Journal of Intelligent Decision Support System (IDSS) 6 (4) (2023), pp. 223-230

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Optimizing Urban Traffic Management Through Advanced Machine Learning: A Comprehensive Study

Fahmi Izhari[1], Hanna Willa Dhany[2]
[1,2] *University of Pembangunan Panca Budi, Medan, Indonesia*

Article Info

Article history:

Received Nov 14, 2023
Revised Nov 20, 2023
Accepted Nov 30, 2023

Keywords:

Accuracy
CatBoost,
Classification,
Prediction,
Traffic.

ABSTRACT

Urban transport networks are vital components of modern societies, influencing efficiency and safety. This research explores the potential of traffic data as a crucial information source for forecasting and interpreting traffic problems. Using advanced data processing, statistical analysis, and classification algorithms, the study aims to identify and forecast traffic scenarios. With an interdisciplinary approach integrating computer science, statistics, and transportation engineering, the research emphasizes a holistic perspective on traffic concerns. The study involves outlier detection, label encoding, and cutting-edge technologies like GridSearchCV and ensemble modeling. Inspired by flash flood susceptibility research, machine learning models, particularly LightGBM and CatBoost, are applied to predict traffic situations. DecisionTreeClassifier and CatBoostClassifier emerge as top performers, achieving remarkable accuracies. The evaluation goes beyond accuracy, emphasizing the nuanced understanding of algorithm strengths and limitations for effective urban transportation network management.

Corresponding Author:

Fahmi Izhari,
Faculty of Science and Technology,
University of Panca Budi,
Jalan Gatot Subroto, Medan Sunggal, Sumatera Utara, Medan, Indonesia.
Email: fahmi_izhari@dosen.pancabudi.ac.id

Introduction

Urban transport networks have evolved into sophisticated systems crucial for societal mobility amidst urbanization and the expansion of information technology(Anthony Jnr, 2023; Pathak & Upadhyay, 2023; Zhan et al., 2023). Traffic, as an integral aspect, significantly influences efficiency and safety, necessitating a thorough understanding of traffic patterns through data analysis for optimal urban transport network management.

The research project aims to explore the potential of traffic data as a crucial information source for forecasting and interpreting traffic problems. Beyond time-related factors, traffic data includes additional variables such as days and traffic-specific features. The study employs data processing techniques and statistical analysis to construct classification algorithms focused on accurately identifying and forecasting traffic scenarios. This study is significant as it contributes to the development of adaptive and responsive traffic management systems. Sophisticated categorization models are believed to enhance traffic management decision-making, ultimately improving the efficiency and safety of urban transportation. This interdisciplinary endeavor integrates fundamental concepts from computer science, statistics, and transportation engineering, offering a holistic perspective on traffic concerns and proposing practical solutions applicable to urban environments.

Preliminary procedures involve outlier detection algorithms and label encoding to assure data quality and applicability. The research goes beyond building classification models, considering the interpretability of models to gain insights useful in decision-making, which involves data categorization into different categories according to the rules(Izhari & Willa Dhany, 2020). This scientific technique heavily relies on recent technologies like GridSearchCV and ensemble modeling with a voting classifier. The project aims to develop optimal models for predicting traffic problems using these methodologies. The findings of this study are expected to make a real difference in the sustainability and efficiency of urban mobility. Transportation resources can be directed more wisely and adaptively with a greater awareness of traffic patterns and the ability to foresee traffic scenarios.

Incorporating previous research led by Mohamed Sabre et al., this study pioneers the use of machine learning models, specifically LightGBM and CatBoost, for predicting flash flood susceptibility in Hurghada, Egypt's Wadi System. With AUROC scores exceeding 97%, CatBoost demonstrated superior performance in classification metrics and processing time compared to LightGBM and random forest. High population density areas were identified as most vulnerable, validating CatBoost's efficacy for flash flood susceptibility mapping and emphasizing its superior performance and efficiency(Saber et al., 2022).

Building upon the foundation of these technological advancements, the research presented here aims to extend the application of data-driven approaches to traffic management(Erfani et al., 2023; Liu et al., 2023; Nallaperuma et al., 2019; Su et al., 2021; Tiong et al., 2023; Xu et al., 2023; Ye et al., 2023). By leveraging insights gained from machine learning models, the study seeks to refine the classification algorithms, ensuring not only accuracy but also interpretability(Arya et al., 2023; Hong et al., 2020; Kigo et al., 2023; Mr.Gugloth Ganesh Dr.Pilla Srinivas Rudrapati Mounika Dudekula Rabiya Begum Kaki Leela Prasad, 2023; Rudin et al., 2022; Su et al., 2021). This emphasis on interpretability aligns with the evolving landscape of urban challenges, where transparent decision-making processes become increasingly essential.

