ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Random Forest Based Traffic Congestion Classification in Pekanbaru, Indonesia

Khoirul Imam [1], Roki Hardianto [*2]

[1,2]Faculty of Computer Science, Universitas Lancang Kuning, Pekanbaru, Indonesia

*Corresponding Author
Email: roki@unilak.ac.id

Received: 01/12/2025 Revised: 15/12/2025 Accepted: 20/12/2025 Published: 29/12/2025

Abstract

Abstract

Traffic congestion in Pekanbaru has become a critical urban challenge due to rapid growth in vehicle volume that is not matched by corresponding road capacity expansion. This study aims to develop a classification model that categorizes traffic congestion levels using original operational data obtained from the local Transportation Agency. The proposed approach employs a Random Forest classifier with three primary features: speed, traffic density, and the volume to capacity (V/C) ratio. Congestion levels are defined in three classes, low, moderate, and high, based on rule based thresholds derived from the Highway Capacity Manual (HCM) traffic engineering standards. The results demonstrate that the developed model achieves excellent predictive performance, reaching 98.86% accuracy on the testing dataset. Feature importance analysis further confirms that speed and the V/C ratio are the most dominant variables in determining congestion severity. Overall, the model is effective as a decision support tool for rapid identification of congestion hotspots and provides a practical foundation for developing a more responsive transportation management system in Pekanbaru.

Keywords: Traffic Congestion Classification, Random Forest, Traffic Flow Analysis, V/C Ratio, Decision Support System

Introduction

Traffic congestion is a multidimensional urban systems problem that continues to intensify across rapidly growing cities, particularly where motorization outpaces network capacity and operational control. Beyond longer travel times, congestion generates measurable economic losses through productivity decline and excess fuel consumption, while also aggravating environmental externalities through higher emissions and degraded air quality (Pang et al., 2023; Shang et al., 2024; Wu et al., 2025). Evidence from recent urban studies further indicates that mobility disruptions and network inefficiencies can materially amplify citywide congestion burdens, reinforcing the need for operationally actionable and data grounded congestion diagnostics (Xu et al., 2024).


ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

In many Indonesian cities, including Pekanbaru, congestion mitigation has relied largely on conventional interventions such as traffic engineering adjustments, corridor management, and selective infrastructure expansion. However, the effectiveness of these measures is often constrained by limited analytical support for rapid, objective, and consistent congestion classification at the segment or corridor level. Recent scholarship emphasizes that modern congestion management increasingly depends on multi source data pipelines and analytic frameworks that can translate observed traffic states into interpretable categories for policy and operational response (Akhtar & Moridpour, 2021; Cvetek et al., 2021; Seong et al., 2023). In parallel, explainable analytics has gained prominence because transportation agencies require not only accurate predictions but also transparent identification of dominant contributing factors to strengthen trust and enable targeted interventions (Dong et al., 2024).

Within this context, machine learning has become central to traffic state estimation and congestion assessment because it can model nonlinear relationships among speed, volume, density, and network constraints at scale. Recent studies show that incorporating congestion patterns and network dependencies can improve the robustness of predictive systems, especially when congestion dynamics propagate spatially across connected road segments (Leiser & Yildirimoglu, 2021; Sun et al., 2022). Complementary work integrates traffic flow theory with data driven learning to strengthen congestion estimation under real world variability (Pan et al., 2022), while comparative evaluations of classification models demonstrate the practical value of supervised learning for detecting congestion and predicting traffic jams across diverse urban environments (Hammoumi et al., 2025; Muktar et al., 2025).

This study specifically proposes a Random Forest based classifier to categorize traffic congestion levels in Pekanbaru using official data from the local transportation authority. Random Forest is selected due to its strong empirical performance for traffic condition classification and its capacity to quantify variable importance, which is valuable for operational interpretation and prioritization (Ahmed et al., 2023; Hammoumi et al., 2025). The model is constructed using three primary features that are widely recognized in congestion measurement and assessment practice, namely speed, density, and the volume to capacity (V/C) ratio (Marazi et al., 2023; Seong et al., 2023; Sokido et al., 2024). Observations are labeled into three congestion categories, low, moderate, and high, consistent with recent empirical uses of level of service oriented congestion assessment for intersections and urban roads (Marazi et al., 2023; Sokido et al., 2024).

