Cluster-based LSTM Network for Short-Term Passenger Flow Forecasting in Urban Rail Transit

Jinlei Zhang 1, Feng Chen 1,*, Qing Shen 2, *

1School of Civil Engineering, Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing 100044, China
2Department of Urban Design and Planning, University of Washington, Seattle, WA, USA
Corresponding author: Feng Chen (fengchen@bjtu.edu.cn); Qing Shen (qs@uw.edu)
This work was supported by the National Natural Science Foundation of China (grant numbers 71871027).

ABSTRACT Short-term passenger flow forecasting is an essential component for the operation of urban rail transit (URT). Therefore, it is necessary to obtain a higher prediction precision with the development of URT. As artificial intelligence becomes increasingly prevalent, many prediction methods including the long short-term memory network (LSTM) in the deep learning field have been applied in road transportation systems, which can give critical insights for URT. First, we propose a novel two-step K-Means clustering model to capture not only the passenger flow variation trends but also the ridership volume characteristics. Then, a predictability assessment model is developed to recommend a reasonable time granularity interval to aggregate passenger flows. Based on the clustering results and the recommended time granularity interval, the LSTM model, which is called CB-LSTM model, is proposed to conduct short-term passenger flow forecasting. Results show that the prediction based on subway station clusters can not only avoid the complication of developing numerous models for each of the hundreds of stations, but also improve the prediction performance, which make it possible to predict short-term passenger flow on a network scale using limited dataset. The results provide critical insights for subway operators and transportation policymakers.

INDEX TERMS LSTM; short-term passenger flow forecasting; urban rail transit; K-Means clustering; deep learning

I. INTRODUCTION

The urban rail transit (URT) is experiencing an explosive development in recent years in China. It plays a critical role in public transportation because of its advantages of large volume, fast speed, high punctuality, high safety, and low pollution [1]. Taking Beijing as an example, there are 22 subway lines operating as of 2018, with a total operating mileage of 637 km and 391 stations [2]. With the gradually increasing passenger flow, there will be more challenges for its operation, especially on how to deal with large passenger flow. Therefore, forecasting the passenger distribution in consecutive short terms is of great importance.

Short-term passenger flow forecasting (STPFF) is an essential component of the intelligent transportation system of the URT [3, 4]. Accurate real-time STPFF results can improve the passenger flow management. For policymakers and operators, they need to clearly understand the real-time spatiotemporal distribution of passenger flow through the STPFF results, so that they can properly limit the tap-in passengers to relieve congestion, as well as reasonably allocate staff members to evacuate passengers. The train timetable can be adjusted based on the STPFF results, such as by shortening or extending the headway, to transport more passengers or save on operation cost. For passengers, they can choose a better travel path to reduce travel time by avoiding crowded stations.

With regard to STPFF, there are many classification methods for the prediction models. According to the predicted objects, it can be divided into road STPFF models [5, 6], bus STPFF models [7], URT STPFF models [3, 4, 8-12], freeway STPFF models [13, 14] and railway STPFF models [15, 16]. According to the proposed time of the prediction model, it can be divided into conventional mathematical statistics-based prediction models[17] and emerging artificial intelligence-based prediction models [18]. According to whether the overall distribution of samples in the prediction models is directly known or not, it can be divided into parametric and non-parametric prediction models[19]. According to whether the prediction method is linear or not, it can be divided into linear prediction models [20, 21] and nonlinear prediction models [22]. According to the number of methods used in the prediction process, it can be divided into single prediction models [10] and hybrid prediction models, which means that
two or more prediction methods are applied to the model [4, 23, 24].

Passenger flow forecasting in URT mainly refers to the four-step traffic demand forecasting model, which comprises trip generation, distribution, mode choice, and route choice, for URT planning and construction [25-28]. While STPFF in URT have been relatively rare to date.

Some researchers have used single prediction models to conduct STPFF. For example, Roos et al. [8] built a prediction model based on a dynamic Bayesian network using multisource data, including on-board counting, ticket validation, and transport service data, which were aggregated in a time granularity (TG) of 2 min. What is distinct is that this model could deal with incomplete data using the expectation-maximization algorithm; however, it can be only applied to Paris due to its data particularity. Li et al. [10] proposed a novel multiscale radial basis function (RBF) network using automatic fare collection (AFC) data aggregated by 15 min to forecast irregular passenger flows, such as when there is a concert around a Beijing subway station. Similarly, to forecast the sudden increase in passenger flow, Feng et al. [11] built a random coefficient model based on a hierarchical Bayesian approach using survey data for tap-in passenger flow aggregated by 5 min. They also considered the surrounding environment of different entrances.

Meanwhile, other researchers have used hybrid models to achieve better results. For example, by using the AFC data in a TG of 15 min, Wang et al. [29] combined the seasonal autoregressive integrated moving average (ARIMA) model, which could capture the periodic features of passenger flow, and wavelet decomposition, which could capture the stochastic changing features, to conduct STPFF in Nanjing, China. This hybrid model has particular advantages when there are sudden flow changes due to special events. Wang et al. [3] also attempted to combine the seasonal ARIMA model and support vector machine (SVM) model to conduct STPFF. Then the average predicted values of the two models were considered as the final prediction value. Li [12] applied a gray back-propagation neural network (BPNN) by integrating a gray system and a BPNN model to forecast the hourly passenger flow in a subway station; however, in their paper, they did not explain the modeling process in detail. Wei and Chen [4] put forward an STPFF approach that combined the empirical mode decomposition and BPNN using AFC data in a 15-minute TG in Taipei. Results indicated that the hybrid model performed better than the corresponding single model.

As mentioned above, many methods for STPFF in URT based on artificial intelligence and machine learning, such as BPNN, SVM, and dynamic Bayesian network, have been applied. With deep learning becoming increasingly prevalent, many studies have attempted to employ the state-of-art long short-term memory network (LSTM), encoder-decoder LSTM, convolutional LSTM (ConvLSTM), and deep residual network (ResNet) into the domain of STPFF. Most of these studies are based on LSTM. The LSTM is included in feedback neural networks and has a history of 21 years [30]. It is a type of recurrent neural network (RNN) used in deep learning. The reason that the conventional RNN has achieved great success is that it can connect previous information to a current task, which cannot be realized by feedforward neural networks, such as the multilayer perceptron. However, when the gap between the relevant information and the point where it is required becomes very large, the RNN is unable to learn to connect the information because of the main problem of long-term dependency, which refers to the vanishing gradient and exploding gradient [31, 32]. Different from the standard RNN, the LSTM can fortunately avoid the long-term dependency problem. It can selectively store and access useful information [30], thereby greatly outperforming the conventional RNN in many specific fields. It has been successfully applied to the fields of natural language processing [33], speech recognition [34], etc.

