

Open Access

<https://doi.org/10.48130/DTS-2023-0017>
Digital Transportation and Safety 2023, 2(3):211–219

Driving risk assessment under the connected vehicle environment: a CNN-LSTM modeling approach

Yin Zheng^{1,2}, Lei Han^{1,2}, Jiqing Yu³ and Rongjie Yu^{1,2*}

¹ The Key Laboratory of Road and Traffic Engineering, Ministry of Education, No. 4800 Cao'an Road, Shanghai 201804, China

² College of Transportation Engineering, Tongji University, No. 4800 Cao'an Road, Shanghai 201804, China

³ Ningbo Hangzhou Bay Bridge Development Co., Ltd., No. 1 Hongqiao Road, Cixi 315300, Ningbo, China

* Corresponding author, E-mail: yurongjie@tongji.edu.cn

Abstract

Connected vehicle (CV) is regarded as a typical feature of the future road transportation system. One core benefit of promoting CV is to improve traffic safety, and to achieve that, accurate driving risk assessment under Vehicle-to-Vehicle (V2V) communications is critical. There are two main differences concluded by comparing driving risk assessment under the CV environment with traditional ones: (1) the CV environment provides high-resolution and multi-dimensional data, e.g., vehicle trajectory data, (2) Rare existing studies can comprehensively address the heterogeneity of the vehicle operating environment, e.g., the multiple interacting objects and the time-series variability. Hence, this study proposes a driving risk assessment framework under the CV environment. Specifically, first, a set of time-series top views was proposed to describe the CV environment data, expressing the detailed information on the vehicles surrounding the subject vehicle. Then, a hybrid CNN-LSTM model was established with the CNN component extracting the spatial interaction with multiple interacting vehicles and the LSTM component solving the time-series variability of the driving environment. It is proved that this model can reach an AUC of 0.997, outperforming the existing machine learning algorithms. This study contributes to the improvement of driving risk assessment under the CV environment.

Keywords: Connected vehicle; Connected vehicle environment; Driving risk assessment; CNN-LSTM; Traffic safety

Citation: Zheng Y, Han L, Yu J, Yu R. 2023. Driving risk assessment under the connected vehicle environment: a CNN-LSTM modeling approach. *Digital Transportation and Safety* 2(3):211–219 <https://doi.org/10.48130/DTS-2023-0017>

Introduction

Connected vehicle (CV) is regarded as one of the most significant features of the future road transportation system. Various countries and regions, including the United States, European Union, Japan, Australia, Korea, etc., are supporting the development of CV technologies through policies, rules and regulations, and key projects. And Original Equipment Manufacturers (OEMs), including Ford, General Motors, etc., are investing heavily in supporting CV technology research to promote CV commercialization. The reason that CV has gained great attention is that it helps to address critical traffic issues, including safety, mobility, and environmental sustainability. Among these, improving traffic safety is the highest priority. For instance, within the pilot deployment program in the US, 80% Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) applications are safety-related^[1]. The EU claims that investments in connected vehicles are focused on V2V and V2I solutions and infrastructure, with a focus on reducing road risk and crashes^[2]. Under the CV environment, a combination of advanced devices, enabling V2V and V2I communications, collects the data necessary to perform evaluations of safety applications^[3]. These applications include Forward Collision Warning (FCW), Emergency Electronic Brake Lights (EEBL), Work Zone Warnings (WZW), etc. To conduct driving risk assessment applied to CV environment, the corresponding algorithms require better accuracy and robustness to cope with complex

and variable driving scenarios, especially the spatio-temporal interactions with neighboring vehicles.

The current studies mainly provide driving risk assessment from three approaches: (1) Surrogate Safety Measures (SSMs)^[4], (2) models based on physical field theory^[5,6], and (3) machine learning methods. To be specific, firstly, the SSMs including Time-to-Collision (TTC), Deceleration Rate to Avoid the Crash (DRAC), DeltaV, etc., are utilized to measure traffic safety and then compared with pre-determined thresholds to identify traffic conflicts. However, the SSMs face the problems that only a single interactive object is considered, some scenarios are not applicable (e.g., TTC in a high-speed car-following scenario), and theoretically, the validation of SSMs requires a large amount of crash data. Secondly, as for the models based on physical field theory, elements of the entire traffic system, including the driver, vehicle, and environment, are considered in a general model. For example, based on artificial potential field theory, driver-vehicle-road interactions were utilized to calculate virtual mass, as well as the field strength and field force, and hence a driving safety field model was proposed^[5]. However, these models only utilized cross-sectional data to describe the situation at each moment but ignored that the traffic data belong to time series data, raising the time series classification problem^[7]. Thirdly, the machine learning methods, including Classification and Regression Tree (CART)^[8], deep learning^[9], Bayesian Network^[10], etc., are adopted to evaluate and predict driving risks in a more reliable and robust manner.

And some of these methods can overcome the time series classification problem, e.g., Recurrent Neural Network (RNN). Therefore, this kind of machine learning method can comprehensively solve the problems in driving risk assessment research.

Besides, existing driving risk assessment studies mainly employed three kinds of data: (1) driver behavior information^[11], (2) kinematic data^[12,13], e.g., velocity, acceleration, tire forces, (3) contextual features of the neighboring environment^[10], e.g., road conditions, dynamic traffic flow information. These data include influencing elements from all aspects of the entire transportation system. However, these data cannot provide a timely and comprehensive picture of how the subject vehicle interacts with other traffic participants (e.g., neighboring vehicles), which contributes significantly to driving safety. Fortunately, the CV environment data provide a promising way to address this issue. Under the CV environment, detailed information of each vehicle's driving environment, like the kinematics and spatial information of the neighboring vehicles, is available via V2V communication technologies^[14,15]. However, the detailed information, which was always not accessible and hence left out in the previous studies, can be high-resolution and multi-dimensional^[16]. To describe the detailed information in a driving scenario, constructing structured data (e.g., text independent variables), which is commonly used in traffic safety assessment analysis^[17–19], can be clumsy and unattainable. More specifically, some important characteristics (e.g., time series feature) can be obscured due to the data aggregating process from the spatial and temporal dimensions. Meanwhile, it is difficult to describe the neighboring environmental information based on structured data using the exhaustion method. Therefore, it is necessary to explore an effective and comprehensive description of the CV environment data.