Accordingly, the objectives of this research are as follows. First, to determine the most influential variables associated with congestion level in Pekanbaru using model based importance analysis. Second, to develop and validate a Random Forest classification model that can reliably categorize traffic conditions into operationally meaningful congestion levels. Third, to provide a functional analytic basis for decision support, enabling faster identification of congestion hotspots and supporting more responsive traffic management strategies, aligned with recent work on hotspot identification and applied congestion mapping in urban contexts (Jha et al., 2025).

Materials and Methods

The research pipeline consists of problem formulation, data acquisition, data processing and labeling, algorithm selection, model training, model evaluation, and conclusion.

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

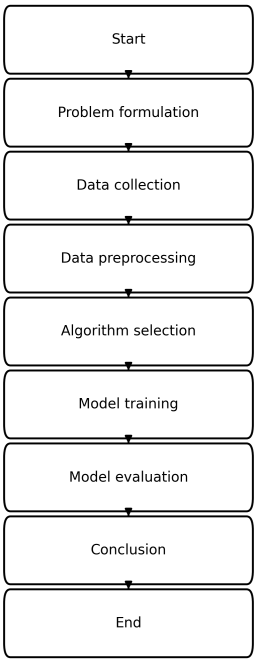

Figure 1. Research workflow

Problem Formulation

The study addresses the need for an objective and efficient mechanism to classify traffic conditions in Pekanbaru, where congestion levels must be identified consistently across multiple road functional classes. The core methodological challenge is transforming routine traffic performance indicators into standardized congestion labels and training a supervised classifier that generalizes to unseen observations. Recent congestion analytics literature recommends combining operational traffic measures with interpretable model structures to support adoption in agency decision making contexts.

Data Acquisition

This research uses secondary data formally obtained from the Pekanbaru Transportation Agency. The dataset is provided in spreadsheet format and contains results of traffic performance surveys conducted in 2024 across major road segments, including arterial, collector, and local roads. The main attributes utilized in this study are road segment name, average speed (km per hour), traffic density (vehicles per km), and the V/C ratio as a capacity utilization proxy.

Data Processing and Label Construction

Data processing is conducted to ensure reliability for supervised learning. This stage includes consistency checks for units, handling missing or anomalous records, removing duplicated entries if present, and preparing variables

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

for modeling. Road segment name is treated as a categorical attribute and is encoded to allow integration with numerical predictors.

For supervised classification, labels are constructed using traffic engineering service level concepts commonly linked to capacity utilization and operational conditions. In line with recent studies that define congestion severity using capacity driven ratios and service quality logic, observed traffic conditions are mapped to service level categories and then consolidated into three congestion classes: low, moderate, and high. This approach reduces subjectivity by grounding class definitions in engineering standards while still enabling machine learning based pattern recognition.

Algorithm Selection

Random Forest is selected as the primary classification algorithm due to its strong performance on complex and non linear relationships, its stability under heterogeneous predictors, and its ability to quantify feature influence through embedded importance measures and permutation based explanations. These properties have been repeatedly emphasized in recent intelligent transportation research as valuable for operational deployment because they combine accuracy with interpretability.

Model Training

The dataset is split into training and testing subsets using an 80 to 20 proportion with stratification to preserve the class distribution across splits. Model training is performed on the training subset, where the classifier learns the relationship between input predictors (speed, density, V/C ratio, and encoded road segment) and the congestion label. Key hyperparameters such as the number of trees and depth related controls can be determined through cross validation within the training data to balance bias and variance, as recommended by recent methodological syntheses in congestion forecasting and classification.

