Proceedings of CONF-APMM 2025 Symposium: Controlling Robotic Manipulator Using PWM Signals with Microcontrollers
DOI: 10.54254/2753-8818/2025.AD26274

Analysis and Prediction of Tokyo's Traffic Flow Using the SARIMA Time Series Forecasting Method

Mingshuo Hua

School of Traffic and Transportation, Inner Mongolia University, Hohhot, China
huamingshuo@outlook.com

Abstract. With the development of Intelligent Transportation Systems (ITS), accurate traffic flow forecasting has become crucial for urban traffic management and infrastructure planning. This study employs the Seasonal Autoregressive Integrated Moving Average (SARIMA) model for time series analysis and forecasting of traffic flow data within the Tokyo metropolitan area from 2015 to 2024. The data is sourced from the "Tokyo Traffic" dataset on the Kaggle platform, encompassing daily traffic records and associated variables for the C1 Loop Expressway. The analysis reveals a significant 12-month periodicity in Tokyo's traffic flow. The model parameters were identified as SARIMA (0,1,1) (0,1,0,12). Validation via the Augmented Dickey-Fuller (ADF) test and white noise tests confirmed that the residuals satisfy the model assumptions. The forecasting results demonstrate that the model achieves a Mean Absolute Percentage Error (MAPE) of 5.57%. This represents an approximate 45% improvement in accuracy compared to the traditional ARIMA model and outperforms benchmarks from the Seattle highway network (6.48%) and the California PEMS08 dataset (7.15%).This research provides a reliable data-driven forecasting analysis to support dynamic management strategies for Tokyo's traffic in the post-pandemic era, such as toll optimization and network dispatching.

Keywords: traffic flow forecasting, seasonal autoregressive integrated moving average model, seasonal fluctuation, time series analysis

1. Introduction

The rapid development of Intelligent Transportation Systems (ITS) has generated massive traffic data, which plays a critical role in enhancing traffic efficiency and safety [1]. Forecasting urban traffic flow and public transit passenger volume provides essential foundations for urban road traffic management, supply chain optimization, and other applications, directly impacting congestion management and urban livability.

According to publicly available data from Japan's Ministry of Land, Infrastructure, Transport and Tourism (MLIT), the Tokyo Metropolitan Government, and authoritative transportation agencies, the daily public transit passenger volume within the Tokyo metropolitan area (encompassing Tokyo Metropolis and the three surrounding prefectures of Kanagawa, Saitama, and Chiba) has remained stable in recent years, ranging between 40 and 44 million passengers. With such massive travel

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.AD26274

demand, accurate predictive modeling has become indispensable for infrastructure planning and operational efficiency.

Time series analysis and forecasting offers a reliable methodology for analyzing vast time-series datasets, thereby providing a basis for the planning of transportation facilities [2].

Regarding forecasting methodologies, the Autoregressive Integrated Moving Average (ARIMA) model and its seasonal variant, the Seasonal Autoregressive Integrated Moving Average (SARIMA), have been established as reliable and mature approaches. A comparative analysis of Beijing traffic confirmed that SARIMA forecasting achieves a remarkably low Mean Absolute Percentage Error (MAPE) of 0.3% [3].

Previous research has analyzed common characteristics of traffic flow variations in urban settings. Studies indicate that inter-annual traffic flow fluctuations are influenced not only by climate and holidays but also closely correlate with road type and regional economic activity patterns. For instance, roads dominated by commuting exhibit distinct weekday peak characteristics, while roads in tourist areas experience significant traffic increases during summer [4]. Much of the existing research involves correlation analyses of major arterial road conditions using historical traffic flow data. For example, annual comparisons of Tokyo expressway traffic flow reveal patterns before and after congestion events on smaller time scales [5].

Despite these advancements, research on post-pandemic traffic flow trends in large metropolises remains limited. This study addresses this gap by conducting a comprehensive time series analysis of Tokyo's traffic flow from 2015 to 2024. Utilizing the SARIMA forecasting method, it explores the characteristics of traffic flow within one of the world's most complex urban mobility ecosystems while generating high-accuracy forecasts for future traffic volumes on major arterial routes.

