Short-Term Passenger Flow Prediction of Suzhou Metro Using Random Forest and LSTM

Kai Zhang[1,*] and Lixun Zhuang[2]
[1]Computer Science, University of Birmingham, United Kingdom
[2]Guangdong Experimental High School International Department (AP), Guangzhou, Guangdong, China
*kxz419@student.bham.ac.uk

Abstract. In the operation of urban public transportation, the volatility of subway passenger flow stands as a key issue affecting operational efficiency and passenger experience. Congestion during peak passenger flow periods, wasted transport capacity during off-peak hours, and sudden changes in passenger flow under special scenarios such as holidays not only increase the difficulty of operational dispatching, but also easily lead to extended passenger waiting times and reduced travel comfort. To address the fluctuation of metro passenger flow and enhance operational efficiency and passenger experience, this paper develops a short-term prediction model for Suzhou Metro passenger flow in 2025 by integrating Random Forest with Long Short-Term Memory (LSTM). Random Forest is employed to select key features, including historical passenger volume, holidays, weather conditions, line attributes, and social events, which significantly influence flow variations. The LSTM component then effectively captures complex temporal dependencies and nonlinear patterns in the data, improving prediction accuracy. Experimental results demonstrate that the proposed hybrid model exhibits strong generalization capability and reliable prediction performance under various conditions, such as peak hours, holidays, and emergencies. This approach provides effective data-driven support for dynamic metro scheduling, resource allocation, and operational management, contributing to reduced congestion and improved service quality for passengers.

Keywords: Suzhou; Metro; Random Forest; LSTM

1 Introduction

With rapid urbanization and the growing demand for public transportation, metro systems have become an essential component of modern cities. Suzhou, as one of the fastest-growing cities in China, has seen the rapid expansion of its metro network. By the end of 2022, the system included five lines, 168 stations, and a total operating length of 210 km [1].

The uneven spatiotemporal distribution of passenger flow often causes congestion during peak hours, insufficient transport capacity, and reduced operational efficiency. If passenger flow predictions are significantly overestimated, it can lead to substantial waste of human, material, and financial resources; if predictions are considerably underestimated, it may result in overcrowded stations and carriages, which in minor cases cause passenger dissatisfaction and in severe cases lead to safety incidents such

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_15

as injuries [2]. Accurate forecasting of passenger flow patterns is crucial for improving traffic management, capacity allocation, and service quality [3]. Analyzing metro passenger flow characteristics and volume, Youdaoplaceholder0 historical passenger load data from existing lines to predict temporal variations in passenger flow at individual stations, can assist metro operators in implementing appropriate measures. This enables passengers to make more informed travel plans and supports operators in scheduling and planning, which is of significant importance for the overall efficiency of metro operations and passenger safety [4].

In the field of intelligent transportation, machine learning offers an effective means of analyzing traffic flow data and predicting future traffic conditions. Two commonly employed methods are the Random Forest and Long Short-Term Memory (LSTM) algorithms. Random Forest is applicable to both classification and regression tasks. It operates by aggregating multiple decision trees to form a "forest," with prediction outcomes determined based on the input features [5]. The Random Forest model, which is widely adopted in machine learning, has been utilized to forecast passenger flow at transportation hubs, yielding relatively high overall prediction accuracy [6].

The Long Short-Term Memory (LSTM) network, a recurrent neural network designed for sequence prediction, is particularly effective in modeling long-term dependencies [7]. However, LSTM processes data in a single temporal direction and struggles with capturing dependencies across distant time steps [8]. In short-term prediction tasks, where data are segmented into minutes or hours, the influence of factors such as holidays and weather diminishes, making the extraction of short-term features increasingly important [8]. Furthermore, this study establishes LSTM-based short-term passenger flow prediction models for different categories of metro stations. By comparing the prediction results across station categories, this paper analyze and evaluate the differences in prediction accuracy when using LSTM models, thereby providing a refined understanding of their applicability in various operational contexts [9].

This study develops a hybrid Random Forest–LSTM model for short-term passenger flow prediction of Suzhou Metro, using 2024–2025 operational data. Random Forest is used to select effective features, while LSTM models temporal dependencies, providing reliable prediction results for metro operations, scheduling, and passenger flow management.

2 Dataset and Methods

2.1 Dataset

The dataset was obtained from Suzhou Rail Transit Company and covers the period from July 14 to July 22, 2025, including daily inflow and outflow data from 248 stations across nine metro lines. Additional factors such as holiday indicators, calendar features, and line differences were integrated into the model to enhance predictive accuracy. All data were anonymized and standardized before model training to ensure consistency and privacy. Figure 1 shows the daily passenger flow of Suzhou Metro/10,000.

