Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6625435, 11 pages
https://doi.org/10.1155/2020/6625435

Research Article

Forecast and Early Warning of Regional Bus Passenger Flow Based on Machine Learning

Wusheng Liu,[1] Qian Tan ,[2] and Wei Wu[3]

[1]*Engineering Research Center of Catastrophic Prophylaxis and Treatment of Road & Traffic Safety of Ministry of Education, Changsha University of Science & Technology, Changsha 410114, China*
[2]*School of Traffic & Transportation Engineering, Central South University, Changsha 410075, China*
[3]*School of Traffic and Transportation Engineering, Changsha University of Science & Technology, Changsha 410114, China*

Correspondence should be addressed to Qian Tan; tq.helen@csu.edu.cn

Received 3 November 2020; Revised 25 November 2020; Accepted 4 December 2020; Published 15 December 2020

Academic Editor: Hussein Abulkasim

This paper mainly forecasts the short-term passenger flow of regional bus stations based on the integrated circuit (IC) card data of bus stations and puts forward an early warning model for regional bus passenger flow. Firstly, the bus stations were aggregated into virtual regional bus stations. Then, the short-term passenger flow of regional bus stations was predicted by the machine learning (ML) method of support vector machine (SVM). On this basis, the early warning model for regional bus passenger flow was developed through the capacity analysis of regional bus stations. The results show that the prediction accuracy of short-term passenger flow could be improved by replacing actual bus stations with virtual regional bus stations because the passenger flow of regional bus stations is more stable than that of a single bus station. The accurate prediction and early warning of regional bus passenger flow enable urban bus dispatchers to maintain effective control of urban public transport, especially during special and large-scale activities.

1. Introduction

Urban public transport is a traffic mode to alleviate traffic congestion and make efficient use of road resources. To realize intelligent dispatching of buses, it is important for decision makers to know well the change law of bus passenger flow and accurately predict the passenger flow in the short term. Burst passenger flows cause a huge amount of traffic demand in a short time, which may bring great pressure to public security. Early warning of passenger flow is required to have some preparation by the administration. However, the urban bus stops lack real-time early warning tools with good accuracy nowadays.

The short-term bus passenger flow is affected by various random, complex, and space-varying factors. In areas with large passenger flow, it is difficult to forecast the short-term change of bus passenger flow. Currently, short-term traffic is generally predicted by traditional statistical methods, novel intelligent methods, and hybrid methods.

Originating from time series analysis in the 1980s, the traditional statistical methods are built on the data collected by manual surveys. Over the years, these methods have been evolving and intellectualized in the context of traffic flow prediction. In the forecast of short-term passenger flow at a bus station, the entry and exit volumes mainly come from the card swiping records of the automatic fare collection (AFC) system. Taking such continuous data as a time series, many models have been designed for traffic flow prediction based on statistical principles, including autoregressive (AR) model, moving average (MA) model, autoregressive integrated moving average (ARIMA) model, and seasonal ARIMA (SARIMA) model [1–4]. In addition, many have introduced k-nearest neighbors (k-NN) [5], nonparametric regression [6, 7], and Kalman filter [8] to predict short-term traffic flow.

Since the dawn of the big data era, the amount of public transport data has exploded with the application of novel techniques such as the integrated circuit (IC) card. The traditional statistical methods could no longer adapt to the

complex environment and changing passenger flow at bus stations. With the rapid development of computational intelligence (AI) and data mining, data-driven intelligent methods have become popular in the prediction of short-term passenger flow [9]. The novel intelligent methods include long short-term memory (LSTM) [10–12], neural network (NN) [13–20], random forest (RF) [21, 22], support vector machine (SVM) [23, 24], fusion convolutional LSTM (FCL Net) [25], agent-based model (ABM) [26], and Bayesian network [27].

Recent years have witnessed the emergence of many hybrid forecast methods, most of which are combinations of novel intelligent methods such as NN and ML. For instance, Ke et al. [25] fused the FCL Net into a new deep learning (DL) method for the projection of short-term passenger demand. Xiao et al. [28] developed a new hybrid forecast strategy for air transport demand, which couples singular spectrum analysis (SSA), adaptive network fuzzy inference system (ANFIS), and optimized particle swarm optimization (OPSO). Sun et al. [29] combined wavelet transform (WT) with SVM into a hybrid prediction model for passenger flow; the model decomposes, predicts, and reconstructs the data on passenger flow in three stages and inherits the merits of both WT and SVM. Tan et al. [30] put forward a total traffic flow prediction method based on NN, MA, exponential smoothing (ES), and autoregressive MA (ARIMA).

In addition, Hinton and Salakhutdinov [31] applied DL to solve short-term prediction. Hu et al. [32] optimized the parameters of support vector regression (SVR) through particle swarm optimization (PSO), introduced historical momentum to reduce the impact of noise in traffic flow data, and then established a PSO-SVR model for the forecast of short-term traffic flow. Doğan [33] designed the periodic outstaring and prediction (PCP) algorithm, adopted the algorithm to improve the training set of artificial neural network (ANN), and proved that the improved ANN could predict short-term traffic flow based on selected clusters. Bagloee et al. [34] proposed a hybrid ML-based method to solve the bilevel optimization problem. Han et al. [35] derived a hybrid, optimized LSTM from Nesterov accelerated adaptive moment estimation (Nadam) and stochastic gradient descent (SGD).

To sum up, the traditional statistical methods mostly treat the current traffic state as a linear combination of the previous states and errors. On the upside, the traffic flow can be predicted simply by mathematical statistics, with relaxed data requirements. On the downside, the traditional methods fail to reflect the randomness and nonlinearity of traffic flow, consume too much manpower and financial resources in data acquisition, and have a low accuracy in the prediction of traffic flows, especially that with sudden changes. The novel intelligent methods are better than the traditional methods in data fitting and prediction accuracy, but the computing process is much more complex. The hybrid methods generally consider the features of actual traffic flow. The prediction accuracy of these methods varies with the coupled algorithms. Compared with the intelligent methods, the hybrid methods are highly complicated. The advantages, disadvantages, and applicability of common short-term prediction methods are shown in Table 1.

