JAISEN: Journal of Advanced Information Systems and Engineering
Vol. 1, 2 (February, 2026), pp 32-41
ISSN: - EISSN: - , DOI: -

Hybrid K-Means-LSTM Model for Traffic Volume Prediction on Pekanbaru Arterial Roads

Maya Ramadhani[1], Susandri Susandri[2], Ahmad Zamsuri[3]

[1] Master of Computer Science, Lancang Kuning University , maya@unilak.ac.id
[2] Master of Computer Science, Lancang Kuning University , susandri@unilak.ac.id
[3] Master of Computer Science, Lancang Kuning University , ahmadzamsuri@unilak.ac.id

ARTICLE INFO

Keywords:

Traffic Congestion;
Time Series Prediction;
K-Means Clustering;
LSTM;
Hybrid Model.

Article history:

Received 2026-01-21
Revised 2026-02-12
Accepted 2026-02-25

ABSTRACT

Traffic congestion remains a critical issue in rapidly growing urban areas, including Pekanbaru City, Indonesia. Accurate traffic volume prediction is essential to support effective traffic management and proactive decision-making. This study proposes a hybrid clustering–sequence model by integrating K-Means Clustering and Long Short-Term Memory (LSTM) to improve urban traffic volume prediction on arterial roads. Hourly traffic volume data collected from Jalan Jenderal Sudirman were used as the primary indicator of congestion due to their strong relationship with traffic density and road capacity utilization. The research framework consists of data preprocessing, traffic pattern clustering using K-Means, and time-series prediction using LSTM, where cluster labels are incorporated as additional input features through one-hot encoding. The optimal number of clusters was determined using the Elbow and Silhouette methods, while prediction performance was evaluated using error-based metrics. The experimental results demonstrate that the hybrid K-Means-LSTM model outperforms the standalone LSTM model, particularly during peak traffic periods, by reducing prediction error and improving temporal pattern recognition. Furthermore, the clustering results provide meaningful interpretations of traffic conditions, categorized into low, medium, and high congestion levels. These findings indicate that integrating traffic pattern segmentation into sequence learning enhances both predictive accuracy and model interpretability. The proposed approach offers practical insights for data-driven urban traffic management and supports the development of intelligent transportation systems.

Corresponding Author:
Maya Ramadhani
Master of Computer Science, Lancang Kuning University; maya@unilak.ac.id

1. INTRODUCTION

Urban traffic congestion has become a persistent challenge in many developing cities as a consequence of rapid population growth, urbanization, and the increasing number of private vehicles (Subair, Ibitoye, & Kuranga, 2024). In Pekanbaru City, traffic congestion frequently occurs on major arterial roads, particularly Jalan Jenderal Sudirman, which functions as a key urban corridor connecting government offices, business districts, and educational institutions (Hasan, Albafery, & Mulyadi, 2025). High traffic volume during peak hours not only reduces travel efficiency but also contributes to increased fuel consumption, air pollution, and broader economic costs associated with delays and reduced productivity (T. Li, Song, & Yang, 2022).

In the context of urban transportation systems, traffic volume is widely recognized as a fundamental indicator for assessing congestion levels and roadway performance (Khadka, Li, & Wang, 2022). Unlike speed-based indicators, traffic volume directly reflects demand intensity and the level of infrastructure utilization, making it particularly relevant for traffic control, planning, and policy formulation. Accurate traffic volume prediction enables transportation authorities to anticipate congestion patterns, optimize traffic signal timing, and support proactive traffic management strategies, especially on arterial roads that serve as primary mobility corridors in urban areas (Chawla et al., 2024).

Recent advancements in intelligent transportation systems (ITS) have encouraged the adoption of machine learning and deep learning techniques for traffic prediction (Yuan et al., 2022). Among these approaches, recurrent neural networks, particularly Long Short-Term Memory (LSTM) models, have demonstrated strong performance in modeling temporal dependencies inherent in traffic data (Khan, Fouda, Do, Almaleh, & Rahman, 2023). LSTM-based models are capable of capturing sequential patterns and long-term dependencies, making them suitable for time-series forecasting tasks (Kumar, Tripathi, & Singh, 2023). However, urban traffic flow characteristics are inherently heterogeneous, as traffic conditions during peak hours differ substantially from those observed during off-peak periods (Anna, Chand, Alsultan, & Dixit, 2026).

