# Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on LSTM and MLP

*Weiwei Zhu*
Shanghai Municipal Road Transport Development Center, Shanghai, China

*Jinglin Wu*
Institute of Urban Risk Management, Tongji University, Shanghai, China, and

*Ting Fu, Junhua Wang, Jie Zhang and Qiangqiang Shangguan*
College of Transportation Engineering, Tongji University, Shanghai, China

## Abstract

**Purpose** – Efficient traffic incident management is needed to alleviate the negative impact of traffic incidents. Accurate and reliable estimation of traffic incident duration is of great importance for traffic incident management. Previous studies have proposed models for traffic incident duration prediction; however, most of these studies focus on the total duration and could not update prediction results in real-time. From a traveler's perspective, the relevant factor is the residual duration of the impact of the traffic incident. Besides, few (if any) studies have used dynamic traffic flow parameters in the prediction models. This paper aims to propose a framework to fill these gaps.

**Design/methodology/approach** – This paper proposes a framework based on the multi-layer perception (MLP) and long short-term memory (LSTM) model. The proposed methodology integrates traffic incident-related factors and real-time traffic flow parameters to predict the residual traffic incident duration. To validate the effectiveness of the framework, traffic incident data and traffic flow data from Shanghai Zhonghuan Expressway are used for modeling training and testing.

**Findings** – Results show that the model with 30-min time window and taking both traffic volume and speed as inputs performed best. The area under the curve values exceed 0.85 and the prediction accuracies exceed 0.75. These indicators demonstrated that the model is appropriate for this study context. The model provides new insights into traffic incident duration prediction.

**Research limitations/implications** – The incident samples applied by this study might not be enough and the variables are not abundant. The number of injuries and casualties, more detailed description of the incident location and other variables are expected to be used to characterize the traffic incident comprehensively. The framework needs to be further validated through a sufficiently large number of variables and locations.

**Practical implications** – The framework can help reduce the impacts of incidents on the safety of efficiency of road traffic once implemented in intelligent transport system and traffic management systems in future practical applications.

**Originality/value** – This study uses two artificial neural network methods, MLP and LSTM, to establish a framework aiming at providing accurate and time-efficient information on traffic incident duration in the future for transportation operators and travelers. This study will contribute to the deployment of emergency management and urban traffic navigation planning.

**Keywords** Prediction of traffic incident duration, Long short-term memory, Multi-layer perception, Deep learning

**Paper type** Research paper

## 1. Introduction

Traffic incidents cause casualties, direct economic losses and traffic congestions which have been studied over the years (Adler *et al.*, 2013; Hojati *et al.*, 2016). For instance, Skabardonis *et al.* (2003) found out that in California, about 72% of non-recurrent congestions and 13%–30% of traffic delays in peak hours are associated with traffic incidents. In addition, traffic incidents lead to a high probability of

secondary incidents. The risk of a second incident has been estimated to be six times greater than that of a primary incident (Wang *et al.*, 2019). These factors highlight the importance of implementing proper and timely countermeasures for traffic incidents such as traffic flow control and incident response resource allocation (Haule *et al.*, 2019).

Accurate and time-efficient prediction methods for the traffic incident duration are required for formulating and implementing traffic incident countermeasures (Chung, 2010). Traffic incident duration is defined as the period between the time when an incident occurs to the time when traffic recovers to normal (Hojati *et al.*, 2014; Highway Capacity Manual, 2010). Large number of studies have been devoted to the prediction of traffic incident duration. Dimitriou and Vlahogianni (2015) proposed a fuzzy rule-based system to estimate highway traffic incident durations. Lin *et al.* (2016) proposed an improved M5P model by combining a hazard-based duration model to minimize data heterogeneity in traffic incident duration prediction.

However, most studies used the total incident duration as the prediction object, which means that the prediction results are given at the time when the incident occurred and will not update over time. In fact, the residual incident duration, i.e. how long the negative impact of the incident will continue in the future, has the most practical application value. Given a real-time prediction, transportation operators could adjust and optimize the countermeasures for traffic incidents (Adler *et al.*, 2013). Travelers on the way could also decide whether to choose an alternate route.

In addition, few (if any) studies have taken into account the dynamic traffic flow parameters. Dynamic traffic flow parameters refer to the real-time temporal sequences of traffic parameters including speed, traffic volume and time occupancy during the duration of the traffic incident. It is generally accepted that the real-time traffic flow status has an effect on traffic incident duration (Ma *et al.*, 2017; Ru *et al.*, 2017).

Given the research gaps in the real-time prediction and the employment of dynamic traffic flow parameters, this study uses two artificial neural networks (ANN) methods, multi-layer perception (MLP) and long short-term memory (LSTM), to establish a framework aiming at providing accurate and time-efficient information on traffic incident duration in the future for transportation operators and travelers. This study will contribute to the deployment of emergency management and urban traffic navigation planning.

## 2. Literature review

Different statistical methods and machine learning methods have been applied in traffic incident duration prediction, including tree-based method (Lin *et al.*, 2016; Weng *et al.*, 2015), Bayesian classifier (Cong *et al.*, 2018; Ozbay and Noyan, 2006; Zou *et al.*, 2021), hazard-based method (Haule *et al.*, 2019; Li *et al.*, 2017; Li *et al.*, 2015) and ANN (Lee *et al.*, 2017). Among these methods, the accelerated failure time (AFT) model is the most widely used hazard-based method (Li *et al.*, 2017; Li *et al.*, 2015). It assumes that the factors related to the incident will accelerate or decelerate the incident duration; thus, it is easily interpreted (Kay and Kinnersley, 2002). Chung (2010) has established a log-logistic AFT metric model based on the Korean Freeway accident data and the

results yielded a reasonable prediction. Hojati *et al.* (2013) proposed a series of parametric AFT survival models of incident duration based on three common distributions, i.e. log-logistic, lognormal and Weibull. Zou *et al.* (2021) proposed a Bayesian Model Averaging model to predict traffic incident clearance time. Besides, decision tree models were also widely applied to predict incident clearance time due to it can determine the importance of explanatory variables (Weng *et al.*, 2019). Ma *et al.* (2017) developed a novel approach, gradient boosting decision trees, to predict incident clearance time using different traffic parameters.

However, the structures of these machine learning methods are limited in applying temporal sequences and dynamic traffic flow parameters are not suitable as inputs. Previous studies have used traffic status and other variables in describing traffic flow conditions as substitutes. Lin *et al.* (2016) applied "congestion or not" as a variable of the prediction model for traffic incident duration. Zou *et al.* (2018) adopted "peak hour or not" along with 13 other variables to investigate the dependence between incident clearance and response time. Some studies also used traffic flow characteristics. For example, Ghosh *et al.* (2014) applied the "85th/15th percentile speed" and "peak hour volume" to examine the impact of influence factors on the clearance time of incidents. Hojati *et al.* (2016) considered posted speed limit, road capacity, recurrent flow, the ratio of speed before and after the incident, etc., as variables to model travel time reliability. However, as traffic incidents are generally a sustaining process and traffic flow status changes over time, the application of stationary traffic flow parameters cannot provide an accurate picture of the situation. Therefore, a new method integrated with dynamic traffic flow information is required.