Furthermore, the synergistic amalgamation of computer science, statistics, and transportation engineering within this interdisciplinary initiative strategically situates it at the nexus of cutting-edge research in urban mobility. By fostering a collaborative environment that taps into the expertise of diverse fields, the initiative aims to cultivate a nuanced understanding of the intricate dynamics shaping urban traffic. This holistic approach is not only vital for unraveling the complexities of urban environments but also serves as the bedrock for crafting innovative and pragmatic strategies. These strategies are meticulously tailored to overcome the distinctive challenges posed by urban settings, promising to usher in a new era of elevated traffic management and enhanced urban mobility.

The early stages of this research involve an intricate dance with data, employing sophisticated techniques such as outlier detection algorithms and label encoding for comprehensive data processing. This meticulous commitment to data quality ensures the precision and relevance of the dataset within the specialized context of traffic analysis. The significance of this groundwork extends to the construction of robust models, where advanced technologies like GridSearchCV and ensemble modeling with a voting classifier take center stage. This technological prowess not only underscores the initiative's commitment to methodological excellence but also pushes the boundaries of what is achievable in the realm of traffic prediction.

The ultimate goal is not just to develop models but to create optimal models that transcend the conventional limitations of traffic prediction. This forward-thinking approach aligns seamlessly with the broader mission of enhancing the sustainability and efficiency of urban mobility. The initiative, in its essence, is a pioneering force that goes beyond statistical comprehension; it aspires to catalyze tangible improvements across the vast and dynamic landscape of urban transport networks.

By laying this foundation, the initiative doesn't merely seek to understand traffic; it aspires to be a catalyst for transformative advancements in effective and sustainable traffic understanding and management. The research endeavors to shape the trajectory of urban development, aligning with the evolving needs of modern societies. This holistic perspective underscores the initiative's significance in propelling urban mobility towards a future that is not only efficient but also inherently sustainable and responsive to the ever-evolving challenges of urban living.

Method

In this research endeavor, a meticulous and systematic approach was undertaken to optimize urban traffic management through advanced machine learning techniques. The methodological progression unfolded as follows:

```
Load Dataset
Preprocess Data
Explore Data (Visualize)
Detect Outliers
Train-Test Split and Label Encoding
Standardize Features
Model Selection and Hyperparameter Tuning
 For each model:
   - Use GridSearchCV for hyperparameter tuning
   - Store best estimator and its score
Model Training and Evaluation
 For each model:
   - Train and evaluate on the test set
   - Store prediction results and time
Ensemble Model (Voting Classifier)
 - Combine selected models into a voting ensemble
 - Train and predict using the ensemble
 - Evaluate ensemble model accuracy
Find Best Model
 - Identify the index of the model with the highest accuracy
 - Retrieve information about the best model and its accuracy
Plot Accuracy vs Prediction Time
 - Plot a bar chart of prediction times for each model
End
```

Figure 1. Pseudocode

The process initiates by loading and preprocessing the dataset, addressing missing values and duplicates. Visual exploration via histograms and outlier detection enhances dataset reliability. Splitting into training and testing sets, label encoding, and feature standardization follow. Model selection involves various algorithms, and GridSearchCV refines hyperparameters, storing the best estimator and its score. Models are then trained and evaluated, recording prediction results and time. An ensemble model, created with a Voting Classifier, combines models for enhanced performance. It is trained and evaluated for accuracy. Identifying the best model involves pinpointing the index with the highest accuracy score. A bar chart visually compares prediction times, concluding the process.

The meticulously crafted workflow, as outlined in the pseudocode, guides a comprehensive analysis of an urban traffic dataset. This dataset, derived through computer vision, provides a detailed snapshot of traffic dynamics, capturing hourly counts of various vehicle types and their classifications into distinct traffic situations. The research focuses on contributing to urban traffic management, emphasizing sustainability and efficiency in urban mobility. This dataset, featuring hourly counts of vehicles, aligns with the workflow's steps, from preprocessing to ensemble model evaluation. The classifications, spanning from Heavy to Low congestion, enable a nuanced analysis, supporting data-driven decisions for sustainable urban mobility.