Previous studies have widely applied time-series prediction models such as LSTM and Bidirectional LSTM (BiLSTM) to forecast traffic conditions, showing promising results in learning general traffic trends(Abduljabbar, Dia, & Liyanage, 2025). Nevertheless, most existing approaches construct prediction models directly from historical data without explicitly accounting for the variability of traffic patterns across different congestion states (Sayed, Abdel-Hamid, & Hefny, 2023). As a result, prediction performance may degrade when traffic dynamics vary significantly between peak and non-peak conditions, particularly when abrupt changes occur due to commuting behavior and fluctuations in road usage intensity (S. Li, Magli, Francini, & Ghinamo, 2024).

On the other hand, clustering-based approaches, such as K-Means, have been proven effective in identifying traffic patterns and categorizing congestion levels based on data similarity (Gupta, Kumar, & Kumar, 2025). These methods enable the segmentation of traffic data into meaningful groups that represent different traffic states. However, clustering techniques are typically employed for descriptive analysis and do not inherently provide predictive capability (Jaeger & Banks, 2023). This limitation highlights a research gap between traffic pattern identification and temporal traffic prediction.

To address this gap, this study proposes a hybrid K-Means–LSTM framework that integrates traffic pattern clustering into time-series prediction. The main contribution of this study lies in utilizing cluster labels as contextual features for the LSTM model, allowing it to learn temporal patterns under different traffic conditions more effectively. By combining unsupervised traffic pattern discovery with sequence-based prediction, the proposed approach is expected to improve prediction accuracy while providing interpretable insights into urban traffic dynamics, thereby supporting data-driven traffic management and congestion mitigation strategies.

2. METHODS

This study adopts a quantitative applied research approach by utilizing data mining and deep learning techniques to analyze and predict urban traffic dynamics. The dataset consists of 812 hourly traffic observations collected from Jalan Jenderal Sudirman, Pekanbaru, covering the observation period from November 2024 to May 2025. The hourly aggregation of data enables the model to capture temporal variations in traffic flow while maintaining sufficient granularity for time-series analysis. Traffic volume was selected as the primary variable to represent congestion levels, as it directly reflects roadway demand intensity and provides a reliable indicator of congestion conditions on arterial roads.

The selection of Jalan Jenderal Sudirman as the study location is based on its strategic function as one of the main arterial corridors in Pekanbaru City, serving as a critical link between governmental, commercial, and residential areas. This corridor accommodates a heterogeneous mix of traffic, including private vehicles, public transportation, and commercial traffic, resulting in substantial variability in traffic volume across different time periods. Such variability introduces diverse traffic patterns between peak and off-peak hours, making the location particularly suitable for evaluating the robustness of time-series prediction models under heterogeneous traffic conditions. Consequently, the characteristics of this roadway provide a representative and challenging environment for assessing the effectiveness of the proposed hybrid prediction approach.

2.1 Data Preprocessing

Data preprocessing includes several essential steps to ensure data quality and suitability for subsequent analysis and modeling. The process begins with data cleaning, which involves removing missing values and duplicate records to eliminate inconsistencies that may negatively affect model performance. This step ensures that the dataset accurately represents actual traffic conditions and provides a reliable foundation for further analysis.

Subsequently, temporal feature transformation is applied by extracting time-related attributes such as date, day of the week, and hourly intervals. These temporal features enable the models to capture recurring daily and weekly traffic patterns that are inherent in urban traffic systems. Finally, Min–Max normalization is performed to scale numerical variables into a uniform range, preventing variables with larger magnitudes from dominating the learning process. This normalization step is particularly important for distance-based clustering algorithms and neural network training, as it contributes to improved model convergence and stability.

Table 1. Dataset characteristics.

Attribute	Description	Type
Date	Observation date	Temporal
Day	Day of the week	Categorical
Time Interval	Hourly interval	Temporal
Vehicle Volume	Number of vehicles	Numeric
Avg Speed	Average vehicle speed	Numeric

Table 1 summarizes the main attributes of the traffic dataset used in this study, including temporal, categorical, and numerical variables that characterize traffic conditions on Jalan Jenderal Sudirman. The attributes presented in the table provide a structured overview of the data components employed in the analysis and highlight the role of each variable in capturing temporal variations and traffic volume characteristics. This summary supports the data preprocessing and modeling stages by clarifying the types of information utilized for traffic pattern segmentation and time-series prediction.