In recent years, ANN methods have been shown to perform well in short- and long-term forecasting applications with steady data-driven capabilities (Liu *et al.*, 2019). LSTM neural network is one of ANN and performs well in dealing with temporal sequences. LSTM has been widely used in short-term traffic flow forecasting and could be used as a reference in processing dynamic traffic flow parameters (Polson and Sokolov, 2017). Short-term traffic flow forecasting applies existing traffic flow data to continuously predict the traffic flow and travel time for a period in the future (usually within 15 min). Ma *et al.* (2015) used LSTM to predict the speed of the next 2 min by applying data from microwave traffic detectors and the mean absolute percentage error of the model applying both speed and volume as inputs is under 5%. Gu *et al.* (2019) proposed a two-layer deep learning framework based on LSTM and gated recurrent unit neural network to predict lane-level traffic speed. The detailed methodological review for clearance time prediction of road incidents can refer to Tang *et al.* (2020). Nevertheless, LSTM neural network has not yet been used to predict the incident duration, which also motivated this study.

The LSTM model provides new insights on using real-time traffic flow parameters. Moreover, traffic incident-related factors should be taken into account, which calls for a hybrid model in addition to the LSTM. MLP is a standard ANN model and it is capable of dealing with classification problems. In this study, the LSTM and MLP methods are combined to build a framework to use two forms of variables as inputs to predict the short-term traffic incident duration.

**Figure 1** Incident duration decomposition
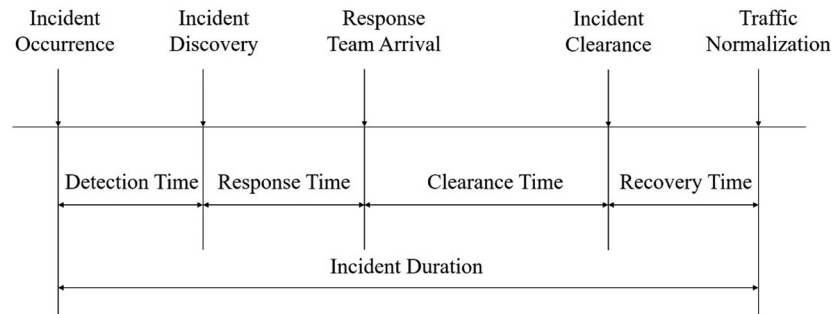


# 3. Methodology

The methodology section provides details about concepts and definitions; key parameters and factors and deep learning modeling approach.

## 3.1 Concepts and definitions

### 3.1.1 Decomposition of traffic incident
The process of an incident is generally divided into four parts: detection time (time duration between incident occurrence and incident discovery), the response time (time duration between incident discovery and response team arrival), clearance time (time duration between response team arrival and incident clearance) and recovery time (time duration between incident clearance and traffic normalization) (Zhang *et al.*, 2019; Highway Capacity Manual, 2010), as shown in Figure 1. To provide timely decision support to road traffic managers and drivers, this study proposes a real-time method to predict the total traffic incident duration. This study uses the velocity thermogram method, which compares the velocity of the vehicle under the influence of the accident with the historical average speed of a certain road section to determine the impact range of an accident. The detailed process of how to obtain the accurate incident duration can refer to Zhang *et al.* (2019).

### 3.1.2 Dynamic prediction for traffic incident duration
The traffic incident management handbook suggested that systems with recorded information should be updated at least every 5 to 10 min during peak periods in urban areas (Farradyne, 2000). This model could provide the information dynamically every 1 min. The objective of this prediction model is the residual incident duration at each update moment. The residual incident duration in this study was classified into three categories based on two levels of incident duration. The two incident levels are 5 and 10 min. The reason for choosing 5 and 10 min as incident duration levels are as follows:
- The levels of 5 and 10 min are more practical. In urban expressways, traffic incident duration is mostly short. In the data set this study used, 35% is less than 5 min while 70% is less than 15 min. Too long intervals may lose practical significance.
- The levels of 5 and 10 min are sufficient for upstream drivers and road managers to make decisions and preparations. Speed limit in urban expressways is usually 80 km/h and cars can drive more than 5 kilometers in 5 min, more than 10 kilometers in 10 min. This distance is

enough for the driver to leave the expressway at the previous exit and choose another route.

## 3.2 Key parameters and factors

### 3.2.1 Traffic flow parameters
In the four parts of the incident duration, traffic flow status is changing constantly. During the detection time and response time, the crash vehicles may occupy one or more lanes and decreasing traffic capacity, which will cause a decline in upstream traffic speed and traffic flow. During the clearance time, the impact on traffic congestion will change according to the processing methods. If medical response, police or towing cars are required, the congestion may get worse. After the incident is cleared, traffic congestion will gradually ease. Consequently, dynamic traffic flow parameters will reflect the process of traffic incident handling. This study uses dynamic traffic flow parameters as the inputs. Comparison is provided in the case study section.

Referring to the previous studies on traffic incident risk estimation with traffic flow conditions (Oh *et al.*, 2005; Fang *et al.*, 2016), the traffic flow parameters (i.e. traffic flow and traffic speed) 5–30 min before the traffic incident to the end of traffic impact have been applied in this study.

### 3.2.2 Incident-related factors
Previous studies indicated that incident duration has a strong correlation with incident characteristics, e.g. incident type and incident severity (Adler *et al.*, 2013). Traffic incident-related factors could be categorized as temporal factors, spatial factors, environmental factors, incident detail factors and operational factors (Abouaïssa *et al.*, 2016). These factors are summarized in Table 1.

## 3.3 Deep learning modeling approach based on long short-term memory and multi-layer perception network
As shown in Figure 2, the proposed framework contains four parts: data processing, LSTM neural network for incident clearance prediction and MLP network for incident clearance prediction. Two kinds of variables, traffic-related factors and traffic flow data, are applied as inputs of the framework to predict the short-term traffic incident duration. Each part of the framework will be detailed below.

