Accident Analysis and Prevention 144 (2020) 105628

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Automated traffic incident detection with a smaller dataset based on generative adversarial networks

Yi Lin[a], Linchao Li[b,*], Hailong Jing[a], Bin Ran[c], Dongye Sun[d]

[a] National Key Laboratory of Fundamental Science on Synthetic Vision, College of Computer Science, Sichuan University, China
[b] Institute of Urban Smart Transportation & Safety Maintenance, Shenzhen University, China
[c] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, United States
[d] National Engineering Laboratory of Transportation Safety and Emergency Informatics, China

ARTICLE INFO

Keywords:
GANs
Imbalance training samples
Incident detection
Spatial and temporal rules

ABSTRACT

An imbalanced and small training sample can cause an incident detection model to have a low detection rate and a high false alarm rate. To solve the scarcity of incident samples, a novel incident detection framework is proposed based on generative adversarial networks (GANs). First, spatial and temporal rules are presented to extract variables from traffic data, which is followed by the random forest algorithm to rank the importance of variables. Then, some new incident samples are generated using GANs. Finally, the support vector machine algorithm is applied as the incident detection model. Real traffic data, which were collected from a 69.5-mile section of the I-80 highway, are used to validate the proposed approach. A total of 140 detectors are installed on the section enabling traffic flow to be measured every 30s. During 14 days, 139 incident samples and 946 nonincident samples were extracted from the raw data. Five categories of experiments are designed to evaluate whether the proposed framework can solve the small sample size problem, imbalanced sample problem, and timeliness problem in the current incident detection system. The experimental results show that our proposed framework can considerably improve the detection rate and reduce the false alarm rate of traffic incident detection. The balance of the dataset can improve the detection rate from 87.48% to 90.68% and reduce the false alarm rate from 12.76% to 7.11%. This paper lends support to further studies on combining GANs with the machine learning model to address the imbalance and small sample size problems related to intelligent transportation systems.

1. Introduction

The effective and efficient management of limited traffic resources is essential to saving travel time for travelers, which increases the necessity of a more accurate and stable travel service. A traffic incident is defined as a sudden event, that causes a reduction in capacity and results in traffic congestion. It can breakdown traffic flow, disturb traffic operation, and cause traffic delays as well as additional pollution. Due to the mentioned impacts, minimizing the effect of traffic incidents has become a major focus. Therefore, developing a real-time traffic incident detection algorithm has become one of the most important research areas.

Due to the dynamic characteristics of traffic flows, immediately and accurately detecting a traffic incident is not a straightforward task. Due to the common installation of traffic sensors on highways, tremendous traffic flow data with wide coverage are available. Three typical sources of traffic flow data are commonly used to detect traffic incidents, including fixed detectors (such as loop detectors), GPS devices (such as personal navigators), and automatic vehicle identification detectors (such as radio frequency identification) (Piccoli et al., 2015). The fixed detector installed to monitor traffic flow, speed, and occupancy at a certain location has been widely used as a data source in previous studies because of its wide coverage and reliable quality. Using data from fixed detectors, various techniques have been developed and applied for traffic incident detection. As demonstrated by Ma et al. (2015a), these methods can be divided into three categories: comparative algorithms, statistical algorithms, and artificial intelligence algorithms (Ghosh and Smith, 2014). A brief discussion of each category is given as follows.

1. Comparative algorithms: In this type of method, the difference in the traffic flow parameters between two adjacent fixed detectors is

* Corresponding author.
E-mail address: lilinchao@szu.edu.cn (L. Li).

https://doi.org/10.1016/j.aap.2020.105628
Received 25 September 2019; Received in revised form 31 May 2020; Accepted 2 June 2020
Available online 20 June 2020

calculated and compared. If the difference is an outlier, an incident might occur in the segment between the two detectors. For example, Payne et al. proposed the popular California algorithm (Samant and Adeli, 2000) and Persaud et al. built the McMaster algorithm based on catastrophe theory (Hall et al., 1993). The two algorithms were widely applied in the early stage of incident detection because they can be easily implemented and consume fewer computing resources. However, the main drawback of the algorithms is that their accuracy cannot meet the requirements of practical traffic management (Samant and Adeli, 2000).

2. Statistical algorithms: This type of method commonly utilizes the temporal characteristics of traffic flow data to build models based on the given statistical theory. Then, the traffic state can be predicted quickly, which is compared with the real value measured by detectors. If the difference between the real value and predicted value is higher than a preset threshold, an incident might occur. Time series algorithms are commonly used to capture the temporal characteristics in previous studies. For example, Ahmed et al. first studied the performance of the autoregressive integrated moving average model to detect traffic incidents using real-world data from the Lodge Freeway in Detroit (Ahmed and Cook, 1979). Chassiakos et al. proposed a detection algorithm called detection logic with smoothing based on a moving average filter (Chassiakos and Stephanedes, 1993). The transferability of the incident detection model has been studied and evaluated (Stephanedes and Hourdakis, 1996). Wang et al. converted traffic incident detection into a traffic state estimation problem that can be solved using a multiple model particle smoother (Wang et al., 2016a,b). To implement statistical algorithms, the threshold should be set manually, which highly depends on the experience of the operator. Moreover, it is difficult to consider both the spatial and temporal characteristics of traffic flow (Mak and Fan, 2007; Li et al., 2019, 2018) in this type of model.
3. Artificial intelligence algorithms: Recently, artificial intelligence-based techniques have experienced rapid application in transportation engineering, especially in traffic incident detection because they can be defined as a binary classification problem (Tang and Gao, 2005). The widely used algorithms include artificial neural networks (ANNs), support vector machines (SVMs), random forests (RFs), and their extended variations. For example, Ma et al. applied a deep neural network to detect traffic congestion of a highway network by considering both the temporal and spatial characteristics of the traffic flow (Ma et al., 2015b). Li et al. extended the traditional SVM based on the ensemble method to improve the performance of the incident detection model (Li et al., 2016a). Li et al. compared some widely implemented machine learning models to detect traffic incidents (Li et al., 2016b). The ensemble method can enhance the accuracy of incident detection models. For example, extreme gradient boosting was applied to detect accidents and analyze factors contributing to traffic accidents (Wu et al., 2020; Parsa et al., 2020). Artificial intelligence algorithms are flexible and data-driven, which means that the algorithms can fit a large quantity of traffic flow data to mine their intrinsic patterns (Huang et al., 2020). Moreover, this type of algorithm can capture both temporal and spatial information of traffic flow, which has been proven important to improve the accuracy in incident detection (Pan et al., 2013; Lv et al., 2015).

In summary, numerous traffic incident detection models have been built based on theory from different research areas. Artificial intelligence algorithms, which can obtain temporal and spatial dependencies simultaneously from traffic flow data, have been popular in recent years. Although incident detection models have developed rapidly, few studies have focused on traffic incident data extraction. In previous literature, three different datasets were always used by researchers to train and test models, including the I-880 dataset, AYE dataset, and simulated dataset. Some problems still need to be solved in relation to present traffic incident datasets, which are shown as follows.