Despite the abundant results on short-term forecast, only a few scholars have explored the prediction or early warning of the passenger flow at urban bus stations. Gong et al. [36] proposed an ARIMA model and a Kalman filter to predict the number of passengers waiting at a bus station. Han et al. [35] created a hybrid and optimized LSTM to project the bus passenger flow. Van Oort et al. [37] converted the IC card data to the number of passengers per line, constructed an origin-destination (OD) matrix between stations, and assigned the matrix to the network to reproduce the measured passenger flow. Kumar et al. [38] developed a bus travel-time prediction method that considers both spatial and temporal variations in travel time. Wu et al. [39] built a convolutional LSTM (ConvLSTM) model with a self-attention mechanism, which accurately predicts the travel time on each segment of a trip and the waiting time at each station. Considering the small size, strong time-variation, and extraction difficulty of short-term passenger flow at bus stations, some scholars have taken account of connected and autonomous vehicles to alleviate the variability of travel time [40, 41]. Albeit these efforts, it is still difficult to make realistic forecast of the short-term passenger flow at bus stations. Wang et al. [42] designed a new framework to solve the problem of sudden passenger flow early warning. Pereira et al. [43] detected overcrowding with a threshold-based method and defined the point whose arrivals exceed the 90% percentile as overcrowding point. Bai et al. [44] monitor passenger flow to display the distribution trend of real-time passenger flow with GIS technology. To sum up, the current research mainly focuses on the early warning of traffic flow, and there is little research on the monitoring and early warning methods of passenger flow at bus stops.

We firstly analyze the features of passenger flow at bus stations and propose a novel concept called regional bus station, which is aggregated from actual bus stations in this paper. Then, the SVM was introduced to predict the short-term passenger flow at regional bus stations. The results show that the prediction accuracy of short-term passenger flow could be improved by replacing actual bus stations with virtual regional bus stations. On this basis, we designed an early warning model for regional bus passenger flow, which monitors the passenger flow in important areas during the period of special activities (e.g., large events) and takes control measures in advance to ensure the smooth progress of these activities.

The remainder of this paper is organized as follows: Section 2 puts forward the concept of regional bus stations by analyzing the IC card data of Shenzhen from November 7 to December 4, 2016, and summarizes the features of regional bus passenger flow; Section 3 selects and trains the prediction variables and employs the SVM to make short-term prediction of the passenger flow at regional bus stations; Section 4 analyzes the accuracy of prediction results; Section 5 carries out the capacity analysis and derives the early warning model for regional bus passenger flow; Section 6 puts forward the conclusions.

Table 1: Advantages, disadvantages, and applicability of common short-term prediction methods.

Prediction method	Advantages	Disadvantages	Applicable conditions
ARIMA model	The model is simple and has the ability to correct local data trend	It is difficult to fit nonlinear problems	Medium- and short-term forecast
Exponential smoothing method	Flexible and simple operation	Less variables are considered and the accuracy of smoothing number is low	Medium- and short-term forecast
Trend extrapolation	The operation is simple and the fitting effect is good	It is difficult to guarantee the accuracy due to less variables	Short-term forecast
Multivariate regressive method	Multiple factors can be considered	Large amount of calculation and high requirement for data	Medium- and short-term forecast
SVM	The model is simple and the results need not be modified	It is difficult to consider the comprehensiveness of indicators	Medium- and short-term prediction of small samples
RF	It can process high-dimensional data without feature selection	The reliability of the attribute weight on the data is not high	Medium- and short-term prediction of small samples
LSTM	Strong nonlinear fitting ability	A large amount of data is needed for network training	Nonlinear prediction

2. Bus Station Distribution and IC Card Data Processing

2.1. Bus Station Data and IC Card Data. The IC card used in this paper is Shenzhen Tong. It is a kind of stored value card for consumption by Shenzhen bus and Shenzhen Metro, which is manufactured under the supervision of Shenzhen Transport Bureau and issued by Shenzhen public transport settlement management center. The bus data only displays the valid information such as the user card number, card swiping time, and bus license plate number. It is shown in Table 2.

We extracted the GPS data and the card data of passengers and mapped the bus stops according to the time-space relationship.

(1) The GPS records of each line for 20 days are selected, and the GPS records are sorted according to the license plate number and the return time of GPS to get the track of the bus.

(2) Select the track point closest to the card swiping time as the boarding point, calculate the distance between the boarding point and each bus stop on the bus operation line, and select the nearest bus stop with the actual distance less than 100 as the actual bus stop, as shown in Figure 1.

The GPS data of 28 days from November 7 to December 4, 2016, are used in this paper, which contain about 541,115,294 pieces of time and location information of 10,314 vehicles. Therefore, we can study the characteristics of working day and non working bus passenger flow.

The passenger flows of bus stations in Shenzhen were acquired from the IC card data during the bus operation time (6:00–22:00) from November 7 to December 4, 2016. Figure 2 shows the distribution of the number of card swipes on buses in Shenzhen on a typical day.

To pinpoint the bus line number of each card swipe, the geographical positioning system (GPS) data were matched with the location data of bus stations, as shown in Figure 3. Through the matching, the bus arrival time was obtained for each bus station. Then, the boarding station of each card user was identified by acquiring the boarding time and bus line number from his/her IC card and matching the bus line number with the arrival time obtained in the previous step.

According to the time-varying distribution of bus trips in many days, it is found that the bus trips have obvious peak characteristics, and the peak hours are concentrated at 7:00 a.m. and 18:00 p.m. 7:00–9:00 and 17:00–19:00 are selected as the morning and evening peak hours of bus travel, as shown in Figure 4.

2.2. Regional Bus Station. To facilitate the passenger flow prediction of all bus stations in a region of Shenzhen (E: 113.76°–114.62°; N: 22.45°–22.87°), the road network in the region was meshed into 3,612 1 km × 1 km grids. 933 grids were found to have bus stations. In this way, the 53,914 bus stations in the region were aggregated into 993 regional bus stations (the black spots in the lower part of Figure 5).

3. Short-Term Passenger Flow Prediction at Regional Bus Stations

3.1. Principles. The short-term regional passenger flow was predicted in the following steps:

Step 1. Encode all the grids, and count the number of stations in each grid.

Step 2. Taking 1 h as the time window, count the number of passengers boarding buses at each station from 6:00 to 23:00 in the four days from December 1 to December 4, 2016.

Step 3. Allocate the data in four time windows (19:00-20:00, 18:00-19:00, 17:00-18:00, and 16:00-17:00) of the first 3 days (December 1–3, 2016) to the training set and the data in the same time windows of the last day (December 4, 2016) as the test set. Train the data by the ML-based forecast procedure.

Table 2: Data format of Shenzhen Tong (bus).

User number	Card bill number	Time	License plate number
4742502	666759180	2016-08-06T11:24:31.000Z	CB372
4742503	667174358	2016-08-06T11:39:09.000Z	CB372
4742504	667290097	2016-08-06T10:41:19.000Z	CB372

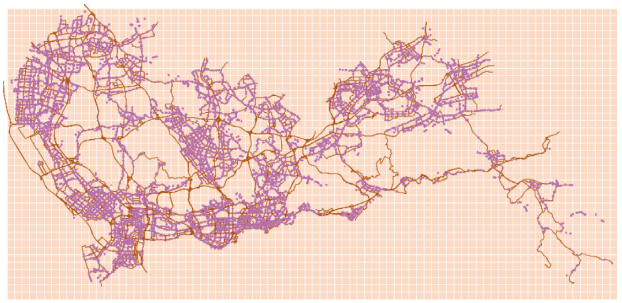

Figure 1: Location information of bus network and stations in Shenzhen.