In the domain of road traffic, there are many relevant studies than can give valuable insights. The existing studies include but are not limited to following. Ma et al. [35] employed a three-layer LSTM to predict urban travel speed using remote traffic microwave sensor data. To predict the traffic speed, they used speed together with traffic volume as input data. Results indicated that more input information could achieve a higher precision. Further, they also treated traffic as images and proposed a convolutional neural network (CNN)-based method to predicts large-scale, network-wide traffic speed with a high accuracy [36]. Fu et al. [37] utilized the LSTM and gated recurrent unit (GRU) network to predict traffic volume. They used the past six data points to predict the next data point with a five-minute time span. Azzouni and Pujolle [38] proposed an LSTM-RNN framework to predict the traffic matrix in a large network. Results also showed a much higher precision than the benchmark methods in the domain of traffic matrix prediction. After the introduction of convolutional LSTM (ConvLSTM) used for precipitation nowcasting [39], many researchers began to consider spatial characteristics. Lv et al. [40] built a stacked autoencoder (SAE) model considering spatial and temporal correlations to predict short-term traffic flow. Zhao et al. [41] built a two-dimensional LSTM to predict traffic volume. In their study, the origin-destination matrix was used in the model to capture the temporal and spatial correlation of traffic volume in different observation stations. They predicted the traffic flow in TG values of 15, 30, 45, and 60 min, respectively. Similarly, to capture spatiotemporal features, Liu et al. [42] built a ConvLSTM network combined with the bidirectional LSTM (Bi-LSTM) to implement traffic flow forecasting. Since the deep residual network (ResNet) was proposed for image recognition, many studies also tried to apply the ResNet in road traffic time series prediction, such as citywide crowd flows prediction [43] and hotspot traffic speed prediction [44].

In the field of URT, there are also some latest applications to conduct STPFF based on deep learning methods. Guo et al. [45] built a support vector regression (SVR) and LSTM based
fusion model to predict abnormal passenger flows in URT. Tang et al. [46] proposed a ST-LSTM model combining passenger flow, time cost matrix and spatial correction matrix to conduct STPFF in URT. Liu et al. [47] built a model based on LSTM model involving environmental factors and temporal and spatial factors.

As we can observe, there are many applications of LSTM and other deep learning methods (e.g. ConvLSTM, ResNet) to the traffic prediction field. Most of them are divided into complex network architectures to gain better performance. However, how to improve the prediction precision based on simple network architecture is also an important research question. Moreover, most of them integrate spatial correlations into their models. The proposed model in our study is based on the clustering results, and in each cluster, stations tend to have high similarity in spatial location. Therefore, we also considered not only temporal correlation but also local spatial correlation to some extent.

To sum up, there are some common shortcomings in most of the existing studies for STPFF in URT. First, most studies only took a few stations or a line as examples to conduct the prediction and they did not take a global perspective to conduct STPFF for all stations in a metro network. If the model would be applied to conduct STPFF for all stations, that model may achieve a higher precision in the testing candidate stations; however, it might cause a lower precision for many other stations because of ignoring the substantial differences between stations. Furthermore, as already known, there are always hundreds of stations in a URT network. Therefore, it is unrealistic to develop numerous models for each of the stations. Second, the TG, which is the time interval that passenger flows are aggregated, used in the existing studies varied from several minutes to several days, which means that they did not consider how long it was reasonable to predict passenger flow in the future. However, choosing a reasonable TG is very important because a small TG can capture the detailed passenger flow information but will be more difficult to predict because of the considerable flow fluctuation. In contrast, a large TG can ensure more prediction precision but will lose more detailed passenger flow information. Third, the data is not completely open to the public and is hard to obtain nowadays. As already known, it is necessary to acquire more training dataset in the field of machine learning to improve the prediction performance. Therefore, how to improve the prediction performance using limited dataset is also an existing problem.

To compensate for the shortcomings of the existing research, we first propose a novel two-step K-Means clustering method to cluster subway stations based on passenger flow characteristics. Within each subclass, passenger flows not only have similar variation trends but also have similar ridership volumes. Besides, stations in a class always have similar spatial locations. Therefore, local spatial correlations were also considered in the proposed model to some extent. In addition, a predictability assessment combined model, including the similarity measurement model and stationarity test model, is described to assess the predictability of passenger flows aggregated by different TGs. Then, a reasonable TG interval for every station cluster will be recommended and used as inputs for the prediction process. This model can demonstrate the rationality of TG used in this study theoretically. Finally, the cluster based LSTM model (CB-LSTM) is applied to conduct the STPFF in URT to obtain a higher global prediction precision. Conducting STPFF based on subclasses can allow more training dataset to be used for model development, make it realistic for the prediction on a network scale, as well as avoid developing numerous models for each of the hundreds of stations.

The study framework is shown in Figure 1, and the remainder of this paper is organized as follows. In section 2, we provide the data source and preprocessing. In section 3, the methodology, including the two-step K-Means clustering model, the predictability assessment combined model, and the LSTM, is described in detail. A case study and its results are presented in section 4. The conclusion is summarized in section 5.

**FIGURE 1 Framework of this study**

**II. DATA DESCRIPTION AND PREPROCESSING**

The AFC data from February 29, 2016 to April 3, 2016 are collected from Beijing subway. They contain consecutive five-week datasets from 05:00 to 24:00 (19 h or 1140 min). The useful fields extracted from the original dataset are the card number, entering station number, exiting station number, entering time, exiting time, entering station name, and exiting station name. In March 2016, there were 17 lines and 276 stations in total (No repeated counting for interchange stations and excluding airport express lines). Total records reach up to 130 million for 25 weekdays. For interchange stations in two or more lines, they have two or more station numbers. To improve the extraction of the passenger flow time series, we give a unique station number for all stations. Moreover, the
Entering and exiting time are measured in minutes from 0 to 1140, representing 05:00 to 24:00 in a day. An example of AFC data before and after processing is shown in TABLE I and TABLE II, respectively.

