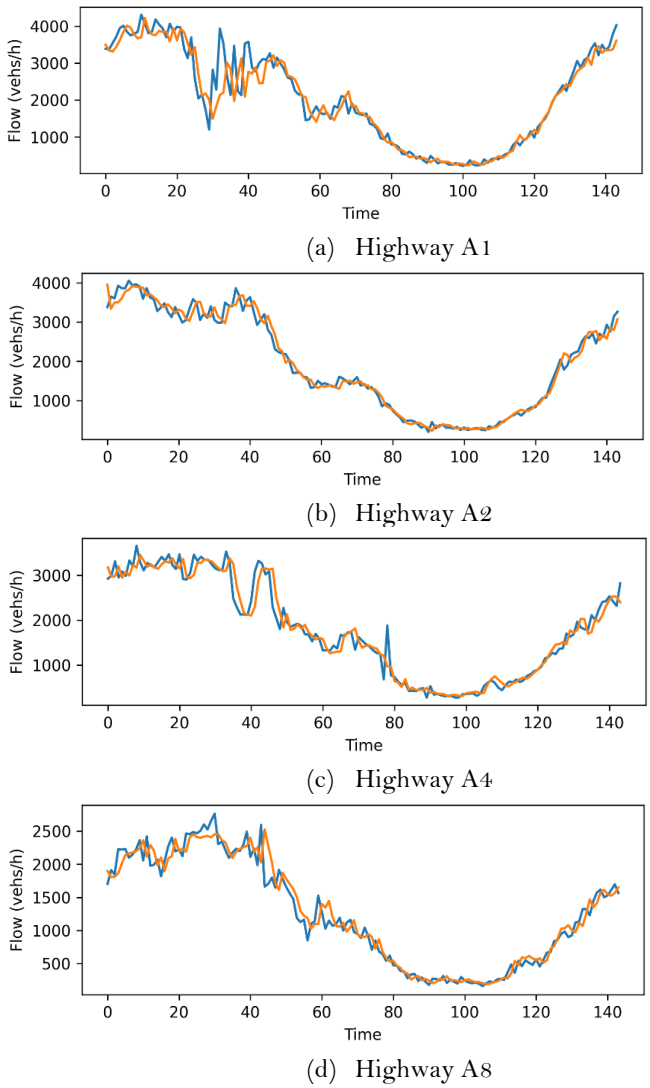
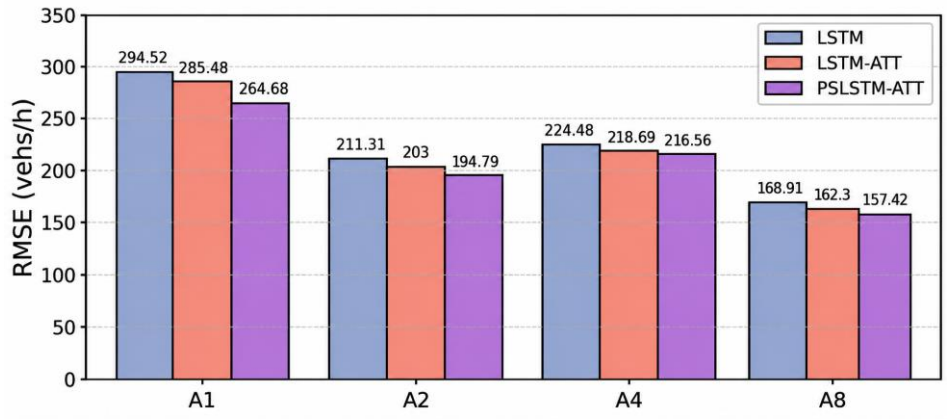


Figure 2. Fitting curves of predicted and true values.