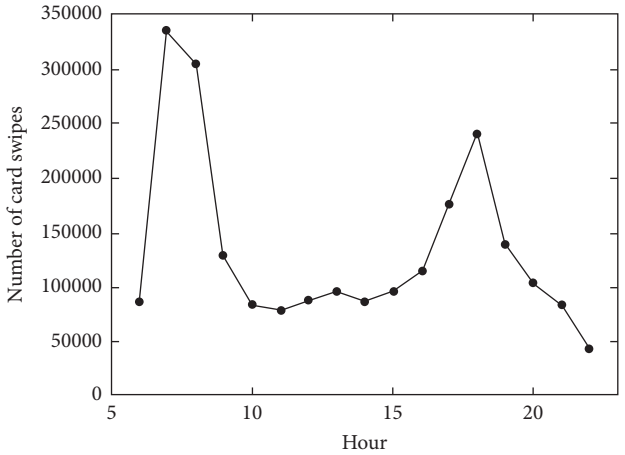

Figure 2: The distribution of the number of card swipes on buses in Shenzhen on December 4.

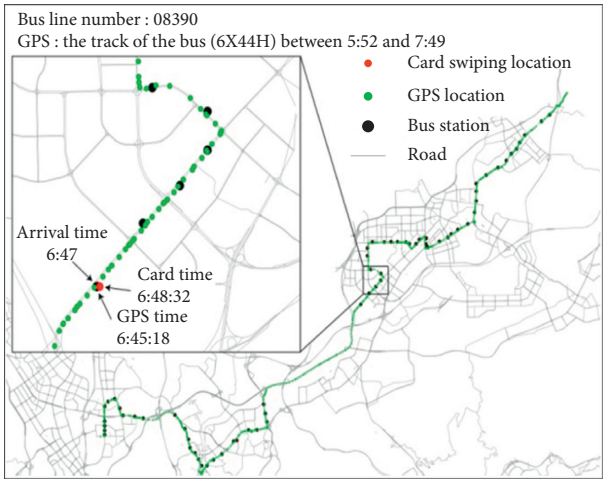

Figure 3: The acquisition of IC card information.

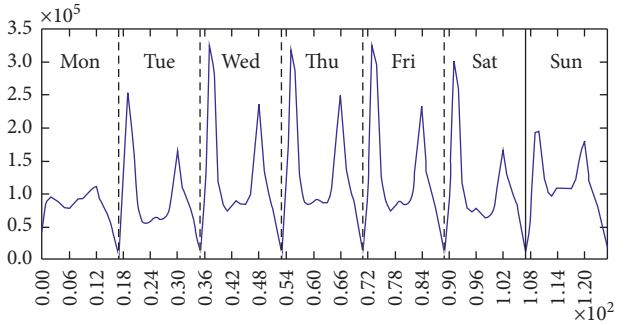

Figure 4: Time-varying distribution of bus travel.

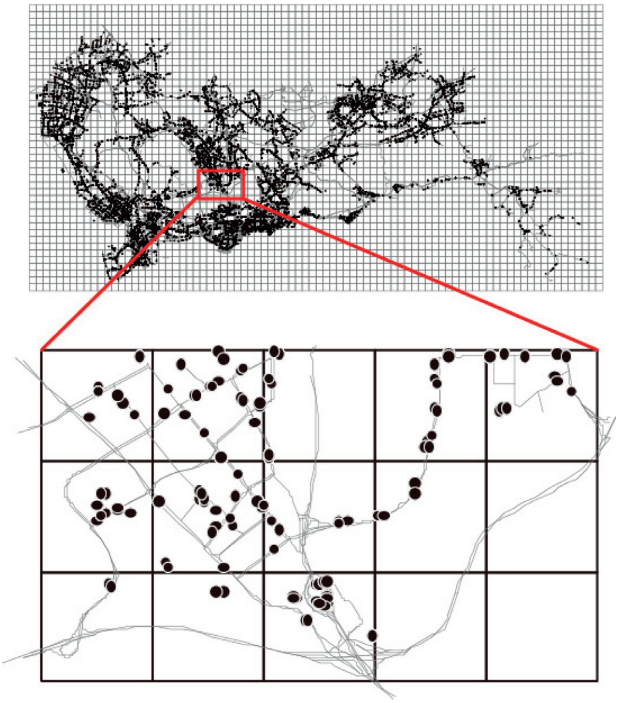

Figure 5: The distribution of bus stations in the region.

Step 4. Select the features of short-term regional bus passenger flow through ML, producing a set of valid features R.

Step 5. Select the top N features that influence the target period the most from the feature set R, and take them as the input of the regression model for SVM-based prediction.

3.2. Feature Training. Let $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, $x \in R^d$, $y \in R$ be the collected data, where (x_l, y_l) is

the training data of the lst day, d is feature dimension, x_l is the set of passengers boarding on the lst day under feature d, and y_l is the test data of the lst day.

Details of alternative features are shown in Table 3. The alternative feature for a target grid is the 1 km × 1 km grids around that grid, and the alternative feature set for the 993 grids in the region is whether the target grids are surrounded by 1 km × 1 km grids. If there are i grids in the set, then the number of boarding passengers should be counted in the previous j time windows of each grid in the set. Therefore, the feature dimension can be expressed as $d = i \times j$.

The number of alternative features varies from grid to grid, that is, the d value is a variable in $X^{(l\times d)} = (x_1, x_2, \ldots, x_l)^T$. Here, the value of j is set to 3. The details of alternative features are presented in Table 3.

Let T be the target time window. Then, the $N = 3$ most important features were selected through recursive feature elimination (Algorithm 1).

3.3. SVM-Based Regression Prediction. The regression prediction takes the N features that influence the target period the most as the inputs. Thus, the dimension of sample x_i is equal to N.

Let $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}, x \in R^N, y \in R$ be the training set. Then, the training samples were mapped into a higher dimension by function Φ, turning nonlinear regression inputs into high-dimensional inputs for linear regression.

The nonlinear mapping function can be expressed as $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. Then, the SVM-based nonlinear regression can be defined as

$$\begin{aligned} \min_{\alpha^* \in R^{2l}} \quad & \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)\left(\alpha_j^* - \alpha_j\right)k\left(x_i \cdot x_j\right) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) \\ & -y_i\sum_{i=1}^{l}(\alpha_i - \alpha_i^*), \\ \text{s.t} \quad & \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \alpha_i, \alpha_i^* \geq 0 \\ & i = 1, 2, \ldots, l, \alpha_i^* \leq C, \end{aligned} \tag{1}$$

where C is the penalty parameter and ε is the error term.