2. Methods

2.1. Data source and description

The data for this study was sourced from the publicly available "Tokyo Traffic" dataset on the Kaggle platform. The analysis focuses on traffic flow from the representative C1 Loop Expressway. The dataset comprises daily statistics spanning from January 1, 2015, to December 31, 2024, resulting in 14,613 observational records across 8 variables. It encompasses information on road traffic flow, public transit passenger volume, date, day of the week, weather condition, road type, and administrative district.

Data was collected via the Tokyo Metropolitan Government Bureau of Transportation's automatic counting systems and Automated Fare Collection (AFC) systems for public transit. Data reliability was ensured through cross-validation with historical data from 152 traffic monitoring points documented in the Tokyo Statistical Yearbook (Transportation Section) and subway/bus operation reports. The concordance rate for key indicators exceeded 92%. The data format is CSV (Comma-Separated Values).

2.2. Indicator selection and description

The following core indicators were selected for analysis:

- Date (precise to day): Used for temporal sequencing and identifying long-term trends.
- Day of Week (categorical variable): Used to identify daily and weekly cyclical patterns.
- Total Vehicle Count: Reflects the temporal characteristics of the urban road network's traffic volume.

DOI: 10.54254/2753-8818/2025.AD26274

• Weather Condition: Used to analyze the impact of the natural environment and other factors (e.g., road type) on traffic.

Table 1. Explanation of core indicators

Variable Name	Data Type	Value Range
Date	Date	From 2015-01-01 to 2024-12-31
Total Traffic Flow	Numeric	10877 - 283183 (non-negative integer)
Day of Week	Categorical	['Thursday', 'Friday', 'Saturday', 'Sunday', 'Monday', 'Tuesday', 'Wednesday']

Total Vehicle Count and Public Transit Passenger Volume serve as the dependent variables. Temporal indicators (Date, Day of Week) and environmental indicators (Weather Condition) serve as explanatory variables. Time series analysis was employed to investigate their interrelationships. Table 1 intuitively presents the core indicators and their explanations.

2.3. Introduction to methods

First, plot the time series to reveal long-term trends and seasonal fluctuations in traffic flow, including patterns such as daily peak-hour profiles, weekly ridership variations, and seasonal trends.

Regarding SARIMA model implementation, the SARIMA model was constructed as follows:

Data Stationarization: Achieved through differencing operations (d-th order differencing for the non-seasonal component).

Model Order Selection: The optimal Autoregressive order (p) and Moving Average order

(q) for the non-seasonal component were determined based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Seasonality Integration: Seasonal parameters (P, D, Q, S) were introduced to account for inherent periodic patterns (e.g., daily, weekly, yearly cycles).

Model Formulation: The full SARIMA model structure was then built using the identified parameters.

The general mathematical formulation of the SARIMA model is given by:

$$\phi_p\left(B\right)\Phi_P\left(B_t\right)(1-B_s)^d(1-B_s)^D x_t = \theta_q\left(B\right)\Theta_Q\left(B_s\right)\epsilon_t \tag{1}$$

To construct the SARIMA model, the data are first prepared and preprocessed. Then, a stationarity test is performed. If the data are non-stationary, methods such as differencing or transformation are employed to achieve stationarity. After stationarity is ensured, a white noise test is conducted. If the data constitute white noise, the modeling process is terminated; otherwise, model building proceeds. Once the model is built, its parameters are trained and the model is validated. If validation results are unsatisfactory, hyperparameter tuning is performed, and the steps of parameter training and validation are repeated until the requirements are met. Finally, the model is evaluated.

