Transportation Research Part C 119 (2020) 102740

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Convolutional neural networks with refined loss functions for the real-time crash risk analysis

Rongjie Yu [a,b], Yiyun Wang [a,b], Zihang Zou [c,*], Liqiang Wang [c]

[a] College of Transportation Engineering, Tongji University, 4800 Cao'an Road, 201804 Shanghai, China
[b] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 4800 Cao'an Road, 201804 Shanghai, China
[c] Department of Computer Science, University of Central Florida, Orlando, FL 32816, United States

ARTICLE INFO

Keywords:
Real-time crash risk analysis
Convolutional Neural Network (CNN)
Focal loss
Temporal and spatial operational features
Imbalanced data issue

ABSTRACT

The real-time crash risk analyses were proposed to establish the relationships between crash occurrence probability and pre-crash traffic operational conditions. Given its great application potentials that link with Active Traffic Management System (ATMS) for proactive safety management, it has become an important research area. Currently, researchers mainly developed the real-time crash risk analysis models with traffic flow descriptive statistics employed as explanatory variables and with re-sampled balanced dataset, which hold the limitations of insufficiently capturing the temporal-spatial traffic flow characteristics and failing to provide classification capabilities when deal with the imbalanced datasets. In this study, a Convolutional Neural Network (CNN) modelling approach with refined loss functions has been first time introduced to the real-time crash risk analyses. The primary objectives of the proposed CNN models are: (1) utilizing the tensor-based data structure to explore the multi-dimensional, temporal-spatial correlated pre-crash operational features; and (2) optimizing the loss functions to overcome the low classification accuracy issue brought by the imbalanced data. Data from the Shanghai urban expressway system were utilized for the empirical analysis. And a total of three types of loss functions, including traditional binary cross entropy, the α-weighted cross entropy and the focal loss, were introduced and being tested with varying ratios of crash and non-crash datasets. The modeling results show that the CNN model has better classification performance compared to the traditional Multi-layer Perceptrons (MLP) model with the tensor-based structure data. Besides, the developed CNN model with focal loss function has substantial classification enhancement under the imbalanced datasets. Finally, the distributions of predicting probabilities for balanced and imbalanced datasets were plotted to understand the effects of the imbalanced dataset and revealed how the proposed CNN model with focal loss function improves the model performance.

1. Introduction

Safety is the most critical issue for the transportation system as it was reported that there were about 1.35 million people died annually due to traffic crashes, which has ranked as the eighth leading cause of death in the world (World Health Organization, 2018). To improve the traffic safety statuses, tremendous efforts have been investigated to develop safety analysis models to understand the

* Corresponding author.
E-mail addresses: yurongjie@tongji.edu.cn (R. Yu), wangyiyun@tongji.edu.cn (Y. Wang), zzz@knights.ucf.edu (Z. Zou), lwang@cs.ucf.edu (L. Wang).

https://doi.org/10.1016/j.trc.2020.102740
Received 20 April 2020; Received in revised form 18 July 2020; Accepted 28 July 2020
Available online 6 August 2020

influencing factors and further implement improvement countermeasures. Among which, an emerging approach is to conduct real-time crash risk analyses with the recent developments of advanced traffic sensing and operation management techniques (Oh et al., 2001; Ahmed and Abdel-Aty, 2011; Shi and Abdel-Aty, 2015). The real-time crash risk analyses try to establish the relationships between crash probability and the pre-crash traffic operational conditions (Hossain and Muromachi, 2012), which could be used to provide proactive warnings to trigger Active Traffic Management System (ATMS) and further reduce the crash potentials (Ahmed and Abdel-Aty, 2011). Besides, the real-time crash risk estimations also hold potentials for the connected and autonomous vehicle (CAV) application scenarios to provide network-level crash prediction for the real-time motion planning (Katrakazas et al., 2015).

Within the existing crash risk analysis studies, researchers mainly adopt the descriptive statistics of traffic flow parameters to represent the traffic operational conditions (Hossain et al., 2019), and further employ binary classification models (e.g. logistic regression model) to develop the functional relationships between the traffic flow parameters and binary crash outcomes (crash vs. non-crash) (Roshandel et al., 2015). Although decent estimation accuracies have been obtained, there are still several issues during the modeling procedure remain to be solved. First, during the independent variables formation process, given the unknown essential of traffic crash precursors, the descriptive statistics of traffic flow parameters are mainly adopted by researchers. For instance, mean speed, standard deviation of speed, average volume, etc. However, the crash occurrence is claimed to be impacted by the concurrent impacts of multiple influencing factors (Lee et al., 2003), therefore, the simple descriptive statistics of traffic flow parameter may not be sufficient. Besides, the subjective formed independent variables possess the risk of subjective selection biases (Dabiri and Heaslip, 2018). It is critical to identify a new approach to capture the crash precursor features from the operation data. Recently, in order to explore efficient features from the traffic flow data with consideration of their temporal and spatial internal relations, several studies have tried to utilize higher dimensions of input data when modeling crash data. For instance, Cai et al. (2019) transferred high-resolution transportation and land use data into the form of images while Polson and Zhu et al. (Polson and Sokolov, 2017; Zhu et al., 2018) modeled the spatio-temporal relations lie in the traffic flow and incident data.

Secondly, given the crash occurrence rare event feature, the non-crash samples are greatly outnumbered crash samples within the empirical data (Abdel-Aty et al., 2004). And this leads to the imbalanced data classification issue of crash risk analysis, where the predominance of the majority class would mislead the model's optimization direction (Krawczyk, 2016). Regarding this, current real-time crash risk analyses mostly form equivalent ratios (mostly 1:1 and 1:4) through the under sampling methods (Abdel-Aty et al., 2004; Yang et al., 2018), where the number of non-crash samples is substantially reduced for the data fed into the models by matched case-control or randomly discarding. However, it is claimed that the under-sampling procedure might lose useful information the model need to learn from (Johnson and Khoshgoftaar, 2019). Another approach is the over-sampling method which increases the number of crash samples by randomly duplicating (Yuan et al., 2019), while this method may cause the overfitting problem (Chawla et al., 2004; Longadge and Dongre, 2013). In addition to re-sampling the data before modeling, the other approach is to deal with the imbalanced data issue by adjusting analytical methods settings during the model learning or decision process (Zhou and Liu, 2005; Sun et al., 2007), which lacks of in-depth investigation in the crash risk analysis field.

With the above-mentioned research gaps, this study aims at conducting real-time crash risk analyses with multi-dimensional traffic flow input features, and dealing with the imbalanced data classification issue by exploring a modeling approach during the model learning process. To be specific, we employ the Convolutional Neural Network (CNN) model with refined focal loss functions to perform the real-time crash risk analysis. The main contributions of this study can be summarized as follows:

(1) Utilized a tensor-based structure rather than the traditional matrix-based analysis data to represent crash precursors.
(2) Extracted the multi-dimensional, temporal and spatial correlated pre-crash operational features with the application of CNN model.
(3) Overcome the low classification accuracy brought by the imbalanced data with the optimization of the loss function structure of CNN model.

