Received: 21 February 2019 | Revised: 30 October 2019 | Accepted: 25 November 2019

DOI: 10.1002/for.2639

RESEARCH ARTICLE

WILEY

A predictive model of train delays on a railway line

Chao Wen[1,2,3] | **Weiwei Mou**[1,2] | **Ping Huang**[1,2,4] | **Zhongcan Li**[1,2]

[1]School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China

[2]National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu, China

[3]National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu, China

[4]High-speed Railway Research Centre, University of Waterloo, Waterloo, Ontario, Canada

Correspondence
Ping Huang, School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China.
Email: huangping129@my.swjtu.edu.cn

Funding information
National Nature Science Foundation of China, Grant/Award Numbers: 71871188 and U1834209, U1834209, 71871188; Science & Technology Department of Sichuan Province, Grant/Award Number: 2018JY0567; China Scholarship Council; Doctoral Innovation Fund Program of Southwest Jiaotong University, Grant/Award Number: D-CX201827; Science and Technology Department of Sichuan Province, Grant/Award Number: 2018JY0567

Abstract

Delay prediction is an important issue associated with train timetabling and dispatching. Based on real-world operation records, accurate forecasting of delays is of immense significance in train operation and decisions of dispatchers. In this study, we established a model that illustrates the interaction between train delays and their affecting factors via train describer records on a Dutch railway line. Based on the main factors that affect train delay and the time series trend, we determined the independent and dependent variables. A long short-term memory (LSTM) prediction model in which the actual delay time corresponded to the dependent variable was established via Python. Finally, the prediction accuracy of the random forest model and artificial neural network model was compared. The results indicated that the LSTM model outperformed the other two models.

KEYWORDS

delay prediction, LSTM model, railway, real-world data

1 | INTRODUCTION

Delays may spread out along time and space orientations, leading to delay propagation on the line or even on the network and contributing to the complexities of train operations. Train operations are highly dependent on running and dwell time variations (Chang & Thia, 1996). Dispatchers need a continuous estimation of succeeding train status, including the arrival and departure time at stations, running time in sections, and delays at stations and sections. Delay propagation is a function of delay aggravation caused by disturbances and delay recovery activities conducted by dispatchers. Delay propagation has been the main source of displacements in the railway system; thus minimizing delay propagation takes high priority. Analyses of microscopic and macroscopic approaches show that most of the studies consider the railway system at a microscopic rather than at a macroscopic level, and almost all papers have focused on minimizing delays of passengers or freight.

The estimation of running times requires predicting the effect of disturbances and subsequent buffer time adjustments that may be experienced during their operations. Delay prediction is a process of estimating delay probability based on known data at a given checkpoint and is typically measured via arrival (departure) delay. The key to making a delay prediction based on operational data involves establishing the relationship between train delays and various characteristics of a railway system. This provides a basis for the operator's scheduling decision. Delay prediction is a typical data-driven process because the following arrival or departure time is subject to its current status and the adjacent leading train. Thus dispatchers can set up routes and determine the arrival and departure times one after another. At a strategic level, accurate train delay prediction is conducive to the analysis of railway capacity and the effectiveness of railway route planning. It is well known that operators tend to reduce train delays by investing in infrastructure. Accurate delay prediction can detect habitual delays in railway routes and potential conflicts in train operation promptly, and this enables operators to improve infrastructure for specific routes, and thereby improves the overall transport efficiency of the railway system. With respect to the tactical level, accurate delay prediction is of tremendous significance in the establishment of a flexible and stable train timetable and aids in improving the stability of the train operation plan. Timetables are tested for robustness via probability distributions of process durations that are derived from historical traffic realization data. Conclusions from the tests are subsequently used to improve timetable robustness (Medeossi, Longo, & Fabris, 2011).

In this study, we attempt to predict train delays using a deep learning model while considering the interactions and delay propagation among trains in a group. Based on the actual running data of the Dutch railway Rotterdam Central to Dordrecht section, we use the long short-term memory (LSTM) model to predict the train arrival delay, which could be used as decision support for dispatchers. The main structure of the study is as follows: Section 2 presents the problem statements and state of the art on delay prediction; Section 3 introduces the LSTM model for arrival delay prediction; Section 4 presents a model forecast accuracy analysis and model evaluation with an experiment using the real-world train operation records; Section 5 discusses the main conclusions and applications.

2 | PROBLEM STATEMENT AND LITERATURE REVIEW

2.1 | Problem statement

Figure 1 shows that determining train status is a recursive, iterative process. The time axis (red line) denotes the current time, and dispatchers need to use known data on the left of the time axis to predict unknown events on the right of the time axis. The origination departure time p_4 of train $i + 1$ is determined by p_1, and p_5 is derived by p_2 and p_4. Similarly, p_7 is derived from p_5 and p_6. Train $i + 1$ is mostly subject to the status of train i, whereas train $i + 2$ is mostly subject to the status of train $i + 1$, whose predicted points of p_4, p_5, and p_7 are considered historical data. Major disturbances can propagate to other trains in the network, thereby requiring short-term adjustments in the timetable to limit delay propagation.

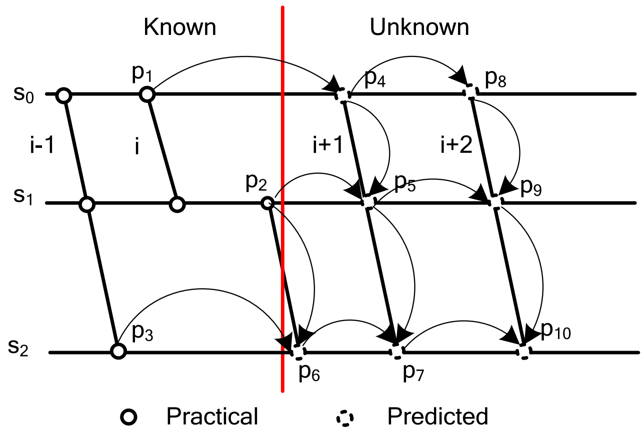


FIGURE 1 Recursive iterative processes of delay prediction [Colour figure can be viewed at wileyonlinelibrary.com]

2.2 | Literature review

Traditional statistical machine learning methods consider train operation performance as model-driven data to update algorithm structure and parameters in time, such as delay probability updating in the Bayesian network and pruning of a decision tree. Based on the train operation data of the Netherlands railway network, Huisman, Boucherie, and Van Dijk (2002) propose a solvable queueing network model to compute performance measures of interest without requiring train schedules (timetables). A new analytical stochastic model of train delay propagation in stations is proposed, which estimates the knock-on delays of trains caused by route conflicts and late transfer connections realistically (Yuan & Hansen, 2007). Berger, Gebhardt, Müller-Hannemann, and Ostrowski (2011) proposed a stochastic model of delay propagation to predict train arrival and departure delay events in large transportation networks. The model is suitable for all public transportation systems and requires online prediction. The actual delay data of the train should be updated in real time. The results obtained by Olsson and Haugland (2004) indicate that passenger management is an important factor that affects train punctuality in congested areas, whereas the management of train crossings is the key factor that affects train punctuality in noncongested areas. Daamen, Goverde, and Hansen (2009) describe a software tool that automatically identifies knock-on delays in an accurate and nondiscriminative way. The tool is based on the logic of the dynamic system, which has been described with a CPN model and can distinguish between two main classes of knock-on delay: hindrance at conflicting track sections and waiting for scheduled connections at stations. Based on a timed event graph representation of a scheduled railway system, Goverde (2010) introduced a max-plus delay propagation model and a bucket-based delay propagation algorithm to compute the propagation of initial delays over a periodic railway timetable. The author modeled the railway system as a linear system in max-plus algebra including zero-order dynamics corresponding to the delay propagation within a timetable period. The behavior of the delay propagation and the convergence of the algorithm were analyzed depending on timetable properties such as realizability and stability.

Statistical distribution and regression models are usually used to interpret the distributional characteristics of train delays and the dependency between train delays and their influencing factors. Gorman (2009) used statistical analysis to identify primary congestion predictive factors (meets, passes, overtakes) that are the major contributors to freight rail congestion and used statistical methods to forecast the average monthly train running time, where the average absolute percentage error corresponded to 4.6%. Flier, Graffagnino, and Nunkesser (2009) combined a linear regression and combination model to predict delay based on the online train delay monitoring data of the Swiss railway network. The model tested the regional corridor of Lucerne and achieved good prediction results without considering station capacity constraints. Because it is difficult to analyze stochastic models for the propagation of delays in railway networks, Meester and Muns (2007) discuss a model for delay propagation and show that in a world of so-called phase-type distributions it is possible to derive secondary delay distributions from primary delay distributions.

The train running process is typically considered as a Markov process. Based on the examination of a large set of historical data, Barta, Rizzoli, Salani, and Gambardella (2012) demonstrated how to classify different terminals according to the ability either to absorb or to amplify delays and developed a Markov-chain-based model to evaluate the evolution of train delays as a train visits successive terminal. Because the sequence of train departure and arrival times can be viewed as a stochastic process, the performance of train operations may be analyzed and modeled using Markov chains. Şahin (2017) modeled the train departure and arrival delays at stations as states and analyzed the successive state changes along train paths on a single-track railway. A method for modeling the uncertainty of train delays, based on a Markov stochastic process, was proposed by Kecman, Corman, and Meng (2015). The model was based not just on static, offline collected data, but also included the dynamic characteristics of the ever-changing delays, thus increasing the reliability of prediction by 71%.

Bayesian networks can provide a timely update of the train running status based on new operation data. Zilko, Hanea, Kurowicka, and Goverde (2014) first attempted to apply the nonparametric Bayesian network (NPBN), which represents the joint distribution between variables that describe the nature of the disruption, to predict the disruption length for the Dutch Operational Control Centre Rail. Later, they proposed a new model with copula Bayesian networks that considered the factors influencing the length of the disruptions and modeled the dependence between them (Zilko, Kurowicka, & Goverde, 2016). Lessan, Fu, and Wen (2018) proposed a hybrid Bayesian network model to predict high-speed railway (HSR) delays using train operation records of Wuhan–Guangzhou HSR. The proposed model can achieve over 80% accuracy in predictions on overage within a 60 min horizon.

Additionally, hybrid models are also widely used in train operation modeling. A joint method employing Bayesian reasoning and Markov model can be used to predict the delay state at different stations (Martin, 2016; Martin & Romanovsky, 2016). It has also been used to establish a delay prediction model (Corman & Kecman, 2018; Kecman & Goverde, 2015b) that utilizes robust linear regression, regression tree, and random forest models to predict the train running time and dwell time. Furthermore, robust linear regression was improved, and a local model was proposed for local routes and sections. The results indicated that the local model exhibited higher prediction accuracy. A hybrid approach that combined decision tree and random forest regression have also been used to predict the running time, dwell time, train delay, and penalty costs, and merges the data-driven model and experience-based model approaches (Lulli, Oneto, Canepa, Petralli, & Anguita, 2018).

