Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.25252

Research on Short-Term Passenger Flow Prediction in Jiuzhaigou Based on LSTM

Yu Pan

School of Software, Beijing Union University, Beijing, China
2022250360019@buu.edu.cn

Abstract: With the rapid growth of tourism, short-term passenger flow prediction has emerged as a crucial research area for enhancing scenic spot management and improving the tourist experience. Utilizing a long short-term memory (LSTM) network, this study analyzes daily passenger flow data from the Jiuzhaigou Scenic Area over recent years to address issues related to missing data and prediction accuracy. By employing data preprocessing, parameter tuning, and feature engineering, the study examines the effects of various model configurations on predictive performance. Ultimately, a high-accuracy LSTM model is developed. Experimental results demonstrate that the time series model based on LSTM effectively captures the dynamic patterns of Jiuzhaigou's short-term passenger flow, providing valuable data support for the scientific management of scenic spots and the development of emergency response plans.

Keywords: long short-term memory network, Jiuzhaigou passenger flow prediction, data preprocessing, neural network model

1. Introduction

With China's significant progress toward socialism, the tourism market has exhibited rapid growth. However, the surge in tourist numbers has led to overcrowding in certain scenic spots, diminishing the travel experience and affecting the reputation of these locations. This article focuses on employing the long short-term memory (LSTM) network method to address the challenge of short-term tourist flow prediction in scenic spots. The LSTM network demonstrates strong performance in handling small sample sizes, high-dimensional classification, nonlinearity, and local minima issues, making it well-suited for addressing short-term passenger flow prediction problems.

From a research perspective, applying the LSTM network to passenger flow prediction enhances the theoretical framework for managing scenic spot tourism. Practically, predicting scenic spot passenger flow enables the preparation of emergency management measures to prevent overcrowding and stampede incidents, assists tourists in organizing rational travel plans, and improves both the tourist experience and the efficiency of scenic spot management.

2. Literature review

A significant number of articles have discussed the use of long short-term memory (LSTM) networks for predicting short-term passenger flow. The main research advancements can be categorized into three stages: the verification stage, the improvement stage, and the combination model stage. During

© 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: 10.54254/2755-2721/2025.25252

the verification stage, Zhou Ying [1] employed logistic regression to model the relationship between the spatial correlation of scenic spots and influencing factors. Based on the spatial correlation results obtained from the model, an LSTM model was established. The LSTM model was trained using data from the Niu Shou Mountain Cultural Tourism Area and two other spatially correlated tourist attractions. This approach to predicting short-term passenger flow across different scenic spots validated the feasibility of the LSTM model for this purpose. In the improvement stage, numerous scholars proposed enhancement strategies to address the prediction accuracy loss issues of the LSTM model when forecasting short-term passenger flow. For instance, sparse principal component analysis [2] was incorporated into the LSTM model to examine the differences between the optimized LSTM model and other prediction models. Results indicated that the optimized LSTM model outperformed other models in predicting short-term passenger flow [2-9]. In the combination model stage, researchers applied hybrid prediction models to forecast short-term passenger flow, demonstrating that these combination models achieved higher prediction accuracy compared to single prediction models [10-12]. The key research ideas are summarized as follows:

1. Using the short-term passenger flow data from multiple tourist attractions as the research object, the relationship between the spatial correlation of scenic spots and influencing factors was modeled, and an LSTM model was established. The prediction performance was tested using data from different scenic spots [1], demonstrating that the LSTM model can effectively predict short-term passenger flow.

2. The LSTM model was trained using data from a single tourist attraction, and optimization strategies were proposed. For example, Liu Xinyue [3] developed an improved LSTM model for short-term passenger flow prediction in the Terracotta Warriors and Horses Scenic Area, namely the GBDT-LSTM model based on the gradient boosting decision tree and the PCA-LSTM model based on principal component analysis. Alternatively, other improved LSTM models were proposed to address cases where some prediction data showed low accuracy. For instance, Ni Tian [2] used convolutional neural networks to optimize the prediction model for holiday passenger flow data and established the SPCA-CNNLSTM prediction model. The prediction performance of this model was compared with LSTM and other prediction models, revealing that it delivers superior predictive accuracy [2–9, 13,15].