Model Evaluation and Interpretation

After training, model performance is assessed on the held out test subset. Evaluation uses standard classification metrics, including accuracy, precision, recall, and F1 score, supported by a confusion matrix to diagnose class specific misclassification patterns. In addition, feature importance analysis is conducted to identify dominant predictors. Recent transportation ML studies recommend complementing embedded importance with permutation importance or explainability oriented analysis to strengthen interpretability for practitioners and to validate that the model aligns with traffic engineering intuition.

Results and Discussion

Data Classification Results

After applying the Highway Capacity Manual (HCM) standard rule based classification to 439 valid records, a realistic class distribution was obtained, as illustrated in Figure 2.

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

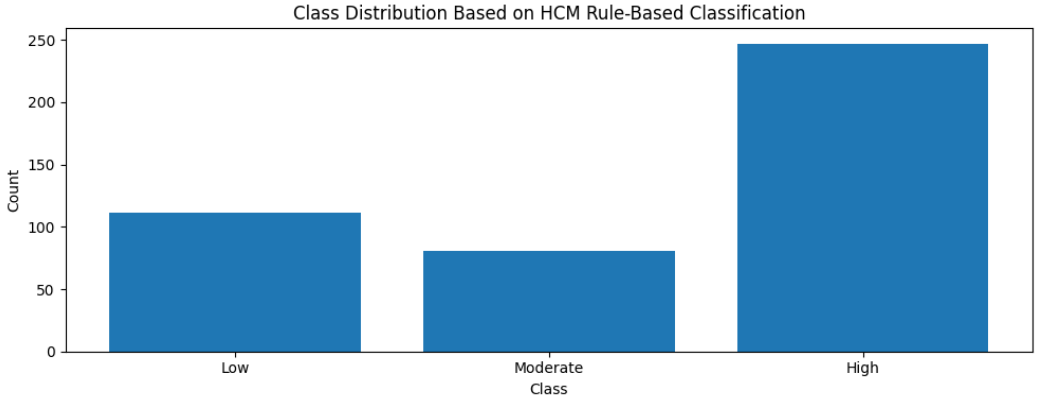

Figure 2. Data Distribution

The distribution (High: 247, Low: 111, Moderate: 81) indicates that, based on the observed data, traffic conditions in Pekanbaru most frequently fall into either high congestion conditions or very free flow conditions, while the transitional "moderate" condition is less commonly observed.

Exploratory Analysis

To visually validate the classification logic, box plot visualizations were generated in Figure 3, Figure 4, and Figure 5 to compare the characteristics of each class. The results confirm that the constructed classes exhibit clearly differentiated patterns. Specifically, the "High" class shows the lowest speeds and the highest V/C ratio, while the opposite pattern is observed for the "Low" class, thereby supporting the logical consistency of the proposed classification method.