Neural network machine learning methods do not need to be based on prior scheduling knowledge, and they predict train delays by learning useful features from massive data. Marković, Milinković, Tikhonov, and Schonfeld (2015) determined the effect of the infrastructure on train delays by experts and then used the support vector machine model to predict the arrival time of a train at a station. When compared with the ordinary artificial neural network model, this indicated that the support vector machine model exhibited a better prediction effect. Based on the actual data of the Wuhan–Guangzhou high-speed railway, Chen, Wang, and Li (2015) proposed three models, namely least squares method, support vector machine, and least squares support vector machine models, to determine train location and predict train delay. Specifically, an artificial neural network (ANN) was used to establish the delay prediction model, and a data-driven model was constructed based on the train operation data in Iran and Germany. The model validation results indicated that the prediction accuracy of the model is high (Peters, Emig, Jung, & Schmidt, 2005; Yaghini, Khoshraftar, & Seyedabadi, 2013).

Most recently, a shallow and deep extreme learning machine (DELM) was proposed in conjunction with the rapid development of big data technologies. Oneto et al. (2016, 2017b) presented a data-driven train delay prediction system for a large-scale railway network to provide useful information on railway traffic control processes by using state-of-the-art tools and techniques. The system extracted information from a large amount of historical train movement data using big data technologies, learning algorithms, and statistical tools. The described approach and prediction system were validated based on real historical data in 6 months. The results revealed that the DELM outperformed the current technique, and this was mainly based on the event graph proposed by Kecman and Goverde (2015a). Oneto et al. (2017a) developed a data-driven dynamic train delay prediction system based on the findings of Oneto et al. (2017b). This integrated heterogeneous data sources to deal with varying dynamic systems via DELM. The system exploited state-of-the-art tools and techniques, was completely data driven, and did not require any prior information on the railway network.

When compared with the traditional statistical machine learning model, deep learning uses deep neural network models for learning. The steps it learns corresponds to signal–feature–value. The first step involves not determining via learning the structure of the input data and not via random initialization. Therefore, the initial value is closer to the global optimum, and the model achieves better results. Overall, it corresponds to a layer-wise training mechanism. If the traditional neural network reaches more than seven layers, then the residual propagation to the foremost layer is extremely low; gradient diffusion occurs, and this affects the accuracy of the model. When compared to traditional neural networks, deep learning reduces the number of neural network parameters and adds new structures (e.g., LSTM and ResNet), a new activation function (ReLU), new weight initialization methods (e.g., layer-by-layer initialization and XAVIEER), new loss functions, and new overfitting methods (e.g., Dropout and BN). It is characterized by a deep neural network selection that overcomes artificial choices.

Finally, the summary and analysis of the reviewed typical literature are listed in Table 1.

3 | TRAIN ARRIVAL DELAY PREDICTION MODEL

3.1 | Selection of characteristic variables

Delay prediction is a process of estimating the probability of train delays at subsequent recording points based on train operation history data, and arrival delays typically determine this. It is assumed that a train is currently located at station s_n, the previous station corresponds to s_{n-1},and the next station corresponds to s_{n+1}. Based on the train delays at s_n and s_{n-1} stations and scheduled running time of trains at sections (s_{n-1}, s_n), (s_n, s_{n+1}), the study predicts the arrival delays of trains at s_{n+1}stations. As shown in Figure 1, the train arrives at the station s_n at time t_n^A on schedule and starts at the same station at time t_n^D. However, in the actual operation process, given various interference factors, the train can deviate from the timetable to generate the actual arrival time $\hat{t}_n^A$ and actual departure time $\hat{t}_n^D$. Figure 2 shows successive stations (s_{n-1}, s_n, and s_{n+1}), with the parameters in parentheses indicating the scheduled time and actual time of the event. The train delay is typically divided into arrival delay and departure delay. The difference between the actual and scheduled times $\left(\hat{t}_n^A - t_n^A\right)$ and $\left(\hat{t}_n^D - t_n^D\right)$ indicate the arrival and departure delays, respectively, of the train at station s_n.

The train can be delayed as a result of various disturbances in the operation process. We select seven parameters via the analysis of the train arrival delays at the station to constitute the feature space (F). The study assumes that the parameters affect the future delay of the train, and thus the future arrival of the train is predicted based on the selected parameters.

The variables included in the feature space (*F*) are as follows:

1. *Train characteristic* (X_1). There are three main characteristics of trains running in the Rotterdam Central to Dordrecht section of the Netherlands railway system, namely regional trains stopping at each station (SPR), intercity trains stopping at large stations (IC), and empty ride trains (LM).
2. *The impact of peak hours* (X_2). We consider the impact of peak hours on train arrival times. A binary variable is created that indicates whether the observed process takes place during a peak hour. Because there are no

TABLE 1 Summary of the main reviewed literature

Literature	Methods	Data of country	Features and contributions
Huisman et al. (2002)	Queueing network	Netherlands	Transform the railway network into a queueing network
Yuan and Hansen (2007)	Stochastic model	Netherlands	Consider knock-on delays caused by route conflicts and late transfer connection
Berger et al. (2011)	Stochastic model	Germany	Realistic and suited for an online scenario
Flier et al. (2009)	Linear regression combination model	Switzerland	Combine linear regression and combination model without considering station capacity constraints
Daamen et al. (2009)	Petri net	Netherlands	Describe a tool that automatically and without discrimination identifies route conflicts and train numbers involved
Gorman (2009)	Statistical analysis; Regression analysis	USA	Use econometric methods to predict congestion delays
Goverde (2010)	Max-plus algebra; timed event graphs.	Netherlands	Introduce an effective delay propagation algorithm considering buffer times
Barta et al. (2012)	Markov chains	Hupac transportation network	Present an algorithm to forecast the propagation of delays in a rail freight network
Kecman et al. (2015)	Markov chains	China	Propose a Markov model whose probability transition matrix changes over time
Zilko et al. (2016)	Bayesian networks	Netherlands	Construct a disruption length prediction model with computational efficiency and fast inference
Lessan et al. (2018)	Bayesian network	China	Develop the hybrid BN structure model
Martin (2016), Martin and Romanovsky (2016)	Markov model	UK	Propose a method in which to formally specify the design and reliability criteria of an advisory system
Kecman and Goverde (2015b)	Linear regression, regression tree, and random forest	Netherlands	Predict the train running time and the dwell time with a variety of methods
Lulli et al. (2018)	Decision tree and random forest	Italy	Propose a hybrid approach to address the limitations of experience-based models and data-driven models
Marković et al. (2015)	Support vector regression	Serbia	Consider the effect of infrastructure on train delays by experts
Chen et al. (2015)	Support vector regression	China	Propose three-position computation models
Yaghini et al. (2013)	Neural network	Iran	Use three different methods to define inputs
Peters et al. (2005)	Neural network	Germany	Use the neural network, which can abstract from known delay constellations
Oneto et al. (2017b); Oneto et al. (2017a).	Shallow and deep extreme learning machines	Italy	Propose a completely data-driven system

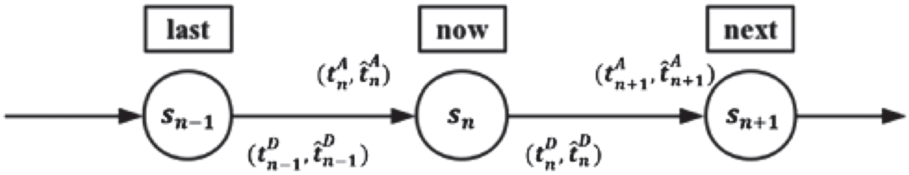


FIGURE 2 General scheme of train movements at three successive stations

data on passengers in this paper, it is difficult to determine the exact time division of peak and off-peak periods. Therefore, the definition of train peak hours and off-peak hours is mainly adopted from that of the peak hours of working days for Dutch national train operators. The peak time for the morning is from 6:30 to 9:00 and for the evening from 16:00 to 18:30.

3. *Departure delay time of the train at the current station* (X_3). The actual departure delay time of the train at the current station s_n denotes the difference between the actual departure time of the train at station s_n and the planned departure time. The equation corresponds to $\hat{t}_n^D - t_n^D$, and this is accurate to seconds.
4. *Arrival delay time of the train at the current station* (X_4). The actual arrival delay time of the train at the current station s_n denotes the difference between the actual arrival time of the train at s_n station and planned arrival time. The equation corresponds to $\hat{t}_n^A - t_n^A$, and this is accurate to seconds.
5. *Departure delay time of the train at the last station* (X_5). The actual departure delay time of the train at the last station denotes the difference between the actual departure time of the train at station s_{n-1} and planned departure time. The equation corresponds to $\hat{t}_{n-1}^D - t_{n-1}^D$, and this is accurate to seconds.
6. *Planned running time of the train in the last section* (X_6). The calculation equation for the planned running time between the last station s_{n-1} and current station s_n corresponds to $t_n^A - t_{n-1}^D$, and this is accurate to seconds.
7. *Planned running time of the train in the next section* (X_7). The calculation equation for the planned running time between the current station s_n and next station s_{n+1} corresponds to $t_{n+1}^A - t_n^D$, and this is accurate to seconds.

The output variable of the delay prediction in the study denotes the arrival delay time (Y) of the train at the next station. The typical delay prediction data based on the aforementioned characteristic variables are shown in Table 2. The expression is detailed as follows:

$$\mathrm{Y} = \varphi(\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_4, \mathrm{X}_5, \mathrm{X}_6, \mathrm{X}_7), \tag{1}$$

where denotes the train arrival delay (output variable), X_1, X_2, X_3, X_4, X_5, X_6 and X_7 denote the train delay influence factors (input variables), and φ denotes the machine learning algorithm model.

3.2 | LSTM model

The LSTM model was proposed by Hochreiter and Jürgen (1997) to improve the model based on the recurrent neural network (RNN). In a conventional RNN, the hidden layer generally corresponds to an extremely simple node such as tanh. The LSTM improves the simple node of the hidden layer into a storage unit. The basic structure of the storage unit is shown in Figure 3. The storage unit is composed of an input gate *i*, an output gate *o*, a forgetting gate *f*, and a memory cell *c*. In forward propagation, the input gate determines when to activate the incoming storage unit, while the output gate determines when to activate the outgoing storage unit. In reverse propagation, the output gate determines when to allow errors to flow into the storage unit, and the input gate determines when to let it flow out of the storage unit. The input gate, output gate, and forgetting gate constitute keys to control information flow. The operation principle of the storage unit is expressed in terms of Equations (2)–(6) (Bengio, Simard, & Frasconi, 2002; Gers, Schraudolph, &

TABLE 2 Modeling data

Date	Train number	Last station	Current station	Next station	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*Y*
2017/9/4	5,139	Brd	Zwd	Ddr	SPR	0	113	17	−7	180	300	140
2017/10/2	2,216	Sdm	Rtd	Rtb	IC	1	106	124	138	132	300	124
2017/10/26	2,214	Rtd	Rtb	Rtz	IC	1	819	809	781	120	132	812
2017/11/7	2,218	Rlb	Brd	Zwd	IC	0	215	215	190	60	240	179
2017/12/8	5,027	Dtz	Sdm	Rtd	SPR	1	128	94	116	378	300	80