3. Various types of data impacting short-term passenger flow were analyzed to construct features capable of predicting scenic spot passenger flow. Using multiple methods, relevant data was collected to establish predictive models. For instance, Yang Ruoyu [8] designed a portable device for tourists and employed multi-source heterogeneous tourism data in conjunction with data obtained from the portable devices. A graph convolutional neural network (GCN) was then used to learn spatial correlation features between regions. The processed data was applied to develop an improved GCN-LSTM model. Repeated verification of the research data demonstrated that the optimized parameters significantly enhanced the model's predictive performance [4, 8, 13,14].

4. Perform technical processing on the obtained data from tourist scenic spots. For example, Deng Yufei [6] used the daily passenger flow data of Jiuzhaigou and Siguniang Mountain as research objects, extracting initial core keywords through technical word extraction and direct word extraction systems. The core keywords were further screened and processed, and the attention mechanism was applied to optimize the LSTM neural network. Ultimately, a combined model, ARIMA-ATT-LSTM, incorporating network search data, was established. Similarly, Xu Shuo [15] incorporated factors such as scenic spot temperature into the research, alongside static historical data. He applied the latest hybrid models and improved the network architecture by using the preprocessing model Tr to develop two hybrid network prediction models: Tr-BiLSTM-CNN and A-Tr-BiLSTM-CNN. It was concluded that data preprocessing significantly enhances the model's prediction accuracy.

DOI: 10.54254/2755-2721/2025.25252

Existing research by Chinese scholars demonstrates that the long short-term memory network (LSTM) has achieved significant advancements in short-term passenger flow prediction, showcasing strong application potential. From initial model verification to model improvements and the construction of combined models, researchers have progressively deepened their understanding and application of the LSTM model in predicting short-term passenger flow. This has provided a powerful tool for the management and planning of tourist attractions. Overall, the work of Chinese scholars has not only promoted the application of LSTM in this field but also laid a solid foundation for future research directions and technological innovation.

3. Methodology

This study primarily employs machine learning and data science methods. To address the issue of missing data in the daily passenger flow records of the Jiuzhaigou Scenic Area, the average values from the previous two years are used to fill the gaps. Subsequently, a long short-term memory (LSTM) network model is developed to predict the data.

This study primarily employs machine learning and data science methods. To address the issue of missing data in the daily passenger flow records of the Jiuzhaigou Scenic Area, the average values from the previous two years are used to fill the gaps. Subsequently, a long short-term memory (LSTM) network model is developed to predict the data.

$$z = tanh(W \cdot [x_t, h_{t-1}]) \tag{1}$$

$$z_i = \sigma(w_i \cdot [x_t, h_t - 1]) \tag{2}$$

$$z_f = \sigma(w_i \cdot [x_t, h_t - 1]) \tag{3}$$

$$z_o = \sigma(w_i \cdot [x_t, h_t - 1]) \tag{4}$$

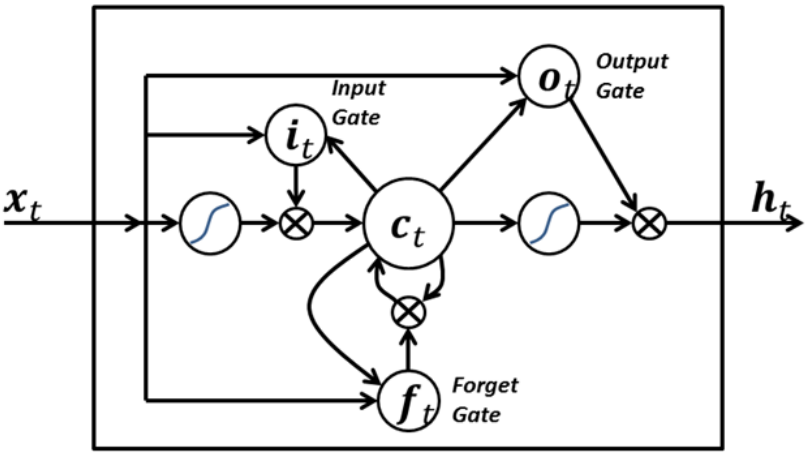

Figure 1: Classic LSTM structure [16]