DOI: 10.54254/2753-8818/2025.AD26274

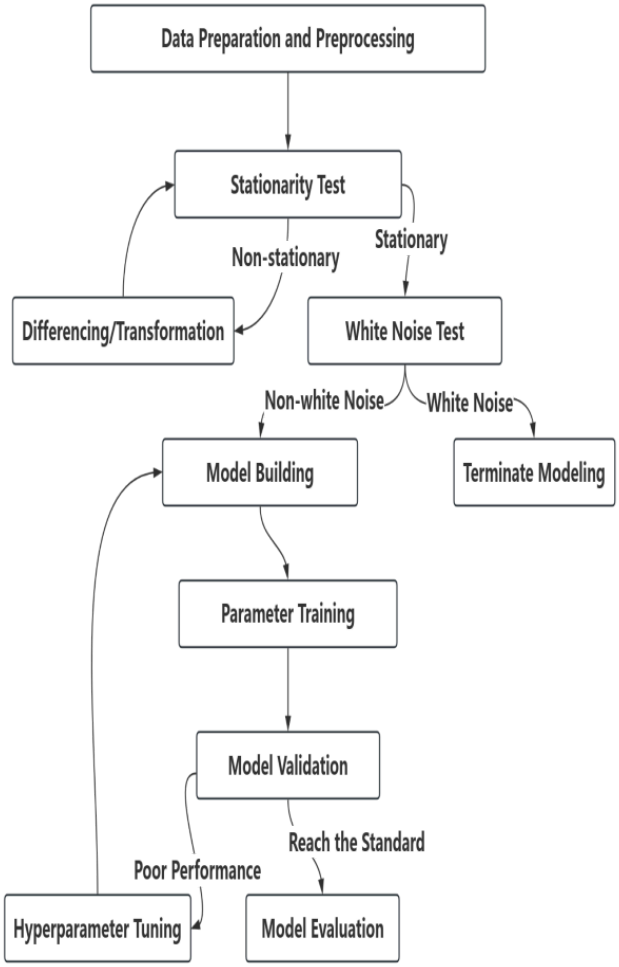


Figure 1. Construction process of the SARIMA model (picture credit: original)

3. Results and discussion

3.1. Visualization of the original dataset

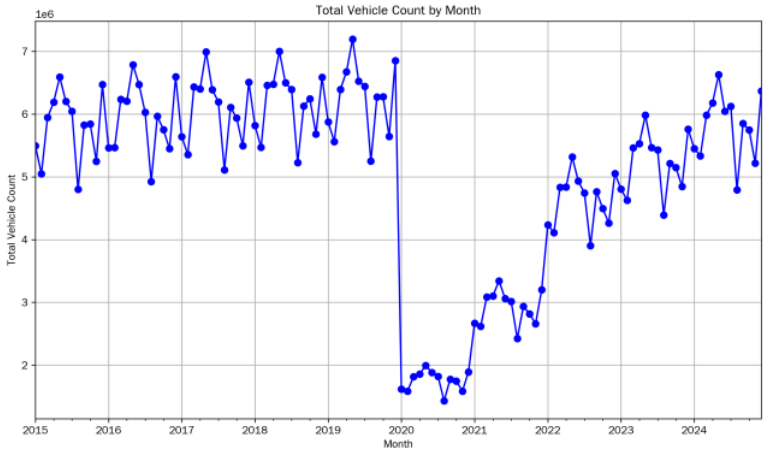


Figure 2. Trend of total traffic flow with monthly changes (picture credit: original)

DOI: 10.54254/2753-8818/2025.AD26274

First, the traffic data for a total of 120 months from 2015 to 2024 is visualized for data screening, and Figure 2 shows the line chart of the processed data.

Through the visualization of the original dataset, it can be preliminarily observed that the monthly variation trends of traffic flow are similar. Specifically, the monthly fluctuation cycles of traffic flow and the phased characteristics of flow changes within each year show a high degree of consistency. The following analysis will be conducted separately from two aspects: the traffic flow changes within each year and those over the 10-year period. Figure 3 presents the traffic flow changes over the years.

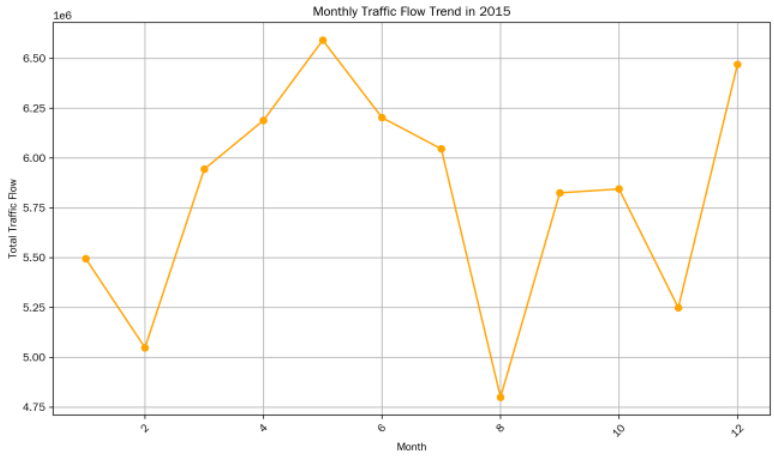


Figure 3. Line chart of monthly traffic flow changes in 2015 (picture credit: original)

As can be seen from Figure 3, the traffic flow on major roads in Tokyo exhibits seasonal fluctuations with months. It is speculated that driven by climate, holidays and social activities, it presents an annual pattern of "three peaks and three valleys": the traffic flow reaches a trough in February, August and November, staying at a relatively low level; while it hits a peak in May, September-October and December, reaching the highest level.