The optimum solution can be calculated as $\overline{\alpha}^{(*)} = (\overline{\alpha}_1, \overline{\alpha}_1^*, \ldots, \overline{\alpha}_l, \overline{\alpha}_l^*)$. The positive subsector $\overline{\alpha}_j > 0$ of $\overline{\alpha}$ or $\overline{\alpha}_j^* > 0$ of $\overline{\alpha}^*$ was chosen. Then, the $\overline{b}$ can be calculated by

$$\overline{b} = y_j - \sum_{i=1}^{l}(\overline{\alpha}_i^* - \overline{\alpha}_i)K\left(x_i, x_j\right) - \varepsilon. \tag{2}$$

Then, the decision function can be described as

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k\left(x_i \cdot x\right) + \overline{b}. \tag{3}$$

Table 3: The details of alternative features.

Grid number	$t-3$	$t-2$	$t-1$	T
1	Pop_1_$t-3$	Pop_1_$t-3$	Pop_1_t_1	Grid
2	Pop_2_$t-3$	Pop_2_$t-3$	Pop_2_$t-1$	
...	...	...	...	
i	Pop_i_$t-3$	Pop_i_$t-3$	Pop_i_$t-1$	

For target grids, the number of passengers boarding in time window t can be forecasted based on the same N most important features on the given day. The radial basis function (RBF) was chosen as the kernel function: $k(x_i \cdot x) = \exp(-(\|x - x_i\|^2/2\sigma^2))$.

4. Prediction Results and Model Test

4.1. Prediction Results. Two time windows were randomly selected, namely, 11:00-12:00 and 19:00-20:00 on December 4, 2016, and the number of boarding passengers in the 993 grids was counted. The results predicted by our method are compared with the actual data in Figures 6 and 7.

4.2. Verification of Time Window. The target time window for prediction was set as 1 h. To verify the correctness of the time window, the real number of passengers boarding in 19:00-20:00 on December 4, 2016, was selected as the current period and compared with the real number of the previous period (18:00-19:00), the real number of the subsequent period (20:00-21:00), and the predicted value of the current period (19:00-20:00) (Figure 7).

As shown in Figures 8 and 9, the points were scattered, and the degree of linear fitting was poor, indicating that the number of passengers at each station in the grid fluctuates significantly within the time window of 1 h. Hence, it is necessary to set 1 h as the time window.

Moreover, the regression result of Figure 8 was $\beta_1 > 1$ (β_1 is the slope of the linear fitting line), that is, the result is greater than the actual boarding number in 19:00–20:00 of all grids. The regression result of Figure 9 is $\beta_1 < 1$, indicating that the actual boarding number in 19:00–20:00 of all grids is more than that in the following hour. Therefore, the data in the current time window cannot be replaced by the data of the previous or subsequent time window. In addition, in Figures 6 and 7, the slope β of linear fitting approaches to 1, indicating that it is better to use 1 hour as the time window. This further confirms the rationality of the 1 h time window.

4.3. Verification of Grid Size. We cross-grained the road network into 3,612 square grids with an area of 1 × 1 sq. km. Within the partitioned 3,612 square grids, 993 had at least one bus station inside. We cross-grained the road network with an area of 0.5 × 0.5 sq, and we found that 10% of the area had no bus stops. Moreover, the division of regional bus stations should be combined with the average length of urban road sections, and the area enclosed by each road section should be taken as the regional bus station as far as possible, and the average length of the road section in this paper is 0.78 km. In addition, we forecast the passenger flow

Input: training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}, x \in R^d, y \in R$ and linear regression model
(1) Initialization: original feature set $S = \{1, 2, \ldots, D\}$, and feature sort set $R = \varnothing$
(2) While $S \neq \varnothing$ do
(3) Obtain the training samples of candidate feature set
(4) Obtain the weight of each feature by linear regression model, i.e., the coefficient of linear regression model $\omega_k, k = 1, 2, \ldots, |S|$
(5) Find out the features of the minimum score of sorting criteria: $p = \arg\min_k \omega_k$
(6) Update feature set $R = \{p\} \cup R$
(7) Exclude feature $S = S/p$ in S
(8) End while
Output: feature sort set R

Algorithm 1: Recursive feature elimination.

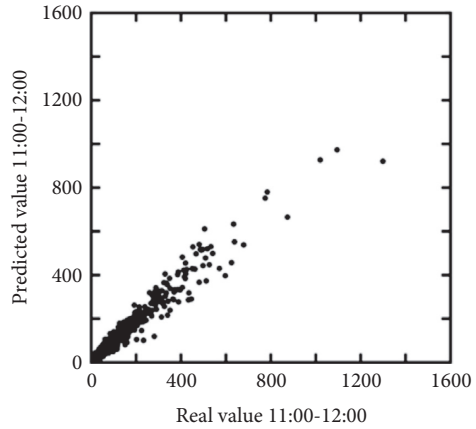

Figure 6: The predicted value vs. real value of the number of regional bus passenger flow (11:00-12:00).

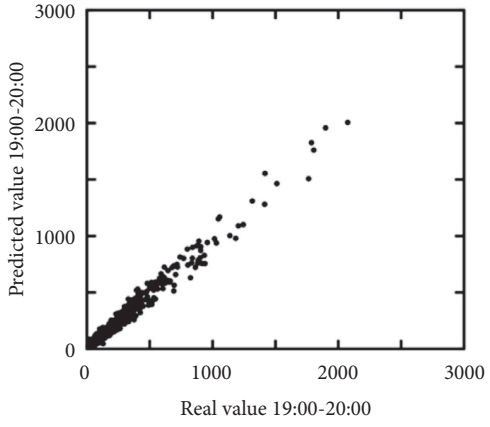

Figure 7: The predicted value vs. real value of the number of regional bus passenger flow (19:00-20:00).

of regional bus stations with different flows in different regions. We find that, with the increase of passenger flow of regional bus stations, the prediction accuracy increases, as shown in Figure 10. Therefore, it is appropriate to select 1×1 sq as the size of regional bus stations.

4.4. Model Test. It can be directly inferred from Figure 6 that the actual value and the predicted value follow a linear

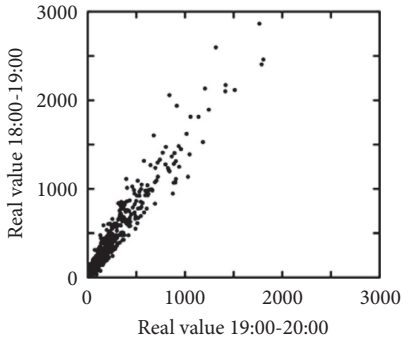

Figure 8: The real value vs. real value of the boarding number in different time windows (a).