1. **Problem 1: The most serious problem is the real-time characteristic of samples.** In the majority of previous studies, all samples during the incident were extracted and labeled as incident samples. For example, the duration of an incident is from 12:00:07 to 12:05:00, and the sample interval of the detector is 30 s. Then, 10 samples at 12:00:30, 12:01:00, 12:01:30, 12:02:00, 12:02:30, 12:03:00, 12:03:30, 12:04:00, 12:04:30, and 12:05:00 are labeled as incidents. However, to train a real-time incident detection model, only the sample immediately after the incident (sample at 12:00:30) should be used. From the perspective of the shock wave theory, some time is needed to spread the influence of the incident to distant sections. When the incident first occurs, the characteristics of traffic flow change very little, and thus, the data from the detector are only slightly different from the previous normal data. The difference in traffic flow data will increase as time elapses. Therefore, if all samples during the incident are used to train the model, the model might fail to detect the incident immediately.
2. **Problem 2: The small sample size affects the accuracy of incident detection models.** In the current traffic operation center, no specific system is developed to map traffic flow data to incident samples. These two types of data should be extracted manually from different data collection systems and then joined together by certain information. Therefore, it is time-consuming and laborious to extract necessary variables and further build a traffic incident dataset.
3. **Problem 3: The imbalance of nonincident samples and incident samples in the training dataset reduces the performance of machine learning models in incident detection.** In the real world, traffic data suffer from an imbalance that typically contains many more nonincident samples than incident samples. When facing this scenario, incident detection models often fail to provide the optimal classification result. Nonincident samples can be classified accurately, while incident samples may be distorted.

Deep learning is a rapidly developing area at an intersection of research into artificial intelligence, optimization, and pattern recognition. Generative models are also a class of deep learning models that have been thoroughly studied for many years, but they were not realized until deep learning was introduced to generate image samples and interpolations of high visual fidelity. The proposed generative adversarial networks (GANs) make it possible to generate new image samples based on the features of real data, which are almost the same as the raw images (Goodfellow et al., 2014; Radford et al., 2015; Mirza and Osindero, 2014). The GAN model also inspires us to reconsider the issue of traffic incident detection. Some real-time traffic incident samples can be generated based on real data with the GAN model, which can compensate for the scarcity of incident samples. Moreover, the generated incident samples can balance the dataset and deal with the imbalanced data problem. In previous studies, random oversampling and undersampling were widely used to generate a balanced dataset (Ozbayoglu et al., 2016; Mujalli et al., 2016). However, there are some issues with traditional resampling methods. On the one hand, oversampling can result in overfitting since samples of a minority class are duplicated. On the other hand, undersampling may cause some critical samples of the majority class to be deleted from the final dataset (Parsa et al., 2019). To cover the shortcomings of the methods, some extended oversampling and undersampling methods were proposed in previous studies. As presented in Basso et al. (2020), a synthetic minority oversampling technique was applied that oversampled the minority class by randomly generating synthetic samples among the minority class samples and their k nearest neighbors. In Kitali et al. (2019), random oversampling examples were applied that generated new samples by a smoothed bootstrap approach. All the above methods use part of the samples to generate new samples. In contrast, the GANs can

Table 1
Variables selected by the spatial and temporal rules (Li et al., 2020).

Variable	Notation	Index
Volume of upstream detector t before the incident	v_up_t	$(300 - t)/10 + 1$
Occupancy of upstream detector t before the incident	o_up_t	$(300 - t)/10 + 2$
Speed of upstream detector t before the incident	s_up_t	$(300 - t)/10 + 3$
Volume of the upstream detector just after the incident	v_up_0	31
Occupancy of the upstream detector just after the incident	o_up_0	32
Speed of the upstream detector just after the incident	s_up_0	33
Volume of the downstream detector t before the incident	v_dn_t	$(300 - t)/10 + 34$
Occupancy of the downstream detector t before the incident	o_dn_t	$(300 - t)/10 + 35$
Speed of downstream detector t before the incident	s_dn_t	$(300 - t)/10 + 36$
Volume of the downstream detector just after the incident	v_dn_0	64
Occupancy of the downstream detector just after the incident	o_dn_0	65
Speed of downstream detector just after the incident	s_dn_0	66
Mean traffic volume during 5 min of the upstream	m_v_up	67
Mean occupancy during 5 min of the upstream	m_o_up	68
Mean traffic speed during 5 min of the upstream	m_s_up	69
Standard deviation of the traffic volume during 5 min of the upstream	s_v_up	70
Standard deviation of the occupancy during 5 min of the upstream	s_o_up	71
Standard deviation of the traffic speed during 5 min of the upstream	s_s_up	72
Mean traffic volume during 5 min of the downstream	m_v_dn	73
Mean occupancy during 5 min of the downstream	m_o_dn	74
Mean traffic speed during 5 min of the downstream	m_s_dn	75
Standard deviation of the traffic volume during 5 min of the downstream	s_v_dn	76
Standard deviation of the occupancy during 5 min of the downstream	s_o_dn	77
Standard deviation of the traffic speed during 5 min of the downstream	s_s_dn	78
Difference of volume between the upstream and downstream detector just after the incident	v_up_dn	79
Difference of occupancy between the upstream and downstream detector just after the incident	o_up_dn	80
Difference of speed between upstream and downstream detector just after the incident	s_up_dn	81

In this table, t is selected from the following values: 30, 60, 90, 120, 150, 180, 210, 240, 270, and 300.

capture the distributions of all samples to generate new samples for the minority class, which can improve the diversity of samples.

In this paper, a novel traffic incident detection framework is proposed using random forest, a generative adversarial network, and support vector regression. The basic GANs are implemented to generate incident samples. Combined with numerous real nonincident samples, better performance is expected to be obtained for incident detection. Three major contributions of this paper are shown as follows:

1. The study attempts to build timely models for traffic incident detection. Unlike previous studies, the incident detection model in this paper is trained using the data extracted just after it happens. Therefore, the model is sensitive to the incident, which can improve the accuracy of the model.
2. To the best of our knowledge, this study is the first to use GANs to address sample scarcity in traffic incident detection studies. The experimental results show that the statistical characteristics of real samples and newly generated samples are similar, which proves the effectiveness of the proposed GANs.
3. The incident samples generated by the GANs are applied to eliminate the influence of the class imbalance. In real data, several nonincident samples can be obtained for one incident sample, which will impact the performance of the incident detection. Using generated incident samples, the imbalance of the raw dataset can be solved.

The rest of this paper is organized as follows. Section 2 illustrates the variable selection method, the SVM and the traffic incident data generation model. In Section 3, real-world data are introduced, and several experiments are designed to validate our proposed model. Section 4 examines the experimental results. Finally, in section 5, the study is concluded, and some limitations of this study are presented.

