International Journal of Science, Strategic Management and Technology
Volume **02** Issue **03** March-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

# Multi-Model Predictive and Explainable Public Transport Overcrowding System

**N.Manikandan[#1], Mrs. B. Revathi[*2]**

[1]*Department of Artificial Intelligence and Data Science, Ramco Institute of Technology,Rajapalayam,India-626117*

[1]manikandan76112@gmail.com

[2]*Department of Artificial Intelligence and Data Science, Ramco Institute of Technology,Rajapalayam,India-626117*

[2]revas85@gmail.com

**https://doi.org/**10.55041/ijsmt.v2i3.402

**Cite this Article**: N.Manikandan, & Revathi, B. (2026). Multi-Model Predictive and Explainable Public Transport Overcrowding System. International Journal of Science, Strategic Management and Technology, *02*(03). https://doi.org/10.55041/ijsmt.v2i3.402



***Abstract***— This paper presents a machine learning–based framework for predicting overcrowding levels in public transport systems. The proposed system analyzes historical passenger flow data and key operational parameters to classify crowd levels into LOW, MEDIUM, and HIGH categories. Random Forest and XG Boost algorithms are employed as base classifiers, and their outputs are combined using an ensemble strategy to enhance prediction accuracy and robustness. The framework further integrates SHAP-based interpretability to explain model decisions and identify the most influential features affecting crowd levels. In addition, a real-time simulation module is incorporated to compare predicted and observed crowd conditions, enabling deviation detection and adaptive monitoring. Experimental evaluation demonstrates strong classification performance across all crowd categories, confirming the reliability and effectiveness of the ensemble approach. The proposed model supports intelligent transport management by providing accurate crowd predictions and interpretable insights, thereby assisting authorities in proactive decision-making, resource allocation, and passenger information systems. The overall system contributes to improving operational efficiency and enhancing commuter safety within urban public transportation networks

***Keywords***— Machine Learning, Random Forest, XG Boost, SHAP Interpretability, Real-Time Deviation Detection.

## 1. INTRODUCTION

Public transportation plays a very important role in daily life, especially in developing regions and densely opulated cities. Buses are one of the most widely used transport modes because they are affordable and accessible. However, overcrowding in buses remains a major challenge. During peak hours, weekends, holidays, or special events, passenger demand increases suddenly. This leads to congestion, increased travel delays, discomfort, and sometimes safety concerns.

In many cases, passengers decide their travel time without knowing the expected crowd level. As a result, they may experience unexpected congestion. Similarly, transport authorities often respond reactively instead of proactively because they lack predictive systems that estimate demand in advance. Therefore, there is a strong need for intelligent systems that can forecast crowd levels before the bus arrives at a stop.

Recent advancements in machine learning and data-driven modelling provide an opportunity to solve such real-world problems. Machine learning models can analyze patterns

from historical data and identify relationships between time, passenger demand, delays, and external factors such as weather or seasonal trends. By learning these patterns, models can predict future crowd conditions with reasonable accuracy.

However, building such a system is not straightforward. Crowd behaviour is influenced by multiple dynamic factors:

a. Time of the day (morning and evening rush hours)

b. Day of the week (weekday vs weekend travel)

c. Seasonal variations (vacation months or festival periods)

d. Weather conditions such as rain or fog

e. Unexpected events causing sudden demand spikes

f. Delay propagation due to traffic congestion

All these variables interact with each other. For example, heavy rainfall during evening peak hours can significantly increase delay and passenger buildup at bus stops. Therefore, a simple rule-based system is not sufficient. A predictive model must consider temporal, environmental, and behavioural patterns together

In this project, a multi-city public transport dataset was generated using a temporal simulation approach. Instead of random values, the dataset includes autoregressive passenger buildup, delay propagation, event-based surges, seasonal multipliers, and bus capacity constraints. This makes the dataset more realistic and suitable for training machine learning models.

To perform classification, ensemble learning methods are adopted. Random Forest is used because of its robustness and ability to handle nonlinear relationships. XG Boost is used due to its strong gradient boosting performance and capability to capture complex feature interactions. The outputs of these models are combined to improve prediction stability. Crowd levels are categorized into LOW, MEDIUM, and HIGH classes, making the output simple and understandable for users.