In this study, a driving risk assessment framework under the CV environment is proposed, and two advances are highlighted as follows: (1) A novel form of describing the detailed information of the vehicles neighboring the subject vehicle, i.e., time series top view set, was proposed to describe the CV environment data; (2) Developed a hybrid Convolutional Neural Network and Long Short-term Memory (CNN-LSTM) model to analyze both the spatial and temporal features, respectively, i.e., the CNN component considers various elements of the driving environment and the LSTM component solves the time series classification problem. The performance outperformed the existing machine learning algorithms with an AUC of 0.997.

The rest of this paper is structured as follows. In the background section, the driving risk assessment data and approaches are reviewed briefly. In the data preparation section, the identification process of high-risk and non-high-risk events is described in detail. In the methodology section, the CNN-LSTM model structure is introduced, as well as the evaluation metrics. In the modeling results section, the experiment design and the modeling results are clarified. Finally, the summary and discussions section presents the summary of this study and the discussions on the application scenarios of the proposed model.

Background

Traffic safety studies related to CV

In recent years, there has been an increasing amount of traffic safety studies related to CV. According to the objectives of

these studies, four main categories can be summarized as follows.

(1) First, to evaluate the human aspects that affect the safety of CVs, such as driver compliance^[20,21]. Sharma et al.^[21] thoroughly investigated the effect of driver compliance on the mixed traffic environment of CVs and traditional vehicles including both high-compliance and low-compliance drivers.

(2) To calculate the safety implications of CVs in a mixed traffic environment while taking into account various CV Market Penetration Rates (MPRs)^[22–24]. By using a meta-analysis methodology, Xiao et al.^[24] estimated the safety impacts of CVs using MPR and discovered that safety is increased by 4% with 10% MPR, and by 43% with 90% MPR.

(3) To develop traffic flow models for the CV environment, such as platoon-based cooperative driving models^[25], merging advisory models^[26], alarm algorithms^[22], etc. Jia & Ngoduy^[25] introduced a platoon-based car-following model for CVs and used numerical simulations to validate the proposed model.

(4) In addition, to implement traffic safety strategies and management improvements for the CV environment, such as managed lanes^[27,28], improved intersections^[29,30], etc. Rahman & Abdel-Aty^[27] suggested an algorithm for controlling CVs to form platoons in managed lanes, and the longitudinal safety of these platoons was evaluated. Regarding studies on estimating CV-related risk, there are several that concentrate on road segment crash potential prediction^[31], crash-prone intersection identification^[30,32], and risky driving pattern detection^[33].

There are, however, few studies on the real-time driving risk assessment for the CV environment, particularly for the CV environmental characterization data.

Time series classification

As for the time series classification problem, there are some basic deep learning approaches^[7], e.g., Multi-Layer Perceptron (MLP), CNN, RNN, their variants, e.g., LSTM, Echo State Networks, which is a type of RNN. The practices of adopting these approaches to solve time series classification problems in the traffic field are summarized below.

(1) First, traffic flow prediction is one of the most crucial applications^[34,35]. In order to extract the spatial and temporal characteristics, Wang et al.^[35] developed a traffic flow prediction model based on the 1DCNN-LSTM network. This model was demonstrated to display a faster convergence speed and higher prediction accuracy when compared to typical neural network models. With the help of Graph Convolutional Network (GCN) and the Gated Recurrent Unit (GRU)'s sequence-to-sequence structure, Boukerche & Wang^[36] created a hybrid deep learning model that increased the predictability and effectiveness of traffic flow.

(2) Second, the time series classification methods are also widely employed in the field of predicting driver behaviors, such as lane change intention recognition. To recognize the desire to change lanes, Guo et al.^[37] developed a bi-directional long and short-term memory network based on the attention mechanism (AT-BiLSTM). The suggested approach surpassed the current machine learning techniques with an accuracy of 93.33% at 3 s prior to lane change.

(3) Third, increasing research are treating the classification of time series as a challenge in driving risk assessment. To generate driving risk scores, Hu et al.^[9] integrated a convolutional neural network and a long short-term memory encoder/decoder network into a semi-supervised framework. These

A CNN-LSTM modeling approach

scores represented the best comprehensive performance among the available machine learning techniques.

However, to the best of our knowledge, there is no driving risk assessment study that considers time series classification problem for the CV environment.

Data preparation

The modeling data used in this study are derived from the highD dataset^[38], which records the realistic vehicle trajectories on German highways. This massive dataset consists of six separate routes, 110,500 vehicles, 44,500 kilometers travelled, and 147 driving hours. HighD dataset contains nearly twelve times as many vehicles as the well-known NGSIM dataset. And its detecting algorithm could detect about 99% of vehicles while keeping the false positive rate at 2% and position error less than 10 cm^[38]. The overhead viewpoint without blind spots is used to gather traffic data when a drone is used, as illustrated in Fig. 1. The data have a high level of precision with a collection frequency of 25 Hz, and the positioning error is often less than 10 cm. Vehicle type, size, and maneuvers are all included in each vehicle's trajectory. These benefits allow the highD dataset to be utilized to represent the CV environment data since it provides real-time access to precise information about the neighboring environment for each vehicle.

High-risk and non-high-risk events identification

The Modified Time to Collision (MTTC) was employed in this study to identify high-risk events, and 2 s was chosen as the threshold based on earlier research^[39]. When the MTTC value of vehicles reaches 2 s, it was set as the zero time to further extract data to characterize the event. When the MTTC value of a vehicle is lower than 2 s, it may be inferred that the vehicle is in a high-risk event. The case-control data structure was used to identify non-high-risk events, and a 1:4 ratio between the number of high-risk and non-high-risk events was chosen. You may refer to the earlier work^[40] for further information on exactly how high-risk and low-risk events are identified.