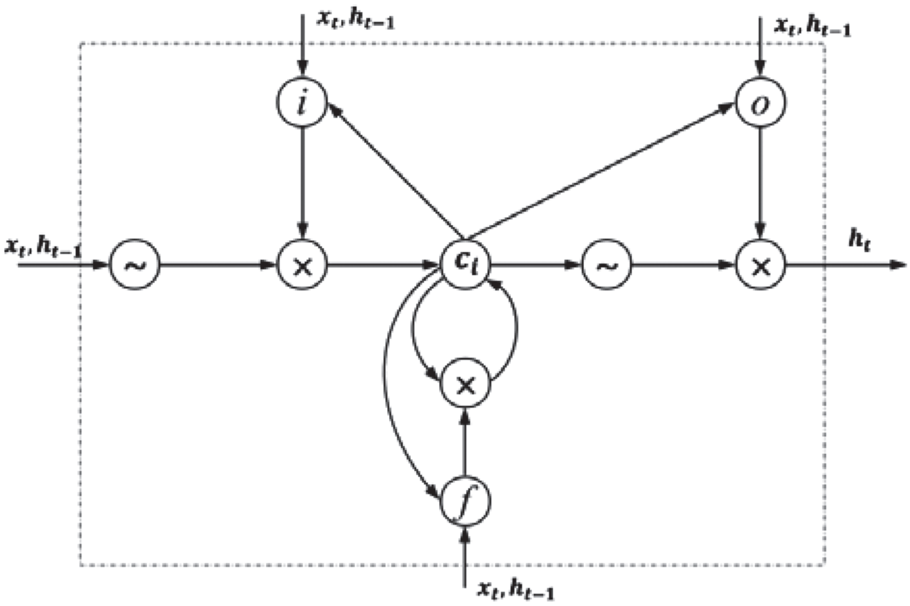


FIGURE 3 Basic structure of the LSTM storage unit

Schmidhuber, 2002; Greff, Srivastava, & Koutník, 2016) as follows:

$$i_t = \delta(\boldsymbol{W_i} x_t + \boldsymbol{U}_i h_{t-1} + \boldsymbol{V}_i C_{t-1} + b_i), \tag{2}$$

$$f_t = \delta\left(\boldsymbol{W_f} x_t + \boldsymbol{U}_f h_{t-1} + \boldsymbol{V}_f C_{t-1} + b_f\right), \tag{3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(\boldsymbol{W_c} x_t + \boldsymbol{U}_c h_{t-1} + b_c), \tag{4}$$

$$o_t = \delta(\boldsymbol{W_o} x_t + \boldsymbol{U}_o h_{t-1} + \boldsymbol{V}_o C_t + b_o), \tag{5}$$

$$h_t = o_t \cdot \tanh(c_t), \tag{6}$$

where c_t denotes the calculation method of memory cells at time t; h_t denotes all outputs of LSTM units at time t; $\boldsymbol{W}$, $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{b}$ denote the matrices of coefficients and vector of offset; δ denotes the activation function sigmoid; '·' denotes a point multiplication operation; and i_t, f_t, and o_t denote the calculation methods of the input gate, forgetting gate, and output gate at time t, respectively. As shown in Figure 3, the outputs of the three gates of the input gate, forgetting gate, and output gate are connected to a multiplier element to control the input and output of information flow and the status of cell units, respectively.

In actual operations, given the mutual restriction between trains, a delay of the forward train can affect the backward train and result in delay propagation along the time axis. The LSTM model assumes time series format data as input, and its results at any t-time are based on the results at the previous time and input data at the current time. This mechanism enables the preservation and reuse of time series information in the model for a long period such that it learns the knowledge of time series correlation in time series data.

The LSTM model for delay prediction is constructed as follows:

1. A number of subsequent stations on a railway line are selected to extract the arrival delay time Y and corresponding feature space (F). All train delays and their extraction attributes are sorted based on the actual departure time of the train at the current station, and training data sets and test data sets are divided. The trains are sorted according to their actual departure time at the current station, and the data include overtaking of trains at stations before (and including) the current station. After the train departs from the current station, it runs through a section to reach the forecasting station, and because it cannot overtake or be overtaken in this section the model can avoid considering the overtaking information of trains, which makes the model easier to use in practice. As shown in Figure 4, the first row in the figure indicates the train arrival delay time (Y), and the second row indicates the characteristic space (F) of the influence factors of the delay time. Specifically, i denotes the train number; s_n denotes the station number, and the sliding window length l denotes the number of trains that are predicted to be entered each time. Hence we consider the effect of the previous l trains on the current train delay.
2. Determination of parameter l: The delay time and influencing factors of each train are treated as time series. The model considers the interaction relationship between different train numbers by inputting multiple trains each time. As shown in Figure 5, with the increase in l the prediction effect of the model has not significantly improved, so we finally chose $l = 1$.

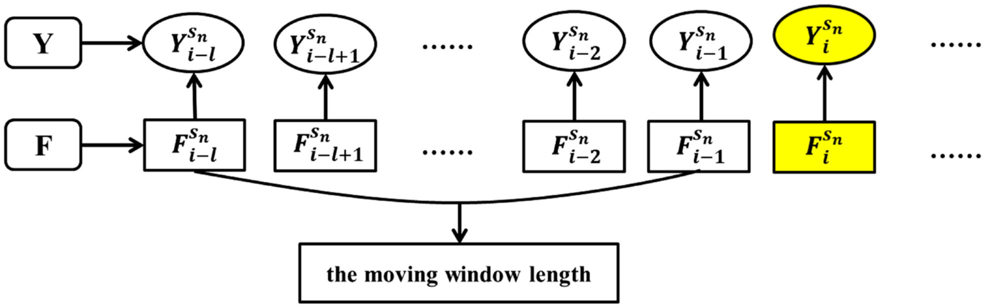


FIGURE 4 LSTM input data format [Colour figure can be viewed at wileyonlinelibrary.com]

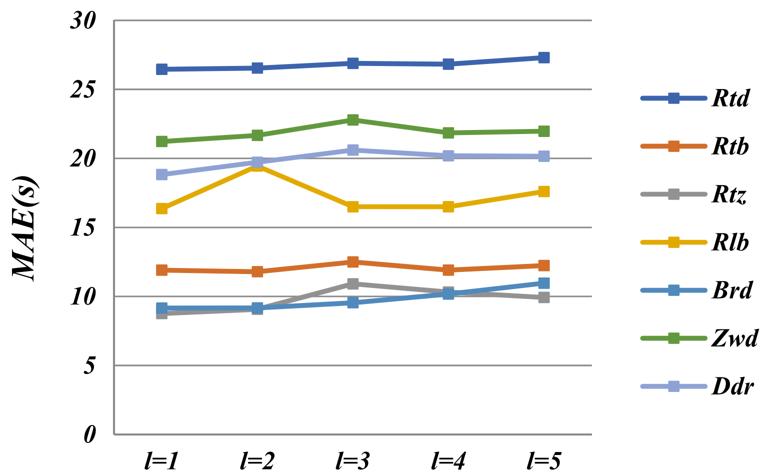


FIGURE 5 Moving window length selection [Colour figure can be viewed at wileyonlinelibrary.com]

Thus only the effect of the previous train delay on the arrival of the current train is considered.

3. After determining the optimal number of input trains, the model structure and parameters (e.g., hidden layer number, neuron number, learning rate, optimizer, and dropout rate) are optimized to obtain the optimal parameters and structure of the model and predict the arrival delay of the train at the station. Finally, the LSTM model with the time series input form is shown in Figure 6. The arrival delay time ($Y_i^{S_n}$) of the current train is predicted based on the feature space ($F_i^{S_n}$) of the current train and the effect of only the previous train ($F_{i-1}^{S_n}$). The aforementioned step is repeated to finally realize the prediction for all stations from the Rotterdam Central to the Dordrecht section.

4 | EXPERIMENTAL RESULTS

4.1 | Data description

The actual data of the train operations in the study were collected from the Rotterdam Central to Dordrecht section of the Dutch railway system, which contains seven stations, namely Rotterdam Central (Rtd), Rotterdam Blaak (Rtb), Rotterdam Lombardijen (Rlb), Barendrecht (Brd), Zwijndrecht (Zwd), and Dordrecht (Ddr). This is a busy section of the Dutch railway line, with higher train frequency and capacity utilization. There are three different train characteristics in this section, namely IC, SPR, and LM, so this section was selected for our train arrival delay study. Basic information about the stations, sections, and train operations is shown in Table 3.

The data include the 66,178 train operation records at seven stations (six sections) in total. The train records in the peak hours accounted for 26.5% and in other periods for 73.5%, for three types of trains: IC (49.2%), SPR

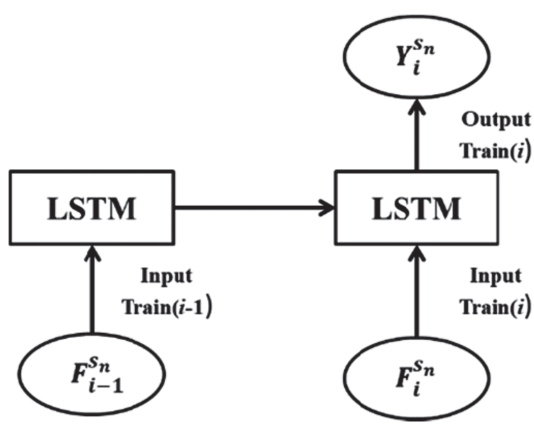


FIGURE 6 LSTM prediction model

TABLE 3 Station infrastructure information

		Rtd	Rtb	Rtz	Rlb	Brd	Zwd	Ddr
Track layout	Tracks	16	2	4	8	4	5	14
	Platforms	8	2	1	2	3	3	3
Scheduled headway times (min)	IC	14	9	10	9	8	6	6
	SPR	15	11	11	11	11	11	5

		Sdm-Rtd	Rtd-Rtb	Rtb-Rtz	Rtz-Rlb	Rlb-Brd	Brd-Zwd	Zwd-Ddr
Distance (km)		3.4	2.0	2.4	2.8	3.1	7.6	2.1
Allowed speed (km/h)		100	100	100	100	100	100	100
Average scheduled running times (min)	IC	5	2.3	2.5	2	2.4	4.6	3.4
	SPR	5	2.3	2.5	2	2.4	4.6	3.4

TABLE 4 Part of the original data

Date	Train number	Train characteristic	Location	Activity	Planned time	Actual time
2017/9/4	5,195	SPR	Zwd	K_A	1:14:18	1:14:43
2017/9/21	2,274	IC	Brd	K_A	22:59:00	22:59:10
2017/9/29	5,131	SPR	Ddr	A	9:20:00	9:20:21
2017/11/13	5,025	SPR	Rtd	A	7:39:00	7:39:15

Note. A stands for an arrival. K_A indicates a short stop, which is an arrival planned within the same minute.

(48.4%), and LM (2.4%). The time span was 66 working days from September 4, 2017, to December 8, 2017, excluding weekends. Data records include the date, train number, train characteristic, location, train activity, planned time, realization, delay jump, and delay cause. A few examples of the data are shown in Table 4.