2. Methodologies

2.1. Variable selection of input data for incident detection models

Traffic incidents on highways not only cause severe damage but also result in vehicle delays in the upstream direction. Considering the change in traffic flow parameters, such as traffic volume, traffic speed, and occupancy, is fundamental to building automatic incident detection models. When an incident occurs, measurements of traffic detectors both upstream and downstream will experience a sudden change, but the upstream and downstream changes are different. In the upstream, the traffic volume and traffic speed will decrease, while the occupancy will increase. However, downstream, the traffic volume and occupancy will decrease and the traffic speed will increase (Karim and Adeli, 2002a,b). Therefore, both the spatial and temporal effects of incidents on traffic flow should be considered in automatic incident detection models. There is a need to select parameters on the upstream and downstream segments from temporal and spatial perspectives.

Variables of the dataset to train incident detection models are selected based on the following specific temporal and spatial rules. **Spatial rule 1:** traffic flow parameters extracted from the adjacent upstream detector and downstream detector are considered. These two detectors are more sensitive than distant detectors in that detection models can recognize incidents with less delay. It can be inferred from the shockwave theory that some time is needed to spread the influence of the incident to distant detectors. Therefore, the times of traffic flow parameters of adjacent detectors change earlier than the times of distant detectors. The time interval of the data is 30 s in this study, which is so short that the influence of the incident cannot spread to distant detectors. **Spatial rule 2:** combinations of traffic flow parameters extracted from upstream and downstream detectors are also selected as potential variables. It has been proven that combinations can contribute to the accuracy of the detection model; for example, California algorithms apply the difference between the occupancy of two adjacent detectors as one of the variables. The temporal correlation of the traffic flow that has always been ignored in previous studies is also considered to generate the training dataset in this paper. **Temporal rule 1:** traffic

flow parameters of adjacent detectors, which are collected 5 min before the incident, are selected as variables. For example, if an incident occurred at 14:05:14, the traffic flow characteristics at 14:05:30, 14:05:00, 14:04:30, 14:04:00, 14:03:30, 14:03:00, 14:02:30, 14:02:00, 14:01:30, and 14:01:00 will be considered. Three traffic flow parameters, including the traffic flow, the traffic speed, and the occupancy used in this paper, are observed every 30 s, and thus, $3 \times 10 = 30$ variables can be obtained for each sample by this rule. The interval of 5 min is selected because in previous studies, it was proven that traffic conditions begin to change 5 min before the incident (Abdel-Aty et al., 2004). **Temporal rule 2:** the mean and standard deviation of the traffic volume during 5 min, the mean and standard deviation of the traffic speed during 5 min, and the mean and standard deviation of the occupancy during 5 min are calculated and selected as potential variables (Abdel-Aty et al., 2004).

As shown in Table 1, 81 potential variables are selected to build detection models by using the four proposed rules. However, most machine learning and statistical models may not be effective for high-dimensional data. In this work, random forest is proposed to rank the importance of potential variables, and then some variables with less importance are deleted. The criteria for selecting variables are introduced in the following section.

Random forest is a widely used ensemble learning algorithm for classification and regression in traffic engineering. It is extended from the classification and regression tree (CART) and can rank the importance of variables. In this study, traffic incident detection can be treated as a classification problem. Assume a learning set $DS = (X_1, Y_1)$, .. ., (X_n, Y_n) is the observations of a random vector (X, Y). Vector $X = (X^1, .. ., X^{81})$ is the explanatory variable selected by our defined rules. $Y = (0, 1)$ is the class label where 1 represents the incident sample and 0 represents the nonincident sample. The principle of the random forest is to combine a series of binary decision trees. The difference of the random forest from CART is that the random forest uses bootstrap samples of DS and selects a subset of X randomly at each node which makes the random forest more robust and is stable as a perturbation of the training sample, and thus, it can handle the small sample set. Before the generation of the training dataset, the sample size is small, and thus, the random forest is selected to quantify the importance of variables. For each tree t of the forest, data not used in the bootstrap sample is defined as the OOB_t sample. Let $errOOB_t$ denote the error of tree t on this OOB_t sample. Then, randomly permute the values of vector $X^i (i = 1, \ldots, 81)$ in OOB_t to obtain a perturbed sample OOB_t^i and calculate errOOB_t^i. Finally, the variable importance of X^i is defined as:

$$\mathrm{VI}(X^i) = \frac{1}{N}\sum_t (\mathrm{errOOB}_t^i - \mathrm{errOOB}_t) \tag{1}$$

where N is the number of trees of the random forest. After the calculation, scores of variable importance are sorted in decreasing order, as shown in Fig. 1. The importance of the variable is so small after the 28th variable, and thus, these variables can be ignored when building incident detection models. Therefore, the top 28 variables with high importance are selected to generate incident samples and build the incident detection model.

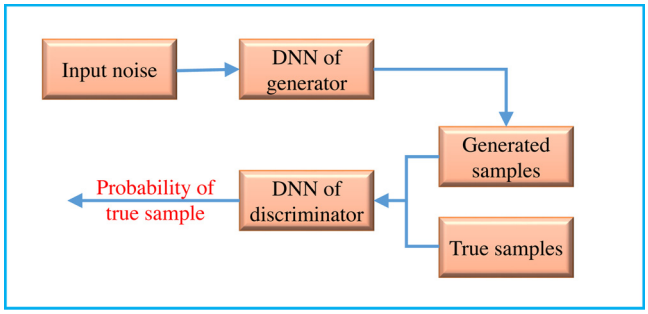

Fig. 2. General structure of GANs.

2.2. GANs

The GAN is a framework for training generative models, which was first proposed to solve the problem of training sample scarcity. The core idea of GANs is to set up a game between two adversarial models. One of them is called the generator (G), and the other is the discriminator (D). As the name suggests, the generator generates the desired data from the input noise, which may be sampled from an arbitrary random distribution or prior distribution of certain tasks. The generator attempts to make the output samples that are intended to come from the same distribution of the training data. The purpose of the discriminator is to classify the input data into two categories: real data and fake data. The training procedure of the GANs is to adjust the parameters of the model that represent the most likely distribution of training data by the competition and balance of two models. The general structure of the GANs is shown in Fig. 2. The input noise is converted into generated samples by the deep neural network of the generator. Both real samples and generated samples are fed into the discriminator to compute a scalar that denotes the probability of given samples drawn from the distribution of real samples.

