JIVSC

Journal on Informatics Visualization and Social Computing

Volume 2, Issue 1, 11-24.
e_ISSN: E-ISSN 3123-7002
https://e-journal.gomit.id/jivsc
DOI : 10.58723/jivsc.v2i1.192

Spatial Analysis and Classification of Traffic Congestion Levels Using a Hybrid Machine Learning Approach in Badung Regency, Bali

Rizal Wahyu Pratama
Telkom University,
INDONESIA

Mikhael Setia Budi
Telkom University,
INDONESIA

Eka Sahputra*
Telkom University,
INDONESIA

Khulika Malkan
Telkom University,
INDONESIA

Wisnu Aji Sanjaya
Telkom University,
INDONESIA

Yoka Romadani
Telkom University,
INDONESIA

Article Info

Article history:

Received: February 21, 2026
Revised: March 04, 2026
Accepted: May 02, 20XX

Keywords:

Machine Learning;
Traffic Congestion;
K-Means, Random Forest;
Spatial Analysis;
Smart City.

Abstract

Background: Traffic congestion in Badung Regency, Bali, presents complex challenges for transportation management in a strategic tourism area, necessitating intelligent monitoring methods that surpass conventional surveys.
Aim: This study develops a Hybrid Machine Learning framework that integrates unsupervised and supervised learning techniques to map spatial patterns and precisely predict congestion anomalies.
Methods: Initially, the K-Means Clustering algorithm validated the segmentation of traffic data into five optimal clusters (k=5) based on Elbow and Silhouette Score tests, which subsequently served as the ground truth for data labeling. Furthermore, a comparative evaluation of Random Forest, XGBoost, and Gradient Boosting models was conducted to determine the best predictive performance.
Result: Experimental results identified Random Forest with strict pruning parameters (max_depth=3) as the optimal model, balancing accuracy and generalization stability while effectively avoiding overfitting through learning curve analysis. This model achieved a testing accuracy of 92.00% and a cross-validation score of 89.69%, with the Travel Time Index (TTI) and average speed identified as the primary determinants.
Conclusion: Spatial visualization demonstrated high consistency between the model's prediction maps and actual field conditions, confirming that this approach is effective as a foundation for an Early Warning System to support data-driven traffic management policies in Bali.

To cite this article: Pratama et al. (2026). Spatial Analysis and Classification of Traffic Congestion Levels Using a Hybrid Machine Learning Approach in Badung Regency, Bali. *Journal on Informatics Visualization and Social Computing, 2*(1), 11-24. https://doi.org/10.58723/jivsc.v2i1.192
This article is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License ©2026 by author/s

INTRODUCTION

Rapid developments in transportation technology and urbanization in tourism areas have triggered complex public mobility in various major cities in Indonesia. As a primary tourism center in Bali Province, Badung Regency has experienced vehicle volume growth that far exceeds the available road infrastructure capacity (Ni Made Prisca Rahmayanti et al., 2024). Congestion negatively impacts economic efficiency, environmental quality, and tourist comfort (Chen et al., 2026). Although local governments have implemented conventional traffic policies, these static approaches fail to keep pace with rapid and dynamic changes in traffic flow. Therefore, the integration of geospatial data and artificial intelligence (AI) is crucial for transforming transportation management into a more responsive Smart City system (J. Zhang et al., 2026). In this research, Big Data analysis of vehicle movement patterns serves as the fundamental basis for resolving transportation issues in the region.

Previous research has extensively explored the use of machine learning algorithms to predict traffic flow in various metropolitan cities (Cai et al., 2025). Several studies have used clustering methods to identify road typologies based on vehicle density at specific times (X. Zhang et al., 2025). Other research has applied supervised learning models, such as Random Forest and XGBoost, to predict travel times with varying levels of accuracy (Shah & Piantanakulchai, 2026; Zhu & Goverde,

*** Corresponding author:**
Eka Sahputra, Telkom University, INDONESIA. ✉ekasahputra@telkomuniversity.ac.id

2021). The use of transportation indices, such as the Travel Time Index (TTI), has also been scientifically proven as an effective indicator for measuring congestion intensity in urban areas (Abdolmaleki et al., 2020; Erdelić et al., 2021; Khorshidi et al., 2025). However, current literature is still dominated by studies on cities with industrial characteristics that differ from those of tourism-oriented cities. Most of these studies also tend to focus on theoretical algorithm performance without linking it to detailed administrative boundary visualizations. Literature regarding the sequential integration of unsupervised and supervised learning for nighttime congestion mapping in Bali remains very limited.

Despite various studies on traffic prediction, significant gaps remain in the literature concerning the handling of imbalanced data in high-volatility tourism areas, particularly during nighttime periods (Bazarnovi & Mohammadian, 2024; Zheng et al., 2026). Conventional classification models often fail to accommodate the characteristics of short-distance trips that dominate mobility in the culinary and nightlife centers of Badung Regency. Furthermore, precise administrative boundary-based spatial validation is rarely implemented, which often leads to bias in field model accuracy. These gaps cause prediction models to frequently fail in recognizing congestion anomalies during evening peak hours at critical tourism spots. While many studies have compared classification model stability in depth using learning curves to ensure model generalization (Solís-Martín et al., 2025), the lack of transparency regarding the contribution of dominant features in cluster formation remains a barrier to accurate decision-making (Xu et al., 2025).

The urgency of this research is based on the fundamental need for a traffic classification system that is not only objective but also evidence-based. In contrast to static and aggregate annual statistics, data acquisition through crawling techniques offers temporal granularity capable of representing field dynamics. By combining K-Means and Random Forest algorithms, this study can automatically create data labels prior to classification prediction (Li et al., 2025). This hybrid approach will minimize subjectivity in determining congestion or smooth traffic categories based on rigid thresholds. Additionally, the selection of strict pruning parameters in the classification model aims to produce a stable model that is easily interpreted by relevant authorities. Utilizing GeoJSON files as administrative boundaries will ensure that every prediction has a relevant and accountable geographical context (Bührs et al., 2026). This innovation is expected to enhance the effectiveness of traffic monitoring, especially during nighttime tourism hours, and assist in route optimization for road users in Badung Regency.