4.2 | Prediction accuracy analysis

In order to evaluate the prediction effect of the model, the following analysis is initially performed. As shown in Figure 7, the actual and predicted arrival delays of trains at stations are compared. Second, as shown in Figure 8, the scatter plots of the observed and predicted arrival delays of trains are illustrated. The results indicate that the predicted values of train arrival delays exhibit a good match with the observed values. Specifically, in the interquartile range, the whiskers and right tail closely match in the figures for each station. Furthermore, as shown in Figure 8, the majority of predictions are close to the depicted diagonal lines for arrival events. This implies that the predicted value is extremely close to the observed value.

Figure 9 shows the distribution of predicted residuals for train arrival delays at different stations. The figure assumes the train actual arrival delay time as abscissa and the residuals as ordinate for visualization purposes. As shown in the figure, in the seven stations of Rotterdam Central to Dordrecht section of the Netherlands railway system, all stations (with the exception of the Rtd station) exhibit good prediction results. Increases in the prediction error of the Rtd station can be due to the increasingly significant influence of the outliers. Figure 10 shows the prediction accuracy histogram of the LSTM model for the seven stations. As shown in the figure, the model accuracy corresponds to 86.91% with an allowable error within 30 s, and thus the model exhibits a good prediction effect.

To better demonstrate the prediction effect of the LSTM model, the predicted results are divided into four subcategories according to the train type (IC, SPR) and whether they are located in the peak period, and the prediction results of each category are displayed separately. Figure 11 shows the accuracy of each station's prediction results under different allowable errors. The horizontal axis represents the station, the vertical axis represents the accuracy, and bars with different colors represent the allowable errors (30 s, 60 s, 90 s, etc.). As shown in Figures 11a and 11b, for the IC and SPR trains, when the allowable error of the LSTM model is within 60 s, the prediction accuracy of each station exceeds 90%. Figures 12a and 12b show the residual distribution of the IC and SPR trains for each station. It can be seen from the figures that the residuals are mainly located around zero, which indicates that the model has a good predictive performance for both IC and SPR trains.

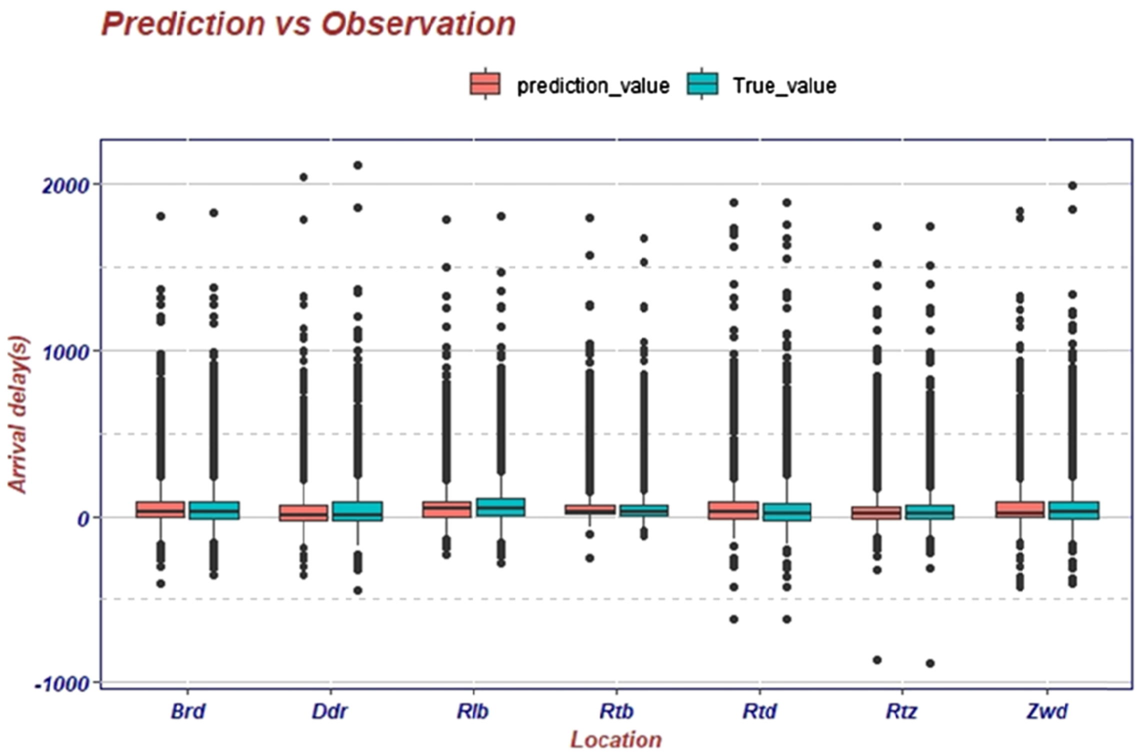


FIGURE 7 Comparison of predicted and observed arrival delay distribution for different stations [Colour figure can be viewed at wileyonlinelibrary.com]

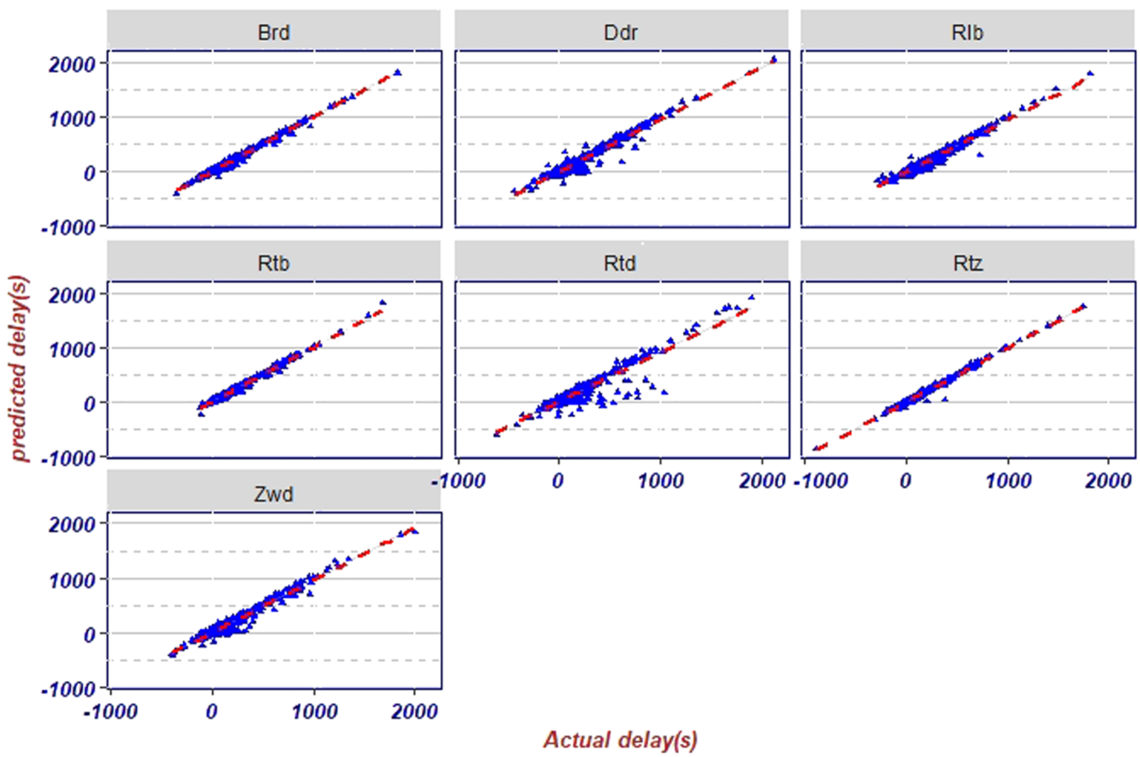


FIGURE 8 Scatter plots of actual delays VS predicted delays [Colour figure can be viewed at wileyonlinelibrary.com]

Figures 13 and 14 show the predicted results of the LSTM model of the trains during peak and off-peak hours. Figure 13 shows the predictive accuracy of each station under different allowable errors. It can be seen from Figures 13a and 13b that the LSTM prediction accuracy of the trains during peak and off-peak hours is over

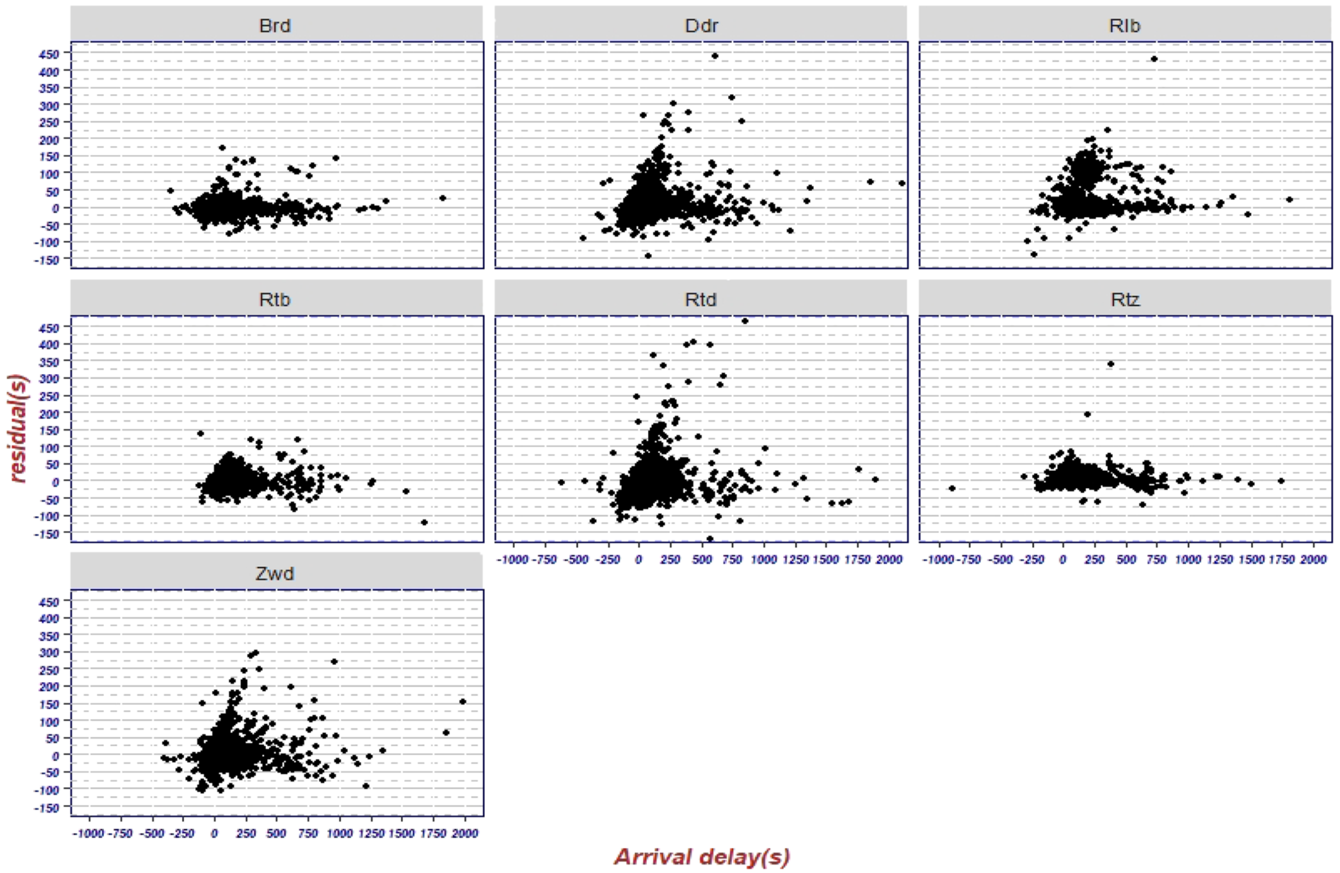


FIGURE 9 Distribution of the residuals of train arrival delays at different stations [Colour figure can be viewed at wileyonlinelibrary.com]