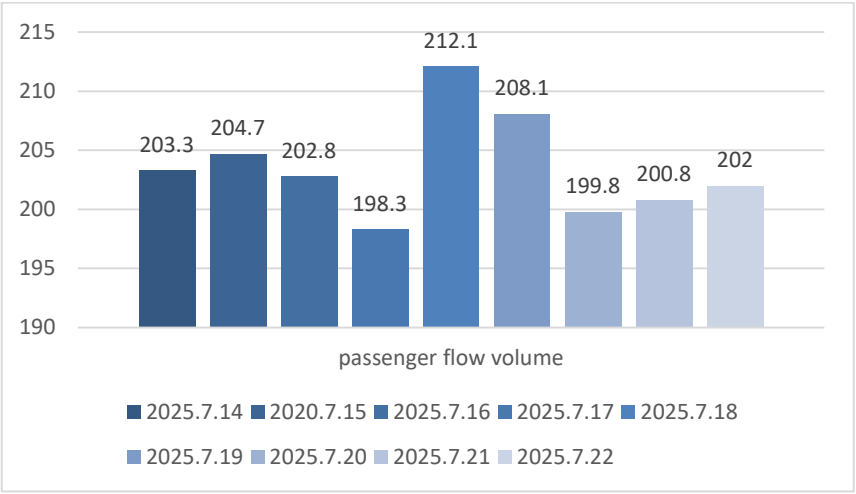

Fig. 1. Daily passenger flow of Suzhou Metro/10,000 (Data from: Suzhou Rail Transit Company)

Data authenticity was guaranteed by sourcing passenger flow records from official government platforms and Suzhou Metro operators. Timeliness was ensured by selecting representative periods from both 2024 and 2025, including May Day holiday data. Predictive precision was improved through the incorporation of auxiliary features such as holiday markers, weekday/weekend distinctions, and line identifiers. Figure 2 shows the average daily passenger flow of the subway during the May Day holiday in 2024/10,000.

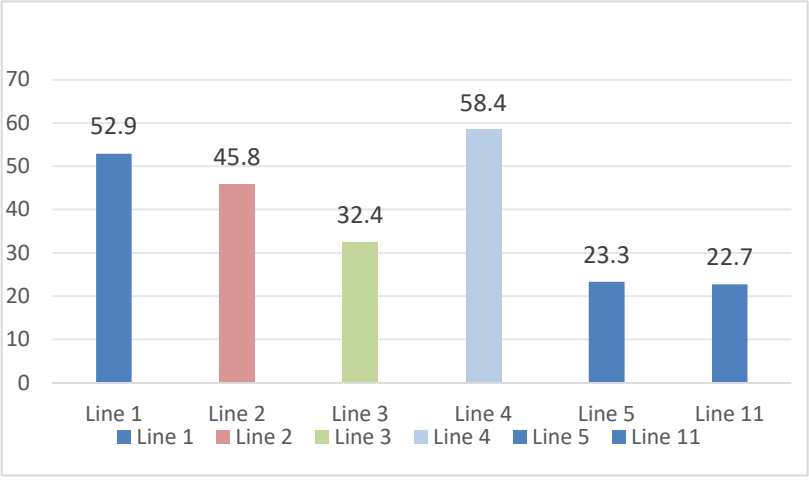

Fig. 2. Average daily passenger flow of subway during May Day holiday in 2024/10,000 (Data from: Suzhou municipal government website and subway operator)

The prediction targets included short-term passenger inflow, outflow, and total flow at specific stations and time intervals. Considering the holiday effect and station heterogeneity, the study emphasized feature selection for more accurate short-term forecasting.

To address the limitations of single-model approaches in both feature selection and temporal dependency learning, this paper proposes a hybrid framework that integrates Random Forest for identifying key predictors and LSTM for capturing sequential patterns in time-series data. Random Forest identifies important non-sequential attributes, while LSTM captures sequential dynamics to improve prediction accuracy.

LSTM can deeply learn and train both long-term and short-term variations in input signals, and efficiently handle complex dynamic dependencies in long- and short-term time-series data. Although urban rail passenger flow fluctuates significantly in the short term, it still depends on both long-term trends and recent passenger flow levels. Therefore, the LSTM model is well-suited for accurately predicting short-term passenger flow in urban rail systems [10]. The LSTM structure (Figure 3) follows standard formulations: the forget gate controls which past information is discarded, the input gate regulates the addition of new information, and the output gate determines the hidden state passed forward. Together, these mechanisms allow the model to retain relevant historical features while adapting to new inputs.