The primary objective of this research is to design an accurate traffic status prediction model by integrating K-Means and Random Forest algorithms in Badung Regency. This study focuses on grouping traffic characteristics into five functional categories based on speed and travel time parameters captured at 19:48 WITA. This research tests the hypothesis that the synthesis of Travel Time Index (TTI) features and average speed can achieve classification accuracy exceeding 90% under nighttime traffic conditions. Furthermore, to ensure model robustness, validation is performed through spatial analysis to verify the consistency between predictions and actual conditions (ground truth). This study will also conduct an in-depth evaluation of the impact of removing specific features through an ablation study to understand the primary determinants of congestion. The final results of this research are expected to provide theoretical contributions to the field of smart transportation and practical contributions to the traffic management system in Bali. Thus, this study represents a strategic step toward achieving sustainable mobility efficiency in an international-level tourism destination.

METHOD

This study adopts a quantitative experimental design utilizing a Hybrid Machine Learning approach that integrates unsupervised and supervised learning methods (Yu et al., 2025). The research framework is systematically designed as illustrated in Figure 1, which maps the data transformation from raw spatial acquisition to a validated predictive model. This hybrid approach was selected as a strategic solution to address the absence of standard class labels in raw traffic data, thereby enabling the development of an objective prediction model based on natural data patterns in the field rather than solely on subjective researcher assumptions (Shang et al., 2022).

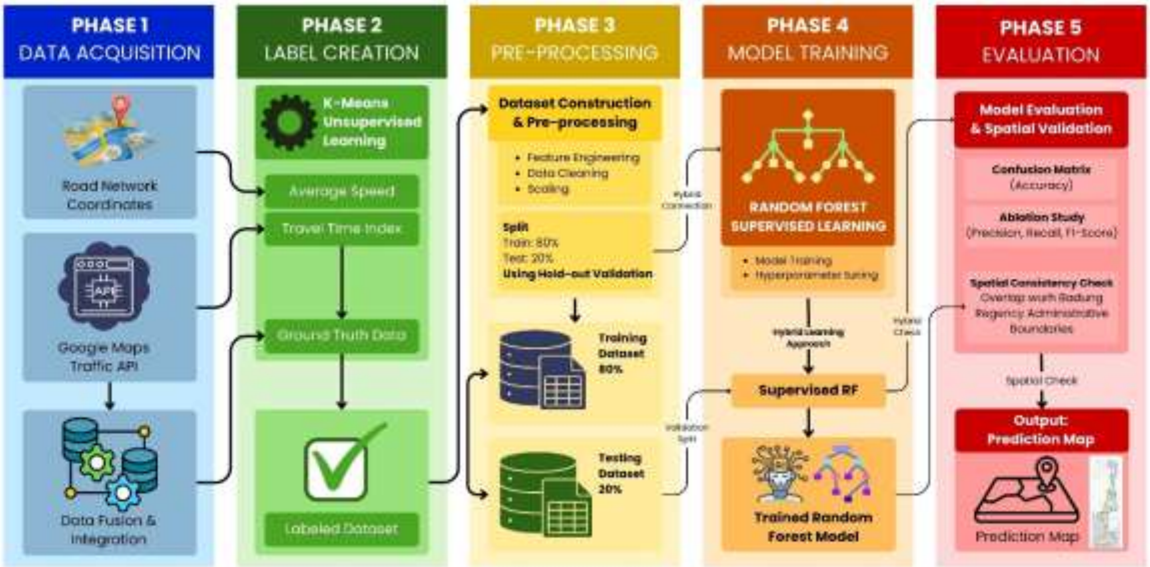


Figure 1. Research Flow Diagram

Operationally, the research sequence begins with the definition of the study area, followed by data collection and preprocessing using the Google Maps Distance Matrix API during the evening peak activity window (19:48 WITA). These raw data are then converted into analytical features, such as the Travel Time Index (TTI) and average speed (Azizah Amelia, 2025). Given the lack of initial class labels, the research proceeds to the Automatic Label Generation stage using the K-Means Clustering algorithm to automatically establish the ground truth or congestion categories (Shang et al., 2024). These labeling results serve as the foundation for the Predictive Modeling stage, where three classification algorithms (Random Forest, XGBoost, and Gradient Boosting) are trained and their performance is compared. The research workflow concludes with the Evaluation & Spatial Validation stage, in which the selected model is not only statistically tested but also visually validated by juxtaposing prediction maps against actual conditions to ensure the geographical consistency of the results.

Study Area

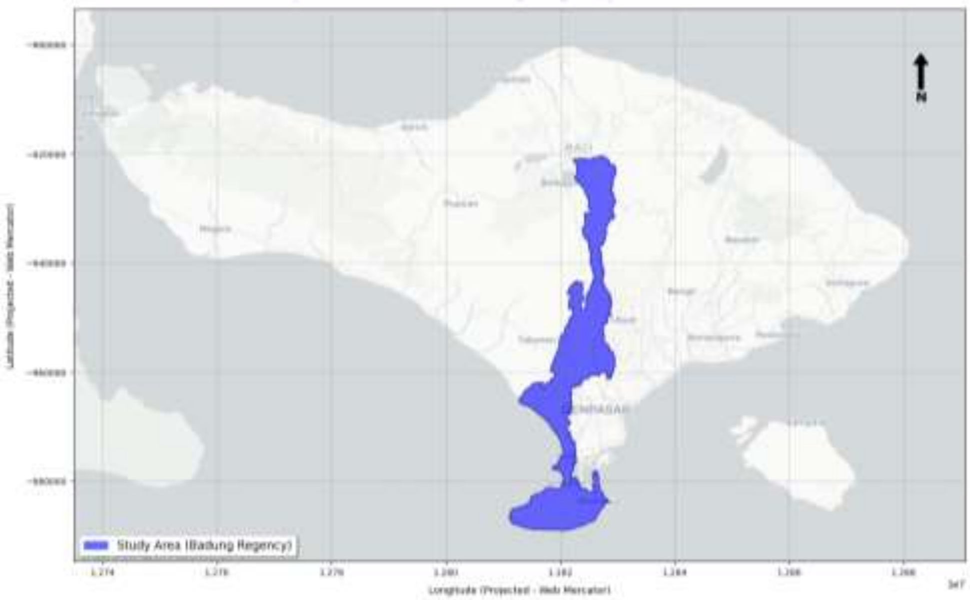


Figure 2. Research Location in Badung Regency, Bali

Geographically, Badung Regency is situated in the center of Bali Island and occupies a strategic position as the primary international tourism gateway via I Gusti Ngurah Rai International

Airport (Rozy et al., 2025). This region is directly bordered by the Indian Ocean to the south and several neighboring regencies, including Tabanan, Gianyar, and Denpasar City, on the other sides. The selection of Badung Regency as the research area is based on the complexity of its road network, which encompasses national, provincial, and local roads interconnected to support the mobility of both local residents and tourists (Soimun et al., 2025). This diversity of road types provides a rich data variation for the machine learning model to objectively recognize various levels of traffic status.