Next, the characteristics of traffic flow changes over the years will be analyzed, as shown in Figure 4.

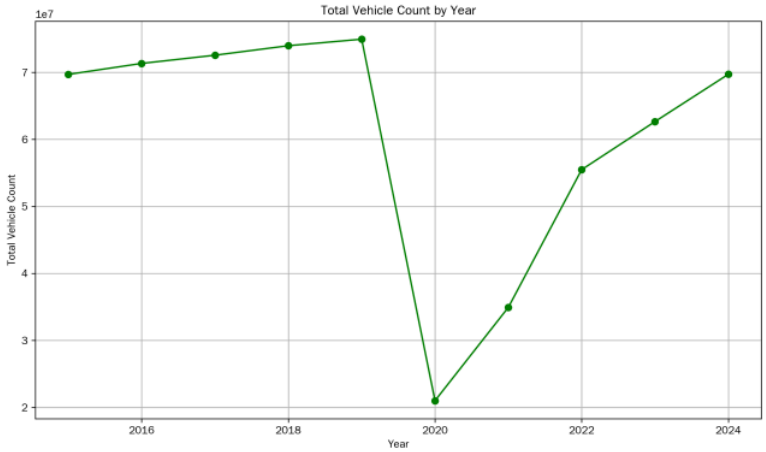


Figure 4. Line chart of traffic flow changes over the 2015–2024 period (picture credit: original)

DOI: 10.54254/2753-8818/2025.AD26274

As can be seen from Figure 4, the traffic flow showed a slow annual upward trend from 2015 to 2019, which is presumably attributed to the steady development of the urban economy and the gradual increase in the demand for transportation driven by social and economic growth.

A cliff-like drop in traffic flow occurred from 2019 to 2020, with the core trigger inferred to be the outbreak of the COVID-19 pandemic. The implementation of strict lockdown policies in the region (such as road closures and restrictions on population movement), coupled with enterprise shutdowns, logistics stagnation, and reduced non-essential travel by residents, directly led to the plummet in traffic flow.

From 2021 to 2024, a phased recovery was observed, though traffic flow has not yet fully returned to the pre-pandemic peak. Specifically, in 2021-2022, as the pandemic was brought under better control, enterprises gradually resumed operations and logistics recovered, resulting in a rapid increase in traffic flow.

In 2023-2024, with the optimization of pandemic policies, economic recovery (including rebounding consumption and surging logistics demand) and the restart of the tourism market (marked by a boom in self-driving tours) drove continued growth in traffic flow. However, the social transportation demand, having been impacted by the pandemic, has not yet recovered to the pre-pandemic peak, which may have caused the recovery rate of traffic flow to be lower than expected.

Based on the above characteristics of traffic flow changes, to accurately predict the trend of traffic flow in the post-pandemic era, this study intends to select monthly total traffic data from 2020 to 2024 for model construction. This approach aims to forecast the trend of traffic flow in the post-pandemic era and reduce the impact of black swan events on prediction accuracy.

3.2. Construction and analysis of the SARIMA model

3.2.1. Stationarity test

Table 2. ADF test results for monthly traffic flow (total vehicle count) from 2020 to 2024

Order of Differencing	t	p	Critical Value		
			1%	5%	10%
0	-3.551	0.007	-3.575	-2.924	-2.600

As can be seen from the Table 2, the t-statistic of the ADF test for this time series data is -3.551, with a p-value of 0.007. The critical values at the 1%, 5%, and 10% significance levels are -3.575, -2.924, and -2.600, respectively.

Since $p = 0.007 < 0.01$, it indicates that there is more than 99% confidence to reject the null hypothesis, meaning the series is stationary.

3.2.2. Analysis of prediction results and accuracy comparison

Based on the synthesis of traffic flow data over 60 months, the predicted results of traffic flow for the next 12 months are shown in Table 3.