The remainder of this paper is organized as follows: in the *Methodology* section, the structure of our proposed CNN model and the modified loss function methods that deal with the imbalanced data issue are presented. In the *Data Preparation* section, we illustrate how crash data were established and the form of the proposed tensor-based data. In the *Modeling Results* section, the experimental results are presented and in the *Discussions* section, the results are analyzed and discussed. Finally, conclusions and future work outlook are provided in the *Conclusion* section.

2. Methodology

2.1. Convolutional Neural Network (CNN)

The Convolutional Neural Networks are biologically-inspired variants of MLPs (Multi-layer perceptrons), which was firstly proposed to deal with image recognition issues (LeCun et al., 1989). CNN differs from the previous MLP as that for MLP each node is fully connected to nodes in the previous layer (Dabiri and Heaslip, 2018). The unique characteristic of CNN is due to its convolution layers, which contain learnable kernel filters as its parameters and take tensor as input and output, so that it can obtain features with spatially local correlations from a small region of the preceding layers. Output element is computed as the dot product between adjacent input elements among channels and kernel filter with respect to one output channel. The intermediate output of convolution layers is interpreted as "features" and can be applied as the input of next convolution layer. Specifically, given an image with w width, h height, and C_{in} color channels, it can be considered as a $C_{in} \times w \times$ h tensor. The 2d convolution is then applied on a region of neighboring pixels

and have C_{out} channels for the output. Then the output tensor is featured with size $C_{out} \times w_{out} \times h_{out}$ as,

$$out(C_{out_j}) = bias(C_{out_j}) + \sum_{k}^{C_{in}-1} weight(C_{out_j}, k) \star input(k) \tag{1}$$

where $\star$ is the valid 2d cross-correlation operator.

The advantage of CNN dealing with the image has been realized that it can exploit more complex architectures and extract high dimensional features of input data. It has been proved to have better performance and more flexibility for other research topics as its applications are not limited for computer vision problems, for example, CNNs have been employed to perform traffic signs recognition (Jin et al., 2014) and pedestrian detection (Szarvas et al., 2005) in transportation field and autonomous driving (Wu et al., 2017; Tian et al., 2018).

Considering the features of crash risk analysis data, the proposed CNN model structure follows the design of Alexnet (Krizhevsky et al., 2012), which is constructed with a sequence of layers, which are the input layer, hidden layers (including the convolution layer, the response-normalization layer, and the activation layer), and the fully-connected layer. Each convolution layer is followed by one batch normalization layer and RELU as nonlinear activation layer. The proposed model is shown in Fig. 1. The illustration of layer settings are as follows:

- The input layer of CNN is a $6 \times 3 \times 6$ tensor, the formation process and the structure will be illustrated in the *Empirical Data* section.
- For the convolution layers, in order to explore the hidden relations temporally and spatially with 2d convolution, the kernel size is set to 3x3 with stride as 2, and 1 as padding. The neighboring data would be calculated for every 2 time slices and 2 traffic flow parameters at a time; and the model is, therefore, capable of extracting high-level combinations of traffic flow variables.
- RELU is used as non-linear activation layer and Sigmoid is used in the last layer for binary classification purpose.
- Normalization Layer is utilized to resolve the internal covariate shift issue (where input distribution changes during training, which slows down the convergence process), by normalizing layer inputs into zero mean and unit variance with the following formula:

$$x^{'} = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} \tag{2}$$

where x is the input and $x^{'}$ is the output, ϵ is a very small constant to validate above formula. Expectation and variance are computed among all elements in a mini batch.

The loss function for CNN is usually the binary cross entropy, the formula is as follows:

$$loss = \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{3}$$

where p_i is the predicted possibility and y_i is the ground truth label with 1 as crash and 0 as non-crash. The loss function is trained by back-propagation. Back-propagation is an algorithm to calculate the gradient of the loss function with respect to the input variables (LeCun et al., 1990). By iteratively accumulating gradients to the weights, the loss function can be minimized and thus the local optimal of the CNN can be obtained. In this study, the network is trained with Stochastic Gradient Descent (SGD) using back propagation as a gradient computing technique (Bottou, 2010). Specifically, the weights are being updated as follow:

$$w_i^{'} = w_i - \eta \frac{\partial E}{\partial w_i} - \eta \lambda w_i \tag{4}$$

where the learning rate η was set as 1×10^{-4}, learning rate decay was set as 0.9 and weight decay λ was set as 1×10^{-1}. Batch size was set as 64. Finally, the model is trained for 100 epochs (an epoch is a single step that a neural network is trained on every sample in one

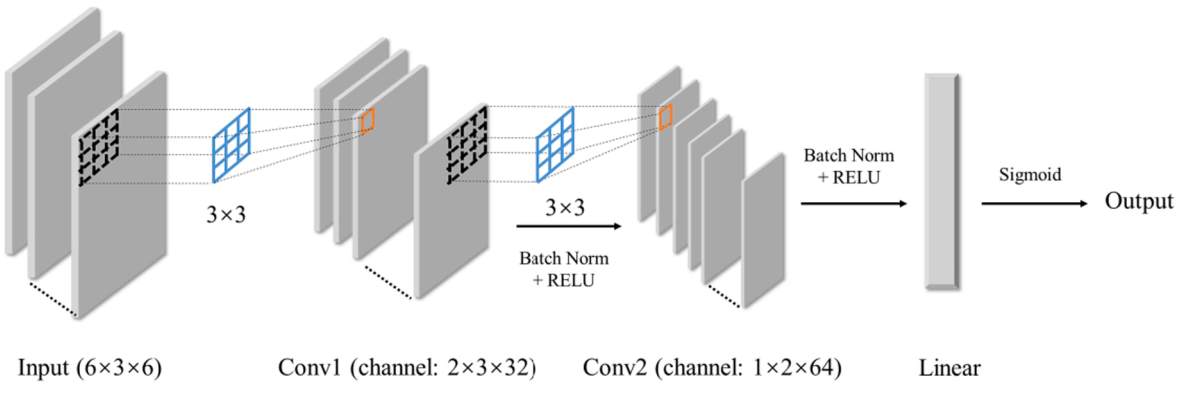


Fig. 1. Proposed model structure.

pass).

2.2. Modified loss function of CNN

Referring to previous studies (Hensman and Masko, 2015; Yan et al., 2015), CNN has issues dealing with imbalanced data since highly imbalanced data would cause the model classification leaning towards the majority class, while the minority class is often the class of primary interest. To solve this problem, researchers try to alter the structure of loss function to achieve a balanced weight for each class.