<table>
<thead>
<tr>
<th>Index</th>
<th>Card number</th>
<th>Entering station number</th>
<th>Exiting station number</th>
<th>Entering time</th>
<th>Exiting time</th>
<th>Entering station name</th>
<th>Exiting station name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74873***</td>
<td>121</td>
<td>203</td>
<td>20160309190900</td>
<td>20160309193524</td>
<td>Yonganli</td>
<td>Fuchengmen</td>
</tr>
<tr>
<td>2</td>
<td>19727***</td>
<td>643</td>
<td>210</td>
<td>20160309123200</td>
<td>20160309124606</td>
<td>Chaoyangmen</td>
<td>Beijingshan</td>
</tr>
<tr>
<td>3</td>
<td>42656***</td>
<td>9708</td>
<td>20160309115600</td>
<td></td>
<td></td>
<td>Tongzhoubaiyuan</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>Card number</th>
<th>Entering station number</th>
<th>Exiting station number</th>
<th>Entering time</th>
<th>Exiting time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>74873***</td>
<td>19</td>
<td>37</td>
<td>849</td>
<td>875</td>
</tr>
<tr>
<td>2</td>
<td>19727***</td>
<td>44</td>
<td>43</td>
<td>452</td>
<td>466</td>
</tr>
<tr>
<td>3</td>
<td>42656***</td>
<td>29</td>
<td>43</td>
<td>356</td>
<td>464</td>
</tr>
</tbody>
</table>

### III. METHODOLOGY

In this section, we formulate the methodological framework of this study. First, a two-step K-Means clustering model is introduced based on the passenger flow characteristics. Second, the predictability assessment model is built to recommend a reasonable TG interval for conducting the STPFF. Third, we describe the LSTM model in detail.

#### TWO-STEP K-MEANS CLUSTERING MODEL

In existing works, there are many clustering methods, among which the K-Means is greatly popular owing to its simplicity, understandability, and fast convergence speed [48]. For clustering subway stations, there are two prevailing methods to define the variables used in the clustering process [49]. One is using ridership indicators (RI), such as the specific ratio of the hourly ridership over the daily ridership or the maximum hourly ridership [50], as variables, which can capture passenger flow variations throughout a day while ignoring the ridership volume to some extent. The other is directly using the hourly ridership values as variables [49, 51], which can capture the ridership volume throughout a day while ignoring the passenger flow variation trend to some extent.

To capture both the variation information and volume information, we introduce a two-step K-Means clustering model. The first step is to obtain the main classes using RI, within which stations have similar passenger flow variation trends. However, the ridership volume greatly varies within each cluster, which will be harmful if we use all datasets to develop one STPFF model for all stations within each cluster. Thus, the second step is to obtain subclasses within every main class using the hourly ridership directly, within which every subclass has a similar ridership volume. Therefore, in every subclass, there are similar passenger flow variation trends as well as ridership volumes, which will enhance the STPFF when using LSTMs. Compared with the ordinary K-Means, this two-step K-Means is more refined and focused, especially in the second step. The clustering process of the two-step K-Means is as follows.

**Step (1.1):** Extract the hourly ridership of subway stations as indicated in Equation (1), where $R_{s j}$ is the $j^{th}$ hourly ridership of station $S$. Then, calculate the $RI$ value as indicated in Equations (2) and (3), where $RI_{s j}$ represents the $j^{th}$ RI of station $S$.

$$
R_{s j} = \begin{pmatrix}
R_{s 1} & R_{s 2} & \cdots & R_{s j} \\
R_{s 1} & R_{s 2} & \cdots & R_{s j} \\
\vdots & \vdots & \ddots & \vdots \\
R_{s 1} & R_{s 2} & \cdots & R_{s j}
\end{pmatrix}
$$

(1)

$$
RI_{s j} = \frac{R_{s j}}{\sum_{i} R_{s i}}
$$

(2)

**Step (1.2):** Conduct the K-Means algorithm using the $(RI_{s j})$ matrix. During this process, we use the K-Means++ algorithm to initiate the clustering centroids. The loss function (Loss) is defined as

$$
Loss = \sum_{i=1}^{n} \sum_{j=1}^{m} \min_{C_j \in \mathbb{C}} \left( \left\| X_{ij} - C_{j} \right\|^2 \right)
$$

(4)

where $X_{ij}$ are the samples, $C_j$ are the clustering centroids, and $n_k$ is the sample size in each class $k$. Furthermore, we use a decrement of this loss function as the clustering criterion to choose the proper number of clusters.

$$
Decrement = Loss_k - Loss_{k-1}
$$

(5)

**Step (2):** For each main cluster, we then repeat step (1.2) using the $(R_{s j})$ matrix to obtain the corresponding subclasses.

After this process, we obtain the station clusters capturing not only similar passenger flow variation trends but also ridership volume characteristics.

#### PREDICTABILITY ASSESSMENT COMBINED MODEL

A reasonable TG is the basis of STPFF. However, in most previous studies, scholars seldom considered whether the selected TG for conducting STPFF is theoretically reasonable
or not. Choosing a reasonable TG to conduct STPFF can not only capture more detailed passenger flow information but also contribute to obtain an acceptable prediction precision. Therefore, the chosen TG is not “the smaller the better,” nor is it “the larger the better.”

Higher similarity and stationarity of passenger flow with historical data contribute to a higher prediction precision. Based on this, Zhang et al. [52] measured the similarity of passenger flows over the same historical period (e.g., the similarity of passenger flow between this Monday and last Monday) using a correction coefficient matrix, and attempted to explore the best TG for STPFF. Wang [53] measured the passenger flow similarity and tested the passenger flow stationarity using the augmented Dickey–Fuller (ADF) test, which examines the stationarity of passenger flow in a specific day (e.g., the stationarity of passenger flow in Monday) to find the best TG for STPFF. In their study, they just chose several specific TG values, such as 1 min and 5 min. Considering different operational requirements, we attempt to obtain a reasonable TG interval rather than only one specific TG. Furthermore, we combine the similarity measurement method and the ADF test as a more precise combined model to recommend more reasonable TG intervals as shown in Figure 2. The methods are described in detail as follows.

![FIGURE 2 Framework for predictability assessment combined model](image)

1) SIMILARITY MEASUREMENT MODEL
To measure the similarity, we use Pearson’s correlation coefficient as given by Equation (6) [53].

\[ r_{S}(i_{D}, j_{D}) = \frac{\sum_{t=1}^{n} (S_{i_{D}(t)} - \bar{S}_{i_{D}})(S_{j_{D}(t)} - \bar{S}_{j_{D}})}{(\sum_{t=1}^{n} (S_{i_{D}(t)} - \bar{S}_{i_{D}})^2)(\sum_{t=1}^{n} (S_{j_{D}(t)} - \bar{S}_{j_{D}})^2)^{1/2}} \]  

(6)

where \( S_{t} \) is the station \( S \) indexed with \( t \), \( i_{D} \) is the day \( D \) in the week indexed with \( i \), \( j_{D} \) is the day \( D \) in the week indexed with \( j \), and \( x_{S(t)} \) is the passenger flow vector.