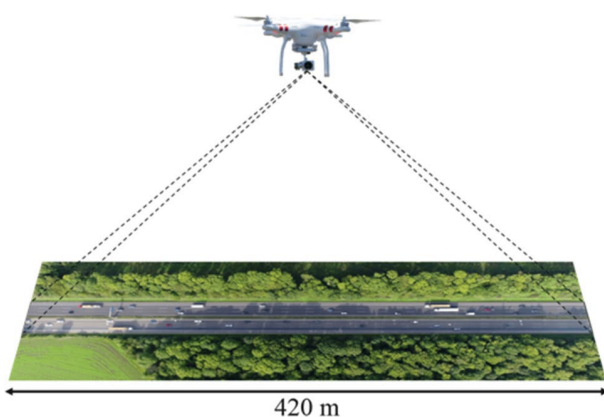


Fig. 1 A drone recording the motion of vehicles along a 420-meter stretch of highway from an overhead perspective^[38].

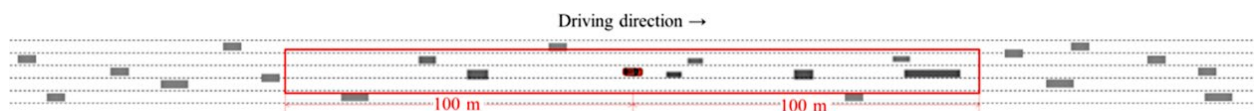


Fig. 2 The illustration of spatial range from a top view.

The driving risk prediction requires the extraction of crash precursor features based on the traffic data in a specific spatial and temporal range before the high-risk and non-high-risk events. For each high-risk or non-high-risk event, the spatial range and temporal range are defined as follows.

As for the spatial range, the lateral range is set as the driving lane of the subject vehicle and the two adjacent lanes in the same direction, which have a greater contribution to the driving risk than other lanes. For the side lane vehicles next to the median barrier or roadside shoulder, the lateral range is adjusted to contain one adjacent lane only. If the subject vehicle is in lane change activities, the lateral range will undergo corresponding dynamic changes. In addition, the longitudinal range is set to be 100 m in both the front and rear of the subject vehicle along the driving direction, and this is an empirical value based on data analysis, for this range relatively comprehensively includes the vehicles that affect the driving risk of the subject vehicle, but not too large. In the top view of Fig. 2, for the red vehicle in the center, the spatial range can be illustrated by the red rectangle.

As for the temporal range, for real-time prediction, the starting time of the temporal range should not be too early, and hence 5 s before the zero time is determined. Besides, considering the response time lag of human drivers, approximately 1 to 2 s^[41], the end time of the temporal range is chosen to be 2 s before the zero time.

Time series top view set

In previous studies, the extraction of crash precursor features is based on text data to aggregate variables (such as average velocity, average flow, velocity variance, etc.), and then use variable selection techniques (such as backward method, random forest, etc.) to obtain key variables. And finally, the input of the crash risk prediction model can be acquired. However, this process is prone to cause information redundancy and key information missing, which will seriously affect the accuracy of prediction. Furthermore, the information redundancy can be reflected by the multicollinearity and coupling relationship of variables. While the missing of key information is mainly due to data collection incompleteness and variable screening process. And it can be reflected by the fact that the above-mentioned variables cannot describe the spatial relationship between the subject vehicle and the neighboring vehicles, and cannot reflect the temporal change characteristics of traffic flow due to the aggregation process.

Therefore, in this study, the feature extraction process is omitted, or replaced by a simpler but more effective way, that is, to construct a time series image (top view) set, which is used as the input of the deep learning model to predict the crash risk. And the time series top view set construction process can be described as follows.

For each moment of the event, there is a corresponding top view as shown in Fig. 3, where (a) is an illustration of the spatial range, (b) is a zoomed-in version of the study area, and (c) is a simplified version as input to the model. More specifically, in Fig. 3b, the dashed lines represent lane lines, and the center of

the image is where the subject vehicle is located, which is true for every image. The other grayscale rectangles represent the vehicles around the subject vehicle. Among them, the relative position of the rectangle is consistent with that of the actual vehicle, the size of the rectangle is proportional to the actual vehicle, and the color is scaled to the velocity of the actual vehicle, that is, the higher the velocity, the darker the rectangle. While in Fig. 3c, only the rectangles representing the neighboring vehicles in Fig. 3b are left, because they are worth researching. This image includes almost all neighboring traffic flow information, such as the number of neighboring vehicles, their spatial location, and their velocity, etc.

For each event, whatever high-risk or non-high-risk, a time series image (top view) set can be built. Each image set is composed of 75 above-mentioned top views, as shown in Fig. 4, respectively representing 75 moments at equal intervals from 5 s (frame000001) to 2 s (frame000075) before the zero

time, which is consistent with the above-mentioned time range and the data collection frequency (25 Hz).

Methodology

In this study, a hybrid CNN-LSTM model structure was adopted, where the CNN model was utilized to capture the spatial characteristics of each top view, and the LSTM model was used to learn the time series features of each driving scenario and further predict the driving risk.

Convolutional Neural Network

The CNN is a kind of feedforward neural network. In this experiment, the original top views were taken as inputs, and after processed by the convolutional layer, pooling layer, and fully connected layer, they are finally input into the LSTM network. The employed CNN structure is shown in Fig. 5, and the detailed parameters with values are listed in Table 1.