One common approach is to introduce an adjusting weight parameter α to the loss function on the basis of binary cross entropy so that the weight of minority class is enhanced. In this study, we call it α-weighted cross entropy. The formula is as follows:

$$loss = \sum_{i=1}^{N} \alpha y_i \log(p_i) + (1-\alpha)(1-y_i)\log(1-p_i) \tag{5}$$

where y_i is the safety situation for the i^{th} sample, $y_i = 1$ indicates crash occurrence and $y_i = 0$ indicates non-crash. p_i is the predicting probability of i^{th} sample being crash state. $\alpha \in (0,1)$ is the weight for the positive sample and $(1-\alpha)$ for the negative. When $\alpha = 0.5$, it is a traditional binary cross entropy.

Another method is proposed by Lin et al. (2017) which is called focal loss. In this study, a modified focal loss formula was defined as:

$$loss = \sum_{i=1}^{N} \alpha(1-p_i)^{\gamma 1} y_i \log(p_i) + (1-\alpha) p_i^{\gamma 2}(1-y_i)\log(1-p_i) \tag{6}$$

where y_i and p_i have the same meaning as formula (5). $\gamma 1$ (≥ 0) and $\gamma 2$ (≥ 0) are the tunable focusing parameters for the positive and negative samples respectively, while $\gamma 1 = \gamma 2 = 0$ the focal loss is the α-weighted cross entropy. The focal loss function adaptively alters the learning focus of easy samples (where p_i is close to 0 for the negative ground truth or p_i is close to 1 for the positive ground truth) and the hard samples (the opposite of the easy sample) based on the predicting probability p_i for each learning stage. Both the α-weighted cross entropy function and the focal loss function were investigated in this study to deal with the imbalanced dataset. Table 1 presents the parameters for the experimented loss functions in this study.

2.3. Model performance evaluation

Finally, to evaluate the CNN model classification performance, sensitivity and false alarm rate (FAR) are utilized which are usually adopted in crash risk analyses (Li et al., 2020). The calculating equations are presented as formula (7) and (8). Sensitivity is the total correct prediction counts on crash among all crash samples. FAR is the total false prediction counts on non-crash among non-crash samples. The evaluation matrix is explained in Table 2. In addition, Area under the ROC (AUC) index is also adopted to evaluate the performance of the binary classifier.

$$Sensitivity = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative} \tag{7}$$

$$False\ \ Alarm\ \ Rate = \frac{Flase\ \ Positive}{False\ \ Positive + True\ \ Negative} \tag{8}$$

As the models' estimation results are posterior crash occurrence probabilities range from 0 to 1. Therefore, in order to compare the model classification accuracy, a threshold (or called cut-off point) needs to be selected. In this study, a fixed threshold of 0.5 was adopted, which was widely adopted in the literatures (Abdel-Aty and Pande, 2005; Jiang et al., 2020). The model performance comparison procedure is shown in Fig. 2. The crash occurrence classification sensitivity was treated as the most critical evaluation index, therefore, within the accepted FAR level (0.1 in this study), models with higher sensitivity are preferred.

3. Data Preparation

In this study, empirical data from the Shanghai urban expressway system were utilized. A total of three datasets were used to obtain

Table 1
Parameters for the experimented loss functions.

Loss function	α	γ1	γ2
α-weighted cross entropy	> 0.5	–	–
Normal focal loss	> 0.5	> 0	> 0
N-balanced focal loss	> 0.5	= 0	> 0
P-balanced focal loss	> 0.5	> 0	= 0

Table 2
Binary classification evaluation matrix.

Ground truth\prediction	Class = Crash	Class = Non-Crash
Class = Crash	True positive *(Sensitivity)*	False negative
Class = Non-Crash	False positive *(False Alarm Rate)*	True negative

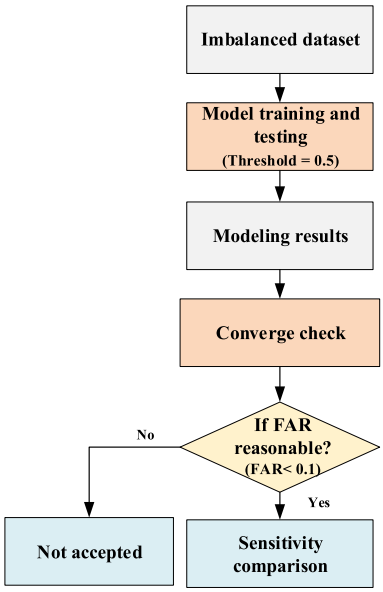


Fig. 2. Model comparison procedure with fixed threshold.

the real-time crash risk analysis data, which are (1) crash data that occurred in Apr 2014. The crash data were provided by the Shanghai Traffic Information Center, which have high quality information of crash location and time. As for the crash location, the information was recorded with stake numbers. Stake numbers are marked along the Shanghai urban expressway, which are consisted of letters or Chinese characters and numbers. They are ordered with non-repetitive numbers of designed foundation piles when constructed. Therefore, the traffic crashes occurred on the urban expressway system hold accurate crash locations. Besides, for each crash occurrence, the crash time was checked based on the full-coverage video surveillance system; (2) roadway section geometry data collected manually from the online street-view map, and was checked with the detailed design files (the expressway was split into 206 roadway sections using on-ramps and off-ramps as dividing points); and (3) raw traffic data (update at 20 s frequency) detected by loop detectors (LDs) at roadway section level. The dense of loop detectors on Shanghai urban expressway systems is relatively high with an average spacing distance of 650 m, compared to an average of around 800 m in the US freeways (e.g. Xu et al., 2013; Abdel-Aty et al., 2005), and an average of 1064 m in South Korea (Kwak and Kho, 2016) and other countries. Thus, the high dense loop detector data could provide detailed traffic flow data for the empirical analyses.

During the data process procedure, raw traffic data have first been aggregated at 5-minute interval (i.e., 9:00–9:05, 9:05–9:10). Therefore, each roadway section would have 288 data points ((24 h × 60 min)/5 min = 288) per day. As for the crashes, based on the crash locations, a total of three adjacent roadway sections were identified, which were named as crash (C, section that the crash occurred), upstream (U, upstream section of the crash section), and downstream (D, downstream section of the crash section) sections; and referencing to the crash occurrence time, a 30-mintue interval traffic data prior to each crash for the corresponding three sections were identified.

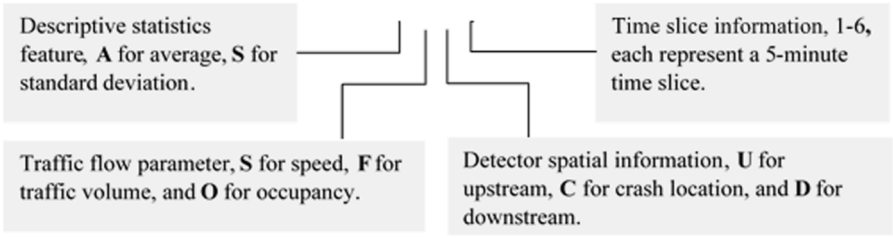


Fig. 3. Nomenclature rule for traffic variables.

