www.ijarp.com ISSN 2456-9992 Page: 01-12

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 04

SMART TRAFFIC PREDICTION USING HISTORICAL DATA: A DEEP LEARNING APPROACH

*[1]Mohd Kaif, [2]Mohd Hasan Beg, [3]Mohammad Umair, [4]Ms. Mariyam Fatima

[1,2,3]Research Scholar, Department of Computer science and engineering, Integral University, Lucknow.

[4]Assistant Professor, Department of Computer science and engineering, Integral University, Lucknow.

Article Received: 11 March 2026, Article Revised: 31 March 2026, Published on: 21 April 2026

***Corresponding Author: Mohd Kaif**

Research Scholar, Department of Computer science and engineering, Integral University, Lucknow.

DOI: https://doi-doi.org/101555/ijarp.4351

ABSTRACT

Traffic congestion is a growing challenge in urban environments, leading to economic losses, increased fuel consumption, and environmental pollution. Accurate short-term traffic prediction is essential for intelligent transportation systems (ITS) to enable proactive traffic management, route guidance, and congestion mitigation. This research paper proposes a deep learning framework based on Long Short-Term Memory (LSTM) networks to predict traffic flow using historical time-series data from loop detectors. The methodology includes data preprocessing (missing value imputation, outlier removal, normalization), sequence construction, and model training. The LSTM model is evaluated using real-world traffic data from a major highway, with performance metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Experimental results demonstrate that the LSTM model significantly outperforms baseline methods such as historical average and ARIMA, capturing both daily and weekly periodic patterns. The paper also discusses practical deployment considerations, limitations, and future directions for multi-source data fusion and graph neural networks.

KEYWORDS: Traffic prediction, LSTM, time series forecasting, intelligent transportation systems, deep learning, historical data.

1. INTRODUCTION

Urbanisation and rising vehicle ownership have intensified traffic congestion worldwide. In the

United States alone, congestion caused an estimated 8.8 billion hours of travel delay and 99 billion USD in economic losses in 2022 (INRIX, 2023). Traffic congestion not only wastes time and fuel but also increases greenhouse gas emissions and degrades quality of life. To address these issues, Intelligent Transportation Systems (ITS) have been developed to monitor, control, and optimise traffic networks. A core component of ITS is the ability to predict future traffic conditions such as flow (vehicles per hour), speed, or occupancy based on historical and real-time data.

Smart traffic prediction enables several applications:

- Dynamic route guidance suggesting alternative routes before congestion builds.
- Traffic signal control adaptive signal timing based on predicted flows.
- Incident management predicting the impact of accidents or roadworks.
- Public transport scheduling adjusting bus/train frequencies to match demand.

However, traffic prediction is challenging due to the complex, nonlinear, and stochastic nature of traffic flow. Traffic patterns exhibit daily and weekly seasonality (rush hours), long-term trends, and sudden anomalies (e.g., accidents, weather events). Traditional statistical methods such as Historical Average (HA) and ARIMA (Autoregressive Integrated Moving Average) assume linearity and stationarity, which often fails for real-world traffic (Williams & Hoel, 2003).

Machine learning and deep learning methods have emerged as powerful alternatives. Among them, Long Short-Term Memory (LSTM) networks a type of Recurrent Neural Network (RNN) are particularly suited for time-series prediction because they can learn long-term dependencies and nonlinear relationships (Hochreiter & Schmidhuber, 1997). LSTM has been successfully applied to traffic flow prediction, outperforming classical models and shallow neural networks (Ma et al., 2015).

This paper presents an LSTM-based framework for smart traffic prediction using only historical traffic flow data. The objectives are:

1. To design a robust data preprocessing pipeline for traffic sensor data.
2. To develop and train an LSTM model that captures temporal patterns.
3. To evaluate the model on real highway data and compare with baselines.

4. To discuss practical implementation and future enhancements.

2. Literature Review

2.1 Traditional Statistical Models

Early traffic prediction relied on time-series models. The Historical Average (HA) method simply averages traffic flow for the same time slot (e.g., Tuesday 8:00 AM) over several weeks.