Table 1. Dataset

	Time	Date	Day of the week	CarCount	BikeCount	BusCount	TruckCount	Total	Traffic Situation
0	12:00:00 AM	10	Tuesday	31	0	4	4	39	low
1	12:15:00 AM	10	Tuesday	49	0	3	3	55	low
2	12:30:00 AM	10	Tuesday	46	0	3	6	55	low
3	12:45:00 AM	10	Tuesday	51	0	2	5	58	low
4	1:00:00 AM	10	Tuesday	57	6	15	16	94	normal
...	...	...	...	...	...	...	...	...	...
2971	10:45:00 PM	9	Thursday	16	3	1	36	56	normal
2972	11:00:00 PM	9	Thursday	11	0	1	30	42	normal
2973	11:15:00 PM	9	Thursday	15	4	1	25	45	normal
2974	11:30:00 PM	9	Thursday	16	5	0	27	48	normal
2975	11:45:00 PM	9	Thursday	14	3	1	15	33	normal

Results and Discussions

The research outcomes, driven by advanced machine learning, present a thorough exploration of urban traffic dynamics. Commencing with dataset preprocessing, including handling missing values and duplicates, the methodology utilizes visual exploration, such as histograms and outlier detection. Subsequent steps involve dataset splitting, label encoding, and feature standardization. Model selection, refined through GridSearchCV, captures the best estimator and its score. After training and evaluation, an ensemble model, formed with a Voting Classifier, undergoes training and evaluation for accuracy.

The intricate hyperparameter optimization process unfolded with multiple model candidates, each subjected to 10-fold cross-validation. The notation "Fitting 10 folds for each of [n] candidates, totalling [total fits]" denotes the extensive search across hyperparameter configurations for each model candidate. In this specific context, the process involved evaluating 18, 3, 28, 375, 100, 5, 6, and 9 different candidate hyperparameter sets, respectively, with each undergoing 10-fold cross-validation. The varying numbers of candidates reflect the diverse sets of hyperparameters explored, resulting in a cumulative count of fits for each model—180, 30, 280, 3750, 1000, 50, 60, and 90 fits, respectively. This exhaustive process aims to identify the optimal hyperparameter configuration for each model, ensuring robust performance on the urban traffic dataset.