There are 5 weeks of AFC data. Thus, after calculating the correlation coefficients, we will obtain a correlation coefficient matrix of the fifth order \( (R_{S}(i_{D}, j_{D})) \) as indicated in Equation (7). Then the average correlation coefficient for day \( D \) of station \( S \) \( (C_{S,D}) \) will be calculated by Equation (8) [53], where \( l \) is the order number. Because all analyses are based on subway station clusters, there will be an integrated correlation coefficient calculated by Equation (9) for the class indexed with \( k \) \( (C_{k,D}) \).

\[ R_{S}(i_{D}, j_{D}) = \begin{bmatrix} 1, r_{S}(2_{D}, 1_{D}), r_{S}(3_{D}, 1_{D}), r_{S}(4_{D}, 1_{D}), r_{S}(5_{D}, 1_{D}) \\ r_{S}(1_{D}, 2_{D}), 1, r_{S}(2_{D}, 2_{D}), r_{S}(4_{D}, 2_{D}), r_{S}(5_{D}, 2_{D}) \\ r_{S}(1_{D}, 3_{D}), r_{S}(2_{D}, 3_{D}), 1, r_{S}(4_{D}, 3_{D}), r_{S}(5_{D}, 3_{D}) \\ r_{S}(1_{D}, 4_{D}), r_{S}(2_{D}, 4_{D}), r_{S}(3_{D}, 4_{D}), 1, r_{S}(5_{D}, 4_{D}) \\ r_{S}(1_{D}, 5_{D}), r_{S}(2_{D}, 5_{D}), r_{S}(3_{D}, 5_{D}), r_{S}(4_{D}, 5_{D}), 1 \end{bmatrix} \]  

(7)

\[ C_{S,D} = \sum_{i=1}^{l} \sum_{j=1}^{n} \frac{r_{S}^{2}(i_{D}, j_{D})}{(i(i-1)/2)} \]  

(8)

\[ C_{k,D} = \frac{1}{S} \sum_{S} C_{S,D} \]  

(9)

After completing all the above calculations, we will obtain the \( C_{k,D} \) for different clusters in different TGs as given by Equation (10), where \( i \) denotes different TG values.

\[ C_{k,D} = (C_{k,D}^{1}, C_{k,D}^{2}, C_{k,D}^{3}, \cdots, C_{k,D}^{l}) \]  

(10)

The coefficient varies from 0 to 1. The larger the coefficient, the more similar the passenger flow series between weekdays and the more suitable it will be for prediction. If the coefficient is larger than a given similarity level, we can determine that passenger flows aggregated in that specific TG have a good similarity and can be used for prediction under the given level.

2) STATIONARITY TEST MODEL (ADF TEST)
The ADF test, which is popular in statistics and econometrics, is used to test the stationarity of passenger flow series. It is an augmented version of the Dickey–Fuller test and is more suitable for larger and complicated datasets. Compared to the Dickey–Fuller test, it allows the autocorrelation of the original series to have more lag orders. The autoregressive model in the test is given by Equation (11).

\[ \Delta y_{t} = a + bt + cy_{t-1} + \sum_{i=1}^{n-1} d_{i}\Delta y_{t-i} + \epsilon_{t} \]  

(11)

where \( a \) is a constant, \( b \) is the coefficient on a time trend, and \( n \) is the lag order of the autoregressive process. Before carrying out the ADF test, we need to first determine whether the constant and time trend will be included in the test regression. It is extremely obvious that the passenger flow series changes regularly with time to some extent. Additionally, there are also some unexpected fluctuations. Therefore, we choose a more general specification, including the constant and time trend in our model. Moreover, the Akaike information criterion is applied to specify the lag length of \( n \).

After completing the above specifications, the ADF test will be conducted for every subway station using the AFC data.
from Monday to Friday. The \textit{p-value} will be obtained for every station in a specific TG as indicated in Equation (12).

\[
P_s = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_s \end{pmatrix}
\]

(12)

Because all the analyses are based on subway station clusters, there will be an integrated \textit{p-value} calculated by Equation (13) for the class indexed with \( k \) (\( P_k \)).

\[
P_k = \frac{1}{s} \sum_s P_s
\]

(13)

After completing all the above calculations, we will obtain the \( P_k \) for different clusters in different TGs as indicated in Equation (14), where \( i \) denote different TGs.

\[
P_k = \{ p^1_k, p^2_k, p^3_k, \ldots, p^i_k \}
\]

(14)

The \textit{p-value} varies from 0 to 1. The smaller the \textit{p-value} obtained from the result of the ADF test, the more stationary the time series and the more suitable will it be for prediction. If the \textit{p-value} is smaller than a given significance level, we can consider that the test result is statistically significant, which means that the time series is stationary. Therefore, the passenger flow in a specific TG can be used for prediction under the given significance level.

3) \textbf{PREDICTABILITY ASSESSMENT COMBINED MODEL}

As mentioned above, the passenger flow series aggregated by a specific TG with more similarity as well as more stationarity is more suitable for conducting STPFF. If the original passenger flow in a specific TG has a higher similarity with historical data but lower stationarity, or vice versa, it is not suitable for conducting STPFF using historical data because it may not achieve an acceptable prediction precision theoretically. Thus, it is not enough to use one of the above two methods. Because of this, we determine reasonable TG intervals by combining the results of the similarity measurement model and stationarity test model; this is called the predictability assessment combined model. In the context of the given similarity level and significance level, the final TG interval is determined among ones that pass both the similarity measurement and stationarity test.

\textbf{LSTM MODEL}

The LSTM is distinguished from RNN for its memory block rather than the original neurons as shown in Figure 3. They are just similar to many copies of the same neural networks, but what is distinct is that every copy of the neural network passes information to its successor after every time step. This means that they can connect past information to the present. The biggest difference between RNN and LSTM is the architecture of neurons distributed in the hidden layer. The neurons of the conventional RNN have only a simple structure, such as only having a sigmoid layer. The recurrent process can be calculated by equation (15) and (16).

\[
h_t = f(Wx_t + Uh_{t-1})
\]

(15)

\[
y_t = Vh_t
\]

(16)

where \( f(\cdot) \) is the activation function, and \( W, U, \) and \( V \) denote the weight matrices.