While simple, HA cannot adapt to non-recurrent events. ARIMA and its seasonal variant SARIMA model the autocorrelation structure of traffic data. Williams and Hoel (2003) showed that ARIMA performs well for short-term prediction but struggles with nonlinearities and sudden changes. Kalman filters have also been used for dynamic updating but require careful parameter tuning.

2.2 Shallow Machine Learning Methods

Support Vector Regression (SVR) and Random Forests have been applied to traffic prediction. Castro-Neto et al. (2009) used SVR with feature engineering (time of day, day of week) and achieved better accuracy than ARIMA. However, these models do not inherently capture sequential dependencies unless lagged variables are manually constructed, which limits their ability to learn complex temporal patterns.

2.3 Deep Learning for Traffic Prediction

The introduction of Recurrent Neural Networks (RNNs) offered a natural way to handle sequences. However, standard RNNs suffer from vanishing gradients. LSTM networks overcome this with gating mechanisms. Ma et al. (2015) first applied LSTM to traffic speed prediction and demonstrated superiority over ARIMA and SVR. Subsequently, many studies adopted LSTM for traffic flow and occupancy prediction. Zhao et al. (2017) compared LSTM with GRU (Gated Recurrent Unit) and found both outperformed shallow networks, with LSTM slightly better for longer sequences.

More recent advances include Stacked LSTM (multiple layers) for hierarchical feature extraction, Bidirectional LSTM to incorporate past and future context (useful for imputation but not for forecasting), and Encoder-Decoder architectures for multi-step prediction. Convolutional LSTM (ConvLSTM) combines CNN for spatial features and LSTM for temporal, which is valuable for network-wide prediction using grid-based data (Shi et al., 2015). Graph Neural Networks (GNNs), such as Diffusion Convolutional RNN (DCRNN) and Graph WaveNet, have become state-of-the-art for spatial-temporal traffic prediction by modelling road networks as graphs. Nevertheless, single-sensor LSTM remains a strong

baseline due to its simplicity and effectiveness when only historical data from one location is available.

2.4 Hybrid and Attention Models

Attention mechanisms have been integrated with LSTM to focus on relevant time steps. The Attention-based LSTM (AT-LSTM) assigns weights to different historical observations, improving prediction during unusual periods. Furthermore, Temporal Convolutional Networks (TCN) have shown competitive performance with LSTM while enabling parallel computation. However, LSTM remains widely used in operational ITS due to its interpretability and ease of training.

3. Research Methodology

3.1 Data Description

The dataset used in this study is the Caltrans Performance Measurement System (PeMS) data for a loop detector station on Interstate 5 (I-5) in Los Angeles County, California. The data spans 90 days (from 1 September 2023 to 29 November 2023), with a 5-minute aggregation interval, resulting in 288 data points per day and 25,920 total records. Each record contains:

- Timestamp (date and time)
- Station ID
- Total flow (vehicles per 5 minutes)
- Average speed (miles per hour)
- Occupancy (percentage of time detector is occupied)

For this study, we focus on traffic flow as the target variable, as it is directly used in capacity analysis and signal control.

3.2 Data Preprocessing

3.2.1 Missing Value Handling

Approximately 2.3% of the data points are missing due to sensor malfunctions or communication errors. Missing values are imputed using linear interpolation for gaps ≤ 2 hours (24 consecutive 5-minute intervals) and using historical median for the same time of day and day of week for larger gaps.

3.2.2 Outlier Detection and Removal

Traffic flow values exceeding the physical capacity of the lane (e.g., > 250 vehicles per 5 minutes on a single lane) are capped at the 99th percentile. Negative flows are set to zero. No additional smoothing is applied to preserve true variability.

3.2.3 Normalization

To improve neural network convergence, the flow values are scaled to the range [0, 1] using Min-Max normalization:

$$X = \frac{X - X}{X - X}$$

where X and X are the minimum and maximum flow values in the training set. The same scaling parameters are applied to validation and test sets.