Another important aspect of this system is interpretability. In real-world applications, especially in transportation management, it is not enough to simply predict a label. Authorities need to understand *why* a certain crowd level was predicted. For this purpose, SHAP-based explanation techniques are integrated to identify the most influential features for each prediction, such as high passenger demand, peak hour indicator, or increasing delays.

The system is implemented as a Flask-based REST API, allowing frontend integration for user interaction. When a user selects route, stop, travel date, and time, the system processes temporal features, performs prediction, detects deviations, and generates suggestions. It also provides possible authority-level actions, such as deploying extra buses during high crowd conditions

Overall, this project aims to demonstrate how machine learning, temporal modelling, and explainable AI techniques can be combined to create a practical crowd prediction framework for public transport systems. The goal is not only to improve passenger experience but also to support smarter and more proactive transport management strategies.

## 2. LITERATURE REVIEW

Over the last few decades, research in intelligent transportation systems (ITS) and traffic forecasting has grown rapidly due to the increasing availability of urban mobility data and smart sensing technologies. Accurate prediction of traffic flow, passenger demand, congestion patterns, and urban mobility trends has become essential for sustainable city planning and efficient transportation management. Early forecasting approaches were primarily based on classical statistical time-series models. The foundational work of George Box and Gwilym Jenkins [4] introduced ARIMA-based forecasting methods, which became widely used for short-term traffic prediction due to their mathematical rigor and interpretability. However, these models often struggled to capture nonlinear and highly dynamic traffic behaviours in large urban networks.

With the advancement of machine learning, researchers began adopting ensemble-based techniques for traffic modelling. Leo Breiman [1] proposed Random Forests, which improved prediction robustness by combining multiple decision trees. Similarly, Chen and Guestrin [2] introduced XGBoost, a scalable gradient boosting framework capable of handling large-scale transportation datasets with high efficiency and accuracy. These models significantly enhanced forecasting performance compared to traditional statistical approaches.

The emergence of deep learning further transformed traffic prediction research. Foundational works such as *Deep Learning* by Ian Goodfellow et al. [22] and *Pattern*

*Recognition and Machine Learning* by Christopher M. Bishop [23] provided theoretical frameworks that influenced modern transportation analytics. Lv et al. [7] demonstrated the application of deep neural networks for traffic flow prediction using large-scale datasets, while Ma et al. [20] applied deep learning models to predict congestion evolution in complex transportation networks. These approaches effectively captured nonlinear patterns and temporal dependencies in traffic data.

To better model spatial and temporal correlations simultaneously, researchers introduced convolutional and recurrent neural network architectures. Ma et al. [8] treated traffic states as spatial images and applied convolutional neural networks (CNNs) for large-scale speed prediction. Li et al. [10] proposed Diffusion Convolutional Recurrent Neural Networks (DCRNN), integrating graph diffusion mechanisms with recurrent units to model traffic dynamics more accurately. Further advancements were made through spatio-temporal graph-based models. Yu et al. [15] developed Spatio-Temporal Graph Convolutional Networks (STGCN), while Wu et al. [18] proposed Graph WaveNet to learn adaptive graph structures for traffic forecasting. These graph-based frameworks significantly improved prediction accuracy in interconnected road networks.

Urban computing and big data analytics have also played a crucial role in intelligent transportation research. Zheng [6] highlighted the integration of heterogeneous urban data sources for smarter city management. Zhu et al. [25] and Nguyen et al. [21] provided comprehensive surveys on big data analytics and deep learning applications in ITS, identifying challenges such as scalability, real-time processing, and data heterogeneity.

Passenger demand forecasting has become another important research area, particularly in urban rail and taxi systems. Moreira-Matias et al. [13] utilized streaming data for taxi demand prediction, while Yao et al. [16], [17] proposed deep multi-view spatio-temporal networks for improved taxi demand forecasting. Yang et al. [14] applied deep learning models to short-term passenger volume prediction in urban rail systems using smart-card data. Additionally, Zhang et al. [11] examined the environmental impact of urban rail transit based on passenger demand forecasts, linking predictive modelling with sustainability objectives.

Interpretability and transparency of predictive models have also gained importance in recent years. Lundberg and Lee [3] introduced a unified framework for interpreting machine learning predictions, enhancing the reliability of intelligent transportation systems. Furthermore, Rasouli and Tsotsos [24] reviewed autonomous vehicle interactions with pedestrians, emphasizing the importance of predictive accuracy and safety in modern mobility systems.

