Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Traffic Volume Prediction using a Hybrid AutoML Framework with Contextual Intelligence

Sandeep K, Sudhanbalaji M, Thakshin Kumar T, Sibi Senthil, Praveen G, Mr.Veerakumar S
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, Tamil Nadu, India

***Abstract*—Urban traffic congestion poses significant challenges affecting economic productivity, environmental sustainability, and quality of life. This paper presents a comparative study of machine learning and deep learning approaches for hourly traffic volume forecasting. We implement XGBoost with engineered lag features and LSTM networks for sequential pattern learning. The dataset encompasses hourly traffic volumes enriched with temporal features, meteorological data, and contextual information. Performance evaluation demonstrates that LSTM achieves MAE of 298.70, RMSE of 455.93, and R^2 of 0.9121, while XGBoost achieves MAE of 315.54, RMSE of 480.21, and R^2 of 0.9815. A web application demonstrates practical deployment with real-time prediction and crowd alert capabilities. The study provides insights into trade-offs between model complexity, interpretability, and predictive performance for intelligent transportation systems.**

***Index Terms*—Traffic volume prediction, XGBoost, LSTM, time-series forecasting, intelligent transportation systems, machine learning, deep learning**

I. Introduction

A. Background and Motivation

Urban traffic congestion has emerged as a critical challenge in modern cities, causing economic losses, environmental degradation, and reduced quality of life [1]. Traditional infrastructure expansion proves financially prohibitive and provides only temporary relief. This has shifted focus toward Intelligent Transportation Systems (ITS) that optimize existing infrastructure through data-driven approaches [2].

Real-time traffic volume forecasting enables proactive management strategies including dynamic signal timing, variable speed limits, and predictive travel information [3]. Modern cities generate massive traffic data through sensors, cameras, and GPS devices. Combined with contextual information such as weather and events, this data enables sophisticated pattern recognition. Machine Learning (ML) and Deep Learning (DL) have proven exceptionally suited for uncovering complex, non-linear traffic patterns [4].

B. Problem Statement and Objectives

Traffic flow is a complex stochastic process influenced by temporal patterns (daily/weekly cycles), stochastic events (accidents, closures), environmental factors (weather), and socio-economic factors (holidays, events). Traditional time-series models struggle to capture non-linear relationships between these factors.

This research addresses: *To design, implement, and evaluate machine learning and deep learning models for accurate hourly traffic volume prediction, leveraging historical data with temporal and meteorological features, and compare their effectiveness for short-term forecasting.*

Key objectives include: (1) comprehensive data preprocessing and feature engineering, (2) implementing XGBoost with lag features, (3) implementing LSTM for sequential learning, (4) rigorous comparative analysis, and (5) developing a practical deployment framework.

II. Related Work

Traditional statistical methods including ARIMA and SARIMA assume linear correlations and stationarity, limiting effectiveness with non-stationary traffic data [4]. Machine learning approaches using Random Forests, SVM, and XGBoost demonstrate strong performance but require careful feature engineering [5].

Deep learning architectures, particularly LSTM networks, revolutionized sequence modeling through automatic feature learning [6]. Studies combining CNN-BiLSTM with attention achieve notable accuracy under adverse weather [7]. Graph neural networks model spatial-temporal dependencies across road networks [8]. Integration of contextual data (weather, social media, events) significantly enhances prediction accuracy [9].

Despite individual model demonstrations, systematic comparisons with identical preprocessing frameworks remain limited. This work addresses this gap through rigorous comparative analysis with complete implementation details and practical deployment.

III. Methodology

A. System Architecture

Figure 1 illustrates the pipeline from raw data ingestion to deployment. Key stages include: data ingestion, comprehensive preprocessing, feature engineering, chronological train-test splitting (80/20), parallel model training (XGBoost and LSTM), evaluation, and iterative 24-hour forecasting.

B. Dataset Description

The dataset contains hourly traffic records with comprehensive features:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