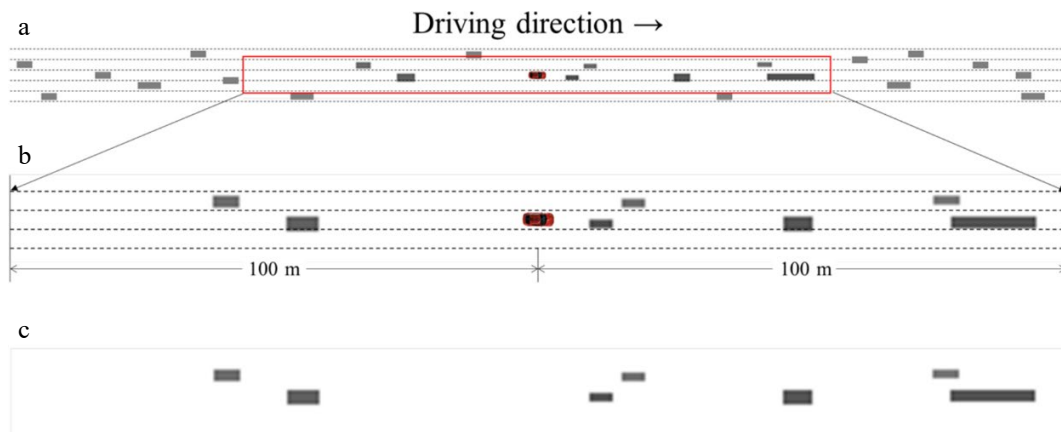


Fig. 3 The detailed and simplified top views for each moment with the subject vehicle in the center. (a) is an illustration of the spatial range, (b) is a zoomed-in version of the study area, and (c) is a simplified version as input to the model.



Fig. 4 The time series top view set describing an event.

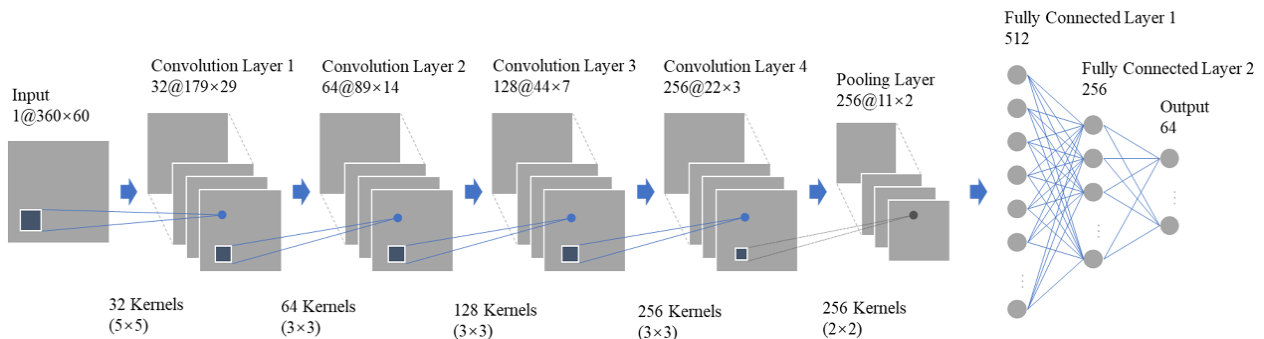


Fig. 5 The structure of CNN model.

A CNN-LSTM modeling approach

Table 1. Parameters with values in the CNN-LSTM modeling process.

Modeling process	Parameters with values
Input of CNN	75 top views with both the front and rear of the subject vehicle: 360 × 30 (or 75 top views with only the front of the subject vehicle: 360 × 60)
Convolution layer	No. of layers: 4 No. of kernels: 32, 64, 128, and 256 Kernel size: (5 × 5), (3 × 3), (3 × 3), and (3 × 3) Stride: (2, 2), (2, 2), (2, 2), and (2, 2) Padding: (0, 0), (0, 0), (0, 0), and (0, 0) Activating function: ReLU
Pooling layer	No. of layers: 1 Kernel size: (2 × 2) Stride: (2, 2) Padding: (0, 0)
Fully connected layer	No. of layers: 2 Hidden neurons: 512 and 256 Activating function: ReLU
Output of CNN model/ Input of LSTM model	No. of features: 64 (for each top view) 75 top views for each event
LSTM	No. of layers: 3 Hidden neurons: 512, 512, 512
Fully connected layer	No. of layers: 1 Hidden neurons: 256 Activating function: ReLU
Output of LSTM model	Binary classification result: high-risk or non-high-risk
Training process	Backpropagation Learning rate: StepLR (lr = 1e-3, γ = 0.3) Loss function: Cross-entropy Mini-batch size: 128 Epochs: 50

Specifically, the convolutional layer, designed with the conv2D layer in this study, consists of several different kernels, and each kernel is defined by a set of shared weights and biases. The result of the convolutional layer is that the same number of features as the kernels are detected, with each feature being detectable across the entire image. The pooling layers is utilized to simplify the information in the output from the convolutional layer. The fully connected layer is used to classify by connecting every neuron from the last layer to every one neuron of the next layer.

Besides, the activation function employed in this study is ReLU, which follows the Eqn. (1) below.

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (1)$$

where x is the input value.

The Adam optimizer^[42] is chosen and cross entropy error function is employed, the calculation formula of cross entropy is shown in Eqn. (2).

$$L = -\frac{1}{N} \sum_i [y_i \times \log(p_i) + (1 - y_i) \times \log(1 - p_i)] \quad (2)$$

where y_i is the label of sample i , high-risk class is 1 and non-high-risk class is 0, p_i indicate the positive possibility of sample i .

Long Short-term Memory

LSTM, introduced by Hochreiter & Schmidhuber^[43], is a special kind of RNN. LSTM is capable of learning long-term dependencies, and hence it can be used to extract information from a sequence. In this study, the feature data identified by CNN model is the input of the LSTM model, and the output is a binary value, which indicates whether an event is high-risk or non-high-risk. The LSTM model structure is mainly composed of forget gate, input gate, and output gate, as shown in Fig. 6.

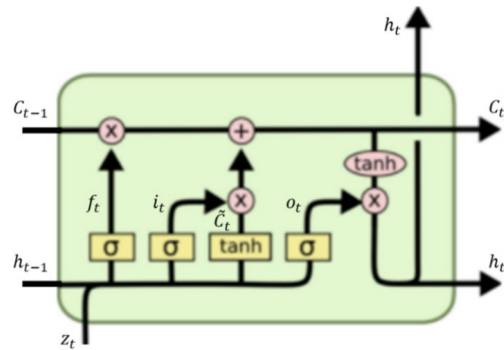


Fig. 6 The basic unit of LSTM model^[43].