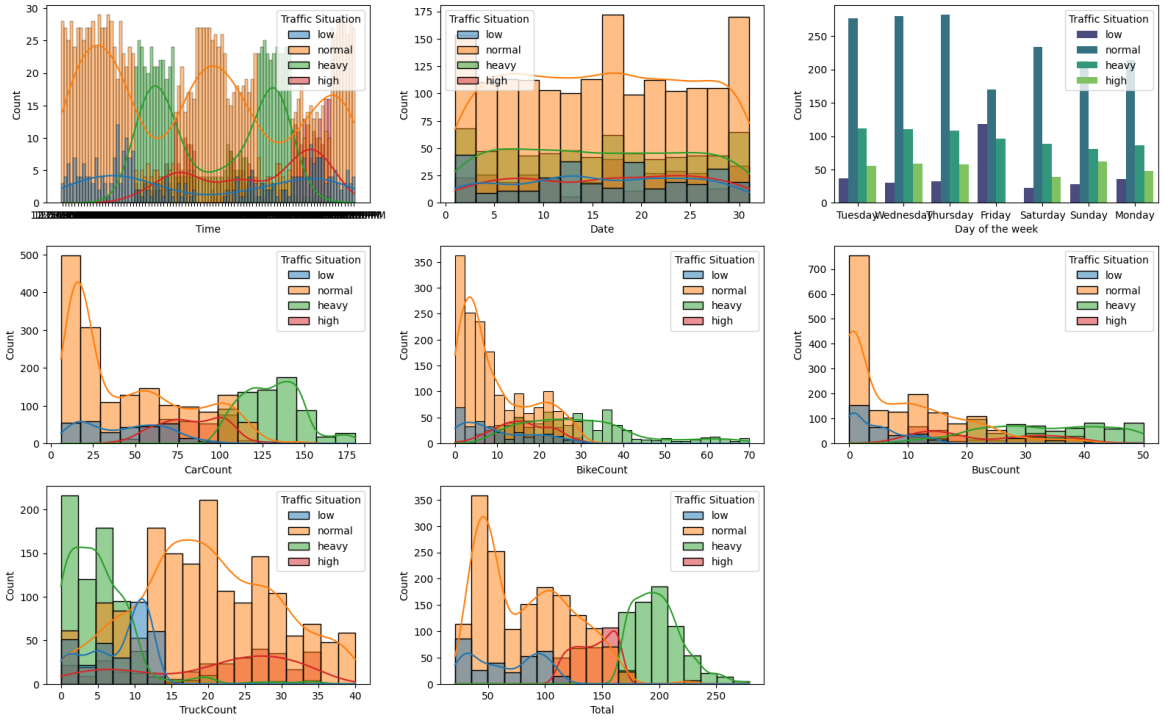

Figure 2. Traffic Situation Distribution

The output of the provided code showcases a series of visualizations illustrating the relationship between various factors, such as time, date, day of the week, car count, bike count, bus count, truck count, and total vehicle count, with different traffic situations. Each subplot within the larger figure presents a distinct perspective on these relationships, employing color differentials to represent different categories of traffic situations. Collectively, these visualizations serve as a powerful exploratory data analysis (EDA) tool, facilitating a nuanced understanding of how different factors contribute to and correlate with various traffic situations. The color-coded distinctions enhance interpretability, allowing for a comprehensive analysis of the dataset.

In the course of this study, an extensive evaluation was conducted to assess the performance of various classification algorithms in predicting traffic situations using the provided dataset. The algorithms subjected to scrutiny encompassed a broad spectrum, including RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, DecisionTreeClassifier, KNeighborsClassifier, XGBClassifier, AdaBoostClassifier, and CatBoostClassifier. Each algorithm underwent a rigorous optimization process through GridSearchCV, allowing for the identification of optimal parameter configurations to enhance their predictive capabilities. The ensuing table serves as a comprehensive repository of accuracy outcomes for the various algorithms employed in our study. The presentation is meticulously structured, adhering to a scientific approach that facilitates clear comprehension and comparison of the predictive performance of each algorithm. This approach aids in distilling valuable insights from the dataset and contributes to the broader understanding of algorithmic efficiency in the domain of traffic situation prediction.

Table 2. Algorithms Accuracy in Predicting Traffic Conditions

Algorithm	Accuracy
Random Forest Classifier	0.991011
Gradient Boosting Classifier	0.993258
Logistic Regression	0.818427
Decision Tree Classifier	0.994382
Kneighbors Classifier	0.886754
XGB Classifier	0.993258
AdaBoost Classifier	0.820649
CatBoost Classifier	0.996629

The results of this systematic experimentation revealed intriguing insights into the efficacy of the evaluated models. DecisionTreeClassifier and CatBoostClassifier emerged as the top-performing models, achieving remarkable accuracies of 99.44% and 99.66%, respectively. The DecisionTreeClassifier demonstrated robust performance, leveraging decision tree decisions with the 'entropy' criterion for effective data separation. Conversely, CatBoostClassifier, specifically designed to handle categorical data, exhibited exceptional accuracy, outperforming all other algorithms in the test suite.