The hidden layers of the deep neural network (DNN) for both the generator and the discriminator in this paper are fully-connected (FC) layers. The feedforward inference rules of the FC layer can be expressed mathematically in the following equations.

$$z_{i,j} = \sum_{l=1}^{L_{j-1}} w_{i,l} a_{l,j-1} + b_{j-1} \tag{2}$$

$$a_{i,j} = f(z_{i,j}) \tag{3}$$

where the notations $z_{i,j}$ and $a_{i,j}$ are the ith activation and output neurons in the jth layer, respectively. L_{j-1} saves the total number of neuron

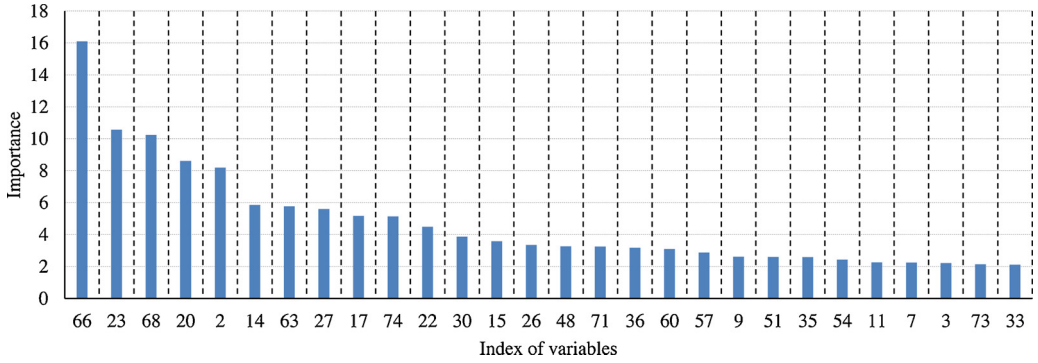

Fig. 1. Top 28 variables selected by random forest.

units in the $(j-1)$th layer. $w_{i,l}$ denotes the connection weight from the lth neuron of the $(j-1)$th layer to the ith neuron of the jth layer. b_{j-1} is the bias of the $(j-1)$th layer. $f(\cdot)$ is the nonlinear activation function, which is typically sigmoid, hyperbolic tangent or rectified linear units.

The optimization rules to train the GANs with the backpropagation algorithm (Rumerlhar, 1986) can be summarized as follows. The first and second terms of Equation (4) represent the optimization goal of the discriminator and the generator, respectively. Equations ((5)-(6)) shows the stochastic gradient descent calculation of the backpropagation algorithm.

$$\min_G \max_D L(G, D) = E_{x\sim p_{\text{data}}(x)} \log(D(x)) + E_{z\sim p_z(z)} \log(1 - D(G(z))) \tag{4}$$

$$\nabla\varphi_d \frac{1}{m} \sum_{i=1}^{m} [\log(D(x^{(i)})) + \log(1 - D(G(z^{(i)})))] \tag{5}$$

$$\nabla\varphi_g \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)}))) \tag{6}$$

where m is the minibatch size during the model training. φ_d and φ_g are parameters of the proposed discriminator and the generator that need to be optimized by the training dataset. With the training process, the optimization makes the distribution of the generative data $p_G(x)$ close to the distribution of the ground truth data $p_{data}(x)$. In other words, the ideal output probability of the discriminator is 0.5 regardless of whether the input sample is real or generated. In this paper, noise inputs of the generator are sampled from a uniform distribution with 100 dimensions, while the output is a generated fake sample whose dimension is the same as the result of the number of selected variables (28 dimensions). The generator consists of several FC layers with rectified linear unit (ReLU) activation. Input vectors of the discriminator are real incident samples and generated incident samples, and the output of the discriminator is 2D one-hot vectors that represent the probability of given samples from the distribution of collected data. The discriminator contains FC layers with batch normalization and dropout layers to speed up the training process and prevent overfitting.

The implementation of the GANs is based on the open-source deep learning framework Keras 2.0.4 with the TensorFlow1.0.0 backend. All parameters (weights and biases) are initiated from the Glorot uniform distribution. An Adam optimizer implemented by Keras is used to optimize the generative model, whose initial learning rate, beta_1, beta_2, and decay of learning rate are 0.0001, 0.9, 0.99, and 0.99, respectively. The configuration of the optimizer for the discriminator is the same as the generative model except for the 0.001 initial learning rate. According to the following common skills for selecting the number of neurons in hidden layers, the detailed configurations of the proposed GANs are listed in Table 2. Regarding the selection of the number of neurons in the classification model (D), $h = \sqrt{N_i N_o}$ is used, where h denotes the number of hidden units in a hidden layer and N_i and N_o are the input and output dimensions of this hidden layer. $h_l = 0.75 \times h_{l-1}$ is used for regression model (G), where h_l saves the number of units in the l^{th} layer. Each hidden layer of the discriminator is followed by a batch normalization layer to speed up the training process. A dropout layer with a 0.25 dropout rate is also appended to alleviate the overfitting problem. The loss functions of the generator and discriminator are the mean squared error (MSE) and binary cross-entropy, respectively. Additionally, the batch size is selected as 64 during the training process.

Table 2
Parameters of GANs.

Model	Input	Output	Neurons
G	(None,100)	(None,28)	[80, 64, 40]
D	(None,28)	(None,2)	[8, 4]

G: Generator, D: Discriminator. None is the batch size. Length of neurons is the number of hidden layer while digitals in neurons denote the number of neurons in different hidden layers.

2.3. Support vector machine

SVM is a powerful method for solving a nonlinear classification problem. SVM was selected because it has proven to be highly effective in previous studies (Chen et al., 2016, 2009; Yuan and Cheu, 2003). In this section, a simple introduction is given. A detailed introduction of SVM can be found in previous studies (Chen et al., 2016, 2009; Yuan and Cheu, 2003).

SVM attempts to find a function $f(\mathbf{x})$ that can best approximate the relationship between the input variables and the output variable. The function can be written as (Noble, 2006):

$$f(\mathbf{x}) = w^T \Phi(\mathbf{x}) + b \tag{7}$$

where w and b are coefficients. The coefficients can be solved by the following optimization:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \tag{8}$$

$$\text{Subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{9}$$

where ξ_i is the slack variable and C is the weight of the penalty. The above function can be solved by Lagrange multipliers, and the procedures can be found in Noble (2006).

3. Experimental design

3.1. Data description

In this study, the raw traffic flow and information of incident data were extracted from the Caltrans Performance Measurement (PeMS) following the rules introduced in the methodologies section. Interstate-80 (I-80) in the California area under consideration is 69.5 miles and has a total of 140 mainline loop detector stations in each direction. Consecutive stations have been installed at approximately 0.5-mile intervals. Loop detector stations measure the traffic volume, traffic speed, and occupancy of corresponding lanes. All reported incidents on I-80, from 01/01/2018 to 01/13/2018, were archived in the PeMS incident database. The average distance between incidents and upstream detectors is approximately 0.247 miles, and the average distance between incidents and downstream detectors is approximately 0.244 miles. The corresponding parameters of the traffic flow were obtained by the data Clearinghouse tool. Considering that this study aims to investigate the relationship between the characteristics of traffic flow and traffic incidents, incidents associated with work zones were removed from the raw data. Moreover, some traffic flow data were missed when some incidents occurred. These incidents were also deleted from the raw data. After the above two steps, 139 incident samples were extracted and acted as real incident samples in this work.