The forget gate is to decide what information to throw away from the old cell state C_{t-1} , which is made by a sigmoid layer:

$$f_t = \sigma(W_f \times [h_{t-1}, z_t] + b_f) \quad (3)$$

where z_t is the input of the LSTM model, which is the feature data identified by CNN model in this study. h_{t-1} is the hidden state of the old cell state C_{t-1} . f_t outputs a number between 0 and 1 in the old cell state C_{t-1} based on h_{t-1} and z_t , and 0 means 'completely forget the old cell state' while 1 means 'completely keep it'.

The input gate is to decide what new information to store in the new cell state C_t . Firstly, a sigmoid layer decides which values to update. Secondly, a tanh layer creating a vector of new candidate values to be added to the cell state. Third, update the old cell state C_{t-1} into the new cell state C_t .

$$i_t = \sigma(W_i \times [h_{t-1}, z_t] + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(W_C \times [h_{t-1}, z_t] + b_C) \quad (5)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (6)$$

where $f_t \times C_{t-1}$ means forgetting the information of old cell state, and $i_t \times \tilde{C}_t$ means adding the new candidate values, scaled by how much we decided to update each state.

The output gate is to decide what part to output to the next cell state C_{t+1} . Firstly, a sigmoid layer decides what parts of the cell state C_t to output. And then the cell state C_t is put through a tanh function and multiplied by the output of the sigmoid layer.

$$o_t = \sigma(W_o \times [h_{t-1}, z_t] + b_o) \quad (7)$$

$$h_t = o_t \times \tanh(C_t) \quad (8)$$

where the tanh function is to push the values to be between -1 and 1. h_t is the hidden state of the cell state C_t , which is to output to the next cell state C_{t+1} .

CNN-LSTM model structure

The final CNN-LSTM model structure is shown in Fig. 7, where the input of the CNN model is the time series top views, and the output of the CNN model, which is also the input of LSTM model, is the time series spatial features. And finally, the output of the LSTM model is the predicting result indicating whether an event is high-risk or non-high-risk. More specifically, the CNN model is with four convolution layers, one pooling layer, and two fully connected layers, and the LSTM model is with three layers and one fully connected layer. The detailed parameters with values of the CNN-LSTM model are listed in Table 1. The training and evaluation process of the CNN-LSTM model was implemented on a workstation equipped with NVIDIA GTX 2080Ti GPU and Intel Core i7 processors.

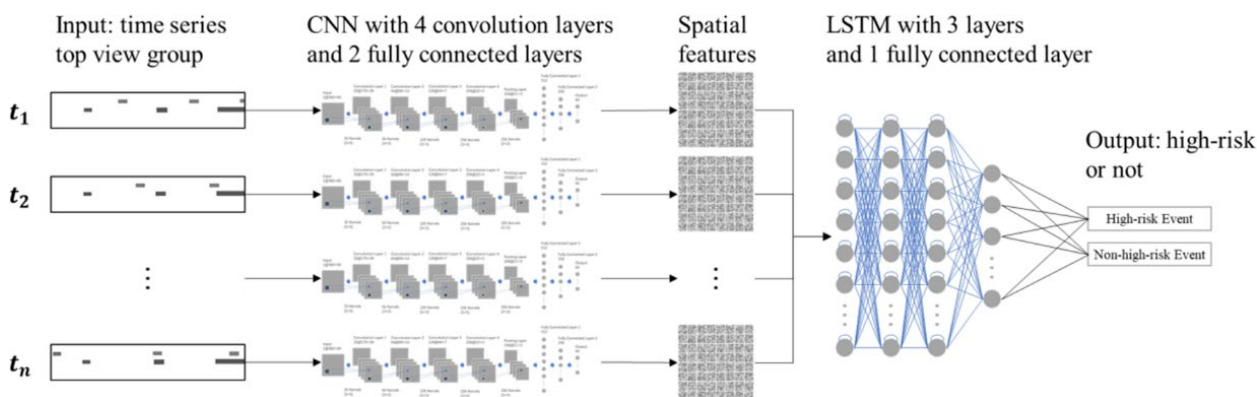


Fig. 7 The proposed CNN-LSTM model.

Evaluation metrics

In this study, accuracy is employed as the evaluation metric in the training process. The receiver operating characteristic (ROC) and area under the ROC curve (AUC) statistic is utilized to evaluate the performance of the proposed CNN-LSTM model. Specifically speaking, the ROC curve illustrates the performance of the binary classification model as its discrimination threshold is varied. The Y-axis of the ROC curve is the true positive rate (TPR), also known as sensitivity, and the X-axis is the false positive rate (FPR), also known as False Alarm Rate, as shown in the Eqn. (9) and (10), respectively. The AUC is utilized for model evaluation and comparison, and its value varies between 0 and 1, with 0.5 representing an uninformative classifier and 1 representing perfect performance. Besides, the accuracy is employed as an overall evaluation metric, as shown in Eqn. (11).

$$TPR = Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$FPR = False\ Alarm\ Rate = \frac{FP}{FP + TN} \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

where *TP* is the True Positive from the confusion matrix, as shown in Table 2, *FN* is the False Negative, *FP* is the False Positive, and *TN* is the True Negative, respectively.

Modeling results

In this study, the modeling dataset, consisting of 255 high-risk events and 1,025 non-high-risk events, was randomly divided into two parts, where 75% was used for training and 25% for validation. To clarify which spatial range and temporal range contribute more to the modeling result, a total of six experiment designs were proposed, considering two variables, i.e., the spatial range and the temporal range. The spatial range is designed into two conditions, the front and rear of the subject vehicle and only the front of the subject vehicle. Besides, as for the temporal range, there are three conditions, namely 5 s to 2 s before the zero time, 5 s to 3 s before the zero

Table 2. The confusion matrix.

Actual condition	Predicted condition	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

time, and 4 s to 2 s before the zero time. And hence, the descriptions of the six experiment designs are shown in Table 3.