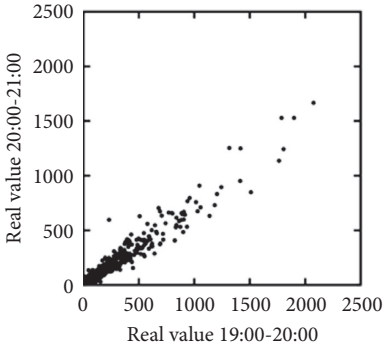

Figure 9: The real value vs. real value of the boarding number in different time windows (b).

relationship. Thus, unary linear regression was adopted for data fitting:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \sigma^2, \tag{4}$$

where β_0 and β_1 both obey the normal distribution. If β_0 approaches to 0 and $\beta_1$approaches to 1, then the model has a high degree of regression. The dataset to be regressed was defined as $(x_i, y_i)(i = 1, 2, \ldots, n, n = 993)$.

Next, the linear regression problem was solved by the least squares (LS) method. The results of time window 11:00-12:00 were $\beta_0 = 4.20$ with its confidence being (2.42, 5.94) on the level of 95%, $\beta_1 = 0.90$ with its confidence being (0.885,

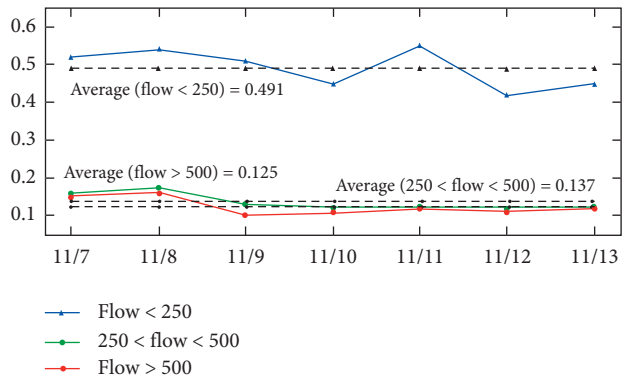

Figure 10: Prediction accuracy of station flow in different grids.

0.907) on the level of 95%, and coefficient of correlation $R^2 = 0.96$.

The predicted results were measured by three metrics: the mean absolute percentage error (MAPE), mean absolute error (MAE), variance of absolute percentage error (VAPE), and root mean square error (RMSE):

$$\begin{aligned} \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \times 100\%, \\ \text{MAE} &= \frac{1}{n}\sum_{n=1}^{n}\left|\widehat{y}_i - y_i\right|, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_i - y_i\right)^2}, \\ \text{VAPE} &= \text{Var}\left(\frac{\left|y_i - \widehat{y}_i\right|}{y_i}\right) \times 100\%. \end{aligned} \tag{5}$$

For the predicted results on the time window 11:00-12:00, the MAPE, MAE, VAPE, and RMSE were 0.18, 35.16, 0.12, and 65.42, respectively.

The results of time window 19:00-20:00 is that $\beta_0 = 2.23$ with its confidence being (−0.08, 4.55) on the level of 95%, β_1 =0.96 with its confidence being (0.95, 0.97) on the level of 95%, and coefficient of correlation $R^2 = 0.98$. The MAPE, MAE, and RMSE of these results were 0.16, 26.29, and 54.29, respectively.

The R^2 values of the two time windows demonstrate the significant correlation between variables X and Y. In both time windows, the β_0 value approached to 0, and the β_1 values approached 1. The error metrics of the results on the two time windows were both satisfactory. Therefore, the proposed ML method is favorable for predicting bus passenger flow in the short term.

4.5. Comparative Analysis of Prediction Accuracy between Single Bus Stop and Regional Bus Stop. In order to analyze the prediction accuracy of single bus stop and regional bus stop, we select the window of the world regional station, in which there are three bus stops: world window stop, Baishizhou stop, and Meilu Jinyuan stop. Bus passenger flow of November 28 is predicted, and the prediction results are shown in Figures 11 and 12.

The accuracy of regional bus station prediction results is significantly higher than that of single stops, especially for stops with relatively little passenger flow, the prediction accuracy is obviously low, and the prediction results have no practical application value.

5. Early Warning Model for Regional Bus Passenger Flow

5.1. Model Construction. The passenger capacity of a regional bus station can be calculated by

$$W = \max\left(C, P_j\right), \tag{6}$$

where C is the capacity of the regional bus station and P_j is the residual capacity of the regional bus line.

According to the *Traffic Engineering Manual*, the capacity C_i of station i can be computed by

$$C = \sum_{i=1}^{m} C_i = \sum_{i=1}^{m} \frac{3600 \times (g/c)_i \times R \times N_i}{\left[(g/c)_i \times D + t_c\right]}, \tag{7}$$

where $(g/c)_i$ is the green light ratio of the intersection in front of station i; R is the adjustment coefficient reflecting the degree of impact from bus arrival time and boarding time on station capacity (the empirical value is 0.833); D is the average boarding/alighting time (the value is generally 20–50 s); t_c is the average clearance time of bus station, i.e., the time required for the former bus to depart from and the current bus to arrive at the same position at the station (the value is generally 9–20 s); m is the total number of regional bus stations; N_i is the effective berth on the ith station.

The residual capacity P_j of regional bus line j can be calculated by

$$P_j = \sum_{j=1}^{n}\left[\frac{\left(S_j - Q_j\right) \times 60 \times H_j}{f_j}\right], \tag{8}$$

where S_j is the load factor of line j at acceptable service level (%), i.e., 80% of the rated passenger capacity; f_j is the departure interval of line j (min); H_j is the capacity of single bus online j (person) (the value is generally 60 persons); Q_j is the actual capacity of line j (%).

The early warning coefficient K can be calculated by

$$K = \frac{f(x)}{W \times \alpha}, \tag{9}$$

where $f(x)$ is the predicted short-term passenger flow of a regional bus stop and α is the proportion of IC card swiping passengers in the total number of bus passengers in the region.

5.2. Value of Early Warning Coefficient. Early warning generally refers to the prewarning of potential dangerous situation information. The early warning coefficient is used to judge what kind of situation needs to be issued and what