To extract nonincident data, the case-control strategy was applied in this study (Abdel-Aty et al., 2004). Nonincident samples were extracted at the same locations under similar weather conditions as incident samples. Moreover, the time of the nonincident sample was also considered. The nonincident samples were selected a day or days ahead of the incident sample, and the highest number of days ahead was 14. The difference in traffic conditions on weekdays and weekends was also considered. If an incident occurred on a weekday, the corresponding nonincident sample should also be on a weekday. Several nonincident samples could be extracted from the raw data for one corresponding incident sample. Finally, a total of 946 nonincident samples were obtained from the raw data. In the raw training dataset, the ratios of incident samples and nonincident samples were not equal, which may

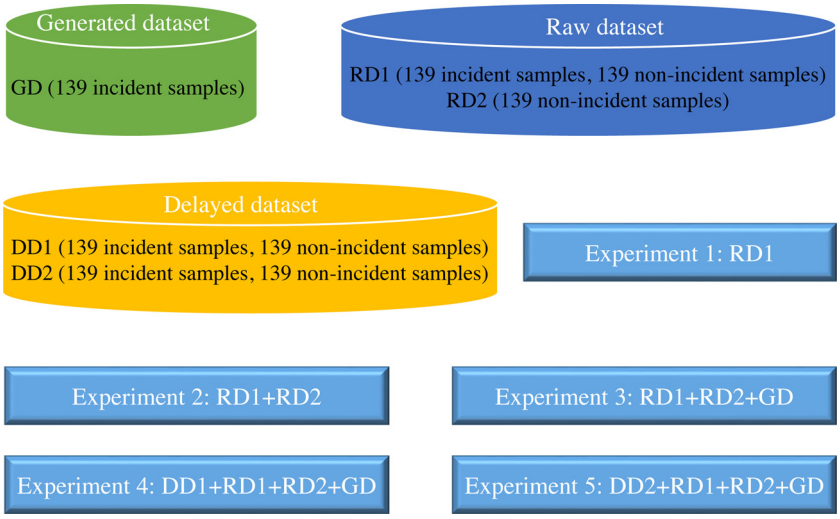

Fig. 3. Data description and experimental design.

confuse the machine learning-based incident detection models. Therefore, GANs were developed to generate incident samples to handle the imbalance problem of the training dataset. As shown in Fig. 3, 139 real incident samples and corresponding nonincident samples were defined as RD1. Another 139 real nonincident samples were defined as RD2. The final dataset was the delayed dataset containing DD1 and DD2. DD1 contains 139 one-minute delayed incident samples and 139 non-incident samples. DD2 contains 139 two-minute delayed incident samples and 139 nonincident samples. The one-minute or two-minute delayed dataset means its samples were extracted one or two minutes after incidents. In previous studies, delayed samples were extracted to increase the sample size. In this study, these samples were not used to improve the real-time ability of the model.

3.2. Evaluation criteria

In this study, the performance of traffic incident detection models was evaluated by three measures: accuracy (ACC), detection rate (DR), and false alarm rate (FAR). Their equations are given by:

$$\mathrm{ACC} = \frac{\text{Number of correct classifications}}{\text{Number of all training samples}} \times 100\% \tag{10}$$

$$\mathrm{DR} = \frac{\text{Number of incident samples detected}}{\text{Number of incident samples}} \times 100\% \tag{11}$$

$$\mathrm{FAR} = \frac{\text{Number of false classifications}}{\text{Number of incident samples}} \times 100\% \tag{12}$$

ACC denotes the percentage of all correct classifications for both incident and nonincident samples. DR represents the percentage of incidents that were precisely detected by the models. When DR is close to 100%, the model performs well. However, a higher DR may also indicate that the model is sensitive, which may cause false alarms. The denominator of the FAR criterion is the total number of detections occurring during a given period. It is calculated to evaluate how many incident alarms are falsely set.

3.3. Experiments

In this study, five experiments were designed to evaluate the proposed traffic incident detection framework shown in Fig. 3.

1. **Experiment 1:** Only RD1 was used as the input of the proposed model.
2. **Experiment 2:** The raw dataset containing RD1 and RD2 was used as the input of the proposed model.
3. **Experiment 3:** In addition to the raw dataset, the generated dataset was also selected as the input.
4. **Experiment 4:** DD1, RD1, RD2, and GD were used as the input.
5. **Experiment 5:** DD2, RD1, RD2, and GD were used as the input.

4. Results

4.1. Experimental results of GANs

According to the experimental design and implementation details of our proposed GANs, the proposed GAN model with 100 epochs was developed, and 139 samples were generated. To eliminate the randomness of the neural network, the model was trained five times, and five different generated incident datasets containing $5 \times 139 = 695$ generated samples were obtained. The five datasets were averaged as final results. To evaluate the generated dataset, a box plot was used to describe the distribution of given variables, including the minimum, lower quartile, median, upper quartile, and maximum. Since different variables have different units in this research, all values were normalized based on Eq., in which i is the index of the data record, which ranges from 1 to 139, and j is the index of the variable selection result, which ranges from 1 to 28 (Fig. 1). $x^j_{\max}$ and $x^j_{\min}$ denote the maximum and minimum values of the jth variable, respectively.

$$\hat{x}_i^j = \frac{x_i^j - x_{\min}^j}{x_{\max}^j - x_{\min}^j} \tag{13}$$

The box plots of real data and generated data are displayed in Figs. 4 and 5 , respectively. From the two figures, it can be seen that their distributions are similar to each other, which means that the generated data pose inherent patterns of real data. The analysis of variance (ANOVA) was also applied to evaluate whether the data were significantly similar. We found that most of the variables (except for s up 90, o up 150, s up 300, and s up 0) in the raw dataset were similar to their corresponding variables in the generated dataset at the 0.05 level. The mean and median difference rates between real data and generated data of given variables were also measured to evaluate the results of our proposed GAN model. The different rates of variables are reported in Fig. 6. The mean and median difference rates of occupation-related variables (such as o_up_90, o_up_120) were higher than those of others since those variables were all very small (basically 10^{-2}), which was very sensitive to any data fluctuation (corresponding to random input noises) during the data generation. However, from the perspective of sample diversity, differences between ground truth data and generated