Badung Regency has a vertically elongated territorial configuration extending from the mountainous regions in northern Bali, as illustrated in Figure 2. The southern part of the region serves as a highly dense international tourism hub, covering the areas of Kuta, Seminyak, and Uluwatu, whereas the northern part is dominated by agricultural zones and nature conservation. These varied topographical conditions, ranging from coastal lowlands to highlands, present significant infrastructure challenges, making it a representative study location for testing the complexities of modern transportation management.

Population and Methods of Sampling

The population in this study is defined as the entirety of highway segments accessible to four-wheeled vehicles within the administrative region of Badung Regency, Bali (Joelianto et al., 2022). Given the extensive scope of the road network and the constraints of computational resources required to map every meter of road in real-time, this research does not employ a total population census (Ranjan et al., 2023). The population focus is directed toward strategic routes connecting economic centers, tourism hubs, and residential areas characterized by dynamic traffic fluctuations during the evening.

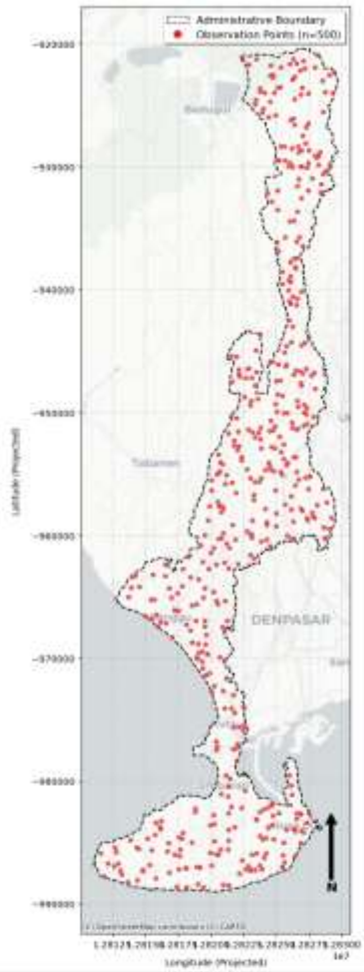


Figure 3. Spatial Distribution of 500 Traffic Observation Points

The sampling technique utilized is Purposive Random Spatial Sampling with strict geographical boundaries. The researchers generated 500 random route coordinates, as illustrated in Figure 3, while applying algorithmic filters to ensure that these points remained within the study area. Data samples were acquired at the specific time of 19:48 WITA to capture the characteristics of nighttime traffic loads, representing the peak of culinary tourism activities and commuting patterns in the Badung region (Imoh & Rad, 2025).

Data Acquisition

The traffic data acquisition process in this study was implemented programmatically using a Python-based web crawling algorithm integrated with the Google Maps Distance Matrix API. Data collection focused on 500 observation route points scattered randomly yet representatively throughout the primary and arterial road networks of Badung Regency. Each API request was configured using the driving mode with the best_guess traffic model to ensure travel time estimation accuracy that closely approximates real-world field conditions. This method enabled the extraction of precise raw geospatial data, including origin and destination coordinates, along with key travel metrics, the full details of which are presented in Table 1.

Table 1. Traffic Data Variable Specifications

No	Variable Name	Data Type	Description
1	Route_ID	String	Unique identity for each route taken.
2	Timestamp	Datetime	Data collection time (19:48 WITA).
3	Distance_m	Numerik (Int)	Route distance in meters.
4	Duration_Normal_sec	Numerik (Int)	Estimated travel time under normal conditions (seconds).
5	Duration_Traffic_sec	Numerik (Int)	Estimated travel time during traffic jams (seconds).
6	TTI	Numerik (Float)	Travel Time Index.
7	Speed_KMH	Numerik (Float)	Average vehicle speed (km/h).
8	Origin_Lat	Numerik (Float)	Latitude of the starting point of departure.
9	Origin_Lon	Numerik (Float)	Longitude of the starting point of departure.
10	Dest_Lat	Numerik (Float)	Latitude of the destination point.
11	Dest_Lon	Numerik (Float)	Longitude of the destination point.

The resulting final dataset records traffic conditions specifically during the evening peak activity window at 19:48 WITA. This timing was selected based on the urgency to capture congestion patterns triggered by a combination of returning commuters and tourist mobility toward culinary centers in South Badung. These raw data were subsequently enriched through a preprocessing stage by calculating vital derivative variables: the Travel Time Index (TTI) and Average Speed (Speed_KMH). The TTI variable is calculated as the ratio between the duration during traffic (Duration_Traffic_sec) and the normal duration (Duration_Normal_sec), which serves as the primary indicator for labeling the severity of congestion for each observed road segment.