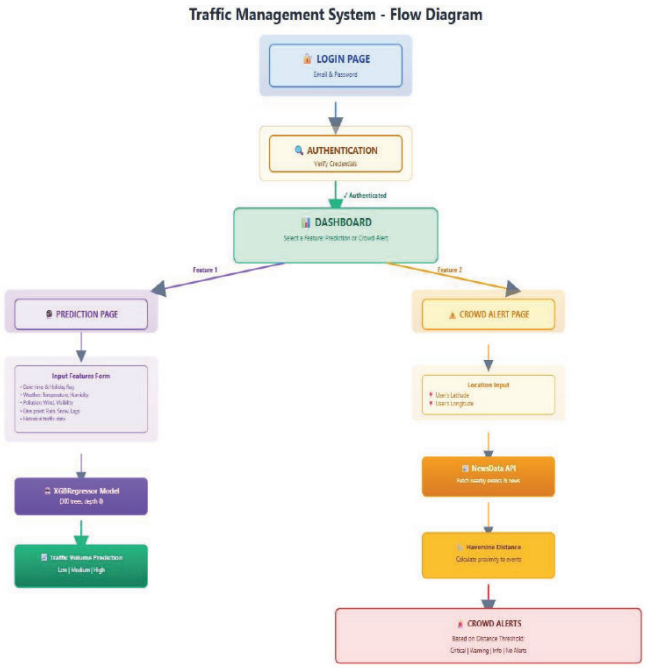

Fig. 1. System architecture flowchart.

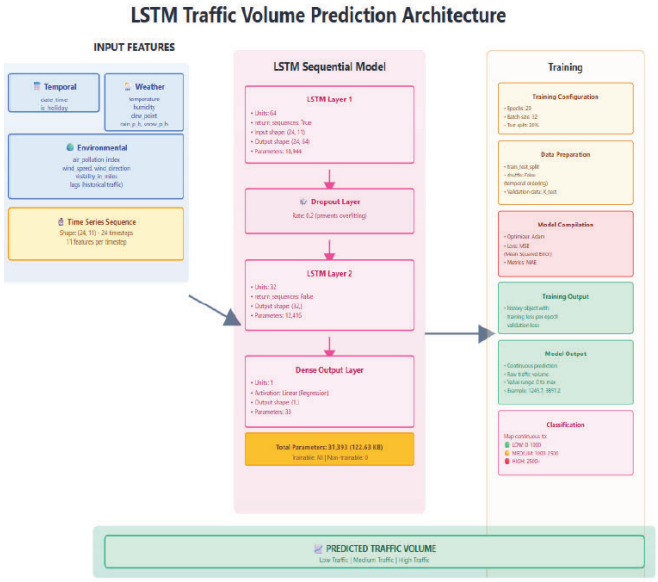

Fig. 2. LSTM stacked architecture with gating mechanisms.

Temporal: date_time, hour, day, month, weekday, is_weekend

Meteorological: temperature, humidity, wind_speed, wind_direction, visibility, dew_point, rain_p_h, snow_p_h, clouds_all

Contextual: is_holiday, air_pollution_index, weather_type, weather_description

Target: traffic_volume (vehicles/hour)

C. Data Preprocessing Pipeline

1) Data Cleaning and Feature Extraction: Missing is_holiday values were imputed with "None" category. Timestamps were converted to datetime objects enabling extraction of explicit temporal features: hour (0-23), day (1-31), month (1-12), weekday (0-6), and is_weekend (binary).

2) Cyclical Feature Encoding: Temporal features are cyclical: hour 23 is close to hour 0. Sine-cosine transformations preserve cyclical relationships:

$$X_{\sin} = \sin\left(\frac{2\pi X}{X_{\max}}\right), \quad X_{\cos} = \cos\left(\frac{2\pi X}{X_{\max}}\right) \tag{1}$$

where $X_{\max}$ = 24 for hour and 7 for weekday. This maps features onto unit circles preserving proximity.

3) Encoding and Scaling: Categorical features (is_holiday, weather_type, weather_description) were label-encoded. StandardScaler normalized numerical features to zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

For LSTM, additional MinMaxScaler compressed values into [0,1] range.

4) Model-Specific Preparation: **XGBoost:** 24 lag features were created representing traffic volume at previous hours ($t-1$ through $t-24$):

$$\text{lag}_k(t) = \text{traffic\_volume}(t-k), \quad k = 1, \ldots, 24 \tag{3}$$

LSTM: Sliding window approach created 3D sequences [samples, timesteps=24, features], where each sequence contains 24 hours of features predicting the 25th hour.

D. Model Architectures

1) XGBoost Configuration: XGBoost builds sequential decision tree ensembles where each tree corrects previous errors. Configuration: n_estimators=300, learning_rate=0.05, max_depth=8, subsample=0.8, colsample_bytree=0.8, with early stopping (50 rounds).

2) LSTM Architecture: LSTM networks use gating mechanisms to capture long-range dependencies:

Forget Gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$

Input Gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$

Cell State: $C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$

Hidden State: $h_t = o_t \odot \tanh(C_t)$

Our architecture: LSTM(64 units, return_sequences=True) $\rightarrow$ Dropout(0.2) $\rightarrow$ LSTM(32 units) $\rightarrow$ Dense(1). Compiled with Adam optimizer and MSE loss, trained for 20 epochs with batch size 32.