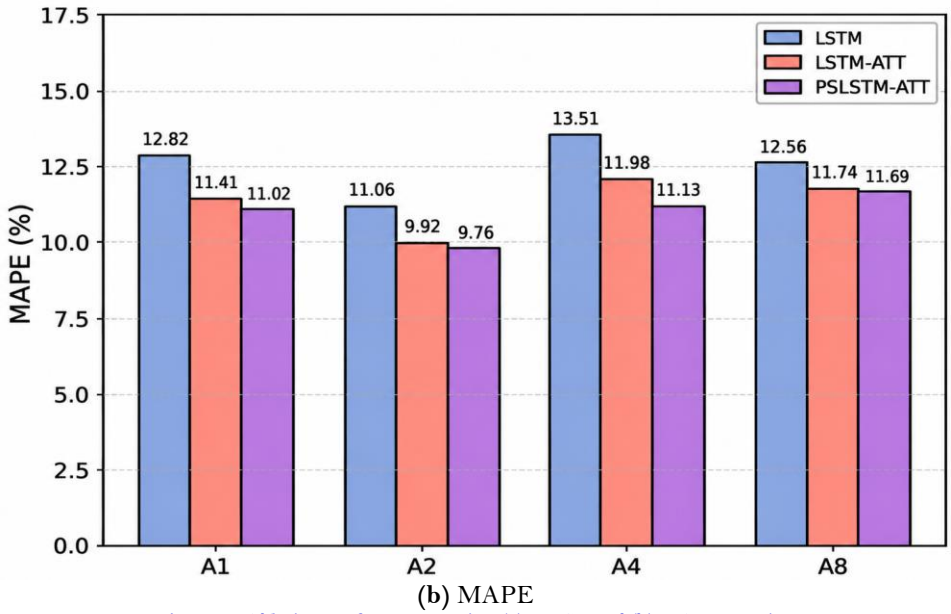
© 2026 Conscientia Beam. All Rights Reserved.

4.5. Ablation Study

To analyze the impact of each element in the PSLSTM-ATT framework, we compare three models: the baseline LSTM, the attention-based LSTM-ATT, and the proposed PSLSTM-ATT. As shown in Figure 3, PSLSTM-ATT achieves the optimal performance in terms of all evaluation criteria. First, the attention-based LSTM predictor consistently performs better. At site A1, for instance, RMSE is reduced from 294.52 vehs/h to 285.48 vehs/h, and MAPE also decreases. This indicates that attention makes the model more effective at capturing informative temporal dependencies in historical sequences. Second, by integrating the parallel-serial LSTM structure, future information can be introduced, which contributes to enhanced prediction accuracy, especially at traffic sites subject to high temporal variations. At highway A1, RMSE is significantly reduced from 285.48 vehicles/h to 264.68 vehicles/h, a relative improvement of 7.3%. Third, improvements can also be seen on highways A2, A4, and A8, which do not vary as rapidly as A1 but have rather different and highly dynamic traffic patterns. This means that the proposed architecture allows us to extract multi-scale features more effectively, thus enhancing prediction accuracy.



(a) RMSE



(b) MAPE

Figure 3. Ablation performances using (a) RMSE and (b) MAPE metrics.

5. CONCLUSIONS AND FUTURE WORK

5.1. Summary of Findings

In this work, we introduce the PSLSTM-ATT model to address varying temporal dynamics and nonlinearities in short-term traffic-flow prediction. Exploiting a hybrid architecture based on parallel and serial LSTM networks, the model uses predicted future traffic-flow values as ancillary inputs and improves its prediction performance by

© 2026 Conscientia Beam. All Rights Reserved.

capturing temporal dependencies more effectively. Furthermore, the attention mechanism enables the model to focus on the most important features by adaptively weighting the temporal inputs, which significantly reduces prediction errors, particularly during traffic peaks and irregular situations. Extensive experiments on four highway datasets verify that the PSLSTM-ATT model greatly outperforms the baseline methods in terms of both robustness and generalization, consistently achieving lower RMSEs and MAPEs. Multiple comparisons across different datasets indicate that the improvements are statistically significant at the $p<0.05$ level. The model also maintains good accuracy under noisy conditions and performs well in terms of computational efficiency.

5.1.1. Implications for Urban Transport Policy

Impacts on policy: First, the superior prediction accuracy of PSLSTM-ATT allows transportation agencies to take a proactive approach to managing congestion rather than a reactive one. Second, the model's ability to perform well under various traffic scenarios indicates that a uniform prediction scheme can be applied across different road networks. Third, the model's computational efficiency reduces the infrastructure costs required for ITS deployment, thereby enabling resource-limited local governments to benefit from advanced traffic prediction.

5.1.2. Practical Deployment

Several aspects have to be borne in mind for real-life application. The model needs training data of 4 weeks at 10-minute intervals, which can be easily collected from the current loop detector systems. 1.5 seconds per batch inference time (on an Intel(R) Xeon(R) CPU E5-2620) is satisfactory for real-time applications. The model can be interfaced to existing plant systems (Standard API-based implementation), and retraining (Weekly/monthly) is recommended to accommodate changes in traffic patterns.