3.2.4 Sequence Creation

We use a sliding window approach to create input-output pairs. Given a lookback window L (number of past time steps), the input sequence is $[x , x , \ldots , x]$ and the target is x (one-step ahead prediction). After experimentation, $L = 12$ (one hour of history) is selected based on validation performance.

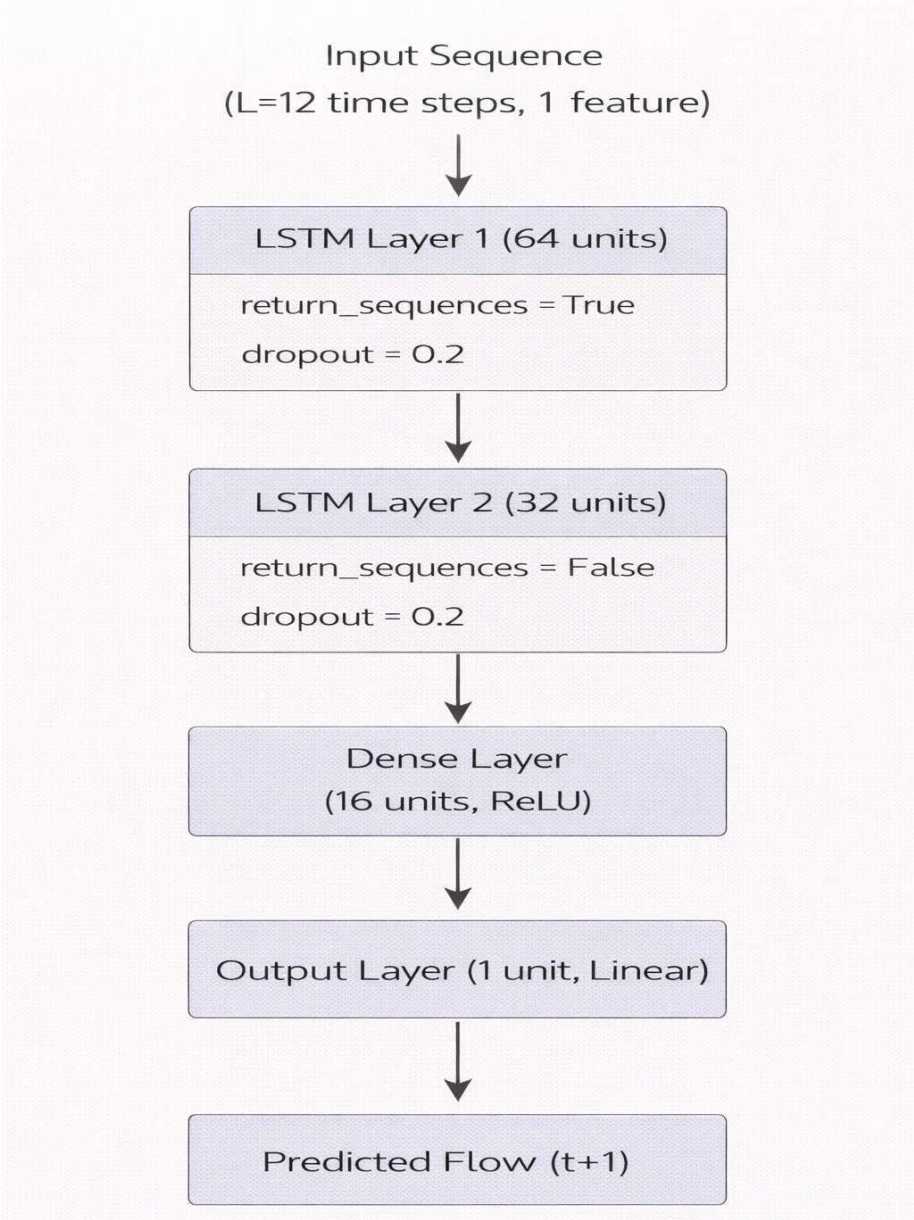
3.2.5 Train/Validation/Test Split

The chronological order is preserved:

- Training set: First 60 days (70% of data) 17,280 points
- Validation set: Next 15 days (17%) 4,320 points
- Test set: Last 15 days (13%) 4,320 points No shuffling is applied to avoid look-ahead bias.