The LSTM block as shown in Figure 4 is much more complicated; however, this complicated and useful calculating mechanism ensures the ability of solving the long-term dependency problem. It is formed by memory cells and three gates, namely the input gate, output gate, and forget gate, all of which are neural network layers consisting of many neurons. The gates can monitor what information will be input, output, and forgotten, respectively. The memory cells are regulated by these gates. They record the past and present system states, which denote the cell state. By imposing the original input \( x_t \), input gate \( i_t \), output gate \( o_t \), forget gate \( f_t \), cell state \( c_t \), and final output \( h_t \), the prediction can be performed by using Equation (17) to (24) [54-56] to train the network and obtain prediction results.

\[
f_t = \sigma(W_f x_t + Uh_{t-1} + V_f c_{t-1} + b_f)
\]

(17)

\[
i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i)
\]

(18)

\[
c_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)
\]

(19)

\[
f_t \odot c_{t-1} + i_t \odot c_t
\]

(20)

\[
o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o)
\]

(21)

\[
h_t = o_t \odot \tanh(c_t)
\]

(22)

\[
s(x) = \frac{1}{1 + e^{-x}}
\]

(23)

\[
tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}
\]

(24)

where \( t \) is the time step, \( \odot \) is the Hadamard product, and the initial values are \( c_t = 0 \) and \( h_t = 0 \). In addition, \( \sigma(x) \) is the logistic sigmoid activation function and \( \tanh(x) \) is the hyperbolic tangent activation function. \( W, U, \) and \( V \) denote the weight matrices and \( b \) is the bias vector parameter. Figure 4 is an enlarged view of the block in Figure 3 (b). The marks \( \otimes \) and \( \oplus \) denote matrix multiplication and addition, respectively.
All the above calculations are called forward propagation. However, the mean square error, which is used as the loss function in this study, will expand with time. Therefore, the back propagation through time (BPTT) algorithm and Adam optimizer are utilized to optimize the weight parameters. Different from the conventional RNN, which will solve the problem of vanishing gradient or exploding gradient, the LSTM block can truncate the errors until they become zero during the back propagation process. Through constant forward propagation of data and back propagation of errors, the network parameters will remain relatively stable and can eventually be used for prediction.

IV. CASE STUDY AND RESULT DISCUSSION

In this section, we will discuss the application results of the methodology. First, the clustering result is presented. Then, based on the clusters, the predictability assessment, namely similarity measurement and stationarity test, is conducted to recommend a reasonable TG interval, which will be used to extract the passenger flow series later. Third, the constructed LSTM model is described in detail and its outcome for STPFF is presented.

TWO-STEP K-MEANS CLUSTERING RESULTS

First, the subway stations in Beijing are clustered using the hourly tap-in ridership indicators as mentioned in section 3.1, so that stations with similar passenger flow variations throughout a day will be gathered together. The hourly tap-in
ridership for every station is the average data extracted from March 7 to 11 (Monday to Friday) to avoid some random factors. The loss of the objective function and its decrement are shown in Figure 5. As clearly shown, when the cluster number is greater than 6, the loss has a relatively small and stable decrement, which means that the value of 6 can be considered as the “elbow”. Thus, the final cluster number is set as 6.

**FIGURE 5 LOSS AND ITS DECREMENT**

The clustering result is shown in Figure 6. The spatial distribution and the station number in each cluster are shown in Figure 7. It is evident that the passenger flows have extremely similar variation trends throughout a day within each cluster in terms of the tap-in ridership. Meanwhile, a large passenger flow variation trend difference exists between clusters.

For **cluster 1** stations, the tap-in ridership in the morning peak periods is extremely high. After the morning peak periods, it decreases to a very low level and remains stable. The trend of tap-out ridership is just the reverse. These stations are mainly distributed in the suburb of Beijing as shown in Figure 7. We call these stations as classical “residential stations.” **Cluster 2** has a similar change trend as cluster 1. The difference is that during evening peak periods, the tap-in ridership has a slight increase compared to that in cluster 1. The tap-out ridership also has a slightly higher increment than that in cluster 1 during the morning peak periods. Compared to cluster 1 stations, these stations are located a little bit further inside of Beijing. We call them as “hybrid-residential oriented stations.”

For **cluster 3** stations, their tap-in ridership climbs to a higher level during evening peak periods while it stays in a relatively lower and stable level in other periods. The tap-out ridership shows an opposite trend. These stations are located right in downtown Beijing as shown in Figure 7. We call these stations as “working stations.” The difference between cluster 3 and **cluster 4** is similar to that between cluster 1 and cluster 2, and we call stations in cluster 4 as “hybrid-working oriented stations,” which are also located in downtown Beijing.

For **cluster 5**, the ridership characteristics present an obvious “double peak” trend as for the tap-in as well as tap-out ridership. They dispersedly distribute in the downtown and the interface between the downtown and suburb. We call these stations as “ordinary stations.”

For **cluster 6**, the ridership remains in a relatively stable level throughout a day, either in a high level or in a low level. Their characteristics of morning and evening peaks are not as obvious as those in cluster 5. Most of them are traffic hubs or major railway stations that are the most dispersedly located throughout Beijing City. We call these stations as “ordinary-traffic-hub oriented stations.”

As is mentioned above, stations in the same cluster have very similar local spatial locations, which means that we also considered not only temporal correlation but also local spatial correlation to some extent in the proposed CB-LSTM model.
FIGURE 6 HOURLY RIDERSHIP DISTRIBUTION OF STATIONS IN DIFFERENT CLUSTERS
As we can observe from Figure 6, although the passenger flow within each cluster has a similar variation trend, the ridership volume varies significantly. Thus, after the first step of K-Means, we carry out the second step using the hourly ridership to divide the main cluster into subclasses. According to the loss decrement of the objective function, we determine the subclass number within each main cluster as shown in Figure 8.

Taking cluster 2 as an example, we divide this main cluster of 71 stations into 6 subclasses. The hourly tap-in ridership variation is shown in Figure 9. Obviously, in each subclass, the passenger flows do not only have more similar variation trends but also have more similar ridership volumes compared to their main cluster. Therefore, the STPFF is based on the subclass clustering result.

Given the paper length, other clustering results are presented in Appendix.
more on variation trends rather than ridership volumes. Based on the methodology stated in section 3.2, we calculate the average coefficient between five consecutive Mondays and the same calculations for Tuesdays to Fridays for each cluster. Furthermore, for each cluster, the ADF test is conducted using the data from Monday to Friday in the second week to obtain the average p-value.
From a real-time and short-term perspective, 30 TGS from 1 min to 30 min in 1-minute intervals are chosen because the headways of URT in Beijing have been shortened to nearly 1 min in specific stations. The similarity level for the similarity measurement and the significance level for the ADF test are assigned the values of 0.9 and 0.05, respectively, in our study. This means that if the coefficient is higher than 0.9 and the \( p \)-value is lower than 0.05 in a given TG, then the TG can be recommended to aggregate the passenger flow series for prediction. The assessment results are shown in Figure 10 and TABLE IV.