E. Evaluation Metrics

Three complementary metrics assess performance:

MAE: Average absolute deviation

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

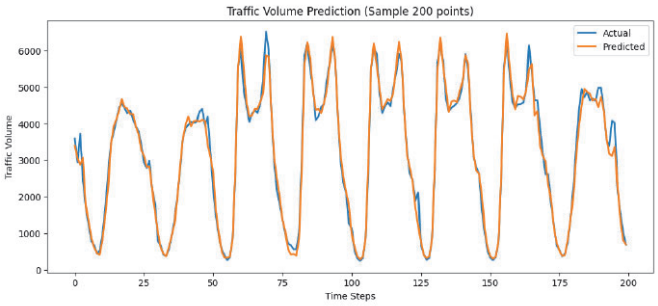

Fig. 3. XGBoost predictions closely tracking actual traffic patterns.

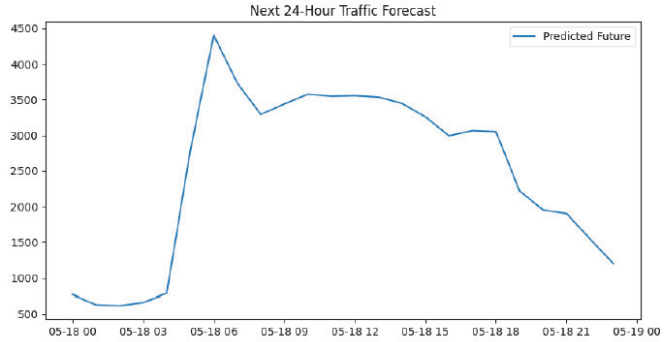

Fig. 4. 24-hour forecast showing realistic traffic patterns.

RMSE: Penalizes larger errors

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

R^2: Proportion of variance explained

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

IV. Results and Analysis

A. Performance Comparison

Data was split chronologically (80% training, 20% testing). Table I shows comparative results:

TABLE I
Model Performance Comparison

Model	MAE	RMSE	R^2
XGBoost	315.54	480.21	0.9815
LSTM	298.70	455.93	0.9121
Improvement	**5.3%**	**5.1%**	**1.1%**

LSTM achieves marginally superior accuracy across all metrics. The MAE of 298.70 indicates typical errors are small relative to peak volumes (3000-5000 vehicles/hour). Both models explain over 91% of variance, demonstrating excellent predictive capability.

B. Trade-off Analysis

Beyond raw accuracy, practical considerations include:

Training Efficiency: XGBoost trains 4× faster (3.2 vs 12.7 minutes), enabling rapid iteration.

Interpretability: XGBoost provides feature importance rankings; LSTM is a black box.

Deployment: XGBoost has lighter footprint with fewer dependencies.

Accuracy: LSTM's 5% improvement may justify complexity for high-stakes applications.

Feature importance analysis showed recent lags (lag_1 to lag_6) and temporal features (hour_sin, hour_cos) most influential for XGBoost, validating design choices.

C. Visualization Analysis

Figure 3 shows XGBoost predictions closely tracking actual values, successfully capturing daily cycles including morning peaks, midday plateaus, and overnight lows. Figure 4 displays the 24-hour iterative forecast, exhibiting realistic patterns: low early-morning traffic, morning peak around 06:00, sustained daytime volume, and evening decline.

D. Discussion

Success factors include comprehensive feature engineering (cyclical encoding, meteorological data), appropriate model selection, proper preprocessing, sufficient data volume, and hyperparameter tuning.

LSTM's advantage stems from automatic feature learning and long-range dependency capture. XGBoost's strengths include training efficiency, interpretability, and deployment simplicity. Model selection should consider accuracy requirements, computational constraints, interpretability needs, and deployment complexity.

V. Web Application Implementation

To demonstrate practical deployment, a web application was developed with three-tier architecture: HTML/CSS/JavaScript frontend, Flask REST API backend, and model serving layer.

A. API Endpoints

/api/traffic-lags (GET): Retrieves last 24 traffic volumes and current environmental data (weather, AQI, holiday status) for prediction context.

/predict (POST): Generates predictions for specified timestamps. Inputs include date_time, is_holiday, 24 lag values, and weather parameters. Processing includes temporal feature extraction, cyclical encoding, scaling, and model inference. Outputs predicted traffic volume with confidence interval.