### 3.3.1 Data processing
Traffic flow data and incident-related factors are two parts of inputs of the framework. They should be processed first before

**Table 1** Incident-related factors and explanations

| Categories | Variables | Values |
|---|---|---|
| **Temporal factors** | Time of the day[a] | 0:00–23:59 |
| | Day of the week | Monday to Sunday |
| | Season | Spring, summer, autumn and winter |
| **Spatial factors** | Location of the incident | – |
| | Type of road | Tunnels, elevated roads and general roads |
| | Distance from CBD | (Continuous variable) |
| **Environmental factors** | Weather | Sunny, rainy, foggy and snowy |
| | Visibililty | 0–10 km |
| | Traffic condition | Congested or not |
| | Pavement condition | Dry or wet |
| **Incident details** | Incident type | – |
| | Occupied lanes[b] | 1, 2, 3 or more and road closed |
| | The number of involved cars | 1, 2, 3 or more |
| | Injuries and deaths | None, 1, 2, 3 or more |
| | Incident number on the same road in one day | None, 1, 2, 3 or more |
| | Secondary incidents | Yes, no |
| | Involvement of a bike | Yes, no |
| | Work zone involved | Yes, no |
| | Incident on HOV lane | Yes, no |
| | Incident severity[c] | A1, A2 and A3 |
| | Shoulder availability | Yes, no |
| **Operational factors** | Police requirement | Yes, no |
| | Towing cars | Yes, no |
| | Medical response | Yes, no |
| | Patrol involved | Yes, no |
| | Traffic control | – |
| | Moving to shoulder | Yes, no |

**Notes:** [a]Similar expression: peak hour or not, daytime or night; [b]Similar expression: capacity reduction; [c]A1: People died during an accident, A2: People injured during an accident or died after an accident and A3: Property damage; HOV = high occupancy vehicle

being imported into the model. Data processing contains two parts: slide the time window to get more samples and data normalization.

*3.3.1.1 Slide the time window to get more samples.* The input variables of an incident are $\{V_1, V_2, V_3, \ldots, V_n, Vol, Spd, Y\}$, where $\{V_1, V_2, V_3, \ldots, V_n\}$ denote incident related factors, $Vol$ and $Spd$ denote traffic volume data and speed data (i.e. Traffic flow data), $Y$ denotes traffic incident duration. In Figure 2, at each updated moment, the length of traffic flow data is fixed, i.e. the length of the time window. The recommended time window length is 5–30 min. Figure 3 shows the process of sliding the time window when the time window length is 5 min.

For each incident, the time window sequences $Vol_{tw1}$, $Vol_{tw2}$, $\ldots$, $Vol_{twm}$ are paired with the same incident-related factors $V_1$, $V_2$, $V_3$, $\ldots$, $V_n$ and incident duration $Y$. $m$ in $Vol_{twm}$ indicates the number of time windows corresponding to the incident. After sliding the time window, these two variables should be added or updated:

- Update incident duration $Y$. At the update moment, the total incident duration $Y$ should be replaced by the residual incident duration $Y'$. As mentioned above, two forms of output results are proposed: **Label 1** – whether $Y'$ is greater than 5 min. Categorical variables $y_{label1} = \{0,1\}$, where $0 - Y' \leq 5\ min$ and $1 - Y' > 5\ min$. **Label 2** – whether $Y'$ is greater than 10 min. Categorical variables $y_{label2} = \{0,1\}$, where $0 - Y' \leq 10\ min$ and $1 - Y' > 10\ min$.

- Add a new variable elapsed time ($V_{n+1}$) and update it at each prediction moment. Elapsed time means how long the incident has lasted. It is a continuous variable and the unit is minutes.

In summary, an incident sample should contain the following variables:

$$\{V_1, V_2, V_3, \ldots, V_{n+1}; Vol_{tw1}; Spd_{tw1}; Y'\},$$
$$\{V_1, V_2, V_3, \ldots, V_{n+1}; Vol_{tw2}; Spd_{tw2}; Y'\},$$
$$\ldots$$
$$\{V_1, V_2, V_3, \ldots, V_{n+1}; Vol_{twm}; Spd_{twm}; Y'\}$$

*3.3.1.2 Data normalization.* In the multi-factor evaluation system, each factor may have different magnitudes or units. The impact of the factor will be enlarged or minified depending on the value of the factor if it is used for analysis directly. Therefore, to ensure the reliability of the results and the equal contribution of each factor, the raw data need to be normalized before being input onto the framework. This study performs min-max normalization on the raw data to make the result fall into the interval [0,1]. The transformation function is as follows:

$$x^* = \frac{x - min}{max - min} \tag{1}$$

where $max$ is the maximum value and $min$ is the minimum value of the sample data.

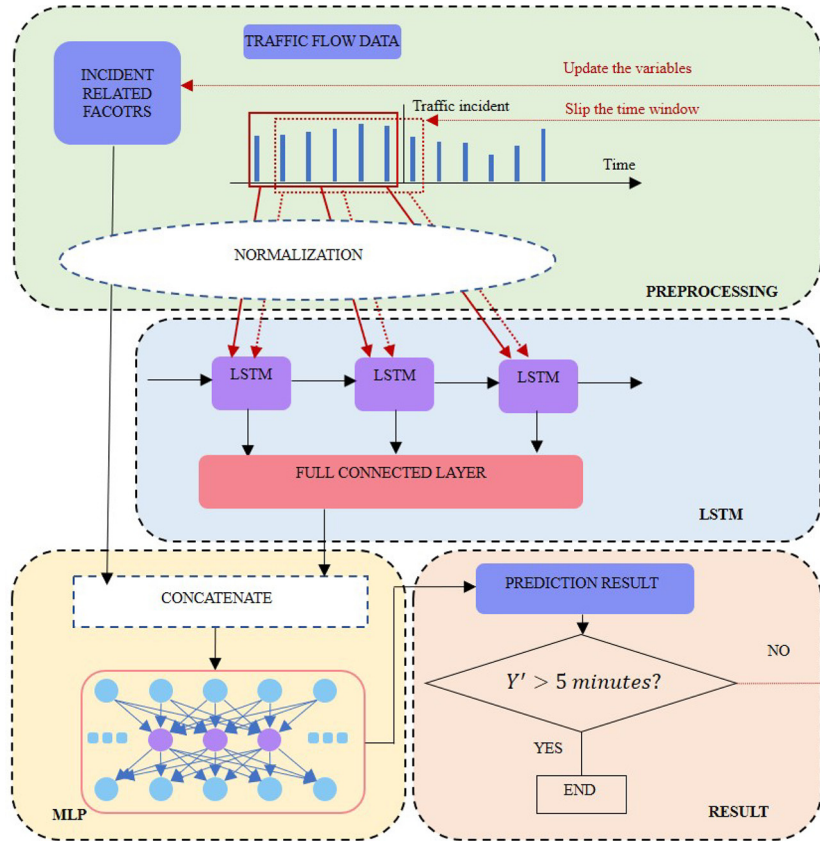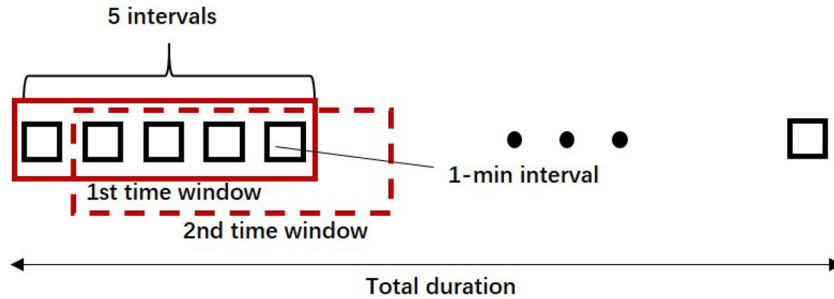**Figure 2** Overview of the proposed framework



**Figure 3** The process of sliding the time window



*3.3.2 Long short-term memory neural network for incident clearance prediction*

LSTM is a powerful type of artificial recurrent neural network (RNN), which is good at dealing with sequential data (Song *et al.*, 2020). As shown in, Figure 4, similar to the traditional MLP, RNN consists of an input layer, hidden layer and output layer. The hidden layers of RNN are more like a block. For each block, the output $h_t$ is calculated with both $x_t$ (input of the model at time $t$) and $h_{t-1}$ (the result of the memory cell at the last time $t-1$):

$$h_t = \varnothing \ (Ux_t + Wh_{t-1} + b) \tag{2}$$

where $U$ and $W$ represent weight coefficients, $\varnothing$ represents activation function and $b$ represents bias.