For cluster 1 to cluster 3, they have all passed the stationarity test because their \( p \)-values are far below the given significance level of 0.05. This means that the passenger flow series are stationary with a high probability of 95%. All the coefficients are greater than the given similarity level of 0.9, which means that the passenger flow series from these clusters maintain a high similarity of 90% or more with historical AFC data in the same period. Furthermore, it is obvious that the coefficient gradually approaches to 1 with TG, because a larger TG will cause more detailed passenger flow information being ignored, and the predictability, therefore, increases.

Therefore, theoretically, the TG interval of 1 to 30 min is recommended at the given similarity level and significance level. As for cluster 4 stations, they pass the stationarity test as well. However, for the similarity measurement, the coefficient of TG under 5 min is lower than the given similarity level of 0.9. Thus, for this cluster, a TG interval of 5 to 30 min is recommended.

In terms of the TG within 30 min for cluster 5 and cluster 6, all the passenger flow series from Monday to Friday cannot pass the similarity measurement under the given similarity level of 0.9 because of the relatively large passenger flow fluctuation. Hence, we extend the TG from 30 to 120 min. According to the results of cluster 5, the coefficient for TG values from 30–120 min is greater than 0.9, while the \( p \)-values from 1–30 and 40–95 min are lower than 0.05. Therefore, a TG interval of 40–95 min for cluster 5 is recommended after taking the intersection. Similarly, according to the results of cluster 6, the coefficient for TG values from 60–120 min is greater than 0.9, while the \( p \)-values from 1–30 and 40–87 min are lower than 0.05. Thus, the TG interval of 40–87 min for cluster 6 is recommended after taking the intersection.
FIGURE 10 PREDICTABILITY ASSESSMENT RESULTS FOR DIFFERENT MAIN CLUSTERS

TABLE IV
RECOMMENDED TG INTERVAL AFTER PREDICTABILITY ASSESSMENT

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Recommended TG Interval</th>
<th>Cluster</th>
<th>Recommended TG Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1–30 min</td>
<td>4</td>
<td>5–30 min</td>
</tr>
<tr>
<td>2</td>
<td>1–30 min</td>
<td>5</td>
<td>40–95 min</td>
</tr>
<tr>
<td>3</td>
<td>1–30 min</td>
<td>6</td>
<td>40–87 min</td>
</tr>
</tbody>
</table>

Note: The given similarity level is 0.9 and the given significance level is 0.05.
CLUSTER-BASED LSTM MODELING RESULTS

Based on results of the two-step K-Means model and the predictability assessment combined model, we take 9 stations in subclass 4 of main class 2 as an example as shown in Figure 11. For this cluster, the recommended TG under the given level is 1 to 30 min. In consideration of the “short-term,” we choose the TG from 1 to 5 min to aggregate the passenger flow series and compare the prediction performance under different TGs.

We apply three indicators to evaluate its performance, namely root mean squared error (RMSE), mean absolute error (MAE), and R-squared ($R^2$) as given by Equations (25) – (27).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$  \hspace{1cm} (25)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$  \hspace{1cm} (26)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$  \hspace{1cm} (27)

where $y_i$ is the prediction value, $\hat{y}_i$ is the actual value, $\bar{y}$ is the average value of all true values, and $n$ is the number of samples.

The data in 25 workdays from the consecutive five-week AFC data are divided into two parts, within which the first 20 workdays are used to train the model and the other 5 workdays are used to test the model. For the 9 stations, there are approximately 1.15 million records in total for the 25 workdays. The average ridership in a workday is listed in TABLE V. The data used in the prediction process varies from 05:00 to 23:00. Therefore, for TG values from 1 to 5 min, the observations in a day are 1,080, 540, 360, 270, and 216, respectively.

Furthermore, we use a moving prediction method as shown in Figure 12. Supposing that there are five time steps in a time window, the first four steps will be input to the model to forecast the last one. After that, the time window will move forward, with the latest actual value being added to the window, and the oldest value being deleted for capturing influences of the latest passenger flow information. After reconstructing the structure of the original passenger flow series, they are rescaled to the range of −1 to 1. After prediction, they will be inversely transformed to the original scale to assess the model performance.

![FIGURE 11 SPATIAL DISTRIBUTION OF 9 STATIONS IN SUBCLASS 4](image)

<table>
<thead>
<tr>
<th>Name</th>
<th>Yuquanlu</th>
<th>Shuangqiao</th>
<th>Jiaomenxi</th>
<th>Gongyixiqiao</th>
<th>Liujiayao</th>
<th>Puhuangyu</th>
<th>Liuliqiaodong</th>
<th>Panjiayuan</th>
<th>Jiaoemendong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station Index</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Ridership</td>
<td>27113</td>
<td>21071</td>
<td>24879</td>
<td>23747</td>
<td>30809</td>
<td>24928</td>
<td>27266</td>
<td>27530</td>
<td>22118</td>
</tr>
</tbody>
</table>
For the 9 stations, we develop 2 methods to carry out the prediction and then compare their performance. Both of them are used to verify the methodology introduced in this study. Method 1, which is the CB-LSTM model, is an innovative way of developing models based on clusters, whereas Method 2 is the ordinary LSTM model, i.e., developing a specific model for each of the 9 stations. They are described in detail as follows.

1) METHOD 1: CLUSTER-BASED LSTM MODEL
CB-LSTM is an innovative model that we introduce in this study, i.e., developing the STPFF model based on clusters as shown in Figure 13. Because the passenger flow series within a subclass do not only have similar variation trends but also have similar ridership volumes, it is feasible to integrate the data in the first 20 workdays of the 9 stations to train the model. Moreover, the data is not completely open to the public and is hard to obtain nowadays. And this method can increase the training dataset volume using limited dataset, which is extremely beneficial to achieve a satisfactory performance in the machine learning field. Then, prediction is conducted for each of the 9 stations and the average model performances, namely MAPE, MAE, and \( R^2 \), are obtained. Furthermore, eight other models in the field of artificial intelligence, namely BPNN, GRU, ordinary RNN, SVM, ARIMA, CNN, ConvLSTM and ResNet are utilized to compare the performance of CB-LSTM.

The CB-LSTM configuration performing best has an input layer, a stateless LSTM block layer, and an output layer. By trial and error during the hyper-parameter tuning process, 100 neurons are put into every layer in the LSTM block. In addition, 30 time steps are used in a time window and the batch size and epochs are set as 100 and 50, respectively, during every training process. Eventually, we only save the best performing model. The prediction process is conducted using the same LSTM configuration for different TGs. Each of the above constructed models is run 10 times to avoid the effects of random initialization of parameters.