2.2 K-Means Clustering

K-Means Clustering was applied to segment traffic volume patterns into several distinct congestion levels based on similarities in traffic characteristics. This unsupervised clustering approach enables the identification of inherent traffic states without prior labeling, allowing the data to naturally form groups that represent different congestion conditions. To ensure the reliability and interpretability of the clustering results, the optimal number of clusters was determined using the Elbow Method and Silhouette Score, which together provide complementary perspectives on cluster compactness and separation. The Elbow Method was employed to identify the point at which increasing the number of clusters yields diminishing improvements in within-cluster variance, while the Silhouette Score was used to evaluate the degree of separation between clusters. The combined use of these evaluation criteria ensures that the selected clustering configuration balances model simplicity, clustering quality, and interpretability, thereby providing meaningful traffic pattern segmentation for subsequent predictive analysis.

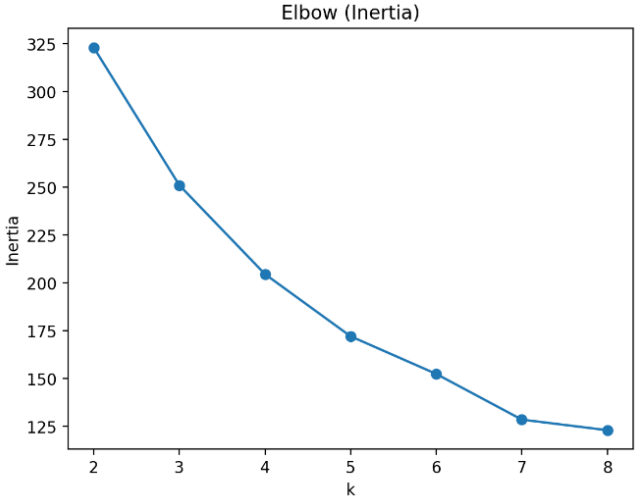

Figure 1. Elbow method for optimal cluster selection.

Figure 1 illustrates the application of the Elbow Method used to determine the optimal number of traffic clusters in the clustering process. The figure shows the relationship between the number of clusters and the within-cluster variance, where a noticeable change in the rate of variance reduction indicates the most appropriate cluster configuration. This point, commonly referred to as the "elbow," represents a balance between model simplicity and clustering effectiveness. The use of the Elbow Method supports the selection of a cluster number that adequately captures traffic pattern variability while avoiding unnecessary model complexity, thereby contributing to meaningful and interpretable traffic pattern segmentation.

2.3 LSTM Prediction and Hybrid Integration

An LSTM (Long Short-Term Memory) model was developed to predict future traffic volume by learning temporal dependencies from historical traffic data. LSTM is particularly well-suited for time-series prediction tasks due to its ability to capture long-term dependencies and mitigate the vanishing gradient problem commonly encountered in traditional recurrent neural networks. In this study, the standalone LSTM model serves as a baseline to evaluate the effectiveness of incorporating additional contextual information derived from traffic pattern segmentation.

For the proposed hybrid model, cluster labels generated by the K-Means algorithm were transformed using one-hot encoding and incorporated as additional input features to the LSTM network. This integration enables the model to condition its sequential learning process on traffic pattern categories, allowing it to differentiate between varying congestion states. By embedding

clustering-based contextual information into the input representation, the LSTM model is provided with an enriched feature space that reflects both temporal dynamics and traffic state characteristics.

The integration of K-Means clustering with LSTM is motivated by the need to incorporate contextual traffic patterns into sequential learning. By grouping traffic observations into clusters representing different congestion levels, the hybrid model is supplied with structured information regarding traffic states that would otherwise be implicitly learned. The use of one-hot encoded cluster labels allows the LSTM model to distinguish temporal dependencies across varying traffic regimes, thereby enhancing its ability to generalize across peak and non-peak traffic conditions. As a result, the hybrid approach improves model adaptability to heterogeneous traffic dynamics while maintaining the temporal modeling strengths of LSTM.