The output at time $t$ is:

$$o_t = Vh_t + c \tag{3}$$

where $V$ represents weight coefficient and $c$ represents bias.

The output of the model is:

$$y_t = \sigma(o_t) \tag{4}$$

where $\sigma$ represents activation function.

Based on the memory of the previous learning content, RNN is mostly used for machine translation and speech recognition (Su *et al.*, 2019). However, during the training of the RNN model, the information at all times before will be traced back when calculating the partial derivative of the loss function to the
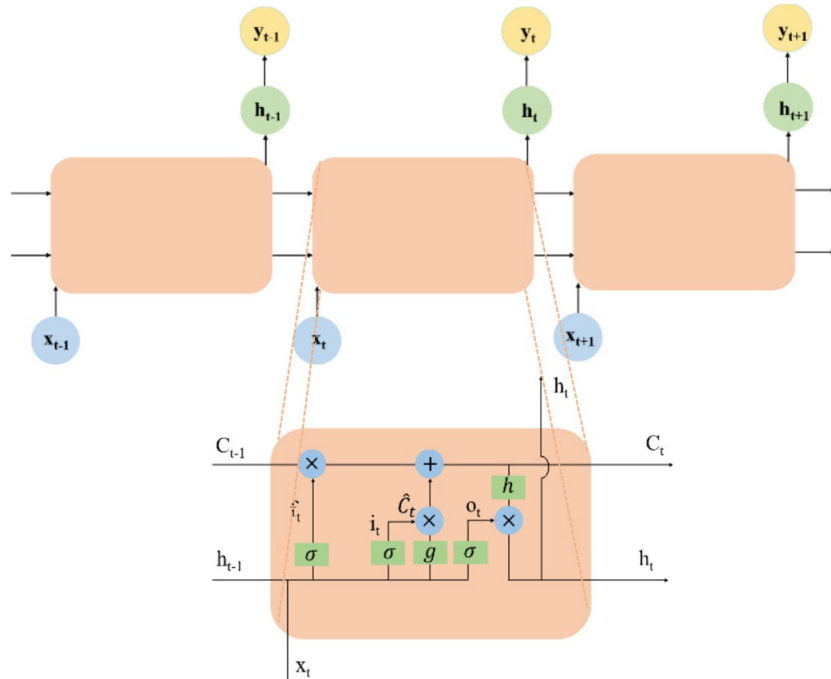
**Figure 4** Structure of a RNN
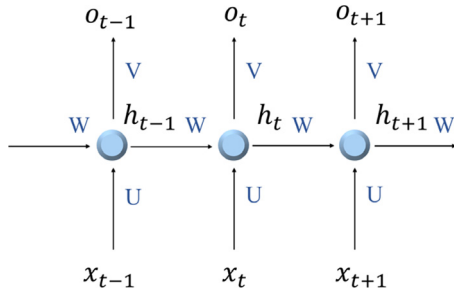


weight coefficient, which leads to the continuous multiplication of the derivative of the activation function. The continuous multiplication will cause the gradient to be too large (named "gradient explosion") or too small (named "gradient disappearance"), therefore the model learning efficiency is unstable or the information may be weakened. Aiming at solving this problem, Hochreiter and Schmidhuber (1997) proposed the LSTM which could overcome long-term dependencies and determine the best time window automatically. As shown in Figure 5, an LSTM neural network consists of one input layer, one hidden layer and one output layer. In the hidden layer, different from the RNN, there are three "gates" in each memory cell, namely, "forget gate," "input gate" and "output gate." The gate can control whether the previous status information passes and affects the subsequent predictions.

- The first gate is the "forget gate" which decides whether the information will be discarded. It reads $h_{t-1}$ and $x_t$ and outputs a value between 0 and 1, where 1 means "completely reserved" and 0 means "completely discarded."

$$f_t = \sigma\left(W_f[\ h_{t-1},\ x_t] + \ b_f\right) \tag{5}$$

- The next step is to determine what new information is stored in the cell state, which will be calculated in the "input gate" by the following functions:

$$i_t = \sigma(W_i[\ h_{t-1},\ x_t] + \ b_i) \tag{6}$$

$$C_t = f_t \ \times \ C_{t-1} + i_t \times g(W_c[\ h_{t-1},\ x_t] + b_C) \tag{7}$$

- Finally, the output value will be calculated in "output gate" by the following functions:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_{t=}o_t \times h(C_t) \tag{9}$$

In the output layer, the output value is calculated as:

$$y_t = W_{ym}\ h_t + b_y \tag{10}$$

where $\sigma$, g and $h$ represent the activation function, $b_f$, $b_i$, $b_C$, $b_o$, $b_y$ representing the bias, $W_f$, $W_i$, $W_c$, $W_o$, $W_{ym}$ represent the weight.

Traffic flow data are input to the part of LSTM after normalization. The LSTM part is implemented by Python's Keras Library. The construction of this part involves the configuration of the following parameters:

- Input_shape. The input dimensions and lengths of the model, determined by the time window length and the number of categories for traffic flow parameters (traffic volume, speed or both volume and speed). For example, if the time window length is 30 min and both two traffic parameters are the inputs, the input_shape format is 30 × 2.

**Figure 5** Structure of a LSTM network

- Units. The number of hidden layer neural nodes in the LSTM neural network. It is determined according to the time window length.