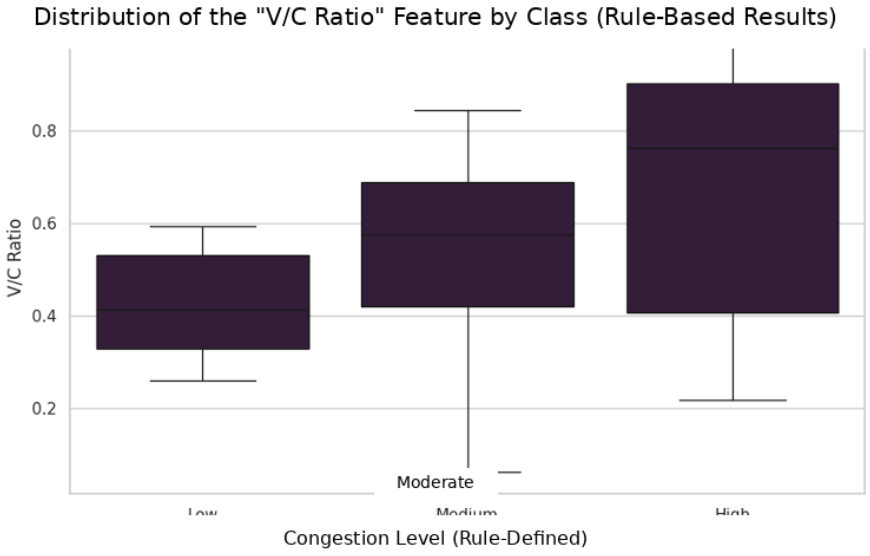

Figure 3. Box Plot of V/C Ratio

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

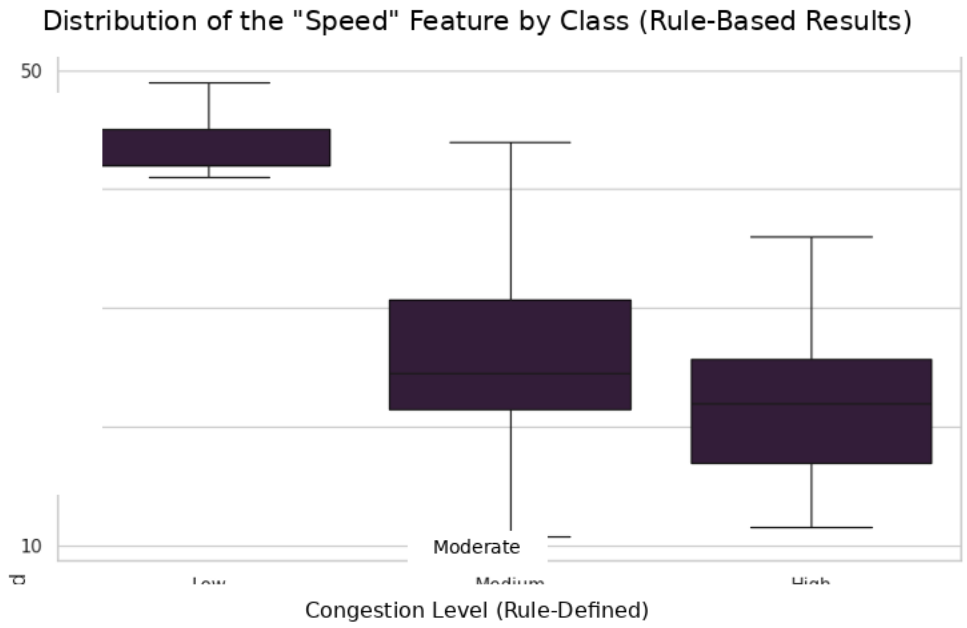

Figure 4. Box Plot of Speed

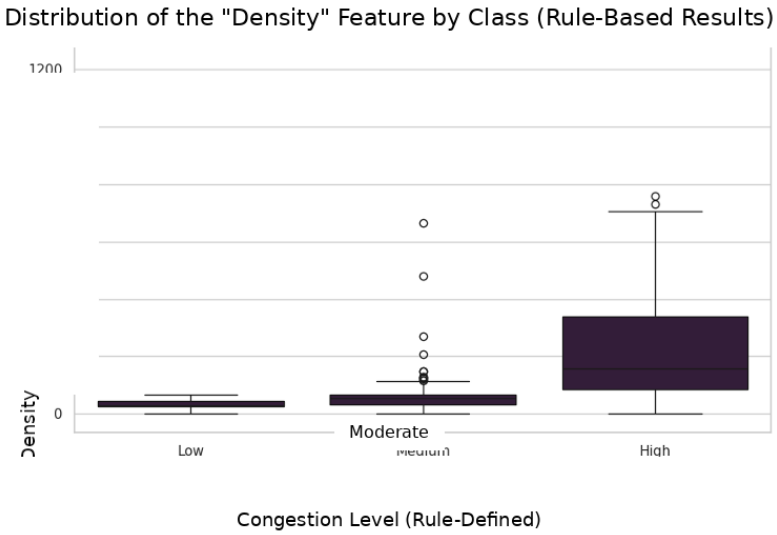

Figure 5. Box Plot of Density

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

Model Results

After training, the Random Forest model was evaluated using 88 test instances. The evaluation demonstrates highly satisfactory performance, with a final accuracy of 98.86 percent. Detailed performance metrics are presented in Table 1 and Figure 6.