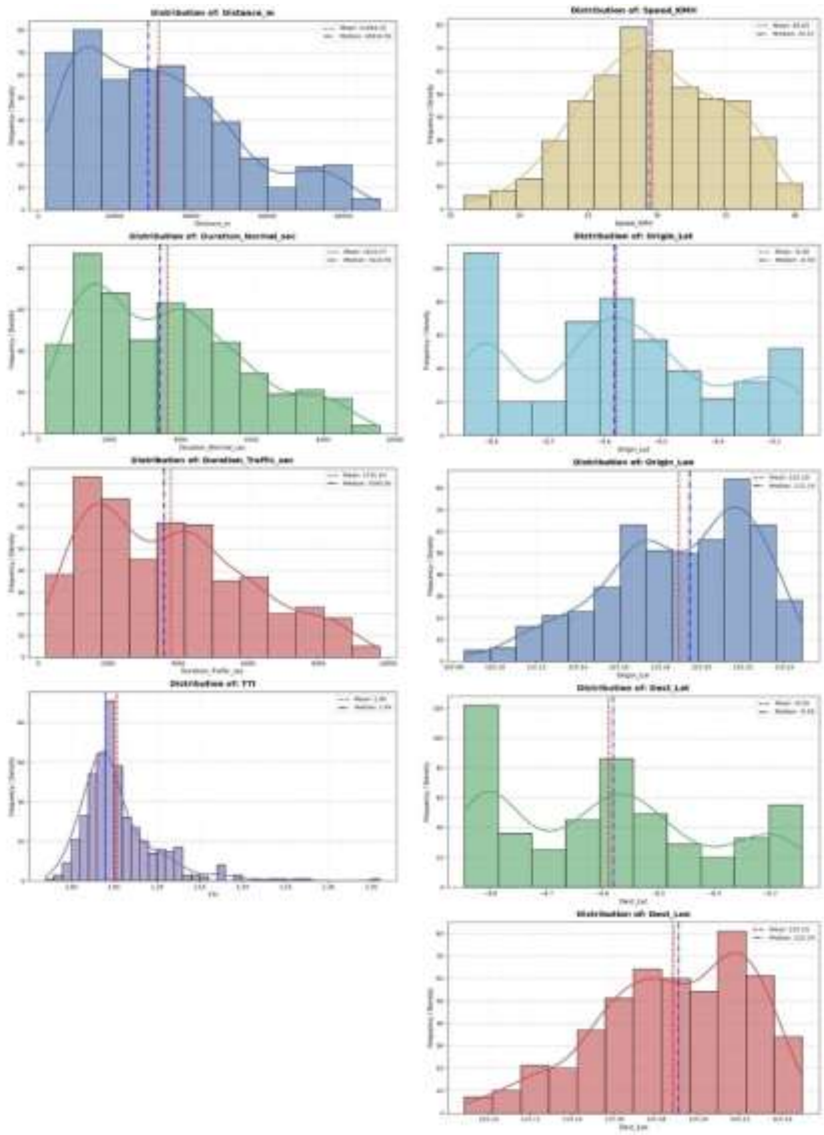


Figure 4. Variable Distribution Graph

Initial validation of the statistical characteristics of the acquired variables is visualized in Figure 4. These distribution graphs demonstrate that the dataset possesses a heterogeneous range of values, representing dynamic traffic conditions ranging from stop-and-go phases (low speed with high TTI) to free-flow conditions. The data distribution for the speed (Speed_KMH) and travel time index (TTI) variables confirms that the data acquisition successfully captured real-world traffic pattern variability rather than mere momentary anomalies. The diversity of the data profile shown in these histograms is a crucial prerequisite for the machine learning algorithm to effectively learn congestion patterns and accurately distinguish between traffic classes in the subsequent modeling stage.

Label Creation

Since the acquired traffic data lacked inherent categorical labels, this study implemented an Unsupervised Learning approach using the K-Means algorithm to generate class labels automatically (Winfrey et al., 2023). This process aimed to group 500 observation data points into categories exhibiting similar congestion characteristics based on the Speed_KMH and Travel Time Index (TTI) features.

This labeling process was established based on two primary features: Average Speed (Speed_KMH) and Travel Time Index (TTI). Both variables were utilized as input dimensions for the algorithm to map traffic patterns ranging from very smooth to total congestion conditions. To ensure the validity of the grouping, the determination of the optimal number of clusters (k) was not conducted arbitrarily instead, a dual heuristic evaluation procedure was employed (Cebolla-Alemany et al., 2024). The Elbow Method was utilized to identify the point where the reduction in intra-cluster variance begins to plateau, while Silhouette Analysis was applied to measure the quality of separation and cohesion among the resulting groups. The final output of this clustering process was subsequently converted into class labels (Ground Truth) to train the classification models in the following stage.

Analysis and Evaluation Plan

Prior to the performance evaluation stage, the dataset consisting of 500 observation points was partitioned using the Hold-out Validation method, with 80% of the data allocated to the training set and 20% to the testing set. This division was conducted using stratified sampling to maintain a balanced proportion of congestion classes across both subsets, ensuring the model remains unbiased toward the majority class. The model's performance was then comprehensively analyzed using a Confusion Matrix to map the distribution of prediction errors. Additionally, standard metrics such as Precision, Recall, and F1-Score with a macro-average were utilized to assess the balance of detection across classes, ensuring the model's ability to recognize even infrequent traffic conditions (Mehavilla et al., 2026). An Ablation Study was also implemented by iteratively removing features to identify the variables that contribute most significantly to the model's predictive accuracy.

The final analysis stage focused on spatial visual validation using interactive maps based on GeoPandas. The prediction maps generated by the model were juxtaposed side-by-side with the original Ground Truth labels and overlaid with the administrative boundaries of Badung Regency. This comparative approach aims to qualitatively demonstrate that the model can replicate geographic and contextual congestion distribution patterns, rather than merely memorizing statistical numerical patterns, thereby ensuring the prediction results offer real practical utility for policymakers (Obaidat & Alomari, 2026).

Scope and Limitations

The scope of this research is confined to congestion analysis based on snapshot data a single point in time at 19:48 WITA, and thus does not encompass long-term time-series predictions or congestion pattern forecasts for morning or afternoon periods. The modeling variables are limited to speed, distance, and travel time data provided by the API, without accounting for external variables such as weather conditions, traditional ceremony schedules, or incidental traffic accidents.

Another methodological limitation lies in the complete dependence on the accuracy of third-party data (Google Maps), where any inherent algorithmic bias from the service provider is beyond the researchers' control (Mirzahossein et al., 2025). Furthermore, while the random sampling technique involving 500 points is statistically adequate for validation, it may not cover all minor residential roads or local shortcuts within Badung Regency. Consequently, the generalization of these results should be interpreted within the context of the main and secondary roads identified by digital mapping services during evening operational hours.

RESULTS AND DISCUSSION

Analysis of Traffic Cluster Characteristica

The determination of the optimal number of clusters (k) in the K-Means algorithm was performed through a rigorous statistical validation procedure to ensure that the resulting labels represent the natural structure of the data. As visualized in Figure 5, the evaluation utilized two complementary heuristic metrics. The graph on the left (Elbow Method) illustrates the downward trend of inertia values or the Sum of Squared Errors (SSE) as the number of clusters increases. A sharp decrease in intra-cluster variance is observed from $k=2$ to $k=4$, followed by a significant deceleration or the formation of an elbow (inflection point) precisely at $k=5$. This point indicates a diminishing returns phenomenon, where adding clusters beyond five no longer significantly contributes to data compaction; instead, it merely increases model complexity, potentially leading to overfitting.

The validity of selecting $k=5$ was further confirmed by the graph on the right in Figure 5, which presents the Silhouette Score analysis. This metric measures the quality of inter-cluster separation and internal cohesion, where values approaching $+1$ signify an ideal grouping. Based on the graph, the average Silhouette Coefficient reached its peak in the $k=5$ scenario, outperforming other cluster counts. This confirms that with a five-category division, the data is partitioned with the most distinct boundaries (well-separated), and the level of ambiguity or overlap between cluster members is at its minimum compared to other grouping options.