The LSTM layer is connected to the DENSE layer. The DENSE layer is the name of the fully connected layer in Python's Keras Library. All neurons in the fully connected layer are connected to each other and have a directional relationship. Configuration of two parameters in this layer are set as follows:
- Units. This is the output dimension of this part, it is set to 1.
- Activation. The activation function of this layer. It is Relu function here:

$$\sigma(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \tag{11}$$

### 3.3.3 Multi-Layer perception network for incident clearance prediction

MLP is a well-known method of ANN. As shown in Figure 6, MLP consists of one input layer, one or more hidden layers and one output layer. Neurons in each layer are fully connected to the next layer. The input of the model is denoted as $x = \{x_1, x_2, x_3, \ldots, x_i\}$. Each node in the hidden layer ($l$) is denoted as $a^l = \{a_1^l, a_2^l, a_3^l, \ldots, 0Ca_m^l\}$ and can be calculated as follows:
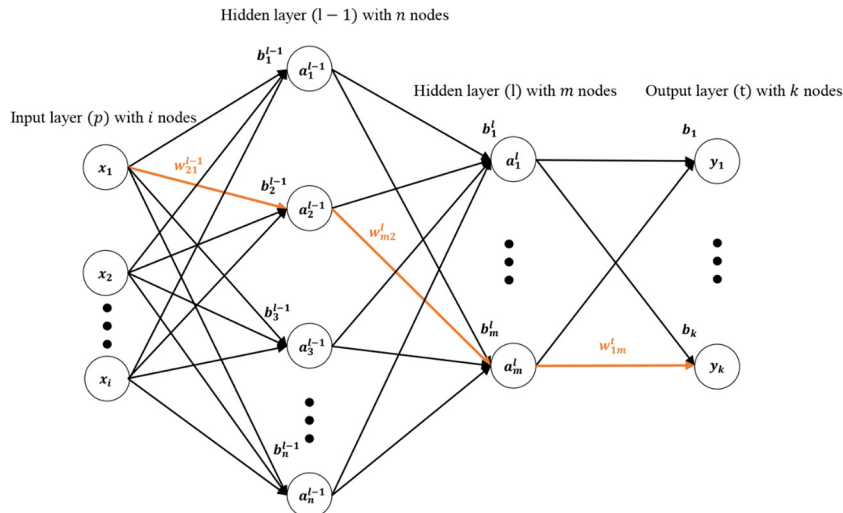
$$a_m^l = \sigma\left(\sum_j^n w_m^l a_j^{l-1} + b_m^l\right) \tag{12}$$

where $w_{mi}^l$ represents the weight coefficient between node $a_i^{l-1}$ and node $a_m^l$. Furthermore, $b_m^l$ represents the bias of the node $a_m^l$ and $\sigma$ represents the activation function.

The output of the model is denoted as $y = \{y_1, y_2, y_3, \ldots, y_k\}$ and can be calculated as:

$$y_k = \varphi\left(\sum_j^m w_{km}^t a_j^l + b_k\right) \tag{13}$$

where $\varphi$ represents the activation function.

Training MLP is based on back propagation by adjusting weight coefficients and bias according to the error gradient descent method. The expression for updating the weight coefficient is as follows:

$$w_i : w_i \ - \alpha \frac{\partial}{\partial w_i} L(W) \tag{14}$$

where $\alpha$ indicates the learning rate, $L(W)$ indicates loss function and $\frac{\partial}{\partial w_i} L(W)$ indicates the partial derivative of the loss function $L(W)$ for the biased $w_i$. Loss function is to evaluate the model prediction results and the target of model training is to get the minimum loss value $L_{min}(W)$. For binary classification problems, the most common loss function is binary cross-entropy:

$$L(W) = -\sum_{i=1}^q y_i log Z_i + (1 - y_i) \log(1 - y_i) \tag{15}$$

where $Z_i$ indicates the true value, $y_i$ indicates the prediction value of the model and $q$ indicates the number of samples.

The output of the LSTM part and the normalized incident-related factors are integrated into this part first. The integration is implemented by function *keras.concatenate* in Python. Then the integrated variables are input into the part of the MLP network of two hidden layers.

For determining the clearance of the incident: if the result of from modeling is $Y' > 5\,min$, slide the time window and get the traffic flow data of the next time window into the next cycle; if $Y' \leq 5\,min$, the clearance of the incident will happen in the coming 5 mins, then the program will be terminated.

In the case study section, the model will be tested and validated. The false-positive rate (FPR) and true positive rate (TPR) are often used to comprehensively evaluate the ability of the prediction model. The receiver operating characteristic (ROC) curve and Kolmogorov-Smirnov (KS) curve is created by plotting the TPR against the FPR at various threshold

**Figure 6** Multi-layer perceptron flowchart

settings. Besides, the area under the ROC curve (AUC), which provides an aggregate measure of performance across all possible classification thresholds, was used in this study to evaluate the prediction performance.

## 4. Case study

### 4.1 Description of sites and data

This study selected the traffic incident data and traffic flow parameters in Shanghai Zhonghuan Expressway as a case study. The total length is 70 km approximately with a speed limit of 80 km/h. The expressway is two-directional with four lanes in each direction. Different data sources are used for incident-related factors and traffic flow information. Incident-related factors are from the Road Network Monitoring Center, while Traffic flow information was collected from the inductive loop detectors along the Zhonghuan Expressway.

A total number of 4,041 indecent records were originally collected, covering the period from April 1, 2017 to October 7, 2017. Table 2 presents detailed descriptions of the indicent-related variables collected in the case study. Traffic flow

information was obtained from 176 inductive loop detector sets (loops at the same detection spot are considered as one set) distributed along the Shanghai Zhonghuan Expressway, with 800 m intervals on average. These loop data contain both speed and traffic volume information with an acquisition frequency of 20 s. To be effectively applied in the modeling framework, traffic data was converted into 1-min intervals by aggregating the traffic volume and averaging the speed. Data cleansing and matching work was conducted to remove vague traffic incident records and to pair up the incidents with the associated traffic flow information. After data cleansing and data matching, 391 incidents with their traffic flow information paired were selected. Note that no secondary incidents were observed, therefore they were not considered in this study.

### 4.2 Data processing

*4.2.1 Comparison groups as modeling input*

To determine the best input parameter combination, two comparative tests were conducted: test with different traffic flow parameter combinations: traffic volume, speed, both traffic