For fairness, the BPNN, GRU, and ordinary RNN also have one input layer, one hidden layer with 100 neurons, and one output layer. The other parameters are tuned to obtain the best performance. For the SVM regression, the kernel is the RBF (RBF-SVR). For ARIMA, we choose the Expert Modeler in SPSS to obtain the final performance. For CNN, we use the Conv1D with one core layer and the filters, kernel size and strides are 8, 5, and 2, respectively. It is noted that our CB-LSTM model is built upon cluster analysis and one dimensional data, while comparison with the ConvLSTM model empirically will require two dimensional data. For comparison, therefore, input data will involve tap-in flows from all stations and output data will be tap-in flows from the nine stations we selected. For fairness, the ConvLSTM also has one core layer and the filters, kernel size and strides are 8, 5, and 2, respectively. The ResNet has two residual units to capture the effect of residual skip connection. It should be noted that this kind of comparison between CB-LSTM and ConvLSTM or ResNet is not sufficient because we use the tap-in flows from all stations to predict the nine stations. However, this can show that the CB-LSTM can perform better than most state-of-the-art models in this case and can be applied to real world application.

The average training time and average prediction time for the CB-LSTM model are summarized in TABLE VI. The average prediction time covers the time from inputting the original passenger flow series to outputting the prediction results for the 5 testing days. As we can observe, it only takes several seconds to accomplish the prediction process, which can meet the requirements of “real time” for STPFF.

<table>
<thead>
<tr>
<th>TG</th>
<th>1 min</th>
<th>2 min</th>
<th>3 min</th>
<th>4 min</th>
<th>5 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Training Time (s)</td>
<td>1971.82</td>
<td>995.07</td>
<td>648.59</td>
<td>512.32</td>
<td>459.66</td>
</tr>
<tr>
<td>Average Prediction Time (s)</td>
<td>4.21</td>
<td>2.58</td>
<td>2.08</td>
<td>1.81</td>
<td>1.62</td>
</tr>
</tbody>
</table>

Note: The experiments are performed using a laptop with Intel i7-8550U
Overall, the evaluation indicators for the different models under the different TGs are shown in Figure 14 and TABLE VII. The forecasting result using CB-LSTM under the TG of 5 min is presented in Figure 15. From these results, we can draw that the CB-LSTM performs better in most cases. Several conclusions are listed as follows.

First, in terms of RMSE, the CB-LSTM model has a marginally better performance than the other models and the gap becomes larger with TG. However, regardless of the TG value, the CB-LSTM has a better performance than any other model, among which, the ARIMA and RBF-SVR perform relatively more poorly in larger TG and the ResNet and ConvLSTM perform relatively more poorly in smaller TG. However, this kind of result cannot show that the CB-LSTM is more advanced than ConvLSTM or ResNet. It can only indicate that CB-LSTM can perform better because it is a kind of cluster based analysis, which is one of the contributions of our study.

Second, in terms of MAE, a similar pattern with that of RMSE is shown. Meanwhile, the performances of GRU and CB-LSTM are close to each other, which may be the reason that the interior architecture of GRU is relatively similar to CB-LSTM.

Third, in terms of \( R^2 \), the indicator gradually increases with TG, which denotes that a larger TG will cause more detailed passenger flow information being ignored, and the predictability, therefore, increases. Furthermore, the CB-LSTM can reach 0.9699 under a TG of 5 min. Even under a TG of 1 min, it can reach up to 0.9286, which shows the favorable prediction performance of CB-LSTM for STPFF in URT.

Finally, Figure 15 shows that the prediction data are favorably anastomotic with the raw data. Moreover, the cluster-based STPFF is feasible because the CB-LSTM can capture most passenger flow characteristics within one subclass.

2) METHOD 2: ORDINARY LSTM MODEL

The performance of CB-LSTM is the average prediction performance for the 9 stations. To verify the specific prediction performance for each of the 9 stations, we introduce Method 2, which involves developing specific LSTM models for the 9 stations using the same architecture with CB-LSTM as shown in Figure 15. Using the model developed in CB-LSTM, we also extract the prediction performance for each of the 9 stations to compare the performances of Method 1 and Method 2. In Method 2, we only take the TG of 5 min as an example. The results are displayed in Figure 16 and TABLE VIII. Several conclusions can be drawn as follows. As is shown, CB-LSTM also performs better than ordinary LSTM in most cases.

First, the developed cluster-based model outperforms the models developed for each of the 9 stations in most cases. This is because the number of training dataset in Method 1 is 9 times the size of that in Method 2, which is critically essential to train models.

Second, most stations achieve acceptable and satisfactory results. Some stations can even reach up to 0.984 in terms of \( R^2 \). Meanwhile, station 2 performs the worst because of the considerable passenger flow fluctuation; this can also be confirmed by observing subgraph 2 in Figure 14.

To avoid randomness and prove the advancement of CB-LSTM model than ordinary LSTM model, we select a subclass in five other main classes to conduct CB-LSTM and ordinary LSTM. Results are shown in Appendix. Similar conclusions can also be drawn.

V. CONCLUSION

This study proposed an innovative method of STPFF in URT. First, we introduced a novel two-step K-Means clustering model, which can cluster subway stations into main classes and subclasses. The subclass can capture not only the passenger flow variation trends throughout a day but also the ridership volume characteristics. Second, we described a predictability assessment model, namely similarity measurement and stationarity test. This model is able to recommend a reasonable TG to aggregate passenger flow in advance, which can ensure an acceptable prediction precision under a given similarity level and significance level. Finally, the CB-LSTM in the deep learning field is introduced to conduct STPFF in URT. The results demonstrate its feasibility and satisfactory performance. Moreover, we innovatively conduct the STPFF based on subclasses, which can increase the training dataset volume and is critically beneficial for prediction precision. The methodology is successfully applied to the URT in Beijing and verified by the favorable prediction results. The main findings are summarized as follows:

1) Conducting the STPFF based on clusters can avoid the complication of developing many specific models for each of the hundreds of subway stations on a network scale. More importantly, this method can increase the prediction precision in most cases when the dataset is limited.

2) The CB-LSTM achieves a satisfactory level of predictability for STPFF in URT. For some specific stations, the prediction performance can reach up to 0.984 in terms of \( R^2 \) under a TG of 5 min, and even under a TG of 1 min, it can reach up to 0.928. This demonstrates the favorable performance of the CB-LSTM as well as its ability to capture fluctuation characteristics of the short-term passenger flow.