Based on the convergence of evidence from both visual methods, this study established five clusters as the definitive Ground Truth for the subsequent classification stage. Semantically, these five clusters are mapped into an operationally meaningful hierarchy of traffic status based on their cluster center characteristics (centroids), ranging from Very Smooth (high speed, low TTI) to Total Congestion (near-zero speed, extreme TTI). This decision provides a robust foundation for the Supervised Learning algorithm to learn granular and realistic congestion patterns, rather than merely distinguishing between congestion and smooth flow in a binary fashion.

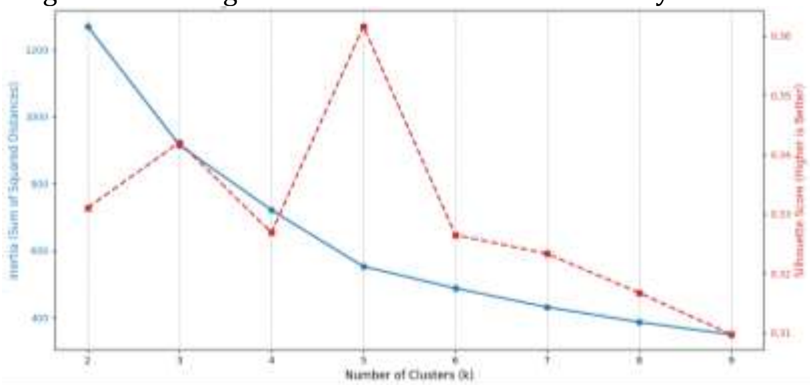


Figure 5. Optimization of k: Elbow Method vs Silhouette Score

Training and Testing Model Classification

The traffic data classification stage was performed by training and testing machine learning models using data labeled with Ground Truth. This process aims to develop a predictive model that is not only accurate in capturing training data patterns but also capable of generalizing effectively to new, unseen data.

Model Comparison and Selection

As an initial step, this study benchmarked three popular ensemble learning algorithms: Random Forest, XGBoost, and Gradient Boosting. To ensure a fair comparison, all three models were trained using controlled hyperparameters (*max_depth=3*). The summary of this comparative evaluation is presented in Table 2.

Table 2. Tabel Komparasi Stabilitas Model

Model Algorithm	Training Score	Validation Score	Gap	Visual
Random Forest	0.924700	0.896900	0.0278	Converging
XGBoost	1.000000	0.931100	0.0689	Flat Line
Gradient Boosting	1.000000	0.932500	0.0675	Flat Line

As shown in Table 2, the boosting-based algorithms (XGBoost and Gradient Boosting) tend to yield perfect training scores of 1.0000. While these results appear impressive, they indicate significant overfitting, where the models aggressively memorize training patterns, resulting in a statistically unrealistic flat line learning curve. In contrast, the Random Forest algorithm demonstrates a more organic learning behavior. Although its Final Training Score is slightly lower at 0.9247, the marginal gap of 0.0278 between the training and validation scores indicates a low risk of overfitting.

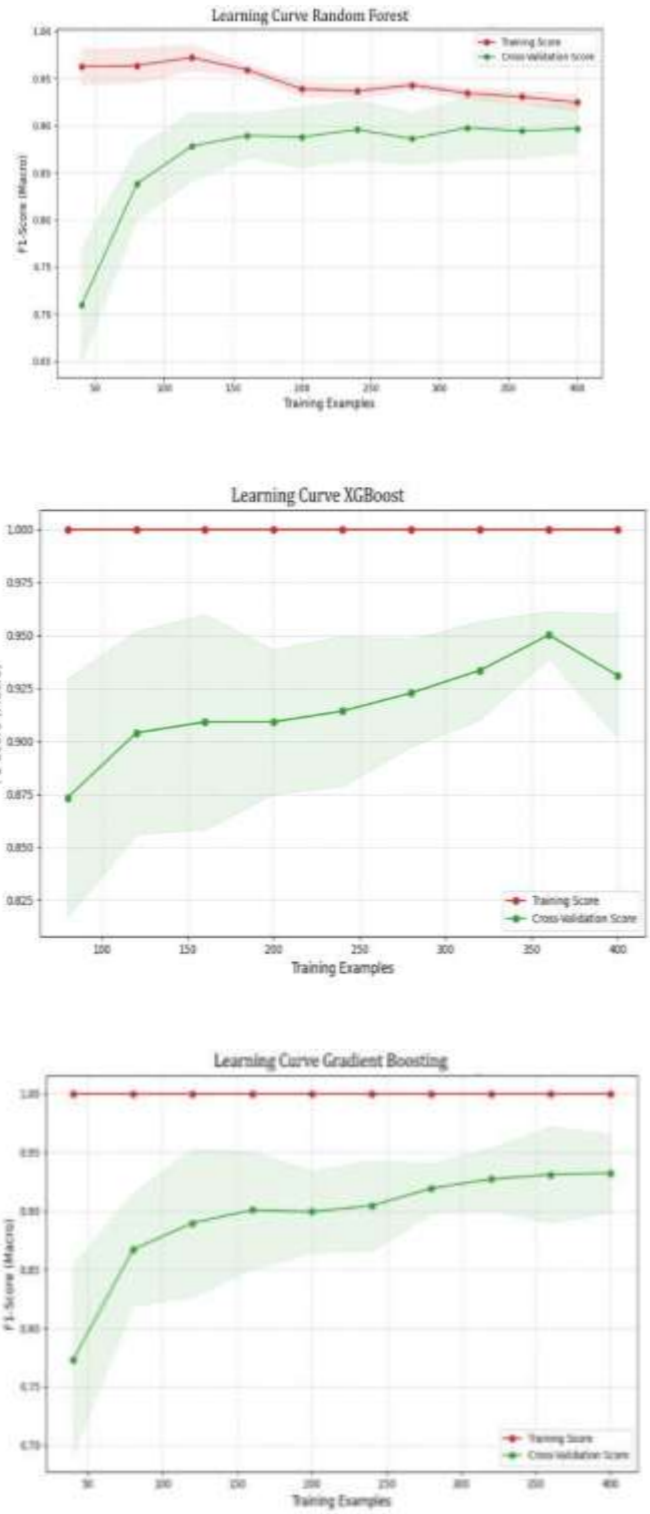


Figure 6. Learning Curve Hybrid Machine Learning

This convergence is further validated through the visualization in Figure 6. The Random Forest graph illustrates that as the number of training samples increases, the training accuracy (red line) gradually decreases, while the validation accuracy (green line) continues to improve until both curves meet and stabilize at a high accuracy level. This asymptotic convergence pattern confirms that the Random Forest model has effectively learned generalizable patterns rather than simply memorizing noise. Consequently, this study identifies Random Forest as the Champion Model for the final prediction phase.