3. FINDINGS AND DISCUSSION

3.1 Traffic Pattern Clustering Results

The clustering process successfully grouped traffic data into three distinct clusters representing low, medium, and high congestion levels. The results indicate clear temporal patterns, with high congestion predominantly occurring during morning and evening peak hours.

The clustering results further reveal that traffic conditions on Jalan Jenderal Sudirman exhibit strong temporal regularities. High congestion clusters are consistently associated with morning and evening commuting hours, while low congestion clusters dominate late-night and early-morning periods. This temporal consistency indicates that the clustering process successfully captures meaningful traffic states rather than arbitrary groupings, thereby validating the use of clustering as a preliminary analytical step.

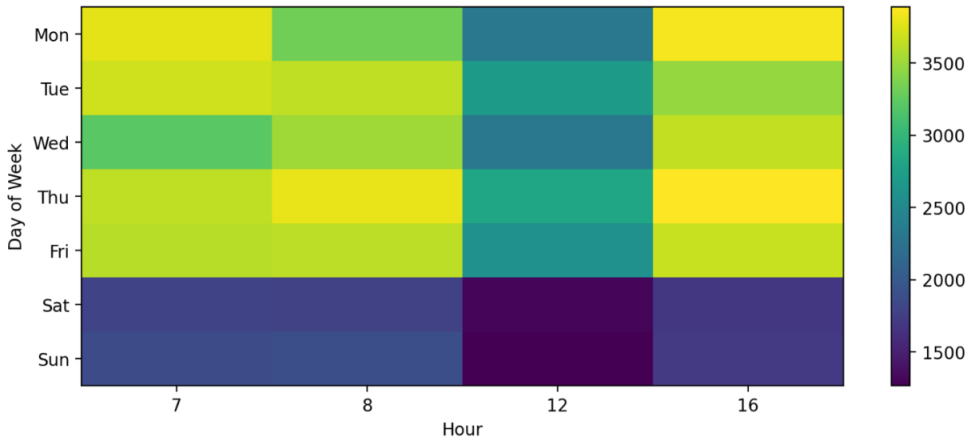

Figure 2. Heatmap of average traffic volume by hour and day.

Figure 2 illustrates the temporal distribution of traffic volume across different hours and days, clearly highlighting peak congestion periods on Jalan Jenderal Sudirman. The heatmap visualization reveals consistent patterns of increased traffic volume during morning and evening commuting hours, while lower traffic intensity is observed during late-night and early-morning periods. This temporal variation indicates the presence of recurrent daily traffic patterns and confirms the heterogeneity of traffic conditions over time. The visualization supports the clustering results by demonstrating that traffic congestion is not uniformly distributed but is strongly influenced by temporal factors, providing an intuitive representation of traffic dynamics for subsequent predictive analysis.

3.2 Performance of Standalone LSTM

The standalone LSTM model was able to follow the general trend of traffic volume fluctuations, indicating its capability to learn temporal patterns from historical data. However, the model showed notable limitations in capturing abrupt changes that frequently occur during peak traffic periods. Although the standalone LSTM demonstrates effectiveness in modeling overall traffic trends, its performance becomes constrained when traffic conditions change rapidly within short time intervals, which is a common characteristic of urban traffic environments.

These limitations suggest that temporal dependency alone may not be sufficient to fully characterize complex urban traffic dynamics. During peak hours, traffic volume is strongly influenced by external and contextual factors such as work schedules, commuting behavior, and variations in road usage intensity, which may introduce sudden fluctuations that are difficult to anticipate using historical sequences alone. As a result, the standalone LSTM model may exhibit reduced responsiveness to high-variance traffic conditions, highlighting the need for additional contextual information to improve predictive robustness under heterogeneous traffic regimes.

3.3 Performance of Hybrid K-Means-LSTM

The hybrid K-Means–LSTM model demonstrated improved prediction accuracy and greater stability compared to the standalone LSTM model. By incorporating cluster information derived from traffic pattern segmentation, the hybrid approach enables the model to more effectively distinguish traffic dynamics across different congestion conditions. This additional contextual information allows the prediction process to adapt according to varying traffic states, rather than relying solely on historical temporal dependencies.