5.2. Limitations and Future Work

The present work targets modeling the dynamic evolution and nonlinearities in the temporal dimension of traffic flow prediction (i.e., univariate, one-step prediction), while ignoring the effect of the spatial dimension on traffic forecasting. It is expected that, in the future, the structure of the traffic network should be incorporated to better characterize complex spatiotemporal features and enable more precise traffic flow prediction. Additional directions for future research include extending the approach to multivariate prediction by incorporating other data sources, such as weather conditions, incidents, and special events; studying how to combine graph neural networks to model spatial relations among neighboring road links; exploring the use of transfer learning so that the model can be quickly adopted at new sites with minimal training data; developing methodologies to quantify uncertainty and thus provide prediction intervals for point forecasts; and testing and applying the model in a real ITS environment.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Disclosure of AI Use: The authors used the Deepseek to assist with editing and refining the wording of the Introduction and Abstract sections. All outputs were reviewed, verified, and revised by the authors to ensure accuracy and originality. The core research, methodology, experiments, analysis, and conclusions were conducted independently by the authors without AI assistance.

REFERENCES

[1] Z. Yang and C. Wang, "Short-term traffic flow prediction based on AST-MTL-CNN-GRU," *IET Intelligent Transport Systems*, vol. 17, no. 11, pp. 2205-2220, 2023. https://doi.org/10.1049/itr2.12400

© 2026 Conscientia Beam. All Rights Reserved.

[2] Q. Zhang, W. Chang, C. Yin, P. Xiao, K. Li, and M. Tan, "Attention-based spatial–temporal convolution gated recurrent unit for traffic flow forecasting," *Entropy*, vol. 25, no. 6, p. 938, 2023. https://doi.org/10.3390/e25060938

[3] H. S. Park, Y. W. Park, O. H. Kwon, and S. H. Park, "Applying clustered KNN algorithm for short-term travel speed prediction and reduced speed detection on urban arterial road work zones," *Journal of Advanced Transportation*, vol. 2022, no. 1, p. 1107048, 2022. https://doi.org/10.1155/2022/1107048

[4] J. Hu, J. J. Lee, J. Bai, and S. Mao, "A simple yet efficient K-nearest neighbor-based method for high-resolution traffic time–space diagram imputation," *Transportation Research Record*, 2026. https://doi.org/10.1177/03611981261419011

[5] S. Sharma, Z. Alkhafajy, A. Choudhary, S. P. Dwivedi, S. Pandey, and P. K. Mishra, "Traffic flow prediction in smart cities using IoT and optimised KNN-based intelligent transportation system," in *2025 7th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, 2025: IEEE, pp. 1-8.

[6] M. Tolani *et al.*, "Machine learning based adaptive traffic prediction and control using edge impulse platform," *Scientific Reports*, vol. 15, no. 1, p. 17161, 2025. https://doi.org/10.1038/s41598-025-00762-4

[7] F. Wang, Y. Liang, Z. Lin, J. Zhou, and T. Zhou, "SSA-ELM: A hybrid learning model for short-term traffic flow forecasting," *Mathematics*, vol. 12, no. 12, p. 1895, 2024. https://doi.org/10.3390/math12121895

[8] S. Arunachalam, "Long-term and short-term traffic flow prediction with different weather conditions using optimized wheel-graph attention based neural network," *Future Generation Computer Systems*, vol. 176, p. 108198, 2026. https://doi.org/10.1016/j.future.2025.108198

[9] X. Ni, J. Duan, X. Li, and X. Zhang, "A novel hybrid neural network for high-accuracy vehicle-to-infrastructure network traffic prediction," *Advanced Engineering Informatics*, vol. 71, p. 104423, 2026. https://doi.org/10.1016/j.aei.2026.104423

[10] C. Ren *et al.*, "Short-term traffic flow prediction: A method of combined deep learnings," *Journal of Advanced Transportation*, vol. 2021, no. 1, p. 9928073, 2021. https://doi.org/10.1155/2021/9928073

[11] H.-S. Chiang, K.-H. Lin, H. Lin, and M.-Y. Chen, "Short-term traffic flow prediction based on deep learning approach with inter-day and intra-day strategies," *Expert Systems with Applications*, vol. 299, p. 130036, 2026. https://doi.org/10.1016/j.eswa.2025.130036