Evaluation of Random Forest Model Performance

Once the champion model was established, a comprehensive performance evaluation was conducted using a testing set of 100 samples, constituting 20% of the total dataset. The quantitative results of this evaluation are summarized in Table 3.

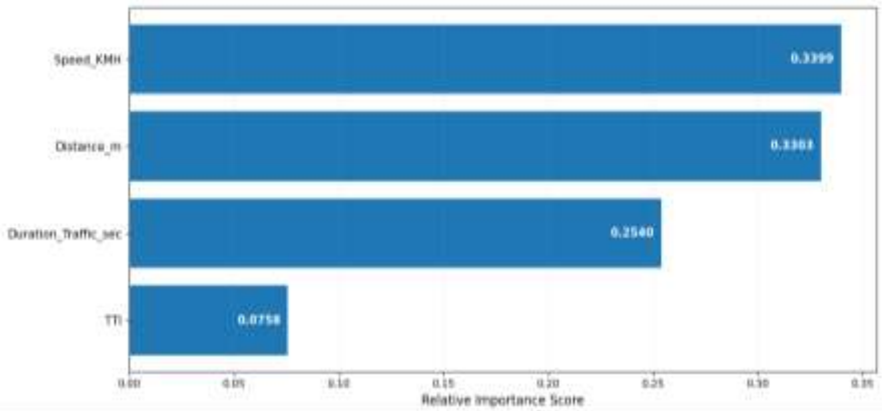
Table 3. Random Forest Model Classification Report

	Precision	Recall	F1-Score	Support
0	1.00	0.75	0.86	8
1	1.00	0.86	0.92	21
2	0.92	0.92	0.92	24
3	0.84	1.00	0.91	31
4	1.00	0.94	0.97	16
Accuracy			0.92	100
Macro AVG	0.95	0.89	0.92	100
Weighted AVG	0.93	0.92	0.92	100

Overall, the Random Forest model achieved a final accuracy of 92.00%, which is considered highly satisfactory for multi-class classification in dynamic traffic environments. A more detailed analysis of Table 3 reveals a balanced performance across all classification categories. Precision metrics reached perfect scores (1.00) for Class 0 (Very Smooth), Class 1 (Smooth), and Class 4 (Total Congestion), demonstrating the model's high reliability in minimizing False Positives under these extreme conditions. While there was a slight reduction in the Recall metric for Class 0 (0.75), the weighted average F1-Score remained robust at 0.92. This level of performance confirms that the model can detect and differentiate between five distinct congestion levels with high consistency, establishing it as a valid instrument for traffic mapping in Badung Regency.

Feature Importance Analysis

This study conducted a Feature Importance analysis to identify the variables contributing most significantly to the classification model's decision-making process. Initial testing indicated that the Travel Time Index (TTI) serves as a dominant variable, with a relative importance score significantly exceeding other features. Average speed (Speed_KMH) ranks second as a crucial supporting indicator in differentiating road density status. These findings theoretically reinforce the position of TTI as the most representative congestion measurement standard because it directly reflects the delay ratio. The dominance of TTI within the model structure ensures that the classification predictions are grounded in robust transportation logic.



Gambar 7. Feature Importance

Based on the visualization in Figure 7, the analysis reveals the following distribution of feature contributions:

Speed_KMH

The average speed variable is the primary determinant, achieving the highest importance score of 0.3399. This confirms that vehicle speed fluctuations are the most sensitive indicator for distinguishing traffic status, particularly in separating free-flow conditions from acute congestion..

Distance_m

Travel distance ranks second with a score of 0.3303, representing a statistically narrow margin compared to average speed. This high value indicates that the model relies heavily on the spatial dimensions of the routes to classify travel patterns, especially in identifying long-distance trips within specific clusters.

Duration_Traffic_sec

Travel time under traffic conditions contributes a score of 0.2540. This variable acts as a crucial temporal weight, allowing the model to interpret real-time travel obstacles experienced by vehicles.

TTI

The Travel Time Index (TTI) recorded the lowest explicit importance score at 0.0758. Although theoretically a robust delay ratio, the information contained within the TTI appears to be largely represented explicitly by the combination of speed and duration variables within this specific Random Forest structure.

The dominance of Speed_KMH and Distance_m, which cumulatively account for over 67% of the model's decision influence, suggests that traffic characteristics in Badung Regency are heavily influenced by the ratio of distance to travel speed. These findings technically support the decision to utilize a model with strict pruning parameters (*max_depth=3*), as the primary features possess sufficient discriminative power to achieve high accuracy without requiring an overly complex tree structure.

The validity of each variable's contribution was further verified through an Ablation Study using an iterative feature removal method. The results of the ablation study demonstrate that removing the TTI variable leads to the most significant decrease in model performance (F1-Score) compared to the removal of distance or duration variables. This phenomenon empirically proves that the TTI feature cannot be substituted by other features in maintaining the accuracy of congestion status predictions. While the distance variable provides a minimal contribution to overall accuracy, it remains essential for identifying long-distance travel within Cluster 4. This series of tests confirms that the constructed model utilizes an optimal and efficient feature set for traffic analysis in Badung Regency.

Spatial Validity and Prediction Visualization

Statistical validation alone is often insufficient to represent model performance within a geographical context. Consequently, this study conducted a spatial validity test by juxtaposing the actual label maps (Ground Truth) and the Random Forest model's prediction maps side-by-side, as presented in Figure 8. This visualization aims to qualitatively demonstrate that the 92% numerical accuracy obtained previously aligns with the consistency of field mapping.

Based on Figure 8, there is a very high level of visual alignment between the left-side map (Actual Data) and the right-side map (Model Prediction). The spatial distribution pattern of congestion was successfully replicated with significant precision. Concentrations of red and orange points are accurately clustered in the South Badung region, which serves as the center for tourism and commercial activities (Kuta, Seminyak, Canggu). This indicates that the model is capable of capturing spatial correlations showing that areas with high activity density have a higher probability of congestion even though coordinates were not explicitly included as training features.