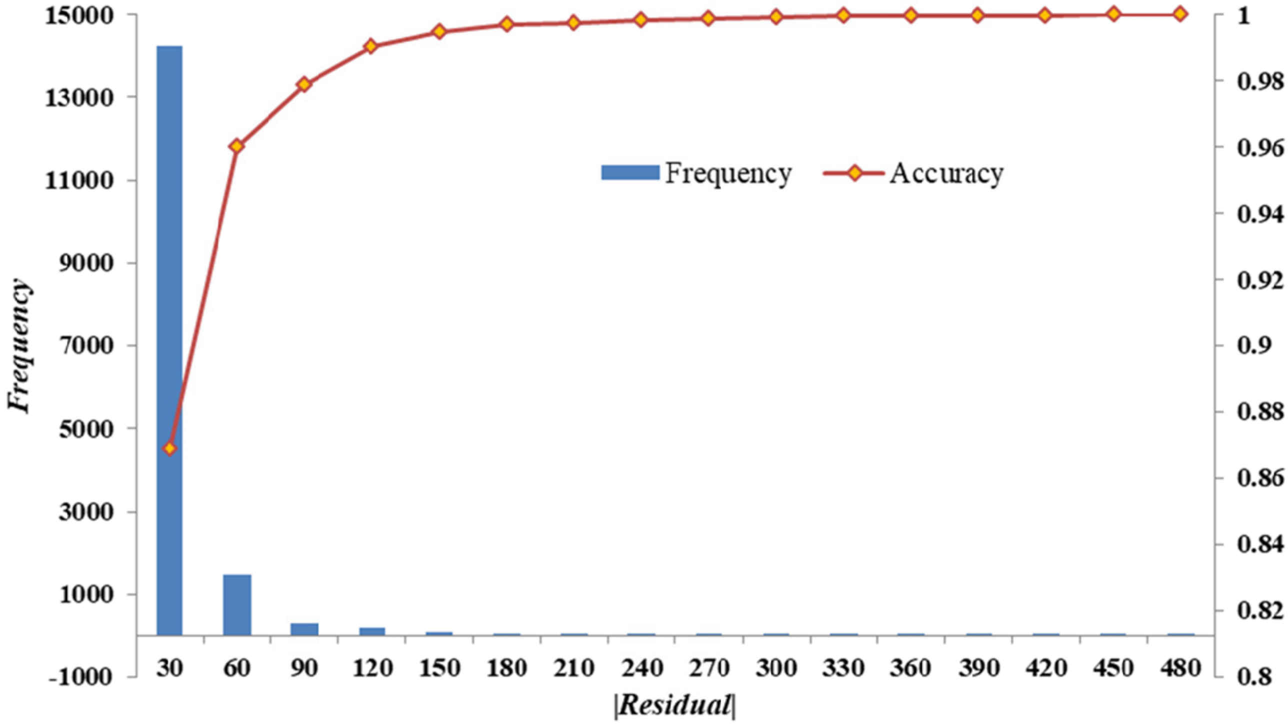


FIGURE 10 LSTM model prediction accuracy [Colour figure can be viewed at wileyonlinelibrary.com]

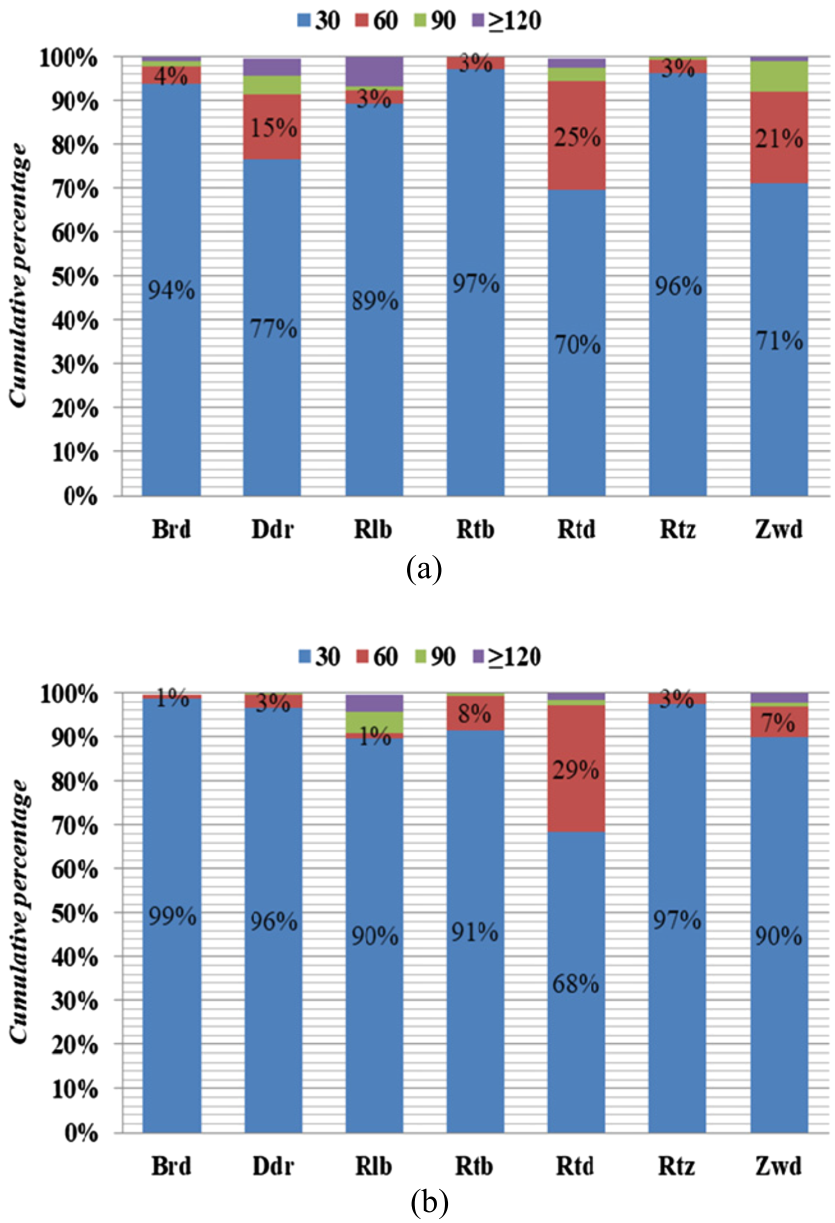


FIGURE 11 Prediction accuracy of (a) IC and (b) SPR trains [Colour figure can be viewed at wileyonlinelibrary.com]

90% within the allowable error of 60 s. Figures 14a and 14b are the residual distribution diagrams of the train arrival delay predictions for different time periods.

4.3 | Model evaluation

4.3.1 | Benchmark model

In order to better evaluate the prediction effect of the model, two benchmark models are selected and compared with the LSTM model: namely the random forest (RF) model and artificial neural network (ANN) model. They are detailed as follows.

Random forest. The RF is a joint prediction model that is composed of multiple decision trees (Cutler, Cutler, & Stevens, 2004; Loh, 2011), and this can be used as a fast and effective classification and prediction model. Each decision tree in RF consists of several forks and nodes. Each decision tree is regressed and predicted. Finally, the predictive effect of the random forest is determined via the predictive effect of multiple decision trees. The RF corresponds to an ensemble learning algorithm, and this belongs to the Bagging type. The final result is voted or averaged by combining multiple weak classifiers, and thus the overall model results exhibit higher accuracy and generalization performance. Thus the model yields good results, and this is mainly due to the "random" and "forest" elements, which make it resistant to overfitting and increase precision.

Artificial neural networks. An ANN is one of the most commonly used train delay prediction models (Malavasi,