Table 1. Model Performance Results

Class	Precision	Recall	F1 score	Support
Low	1.00	1.00	1.00	22
Moderate	1.00	0.94	0.97	16
High	0.98	1.00	0.99	50

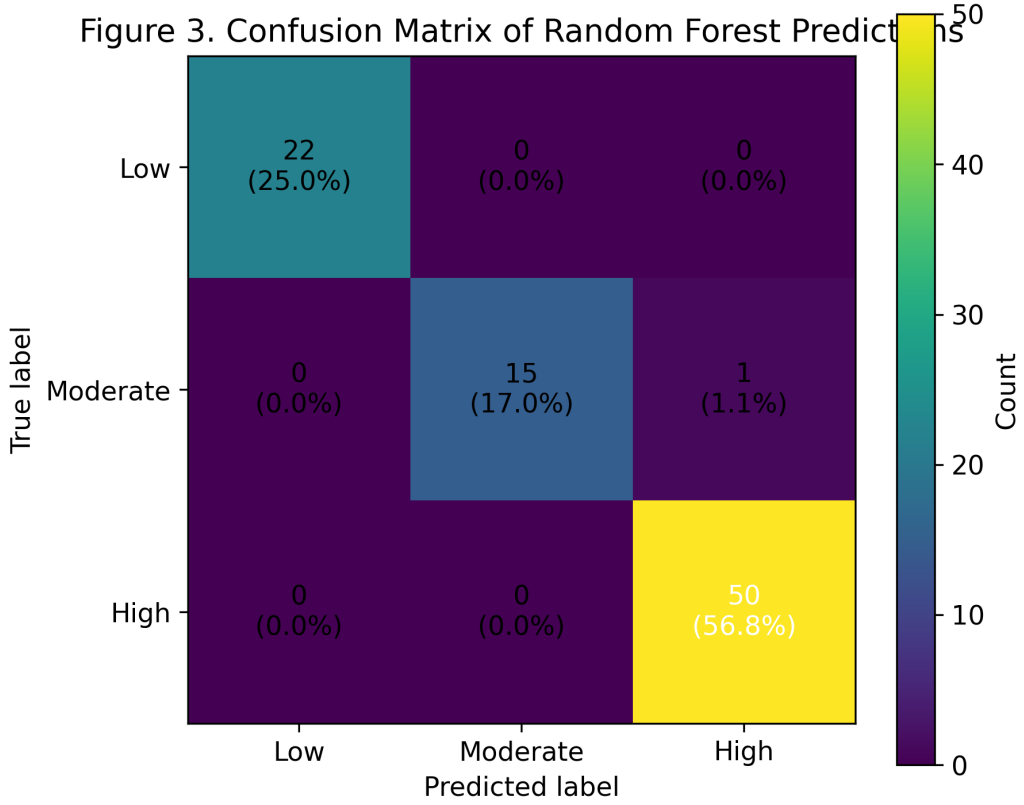

Figure 6. Confusion Matrix

ComniTech: Journal of Computational Intelligence and Informatics
https://dx.doi.org/000000000
P-ISSN: 3063-315X E-ISSN: 3063-0630

As shown in the classification report and confusion matrix, the model achieves near perfect precision and recall across all classes. Only one minor error is observed, where a single "Moderate" instance is misclassified as "High," indicating a very high level of reliability.

Analysis of the Most Influential Factors

The feature importance analysis in Figure 7 shows that the three most significant determinants of congestion classification are Speed, V/C Ratio, and Density. This finding is crucial because it validates the model logic from a traffic engineering perspective. The strong influence of these three features indicates that the model's classification decisions are grounded in appropriate operational principles rather than simply memorizing the dataset [14].