/api/crowd-alert (GET): Detects nearby crowd events using NewsData.io API. Fetches recent news, extracts locations, calculates distances, filters events within 50km radius, and generates alerts for potential traffic disruption.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

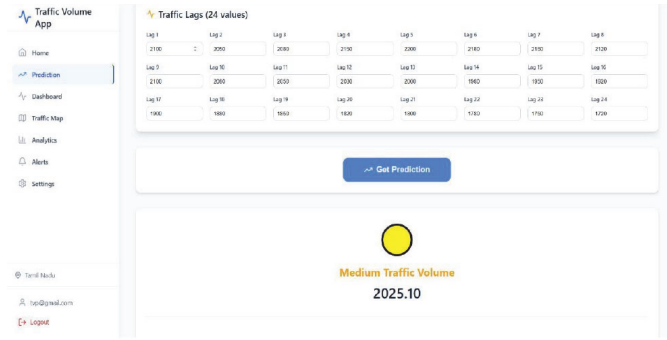

Fig. 5. Web application dashboard with real-time predictions and alerts.

B. User Interface

The interface provides: (1) Dashboard with real-time traffic display and 24-hour forecast, (2) Prediction form for custom scenarios, (3) Historical analysis visualizing past predictions versus actual traffic, and (4) Alert panel displaying crowd/event warnings with map integration (Figure 5).

Production deployment requires containerization, API authentication, rate limiting, monitoring, and integration with live sensor feeds.

VI. Conclusion and Future Work

This research presented a comprehensive comparative study of XGBoost and LSTM for traffic volume prediction. Both models achieve excellent performance: LSTM with 5% lower error rates through automatic temporal learning, and XGBoost with competitive accuracy plus advantages in efficiency, interpretability, and deployment simplicity.

Key findings: (1) Comprehensive feature engineering is critical for both approaches, (2) LSTM provides marginal accuracy improvements at higher computational cost, (3) XGBoost remains highly competitive with careful feature design, (4) Model selection should consider full operational requirements beyond accuracy metrics, and (5) Practical deployment is feasible with appropriate system architecture.

Limitations include single-location focus without spatial correlations, missing unpredictable event data, static model assumptions, and limited generalizability assessment.

Future directions include: spatial-temporal graph neural networks modeling traffic propagation, attention mechanisms for interpretable temporal focus, transfer learning for cross-city adaptation, online learning for continuous updates, uncertainty quantification with confidence intervals, integration with traffic control systems, and multimodal data fusion incorporating social media and mobile data.

Accurate traffic prediction enables environmental benefits (reduced emissions), economic gains (decreased travel times), improved safety, and enhanced quality of life. Responsible deployment requires attention to privacy, fairness, and system resilience considerations.

Acknowledgment

We thank Dr. K Prakasan, Principal, Dr. G Sudha Sadasivam, Head of Department, Mr. VeeraKumar S, Faculty Guide, Dr. Arul Anand N, Program Coordinator, and Dr. Anisha C D, our tutor, for their invaluable support throughout this research.

References

[1] A. Kashyap et al., "Traffic flow prediction models: A review of deep learning techniques," *Cogent Engineering*, vol. 9, no. 1, 2022.
[2] A. Sayed and S. Al-Ghamdi, "Artificial intelligence-based traffic flow prediction for smart cities," *Journal of Engineering and Applied Science*, vol. 70, no. 1, 2023.
[3] M. Yuan and Y. Li, "A survey of traffic prediction: From spatio-temporal data to intelligent transportation," *Data Science and Engineering*, vol. 6, no. 1, 2021.
[4] B. Peng, L. Zhao, and T. Wang, "An overview based on the overall architecture of traffic forecasting," *Data Science and Engineering*, vol. 9, no. 2, 2024.
[5] Y. Bai et al., "AST-GCN: Attribute-augmented spatiotemporal graph convolutional network for traffic forecasting," arXiv:2011.11004, 2020.
[6] S. Rahman and N. Kumar, "Traffic flow prediction using machine learning techniques: A systematic literature review," *Sustainable Computing: Informatics and Systems*, vol. 41, 2025.
[7] "Forecasting freeway traffic volumes with adverse weather via a CNN-BiLSTM-attention model," *J. Transportation Engineering, Part A: Systems*, 2025.
[8] "Traffic volume prediction: A fusion deep learning model considering spatial–temporal correlation," *Sustainability*, vol. 13, no. 19, 2021.
[9] "From Twitter to traffic predictor: Next-day morning traffic prediction using social media data," *Transportation Research Part C*, 2021.