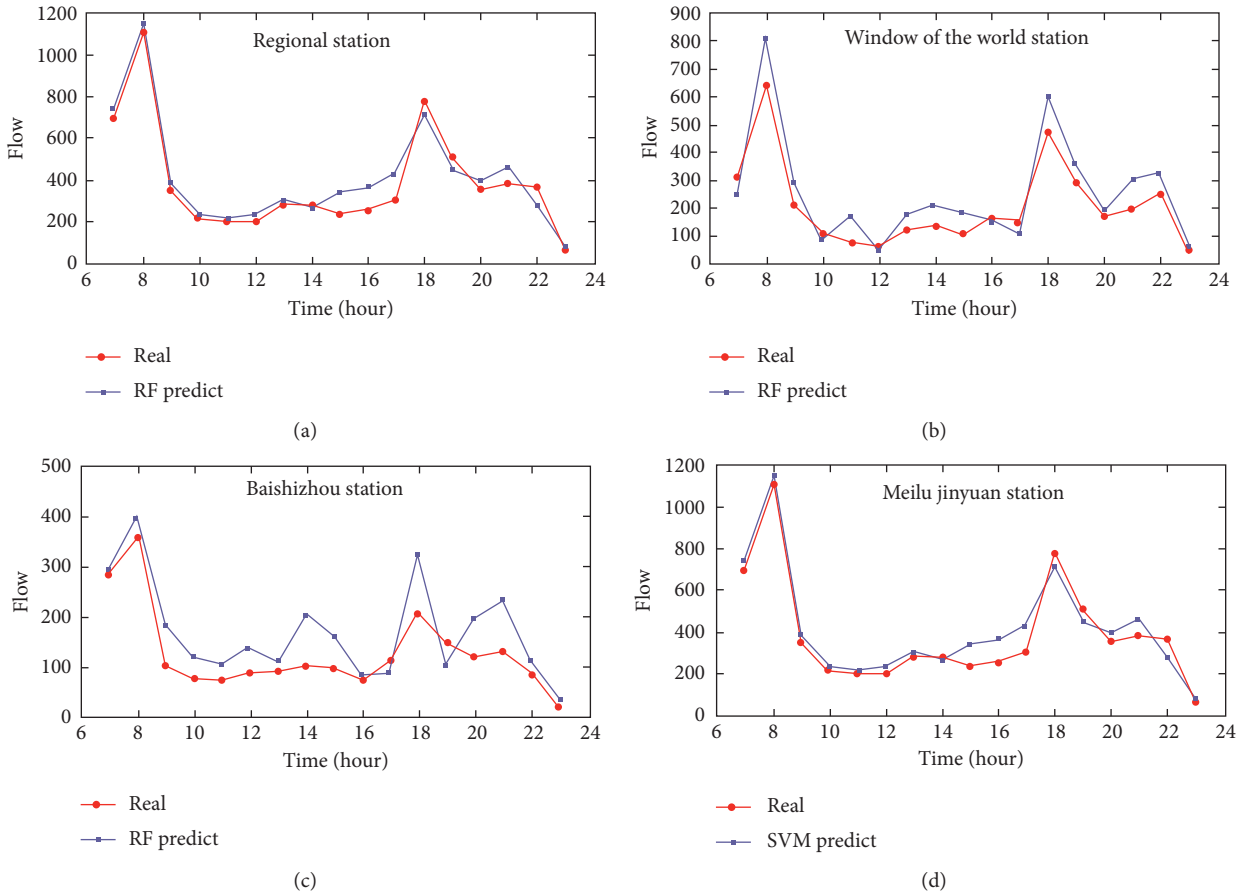

Figure 11: Prediction result between single bus stop and regional bus stop.

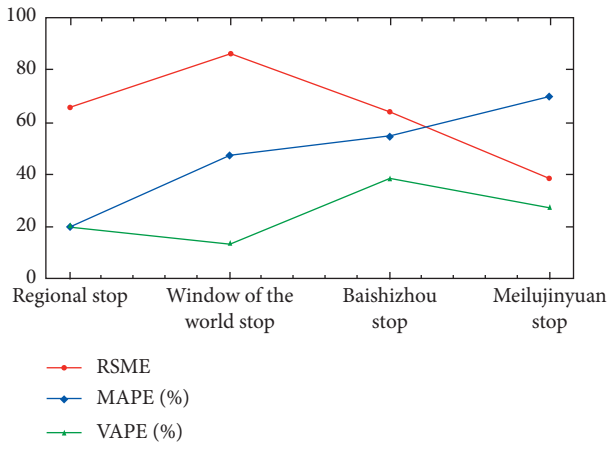

Figure 12: MAPE between single bus stop and regional bus stop.

degree of alarm is issued. In this paper, the early warning method adopts the method of combining the early warning index and the warning limit interval. If the value of the early warning index is in the corresponding warning limit interval, it corresponds to the alarm of this degree. Combined with the setting method of traffic flow warning interval [45], considering the level of public transport service, that is, considering the passenger comfort, reliability, and safety of operation service [46], the early warning interval of regional bus stops is divided into four levels. The first-level warning interval corresponds to the first level of service. The overall full-load rate of regional public transport vehicles is extremely low, and the passengers' comfort is very high, and the safety factor is very high. The second-level warning interval corresponds to the second-level service level. The overall full-load rate of regional public transport vehicles is not high, the waiting time of passengers is small, the congestion is low, and the safety factor is relatively high. The third level of warning interval corresponds to the third level of service, and the public transport vehicles are at the edge of full load. At this time, the safety factor is low, and the bus is crowded, which represents the general safety level; the fourth-level of warning interval corresponds to the fourth level of service, the safety factor is very low, the departure frequency cannot meet the needs of passengers, and the bus congestion is high, which represents the danger. The warning coefficients of the four levels are shown in Table 4.

If K falls in (0, 0.75), the region is safe; if K falls in (0.75, 0.9), the region needs to be monitored; if K surpasses 0.9, the region must be alarmed.

5.3. Case Analysis. The case analysis targets the grid at the junction of Shennan Avenue and Qiaocheng Road

Table 4: Early warning level of regional bus stops.

Early warning level	The first level	The second level	The third level	The fourth level
Warning interval	0–0.25	0.25–0.50	0.50–0.75	>0.75
Explanation	Safe	Relatively safe	Relatively danger	Danger

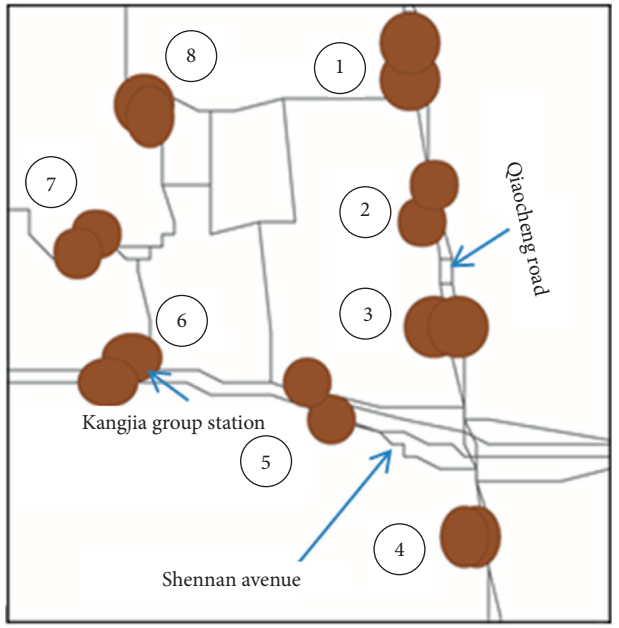

Figure 13: The target grid.