In addition, unlike traditional real-time crash risk analysis that adopted a matched-case control design to identify non-crash cases traffic data, a full set of non-crash cases traffic data were extracted in this study. The full set non-crash traffic data were obtained through the following steps. First, a 1-hour time window (i.e. 30-minute before crash occurrence and 30-minute after) was used to remove the crash related traffic data from the raw traffic data; therefore, the data points within this 1-hour time window were dropped. Then, a 30-mintue interval traffic data for the remaining data points were identified to formulate the non-crash data.

For the identified traffic data (both crash and non-crash cases), the selected 30-minute interval was split into six 5-minute time slices. Then, mean and standard deviation value for speed, volume and occupancy were calculated. A four characters' nomenclature rule (shown in Fig. 3) for the traffic variables was proposed. The first letter stands for the descriptive statistics (mean or standard deviation), and the second letter represents traffic flow parameter (speed, volume or occupancy), the third letter show the section (C, U, or D) while the last numeric characteristic indicates the time slice for which the variable belongs to. Finally, there are a total of 108 traffic flow parameters (6 traffic operation statues variables, 3 roadway sections and 6 time slices) in the final dataset.

Generally, traditional logistic regression models directly utilize the 1 dimensional vector variables as inputs and learn weights from all variables simultaneously via training. However, in this study, more attentions have been paid to explore the high-level, complex combinations of the traffic flow parameters rather than considering them globally. Thus, the explanatory variables were re-aligned into a $6 \times 3 \times 6$ tensor (shown in Fig. 4), where 6 traffic operation statues variables can be regarded as channels and 3 roadway sections $\times$ 6 time slices can be considered as spatial and temporal coordinates. All the elements have been then normalized into zero mean and unit variance (due to the large variable size, the summary statistics were not given), which provided more convenience to apply convolution operation to learn local information from adjacent time slices and nearby traffic operations.

The final prepared dataset contains 1,152 crash cases and 236,716 non-crash cases. In order to investigate the effects of imbalanced data issue, datasets with varying crash and non-crash ratios were created, which are 1:1, 1:4, 1:10, 1:20, and 1:100 datasets. Then after the randomly sampling procedure, each dataset was further split into training and testing data with the corresponding proportion of 3:1

4. Modeling results

4.1. Basic CNN and MLP

A total of 5 models were developed based on different ratios of the crash and non-crash cases as 1:1, 1:4, 1:10, 1:20 and 1:100. Fig. 5 presents (1) the evolvement of loss function values and accuracy values; (2) sensitivity values and precision values; and (3) specificity values with the number of 100 epochs for the individual modeling results of the varying each crash and non-crash ratios. From which, the influence of data imbalance issue on CNN performance can be explored.

In general, back propagation can train the proposed CNN architecture well with stable results of most ratios of crash and non-crash except for 1:100. The increase of accuracy and the decrease of the loss can be seen in all ratios. As the ratio of crash and non-crash ranges from 1:1 to 1:100, the overall accuracy tend to be 1 and the loss tend to be 0, which is especially obvious for model with ratio of 1:100. This phenomenon can be explained by the predominance of non-crash samples, which dominates the loss function and guide its optimization direction rather than crash samples. As the non-crash samples increase within the dataset, the model tends to judge as many as possible samples as negative in order to gain a small loss.

The developed models are meant to evaluate the real-time crash risks and further identify the proactive control scenarios for ATMS interventions. Therefore, the sensitivity and the FAR are more appropriate and critical to evaluate the modeling performance. The definition of sensitivity and FAR were given in the *Methodology* section. In addition, AUC is also employed to measure the overall performance of the model. The values of sensitivity, FAR and AUC for different ratios are presented in Table 3.

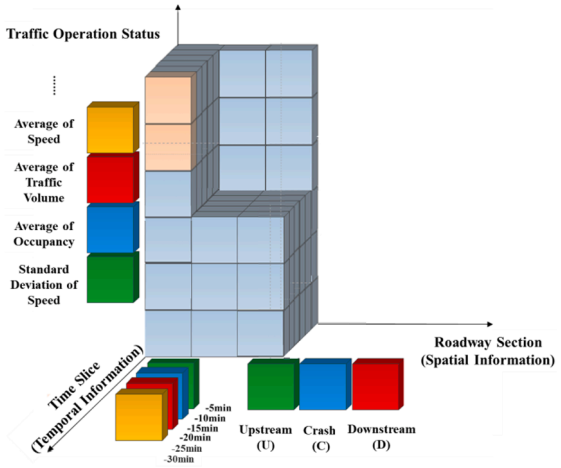
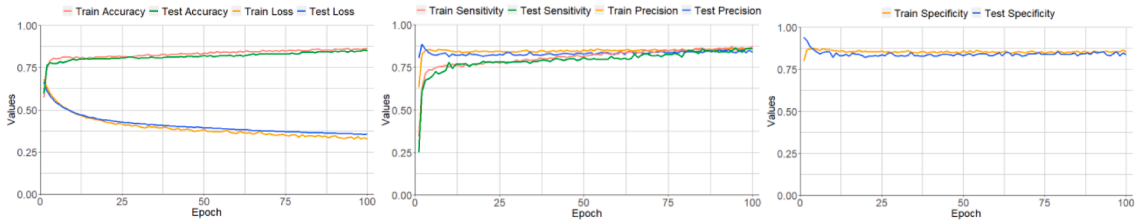
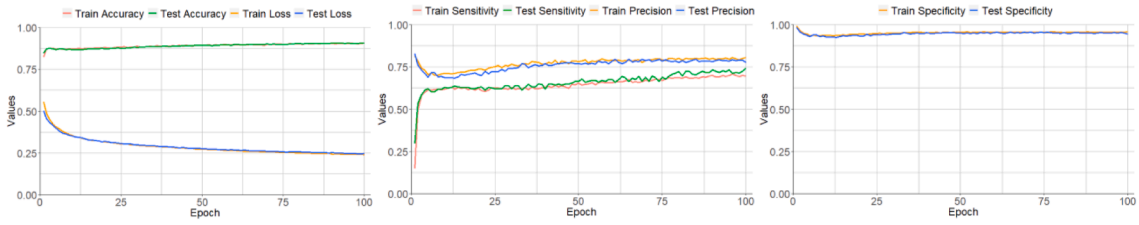


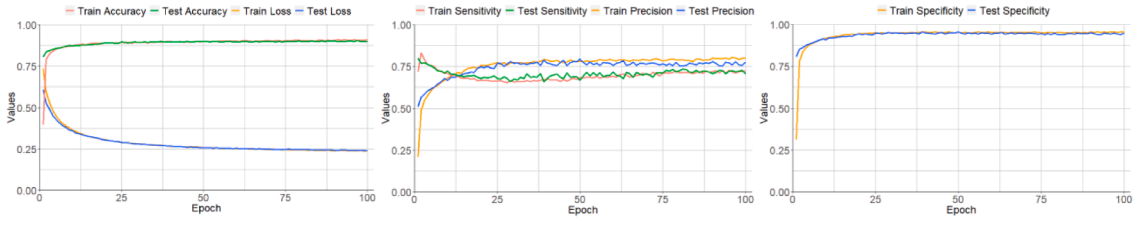
Fig. 4. The tensor structure of the independent variables.