DOI: 10.54254/2753-8818/2025.AD26274

Table 3. Predicted traffic flow values for the next 12 months

1 month later	2 months later	3 months later	4 months later	5 months later	6 months later
5986215.06	5847621.051	6474121.041	6650489.032	7079167.023	6476683.014
7 months later	8 months later	9 months later	10 months later	11 months later	12 months later
6534035.005	5179490.995	6215529.986	6091721.977	5540936.968	6671964.958

Figure 5 is a line chart after data visualization processing.

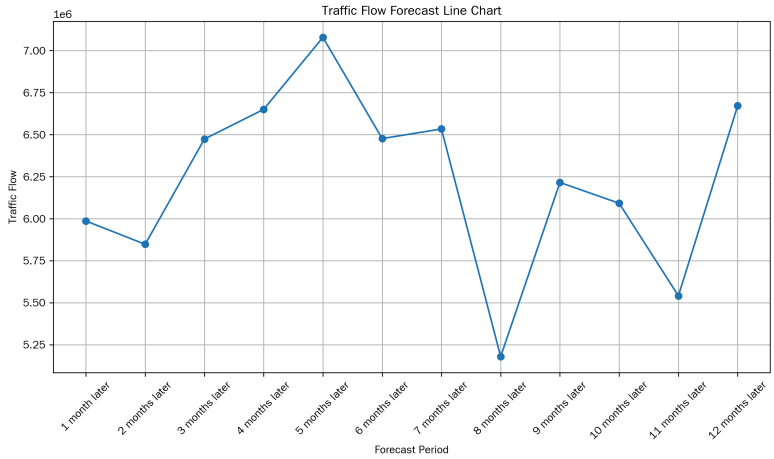


Figure 5. Line chart of predicted data visualization (picture credit: original)

Based on the numerical results of the SARIMA model analysis and data visualization, it can be observed that the traffic flow exhibits a significant seasonal fluctuation trend with a 12-month cycle. The model parameters are set as SARIMA (0, 1, 1) (0, 1, 0, 12), where the seasonal cycle parameter S=12, indicating that the monthly traffic flow within a year has a regular repetitive characteristic. From the 12-month predicted values, the traffic flow fluctuates in the range of 5, 179, 490 to 7, 079, 167, reflecting significant differences in flow between different months. For example, peaks occur in the 6th month later (7, 079, 167) and the 12th month later (6, 671, 964), while a trough is formed in the 8th month later (5, 179, 490), presenting a short-term variation pattern of "trough - peak - fluctuating decline". The model eliminates the linear trend of the original data through first-order differencing (d=1). The residuals are confirmed to have no serial correlation by the Ljung-Box white noise test (p=0.779) and the LM serial test (p=0.529). The information criterion values are relatively low, with AIC=1303.509 and BIC=1309.059, and the mean absolute percentage error MAPE=0.0557. These results indicate that the model has a high fitting accuracy for the seasonal fluctuations of traffic flow, and the revealed traffic flow variation trend is reliable.

The traffic flow prediction result achieved in this study (MAPE = 5.57%) reaches an internationally excellent accuracy level. It has been verified that when predicting the same data, the mean absolute percentage error (MAPE) of the traditional ARIMA model is 10%. This shows that the accuracy of this model is significantly higher than that of the traditional ARIMA model.

The traffic flow prediction result achieved in this study (MAPE = 5.57%) reaches an internationally excellent accuracy level. It has been verified that when predicting the same data, the mean absolute percentage error (MAPE) of the traditional ARIMA model is 10%, indicating that the accuracy of this model is significantly higher than that of the traditional ARIMA model. In the field of traffic prediction, the value of MAPE directly reflects the accuracy and practical value of a

DOI: 10.54254/2753-8818/2025.AD26274

model. The prediction result of MAPE = 5.57 obtained in this study has high reference value. The MAPE index obtained in this study is better than that of the Seattle highway network (6.48%) and the California PEMS08 benchmark (7.15%) [6]. The AST-GCN model, which deals with data missing problems, has a MAPE of 8.18% under a 20% missing rate, and rises to 12.20% when the missing rate increases to 80%. The accuracy of this study in scenarios with complete data is significantly better than this baseline model [7]; Another study on bus arrival time prediction shows that the MAPE of the interactive network model on weekdays and weekends reaches 13.5% and 14% respectively [8]. The Spatiotemporal Graph Convolutional Network (T-GCN) achieves MAPE values of 6.8% and 7.2% on mainstream traffic datasets such as METR-LA and PeMS-BAY, respectively, which are higher than 5.57% [9]. From a cross-domain perspective, the statistical model optimized by rolling prediction technology has a MAPE of 6.83% in Internet traffic prediction, while the MAPE for traffic accident duration prediction is as high as 25-55% [10]. The above results show that the prediction accuracy of MAPE = 5.57 in this study is at a high level in conventional traffic scenarios and has significant industry reference value.