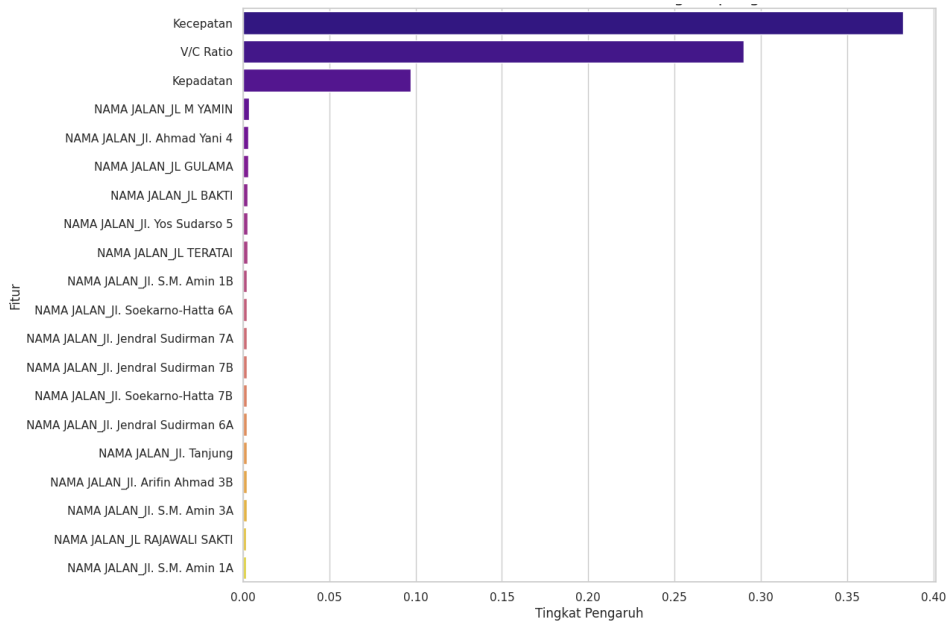

Figure 7. Influential Factors

Conclusion

This study reports comparative classification accuracies of 71.83% for Naive Bayes, 86.89% for Support Vector Machine (SVM), and 87.48% for Random Forest. The dataset comprised 3,355 raw tweets collected from X, and 3,351 tweets remained after preprocessing. Sentiment analysis was conducted using three polarity labels, namely positive, negative, and neutral. Labels were assigned automatically using a rule based classification approach grounded in a predefined keyword lexicon. Based on the experimental evaluation, Naive Bayes achieved the lowest accuracy, whereas Random Forest delivered the strongest performance for sentiment detection related to teacher bullying in Indonesia. These results indicate that Random Forest is more effective in capturing sentiment patterns and class distribution in the dataset than the other two methods.

For future work, it is recommended to expand the volume of tweet data to improve model robustness and potentially increase predictive accuracy. In addition, subsequent studies should consider deep learning approaches

for sentiment analysis on teacher bullying discourse in Indonesia, as such methods are more capable of modeling nuanced linguistic context and learning complex sentiment representations beyond keyword driven signals.

References

Ahmed, U., Tu, R., Xu, J., Amirjamshidi, G., Hatzopoulou, M., & Roorda, M. J. (2023). GPS based traffic conditions classification using machine learning approaches. Transportation Research Record, 2677(2), 1445 to 1454. https://doi.org/10.1177/03611981221111370

Akhtar, M., & Moridpour, S. (2021). A review of traffic congestion prediction using artificial intelligence. Journal of Advanced Transportation, 2021, 8878011. https://doi.org/10.1155/2021/8878011

Cvetek, D., Muštra, M., Jelušić, N., & Tišljarić, L. (2021). A survey of methods and technologies for congestion estimation based on multisource data fusion. Applied Sciences, 11(5), 2306. https://doi.org/10.3390/app11052306