Overall, the literature demonstrates a clear transition from traditional statistical forecasting methods to advanced machine learning, deep learning, and graph-based architectures for traffic and passenger demand prediction. Although significant improvements have been achieved in prediction accuracy and scalability, challenges remain in model interpretability, computational efficiency, and real-time deployment. These studies provide a strong foundation for developing an intelligent, data-driven, and scalable transportation forecasting framework aligned with modern ITS requirements.

## 3. PROPOSED METHODOLOGY

The proposed methodology presents a data-driven predictive analytics framework designed to estimate public transport overcrowding levels using historical and simulated real-time transport data. The system integrates multiple machine learning models with explainable AI techniques to provide accurate crowd level prediction and decision-support insights.

The framework follows a structured pipeline consisting of data acquisition, preprocessing, feature engineering, model training, ensemble prediction, real-time simulation, and interpretability analysis. Tree-based ensemble learning models are used for classification, while explainability techniques provide transparency in prediction outcomes. The system ultimately supports operational decision-making for transport authorities and passengers through predictive alerts and analytical dashboards.

i. Dataset Collection and Cleaning
ii. Data Preprocessing
iii. Feature Extraction
iv. Machine Learning Model Training
v. Ensemble Prediction
vi. Real-Time Simulation and Alerts
vii. Explainability Analysis

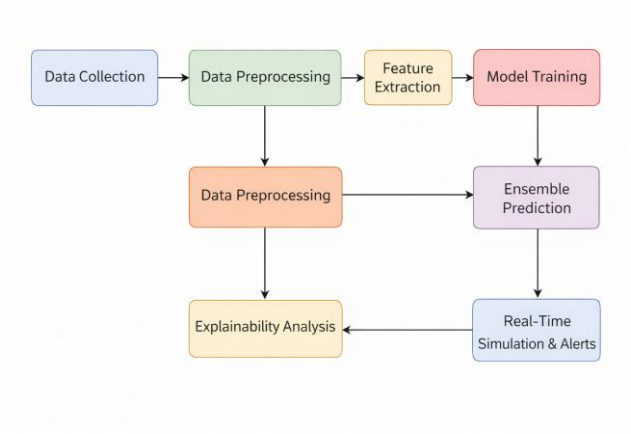


Fig. 1 System Architecture of the Public Transport Overcrowding Prediction System.

### 3.1 Dataset Collection and Cleaning

The system utilizes transport-related datasets containing passenger demand and operational information. Data sources include historical transport records and simulated real-time inputs representing dynamic travel conditions.
Data cleaning ensures quality and consistency through:
1) Removal of missing, duplicate, or inconsistent records
2) Standardization of data formats and units
3) Noise filtering and validation of operational constraints
These steps ensure reliable input for downstream modelling and prediction tasks.

### 3.2Data Preprocessing

Preprocessing transforms raw transport data into structured numerical form suitable for machine learning:
1) Normalization and scaling of numerical features.
2) Encoding of categorical variables such as routes or time categories
3) Temporal structuring of data for trend-based modeling
This stage improves model learning stability and prediction accuracy.

### 3.3Feature Extraction

Feature engineering identifies relevant variables that influence overcrowding patterns.:
Extracted features include:
1) Temporal indicators such as peak hours and travel intervals
2) Operational patterns such as passenger growth trends and service frequency
3) Contextual factors influencing demand variations
These features capture behavioral and temporal dependencies essential for accurate prediction.

### 3.4Machine Learning Model Training

Supervised learning models are trained to classify transport crowd levels into predefined categories.
The system employs:
1)Random Forest for robust nonlinear classification
2) XGBoost for gradient-boosted predictive learning
3) Multi-class classification for crowd level prediction (Low, Medium, High)

### 3.5Ensemble Prediction

Predictions from individual models are combined to improve overall reliability.
The ensemble mechanism performs:
1)Aggregation of model outputs using weighted averaging or voting
2)Final classification of crowd levels
3)Improved stability and reduced prediction variance
This fusion strategy enhances predictive performance compared to individual models.