[12] C. Ren, F. Fu, C. Yin, L. Lu, and L. Cheng, "A combined model for short-term traffic flow prediction based on variational modal decomposition and deep learning," *Scientific Reports*, vol. 15, no. 1, p. 17142, 2025. https://doi.org/10.1038/s41598-025-98496-w

[13] B. Yang, S. Sun, J. Li, X. Lin, and Y. Tian, "Traffic flow prediction using LSTM with feature enhancement," *Neurocomputing*, vol. 332, pp. 320-327, 2019. https://doi.org/10.1016/j.neucom.2018.12.016

[14] S. Atef, S. Zahran, and A. Karam, "Bayesian hyperparameter optimization of GRU and LSTM models for short-term traffic flow prediction: A case study of globe roundabout in Saudi Arabia," *Applied System Innovation*, vol. 9, no. 3, p. 57, 2026. https://doi.org/10.3390/asi9030057

[15] S. M. Abdullah *et al.*, "Optimizing traffic flow in smart cities: Soft GRU-based recurrent neural networks for enhanced congestion prediction using deep learning," *Sustainability*, vol. 15, no. 7, p. 5949, 2023. https://doi.org/10.3390/su15075949

[16] M. Patel *et al.*, "FACT: Autoencoder and attention Conv-LSTM-based collaborative framework for cloud cover prediction," *IEEE Access*, vol. 12, pp. 131488-131504, 2024. https://doi.org/10.1109/ACCESS.2024.3439888

[17] K. I. M. Ata, M. K. Hassan, A. G. Ismaeel, S. A. R. Al-Haddad, T. Alquthami, and S. Alani, "A multi-Layer CNN-GRUSKIP model based on transformer for spatial– TEMPORAL traffic flow prediction," *Ain Shams Engineering Journal*, vol. 15, no. 12, p. 103045, 2024. https://doi.org/10.1016/j.asej.2024.103045

[18] M. Zhang and J. X. Cao, "Dynamic routing spatial-temporal transformer model for traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 7, pp. 10116-10130, 2025. https://doi.org/10.1109/TITS.2025.3552404

© 2026 Conscientia Beam. All Rights Reserved.

[19] V. Chauhan and A. Tiwari, "TimeResLSTM: A hybrid deep learning architecture for short-term traffic flow forecasting," *IEEE Transactions on Computational Social Systems*, vol. 15, no. 5, pp. 6258-6268, 2026. https://doi.org/10.1109/TCSS.2025.3642560

[20] C. Han and S. Song, "A review of some main models for traffic flow forecasting," in *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems, IEEE*, 2003, vol. 1, pp. 216-219.

[21] U. Ali and T. Mahmood, "Using deep learning to predict short term traffic flow: A systematic literature review," presented at the First International Conference on Intelligent Transport Systems, Springer, 2017.

[22] A.-M. Li and J. Li, "A hybrid spatiotemporal LSTM and transformer network for cellular traffic prediction," *Computer Networks*, vol. 272, p. 111669, 2025. https://doi.org/10.1016/j.comnet.2025.111669

[23] H.-T. C. Lin, H. Dai, and V. S. Tseng, "Short-term and long-term travel time prediction using transformer-based techniques," *Applied Sciences*, vol. 14, no. 11, p. 4913, 2024. https://doi.org/10.3390/app14114913

[24] Z. Zhao, Z. Li, F. Li, and Y. Liu, "CNN-LSTM based traffic prediction using spatial-temporal features," *Journal of Physics: Conference Series*, vol. 2037, no. 1, p. 012065, 2021. https://doi.org/10.1088/1742-6596/2037/1/012065

[25] N. Singh, K. Kumar, and B. Pokhriyal, "Attention based spatiotemporal model for short-term traffic flow prediction," *International Journal of System Assurance Engineering and Management*, vol. 16, no. 4, pp. 1517-1531, 2025. https://doi.org/10.1007/s13198-025-02784-7

[26] Y. Wang, J. H. van Schuppen, and J. Vrancken, "Prediction of traffic flow at the boundary of a motorway network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 214-227, 2014. https://doi.org/10.1109/TITS.2013.2278192

[27] Z. Lin, D. Wang, C. Cao, H. Xie, T. Zhou, and C. Cao, "GSA-KAN: A hybrid model for short-term traffic forecasting," *Mathematics*, vol. 13, no. 7, p. 1158, 2025. https://doi.org/10.3390/math13071158

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.

© 2026 Conscientia Beam. All Rights Reserved.