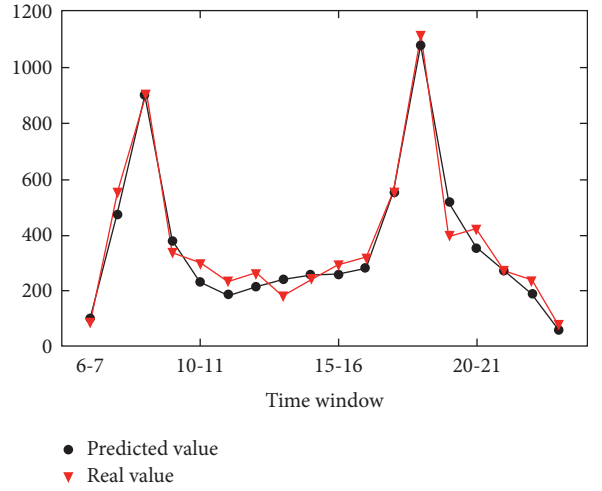

Figure 14: The predicted values vs. real values of the daily number of boarding passengers of 20 days (6:00–23:00).

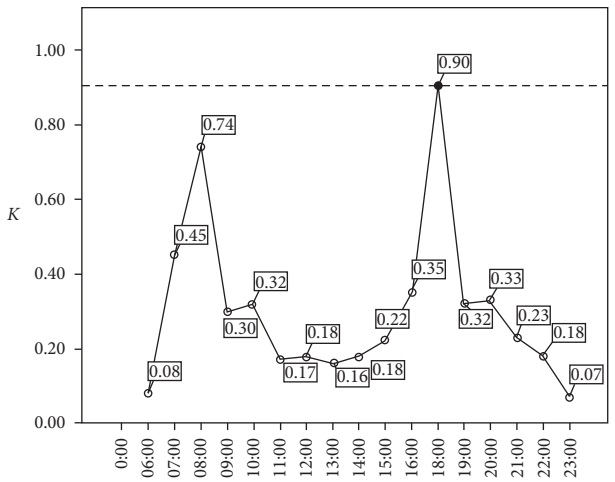

Figure 15: The early warning of passenger flow in the target grid.

(Figure 13). Shennan Avenue is one of the busiest roads in Shenzhen. Involving 16 bus stations and many bus lines, the target grid is highly representative of the public transport in Shenzhen. As shown in Figure 13, there are 35 bus lines at Kongjia Group Station on Shennan Avenue alone.

The proposed SVM-based algorithm was adopted to train and learn the historical boarding number of stations in the target grid and predict the daily number of boarding passengers of 20 days (6:00–23:00). Then, the predicted results were compared with the real values in the grid (Figure 14).

It can be seen from Figure 15 that our algorithm achieved a good prediction effect, as the predicted results followed the same trend as the real values. The MAE and MAPE were small, and the RMSE was smaller than 10%.

Through capacity analysis, the capacity of regional bus stations in 6:00–24:00 was 4,113 person times/h, while the residual capacity of bus lines was 3,840 in 6:00-7:00, 5,760 in 7:00–9:00, 4,032 in 9:00–17:00, 5,760 in 17:00–19:00, and 4,860 in 19:00–24:00. The regional bus capacity and K value in each period are presented in Figure 15.

According to the K values in Figure 15, the early warning coefficient in the target grid was 0.9 in 18:00-19:00. Hence, in this period, the bus capacity in the region is saturated and needs to be monitored. To reduce the K value, it is necessary to increase the regional bus capacity by stepping up departure frequency, improving the capacity of some stations, and expanding the effective parking spaces at stations.

6. Conclusions

This paper analyzes the IC card swiping data of all buses in Shenzhen during November 7–4, 2016 (6:00–22:00), and introduces the ML method of SVM to predict the short-term passenger flow of urban bus stations. The main conclusions are as follows:

(1) The training samples of candidate feature sets and the weight of each feature were through linear regression. The N features with the greatest impact on the target period were selected as the input of the regression model. The SVM-based regression prediction was adopted to predict the bus passenger flow

in the target time window. The model achieved high prediction accuracy when the time window is 1 h. The MAPE, MAE, and RMSE of these results were 0.16, 26.29, and 54.29, respectively.

(2) The capacity and early warning coefficient K of regional bus passenger was analyzed in detail. According to the K values, the capacity of some bus stations and bus lines in the target region should be improved to further optimize the service of public transport.

(3) The concept of regional bus stop is put forward in this paper; a suitable short-term prediction method for the passenger flow of regional bus stops is constructed, and the classification method and early warning coefficient of regional bus stop service level are developed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The project was supported by Open Fund of Engineering Research Center of Catastrophic Prophylaxis and Treatment of Road & Traffic Safety of Ministry of Education (Grant no. kfj180401, Changsha University of Science and Technology); Natural Science Foundation of Hunan Province, China (Grant no. 2019JJ40306); and National Natural Science Foundation of China (Grant no. 61773077).

References

[1] M. S. Ahmed and A. R. Cook, "Analysis of freeway traffic time-series data by using box-jenkins techniques," *Transportation Research Record*, vol. 722, pp. 1–9, 1979.

[2] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, no. 1, pp. 132–141, 1998.

[3] S. Lee and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1678, no. 1, pp. 179–188, 1999.

[4] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 2, pp. 246–254, 2009.

[5] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[6] G. A. Davis and N. L. Nihan, "Using time-series designs to estimate changes in freeway level of service, despite missing data," *Transportation Research Part A: General*, vol. 18, no. 5-6, pp. 431–438, 1984.

[7] B. Yoon and H. Chang, "Potentialities of data-driven non-parametric regression in urban signalized traffic flow forecasting," *Journal of Transportation Engineering*, vol. 140, no. 7, pp. 14–27, 2014.

[8] A. Emami, M. Sarvi, and S. Asadi Bagloee, "Using Kalman filter algorithm for short-term traffic flow prediction in a connected vehicle environment," *Journal of Modern Transportation*, vol. 27, no. 3, pp. 222–232, 2019.

[9] M. Abadi, A. Agarwal, P. Barham et al., "Tensorflow: large-scale machine learning on heterogeneous distributed systems," 2016, https://arxiv.org/abs/1603.04467.

[10] N. C. Petersen, F. Rodrigues, and F. C. Pereira, "Multi-output bus travel time prediction with convolutional LSTM neural network," *Expert Systems with Applications*, vol. 120, pp. 426–435, 2019.

[11] X. Ran, Z. Shan, Y. Fang, and C. Lin, "An LSTM-based method with attention mechanism for travel time prediction," *Sensors*, vol. 19, no. 4, p. 861, 2019.

[12] Y. Hou and P. Edara, "Network scale travel time prediction using deep learning," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 45, pp. 115–123, 2018.

[13] M. Wajeed and V. Sreenivasulu, "Image based tumor cells identification using convolutional neural network and auto encoders," *Traitement du Signal*, vol. 36, no. 5, pp. 445–453, 2019.