**Table 2** Descriptions of incident related variables

| Variable | Description | Codes-values | No. of incidents | Average duration (minutes) |
|---|---|---|---|---|
| Day of the week | Day of the incident | 1-Monday | 58 | 12.7 |
| | | 2-Tuesday | 44 | 12.8 |
| | | 3-Wednesday | 46 | 13.1 |
| | | 4-Thursday | 48 | 13 |
| | | 5-Friday | 74 | 12.9 |
| | | 6-Saturday | 75 | 13 |
| | | 7-Sunday | 46 | 12.9 |
| Hour of the day | Hour of the incident | 0-early morning-00:00∼07:00 | 5 | 13 |
| | | 1-morning peak-7:00∼9:00 | 56 | 13.9 |
| | | 2-morning-09:00∼11:00 | 60 | 12.8 |
| | | 3-noon-11:00∼13:00 | 60 | 12.8 |
| | | 4-afternoon-13:00∼17:00 | 129 | 12.7 |
| | | 5-evening peak-17:00∼19:00 | 48 | 12.8 |
| | | 6-evening-19:00∼24:00 | 33 | 12.8 |
| Weather condition | Weather | 0-sunny | 344 | 12.8 |
| | | 1-rainy | 47 | 12.8, |
| Traffic condition | Traffic condition | 0-unblocked | 187 | 12.9 |
| | | 1-slow passage | 200 | 12.7 |
| | | 2-congestion | 4 | 12.4 |
| Location type | Type of the incident location | 0-ramp | 124 | 12.8 |
| | | 1-roadside | 54 | 12.7 |
| | | 2-middle of the road | 53 | 12.8 |
| | | 3-unknown | 160 | 12.7 |
| Incident type | Incident category | 0-multi vehicle | 48 | 12.8 |
| | | 1-dual vehicle | 226 | 12.7 |
| | | 2-broke down | 113 | 12.8 |
| | | 3-wrong operation | 1 | 8 |
| | | 4-fire | 3 | 12.9 |
| Occupied lanes | Number of lanes occupied by the incident | 0-not occupying lanes | 3 | 7 |
| | | 1- occupying one lane | 365 | 12.8 |
| | | 2- occupying two lanes | 22 | 12.8 |
| | | 3- occupying three lanes | 0 | – |
| | | 4- occupying four lanes | 1 | 2 |
| Police presence | Whether the police are required | 0-no | 308 | 12.7 |
| | | 1-yes | 83 | 12.8 |

volume and speed; test with different time windows lengths: T = (5 min, 10 min, 20 min, 30 min). Meanwhile, to verify the effects of the dynamic traffic flow parameter, a comparative test between the application of dynamic traffic flow parameters and the application of static traffic flow parameters after processed by principal component analysis (PCA) is conducted.

*4.2.2 Preparation for modeling*
With the time window sliding procedure, approximately 4,200 samples are obtained. During this procedure, as the status of a small number of loop detectors was missing or invalid, corresponding samples were deleted. After expanding the sample through the time window sliding procedure, the sample ratio of the two categories in Label 1 is 2:3 and the sample ratio of the two categories in Label 2 is 3:2. As the sample imbalance problem was not significant and no processing was needed. All samples (approximately 4,200) are divided into a training set and test set. The sample ratio is 80% and 20%, therefore 3,360 samples are in the training set and 840 samples are in the test set approximately.

### 4.3 Results and discussions
*4.3.1 Selection of modeling inputs and parameters*
Comparative analysis for modeling input selection was conducted and results are provided in Table 3.

*4.3.1.1 Input of traffic flow parameters.* As provided in Table 3, AUC values are the highest (0.84 for Label 1 and 0.89 for Label 2) when synthesizing both traffic volume and speed as inputs of the model, compared to cases using simply the traffic volume or the speed as input. Therefore, using traffic volume and speed as the modeling inputs has the best performance, as it can fully describe the traffic flow characteristics. The combination of traffic volume and speed is recommended as inputs for predicting incident clearances.

*4.3.1.2 Time window length.* Then, both traffic flow parameters including traffic volume and speed are integrated as the model inputs. Different time window lengths were tested and compared. The results from Table 3 indicated that the

model with 30-min time window performed best with the highest AUC value of 0.86 for Label 1 and 0.94 for Label 2. This is probably because that the longer time series can better reflect the changes in traffic flow during the process of traffic incident development. Therefore, a time window length s set to be 30 min.

*4.3.1.3 Dynamic traffic parameters and static traffic parameters.* The model with the best performance, i.e. the model of 30-min time window length and application of both speed and traffic volume, is applied. Then modeling approach using dynamic traffic parameters was compared to that uses static traffic parameters. As from Table 3, when using dynamic traffic parameters as inputs, AUC values were higher for both prediction tasks represented by Label 1 (if clearance occurs in 5 min) and 2 (if clearance occurs in 10 min). Therefore, dynamic traffic parameters were preferred in the modeling approach. After the process, the cumulative variance contribution of each principal component reached 89.1%, which is consistent with previous studies (Ru *et al.*, 2017).

*4.3.2 Prediction performance with selected parameters*
Finally, It can be clearly found from Ru *et al.* (2017) that the prediction performance of the model applying dynamic traffic flow parameters is significantly better than the model applying traffic flow parameters processed by PCA. The result confirmed that the dynamic traffic flow parameters outperformed static traffic flow parameters.

This paper provides the ROC curve, KS curve and confusion matrix of the proposed model with optimal parameter combination, as shown in Figure 7, Figure 8 and Table 4. It can be seen from Figure 7 that the AUC values of the LABEL 1 and LABEL 2 are 0.86 and 0.94, respectively. As the ROC curve of LABEL 2 is closer to the upper left corner, the predictive performance of this classifier is better than that of LABEL1. Besides, as shown in Figure 8, it can be seen that the prediction performance of the model is better by observing the change of FPR and TPR with the threshold. The confusion matrix of the proposed model with optimal parameter combination is shown in Table 4. The prediction accuracy of LABEL 1 and LABEL 2 are 80.7% and 85.7%, respectively.

**Table 3** Modeling results

| Results – Different traffic flow parameter inputs | | |
|---|---|---|
| **Parameter combination** | | AUC value |
| | *Label 1* | *Label 2* |
| **Traffic volume** | 0.75 | 0.85 |
| **Speed** | 0.79 | 0.83 |
| **Both traffic volume and speed** | 0.84 | 0.89 |
| *Results – Different time window lengths* | | |
| **Time window lengths** | | AUC value |
| | *Label 1* | *Label 2* |
| **5 min** | 0.84 | 0.89 |
| **10 min** | 0.82 | 0.88 |
| **20 min** | 0.80 | 0.90 |
| **30 min** | 0.86 | 0.94 |
| *Results – Dynamic traffic parameters and traffic parameters processed by PCA* | | |
| **Type of traffic flow parameters** | | AUC value |
| | *Label 1* | *Label 2* |
| **Dynamic traffic flow parameters** | 0.86 | 0.94 |
| **Traffic flow parameters** | 0.64 | 0.70 |