In contrast to the standalone LSTM, the hybrid model benefits from the explicit representation of traffic pattern information, which enhances its ability to respond to rapid changes in traffic volume. The model exhibits improved stability and responsiveness during high-variance periods, such as peak traffic hours, indicating that clustering-based contextualization strengthens the learning capability of the LSTM network. These results suggest that combining unsupervised pattern discovery through clustering with sequence-based prediction provides a more robust framework for modeling complex urban traffic dynamics. Consequently, the hybrid K-Means–LSTM approach highlights the importance of integrating traffic state awareness into time-series prediction models for urban traffic analysis.

3.4 Comparative Analysis

Comparative evaluation confirms that the hybrid K-Means-LSTM model consistently produces lower prediction error, particularly during high-variance traffic conditions. This result supports the hypothesis that traffic pattern segmentation enhances sequence-based prediction.

The comparative results emphasize that the performance improvement achieved by the hybrid model is not merely numerical but also conceptual. By explicitly incorporating traffic state information, the hybrid approach aligns prediction mechanisms with real-world traffic behavior. This alignment contributes to improved interpretability, as prediction outcomes can be associated with specific traffic conditions, supporting more informed decision-making for traffic management authorities.

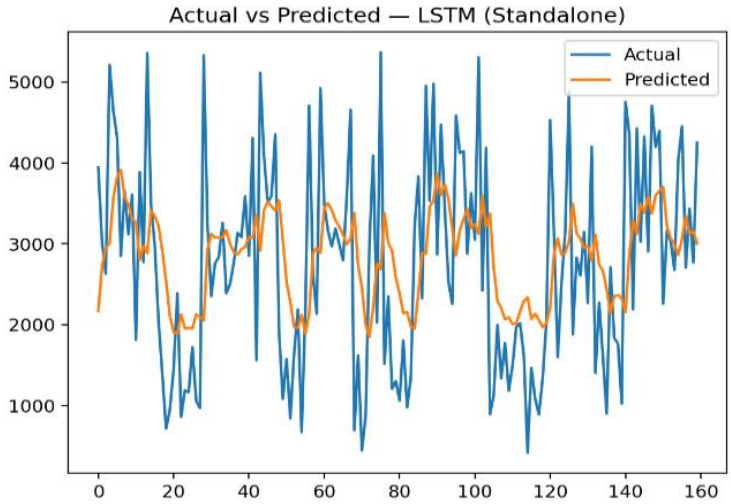

Figure 3. Traffic volume prediction using standalone LSTM.

Figure 3 illustrates the traffic volume prediction results obtained using the standalone LSTM model compared with the actual observed traffic volume. The figure shows that the LSTM model is able to capture the overall temporal trend of traffic volume fluctuations, indicating its effectiveness in learning general patterns from historical data. The predicted values follow the direction of the actual traffic volume, particularly during periods with relatively stable traffic conditions.

However, discrepancies between the predicted and actual values become more apparent during peak traffic periods, where traffic volume exhibits rapid and abrupt changes. In these high-variance intervals, the standalone LSTM model tends to smooth extreme fluctuations, resulting in delayed or less responsive predictions. This behavior suggests that while the LSTM model effectively models long-term temporal dependencies, it encounters limitations in adapting to sudden changes in traffic dynamics that are strongly influenced by contextual factors beyond historical sequences alone.

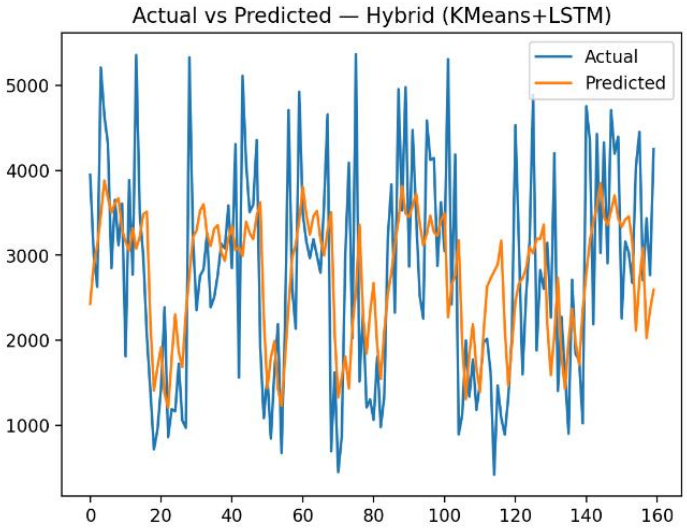

Figure 4. Traffic volume prediction using hybrid k-means-LSTM

Figure 4 presents the traffic volume prediction results generated using the hybrid K-Means–LSTM model in comparison with the actual observed traffic volume. The figure demonstrates that the hybrid model is able to more closely follow fluctuations in traffic volume across different time intervals, particularly during periods characterized by high variability. The predicted values exhibit improved

alignment with the actual traffic data, indicating enhanced responsiveness to changes in traffic conditions.