As for each experiment, the parameters of the CNN-LSTM model are moderately tuned to obtain more optimal prediction performance. The final results of the six optimal models are shown in Table 4, and the loss, accuracy and, ROC curves of these models are shown in Table 5.

Some conclusions can be drawn as follows.

(1) The model in Experiment 4, where the spatial range is only 100 m in the front of the subject vehicle, and the temporal range is from 5 s to 2 s before the zero time, has the best prediction performance with a sensitivity of 0.996, a False Alarm Rate of 0.065, and an AUC of 0.997.

(2) Given the same temporal range value, the models with spatial range only in the front of the subject vehicle generally perform better than the models with a spatial range in both the front and rear of the subject vehicle.

(3) Given the same spatial range value, the models with a temporal range from 5 s to 2 s before the zero time have the best overall performance. And the models with 5 s as the time starting point work better than the models with 4 s as the time starting point, especially in terms of False Alarm Rate.

Summary and discussions

This study is intended to improve the accuracy of the driving risk assessment model under the CV environment. Firstly, a

Table 3. Descriptions of the six experiment designs.

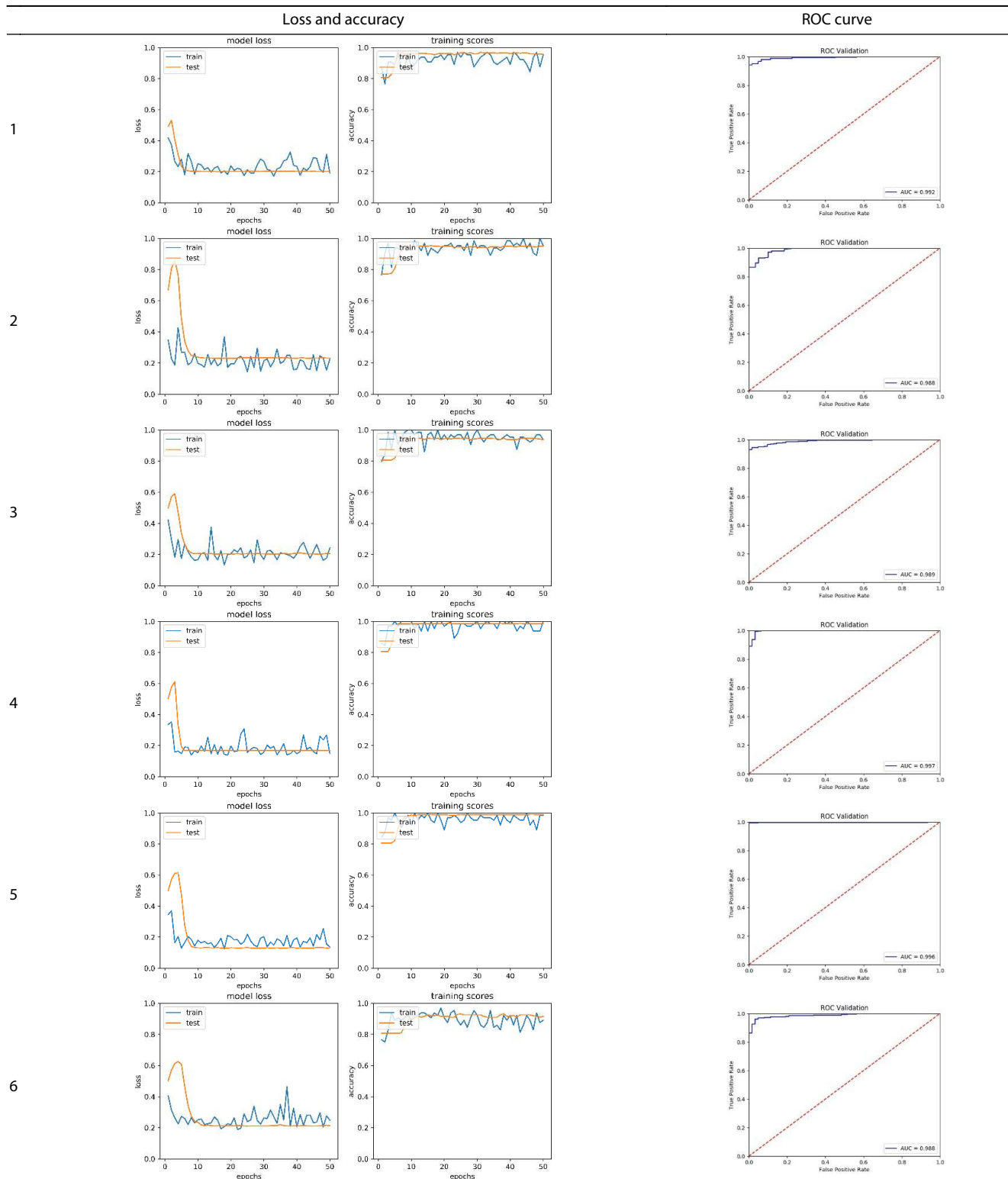
	Spatial range	Temporal range
1	100 m in both the front and rear of the subject vehicle	5 s to 2 s before the zero time
2		5 s to 3 s before the zero time
3		4 s to 2 s before the zero time
4	Only 100 m in the front of the subject vehicle	5 s to 2 s before the zero time
5		5 s to 3 s before the zero time
6		4 s to 2 s before the zero time

Table 4. The prediction performances of the six experiment designs.

	Sensitivity	False Alarm Rate	AUC
1	0.988	0.177	0.992
2	0.977	0.119	0.988
3	0.988	0.274	0.989
4	0.996	0.065	0.997
5	0.992	0.032	0.996
6	0.988	0.387	0.988

A CNN-LSTM modeling approach

Table 5. The loss, accuracy, and ROC curve of the six experimental designs.



novel data form, i.e., time series top view set, was proposed to describe the CV environment data. Then, a hybrid CNN-LSTM model was established to analyze both the spatial and temporal features. More specifically, the CNN component was used to comprehensively consider various elements of the driving environment, and the LSTM component was used to solve the time

series classification problem, which the driving risk assessment can be treated as. Besides, to identify which spatial range and temporal range contribute more to the modeling result, six experiment designs were conducted considering both spatial and temporal variables. Finally, it is proved that the proposed model performed best when the spatial range includes only the

Table 6. Comparison results of modeling performance based on testing data.