**Figure 7** ROC curve of the best performing model
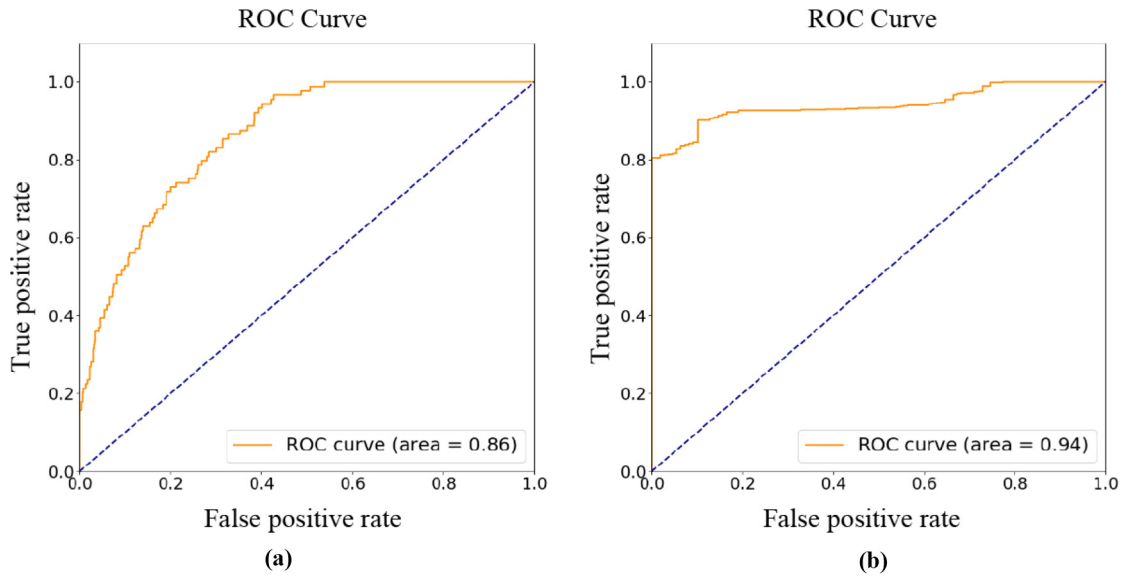


**Notes:** (a) LABEL 1; (b) LABEL 2

**Figure 8** KS curve of the best performing model
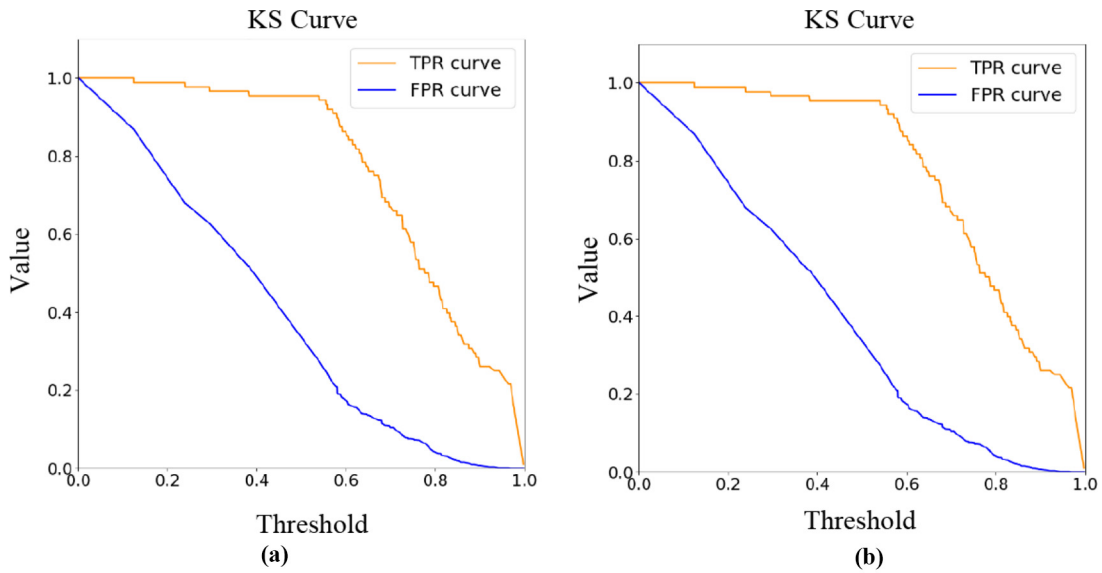


**Notes:** (a) LABEL 1; (b) LABEL 2

**Table 4** Confusion matrix of the best performing model

| | | Results of LABEL 1<br>*Prediction condition* | | | | | Results of LABEL 2<br>*Prediction condition* | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | (%) | | | 0 | 1 | (%) |
| *True condition* | 0 | 261 | 67 | 79.6% | *True condition* | 0 | 412 | 78 | 84.1% |
| | 1 | 96 | 421 | 81.4% | | 1 | 40 | 297 | 88.1% |
| | | *Global accuracy = 80.7%* | | | | | *Global accuracy = 85.7%* | | |

# 5. Conclusions

This paper proposes a framework based on MLP and LSTM to predict the residual incident duration in real-time. The framework uses both real-time traffic flow parameters and traffic incident-related factors. The traffic flow parameters are input into the framework in time series and then processed by LSTM. The traffic incident-related factors are input into the

framework in terms of categorical variables, integrated with the output of the part of LSTM and then processed by MLP. The framework was tested through traffic incident samples recorded by traffic police and traffic flow data obtained by loop detectors in Shanghai Zhonghuan Expressway. The trained framework performed well and shows promise in applying dynamic traffic flow parameters in traffic incident duration prediction.

The main contribution of this study is to propose a framework, which initially applies dynamic traffic flow parameters in traffic incident duration prediction. Dynamic traffic flow parameters better reflect the change of traffic status over traffic incident duration compared with traffic flow characteristics such as posted speed, 85th percentile speed and the ratio of average speed at the time of the incident to that in history. Based on the case study of the Shanghai Zhonghuan Expressway, the impact of both traffic incident-related factors and real-time traffic flow parameters on traffic incident duration are considered in the framework. The results show that the inputs of a 30-min time window, applying both dynamic traffic volume and speed had the best performance and are recommended in future studies.

It is worth mentioning that the framework features high computing power and can provide convenience for road managers and drivers with little delay in future practical applications. However, this model lacks interpretation compared with traditional hazard-based models. Besides, the correlation between input variables is ignored and will be further discussed in future research. The effect of each variable is demonstrated through the statistical method. In addition, the incident samples applied by this study are not enough and the variables are not abundant. The number of injuries and casualties, more detailed description of the incident location and other variables are expected to be used to characterize the traffic incident comprehensively. The framework needs to be further validated through a sufficiently large number of variables and locations.

Further research is desirable to study sequential prediction. Not all traffic incident-related factors can be acquired at the time when the incident is reported, on the contrary, more detailed information is gradually acquired over time. For instance, whether towing cars are required is known as the traffic police arrive at the incident site. A more realistic prediction method that continuously updates model variables and results over time will provide more accurate estimation and reliable references.

## References

Abouaïssa, H., Fliess, M. and Join, C. (2016), "On short-term traffic flow forecasting and its reliability", *IFAC-PapersOnLine*, Vol. 49 No. 12, pp. 111-116.

Adler, M.W., van Ommeren, J. and Rietveld, P. (2013), "Road congestion and incident duration", *Economics of Transportation*, Vol. 2 No. 4, pp. 109-118.

Highway Capacity Manual (2010), *Highway Capacity Manual*, Transportation Research Board, Washington, DC.

Chung, Y. (2010), "Development of an accident duration prediction model on the Korean freeway systems", *Accident Analysis & Prevention*, Vol. 42 No. 1, pp. 282-289.

Cong, H., Chen, C., Lin, P.S., Zhang, G., Milton, J. and Zhi, Y. (2018), "Traffic incident duration estimation based on a dual-learning Bayesian network model", *Transportation*

*Research Record: Journal of the Transportation Research Board*, Vol. 2672 No. 45, pp. 196-209.