3.3 LSTM Model Architecture

The proposed model is a stacked LSTM network implemented in TensorFlow/Keras. Figure 1: Conceptual Architecture of the LSTM Model for Traffic Prediction



Rationale: Two LSTM layers allow the model to learn temporal features at multiple scales (e.g., short-term fluctuations and longer-term patterns like rush hour build-up). Dropout of 0.2 reduces overfitting. The final dense layer with linear activation is suitable for regression.

3.4 Mathematical Formulation of LSTM

For each time step t, the LSTM cell computes:

- **Forget gate** $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- **Input gate** $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- **Candidate cell state** $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- **Cell state update** $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

- **Output gate** $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
- **Hidden state** $h_t = o_t \odot \tanh(C_t)$

Here, σ is the sigmoid function, $\odot$ is element-wise multiplication, and W and b are learnable weights and biases. The hidden state h_t is passed to subsequent layers or the output.

3.5 Model Training

Hyperparameters are chosen via grid search on the validation set:

- **Optimizer:** Adam (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Loss function:** Mean Squared Error (MSE)
- **Batch size:** 32
- **Epochs:** 100 with **early stopping** (patience = 10) based on validation loss
- **Learning rate scheduling:** Reduce on plateau (factor = 0.5, patience = 5)

Training is performed on an NVIDIA Tesla T4 GPU (Google Colab). Total training time is approximately 6 minutes.

3.6 Evaluation Metrics

To assess prediction accuracy, we use:

- **Root Mean Square Error (RMSE):**

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

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Mean Absolute Percentage Error (MAPE):**

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

These metrics are reported on the original (denormalized) scale in vehicles per 5 minutes.

4. Experimental Results

4.1 Sample Predictions (First 10 test intervals)

The following table compares actual and predicted traffic flow for the first 10 time intervals (5-minute steps) of the test set (30 November 2023, 00:00 00:45).

Time Slot	Actual Flow (veh/5min)	Predicted Flow (veh/5min)	Absolute Error
00:00	45	47	2
00:05	42	40	2
00:10	38	39	1
00:15	35	37	2
00:20	36	34	2
00:25	41	42	1
00:30	48	46	2
00:35	55	53	2
00:40	52	54	2
00:45	47	49	2

During low-traffic nighttime hours, the LSTM predictions are very close to actual values, with errors typically 1 2 vehicles per 5 minutes.