Compared to the standalone LSTM model, the hybrid K-Means–LSTM approach shows greater stability and adaptability, especially during peak traffic periods where sudden changes in volume are more prominent. The incorporation of clustering-based traffic pattern information allows the model to adjust its prediction behavior according to congestion levels, thereby reducing prediction lag and smoothing errors observed in the standalone model. These results highlight the effectiveness of integrating traffic pattern segmentation into sequence-based prediction, supporting the hybrid model's ability to better capture complex urban traffic dynamics.

Figures 3 and 4 collectively illustrate the comparative performance between the standalone LSTM model and the hybrid K-Means–LSTM model in predicting traffic volume. While Figure 3 shows that the standalone LSTM is able to capture the general trend of traffic volume, noticeable deviations occur during periods of rapid fluctuation, particularly in peak traffic intervals. These deviations indicate limitations in the model's ability to respond to abrupt changes when relying solely on historical temporal dependencies.

In contrast, Figure 4 demonstrates that the hybrid K-Means–LSTM model achieves closer alignment with the actual traffic volume across varying traffic conditions. The improved correspondence between predicted and observed values, especially during high-variance periods, highlights the contribution of clustering-based traffic pattern information. By incorporating congestion-level context into the prediction process, the hybrid model exhibits enhanced adaptability and stability compared to the standalone LSTM.

The correlation between Figures 3 and 4 underscores the effectiveness of integrating traffic pattern segmentation with sequence-based learning. The observed performance differences suggest that contextual traffic information plays a crucial role in improving prediction robustness under heterogeneous urban traffic dynamics. This comparative visualization confirms that the hybrid approach provides both improved predictive accuracy and a more consistent representation of traffic behavior over time.

Table 2. Model performance comparison

No	Model	RMSE	MAE	MAPE(%)
1	LSTM (Standalone)	1155.388571	917.094727	47.853817
2	Hybrid (KMeans+LSTM)	1109.015047	886.740234	46.064503

Table 2 shows that the hybrid K-Means–LSTM model consistently achieves lower prediction error compared to the standalone LSTM model across all evaluated error metrics. The reduction in error values indicates that incorporating clustering-based traffic pattern information enhances the predictive capability of the model, particularly under varying traffic conditions. This improvement confirms that the hybrid approach provides a more accurate and reliable estimation of traffic volume, supporting the effectiveness of integrating contextual traffic states into time-series prediction.

4. CONCLUSION

This study presents a hybrid K-Means–LSTM model for urban traffic volume prediction on arterial roads, with a specific focus on capturing heterogeneous traffic dynamics in urban environments. The results demonstrate that integrating clustering-based traffic pattern segmentation into the LSTM framework significantly improves both prediction accuracy and model interpretability when compared to standalone prediction models. By combining unsupervised traffic pattern discovery with sequential learning, the proposed approach is able to better represent the complexity of urban traffic behavior across varying congestion conditions.

Beyond predictive accuracy, the proposed hybrid approach demonstrates strong potential in supporting interpretable traffic analytics by explicitly linking prediction outcomes with identifiable traffic patterns. This characteristic is particularly valuable for practical applications, as it enables stakeholders and traffic management authorities to understand not only when congestion may occur, but also under what traffic conditions such congestion is likely to emerge. Such interpretability supports more informed and proactive decision-making processes, including congestion mitigation planning and traffic flow optimization.

Furthermore, the insights generated by the hybrid model contribute to the development of adaptive and data-driven urban traffic management strategies, especially on arterial roads that experience recurring congestion. While this study focuses on temporal traffic patterns using traffic volume as the primary indicator, future research may extend the proposed framework by incorporating spatial features, real-time sensor data, or external influencing factors such as weather conditions. These extensions are expected to further enhance model robustness and applicability within intelligent transportation systems.