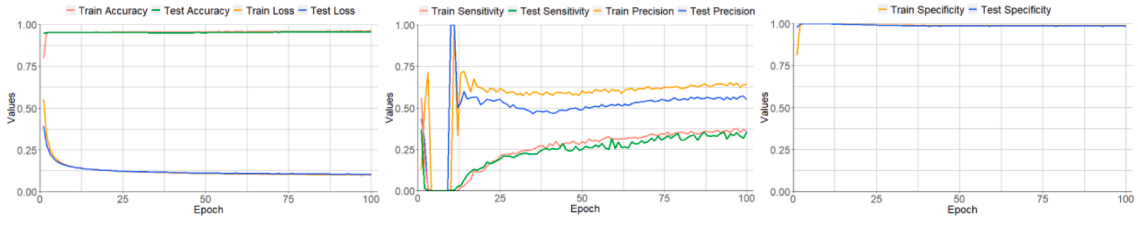
(a) 1:1 ratio



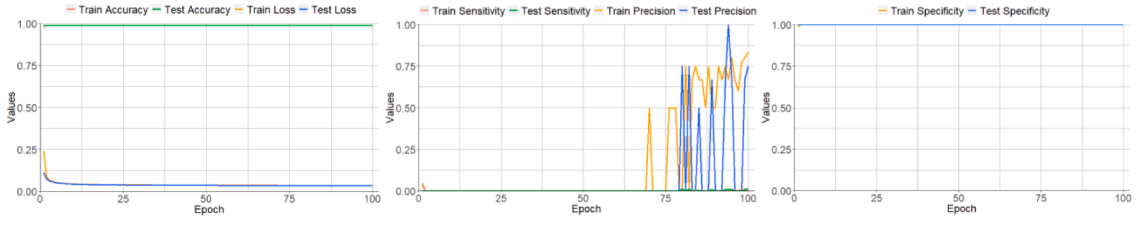
(b) 1:4 ratio



(c) 1:10 ratio



(d) 1:20 ratio



(e) 1:100 ratio

Fig. 5. Modeling results for CNN models with different ratios of crash and non-crash.

As the threshold is fixed to 0.5, which was illustrated in the *Methodology* section, models with FAR under 0.1 are acceptable, and then sensitivity values are used for comparison. It can be seen that when the crash and non-crash ratio is 1:1, although the model achieved the best sensitivity value of 0.861, it has the FAR value of 0.161 which is higher than 0.1 and is not accepted. As for ratio 1:4, the model performed well with a sensitivity of 0.746 and a FAR of 0.055, which means within all crash risk scenario samples, 74.6% of

Table 3
CNN crash risk classification results based on test data.

	CNN			MLP		
Crash/non-crash ratio	**Sensitivity**	**FAR**	**AUC**	**Sensitivity**	**FAR**	**AUC**
1:1	0.861	0.161	0.931	0.832	0.161	0.924
1:4	0.746	0.055	0.951	0.743	0.060	0.944
1:10	0.725	0.059	0.949	0.682	0.053	0.949
1:20	0.343	0.013	0.947	0.354	0.010	0.949
1:100	0.011	0.000	0.953	0.118	0.001	0.959

them were successfully identified; within non-risk scenarios, only 5.5% of them were falsely alarmed as risky. As the ratio of non-crash cases increases, it shows substantial negative influences on the CNN model performance. For instance, with ratio of 1:100, the sensitivity value declines to 0.0107 while the FAR value is almost 0, it implies that the model tends to classify all their inputs into non-crash cases.

Furthermore, a 4-layers MLP model which holds the same layer sequence of the proposed CNN model was then developed for comparison. The different settings between CNN and MLP models are: (1) CNN holds two convolution layers while for MLP they are all fully-connected layers; (2) the input for CNN model is the three-dimensional tensor while for MLP it is the one dimensional. Besides, the number of neurons of each fully-connected layers and the activation functions of the MLP model were optimized. The modeling results for MLP are also included in Table 3. Under the balanced ratio of crash and non-crash cases (i.e. 1:1), the CNN shows better performance than MLP with a higher sensitivity value and a better AUC value. It implies the CNN's capacity of extracting multi-dimensional information. While as the ratio of non-crash samples increasing, the performance of CNN model decays faster than MLP with respect to sensitivity. This phenomenon implies CNN strucutre has the problem of dealing with imbalanced dataset.

4.2. CNN with modified loss functions

Referring to the abovementioned modeling results, it can be seen that CNN models could not provide good fits for the imbalanced data. In this section, two different approaches were employed to improve the modeling performance: (1) α-weighted cross entropy: paying more attention to the positive samples by enhancing the weight of positive samples, i.e. when the positive samples are mistakenly classified, the model would gain more penalty; and (2) focal loss: on the basis of α-weighted cross entropy, introducing adjusting parameters to dynamically alter the weight of hard samples. Specific loss functions structures are exhibited in *Methodology* part. Tables 4 and 5 are the modeling results with the utilization of two approaches:

Table 4 lists the crash risk classification results with different α parameter settings for α-weighted cross entropy. It shows substantial improvement of sensitivity compare to the binary cross entropy. For instance, when α equals to 0.7, from ratio 1:4 to ratio 1:100, the sensitivity values have increased by 9.29%, 9.64%, 22.5% and 12.14% corresponding to their binary cross entropy counterparts. Nevertheless, as the proportion of non-crash samples gets larger, the model still performs poor. For instance, under the ratio of 1:100, the sensitivity is around 0.14, it cannot match with the sensitivity when the dataset is balanced. Therefore, α-weighted cross entropy has limited ability to handle the greatly outnumbered non-crashes.

In addition, different combinations of α and γ were experimented to test focal loss method. It indicated that no significant differences with varying α. Given the limited space issue, only the modeling results with $\alpha = 0.7$ for three different kinds of focal loss structures are presented in Table 5. γ value equals to 0.5, 1, 2, 5 and 10 was tested respectively, given the bad modeling results of 10 and the similar modeling results between 1 and 2, the modeling results of 0.5, 2 and 5 for γ parameter were finally presented.