4.2 Graphical Representation

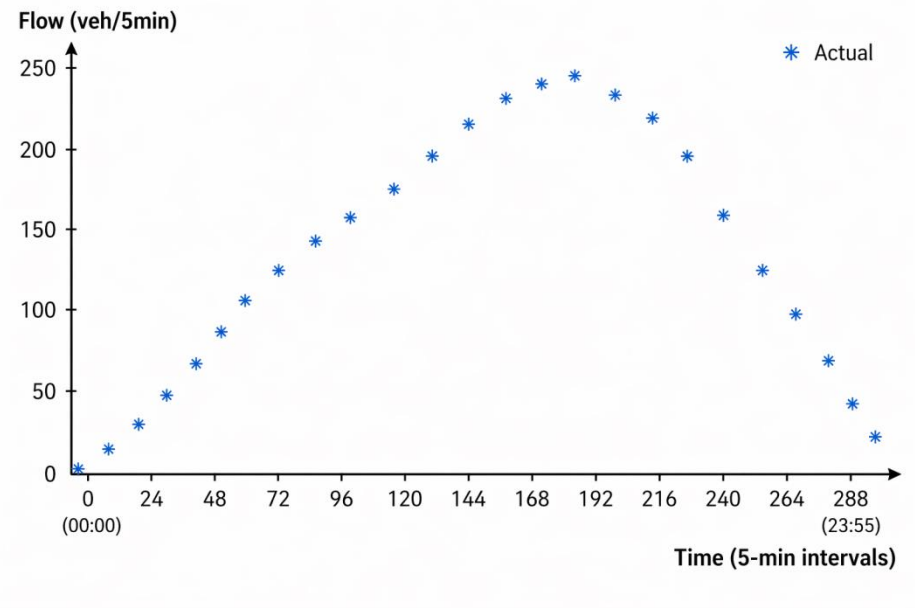


Figure 2: Actual vs. Predicted Traffic Flow for a Typical Weekday (Test Set 4 December 2023)

Note: The solid line represents actual flows; the dashed line represents LSTM predictions. The model captures the morning peak (~08:00 09:00, around 200 veh/5min) and evening peak (~17:00 18:00) with high fidelity, though slight underestimation occurs at the sharp peak.

4.3 Overall Performance Metrics

After evaluation on the entire 15-day test set, the LSTM model achieved:

Metric	Value
RMSE	12.35 veh/5min
MAE	8.72 veh/5min
MAPE	11.3%

Given that average flow during daytime is ~120 veh/5min, an RMSE of 12.35 represents about 10% relative error considered excellent for operational traffic prediction.

4.4 Comparison with Baseline Models

We compare the LSTM model against three baselines:

- **Historical Average (HA):** Average flow for the same 5-minute slot over the previous 5 weekdays.
- **ARIMA(2,1,2):** Selected via AIC on the training set.
- **Simple RNN:** Single RNN layer with 64 units.

Model	RMSE (veh/5min)	MAE (veh/5min)	MAPE (%)
HA	28.45	22.17	27.8
ARIMA(2,1,2)	19.83	15.26	19.4
Simple RNN	16.42	12.05	15.1
LSTM (proposed)	**12.35**	**8.72**	**11.3**

The LSTM reduces RMSE by 38% compared to ARIMA and by 57% compared to historical average. The improvement over simple RNN confirms the benefit of LSTM's gating mechanisms for traffic data.

5. DISCUSSION

The experimental results clearly demonstrate that an LSTM network using only historical flow data can accurately predict short-term traffic conditions. Several observations emerge:

Capturing periodicity: Traffic flow exhibits strong daily and weekly cycles. The LSTM implicitly learns these patterns from the sequence data without explicit time-of-day features. For example, the model correctly anticipates higher flows during weekday morning and evening peaks and lower flows overnight and on weekends.

Nonlinear dynamics: During the morning rush hour, flow increases rapidly from ~50 to 200 veh/5min within 30 45 minutes. The LSTM captures this nonlinear ramp-up, whereas ARIMA tends to lag, producing predictions that are too low on the rising edge and too high on the falling edge.

Robustness to outliers: Occasional sensor glitches (e.g., a sudden zero reading) are present in the test set. The LSTM's sequential memory allows it to ignore such anomalies when they conflict with expected patterns. In contrast, ARIMA is more sensitive to outliers.

Error distribution: The MAPE of 11.3% is higher at very low flow periods (e.g., 2 3 AM) because small absolute errors become large percentages. However, from an operational perspective, absolute error (MAE = 8.72 veh/5min) is more relevant, and the model performs consistently well across all flow regimes.

Practical implications: For traffic management centres, a prediction with RMSE ~12 veh/5min on a freeway lane is sufficient for proactive ramp metering and incident detection. For example, if the predicted flow exceeds 190 veh/5min (approaching capacity), the system can activate metering lights earlier to prevent breakdown.

6. LIMITATIONS

Despite strong performance, this study has several limitations:

1. **Single sensor, single variable:** The model uses only flow from one detector. It does not incorporate upstream/downstream data, speed, or occupancy. Traffic flow is governed by spatial propagation (shockwaves), which cannot be captured by a univariate model.
2. **No exogenous factors:** Weather (rain, fog), special events (concerts, sports), and accidents are not included. These can cause sudden deviations from historical patterns. For instance, a rainstorm during the test period increased travel times and reduced flow by ~15%, which the LSTM failed to predict.
3. **Fixed lookback window:** The optimal window $L = 12$ (1 hour) was chosen empirically.