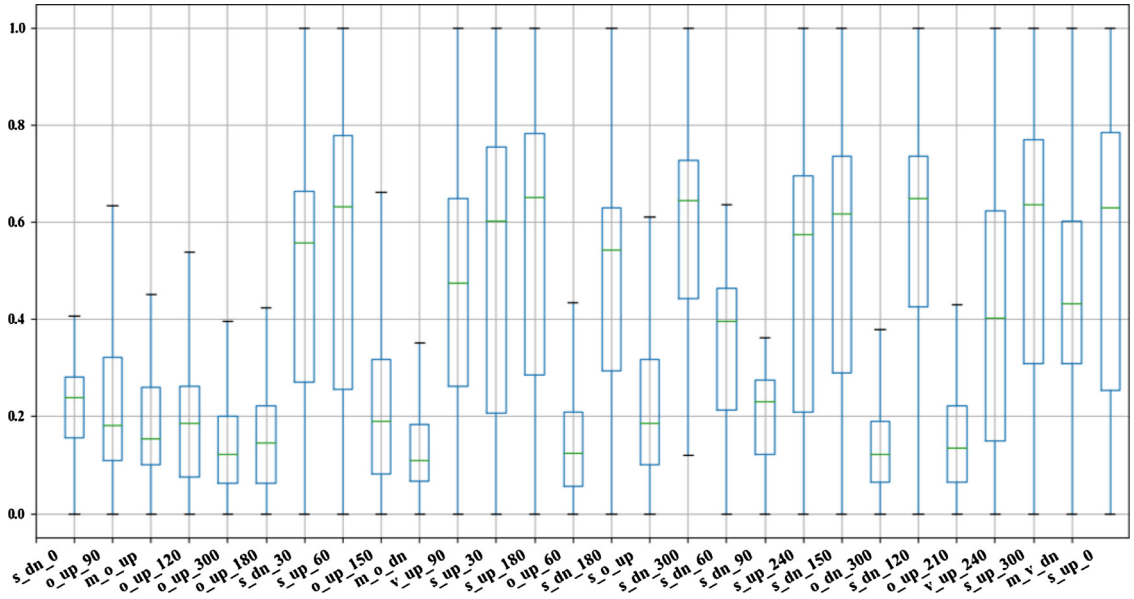

Fig. 4. Box plot of real incident data.

data were expected to provide more distinguishing features of traffic during traffic incidents, which can significantly promote the robustness of the incident detection model when facing unknown traffic patterns. This is also what the GAN model attempts to do.

4.2. Experimental results of the incident detection model

In this section, several experiments were conducted to confirm the important roles of the sample size, sample balance, and timeliness of the traffic data in the incident detection model. In this paper, the support vector machine (SVM) was applied to detect traffic incidents.

The parameters of the incident detection model in different experiments were optimized by the mentioned training datasets, and the reported results were recognized by optimized models. Ten times cross-validation (CV) was implemented in the experiment, and the average value was regarded as the final measurement. The experimental results are shown in Fig. 7, in which all measurements are described as a percentage. The optimized incident detection model during the training was used in all experiments to maintain the fairness of the comparison.

Experiment 1 and Experiment 2 were conducted to confirm the influence of the class imbalance. In Experiment 1, the average detection accuracy, DR, and FAR were 89.89%, 89.78%, and 9.98%, respectively, whereas the measurements in Experiment 2 were 91.53%, 87.48%, and 12.76%, respectively. From the experimental results, it can be seen that the detection accuracy improved by almost 2% when we used an additional 139 nonincident samples to train the model. However, because of the imbalance of incident and nonincident samples, the FAR deteriorated (approximately 3%) in Experiment 2, and the DR also decreased from 89.78% to 87.48% (over 2%). The improved detection accuracy (from 89.89% to 91.53%) can be explained by the increasing number of nonincident samples whose hidden patterns of traffic data are easier to recognize by the incident detection model. However, the imbalanced dataset has considerable negative impacts on the incident detection model, which is reflected by the low DR and the high FAR in Experiment 2.

The comparison of Experiment 1 and Experiment 3 verifies the

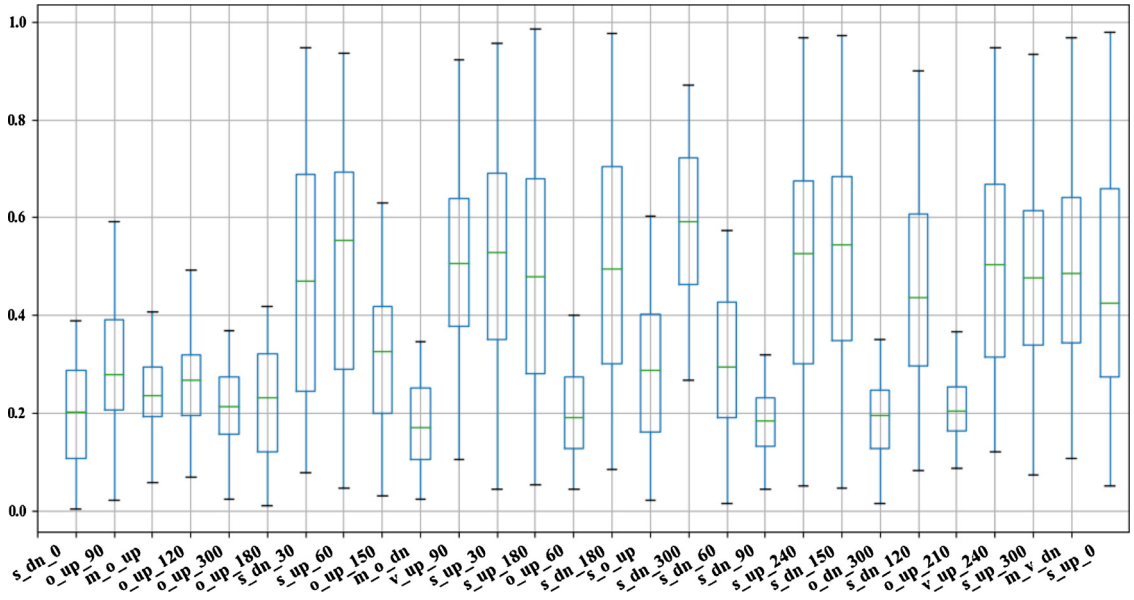

Fig. 5. Box plot of generated incident data.