### 3.6Real-Time Simulation and Alerts

The system simulates dynamic transport conditions to support operational decision-making.
This module performs:
1)Continuous monitoring of simulated passenger occupancy
2)Detection of abnormal crowd buildup or demand surges
3)Generation of alerts and recommendations for service adjustment
It enables proactive response to predicted congestion.

### 3.7Explainability Analysis

Explainable AI techniques are applied to interpret prediction outcomes and identify influencing factors.
The system provides:
1)Feature importance analysis to identify key drivers of overcrowding
2)Model reasoning insights for prediction transparency
3)Interpretable outputs for decision support
This improves trust and usability of the predictive system.

### 3.8 System Output and Decision Support

The framework delivers actionable insights through a monitoring interface that supports transport management and passenger planning.
Outputs include:
1)Predicted crowd levels and alerts
2)Analytical explanations of influencing factors
3)Decision-support recommendations for operational planning
These outputs enable informed interventions to reduce congestion and improve transport efficiency.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **03** March-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

## 4. RESULTS AND DISCUSSIONS

### 4.1 Crowd Prediction Dashboard Module

The input provided to the system consists of user-selected travel details including route (Tirunelveli–Madurai–Chennai), boarding stop (Madurai), travel date, and travel time. These inputs are transmitted to the backend prediction engine, where the trained ensemble model (Random Forest + XGBoost) processes the request and predicts the expected crowd level.

In the below Fig. 1 represents the following concise output:

a. Predicted Crowd Level: MEDIUM
b. Deviation Status: NORMAL
c. Transport Mode: Bus

The result indicates that, based on historical and temporal patterns, the crowd density at the selected stop and time is moderate. The deviation status being "Normal" suggests that the simulated real-time monitoring system did not detect any unexpected surge or sudden fluctuation in crowd levels.

This demonstrates that the system successfully integrates predictive modeling with real-time validation logic to assist passengers in travel decision-making.

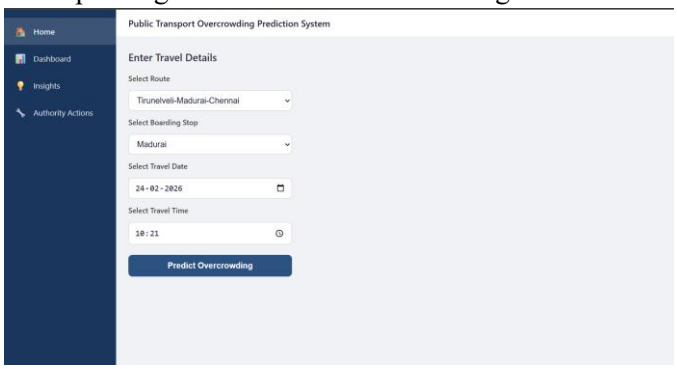


Fig. 2 User Travel Input Interface for Crowd Prediction

### 4.2 Explainable AI (SHAP) Feature Importance Module

The system utilizes SHAP (SHapley Additive exPlanations) to interpret the predictions generated by the ensemble machine learning models. Feature importance visualization provides insight into how different variables contribute to the crowd classification.

In the below Fig. 2 represents the SHAP-based feature importance output.

The graph shows that:

a) Delay Minutes has the highest impact on prediction.
b) Passenger Count significantly influences classification.
c) Rolling Passenger Mean contributes to trend-based forecasting.
d) Peak Hour and Hour moderately affect crowd levels.
e) Passenger Growth and Delay Growth provide short-term dynamic adjustments.
f) Weekday and Weekend indicators have minimal impact compared to operational factors.

The results indicate that operational factors such as delays and passenger accumulation trends are stronger predictors of overcrowding than calendar-based features alone. This confirms that congestion buildup and service delays are primary drivers of crowd intensity.

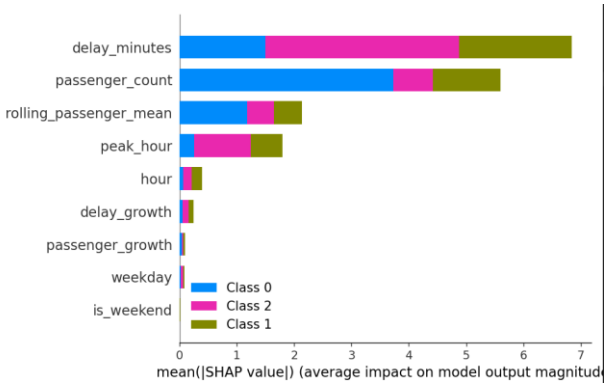


Fig. 3 Real-Time Overcrowding Prediction and Deviation Status Dashboard

### 4.3 Real-Time Deviation Monitoring Module

The system incorporates a simulated real-time monitoring mechanism that periodically re-evaluates crowd predictions. After a passenger checks crowd status and begins traveling toward the bus stop, the system continues polling updated predictions.