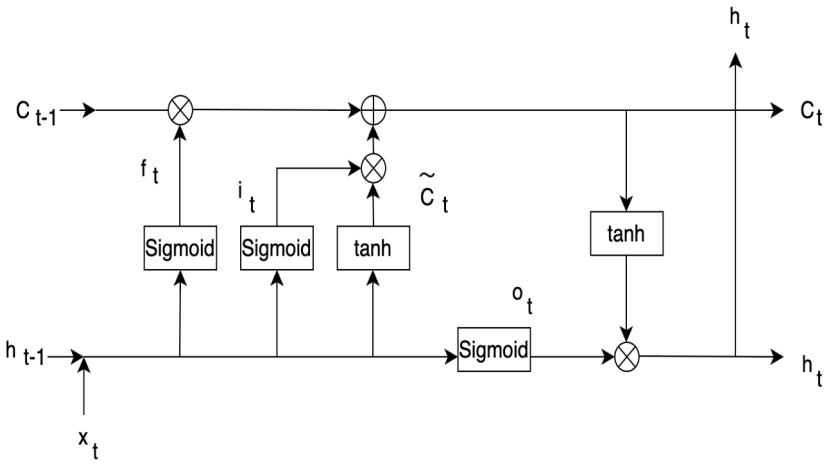

Fig. 3. LSTM assumption diagram [10]

The formula of the forgetting gate in LSTM is:

$$f_t = \sigma\left(W_f\,[\mathrm{h}_{t\text{-}1}, x_t] + b_f\right) \tag{1}$$

The formula of the LSTM input gate is:

$$i_t = \sigma\,(W_i\,[\mathrm{h}_{t\text{-}1}], x_t + b_i \tag{2}$$

$$\widetilde{C}_t = \tanh\ W_c\ [h_{t-1}, X_t] + b_c \tag{3}$$

The formula of the LSTM output gate is:

$$o_{t=}\ \sigma\ (W_O\ [h_{t-1}] + b_o \tag{4}$$

$$h_t = O_t \tanh(C_t) \tag{5}$$

The LSTM follows fixed rules to store, update, and extract information. The forget gate, activated by a Sigmoid function, determines which past information is discarded. The input gate controls how much new information is written, while the candidate cell state represents possible new values. Together, they update the memory cell. The Sigmoid and hyperbolic tangent functions regulate this process. The output gate selects relevant information from the memory and determines the hidden state at the current step, which serves as the final output. The cell state acts as long-term memory and is updated accordingly.

3 Results and Discussion

During the model construction stage, the Random Forest algorithm was first applied to select multi-source features such as holidays and metro lines. By evaluating the impact of different factors on passenger flow, the most valuable subset of features was extracted. For example, the passenger flow on July 16, 2025, reached 2.121 million, significantly higher than 1.983 million on July 19, indicating that holidays and temporal factors have a notable influence on passenger flow fluctuations. Table 1 illustrates a comparison between regular holidays and holiday peaks in Suzhou.

Table 1. Comparison between regular holidays and holiday peaks in Suzhou.

	The daily passenger volume in Suzhou is measured in ten thousand person-times					
	Holiday peak			Regular holidays		
Date	December 31,2024	May2, 2024	June31, 2025	Daily average in 2024	July 10, 2025	July19, 2025
	347.6	336.1	264	178	151.6	158.6

Subsequently, the selected features were input into the LSTM model for time-series modeling. Through the gated mechanism, the LSTM effectively captured long-term dependencies and achieved an accurate prediction of future passenger flow. The model parameters were optimized using backpropagation, and cross-validation was applied to improve generalization. The model demonstrated stable performance and high accuracy across both holiday and weekday scenarios. Figure 4 presents a comparison between the predicted and actual passenger flow