Dimitriou, L. and Vlahogianni, E.I. (2015), "Fuzzy modeling of freeway accident duration with rainfall and traffic flow interactions", *Analytic Methods in Accident Research*, Vols 5/6, pp. 59-71.

Fang, SE., Xie, W., Wang, J. and Ragland, D.R. (2016), "Utilizing the eigenvectors of freeway loop data spatiotemporal schematic for real time crash prediction", *Accident Analysis & Prevention*, Vol. 94, pp. 59-64.

Farradyne, P. (2000), *Traffic Incident Management Handbook*, Federal Highway Administration, Office of Travel Management.

Ghosh, I., Savolainen, P.T. and Gates, T. (2014), "Examination of factors affecting freeway incident clearance times: a comparison of the generalized F model and several alternative nested models", *Journal of Advanced Transportation*, Vol. 48 No. 6, pp. 471-485.

Gu, Y., Lu, W., Qin, L., Li, M. and Shao, Z. (2019), "Short-term prediction of lane-level traffic speeds: a fusion deep learning model", *Transportation Research Part C: emerging Technologies*, Vol. 106, pp. 1-16.

Haule, H.J., Sando, T., Lentz, R., Chuan, C.H. and Alluri, P. (2019), "Evaluating the impact and clearance duration of freeway incidents", *International Journal of Transportation Science and Technology*, Vol. 8 No. 1, pp. 13-24.

Hochreiter, S. and Schmidhuber, J. (1997), "Long short-term memory", *Neural Computation*, Vol. 9 No. 8, pp. 1735-1780.

Hojati, A.T., Ferreira, L., Washington, S. and Charles, P. (2013), "Hazard based models for freeway traffic incident duration", *Accident Analysis & Prevention*, Vol. 52, pp. 171-181.

Hojati, A.T., Ferreira, L., Washington, S., Charles, P. and Shobeirinejad, A. (2014), "Modelling total duration of traffic incidents including incident detection and recovery time", *Accident Analysis & Prevention*, Vol. 71, pp. 296-305.

Hojati, A.T., Ferreira, L., Washington, S., Charles, P. and Shobeirinejad, A. (2016), "Modelling the impact of traffic incidents on travel time reliability", *Transportation Research Part C: emerging Technologies*, Vol. 70, pp. 86-97.

Kay, R. and Kinnersley, N. (2002), "On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza", *Drug Information Journal*, Vol. 36 No. 3, pp. 571-579.

Lee, Y., Wei, C.H. and Chao, K.C. (2017), "Non-parametric machine learning methods for evaluating the effects of traffic accident duration on freeways", *Archives of Transport*, Vol. 43 No. 3.

Li, R., Guo, M. and Lu, H. (2017), "Analysis of the different duration stages of accidents with hazard-based model", *International Journal of Intelligent Transportation Systems Research*, Vol. 15 No. 1, pp. 7-16.

Li, R., Pereira, F.C. and Ben-Akiva, M.E. (2015), "Competing risks mixture model for traffic incident duration prediction", *Accident Analysis & Prevention*, Vol. 75, pp. 192-201.

Lin, L., Wang, Q. and Sadek, A.W. (2016), "A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations", *Accident Analysis & Prevention*, Vol. 91, pp. 114-126.

Liu, Y., Liu, Z. and Jia, R. (2019), "DeepPF: a deep learning based architecture for metro passenger flow prediction",

*Transportation Research Part C: emerging Technologies*, Vol. 101, pp. 18-34.

Ma, X., Tao, Z., Wang, Y., Yu, H. and Wang, Y. (2015), "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data", *Transportation Research Part C: emerging Technologies*, Vol. 54, pp. 187-197.

Ma, X., Ding, C., Luan, S., Wang, Y. and Wang, Y. (2017), "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18 No. 9, pp. 2303-2310.

Oh, J.S., Oh, C., Ritchie, S.G. and Chang, M. (2005), "Real-time estimation of accident likelihood for safety enhancement", *Journal of Transportation Engineering*, Vol. 131 No. 5, pp. 358-363.

Ozbay, K. and Noyan, N. (2006), "Estimation of incident clearance times using Bayesian networks approach", *Accident Analysis & Prevention*, Vol. 38 No. 3, pp. 542-555.

Polson, N.G. and Sokolov, V.O. (2017), "Deep learning for short-term traffic flow prediction", *Transportation Research Part C: emerging Technologies*, Vol. 79, pp. 1-17.

Ru, C., Feng, M., Li, J. and Yu, H. (2017), "Incident duration predication based on traffic boardcasting information", in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, *IEEE*, pp. 804-807.

Skabardonis, A., Varaiya, P. and Petty, K.F. (2003), "Measuring recurrent and nonrecurrent traffic congestion", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1856 No. 1, pp. 118-124.

Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L. and Cheng, Z. (2020), "Time-series well performance prediction based on long short-term memory (LSTM) neural network model", *Journal of Petroleum Science and Engineering*, Vol. 186, p. 106682.

Su, J., Zhang, X., Lin, Q., Qin, Y., Yao, J. and Liu, Y. (2019), "Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding", *Artificial Intelligence*, Vol. 277, p. 103168.

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y. and Huang, H. (2020), "Statistical and machine-learning methods for clearance time prediction of road incidents: a methodology review", *Analytic Methods in Accident Research*, Vol. 27, p. 100123.

Wang, J., Liu, B., Fu, T., Liu, S. and Stipancic, J. (2019), "Modeling when and where a secondary accident occurs", *Accident Analysis & Prevention*, Vol. 130, pp. 160-166.

Weng, J., Feng, L., Du, G. and Xiong, H. (2019), "Maximum likelihood regression tree with two-variable splitting scheme for subway incident delay", *Transportmetrica A: Transport Science*, Vol. 15 No. 2, pp. 1061-1080.

Weng, J., Qiao, W., Qu, X. and Yan, X. (2015), "Cluster-based lognormal distribution model for accident duration", *Transportmetrica A: Transport Science*, Vol. 11 No. 4, pp. 345-363.

Zhang, J., Junhua, W. and Shou'en, F. (2019), "Prediction of urban expressway total traffic accident duration based on multiple linear regression and artificial neural network", in *2019 5th International Conference on Transportation Information and Safety (ICTIS)*, *IEEE*, pp. 503-510.

Zou, Y., Ye, X., Henrickson, K., Tang, J. and Wang, Y. (2018), "Jointly analyzing freeway traffic incident clearance and response time using a copula-based approach", *Transportation Research Part C: emerging Technologies*, Vol. 86, pp. 171-182.

Zou, Y., Lin, B., Yang, X., Wu, L., Muneeb Abid, M. and Tang, J. (2021), "Application of the bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management", *Journal of Advanced Transportation*, Vol. 2021.

## Corresponding author

**Ting Fu** can be contacted at: tingfu@tongji.edu.cn