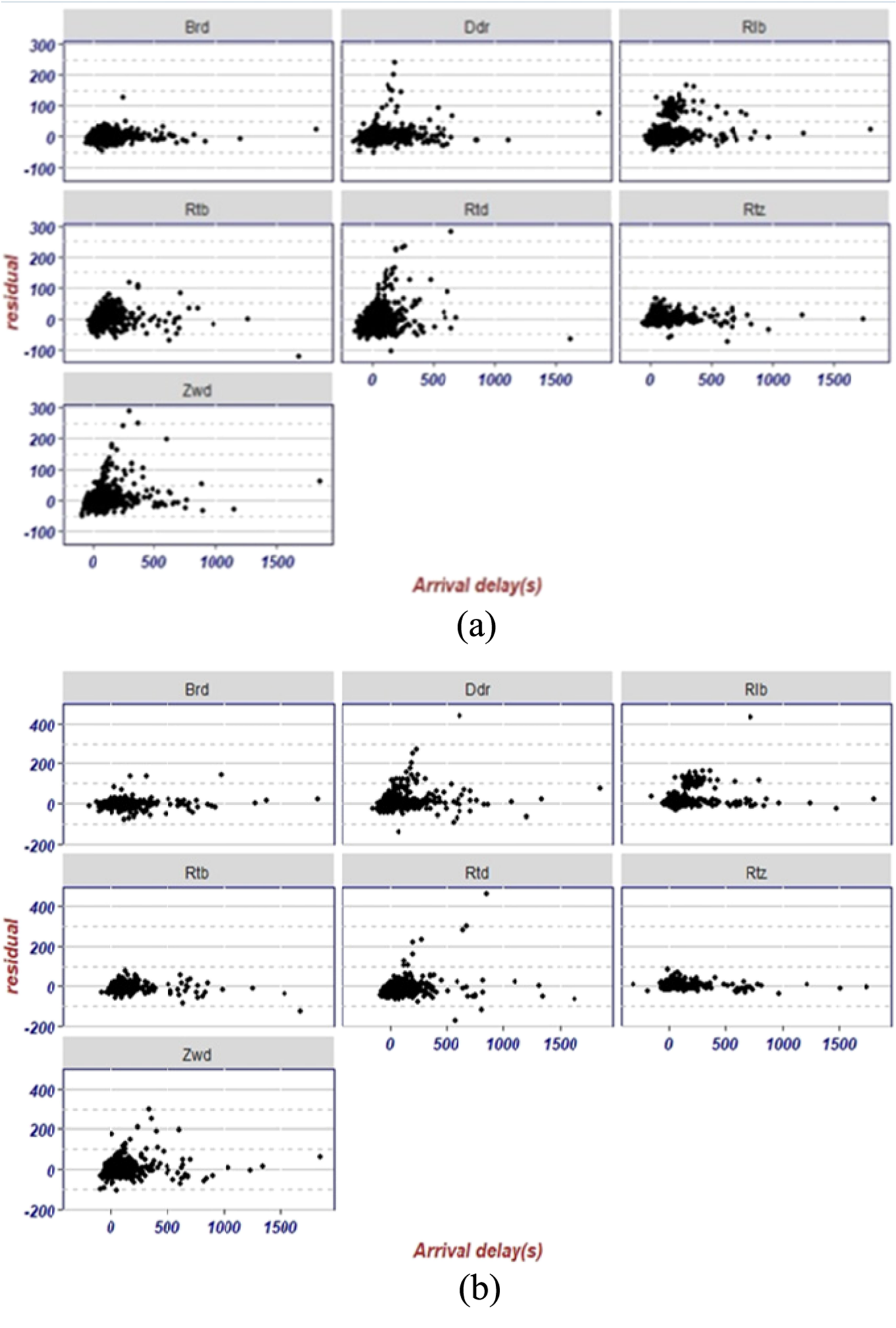
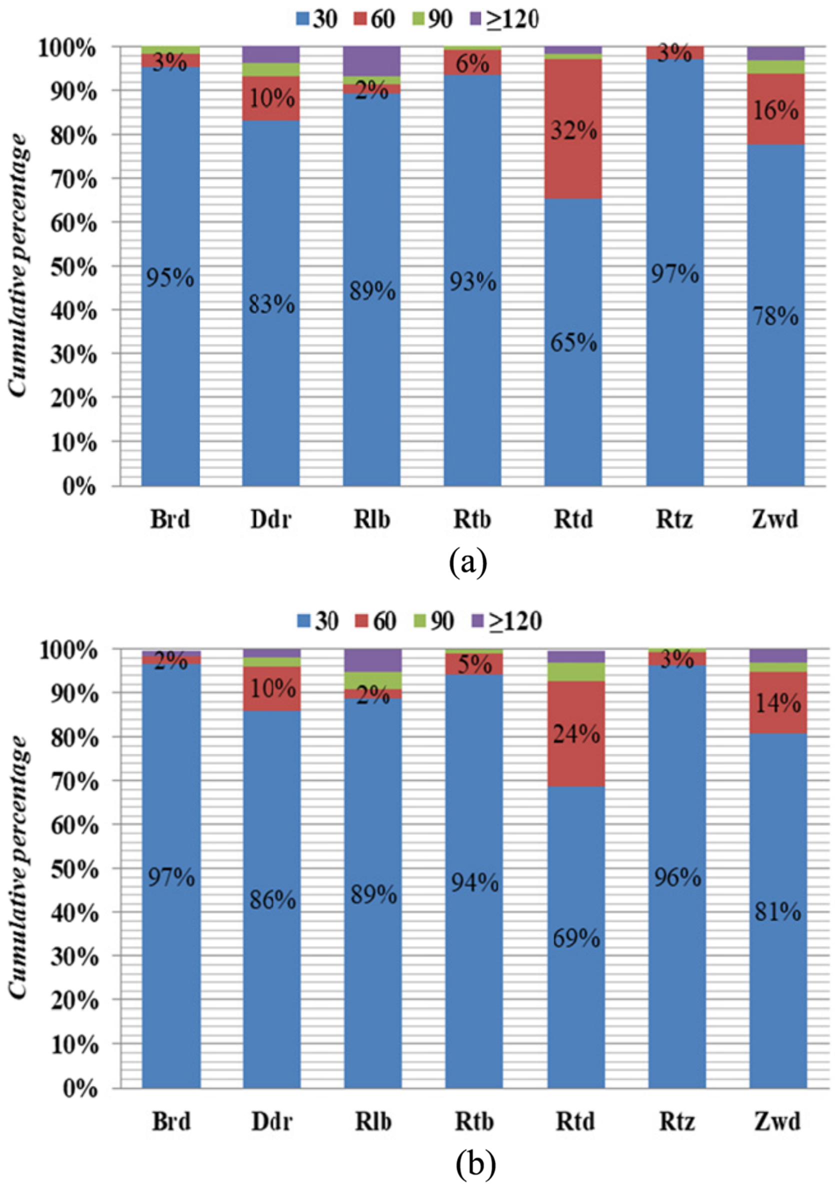


FIGURE 12 Distribution of the residuals of (a) IC and (b) SPR train arrival delays [Colour figure can be viewed at wileyonlinelibrary.com]

2001; Peters et al., 2005; Yaghini et al., 2013). It mainly models the relationship between a set of input signals and a set of output signals. The model is derived from the reaction of the human brain to stimuli from sensory input. Like how the brain uses a network of interconnected cells of a neuron to create a large parallel processor, ANNs use artificial neurons or a network of nodes to solve learning problems. There are three main characteristics of ANNs: (1) an activation function that converts the net input signal of a neuron into a single output signal for further propagation in the network; (2) network topology that describes the number of neurons in the model, number of layers, and the manner in which the layers are connected; and (3) a training algorithm that specifies the setting of the connection weight to suppress or increase the proportion of neurons in the input signal. This model is suitable for situations involving simple input and output data albeit an extremely complex input-to-output process.

FIGURE 13 Prediction accuracy of the trains during (a) peak hours and (b) off-peak hours [Colour figure can be viewed at wileyonlinelibrary.com]



4.3.2 | Model evaluation index

Concerning model evaluation, the study mainly selects mean absolute error (MAE) and root mean squared error (RMSE) as evaluation indexes. The equation to calculate the index is as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |p_i - y_i|, \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{I=1}^{N} (p_i - y_i)^2}, \tag{8}$$

where p_i and y_i denote the predicted and observed delay values for ith arrival events, respectively, and n denotes the total number of observations. The measures quantify the average deviation of the predictions from the observed values. The model's performance level improves when the measures are closer to zero.

Figures 15 and 16 compare MAE and RMSE values for LSTM, RF, and ANN models of different stations. As shown in Figure 15, from the perspective of MAE, the prediction effect of the LSTM model is better than that of the comparative model. Compared with the ANN model, the improvement in the LSTM model is more obvious, and the accuracy of most station models is improved by more than 20%. Compared with the RF model, the improvement in the LSTM model is not obvious, but the prediction effect is still better than the RF model. As shown in Figure 16, from the perspective of RMSE, the LSTM model still has better predictive effects than the ANN and RF models. Compared with the ANN model, the prediction effect of the LSTM model in the seven

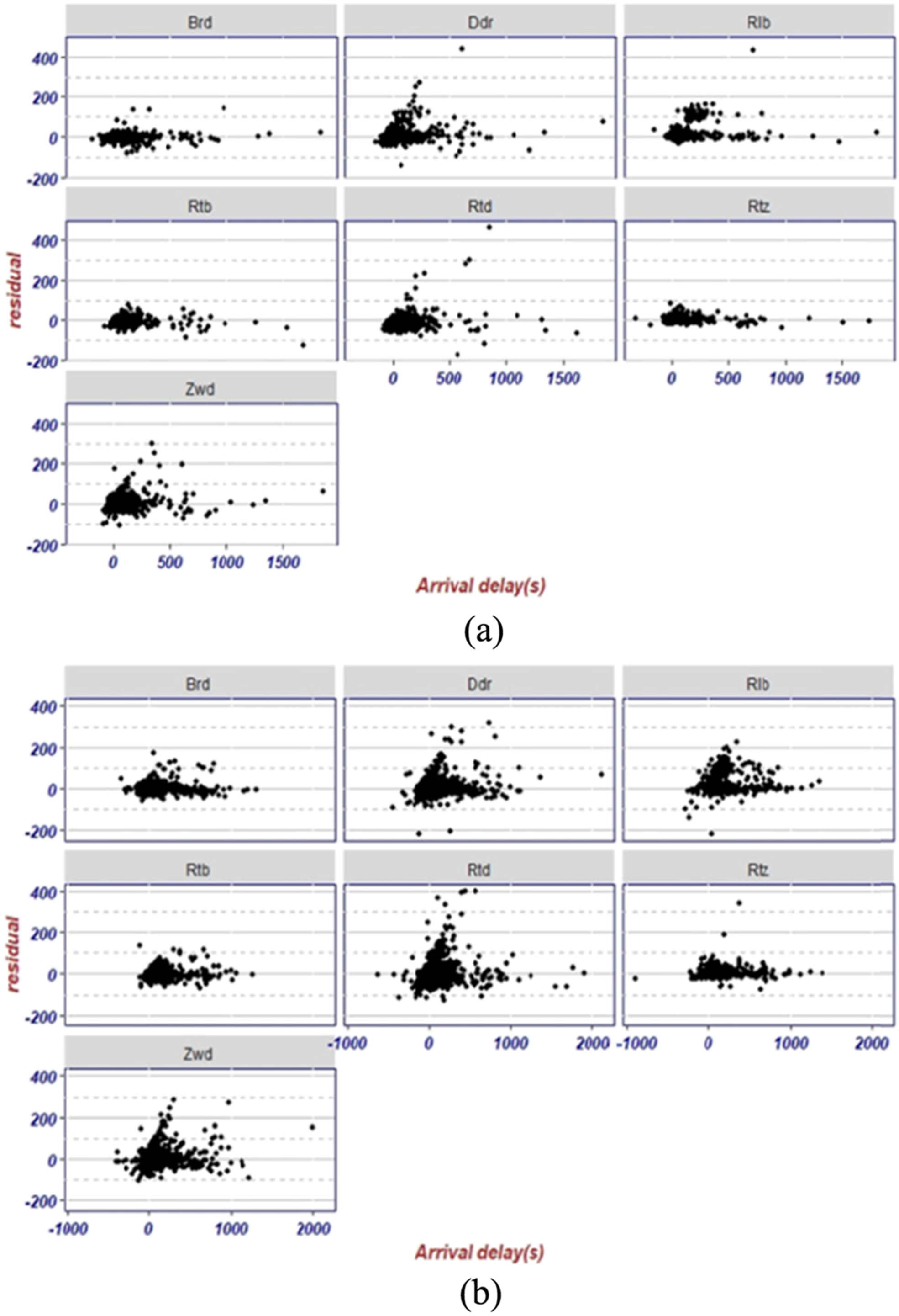


FIGURE 14 Distribution of the residuals of train arrival delays during (a) peak hours and (b) off-peak [Colour figure can be viewed at wileyonlinelibrary.com]

stations is greatly improved, and the average accuracy is improved by more than 20%. Similarly, compared with the RF model, the LSTM model still has a good prediction effect, and the station with the most obvious improvement effect is Rtb, and the accuracy is improved by more than 20%.

To verify the generalization and robustness of the LSTM model on different railway sections, we tested the model performance on another section: the Rilland Bath (Rb)-Vlissingen (Vs) section. This section consists of seven stations, namely Rilland Bath (Rb), Krabbendijke (Kbd), Kruiningen Yerseke (Krg), Goes (Gs), Arnemuiden (Arn), Middelburg (Mdb), and Vlissingen Souburg (Vss). There are fewer train types (IC and LM trains only) and the train frequency is lower. Similarly, RF and ANN were selected as the benchmark, and MAE and RMSE were used as evaluation indicators. The prediction results are shown in Figures 17 and 18.

In the Rb-Vs section, the prediction effect of the LSTM model is still better than the ANN and RF models.