4. Data and conclusions

This study utilizes daily visitor numbers disclosed on the official website of the Jiuzhaigou Scenic Area from May 25, 2012, to November 22, 2024, as the data source for tourist flow, comprising a total of 4,565 records. Due to the impact of earthquakes, heavy rainfall, and policy changes, some data points are missing, which resulted in significant data fitting discrepancies during LSTM model predictions. To address this issue, the missing data were filled by averaging data from the preceding two years, which successfully improved the prediction accuracy of the model.

DOI: 10.54254/2755-2721/2025.25252

The study employs a time-series prediction model based on the LSTM framework for training and prediction. For all datasets, the target prediction column is defined as the daily visitor count. The model's hidden layer size is set to 128, with 2 stacked layers, a dropout probability of 0.2, 95% of the total data used for training, and 15% of the training data reserved for validation. The batch size is 64, the learning rate is 0.01, and the total training epochs are 500. To validate the accuracy of the LSTM model in short-term tourist flow prediction, the dataset is divided into the most recent five years of data and the entire dataset. Additionally, for the missing data filled by averaging the previous two years' values, a label column is added to indicate these entries, forming a new dataset to test whether the label column strategy enhances prediction accuracy.

To determine the optimal training parameters, the patience for early stopping and the time step are set to different values. The model's mean squared error (MSE), loss value, and MSE for specific datasets are used to evaluate the performance of the LSTM model in short-term tourist flow prediction. The following results were obtained, where the blue line represents the validation dataset, and the red line represents the model prediction curve:

When the early stopping patience is set to 10 and the time step to 20, the LSTM model predicts the most recent five years of data with an MSE of 0.015478, a loss value of 0.021785, and an MSE of 0.38706189 for specific datasets.

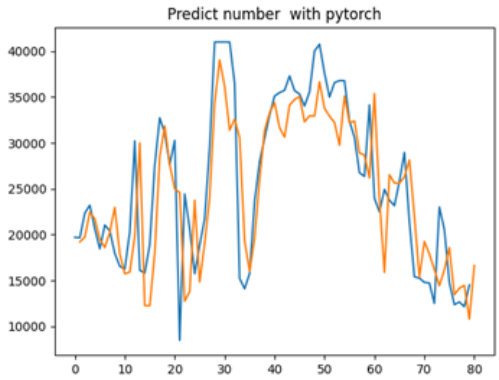

Figure 2: Prediction results for the most recent five years of data

Using the same training parameters, predictions for the entire dataset yield an MSE of 0.021165, a loss value of 0.030537, and an MSE of 0.36993607 for specific datasets.

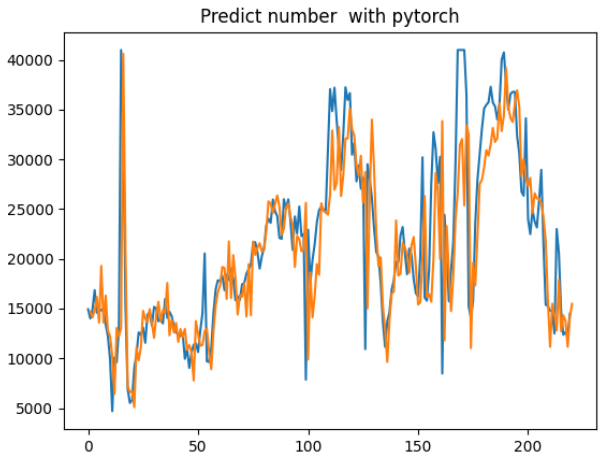

Figure 3: Prediction results for the entire dataset

DOI: 10.54254/2755-2721/2025.25252

Using the same training parameters for the dataset with the added label column, predictions for the most recent five years of data result in an MSE of 0.015071, a loss value of 0.020296, and an MSE of 0.39926783 for specific datasets.