In the below Fig. 3 represents the deviation monitoring output:

a. Predicted Crowd: MEDIUM
b. Observed Crowd: MEDIUM
c. Deviation Detected: No

This module ensures that if crowd conditions unexpectedly increase (e.g., due to events or delays), the user can be notified in near real-time. The implementation demonstrates the feasibility of extending the model into a live monitoring system using polling mechanisms.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **03** March-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

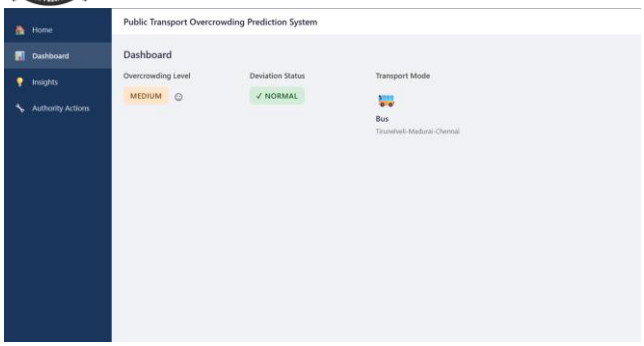


Fig. 4 Explainable AI Feature Importance Analysis Using SHAP

### 4.4Explanation & Insights Module

The system generates human-readable explanations based on SHAP feature contributions. Instead of displaying raw numerical outputs, the system translates influential features into understandable reasoning.

In the below Fig. 4 represents the AI Explanation output:

a. Service delays increasing crowd
b. Consistently high occupancy trend
c. Peak hour travel time

These explanations align with domain knowledge in public transport systems, where delays and peak-hour demand contribute significantly to overcrowding. This improves transparency and user trust in AI-driven predictions.

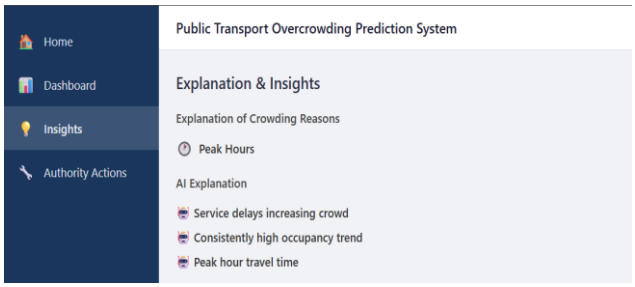


Fig. 5 AI-Based Crowd Reasoning and Explanation Module

### 4.5Ensemble Model Performance Analysis

The system was trained using a six-month synthetic temporal dataset consisting of over 260,000 records. After preprocessing and feature engineering, both Random Forest and XGBoost models were trained and evaluated.

The experimental results showed:

a. Random Forest Accuracy: ~86%
b. XGBoost Accuracy: ~86%
c. Balanced precision and recall across LOW, MEDIUM, and HIGH classes
d. HIGH crowd level classification achieved near-perfect detection

The ensemble approach improves robustness by combining probabilistic outputs from multiple models. The balanced classification metrics indicate that the system performs reliably across all crowd categories.

### 4.6Temporal Dataset Realism and Validation

The dataset generation process incorporated:

a. Seasonal effects (summer and festival months)
b. Weekend travel patterns
c. Peak hour surge effects
d. Event-based crowd spikes
e. Stop-level demand variation
f. Bus capacity constraints (based on Tamil Nadu public bus limits)
g. Delay propagation influenced by congestion

This ensures that the synthetic dataset closely mimics realistic public transportation dynamics, improving the credibility of model predictions.