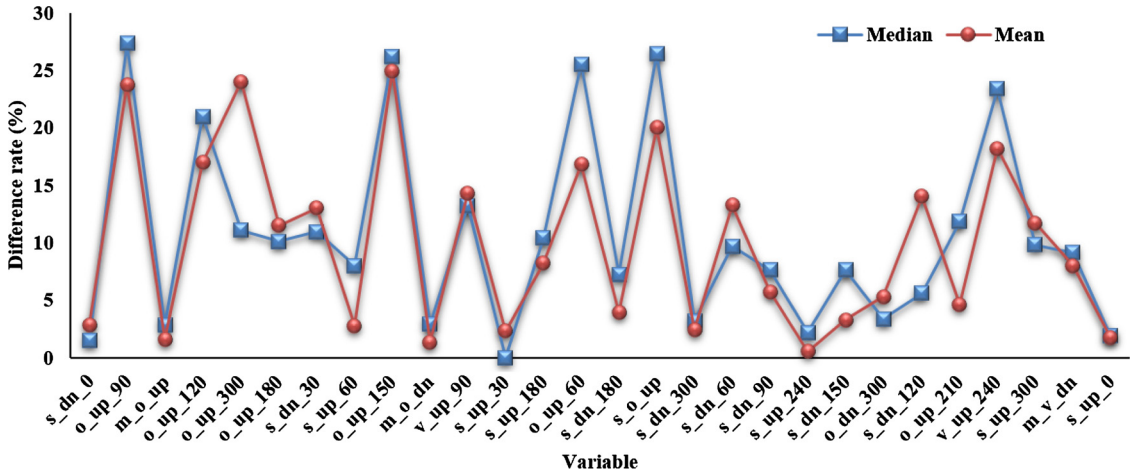

Fig. 6. Difference rate of median and mean values of generated incident data.

importance of the training sample size, which further supports the necessity of our proposed data generation model. In Experiment 3, the average detection accuracy, DR, and FAR were 91.87%, 90.68% and 7.11%, respectively. Therefore, compared with the results of Experiment 1, the overall performance (the average detection accuracy, DR, and FAR) of Experiment 3 improved by approximately 2% since the generated incident samples were used to build a balanced training dataset. The results show that more training samples can improve the performance of the incident model. More importantly, the results prove that incident samples generated by the proposed GANs are conducive to incident detection. From the first 3 experiments, it can be concluded that a balanced training dataset that has the same quantity of incident samples and nonincident samples can improve the performance of the incident detection model compared to using the datasets that only increase the number of training samples.

The delayed data were always used to train the incident detection models. However, a real-time incident detection model was developed in this paper. Thus, Experiments 4 and 5 were conducted to further analyze the effect of delayed data on the accuracy of the real-time model. Training datasets for Experiment 4 and Experiment 5 were generated by combining the training dataset in Experiment 3 with 139 one-minute and 139 two-minute delayed incident samples. In addition, to ensure the balance of training samples, another 139 nonincident samples were also added to datasets (4) and (5) to train the detection model. The comparison results show that detection accuracies decreased by over 2%, from 91.87% in Experiment 3 to 89.40% in Experiment 4 and to 89.22% in Experiment 5. Similarly, the DRs also decreased more than 2%, from 90.68% in Experiment 3 to 88.20% in Experiment 4 and to 88.18% in Experiment 5. Moreover, the FARs also decreased from 7.11% in Experiment 3 to 9.60% in Experiment 4 and to 9.96% in Experiment 5. There were only 556 training samples (278 incident samples and 278 nonincident samples) in Experiment 3, while the number of training samples in Experiment 4 and Experiment 5 was 834 (417 incident samples and 417 nonincident samples). The increase in delayed training incident samples reduces the overall detection performance, which further proves that the immediacy of training incident samples is more important than the number of training samples for the incident detection model. The overall performances in Experiment 4 and Experiment 5 were even worse than the baseline results (Experiment 1), whose training sample number was only one-third (139 incident samples and 139 nonincident samples) of the former experiments. Moreover, the results of both evaluation measurements in Experiment 4 were slightly better than those in Experiment 5, which indicates that the longer the incident data delays, the more detection performance that will be lost. From the experiments, a conclusion can be drawn that the detection model can obtain the best performance when we prepare a balanced training dataset with timely incident samples.

5. Conclusions

Traffic incident detection has been a popular topic in many previous studies and applications in recent decades. However, the small sample size problem still remains because it is difficult to obtain the same number of incident samples as nonincident samples. Moreover, the imbalance of the sample also negatively affects the accuracy of incident detection models. In this study, a method to address the small sample size and the imbalance problem of training samples was proposed. First, several temporal and spatial rules to extract characteristics of the traffic flow were introduced. Then, a variable selection method was built based on the random forest. Finally, the generative adversarial network was applied to produce similar incident samples, which can extract

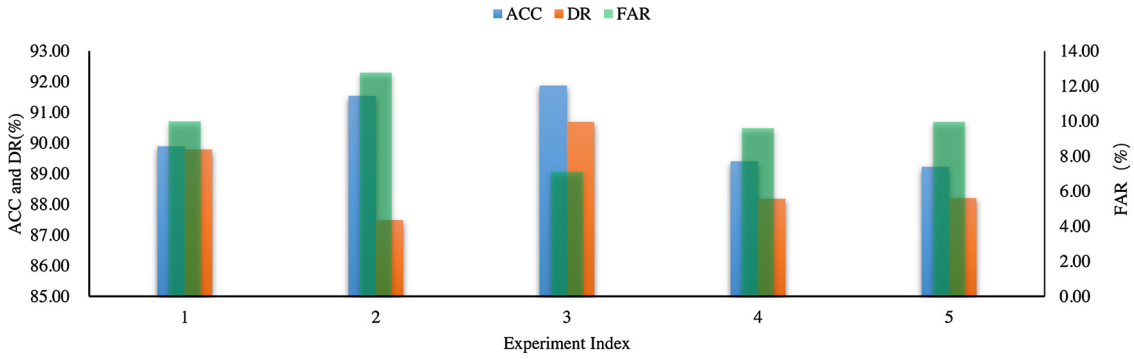

Fig. 7. Performance of incident detection model for different experiments.

characteristics from real incident samples. To validate the proposed approach, real-world traffic data on the I-80 highway were fed into the GAN model, which was further applied to achieve incident detection by an SVM-based classifier. Experimental results demonstrated that the accuracy of the traffic incident detection model can be improved due to the newly generated incident samples of the GAN model. The DR improved from 87.48% to 90.68%, and the FAR decreased from 12.76% to 7.11%.

Notably, only the SVM was applied as the incident detection model to evaluate the proposed method in this paper. In the future, more incident detection models should be implemented to test the proposed method. Moreover, the traffic flow of urban roads is more complex. The application of our proposed method in this area needs to be discussed in the future.

Acknowledgments

This work was supported by the Sichuan scientific and technological transformative project (No. 2017CC0004), the key scientific research of artificial intelligence of Sichuan (No. 2018GZDZX0029), and the Shenzhen science and technology program (No. KQTD20180412181337494).

References

Abdel-Aty, Mohamed, Uddin, Nizam, Pande, Anurag, Abdalla, M. Fathy, Hsia, Liang, 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transport. Res. Rec. 1897 (1), 88–95.

Ahmed, Mohammed S., Cook, Allen R., 1979. Analysis of freeway traffic time-series data by using Box-Jenkins techniques. Number 722.

Basso, Franco, Basso, Leonardo J., Pezoa, Raul, 2020. The importance of flow composition in real-time crash prediction. Accident Analysis & Prevention 137, 105436.

Chassiakos, Athanasios P., Stephanedes, Yorgos J., 1993. Smoothing algorithms for incident detection. Transport. Res. Rec. 1394, 8–16.