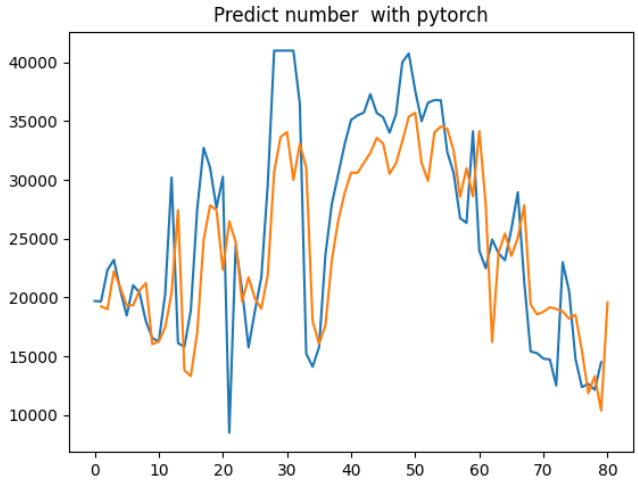

Figure 4: Prediction results for the dataset with label column added

Using the same training parameters, predictions for the entire dataset with the added label column resulted in a mean squared error (MSE) of 0.024966, a loss value of 0.030110, and an MSE of 0.33375474 for specific datasets.

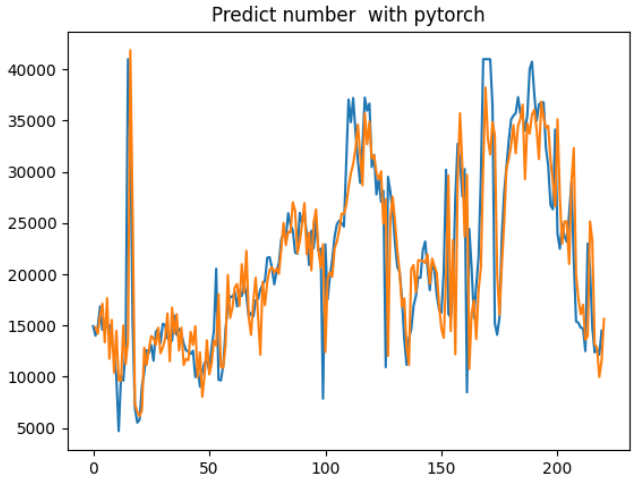

Figure 5: Prediction results for the entire dataset with label column

When the early stopping patience is set to 10 and the time step to 10, the LSTM model predicts the most recent five years of data with an MSE of 0.035288, a loss value of 0.044840, and an MSE of 0.4390218 for specific datasets.

DOI: 10.54254/2755-2721/2025.25252

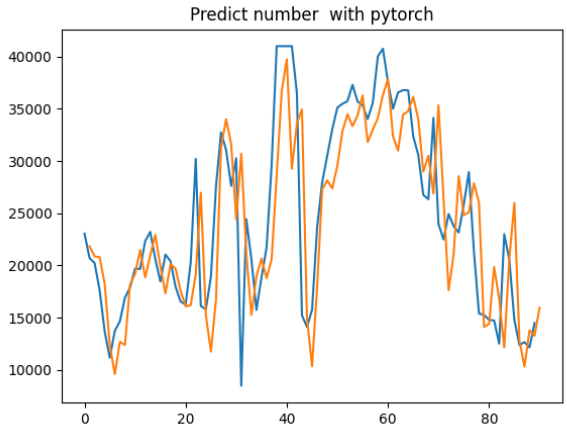

Figure 6: Prediction results for the most recent five years of data

Using the same training parameters for the entire dataset, the model achieved an MSE of 0.046189, a loss value of 0.070091, and an MSE of 0.38061055 for specific datasets.

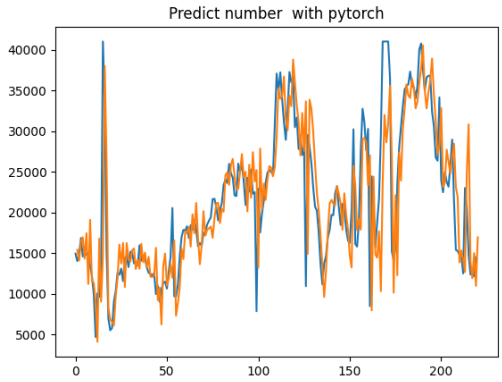

Figure 7: Prediction results for the entire dataset

When applied to the dataset with the added label column for the most recent five years, the LSTM model achieved an MSE of 0.036910, a loss value of 0.036910, and an MSE of 0.40358784 for specific datasets.