Dong, J., Zhang, H., Cui, M., Lin, Y., Wu, H. Y., & Bi, C. (2024). TCEVis: Visual analytics of traffic congestion influencing factors based on explainable machine learning. Visual Informatics, 8(1), 56 to 66. https://doi.org/10.1016/j.visinf.2023.11.003

Hammoumi, A. E., Benabbou, L., & Himmi, M. M. (2025). Leveraging machine learning to predict traffic jams: A comparison of classification models and optimization methods. Results in Engineering, 25, 104241. https://doi.org/10.1016/j.rineng.2025.104241

Jha, M. K., Mishra, A., Saxena, S., & Verma, A. (2025). A machine learning approach to traffic congestion mapping and hotspot identification: A case study of two international cities. Future Transportation, 5(4), 161. https://doi.org/10.3390/futuretransp5040161

Leiser, N., & Yildirimoglu, M. (2021). Incorporating congestion patterns into spatio temporal deep learning algorithms. Transportmetrica B: Transport Dynamics, 9(1), 622 to 640. https://doi.org/10.1080/21680566.2021.1922320

Marazi, N. F., Panda, S., Katiyar, V., & Saha, P. (2023). Traffic congestion assessment tool for urban roads based on traffic and geometric characteristics: A case of Hyderabad, India. Journal of Transportation Engineering, Part B: Pavements, 149(4). https://doi.org/10.1061/JTEPBS.TEENG-7908

Muktar, M. N., Nagandla, S., & Sultana, S. (2025). Data driven modeling of traffic congestion to evaluate the effects of coordinated intersections. Transportation Engineering, 22, 100263. https://doi.org/10.1016/j.treng.2025.100263

Pan, S., Khateeb, Y., Zaidi, A. A., & Shakib, J. (2022). Incorporating traffic flow model into a deep learning algorithm for congestion estimation. Mathematical Problems in Engineering, 2022, 5926663. https://doi.org/10.1155/2022/5926663

Pang, L., Yue, H., Sun, C., & Zhao, J. (2023). Traffic congestion and urban air pollution: Evidence from Beijing. Journal of Environmental Economics and Management, 123, 102913. https://doi.org/10.1016/j.jeem.2023.102913

Seong, J., Kim, H., Kim, D., & Lee, J. (2023). Measuring traffic congestion with novel metrics: A case study of six US metropolitan areas. ISPRS International Journal of Geo Information, 12(3), 130. https://doi.org/10.3390/ijgi12030130

Shang, L., Zhong, L., Zhou, Y., & Shi, Y. (2024). Managing traffic congestion or improving air quality: Low emission zone policy as a multipurpose approach? Sustainable Cities and Society, 109, 105799. https://doi.org/10.1016/j.scs.2024.105799

Sokido, D. L., Worku, H., & Dulo, B. (2024). Measuring the level of urban traffic congestion for sustainable transportation in Addis Ababa, Ethiopia, the cases of selected intersections. Frontiers in Sustainable Cities, 6, 1366932. https://doi.org/10.3389/frsc.2024.1366932

Sun, Y., Jiang, G., Lam, S. K., & He, P. (2022). Learning traffic network embeddings for predicting congestion propagation. IEEE Transactions on Intelligent Transportation Systems, 23(8), 11591 to 11604. https://doi.org/10.1109/TITS.2021.3105445

Wu, X., Schaefer, A., Azevedo, I. L., & others. (2025). Best strategies for air quality and climate mitigation in road transport. Nature Communications, 16, 1714. https://doi.org/10.1038/s41467-025-56701-4

Xu, M., Wu, Y., Li, Y., & Jiang, Y. (2024). Personal mobility of employed residents reduces citywide traffic congestion: An empirical analysis based on public transport strike events. Science of the Total Environment, 943, 173164. https://doi.org/10.1016/j.scitotenv.2024.173164