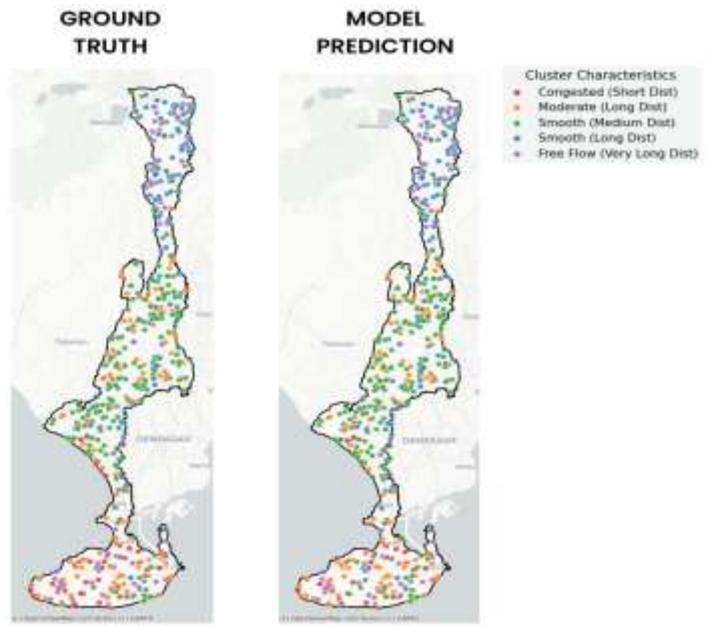


Figure 8. Comparison of Ground Truth with Model Prediction

Conversely, in the North Badung region, which is dominated by hilly topography and agricultural areas, the model consistently predicts these points as Class 0 and Class 1. The minimal presence of random spots (noise) or fatal prediction errors such as predicting total congestion in quiet areas on the right-hand map confirms that the Random Forest algorithm has successfully generalized traffic patterns. Any minor differences observed are localized and typical, resulting from micro-variability in the testing data. Thus, these visualization results confirm that the developed model possesses high reliability for use as an instrument for region-based traffic condition mapping.

CONCLUSION

This study concludes that a Hybrid Machine Learning approach serves as an effective and valid solution for addressing the challenges of absent historical labels in traffic data. The integration of the K-Means Clustering algorithm as an automated label generator successfully identified, objectively, that the traffic structure in Badung Regency is divided into five natural strata (Ground Truth), ranging from Very Smooth to Total Congestion conditions. These findings replace traditional, rigid binary classification assumptions, providing a more granular and representative categorization foundation for actual field congestion dynamics.

In the predictive modeling phase, the Random Forest algorithm demonstrated superior performance compared to other boosting methods, achieving a testing accuracy of 92.00% with high generalization stability and minimal overfitting risk. The model's reliability is reinforced by the spatial visual validation results, where the prediction maps accurately replicate congestion distribution patterns, particularly in detecting congestion concentrations in the South Badung tourism center and smooth flow in the northern region. This consistency between statistical metrics and geographical validity confirms that the developed system is suitable for implementation as a strategic decision-support instrument for traffic management. Future research is recommended to expand the temporal dimension of data acquisition into time-series analysis and integrate dynamic external variables, such as weather conditions and traditional ceremony schedules, to enhance model sensitivity toward traffic anomalies undetected in snapshot data.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the Data Science Undergraduate Program at Telkom University Purwokerto for the academic support and computing facilities provided. These resources were instrumental in facilitating the data analysis process and the successful completion of this research.

AUTHOR CONTRIBUTION STATEMENT

RP initiated the research concept and designed the methodology. RP executed the primary coding and computational tasks and wrote the original draft. MB conducted the spatial data acquisition via API and performed feature engineering. ES supervised the overall research project, provided critical revisions, and validated the methodological approach. KM contributed to the geographic visualization and spatial validity analysis. WS evaluated the model performance and calculated the statistical metrics. YR assisted with the data preprocessing, literature review, and manuscript formatting. All authors have read and approved the final manuscript.

REFERENCES

Abdolmaleki, M., Masoud, N., & Yin, Y. (2020). Transit timetable synchronization for transfer time minimization. *Transportation Research Part B: Methodological*, *131*, 143–159. https://doi.org/10.1016/j.trb.2019.12.002

Azizah Amelia, S. (2025). Utilization Of Google Maps Data And Machine Learning For Traffic Congestion Prediction In Medium-Sized Urban Areas. *Innovative Research in Civil and Environmental Engineering*.

Bazarnovi, S., & Mohammadian, A. (kouros). (2024). Addressing imbalanced data in predicting injury severity after traffic crashes: A comparative analysis of machine learning models. *Procedia Computer Science*, *238*, 24–31. https://doi.org/10.1016/j.procs.2024.05.192

Bührs, M., Rienow, A., & Zepp, H. (2026). Integrating species distribution models (SDM) in spatially explicit GIS-tools to support nature-sensitive urban planning. *Basic and Applied Ecology*, *91*, 61–73. https://doi.org/10.1016/j.baae.2026.01.005

Cai, S., Liu, G., He, J., Du, Y., Si, Z., & Jiang, Y. (2025). Temporal-Spatial Traffic Flow Prediction Model Based on Prompt Learning. *ISPRS International Journal of Geo-Information*, *14*(1). https://doi.org/10.3390/ijgi14010011

Cebolla-Alemany, J., Macarulla Martí, M., Viana, M., Moreno-Martín, V., San Félix, V., & Bou, D. (2024). Optimizing indoor air models through k-means clustering of nanoparticle size distribution data. *Building and Environment*, *266*. https://doi.org/10.1016/j.buildenv.2024.112091

Chen, H., Liu, N. Y., Yang, Y., Wang, Y., Li, Q., & Shan, Y. (2026). Impact Effects of Transport Structure Changes on Urban Traffic Congestion: A Case Study of Core Cities in China. *Transportation Research Part E: Logistics and Transportation Review*, *208*. https://doi.org/10.1016/j.tre.2026.104676

Erdelić, T., Carić, T., Erdelić, M., Tišljarić, L., Turković, A., & Jelušić, N. (2021). Estimating congestion zones and travel time indexes based on the floating car data. *Computers, Environment and Urban Systems*, *87*. https://doi.org/10.1016/j.compenvurbsys.2021.101604