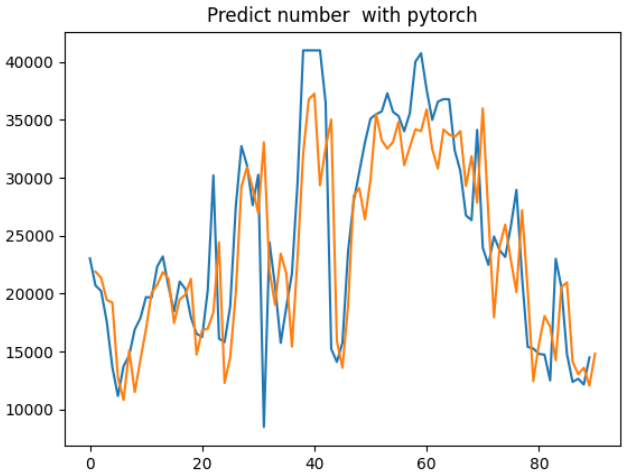

Figure 8: Prediction results for the most recent five years of data with label column

DOI: 10.54254/2755-2721/2025.25252

For the entire dataset with the added label column, the model recorded an MSE of 0.046788, a loss value of 0.074619, and an MSE of 0.3109944 for specific datasets.

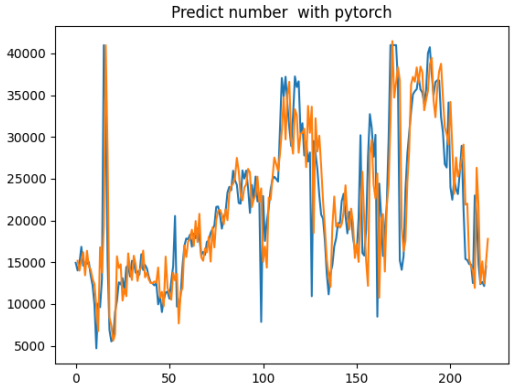

Figure 9: Prediction results for the entire dataset with label column

When the early stopping patience is set to 5 and the time step to 20, predictions for the most recent five years of data resulted in an MSE of 0.016523, a loss value of 0.021515, and an MSE of 0.44203086 for specific datasets.

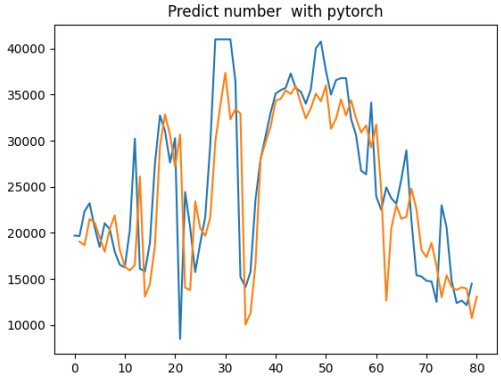

Figure 10: Prediction results for the most recent five years of data

Using the same training parameters for the entire dataset, the model yielded an MSE of 0.025479, a loss value of 0.033006, and an MSE of 0.33110597 for specific datasets.

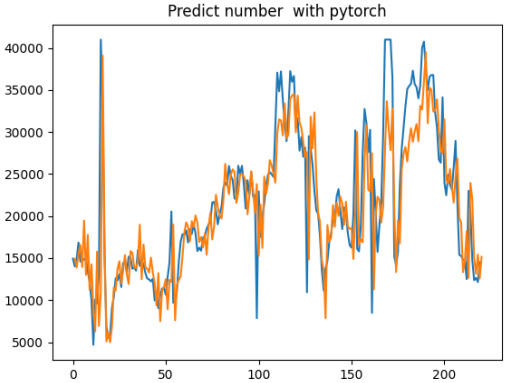

Figure 11: Prediction results for the entire dataset

DOI: 10.54254/2755-2721/2025.25252

Using the same training parameters, predictions for the most recent five years of data with the added label column resulted in a mean squared error (MSE) of 0.015482, a loss value of 0.018977, and an MSE of 0.41650833 for specific datasets.