It can be seen that the coordinated work of α and γ can substantially improve the model performance with higher sensitivity and lower FAR (<0.1) when compares to only employing α parameter. For instance, under 10% FAR, the sensitivity values can reach to around 84%, 74%, 88% and 67% with ratio of 1:4, 1:10, 1:20 and 1:100 respectively, which are 79%, 60%, 43% and 24% employing merely α. Furthermore, P-focal loss works best with the ratio of 1:4, while N-focal loss works better with the ratio of 1:10 and 1:100, and Normal focal loss for 1:20.

Table 4
CNN crash risk classification results under α-weighted cross entropy.

α	**Crash/non-crash ratio**	**Sensitivity**	**FAR**	**AUC**	**Crash/non-crash ratio**	**Sensitivity**	**FAR**	**AUC**
0.5(binary)	**1:1**	0.861	0.161	0.931				
0.5(binary)	**1:4**	0.746	0.055	0.951	**1:20**	0.343	0.013	0.947
0.6		0.825	0.097	0.939		0.589	0.031	0.943
0.7		0.839	0.105	0.939		0.568	0.031	0.946
0.8		0.811	0.098	0.935		0.586	0.029	0.945
0.9		0.821	0.103	0.940		0.572	0.031	0.940
0.5(binary)	**1:10**	0.725	0.059	0.949	**1:100**	0.011	0.000	0.953
0.6		0.818	0.103	0.937		0.153	0.003	0.948
0.7		0.821	0.103	0.936		0.132	0.002	0.953
0.8		0.832	0.105	0.935		0.146	0.003	0.951
0.9		0.821	0.103	0.935		0.268	0.009	0.930

Table 5
CNN crash risk classification results under focal loss.

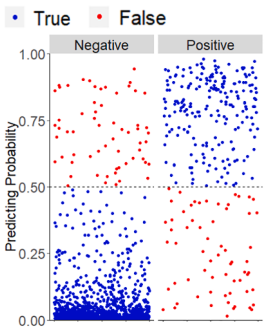
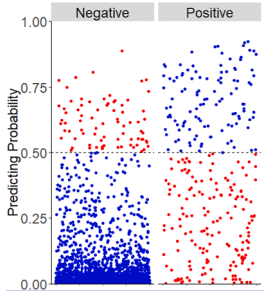
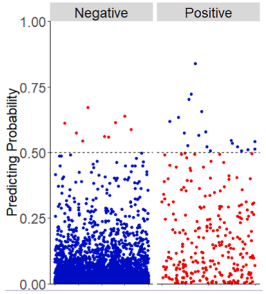
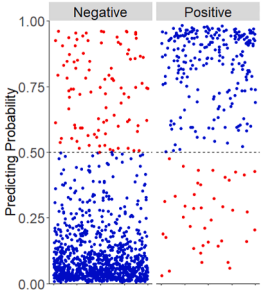
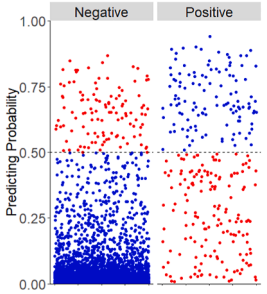
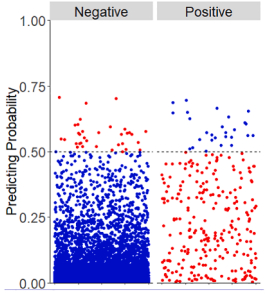
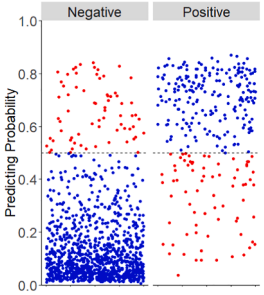
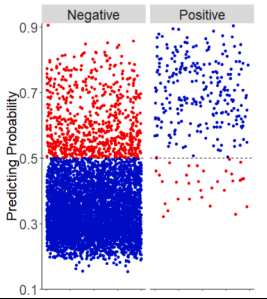
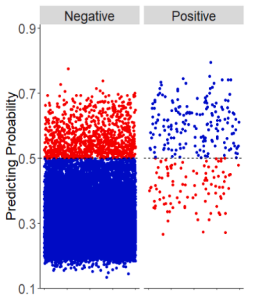
Crash/non-crash ratio	γ	Normal focal loss			N-focal loss			P-focal loss		
		Sensitivity	FAR	AUC	Sensitivity	FAR	AUC	Sensitivity	FAR	AUC
1:4	**0.5**	0.793	0.112	0.925	0.964	0.323	0.937	***0.839***	***0.097***	***0.940***
	2	0.800	0.113	0.927	0.964	0.323	0.942	***0.807***	***0.088***	***0.939***
	5	0.793	0.113	0.924	0.957	0.346	0.934	***0.829***	***0.086***	***0.941***
1:10	**0.5**	0.618	0.044	0.923	***0.732***	***0.046***	***0.937***	0.568	0.032	0.939
	2	0.611	0.048	0.919	***0.736***	***0.050***	***0.939***	0.550	0.032	0.936
	5	0.604	0.048	0.912	***0.700***	***0.049***	***0.934***	0.550	0.028	0.939
1:20	**0.5**	***0.611***	***0.032***	***0.943***	0.882	0.144	0.940	0.154	0.005	0.943
	2	***0.600***	***0.030***	***0.941***	0.882	0.126	0.942	0.161	0.005	0.945
	5	***0.568***	***0.029***	***0.944***	0.839	0.128	0.936	0.175	0.006	0.943
1:100	**0.5**	0.089	0.002	0.933	***0.654***	***0.036***	***0.941***	0.011	0.000	0.950
	2	0.125	0.003	0.931	***0.643***	***0.037***	***0.945***	0.007	0.000	0.947
	5	0.136	0.003	0.937	***0.668***	***0.038***	***0.945***	0.007	0.000	0.948

5. Discussions

5.1. Posterior probability analyses

To better understand the classification performances under the imbalanced datasets, and how the model performances change along with the utilization of refined loss functions, scatter plots for the posterior probabilities were shown in Table 6. Within the

Table 6
Predicting probability values under different CNN loss function structures.

Ratio	1:4	1:20	1:100
Binary cross entropy			
α-weighted cross entropy			
focal loss			

figures, the negative and positive samples are assigned to two vertical strips; the true predictions and the false judgments can be recognized by blue and red plots (which is diverged by 0.5 threshold horizon dashed line within each sample strip). So that the ratio of plots occupying the top right corner reflects **sensitivity** and that of the top left corner exhibits **FAR**. As the crash risk analysis is expected to have high sensitivity while triggering low FAR (Hossain et al., 2019), where the plots would show *as maximize the top right corner plots and meanwhile minimize the top left corner plots.* To be more specific, better model classification capability expects large differences between the posterior probabilities of negative samples and positive samples.