Imoh, U. U., & Rad, M. M. (2025). Analysis and Prediction of Traffic Conditions Using Machine Learning Models on Ikorodu Road in Lagos State, Nigeria. *Infrastructures*, *10*(5). https://doi.org/10.3390/infrastructures10050122

Joelianto, E., Fathurrahman, M. F., Sutarto, H. Y., Semanjski, I., Putri, A., & Gautama, S. (2022). Analysis of Spatiotemporal Data Imputation Methods for Traffic Flow Data in Urban Networks. *ISPRS International Journal of Geo-Information*, *11*(5). https://doi.org/10.3390/ijgi11050310

Khorshidi, N., Rezashoar, S., Amini, P., Zargari, S. A., & Mirzahossein, H. (2025). Metaheuristic-optimized neural networks for travel time index prediction: A comparative study of wild horse and coot optimization algorithms: Optimized MLP for Travel Time Prediction: Comparing WHO and COA. *Transportation Engineering*, *22*. https://doi.org/10.1016/j.treng.2025.100395

Li, J., Wang, L., Fu, S., Fu, W., & Pan, X. (2025). Self-labeled framework with semi-supervised ball K-means clustering-based synthetic example generation for semi-supervised classification in industrial applications. *Engineering Applications of Artificial Intelligence*, *150*. https://doi.org/10.1016/j.engappai.2025.110528

Mehavilla, L., García, J., & Alesanco, Á. (2026). Unveiling user activities on instant messaging platforms: A study of activity fingerprinting through traffic analysis and machine learning techniques. *Knowledge-Based Systems*, *342*. https://doi.org/10.1016/j.knosys.2026.115893

Mirzahossein, H., Nobakht, P., Waller, T., & Lin, D. Y. (2025). Revealing crash hotspots concerning google traffic maps historical data by supervised and ensemble machine learning techniques. *Transportation Engineering*, *20*. https://doi.org/10.1016/j.treng.2025.100326

Ni Made Prisca Rahmayanti, Dewa Ayu Nyoman Sriastuti, & Putu Aryastana. (2024). Kajian Kebutuhan dan Pendapatan Parkir pada Kawasan Objek Wisata Pantai Kelan Kabupaten Badung Bali. *Jurnal Ilmiah Mitsu*. https://doi.org/10.24929/ft.v12i1.2744

Obaidat, M. T., & Alomari, L. D. (2026). Real-time multi-source AI, computer vision, and GIS fusion for data-driven urban traffic management. *Transportation Research Interdisciplinary Perspectives*, *36*. https://doi.org/10.1016/j.trip.2026.101895

Ranjan, S., Kim, Y. C., Ranjan, N., Bhandari, S., & Kim, H. (2023). Large-Scale Road Network Traffic Congestion Prediction Based on Recurrent High-Resolution Network. *Applied Sciences (Switzerland)*, *13*(9). https://doi.org/10.3390/app13095512

Rozy, F., Setia Brata, A., Lestari, W., & Rahmawati, S. (2025). Cluster of Foreign Tourists Using the K-Means Method Based on Arrival Data at I Gusti Ngurah Rai Airport (2018-2024). *International Journal of Management Science and Information Technology*, *5*(2), 362–370. https://doi.org/10.35870/ijmsit.v5i2.5208

Shah, S. M. T., & Piantanakulchai, M. (2026). Ensemble machine learning for travel time prediction: An Active learning-assisted approach. *Results in Engineering*, *29*. https://doi.org/10.1016/j.rineng.2025.108712

Shang, Q., Tang, Y., & Yin, L. (2024). A hybrid model for missing traffic flow data imputation based on clustering and attention mechanism optimizing LSTM and AdaBoost. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-77748-1

Shang, Q., Yu, Y., & Xie, T. (2022). A Hybrid Method for Traffic State Classification Using K-Medoids Clustering and Self-Tuning Spectral Clustering. *Sustainability (Switzerland)*, *14*(17). https://doi.org/10.3390/su141711068

Soimun, A., Sadri, P. D. A., Navianti, D. R., & Fitri, S. N. (2025). The impact of traffic congestion on freight transport efficiency in a tourism corridor: A case study of the Denpasar–Gilimanuk road. *Journal of Applied Sciences in Travel and Hospitality*, *8*(2), 177–198. https://doi.org/10.31940/jasth.v8i2.177-198

Solís-Martín, D., Galán-Páez, J., & Borrego-Díaz, J. (2025). A Model for Learning-Curve Estimation in Efficient Neural Architecture Search and Its Application in Predictive Health Maintenance. *Mathematics*, *13*(4). https://doi.org/10.3390/math13040555

Winfrey, C., Meleby, P., & Miao, L. (2023). Using big data and machine learning to rank traffic signals in Tennessee. *Journal of Traffic and Transportation Engineering (English Edition)*, *10*(5), 918–933. https://doi.org/10.1016/j.jtte.2023.04.005

Xu, S. J., Kong, K. C., Jin, X. B., & Geng, G. G. (2025). Unveiling traffic paths: Explainable path signature feature-based encrypted traffic classification. *Computers and Security*, *150*. https://doi.org/10.1016/j.cose.2024.104283

Yu, T., Huang, W., Tang, X., & Zheng, D. (2025). A hybrid unsupervised machine learning model with spectral clustering and semi-supervised support vector machine for credit risk assessment. *PLOS ONE*, *20*(1). https://doi.org/10.1371/journal.pone.0316557

Zhang, J., Zhang, A., Liu, X., & Luo, P. (2026). Integrating deep learning and big data analytics for enhanced smart city infrastructure and human resource management systems. *Ain Shams Engineering Journal*, *17*(3). https://doi.org/10.1016/j.asej.2025.103962

Zhang, X., Wang, L., Yang, Y., Han, H., Shen, G., Schroepfer, T., & He, J. (2025). Analyzing the typology and livability of 15-minute travel at metro stations in high-density cities: A case study of Singapore. *Cities*, *158*. https://doi.org/10.1016/j.cities.2025.105727

Zheng, Y., Luo, J., & Hu, M. (2026). Spatial-temporal incident-aware dynamic graph convolution networks for traffic flow prediction. *Expert Systems with Applications*, *305*, 130689. https://doi.org/10.1016/j.eswa.2025.130689

Zhu, Y., & Goverde, R. M. P. (2021). Dynamic railway timetable rescheduling for multiple connected disruptions. *Transportation Research Part C: Emerging Technologies*, *125*. https://doi.org/10.1016/j.trc.2021.103080