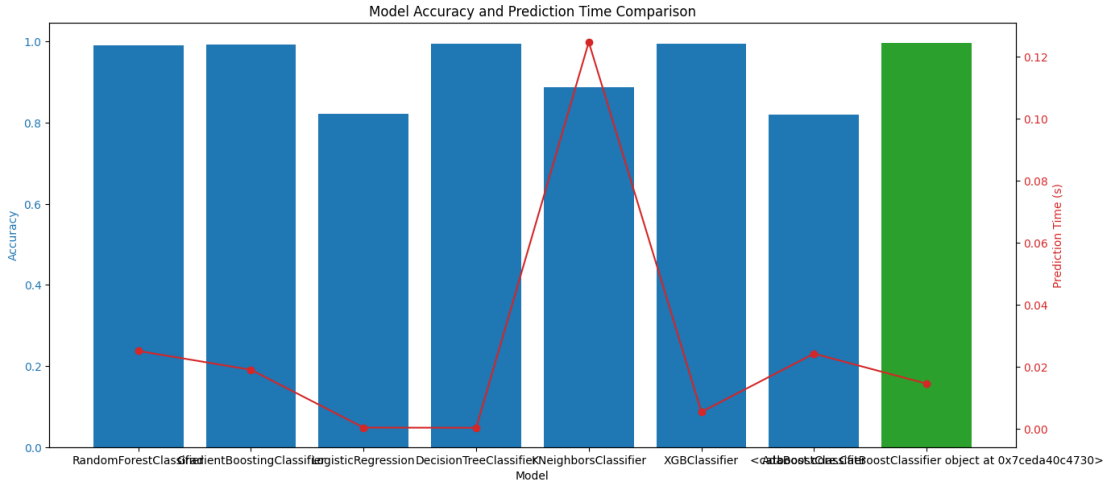

Figure 3. Model Accuracy and Prediction Time Comparison

However, the evaluation went beyond a mere accuracy comparison, the evaluation considered contextual factors. Despite LogisticRegression yielding a lower accuracy of approximately 81.84%, it was deliberately included among the selected models. This strategic decision underscores the recognition that, depending on factors such as interpretability and prediction speed, the optimal model choice may vary. The study significantly contributes to a nuanced understanding of classification algorithm performance in predicting traffic situations, offering valuable guidance for practitioners and researchers. It emphasizes that while accuracy is crucial, a comprehensive exploration of the strengths and limitations of each algorithm is vital for informed decision-making in the realm of urban transportation network management. The results are summarized in Table 2 and visually presented in Figure 3, providing a comprehensive overview of the study's findings.

Conclusions

In summary, this research delves into the intricate dynamics of urban transport networks, emphasizing the pivotal role of traffic in influencing efficiency and safety. Through advanced data processing, statistical analysis, and classification algorithms, the study explores traffic data's potential for accurate forecasting and interpretation, employing an interdisciplinary approach that integrates concepts from computer science, statistics, and transportation engineering.

The project employs outlier detection and label encoding in the early stages to ensure data quality. Emphasizing interpretability in model construction, the research leverages cutting-edge technologies such as GridSearchCV and ensemble modeling for optimal traffic prediction models. Drawing inspiration from flash flood susceptibility research, machine learning models like LightGBM and CatBoost pioneer traffic prediction, with DecisionTreeClassifier and CatBoostClassifier emerging as top performers, achieving accuracies of 99.44% and 99.66%, respectively.

Beyond accuracy, the evaluation deliberately includes LogisticRegression, underscoring the importance of contextual factors in model selection. The study provides valuable insights for practitioners and researchers, emphasizing the nuanced understanding of algorithm strengths and limitations crucial for effective urban transportation network management.

Further studies should focus on improving model interpretability, addressing potential imbalances in the dataset, considering multi-modal transportation for a more comprehensive perspective, and evaluating model generalisation across diverse urban settings. By resolving these issues, future research efforts will be able to move the field towards more accurate, interpretable, and broadly applicable models, thereby contributing to the improvement of urban mobility and traffic management.

References

Agarwal, R., Frosst, N., Zhang, X., Caruana, R. and Hinton, G. E. (2020). Neural additive models: Interpretable machine learning with neural nets. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning.

Aghaei, S., Gomez, A. and Vayanos, P. (2020). Learning Optimal Classification Trees: Strong Max-Flow Formulations. arXiv e-print arXiv:2002.09142.

Anthony Jnr, B. (2023). Sustainable mobility governance in smart cities for urban policy development – a scoping review and conceptual model. *Smart and Sustainable Built Environment*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/SASBE-05-2023-0109

Arya, G., Bagwari, A., Saini, H., Thakur, P., Rodriguez, C., & Lezama, P. (2023). Explainable AI for Enhanced Interpretation of Liver Cirrhosis Biomarkers. *IEEE Access*, *11*, 123729–123741. https://doi.org/10.1109/ACCESS.2023.3329759

Erfani, A., Cui, Q., Baecher, G., & Kwak, Y. H. (2023). Data-Driven Approach to Risk Identification for Major Transportation Projects: A Common Risk Breakdown Structure. *IEEE Transactions on Engineering Management*, 1–12. https://doi.org/10.1109/TEM.2023.3279237

Gunl ¨ uk, O. ¨ , Kalagnanam, J., Li, M., Menickelly, M. and Scheinberg, K. (2021). Optimal decision trees for categorical data via integer programming. Journal of Global Optimization 1–28.

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW1). https://doi.org/10.1145/3392878

Izhari, F., & Willa Dhany, H. (2020). COMPARISON OF AIR QUALITY DATA ACCURATION USING DECISION TREE AND NEURAL NETWORK METHOD. *JURNAL IPTEKS TERAPAN (Research of Applied Science and Education)*, *14*(2), 123–127.