3) A larger TG will result in more detailed passenger flow information being ignored, and the predictability, therefore, increases. Thus, it is important to choose a reasonable TG to aggregate passenger flow according to operation and management requirements of subway companies.
The novel two-step K-Means clustering model is more refined and focused. The passenger flow in each subclass does not only have similar variation trends but also have similar ridership volumes, which ensures the prediction precision.

Overall, the above findings provide critical insights for operators to conduct a more precise STPFF in URT. The study also demonstrates the feasibility of the prevailing deep learning for STPFF in URT on a network scale. However, there are also some limitations. For instance, we only conducted predictions in workdays. In weekends or holidays, there are more factors that should be considered. Moreover, we only considered the characteristics of the passenger flow itself. Some other factors such as the weather or events can be taken as inputs of the model to improve the prediction precision. In addition, we only considered local spatial correlations in the training of CB-LSTM network; the global spatial correlations should be explored using some latest deep learning methods. Future research can focus on overcoming these limitations.
FIGURE 14 COMPARISON OF MODEL PERFORMANCE UNDER DIFFERENT TGS
<table>
<thead>
<tr>
<th>Total testing observations</th>
<th>TG 1 min</th>
<th>TG 2 min</th>
<th>TG 3 min</th>
<th>TG 4 min</th>
<th>TG 5 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>48600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2700</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1800</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1350</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1080</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicators</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA</td>
<td>7.0105</td>
<td>4.9318</td>
<td>0.9136</td>
<td>11.8896</td>
<td>8.0080</td>
<td>0.9319</td>
<td>16.8314</td>
<td>11.0866</td>
<td>0.9370</td>
<td>21.2200</td>
<td>13.7336</td>
<td>0.9426</td>
</tr>
<tr>
<td>RBF-SVR</td>
<td>7.1184</td>
<td>5.2781</td>
<td>0.9201</td>
<td>11.8432</td>
<td>8.4261</td>
<td>0.9433</td>
<td>16.4113</td>
<td>11.5375</td>
<td>0.9512</td>
<td>21.2200</td>
<td>14.5981</td>
<td>0.9540</td>
</tr>
<tr>
<td>BPNN</td>
<td>7.1015</td>
<td>4.9422</td>
<td>0.9205</td>
<td>11.5984</td>
<td>7.7179</td>
<td>0.9457</td>
<td>15.4790</td>
<td>10.0144</td>
<td>0.9566</td>
<td>19.8204</td>
<td>12.4752</td>
<td>0.9598</td>
</tr>
<tr>
<td>CNN</td>
<td>6.9873</td>
<td>4.9877</td>
<td>0.9157</td>
<td>11.5638</td>
<td>7.9409</td>
<td>0.9434</td>
<td>16.1725</td>
<td>10.4959</td>
<td>0.9515</td>
<td>20.3677</td>
<td>12.9770</td>
<td>0.9570</td>
</tr>
<tr>
<td>RNN</td>
<td>6.9395</td>
<td>4.8858</td>
<td>0.9241</td>
<td>11.9410</td>
<td>8.5691</td>
<td>0.9424</td>
<td>16.0953</td>
<td>10.6152</td>
<td>0.9531</td>
<td>19.4983</td>
<td>12.3754</td>
<td>0.9611</td>
</tr>
<tr>
<td>GRU</td>
<td>6.7497</td>
<td><strong>4.7284</strong></td>
<td>0.9282</td>
<td>11.5759</td>
<td>7.4949</td>
<td>0.9459</td>
<td>15.4510</td>
<td>9.7863</td>
<td>0.9568</td>
<td>19.1577</td>
<td><strong>12.0154</strong></td>
<td>0.9625</td>
</tr>
<tr>
<td>ConvLSTM</td>
<td>7.9375</td>
<td>5.5077</td>
<td>0.8940</td>
<td><strong>10.9239</strong></td>
<td><strong>7.4572</strong></td>
<td>0.94681</td>
<td>14.9725</td>
<td>10.2721</td>
<td>0.9551</td>
<td>18.9454</td>
<td>12.6597</td>
<td>0.9587</td>
</tr>
<tr>
<td>ResNet</td>
<td>7.8518</td>
<td>5.3938</td>
<td>0.8951</td>
<td>12.1499</td>
<td>7.9817</td>
<td>0.9354</td>
<td>16.7072</td>
<td>12.1305</td>
<td>0.9440</td>
<td>19.1765</td>
<td>13.0607</td>
<td>0.9585</td>
</tr>
<tr>
<td>CB-LSTM</td>
<td><strong>6.7298</strong></td>
<td>4.7515</td>
<td><strong>0.9286</strong></td>
<td>11.2285</td>
<td>7.4703</td>
<td><strong>0.9491</strong></td>
<td>15.2035</td>
<td><strong>9.7834</strong></td>
<td><strong>0.9581</strong></td>
<td><strong>18.6825</strong></td>
<td>12.0376</td>
<td><strong>0.9643</strong></td>
</tr>
</tbody>
</table>

Note: The total testing observations are the overall number for the 9 selected stations in the 5 test workdays during the testing stage. For fairness, all above neural networks have one core layer.
FIGURE 15 PREDICTION RESULTS UNDER THE TG OF 5 MIN FOR THE 9 STATIONS
FIGURE 16 COMPARISON OF PERFORMANCE BETWEEN DIFFERENT MODELING METHODS FOR THE 9 STATIONS
<table>
<thead>
<tr>
<th></th>
<th>Station 1</th>
<th>Station 2</th>
<th>Station 3</th>
<th>Station 4</th>
<th>Station 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicators</td>
<td>RMSE</td>
<td>MAE</td>
<td>R²</td>
<td>RMSE</td>
<td>MAE</td>
</tr>
<tr>
<td>CB-LSTM</td>
<td>23.2427</td>
<td>15.5528</td>
<td>0.9743</td>
<td>31.5218</td>
<td>16.8641</td>
</tr>
<tr>
<td>LSTM</td>
<td>24.5112</td>
<td>16.3200</td>
<td>0.9714</td>
<td>34.7310</td>
<td>18.6459</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Station 6</th>
<th>Station 7</th>
<th>Station 8</th>
<th>Station 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicators</td>
<td>RMSE</td>
<td>MAE</td>
<td>R²</td>
<td>RMSE</td>
</tr>
<tr>
<td>CB-LSTM</td>
<td>14.8292</td>
<td>10.7181</td>
<td>0.9843</td>
<td>25.4487</td>
</tr>
<tr>
<td>LSTM</td>
<td>14.7408</td>
<td>11.0055</td>
<td>0.9846</td>
<td>26.0438</td>
</tr>
</tbody>
</table>

Notes: There are 1080 observations for each of the 9 stations in the 5 test workdays during the testing stage.
REFERENCES

328.