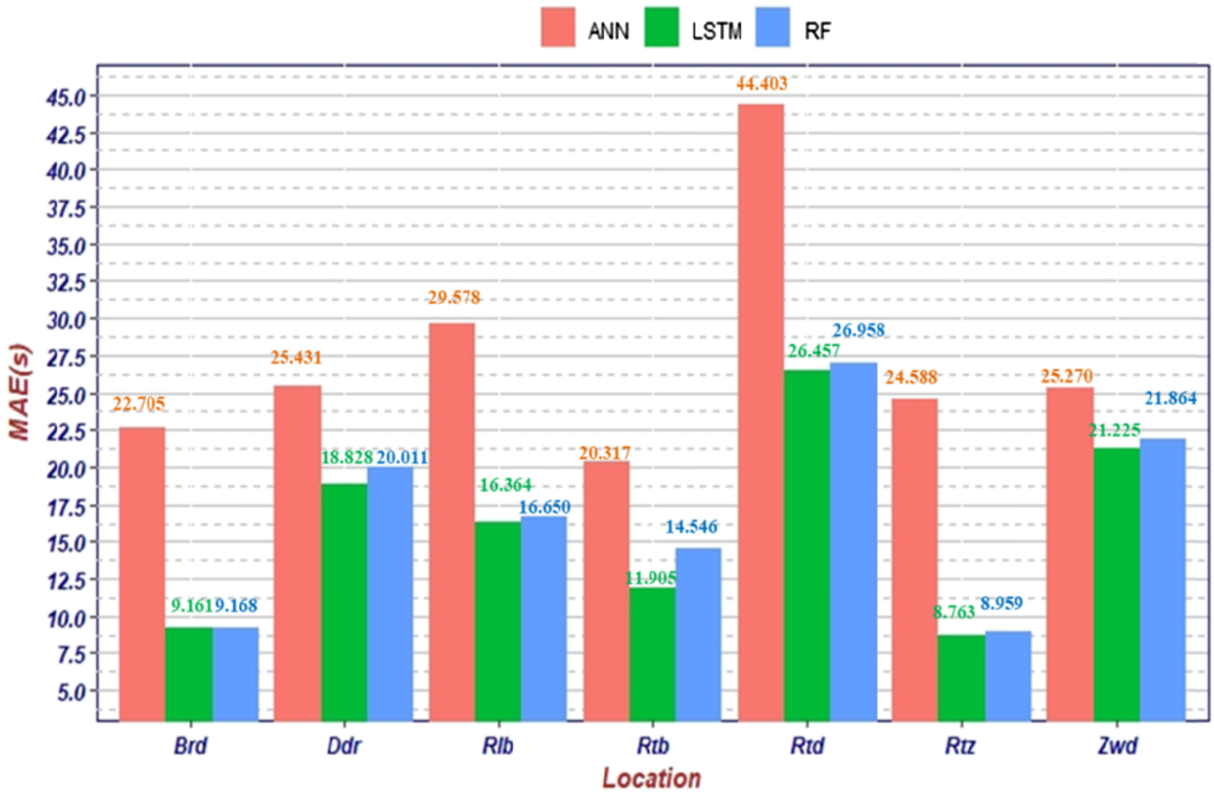


FIGURE 15 Comparison of MAE values at different stations [Colour figure can be viewed at wileyonlinelibrary.com]

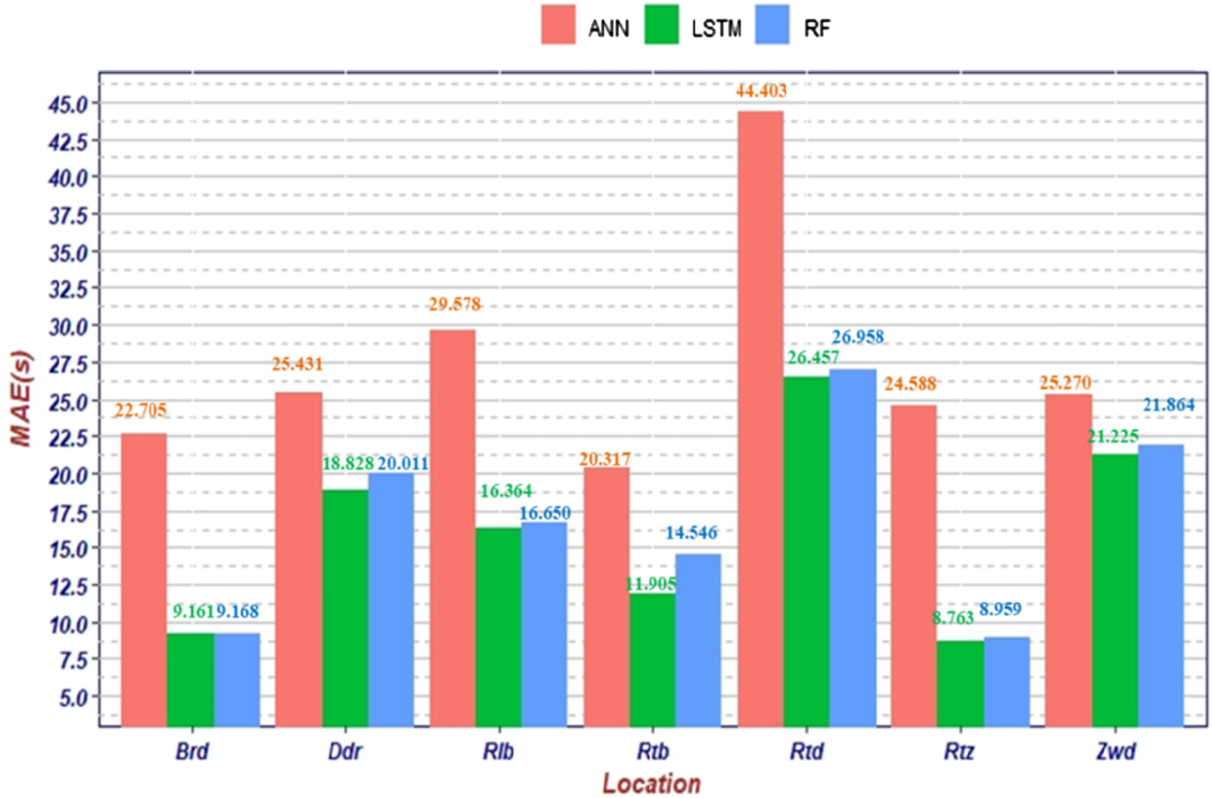


FIGURE 16 Comparison of RMSE values for different stations [Colour figure can be viewed at wileyonlinelibrary.com]

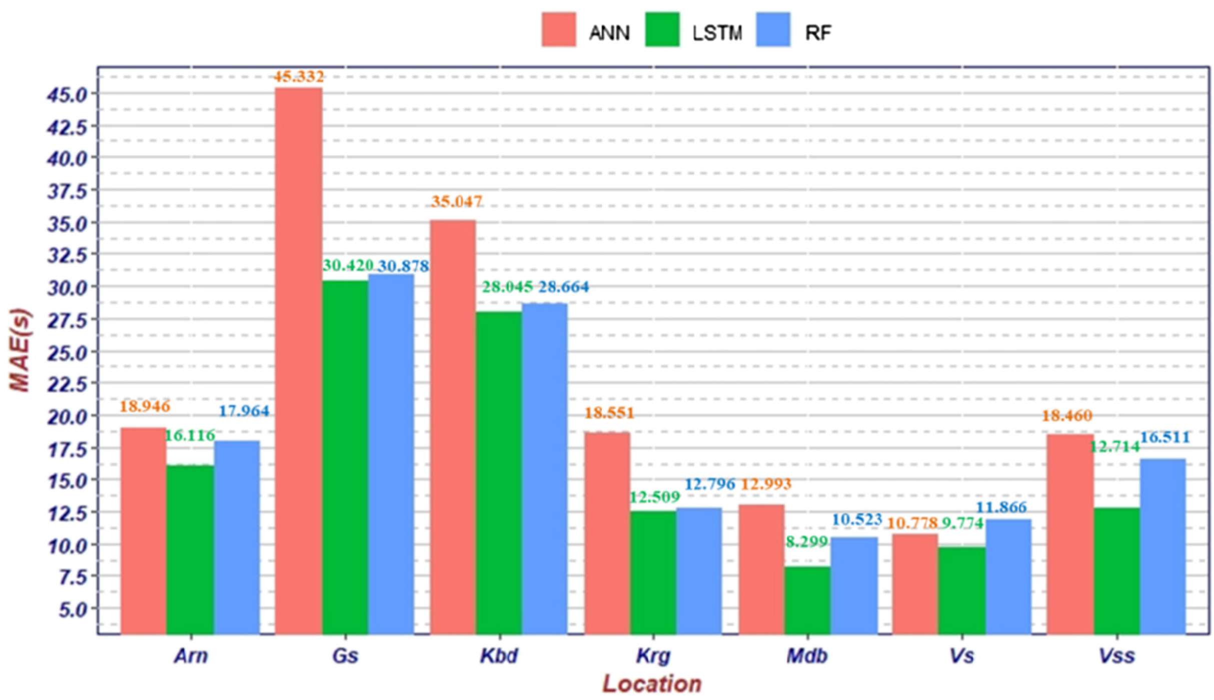


FIGURE 17 Comparison of MAE values at different stations (Rb-Vs) [Colour figure can be viewed at wileyonlinelibrary.com]

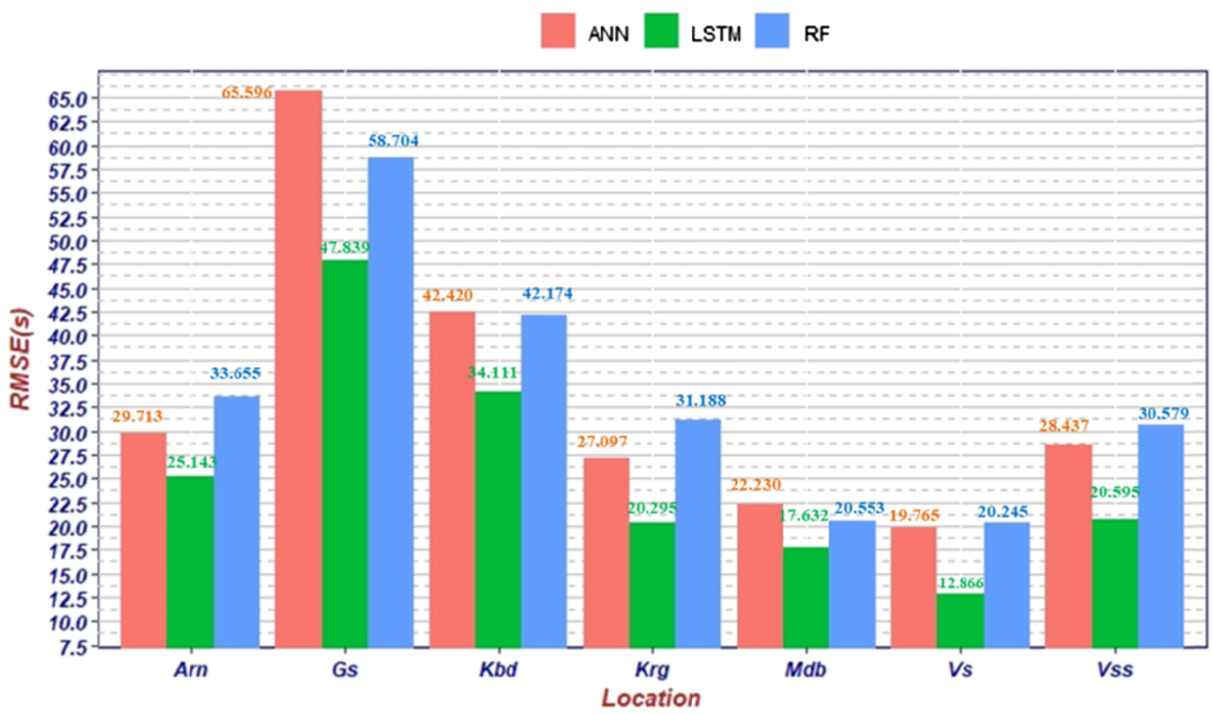


FIGURE 18 Comparison of RMSE values for different stations (Rb-Vs) [Colour figure can be viewed at wileyonlinelibrary.com]

Compared with ANN and RF models, the station with the most obvious improvement in LSTM prediction performance is the Vss station, and the prediction accuracy of the model is improved by more than 20%.