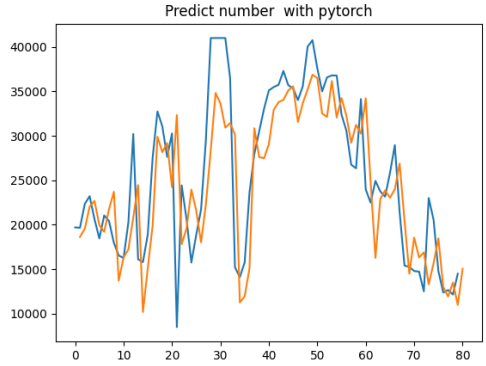

Figure 12: Prediction results for the most recent five years of data with label column

Using the same training parameters, predictions for the entire dataset with the added label column resulted in an MSE of 0.023708, a loss value of 0.028874, and an MSE of 0.3442923 for specific datasets.

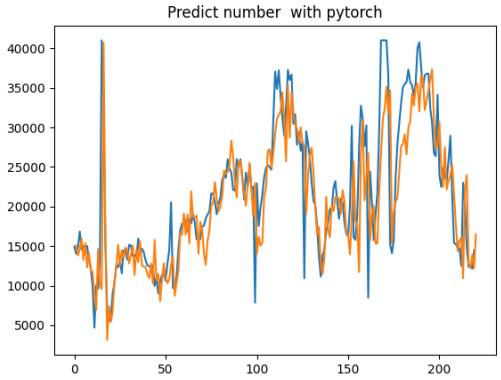

Figure 13: Prediction results for the entire dataset with label column

When the early stopping patience is set to 5 and the time step to 10, predictions for the most recent five years of data using the LSTM model resulted in an MSE of 0.034404, a loss value of 0.033714, and an MSE of 0.47583971 for specific datasets.

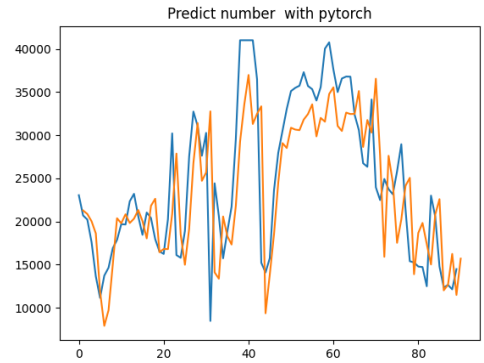

Figure 14: Prediction results for the most recent five years of data

DOI: 10.54254/2755-2721/2025.25252

Using the same training parameters for the entire dataset, the LSTM model achieved an MSE of 0.049147, a loss value of 0.072442, and an MSE of 0.3290522 for specific datasets.

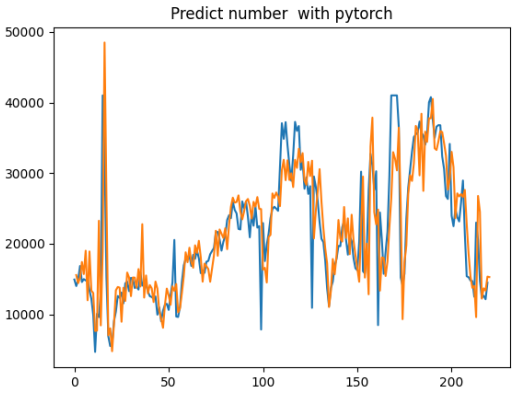

Figure 15: Prediction results for the entire dataset

When applied to the dataset with the added label column for the most recent five years, the LSTM model achieved an MSE of 0.038972, a loss value of 0.037905, and an MSE of 0.41460456 for specific datasets.