Chen, Shuyan, Wang, Wei, Van Zuylen, Henk, 2009. Construct support vector machine ensemble to detect traffic incident. Expert Syst. Appl. 36 (8), 10976–10986.

Chen, Cong, Zhang, Guohui, Qian, Zhen, Tarefder, Rafiqul A., Tian, Zong, 2016. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accident Anal. Prevent. 90, 128–139.

Ghosh, Bidisha, Smith, Damien P., 2014. Customization of automatic incident detection algorithms for signalized urban arterials. J. Intel. Transport. Syst. 18 (4), 426–441.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, Bengio, Yoshua, 2014. Generative adversarial nets. Advances in neural information processing systems 2672–2680.

Hall, Fred L., Shi, Yong, Atala, George, 1993. On-line testing of the mcmaster incident detection algorithm under recurrent congestion. Transport. Res. Rec. 1394, 1–7.

Huang, Tingting, Wang, Shuo, Sharma, Anuj, 2020. Highway crash detection and risk estimation using deep learning. Accident Anal. Prevent. 135, 105392.

Karim, Asim, Adeli, Hojjat, 2002a. Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm. J. Transport. Eng. 128 (1), 21–30.

Karim, Asim, Adeli, Hojjat, 2002b. Incident detection algorithm using wavelet energy representation of traffic patterns. J. Transport. Eng. 128 (3), 232–242.

Kitali, Angela E., Alluri, Priyanka, Sando, Thobias, Wu, Wensong, 2019. Identification of secondary crash risk factors using penalized logistic regression model. Transport. Res. Rec. 2673 (11), 901–914.

Li, Linchao, He, Shanglu, Zhang, Jian, Yang, Fan, 2016a. Bagging-svms algorithm-based traffic incident detection. 16th COTA International Conference of Transportation Professionals.

Li, Linchao, Zhang, Jian, Zheng, Yuan, Ran, Bin, 2016b. Real-time traffic incident detection with classification methods. International Conference on Green Intelligent Transportation System and Safety 777–788.

Li, Linchao, Zhang, Jian, Wang, Yonggang, Ran, Bin, 2018. Missing value imputation for traffic-related time series data based on a multi-view learning method. IEEE Trans. Intel. Transport. Syst. 20 (8), 2933–2943.

Li, Linchao, Qin, Lingqiao, Qu, Xu, Zhang, Jian, Wang, Yonggang, Ran, Bin, 2019. Day-ahead traffic flow forecasting based on a deep belief network optimized by the multi-objective particle swarm algorithm. Knowledge-Based Syst. 172, 1–14.

Li, Linchao, Sheng, Xi, Du, Bowen, Wang, Yonggang, Ran, Bin, 2020. A deep fusion model based on restricted boltzmann machines for traffic accident duration prediction. Eng. Appl. Artif. Intel. 93, 103686.

Lv, Yisheng, Duan, Yanjie, Kang, Wenwen, Li, Zhengxi, Wang, Fei-Yue, 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intel. Transport. Syst. 16 (2), 865–873.

Ma, Xiaolei, Tao, Zhimin, Wang, Yinhai, Yu, Haiyang, Wang, Yunpeng, 2015a. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transport. Res. Part C: Emerging Technol. 54, 187–197.

Ma, Xiaolei, Yu, Haiyang, Wang, Yunpeng, Wang, Yinhai, 2015b. Large-scale transportation network congestion evolution prediction using deep learning theory. PLoS ONE 10 (3), e0119044.

Mak, Chin Long, Fan, Henry S.L., 2007. Development of dual-station automated expressway incident detection algorithms. IEEE Trans. Intel. Transport. Syst. 8 (3), 480–490.

Mirza, Mehdi, Osindero, Simon, 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Mujalli, Randa Oqab, López, Griselda, Garach, Laura, 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accident Anal. Prevent. 88, 37–51.

Noble, William S., 2006. What is a support vector machine? Nature Biotechnol. 24 (12), 1565–1567.

Ozbayoglu, Murat, Kucukayan, Gokhan, Dogdu, Erdogan, 2016. A real-time autonomous highway accident detection model based on big data processing and computational intelligence. 2016 IEEE International Conference on Big Data (Big Data) 1807–1813.

Pan, T.L., Sumalee, Agachai, Zhong, Ren-Xin, Indra-Payoong, Nakorn, 2013. Short-term traffic state prediction based on temporal-spatial correlation. IEEE Trans. Intel. Transport. Syst. 14 (3), 1242–1254.

Parsa, Amir Bahador, Taghipour, Homa, Derrible, Sybil, Mohammadian, Abolfazl Kouros, 2019. Real-time accident detection: coping with imbalanced data. Accident Analysis & Prevention 129, 202–210.

Parsa, Amir Bahador, Movahedi, Ali, Taghipour, Homa, Derrible, Sybil, Mohammadian, Abolfazl Kouros, 2020. Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. Accident Anal. Prevent. 136, 105405.

Piccoli, Benedetto, Han, Ke, Friesz, Terry L., Yao, Tao, Tang, Junqing, 2015. Second-order models and traffic data from mobile sensors. Transport. Res. Part C: Emerging Technol. 52, 32–56.

Radford, Alec, Metz, Luke, Chintala, Soumith, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Rumerlhar, D.E., 1986. Learning representation by back-propagating errors. Nature 323, 533–536.

Samant, A., Adeli, H., 2000. Feature extraction for traffic incident detection using wavelet transform and linear discriminant analysis. Comput.-Aided Civil Infrastructure Eng. 15 (4), 241–250.

Stephanedes, Yorgos J., Hourdakis, John, 1996. Transferability of freeway incident detection algorithms. Transport. Res. Rec. 1554 (1), 184–195.

Tang, Shuming, Gao, Haijun, 2005. Traffic-incident detection-algorithm based on nonparametric regression. IEEE Trans. Intel. Transport. Syst. 6 (1), 38–42.

Wang, Ren, Fan, Shimao, Work, Daniel B., 2016a. Efficient multiple model particle filtering for joint traffic state estimation and incident detection. Transport. Res. Part C: Emerging Technol. 71, 521–537.

Wang, Ren, Work, Daniel B., Sowers, Richard, 2016b. Multiple model particle filter for traffic estimation and incident detection. IEEE Transactions on Intelligent Transportation Systems 17 (12), 3461–3470.

Wu, Weitiao, Jiang, Shuyan, Liu, Ronghui, Jin, Wenzhou, Ma, Changxi, 2020. Economic development, demographic characteristics, road network and traffic accidents in zhongshan, china: gradient boosting decision tree model. Transport. A: Transport Sci. 1–33 (just-accepted):.

Yuan, Fang, Cheu, Ruey Long, 2003. Incident detection using support vector machines. Transport. Res. Part C: Emerging Technol. 11 (3-4), 309–328.