Model	Variable dimension	Time-series variability consideration	Sensitivity	False alarm rate	AUC
Random Forest ^[44]	Cross-section	No	0.517	0.088	0.827
RPLR model ^[40]	Single vehicle	Yes	0.980	0.060	0.960
CNN-LSTM model (this study)	Single vehicle	Yes	0.996	0.065	0.997

front part of the subject vehicle and the temporal range is from 5 s to 2 s before the zero time, with a sensitivity of 0.996, a False Alarm Rate of 0.065, and an AUC of 0.997.

The advantages of the proposed model come to the fore when compared with other studies, where the same dataset is utilized, as shown in Table 6. In terms of the form of data organization, independent variable extraction in a single vehicle dimension, e.g., the time series top view set in this study, the temporal and spatial traffic flow characteristic variables^[40], etc., is superior to aggregation of data in the cross-sectional dimension, e.g., the cross-sectional traffic data^[44]. Besides, considering time-series variability can improve the modeling performance. In this study, the time-series variability is considered in the LSTM component, and in the study by Yu et al.^[40], it is considered in the variable construction process.

This study is a beneficial exploration, particularly in terms of the form of data description for the CV environment and model construction. In terms of driving safety, this proposed model can be used for real-time driving risk prediction under the CV environment, where information about the neighboring environment can be obtained in real-time via an onboard unit. From a more macro perspective, this proposed model can be applied to road segment safety assessment by comprehensively considering the driving risk of vehicles within the road segment, as the operational status of all vehicles is accessible under the CV environment.

However, there are still some limitations in this study, e.g., due to data limitations, the spatial range defined in this study was fixed in shape and size with the subject vehicle in the center. However, the spatial range can be irregularly shaped and the size can vary dynamically with the velocity of the subject vehicle. And the temporal range can also take on more continuous values. Besides, highD trajectory dataset was utilized to represent the CV environment data. In future research, the model performance on the real CV dataset should be verified. Moreover, in future studies, multi-dimensional safety surrogate indicators could be used to extract the multi-vehicle collision and other types of high-risk scenarios. Last but not least, it is essential to consider the influence of CV penetration rate and its impacts of the model performance.

Author contributions

The authors confirm contribution to the paper as follows: Study conception and design: Y. Zheng, R. Yu; Data collection: L. Han, Y. Zheng; Analysis and interpretation of results: J. Yu, R. Yu, Y. Zheng; Draft manuscript preparation: Y. Zheng, R. Yu. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data that support the findings of this study are available in the highD repository. These data were derived from the following resources available in the public domain: www.highd-dataset.com.

Acknowledgments

This study was jointly sponsored by the Zhejiang Province Science and Technology Major Project of China (No. 2021C01011), the National Natural Science Foundation of China (NSFC) (No. 52172349), and the China Scholarship Council (CSC).

Conflict of interest

The authors declare that they have no conflict of interest. Rongjie Yu is the Editorial Board member of *Digital Transportation and Safety* who was blinded from reviewing or making decisions on the manuscript. The article was subject to the journal's standard procedures, with peer-review handled independently of this Editorial Board member and his research groups.

Dates

Received 31 May 2023; Accepted 4 September 2023; Published online 28 September 2023

References

- Bennett R, Marsters R, Szymkowski T, Balke KN. 2018. Connected vehicle pilot deployment independent evaluation: Mobility, environmental, and public agency efficiency (MEP) refined evaluation plan-wyoming.
- Lampropoulos G, Siakas K, Anastasiadis T. 2018. Internet of Things (IoT) in industry: Contemporary application domains, innovative technologies and intelligent manufacturing. *International Journal of Advances in Scientific Research and Engineering (IJASRE)* 4(10):109–18
- Bezzina D, Sayer J. 2015. Safety pilot model deployment: Test conductor team report. Report. No. DOT HS 812 171. Washington, DC: National Highway Traffic Safety Administration.
- Wang C, Xie Y, Huang H, Liu P. 2021. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. *Accident Analysis & Prevention* 157:106157
- Wang J, Wu J, Zheng X, Ni D, Li K. 2016. Driving safety field theory modeling and its application in pre-collision warning system. *Transportation Research Part C: Emerging Technologies* 72:306–24
- Mullakkal-Babu FA, Wang M, He X, van Arem B, Happee R. 2020. Probabilistic field approach for motorway driving risk assessment. *Transportation Research Part C: Emerging Technologies* 118:102716
- Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33:917–63
- Wang J, Zheng Y, Li X, Yu C, Kodaka K, et al. 2015. Driving risk assessment using near-crash database through data mining of tree-based model. *Accident Analysis & Prevention* 84:54–64
- Hu H, Wang Q, Cheng M, Gao Z. 2021. Cost-sensitive semi-supervised deep learning to assess driving risk by application of naturalistic vehicle trajectories. *Expert Systems with Applications* 178:115041