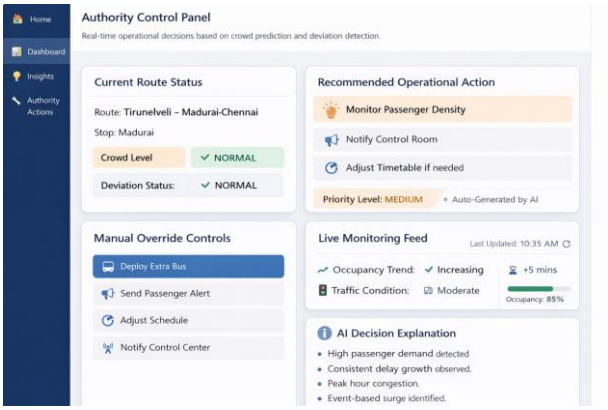


Fig. 6 Authority Monitoring and Operational Decision Support Interface

## 5. CONCLUSION

The project successfully implements an intelligent Public Transport Overcrowding Prediction System that integrates data generation, preprocessing, machine learning models (Random Forest and XGBoost), SHAP-based explainable AI, and real-time monitoring through polling. The system predicts crowd levels (Low, Medium, High) based on temporal, passenger, and delay features while also detecting deviations using simulated live updates. A realistic six-month temporal dataset with seasonal, peak-hour, event, weather, and bus capacity constraints enhances model reliability and practical relevance. The Angular frontend provides user-friendly travel input, live

dashboard updates, AI explanations, and actionable recommendations for both passengers and authorities. By combining predictive analytics, explainability, and real-time monitoring, the project demonstrates a scalable, practical, and decision-support-oriented smart transport solution suitable for real-world public transportation management.

## REFERENCES

[1] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
[2] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
[3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
[4] Box, G., & Jenkins, G. M. (1976). Analysis: Forecasting and control. *San francisco*, *10*.
[5] Vlahogianni, E. I. (2015). Optimization of traffic forecasting: Intelligent surrogate modeling. *Transportation Research Part C: Emerging Technologies*, *55*, 14-23.
[6] Zheng, Y. (2013). Urban computing with big data. *Journal of the Chinese society of computer communication*, *9*(8), 8-18.
[7] Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2014). Traffic flow prediction with big data: A deep learning approach. *Ieee transactions on intelligent transportation systems*, *16*(2), 865-873.
[8] Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., & Wang, Y. (2017). Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *sensors*, *17*(4), 818.
[9] Zhang, J., Zheng, Y., & Qi, D. (2017, February). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).
[10] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*
[11] Zhang, N., Wang, Z., Chen, F., Song, J., Wang, J., & Li, Y. (2020). Low-carbon impact of urban rail transit based on passenger demand forecast in Baoji. *Energies*, *13*(4), 782.
[12] Sun, S., Zhang, C., & Yu, G. (2006). A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on intelligent transportation systems*, *7*(1), 124-132.
[13] Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, *14*(3), 1393-1402.
[14] Yang, X., Xue, Q., Ding, M., Wu, J., & Gao, Z. (2021). Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data. *International Journal of Production Economics*, *231*, 107920.
[15] Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.
[16] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., ... & Li, Z. (2018, April). Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
[17] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., ... & Li, Z. (2018, April). Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).
[18] Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*.
[19] Su, J., Byeon, W., Kossaifi, J., Huang, F., Kautz, J., & Anandkumar, A. (2020). Convolutional tensor-train LSTM for spatio-temporal learning. *Advances in Neural Information Processing Systems*, *33*, 13714-13726.
[20] Ma, X., Yu, H., Wang, Y., & Wang, Y. (2015). Large-scale transportation network congestion evolution prediction using deep learning theory. *PloS one*, *10*(3), e0119044.
[21] Nguyen, H., Kieu, L. M., Wen, T., & Cai, C. (2018). Deep learning methods in transportation domain: a review. *IET Intelligent Transport Systems*, *12*(9), 998-1004.
[22] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.
[23] Bishop, C. M. (2006). *Pattern recognition and machine learning by Christopher M. Bishop* (Vol. 350). Berlin, Germany:: Springer Science+ Business Media, LLC.
[24] Rasouli, A., & Tsotsos, J. K. (2019). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE transactions on intelligent transportation systems*, *21*(3), 900-918.
[25] Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2018). Big data analytics in intelligent transportation systems: A survey. *IEEE transactions on intelligent transportation systems*, *20*(1), 383-398.