REFERENCES

Abduljabbar, R., Dia, H., & Liyanage, S. (2025). Machine Learning Traffic Flow Prediction Models for Smart and Sustainable Traffic Management. *Infrastructures*, *10*(7), 155. https://doi.org/10.3390/infrastructures10070155

Anna, V. A. B. K., Chand, S., Alsultan, A., & Dixit, V. (2026). Investigating the spatial effects of zonal factors on road traffic speed variability during peak hour. *PLOS One*, *21*(1), e0340583. https://doi.org/10.1371/journal.pone.0340583

Chawla, P., Hasurkar, R., Bogadi, C. R., Korlapati, N. S., Rajendran, R., Ravichandran, S., … Gao, J. Z. (2024). Real-Time Traffic Congestion Prediction Using Big Data And Machine Learning Techniques. *World Journal of Engineering*, *21*(1), 140–155. https://doi.org/10.1108/WJE-07-2021-0428

Gupta, S., Kumar, A., & Kumar, A. (2025). Analysis of Traffic Flow Congestion by Integrating Supervised Machine Learning with K-mean Clustering. *SN Computer Science*, *6*(3), 255. https://doi.org/10.1007/s42979-025-03761-4

Hasan, I., Albafery, & Mulyadi, A. (2025). Comparative Study of Vehicle Noise Levels at Different Times in an Urban Area: A Case Study. *Journal of Geoscience, Engineering, Environment, and Technology*, *10*(1), 113–118. https://doi.org/10.25299/jgeet.2025.10.1.21474

Jaeger, A., & Banks, D. (2023). Cluster analysis: A modern statistical review. *WIREs Computational Statistics*, *15*(3). https://doi.org/10.1002/wics.1597

Khadka, S., Li, P. "Taylor," & Wang, Q. (2022). Developing Novel Performance Measures for Traffic Congestion Management and Operational Planning Based on Connected Vehicle Data. *Journal of Urban Planning and Development*, *148*(2). https://doi.org/10.1061/(ASCE)UP.1943-5444.0000835

Khan, A., Fouda, M. M., Do, D.-T., Almaleh, A., & Rahman, A. U. (2023). Short-Term Traffic Prediction Using Deep Learning Long Short-Term Memory: Taxonomy, Applications, Challenges, and Future Trends. *IEEE Access*, *11*, 94371–94391. https://doi.org/10.1109/ACCESS.2023.3309601

Kumar, I., Tripathi, B. K., & Singh, A. (2023). Attention-based LSTM network-assisted time series forecasting models for petroleum production. *Engineering Applications of Artificial Intelligence*, *123*, 106440. https://doi.org/10.1016/j.engappai.2023.106440

Li, S., Magli, E., Francini, G., & Ghinamo, G. (2024). Deep learning based prediction of traffic peaks in mobile networks. *Computer Networks*, *240*, 110167. https://doi.org/10.1016/j.comnet.2023.110167

Li, T., Song, S., & Yang, Y. (2022). Driving Restrictions, Traffic Speeds and Carbon Emissions: Evidence from High-Frequency Data. *China Economic Review*, *74*, 101811. https://doi.org/10.1016/j.chieco.2022.101811

Sayed, S. A., Abdel-Hamid, Y., & Hefny, H. A. (2023). Artificial intelligence-based traffic flow prediction: a comprehensive review. *Journal of Electrical Systems and Information Technology*, *10*(1), 13. https://doi.org/10.1186/s43067-023-00081-6

SUBAIR, S. O., IBITOYE, B. A., & KURANGA, A. T. (2024). Evaluation of Traffic Congestion in an Urban Roads: A Review. *ABUAD Journal of Engineering and Applied Sciences*, *2*(2), 1–7. https://doi.org/10.53982/ajeas.2024.0202.01-j

Yuan, T., Da Rocha Neto, W., Rothenberg, C. E., Obraczka, K., Barakat, C., & Turletti, T. (2022). Machine Learning for Next-Generation Intelligent Transportation Systems: A Survey. *Transactions on Emerging Telecommunications Technologies*, *33*(4). https://doi.org/10.1002/ett.4427