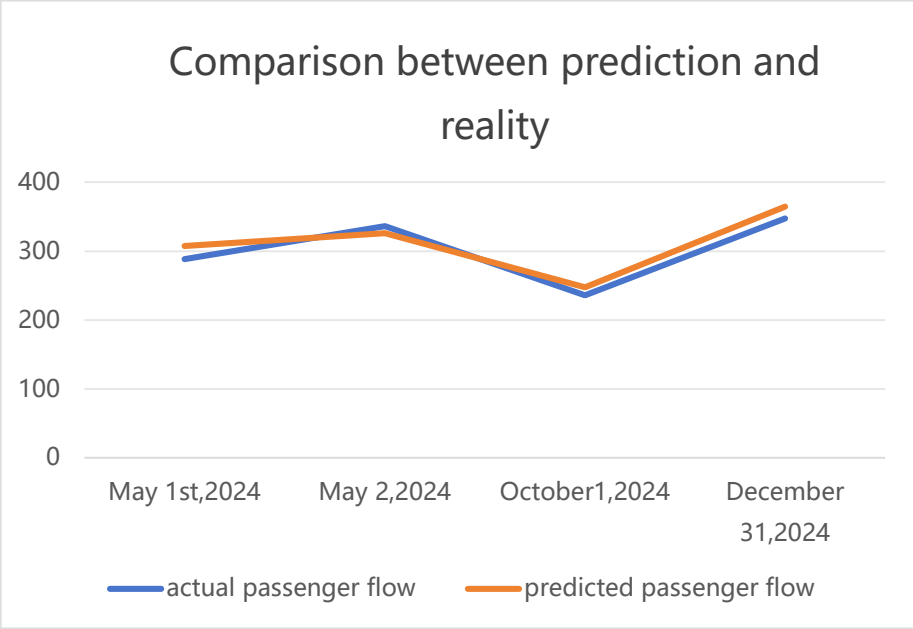

Fig. 4. Comparison between the predicted and actual passenger flow (Picture credit: Original)

The visualization comparing predicted and actual results shows that the model demonstrates strong trend consistency and is suitable for supporting operational scheduling decisions. For example, during the May Day holiday in 2024, Line 4 and its branch recorded an average daily passenger flow of 584,000, which was significantly higher than Line 5 (233,000) and Line 11 (227,000). This finding confirms the critical role of line attributes in passenger flow distribution.

4 Conclusion

This study proposed a short-term passenger flow prediction model for the Suzhou Metro by integrating Random Forest and LSTM. The model successfully identified key influencing factors, accurately captured time-series characteristics, and achieved strong adaptability and generalization across multiple scenarios. The findings provide reliable support for metro operators in passenger flow regulation, resource allocation, and scheduling optimization, contributing to the intelligent development of urban rail transit. However, several limitations remain. Sudden changes caused by extreme weather or large-scale events are difficult to capture. Variations in passenger flow across different time intervals also challenge prediction accuracy. Moreover, the data source was relatively limited, relying mainly on historical records while underutilizing information from surrounding commercial and educational activities. Future work may involve incorporating weather, event data, and commercial activity information, refining temporal granularity, and enhancing external data integration. Such

improvements would lead to more robust and realistic forecasting results, supporting smarter metro operations and better passenger experiences.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. Wang, S.: Research on Strategies to Improve Passenger Satisfaction in Suzhou Rail Transit Company. Master's Thesis, Southeast University (2023)
2. Li, Z., Ni, S., Sun, K., Lü, M.: Short-term passenger flow prediction of urban rail transit based on multi-feature fusion. Journal of Traffic and Transportation Engineering and Informatics 18(4), 93–102 (2020)
3. Liao, G.: Urban rail transit passenger flow prediction based on ARMA and regression correction combined method. Modern Information Technology 9(5), 139–143 (2025). https://doi.org/10.19850/j.cnki.2096-4706.2025.05.026
4. Li, S.: Research on subway passenger flow prediction based on improved LSTM model with attention mechanism. Industrial Control Computer 36(11), 124–128 (2023)
5. Fu, T.: Machine Learning-Based Passenger Flow Prediction and Emergency Evacuation Simulation for Urban Rail Transit Stations. Master's Thesis, Tianjin University of Technology and Education (2023)
6. Zhang, Y.: Research on Short-term Subway Passenger Flow Prediction Based on CEEMDAN-GA-LSTM Combined Model. Master's Thesis, Dongbei University of Finance and Economics (2024)
7. Du, X., Zhao, X., Li, L.: Short-term prediction of urban rail transit station inflow passenger based on LSTM. Journal of Guizhou University (Natural Science Edition) 38(5), 109–118 (2021)
8. Zhao, J., Liu, B., Tian, Z., Wu, W.: Short-term passenger flow prediction for urban rail transit based on CNN-Bi-LSTM network. Journal of Lanzhou Jiaotong University 43(6), 130–137 (2024)
9. Luo, S., Che, C., Zhang, Y.: Data-enhanced short-term passenger flow prediction for rail transit multi-station systems. Transportation Technology and Economics, 1–7 (2025).
10. Zeng, L., Li, Z., Yang, J., Xu, X.: Short-term passenger flow prediction method of urban rail transit based on CEEMDAN-IPSO-LSTM. Journal of Railway Science and Engineering 20(9), 3273–3286 (2023)