5 | CONCLUSIONS

This study presents a machine learning model to analyze the relationship between train arrival delays and various characteristics of a railway system; this is important for planning changes and investments to reduce delays. In the study, the LSTM model is used to construct a prediction model of train arrival delay, and the model is trained and tested based on the historical data of train operation. The results show that the LSTM model exhibits a better predictive effect than RF and ANN models. The performance of the LSTM model is superior, as indicated by the data validation results. Specifically, the LSTM model exhibits better MAE and MSE values, and its prediction accuracy reaches 86.91% within 30 s. In the verification section also, the LSTM model achieved better prediction results than the RF and neural network models.

The LSTM model presents a good measure of the propagation of train delays. This feature ensures that the model can be extended to other conventional railway routes. Additionally, the model exhibits two main advantages, as follows: (a) The model has good interpretability and high efficiency; (b) it includes interrelationships between the various delay factors and the superposition of arrival delays.

The proposed model in this study can be applied to other sections and stations for delay prediction. Currently, the model in the paper does not consider an excessive number of infrastructure factors. For further model expansion, it is possible to consider additional train delay influence factors and extract increasingly accurate feature variables to obtain better prediction results. Owing to data limitations, this paper does not distinguish between the primary delays and knock-on delays. In the future, we will attempt to distinguish between the primary and knock-on delays and integrate LSTM with other models to improve prediction results, and we will try to integrate the LSTM model with other models to improve the prediction accuracy.

ACKNOWLEDGMENTS

This work was supported by the National Nature Science Foundation of China (grant numbers 71871188 and U1834209), the Science and Technology Department of Sichuan Province (grant number 2018JY0567), the Doctoral Innovation Fund Program of Southwest Jiaotong University (grant number D-CX201827), and the China Scholarship Council. We are grateful for the contributions made by our project partners.

DATA AVAILABILITY STATEMENT

The data used in this study are from the 2018 Railroad Problem Solving Competition (https://connect.informs.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey = c09fe8bb-18da-40a6-a909-480c4a72a931).

ORCID

Chao Wen https://orcid.org/0000-0002-3933-2446
Zhongcan Li https://orcid.org/0000-0001-7123-5198

REFERENCES

Barta, J., Rizzoli, A. E., Salani, M., & Gambardella, L. M. (2012). Statistical modelling of delays in a rail freight transportation network. In *Proceedings of the* 2012 Winter Simulation conference, Berlin.

Bengio, Y., Simard, P., & Frasconi, P. (2002). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Berger, A., Gebhardt, A., Müller-Hannemann, M., & Ostrowski, M. (2011). Stochastic delay prediction in large train networks. In *OASIcs—OpenAccess Series in Informatics* (Vol. 20). Schloss Dagstuhl, Germany: Liebnitz Center for Informatics.

Chang, C., & Thia, B. (1996). Online rescheduling of mass rapid transit systems: Fuzzy expert system approach. *IEE Proceedings: Electric Power Applications*, *143*(4), 307–316.

Chen, D., Wang, L., & Li, L. (2015). Position computation models for high-speed train based on support vector machine approach. *Applied Soft Computing*, *30*(C), 758–766.

Corman, F., & Kecman, P. (2018). Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C: Emerging Technologies*, *95*, 599–615.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2004). Random forests. *Machine Learning*, *45*(1), 157–176.

Daamen, W., Goverde, R. M. P., & Hansen, I. A. (2009). Non-discriminatory automatic registration of knock-on train delays. *Networks and Spatial Economics*, *9*(1), 47–61.

Flier, H., Graffagnino, T., & Nunkesser, M. (2009). Finding robust train paths in dense corridors. In *IAROR RailZurich 2009*. Zurich, Switzerland: ETH.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, *3*(1), 115–143.

Gorman, M. F. (2009). Statistical estimation of railroad congestion delay. *Transportation Research, Part E*, *45*(3), 446–456.

Goverde, R. M. P. (2010). A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research, Part C*, *18*(3), 269–287.

Greff, K., Srivastava, R. K., & Koutník, J. (2016). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232.

Hochreiter, S., & Jürgen, S. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Huisman, T., Boucherie, R. J., & Van Dijk, N. M. (2002). A solvable queueing network model for railway networks and its validation and applications for the Netherlands. *European Journal of Operational Research*, *142*(1), 30–51.

Kecman, P., Corman, F., & Meng, L. (2015). Train delay evolution as a stochastic process. In *International Conference on Railway Operations Modelling and Analysis—RailTokyo 2015*, Chiba Institute of Technology, Narashino, Japan.

Kecman, P., & Goverde, R. M. P. (2015a). Online data-driven adaptive prediction of train event times. *IEEE Transactions on Intelligent Transportation Systems*, *16*(1), 465–474.

Kecman, P., & Goverde, R. M. P. (2015b). Predictive modelling of running and dwell times in railway traffic. *Public Transport*, *7*(3), 1–25.

Lessan, J., Fu, L., & Wen, C. (2018). A hybrid Bayesian network model for predicting delays in train operations. *Computers and Industrial Engineering*, *127*, 1214–1222.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Lulli, A., Oneto, L., Canepa, R., Petralli, S., & Anguita, D. (2018). Large-scale railway networks train movements: a dynamic, interpretable, and robust hybrid data analytics system. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 371–380). New York, NY: IEEE.

Malavasi, G. (2001). Simulation of stochastic elements in railway systems using self-learning processes. *European Journal of Operational Research*, *131*(2), 262–272.

Marković, N., Milinković, S., Tikhonov, K. S., & Schonfeld, P. (2015). Analyzing passenger train arrival delays with support vector regression. *Transportation Research, Part C*, *56*, 251–262.

Martin, L. J. (2016). Predictive reasoning and machine learning for the enhancement of reliability in railway systems. In *International Conference on Reliability, Safety and Security of Railway Systems* (pp. 178–188). Berlin, Germany: Springer.

Martin, L. J., & Romanovsky, A. (2016). A formal approach to designing reliable advisory systems. In *International Workshop on Software Engineering for Resilient Systems* (pp. 28–42). Berlin, Germany: Springer.

Medeossi, G., Longo, G., & Fabris, S. D. (2011). A method for using stochastic blocking times to improve timetable planning. *Journal of Rail Transport Planning and Management*, *1*(1), 1–13.

Meester, L. E., & Muns, S. (2007). Stochastic delay propagation in railway networks and phase-type distributions. *Transportation Research, Part B*, *41*(2), 218–230.

Olsson, N. O. E., & Haugland, H. (2004). Influencing factors on train punctuality: Results from some Norwegian studies. *Transport Policy*, *11*(4), 387–397.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., … Anguita, D. (2016). Advanced analytics for train delay prediction systems by including exogenous weather data. In *IEEE International Conference on Data Science and Advanced Analytics 2016* (pp. 458–467). New York, NY: IEEE.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., … Anguita, D. (2017a). Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout. *IEEE Transactions on Systems Man and Cybernetics Systems*, *47*(10), 2754–2767.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., … Anguita, D. (2017b). Train delay prediction systems: A big data analytics perspective. *Big Data Research*, *11*, 54–64.

Peters, J., Emig, B., Jung, M., & Schmidt, S. (2005). Prediction of delays in public transportation using neural networks. In *Computational Intelligence for Modelling, Control and Automation 2005* and *International Conference on Intelligent Agents, Web Technologies and Internet Commerce 2005* (pp. 92–97).

Şahin, İ. (2017). Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances. *Journal of Rail Transport Planning and Management*, *7*(3), 101–113.

Yaghini, M., Khoshraftar, M. M., & Seyedabadi, M. (2013). Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation*, *47*(3), 355–368.

Yuan, J., & Hansen, I. A. (2007). Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research, Part B*, *41*(2), 202–217.

Zilko, A., Hanea, A., Kurowicka, D., & Goverde, R. (2014). Non-parametric Bayesian network to forecast railway disruption lengths. In J. Pombo (Ed.), *Proceedings of the 2nd International Conference on Railway Technology: Research, Development and Maintenance*. Kippen, UK: Civil-Comp Press.

Zilko, A. A., Kurowicka, D., & Goverde, R. M. (2016). Modeling railway disruption lengths with copula Bayesian networks. *Transportation Research, Part C*, *68*, 350–368.

AUTHOR BIOGRAPHIES

Chao Wen received the B.S. degree in Transportation, and the Ph.D. degree in Transportation Planning and Management from Southwest Jiaotong University (SWJTU), Chengdu, China, in 2006 and 2012, respectively. He is currently an Assistant Professor with the School of Transportation and Logistics, SWJTU. He is also a Post-doctoral Fellow with the Department of Civil and Environmental Engineering at the University of Waterloo, Canada. He is the author of more than 50 articles. His research interests include railway time tabling and dispatching, and Railway capacity optimization.

Weiwei Mou received the B.S. degree in Transportation from Southwest Jiaotong University (SWJTU), Chengdu, China, in 2017. She is currently pursuing the M.S. degree in Transportation Planning and Management from the Department of Transportation & Logistics, SWJTU. Her research interests include train delay prediction, data mining, and deep Learning.

Ping Huang received the B.S. degree in Transportation from Southwest Jiaotong University (SWJTU), Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree in Transportation Planning and Management from the Department of Transportation & Logistics, SWJTU, and is also studying as a visiting student in the Department of Civiland Environmental Engineering at the Universityof Waterloo, Canada. His research interests include railway engineering, data mining, and deep Learning.

Zhongcan Li received the B.S. degree in Transportation from Southwest Jiaotong University (SWJTU), Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree in Transportation Planning and Management from the Department of Transportation & Logistics, SWJTU. His research interests include train delay prediction, data mining, and deep Learning.

How to cite this article: Wen C, Mou W, Huang P, Li Z. A predictive model of train delays on a railway line. *Journal of Forecasting*. 2020;39:470–488. https://doi.org/10.1002/for.2639