[14] Z. L. Zhang, Y. F. Wang, and Y. Li, "Inventory control model based on multi-attribute material classification: an integrated grey-rough set and probabilistic neural network approach," *Advances in Production Engineering & Management*, vol. 14, no. 1, pp. 93–111, 2019.

[15] B. S. Kim and T. G. Kim, "Cooperation of simulation and data model for performance analysis of complex systems," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 608–619, 2019.

[16] K. Krishna and B. Prakash, "Intrusion detection system employing multi-level feed forward neural network along with firefly optimization (FMLF2N2)," *Ingénierie des systèmes d information*, vol. 24, no. 2, pp. 139–145, 2019.

[17] Y. Li, D. Shi, and F. Bu, "Automatic recognition of rock images based on convolutional neural network and discrete cosine transform," *Traitement du Signal*, vol. 36, no. 5, pp. 463–469, 2019.

[18] Z. Zhang, Z. L. Guan, J. Zhang, and X. Xie, "A novel job-shop scheduling strategy based on particle swarm optimization and neural network," *International Journal of Simulation Modelling*, vol. 18, no. 4, pp. 699–707, 2019.

[19] K. Gorur, M. Bozkurt, M. Bascil, and F. Temurtas, "GKP signal processing using deep CNN and SVM for tongue-machine interface," *Traitement du Signal*, vol. 36, no. 4, pp. 319–329, 2019.

[20] H. M. Afify, K. K. Mohammed, and A. E. Hassanien, "Multi-images recognition of breast cancer histopathological via probabilistic neural network approach," *Journal of System and Management Sciences*, vol. 10, no. 2, pp. 53–68, 2020.

[21] B. Yu, H. Wang, W. Shan, and B. Yao, "Prediction of bus travel time using random forests based on near neighbors," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 4, pp. 333–350, 2018.

[22] Q. Tan, X. Ling, M. Chen, H. Lu, P. Wang, and W. Liu, "Statistical analysis and prediction of regional bus passenger

flows," *International Journal of Modern Physics B*, vol. 33, no. 11, Article ID 1950094, 2019.

[23] U. Reddy, P. Dhanalakshmi, and P. Reddy, "Image segmentation technique using svm classifier for detection of medical disorders," *Ingénierie des systèmes d information*, vol. 24, no. 2, pp. 173–176, 2019.

[24] S. Deore and A. Pravin, "Histogram of oriented gradients based off-line handwritten devanagari characters recognition using SVM, K-NN and NN classifiers," *Revue d'Intelligence Artificielle*, vol. 33, no. 6, pp. 441–446, 2019.

[25] J. Ke, H. Zheng, H. Yang, and X. Chen, "Short-term forecasting of passenger demand under on-demand ride services: a spatio-temporal deep learning approach," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 591–608, 2017.

[26] L.-M. Kieu, N. Malleson, and A. Heppenstall, "Dealing with uncertainty in agent-based models for short-term predictions," *Royal Society Open Science*, vol. 7, no. 1, Article ID 191074, 2020.

[27] Z. Li, M. N. Janardhanan, Q. Tang, and P. Nielsen, "Mathematical model and metaheuristics for simultaneous balancing and sequencing of a robotic mixed-model assembly line," *Engineering Optimization*, vol. 50, no. 5, pp. 877–893, 2018.

[28] Y. Xiao, J. J. Liu, Y. Hu, Y. Wang, K. K. Lai, and S. Wang, "A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting," *Journal of Air Transport Management*, vol. 39, pp. 1–11, 2014.

[29] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, 2015.

[30] M. C. Tan, S. C. Wong, J. M. Xu, Z. R. Guan, and P. Zhang, "An aggregation approach to short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 60–69, 2009.

[31] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[32] W. Hu, L. Yan, K. Liu, and H. Wang, "A short-term traffic flow forecasting method based on the hybrid PSO-SVR," *Neural Processing Letters*, vol. 43, no. 1, pp. 155–172, 2016.

[33] E. Doğan, "Short-term traffic flow prediction using artificial intelligence with periodic clustering and elected set," *Promet—Traffic & Transportation*, vol. 32, no. 1, pp. 65–78, 2020.

[34] S. A. Bagloee, M. Asadi, M. Sarvi, and M. Patriksson, "A hybrid machine-learning and optimization method to solve bi-level problems," *Expert Systems with Applications*, vol. 95, pp. 142–152, 2018.

[35] Y. Han, C. Wang, Y. Ren, S. Wang, H. Zheng, and G. Chen, "Short-term prediction of bus passenger flow based on a hybrid optimized LSTM network," *ISPRS International Journal of Geo-Information*, vol. 8, no. 9, p. 366, 2019.

[36] M. Gong, X. Fei, Z. H. Wang, and Y. J. Qiu, "Sequential framework for short-term passenger flow prediction at bus stop," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2417, no. 1, pp. 58–66, 2014.

[37] N. Van Oort, T. Brands, and E. de Romph, "Short-term prediction of ridership on public transport with smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2535, no. 1, pp. 105–111, 2015.

[38] B. A. Kumar, L. Vanajakshi, and S. C. Subramanian, "Bus travel time prediction using a time-space discretization approach," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 308–332, 2017.

[39] J. Wu, Q. Wu, J. Shen, and C. Cai, "Towards attention-based convolutional long short-term memory for travel time prediction of bus journeys," *Sensors*, vol. 20, no. 12, p. 3354, 2020.

[40] W. Wu, L. Huang, and R. Du, "Simultaneous optimization of vehicle arrival time and signal timings within a connected vehicle environment," *Sensors*, vol. 20, no. 1, p. 191, 2020.

[41] W. Wu, F. Zhang, W. Liu, and G. Lodewijks, "Modelling the traffic in a mixed network with autonomous-driving expressways and non-autonomous local streets," *Transportation Research Part E: Logistics and Transportation Review*, vol. 134, Article ID 101855, 2020.

[42] H. Wang, L. Li, P. Pan, Y. Wang, and Y. Jin, "Early warning of burst passenger flow in public transportation system," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 580–598, 2019.

[43] F. C. Pereira, F. Rodrigues, E. Polisciuc, and M. Ben-Akiva, "Why so many people? Explaining nonhabitual transport overcrowding with internet data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1370–1379, 2015.

[44] L. Bai, F. Z. Wang, and M. Zhang, "Urban rail transit network passenger flow monitoring and early warning system based on GIS," *Urban Rapid Rail Transit*, vol. 26, no. 6, pp. 56–59, 2013.

[45] L. Cheng, *Research on the Prediction and Early Warning of Short-Term Traffic Flow of Expressway*, Kunming University of Technology, Kunming, China, 2014.

[46] Transportation Research Board (TRB), *Transit Capacity and Quality of Service Manual*, 2nd edition, 2010.