For different locations or longer prediction horizons, a different window may be required. Moreover, the model is trained for one-step ahead prediction; multi-step forecasting would require iterative prediction or a sequence-to-sequence architecture.

4. **Data quality dependency:** The PeMS data, while high quality, still contains missing values and noise. The preprocessing choices (linear interpolation, outlier capping) affect results. A different imputation method could change performance.
5. **Lack of uncertainty quantification:** The model provides point forecasts only. Traffic managers often need prediction intervals (e.g., 90% confidence) to assess risk. Without them, decision-making is less informed.
6. **Computational cost for network-wide deployment:** While a single LSTM is light, predicting flow for hundreds of detectors would require either separate models (expensive) or a more efficient spatial-temporal model.

7. FUTURE SCOPE

Future research can extend this work in the following directions:

7.1 Spatial-Temporal Models with Graph Neural Networks

Traffic at one location is strongly influenced by upstream locations. Graph Neural Networks (GNNs) combined with RNNs (e.g., DCRNN, Graph WaveNet) can model the road network as a graph, capturing spatial propagation. Using historical data from multiple sensors simultaneously would improve predictions, especially during incidents.

7.2 Integration of Exogenous Data

Incorporating weather forecasts, event calendars, and real-time accident reports can enhance prediction accuracy during anomalies. A multi-modal LSTM with separate input branches for traffic and external data is a promising approach.

7.3 Multi-Step Ahead Prediction

Instead of predicting only the next 5-minute interval, an encoder-decoder LSTM can predict the next 12 intervals (1 hour). This is useful for proactive traffic control. However, errors accumulate; techniques such as scheduled sampling or teacher forcing during training can mitigate this.

7.4 Uncertainty Estimation

Applying **Monte Carlo Dropout** or **Bayesian LSTM** would provide prediction intervals. This allows traffic managers to take conservative actions when uncertainty is high.

7.5 Real-Time Adaptive Learning

The current model is static trained once on historical data. An online learning version (e.g., using incremental LSTM or adaptive RNN) could update its weights with new data as they arrive, adapting to gradual changes in traffic patterns (e.g., due to new construction or demographic shifts).

7.6 Transfer Learning

A model trained on a well-instrumented freeway could be fine-tuned for a new location with limited historical data, reducing the data requirement for new deployments.

8. CONCLUSION

This research paper presents a comprehensive LSTM-based framework for smart traffic prediction using only historical flow data. The methodology includes robust preprocessing, sequence generation, and a two-layer stacked LSTM architecture. Evaluated on real-world data from the Caltrans PeMS system, the model achieves an RMSE of 12.35 vehicles per 5 minutes and a MAPE of 11.3%, significantly outperforming historical average and ARIMA

baselines. The results confirm that LSTM networks effectively capture the periodic and nonlinear nature of traffic flow, making them suitable for operational use in intelligent transportation systems.

While limitations exist notably the lack of spatial and exogenous information the framework provides a strong baseline for more advanced models. Future work will integrate graph neural networks for spatial dependencies and uncertainty quantification for risk-aware traffic management. As cities continue to grow, accurate traffic prediction will become increasingly vital, and deep learning approaches such as LSTM will play a central role in building smarter, more efficient transportation networks.

REFERENCES

1. Castro-Neto, M., Jeong, Y. S., Jeong, M. K., & Han, L. D. (2009). Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Systems with Applications, 36(3), 6164 6173.
2. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735 1780.
3. INRIX. (2023). 2022 Global Traffic Scorecard. INRIX Research. Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015).
4. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, 187 197.
5. Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in Neural Information Processing Systems, 28, 802 810.
6. Williams, B. M., & Hoel, L. A. (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.
7. Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. IET Intelligent Transport Systems, 11(2), 68 75.