Under the binary cross entropy loss function, it can be seen that under the well prediction of 1:4 ratio, the predicting probability values for negative and positive samples are scattering away from each other. These two sets probabilities locate at the opposite polar, therefore, the threshold can separate them accurately. As the ratio of negative cases increases, the probabilities for negative and positive samples are both drawing to 0 and locating at the same polar, it confirms the above inference that the model classifies nearly all the samples as negative, thus, the threshold cannot classify them well and the model performs poor.

With the α-weighted cross entropy method which increases the weight of positive samples, there are relatively more plots located at the top right corner compared to the binary cross entropy between the counterpart ratios. Therefore, the sensitivities under α-weighted cross entropy are higher than those with the binary cross entropy. However, the model still cannot handle larger ratio of negative samples, since both positive and negative predicting probabilities are still gathering to 0. Therefore, the ability of α-weighted cross entropy method that deals with the imbalanced data is still limited

Under the focal loss function structure, compared with the other two loss function structures, the distribution of probabilities for negative and positive cases are much more polarized and differed. The threshold line can separate them better to maximize the top right plots and at the same time reduce the top left plots. Therefore, the model performance improves a lot.

To conclude, the focal loss function improves the modeling results by enlarging the differences between the predicting probabilities of negative samples and those of positive samples as much as possible with the benchmark of threshold value.

5.2. Model performance comparisons

Table 7 summarized the modeling results of different models with the crash and non-crash ratios. Through comparing the sensitivity of the CNN (binary cross entropy) model under balanced 1:1 ratio with MLP, it can be concluded that with the additional local and spatial information from the CNN structure, the crash risk prediction accuracy has been enhanced. Besides, through the experimental results of different loss functions, it shows that the α-weighted cross entropy has limited ability to handle the imbalanced data issue while focal loss function could substantially improve the model performance. Even under the imbalanced 1:100 ratio, the proposed CNN (focal loss) model has reached to a 66.8% sensitivity with a 3.8% FAR.

In addition, the proposed CNN models have provided the state-of-art classification performances compared with the literatures utilizing other modeling methods (summarized in Table 8). It can be seen from Table 8 that under relatively balanced dataset (1:4 ratio), the sensitivity of our proposed CNN model enhances 6.6%-27% with a lower FAR value. Aside from the traditional range of under-sampling proportion, even under the 1:100 imbalanced ratio, the proposed CNN model is still competitive that keeps a relatively high sensitivity and at the same time constrains the FAR at the low level.

6. Conclusion

The real-time crash risk analyses hold the benefits of providing deep understandings for the crash precursors and implementing the proactive traffic safety management strategies. In order to obtain a better crash risk predicting performance, tremendous efforts have been investigated from the aspects of various operational sensing data and advanced modeling techniques. However, the majority existing studies established their models based upon basic and manual selected simple descriptive statistics of traffic flow parameters, which may not be sufficient to describe the temporal and spatial traffic operational features. Besides, previous studies mainly adopted re-sampling methods to handle the imbalanced crash risk analysis dataset, which may lose useful information or cause the overfitting issue. Therefore, an analytical scheme that could gain deeper information from the multi-dimensional pre-crash traffic operation data and deal with the imbalanced data issue are needed.

In this study, Convolutional Neural Networks (CNN) models with refined loss functions have been utilized, *for the first time*, to conduct real-time crash risk analysis. The developed modeling scheme holds the advantages of extracting multi-dimensional, temporal and spatial correlated pre-crash operational features and overcoming the low classification accuracy brought by the imbalanced data. The input data were transferred to the tensor-based structure which has three dimensions of time slice, roadway section and traffic operation status. Crash analysis dataset with different ratios of crash and non-crash samples were established and modelled with the utilization of CNNs. Controlled experiments were conducted with MLP and three variation forms of the CNN loss function including the traditional binary cross entropy, the α-weighted cross entropy and the focal loss. The modeling results show that the CNN structure has the ability to extract the additional local and spatial information, and CNNs with focal loss function could substantially improve the model performance even under the imbalanced 1:100 ratio. Besides, further analyses from the aspect of posterior predicting probabilities have revealed how the proposed CNN (focal loss) model improves the model performance. It is shown that the distribution of predicting probabilities for negative and positive cases are much more polarized and differed, rather than classifying nearly all the samples as the negative under the traditional binary cross entropy.

The proposed model was developed and implemented on the basis of Pytorch framework. The experiments were conducted using Python 3.7 under Linux 16.04 operation system with 32 GB RAM, and NVIDA GTX 2080Ti GPU. Under ratio 1:100, the GPU-based training and testing took around 18 s for a single epoch and a total of 100 epochs training were needed to reach convergence

Table 7
Summary of the modeling results with different settings.

Model	Crash/non-crash ratio	Sensitivity	FAR	AUC
CNN (binary cross entropy)	**1:1**	86.1%	16.1%	0.931
MLP		83.2%	16.1%	0.924
CNN (binary cross entropy)	**1:4**	74.6%	5.5%	0.951
	1:10	72.5%	5.9%	0.949
	1:20	34.3%	1.3%	0.947
	1:100	1.10%	0	0.953
CNN (α-weighted cross entropy)	**1:4**	83.9%	10.5%	0.939
	1:10	83.2%	10.5%	0.935
	1:20	58.9%	3.1%	0.943
	1:100	26.8%	0.9%	0.930
CNN (focal loss)	**1:4**	83.9%	9.7%	0.940
	1:10	73.6%	5.0%	0.939
	1:20	88.2%	12.6%	0.942
	1:100	**66.8%**	**3.8%**	**0.945**

Table 8
Crash risk prediction results based on test data in literatures.

Authors	**Modeling algorithm**	**Sensitivity**	**FAR**	**Ratio of crash and non-crash**
Abdel-Aty et al. (2005)	Matched case–control logistic regression	56.0%	20.0%	1:4
Ahmed and Abdel-Aty (2011)	Stratified matched case-control logistic regression	69.1%	45.2%	1:4
Ahmed et al. (2012)	Semiparametric Bayesian modeling	75.0%	45.0%	1:4
Sun and Sun (2015)	Dynamic Bayesian network with time series	76.4%	23.7%	1:5
This study	*CNN with focal loss function*	*83.0%*	*9.70%*	*1:4*
Abdel-Aty and Pande (2005)	Probabilistic neural network (PNN)	73.9%	28.7%	1:8
Xu et al. (2014)	Bayesian updating approach	46.3%	10.0%	1:10
Wang et al. (2015)	Multilevel Bayesian logistic regression model	67.6%	30.0%	1:10
Xu et al. (2016)	Random effect logit model	50.0%	10.3%	1:11
This study	*CNN with focal loss function*	*73.6%*	*5.0%*	*1:10*
Hossain and Muromachi (2011)	Classification and regression trees	63.3%	20.0%	1:92
This study	*CNN with focal loss function*	*66.8%*	*3.8%*	*1:100*

(around 30 min). Thus, the total computing time for the CNN training and testing is acceptable.