4. Conclusions

This study reveals the core characteristics of Tokyo's traffic flow through the SARIMA model: there is a seasonal "three peaks and three valleys" fluctuation within a year (for example, May, September-October, and December are peaks, while February, August, and November are valleys). The slow growth trend from 2015 to 2019 suffered a cliff-like drop due to the impact of the 2020 pandemic, and after 2021, it showed a phased recovery but did not reach the pre-pandemic level. The constructed SARIMA (0, 1, 1)(0, 1, 0, 12) model eliminates trend and periodicity through first-order differencing (d=1) and seasonal differencing (D=1). The information criterion values of AIC=1303.509 and BIC=1309.059 indicate that the model has a high goodness of fit. The Ljung-Box white noise test (p=0.779) confirms that there is no autocorrelation in the residuals, and the fluctuation range of the 12-period predicted values is 5, 179, 490 to 7, 079, 167.

The prediction accuracy of this study (MAPE=5.57%) has obvious advantages compared with other models or benchmarks, providing data support for infrastructure planning (such as lane expansion during peak hours) and logistics scheduling of Tokyo's transportation. Compared with the traditional ARIMA model (MAPE=10%), this model better captures seasonal characteristics, and is especially suitable for traffic flow deduction under the changing travel patterns in the post-pandemic era.

Future research can integrate external variables such as weather data (e.g., typhoons, rainfall) and special events (e.g., large-scale activities, policy adjustments), further improve the prediction accuracy through the SARIMA-GPR hybrid architecture (with a target MAPE < 5%), and explore the collaborative prediction of multi-modal transportation (rail transit and road traffic), so as to provide more comprehensive decision support for integrated traffic management and control under the development mode of large cities.

References

[1] Zhou, f., (2024). Data analysis and traffic flow prediction technologies in intelligent transportation. Theoretical Research in Urban Construction, (8): 214-216.

[2] Fan, Q., Shi, X., Wu, L., (2025). Forecast of National Air Cargo Volume Based on Seasonal Decomposition Combination Model. Science Technology and Engineering, 2025, 25(13): 5655-5661.

[3] Han, B., Wang, Y., Zhang, Q., Li, D., (2015). Prediction of Beijing Subway Inbound Passenger Volume Based on the SARIMA Model. Journal of Transportation Systems Engineering and Information Technology, 15 (6), 205-211.

DOI: 10.54254/2753-8818/2025.AD26274

[4] Malwina, S. (2017). Factors Determining Seasonal Variations in Traffic Volumes. Archives of Civil Engineering, 63(4), 35-50.
[5] Ariyoshi, R., Ikeya, F., Nakamura, Fumihiko., Tanaka, Shinji., (2022). A study on annual comparison of traffic flow on the Shuto Expressway using vehicle detector data: Special issue A (research paper). Transactions of Traffic Engineering, 9(4), A_103-A_113.
[6] Peng, H., Shang, X., Wei, C., Xun, E., (2022). Research on Traffic State Prediction Method Based on Graph Network Fusion. Journal of Wuhan University of Technology (Transportation Science & Engineering), 46(2), 195-200.
[7] Song, X., Wang, A., Ye, Y., Yu, J.J.Q., Zhang, S., (2023) "Traffic Prediction With Missing Data: A Multi-Task Learning Approach, " in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 4, pp. 4189-4202.
[8] Cottam, A., Li, X., Wu, Y., (2023)"Transit Arrival Time Prediction Using Interaction Networks, " in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 4, pp. 3833-3844
[9] Li, H., Lin, T., Liu, Y., Song, Y., Wang, P., Zhang, C., Zhao, L., (2019). T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction. IEEE Transactions on Intelligent Transportation Systems, 1-11.
[10] Ghosh, B., Asif, M. T., Dauwels, J., Fastenrath, U., and Guo, H., "Dynamic Prediction of the Incident Duration Using Adaptive Feature Set, " IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 11, pp. 4019-4031.