Kigo, S. N., Omondi, E. O., & Omolo, B. O. (2023). Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Scientific Reports*, *13*(1), 17315. https://doi.org/10.1038/s41598-023-44326-w

Kursuncu, U., Gaur, M. and Sheth, A. (2020). Knowledge Infused Learning (K-IL): Towards Deep Incorporation of Knowledge in Deep Learning. In Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE 2020) I.

Liu, Z., Lyu, C., Wang, Z., Wang, S., Liu, P., & Meng, Q. (2023). A Gaussian-Process-Based Data-Driven Traffic Flow Model and Its Application in Road Capacity Analysis. *IEEE Transactions on Intelligent Transportation Systems*, *24*(2), 1544–1563. https://doi.org/10.1109/TITS.2022.3223982

Mr.Gugloth Ganesh Dr.Pilla Srinivas Rudrapati Mounika Dudekula Rabiya Begum Kaki Leela Prasad, Dr. K. V. K. (2023). Efficient Classification of Brain Tumor Images Using Neural Network Technique. *Journal of Advanced Zoology*, *44*(S-2), 1381–1395. http://jazindia.com/index.php/jaz/article/view/974

Nallaperuma, D., Nawaratne, R., Bandaragoda, T., Adikari, A., Nguyen, S., Kempitiya, T., Silva, D. De, Alahakoon, D., & Pothuhera, D. (2019). Online Incremental Machine Learning Platform for Big Data-Driven Smart Traffic Management. *IEEE Transactions on Intelligent Transportation Systems*, *20*(12), 4679–4690. https://doi.org/10.1109/TITS.2019.2924883

Pathak, S., & Upadhyay, R. K. (2023). Macro level performance study of Ahmadabad bus rapid transit system: Janmarg. *Green Energy and Intelligent Transportation*, *2*(3), 100093. https://doi.org/https://doi.org/10.1016/j.geits.2023.100093

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*(none), 1–85. https://doi.org/10.1214/21-SS133

Rudin, C. and Schapire, R. E. (2009). Margin-Based Ranking and an Equivalence between AdaBoost and RankBoost. Journal of Machine Learning Research 10 2193–2232.

Saber, M., Boulmaiz, T., Guermoui, M., Abdrabo, K. I., Kantoush, S. A., Sumi, T., Boutaghane, H., Nohara, D., & Mabrouk, E. (2022). Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. *Geocarto International*, *37*(25), 7462–7487. https://doi.org/10.1080/10106049.2021.1974959

Su, Z. C., Chow, A. H. F., & Zhong, R. X. (2021). Adaptive network traffic control with an integrated model-based and data-driven approach and a decentralised solution method. *Transportation Research Part C: Emerging Technologies*, *128*, 103154. https://doi.org/https://doi.org/10.1016/j.trc.2021.103154

Tiong, K. Y., Ma, Z., & Palmqvist, C.-W. (2023). A review of data-driven approaches to predict train delays. *Transportation Research Part C: Emerging Technologies*, *148*, 104027. https://doi.org/https://doi.org/10.1016/j.trc.2023.104027

Xu, H., Sun, Z., Cao, Y., & Bilal, H. (2023). A data-driven approach for intrusion and anomaly detection using automated machine learning for the Internet of Things. *Soft Computing*, *27*(19), 14469–14481. https://doi.org/10.1007/s00500-023-09037-4

Ye, F.-F., Yang, L.-H., Wang, Y.-M., & Lu, H. (2023). A data-driven rule-based system for China's traffic accident prediction by considering the improvement of safety efficiency. *Computers & Industrial Engineering*, *176*, 108924. https://doi.org/https://doi.org/10.1016/j.cie.2022.108924

Zhan, L., Wang, S., Xie, S., Zhang, Q., & Qu, Y. (2023). Spatial path to achieve urban-rural integration development – analytical framework for coupling the linkage and coordination of urban-rural system functions. *Habitat International*, *142*, 102953. https://doi.org/https://doi.org/10.1016/j.habitatint.2023.102953

Zhang, Y., Zong, R., Shang, L., & Wang, D. (2023). A crowd-AI dynamic neural network hyperparameter optimization approach for image-driven social sensing applications. *Knowledge-Based Systems*, *278*, 110864. https://doi.org/https://doi.org/10.1016/j.knosys.2023.110864