Furthermore, within traffic safety analysis field, although there have been several studies tried to apply the emerging deep learning models such as CNN and LSTM to perform crash risk analyses, many of which stayed at the stage of applying the basic method without adjusting and optimizing the model structure according to the specific analysis target. In this study, we have made pioneering efforts altering the input data structure as the tensor-based form aiming at extracting higher dimensional information of traffic operation features to fit the need of the study, and adjusted the loss function structure of deep learning model to handle the imbalanced data issue. The proposed models have achieved decent modeling results, while due to the black box characteristic of deep learning models, it is difficult to interpret the modeling results. The interpretability of the developed models should be investigated to further unveil the crash precursor characteristics. In addition to the adopted CNN model, other deep learning methods (e.g., ResNet) could be further being explored to improve the level of accuracy (Hossain et al., 2019). As existing models are mostly one-way, the methods that have the feedback effect like reinforcement learning can be utilized to satisfy the requirement of proactive traffic safety management application. Furthermore, the application procedure of the proposed modeling methods as well as the model transferability issue are needed in the future.

CRediT authorship contribution statement

Rongjie Yu: Conceptualization, Methodology, Writing - review & editing. **Yiyun Wang:** Investigation, Visualization, Writing - review & editing. **Zihang Zou:** Conceptualization, Software. **Liqiang Wang:** Software, Supervision.

Acknowledgement

This study was jointly sponsored by the Chinese National Natural Science Foundation (NSFC 71771174 and 71531011) and the “Chenguang Program” supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission.

References

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec. 1897 (1), 88–95.
Abdel-Aty, M., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36 (1), 97–108.

Abdel-Aty, M., Uddin, N., Pande, A., 2005. Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Transp. Res. Rec. 1908 (1), 51–58.
Ahmed, M., Abdel-Aty, M., 2011. The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Trans. Intell. Transp. Syst. 13 (2), 459–468.
Ahmed, M., Abdel-Aty, M., Yu, R., 2012. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transp. Res. Rec. 2280 (1), 60–67.
Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, pp. 177–186.
Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., Yuan, J., 2019. Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. Transp. Res. Part A: Policy Practice 127, 71–85.
Chawla, N., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter 6 (1), 1–6.
Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp. Res. Part C: Emerging Technol. 86, 360–371.
Jiang, F., Yuen, K., Lee, E., 2020. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. Accid. Anal. Prev. 141, 105520.
Jin, J., Fu, K., Zhang, C., 2014. Traffic sign recognition with hinge loss trained convolutional neural networks. IEEE Trans. Intell. Transp. Syst. 15 (5), 1991–2000.
Johnson, J., Khoshgoftaar, T., 2019. Survey on deep learning with class imbalance. Journal of Big Data 6 (1), 27.
Hensman, P., Masko, D., 2015. The impact of imbalanced training data for convolutional neural networks. Degree Project in Computer Science, KTH Royal Institute of Technology.
Hossain, M., Muromachi, Y., 2011. Understanding crash mechanisms and selecting interventions to mitigate real-time hazards on urban expressways. Transp. Res. Rec. 2213 (1), 53–62.
Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 45, 373–381.
Hossain, M., Abdel-Aty, M., Quddus, M., Muromachi, Y., Sadeek, S., 2019. Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. Accid. Anal. Prev. 124, 66–84.
Katrakazas, C., Quddus, M., Chen, W., Deka, L., 2015. Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. Transp. Res. Part C: Emerging Technol. 60, 416–442.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Prog. Artificial Intelligence 5 (4), 221–232.
Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097–1105.
Kwak, H., Kho, S., 2016. Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data. Accid. Anal. Prev. 88, 9–19.
LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1 (4), 541–551.
LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1990. Handwritten digit recognition with a back-propagation network. Adv. Neural Inf. Process. Syst. 396–404.
Lee, C., Hellinga, B., Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. Transp. Res. Rec. 1840 (1), 67–77.
Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. Accid. Anal. Prev. 135, 105371.
Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.
Longadge, R., Dongre, S., 2013. Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707.
Oh, C., Oh, J., Ritchie, S., Chang, M., 2001. Real-time estimation of freeway accident likelihood, 80th Annual Meeting of the Transportation Research Board, Washington, DC.
Polson, N., Sokolov, V., 2017. Deep learning for short-term traffic flow prediction. Transp. Res. Part C: Emerging Technol. 79, 1–17.
Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–211.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C: Emerging Technol. 58, 380–394.
Sun, Y., Kamel, M., Wong, A., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 40 (12), 3358–3378.
Sun, J., Sun, J., 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transp. Res. Part C: Emerging Technol. 54, 176–186.
Szarvas, M., Yoshizawa, A., Yamamoto, M., Ogata, J., 2005. Pedestrian detection with convolutional neural networks, IEEE Proceedings. Intelligent Vehicles Symposium, 2005. IEEE, pp. 224–229.
Tian, Y., Pei, K., Jana, S., Ray, B., 2018. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In: Proceedings of the 40th International Conference on Software Engineering, pp. 303–314.
Wang, L., Abdel-Aty, M., Shi, Q., Park, J., 2015. Real-time crash prediction for expressway weaving segments. Transp. Res. Part C: Emerging Technol. 61, 1–10.
World Health Organization, 2018. Global status report on road safety 2018.
Wu, B., Iandola, F., Jin, P., Keutzer, K., 2017. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 129–137.
Xu, C., Tarko, A., Wang, W., Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. Accid. Anal. Prev. 57, 30–39.
Xu, C., Wang, W., Liu, P., Guo, R., Li, Z., 2014. Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models. Transp. Res. Part C: Emerging Technol. 38, 167–176.
Xu, C., Liu, P., Yang, B., Wang, W., 2016. Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. Transp. Res. Part C: Emerging Technol. 71, 406–418.
Yan, Y., Chen, M., Shyu, M., Chen, S., 2015. Deep learning for imbalanced multimedia data classification. In: 2015 IEEE International Symposium on Multimedia (ISM). IEEE, pp. 483–488.
Yang, K., Wang, X., Yu, R., 2018. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. Transp. Res. Part C: Emerging Technol. 96, 192–207.
Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. Transp. Res. Rec. 2673 (4), 314–326.
Zhou, Z., Liu, X., 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 18 (1), 63–77.
Zhu, L., Guo, F., Krishnan, R., Polak, J., 2018. The use of convolutional neural networks for traffic incident detection at a network level. 97th Annual Meeting of the Transportation Research Board, Washington, DC.