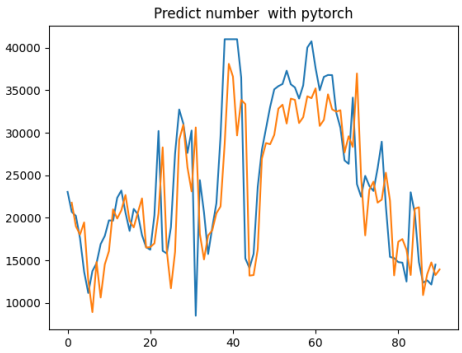

Figure 16: Prediction results for the most recent five years of data with label column

For the entire dataset with the added label column, the model recorded an MSE of 0.048893, a loss value of 0.068511, and an MSE of 0.33927119 for specific datasets.

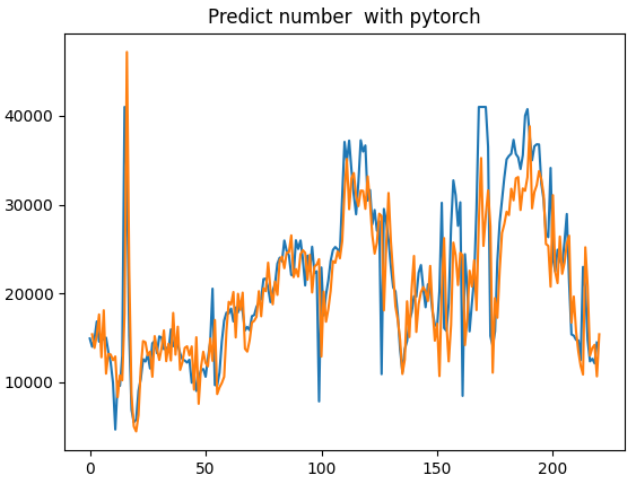

Figure 17: Prediction results for the entire dataset with label column

DOI: 10.54254/2755-2721/2025.25252

A summary of all data is presented in the following table:

Table 1: Summary of all data

Patience	Time Step	Dataset	Train Loss	Valid Loss	Mean Squared Error	label
10	20	5 years	0.015478	0.021785	0.38706189	No
10	20	All	0.021165	0.030537	0.36993607	No
10	10	5 years	0.035288	0.044840	0.4390218	No
10	10	All	0.046189	0.070091	0.38061055	No
5	20	5 years	0.016523	0.021515	0.44203086	No
5	20	All	0.025479	0.033006	0.33110597	No
5	10	5 years	0.034404	0.033714	0.47583971	No
5	10	All	0.049147	0.072442	0.3290522	No
10	20	5 years	0.015071	0.020296	0.39926783	Yes
10	20	All	0.024966	0.030110	0.33375474	Yes
10	10	5 years	0.034038	0.036910	0.40358784	Yes
10	10	All	0.046788	0.074619	0.3109944	Yes
5	20	5 years	0.015482	0.018977	0.41650833	Yes
5	20	All	0.023708	0.028874	0.3442923	Yes
5	10	5 years	0.038972	0.037905	0.41460456	Yes
5	10	All	0.048893	0.068511	0.33927119	Yes

5. Analysis

The Jiuzhaigou tourist dataset collected in this study contains significant missing data due to closures caused by the pandemic, earthquakes, and other factors. The missing data are irregular and lack any discernible pattern. Initially, the missing values were replaced with zeros, and the dataset was fitted using the LSTM model. However, this approach led to lower predicted values and poor prediction accuracy because the extensive missing data should not have been treated as zeros. To address this issue, the study replaced the missing data with the average values of the corresponding dates from the past two years. The dataset was then retrained using the LSTM model, resulting in a substantial improvement in prediction accuracy and better overall performance.

After processing all datasets using this method, the study explored the impact of different training parameters on the LSTM model's prediction accuracy to identify an optimal training strategy. The model's mean squared error (MSE), loss value, and MSE for specific datasets were used as criteria to evaluate prediction accuracy. The study reached the following conclusions:

1. When the training dataset and other training parameters were held constant, the size of the patience parameter and whether labels were used to identify the data had a minimal effect on prediction accuracy.

2. Within the range of time steps used in this study, longer time steps improved the LSTM model's performance for short-term tourist flow predictions.

3. For the dataset used in this study, the prediction accuracy was lower for datasets with a longer time span and higher for those with a shorter time span.