A CNN-LSTM modeling approach

10. Zhu X, Yuan Y, Hu X, Chiu YC, Ma YL. 2017. A Bayesian Network model for contextual versus non-contextual driving behavior assessment. *Transportation Research Part C: Emerging Technologies* 81:172–87
11. Petraki V, Ziakopoulos A, Yannis G. 2020. Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data. *Accident Analysis & Prevention* 144:105657
12. Jiang K, Yang D, Xie S, Xiao Z, Victorino AC, et al. 2019. Real-time estimation and prediction of tire forces using digital map for driving risk assessment. *Transportation Research Part C: Emerging Technologies* 107:463–89
13. Mao H, Guo F, Deng X, Doerzaph ZR. 2021. Decision-adjusted driver risk predictive models using kinematics information. *Accident Analysis & Prevention* 156:106088
14. Ali Y, Zheng Z, Haque MM. 2021. Modelling lane-changing execution behaviour in a connected environment: A grouped random parameters with heterogeneity-in-means approach. *Communications in Transportation Research* 1:100009
15. Lee E-Y, Cho H-J, Ryu K-Y. 2016. A probabilistic approach for collision avoidance of uncertain moving objects within black zones. *Ad Hoc Networks* 52:50–62
16. Lim KL, Whitehead J, Jia D, Zheng Z. 2021. State of data platforms for connected vehicles and infrastructures. *Communications in Transportation Research* 1:100013
17. Abdel-Aty M, Uddin N, Pande A, Abdalla MF, Hsia L. 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board* 1897:88–95
18. Yu R, Quddus M, Wang X, Yang K. 2018. Impact of data aggregation approaches on the relationships between operating speed and traffic safety. *Accident Analysis & Prevention* 120:304–10
19. Yu R, Zheng Y, Abdel-Aty M, Gao Z. 2019. Exploring crash mechanisms with microscopic traffic flow variables: A hybrid approach with latent class logit and path analysis models. *Accident Analysis & Prevention* 125:70–78
20. Sharma A, Ali Y, Saifuzzaman M, Zheng Z, Haque MM. 2017. Human factors in modelling mixed traffic of traditional, connected, and automated vehicles. *AHFE 2017: Advances in Human Factors in Simulation and Modeling*, ed. Cassenti D. Switzerland: Springer, Cham. pp. 262–73. https://doi.org/10.1007/978-3-319-60591-3_24
21. Sharma A, Zheng Z, Kim J, Bhaskar A, Haque MM. 2021. Assessing traffic disturbance, efficiency, and safety of the mixed traffic flow of connected vehicles and traditional vehicles by considering human factors. *Transportation Research Part C: Emerging Technologies* 124:102934
22. Liu H, Wei H, Zuo T, Li Z, Yang YJ. 2017. Fine-tuning ADAS algorithm parameters for optimizing traffic safety and mobility in connected vehicle environment. *Transportation Research Part C: Emerging Technologies* 76:132–49
23. Jo Y, Jang J, Park S, Oh C. 2021. Connected vehicle-based road safety information system (CROSS): Framework and evaluation. *Accident Analysis & Prevention* 151:105972
24. Xiao G, Lee J, Jiang Q, Huang H, Abdel-Aty M, Wang L. 2021. Safety improvements by intelligent connected vehicle technologies: A meta-analysis considering market penetration rates. *Accident Analysis & Prevention* 159:106234
25. Jia D, Ngoduy D. 2016. Platoon based cooperative driving model with consideration of realistic inter-vehicle communication. *Transportation Research part C: Emerging Technologies* 68:245–64
26. Xin Q, Fu R, Ukkusuri SV, Yu S, Jiang R. 2021. Modeling and impact analysis of connected vehicle merging accounting for mainline random length tight-platoon. *Physica A: Statistical Mechanics and its Applications* 563:125452
27. Rahman MS, Abdel-Aty M. 2018. Longitudinal safety evaluation of connected vehicles' platooning on expressways. *Accident Analysis & Prevention* 117:381–91
28. Abdel-Aty M, Wu Y, Saad M, Rahman MS. 2020. Safety and operational impact of connected vehicles' lane configuration on freeway facilities with managed lanes. *Accident Analysis & Prevention* 144:105616
29. Essa M, Sayed T. 2020. Self-learning adaptive traffic signal control for real-time safety optimization. *Accident Analysis & Prevention* 146:105713
30. Hu J, Huang MC, Yu X. 2020. Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. *Accident Analysis & Prevention* 144:105665
31. Li P, Abdel-Aty M, Cai Q, Yuan C. 2020. The application of novel connected vehicles emulated data on real-time crash potential prediction for arterials. *Accident Analysis & Prevention* 144:105658
32. Ma Y, Zhu J. 2021. Left-turn conflict identification at signal intersections based on vehicle trajectory reconstruction under real-time communication conditions. *Accident Analysis & Prevention* 150:105933
33. Ali EM, Ahmed MM, Yang G. 2021. Normal and risky driving patterns identification in clear and rainy weather on freeway segments using vehicle kinematics trajectories and time series cluster analysis. *IATSS Research* 45:137–52
34. Narmadha S, Vijayakumar V. 2023. Spatio-Temporal vehicle traffic flow prediction using multivariate CNN and LSTM model. *Materials Today: Proceedings* 81:826–33
35. Wang K, Ma C, Qiao Y, Lu X, Hao W, et al. 2021. A hybrid deep learning model with 1DCNN-LSTM-Attention networks for short-term traffic flow prediction. *Physica A: Statistical Mechanics and its Applications* 583:126293
36. Boukerche A, Wang J. 2020. A performance modeling and analysis of a novel vehicular traffic flow prediction system using a hybrid machine learning-based model. *Ad Hoc Networks* 106:102224
37. Guo Y, Zhang H, Wang C, Sun Q, Li W. 2021. Driver lane change intention recognition in the connected environment. *Physica A: Statistical Mechanics and its Applications* 575:126057
38. Krajewski R, Bock J, Kloecker L, Eckstein L. 2018. The highD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. *Proc. 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4-7 November 2018*. USA: IEEE. pp. 2118–25. <https://doi.org/10.1109/ITSC.2018.8569552>
39. Meng Q, Qu X. 2012. Estimation of rear-end vehicle crash frequencies in urban road tunnels. *Accident Analysis & Prevention* 48:254–63
40. Yu R, Han L, Zhang H. 2021. Trajectory data based freeway high-risk events prediction and its influencing factors analyses. *Accident Analysis & Prevention* 154:106085
41. Bergasa LM, Almería D, Almazán J, Yebes JJ, Arroyo R. 2014. Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. *2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 8-11 June 2014*. USA: IEEE. pp. 240–45. <https://doi.org/10.1109/IVS.2014.6856461>
42. Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. *arXiv Preprint*
43. Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Computation* 9:1735–80
44. Yuan C, Li Y, Huang H, Wang S, Sun Z, et al. 2022. Application of explainable machine learning for real-time safety analysis toward a connected vehicle environment. *Accident Analysis & Prevention* 171:106681



Copyright: © 2023 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.