6. Conclusion

This study focuses on the short-term prediction of tourist flow in scenic areas, exploring the application of Long Short-Term Memory (LSTM) networks. Using daily tourist flow data from the Jiuzhaigou Scenic Area as a case study, various data preprocessing strategies, model parameter

DOI: 10.54254/2755-2721/2025.25252

settings, and dataset improvement methods (such as adding label columns) were designed and tested. The study systematically analyzed the effects of these combinations on prediction accuracy. It was found that filling missing data with the average values from the same dates in the past two years significantly improved the model's prediction accuracy, effectively addressing the issue of missing data in the dataset. Furthermore, the LSTM model demonstrated strong adaptability for short-term tourist flow predictions. Within the dataset and training parameter range used in this study, longer time spans in the dataset required longer time steps for the model, which resulted in higher prediction accuracy.

References

[1] Zhou, Y. (2019). A study on short-term tourist flow prediction methods considering spatial correlation in scenic areas (Doctoral dissertation, Nanjing Normal University).

[2] Ni, T., & Computer Science and Technology; Computer Application Technology. (n.d.). A study on daily tourist flow prediction methods for scenic areas based on machine learning (Doctoral dissertation, Xi'an University of Technology).

[3] Liu, X. (2021). Tourist short-term flow prediction methods and applications (Doctoral dissertation, Xi'an University of Electronic Science and Technology).

[4] Computer Science and Technology. (n.d.). Research on the application of machine learning in tourism data analysis (Doctoral dissertation).

[5] Hu, Y., Yang, L., & Fang, R. (2022). Data-driven research on tourism demand forecasting. Journal of Jinggangshan University (Natural Science Edition), 4, 7-14.

[6] Deng, Y., & Applied Statistics. (n.d.). Research on the application of ARIMA-ATT-LSTM in tourism flow prediction (Doctoral dissertation, Dalian University of Technology).

[7] Kang, J., Fu, Y., Fang, L., Li, M., Xie, Y., & Zhou, C. (2023). ARIMA-LSTM model for predicting inbound tourist numbers based on least squares weighting: A case study of Shanghai. Journal of Zhejiang University: Science Edition, 50(4), 508-520.

[8] Yang, R. (2023). Research and application of scenic area tourist flow warning system based on STM32 (Master's thesis, Guizhou University). https://link.cnki.net/doi/10.27047/d.cnki.ggudu.2023.001649. DOI: 10.27047/d.cnki.ggudu.2023.001649.

[9] Xie, T., & Liang, Y. (2024). Design of regional big data intelligent prediction algorithm considering complex influencing factors. Electronic Design Engineering, 06, 37-41. DOI: 10.14022/j.issn1674-6236.2024.06.008.

[10] Yu, J., Ye, X., He, X., & Xi, Z. (n.d.). GRU-BiLSTM tourism flow prediction method based on seasonal adjustment. Journal of Xihua Normal University (Natural Science Edition), 1-10.

[11] Li, M. (2024). Research on tourism demand forecasting based on mixed-frequency deep learning (Master's thesis, Guangxi University). https://link.cnki.net/doi/10.27034/d.cnki.ggxiu.2024.001124. DOI: 10.27034/d.cnki.ggxiu.2024.001124.

[12] Dai, Y., & Business Administration; Technology Economics and Management. (n.d.). Research on the impact factors and prediction of tourist flow based on multi-source network data (Doctoral dissertation).

[13] Cao, Y. (2023). Tourist flow prediction based on Baidu Index (Doctoral dissertation, Northeast University of Finance and Economics). DOI: 10.27006/d.cnkI.gdbcu.2023.001976.

[14] Xie, Q., Lu, M., & Xie, C. (2023). High-frequency tourist flow prediction in scenic areas based on LBS and deep learning. Journal of Earth Information Science, 02, 298-310.

[15] Xu, S. (2022). Tourist flow prediction for scenic areas based on multi-feature fusion (Doctoral dissertation, Hainan University).

[16] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2015). LSTM: A search space odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232. PMID: 27411231. arXiv:1503.04069. https://doi.org/10.1109/TNNLS.2016.2582924.