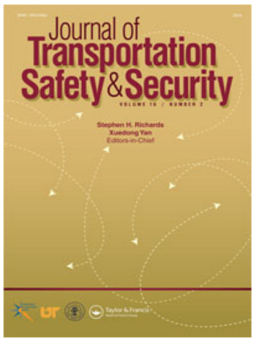

Taylor & Francis
Taylor & Francis Group

Journal of Transportation Safety & Security

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/utss20

# Utilizing angle-based outlier detection method with sliding window mechanism to identify real-time crash risk

Zhen Gao, Jingning Xu, Rongjie Yu & Lei Han

**To cite this article:** Zhen Gao, Jingning Xu, Rongjie Yu & Lei Han (2024) Utilizing angle-based outlier detection method with sliding window mechanism to identify real-time crash risk, Journal of Transportation Safety & Security, 16:2, 157-174, DOI: 10.1080/19439962.2023.2189762

**To link to this article:** https://doi.org/10.1080/19439962.2023.2189762

Published online: 27 Mar 2023.

JOURNAL OF TRANSPORTATION SAFETY & SECURITY
2024, VOL. 16, NO. 2, 157–174
https://doi.org/10.1080/19439962.2023.2189762

# Utilizing angle-based outlier detection method with sliding window mechanism to identify real-time crash risk

Zhen Gao[a*], Jingning Xu[a], Rongjie Yu[b], and Lei Han[b]

[a]School of Software Engineering, Tongji University, Shanghai, China; [b]The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Shanghai, China

**ABSTRACT**

Developing real-time crash risk models has been a hot research topic as it could identify crash precursors and thus triggering active traffic management strategies. Currently, crash risk identification models were mainly developed based upon supervised learning techniques, which requires large sample size of historical crash data. However, crashes are rare events in the real world, where the performance of supervised learning methods can be severely degraded to deal with the imbalanced sample. Besides, the data heterogeneity issue is another critical challenge. In this study, the unsupervised learning approach has been introduced to address unbalanced samples and data heterogeneity issues, and the experimental results has verified the effectiveness of the method. Data from the Shanghai urban expressway system were utilized for the empirical analyses. Several unsupervised learning methods were tested, among which, Angle-Based Outlier Detection (ABOD) model showed the best performance with 80.4% sensitivity and 25.4% false alarm rate (FAR). Considering the varying traffic flow distribution, dynamic ABOD with sliding window is further proposed, which improves the sensitivity by 6.3% and reduces the FAR by 8.1%. Finally, the proposed model is used to construct personalized road-level models, which achieve good performance despite the small sample size and severe sample imbalance.

**KEYWORDS**

Real-time crash risk analysis; Angle-Based Outlier Detection (ABOD); sliding window; temporal and spatial operational features; unsupervised learning

## 1. Introduction

Real-time crash risk identification has been regarded as a crucial procedure to enhance traffic safety in active traffic safety management system (Hossain et al., 2019). By exploring the relationships between the pre-crash traffic flow features and crash potentials, real-time crash risk models were developed to estimate crash probabilities for a given short period (e.g. 15 min or 20 min), the outputs could be used to trigger proactive safety

**CONTACT** Rongjie Yu yurongjie@tongji.edu.cn Zhejiang Scientific Research Institute of Transport; The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Shanghai, China
*Zhejiang Scientific Research Institute of Transport.

control strategies with the purpose of crash risk prevention (Karimpour et al., 2021; Katrakazas et al., 2015; Li et al., 2020). With the development of various advanced traffic sensing and management technologies, extensive real-time traffic data has become available and real-time crash risk analysis has gained increasing attention (Ahmed & Abdel-Aty, 2012; Oh et al., 2001; Shi & Abdel-Aty, 2015).

Recently, studies have mainly focused on developing crash risk models using supervised learning methods, including logistic regression (Abdel-Aty et al., 2004; Roshandel et al., 2015; Wang et al., 2015; Yuan et al., 2019), random forest (Lin et al., 2017), SVM (Yu & Abdel-Aty, 2013; Basso et al., 2018), MLP (Gao et al., 2018), Generalized Additive Model (Khoda Bakhshi & Ahmed, 2022), BP neural network (Ma et al., 2022), etc. The supervised learning approach has obtained decent estimation accuracies, but is very sensitive to unbalanced samples. Researches have shown that the model performance will significantly degrade or even fail completely in the case of insufficient historical crash risk samples (Krawczyk, 2016). To deal with the problem of imbalanced sample, some up-sampling and down-sampling methods such as "case-control" undersampling (Abdel-Aty et al., 2004), random undersampling (Gao et al., 2018), SMOTE (He and Garcia, 2009; Basso et al., 2018) and generative adversarial networks (Cai et al., 2019) have been used in the improvement of samples. The traditional range of under-sampling proportion is usually set to be 1:4 or 1:5 (Abdel-Aty et al., 2005; Ahmed et al., 2012; Ahmed & Abdel-Aty, 2012; Sun & Sun, 2015; Yang et al., 2018). Recent studies have also begun to experiment with more imbalanced ratios around 1:10 (Wang et al., 2015; Xu et al., 2014; 2016). It can be seen from existing studies that the performance of the model tends to get worse as the proportion of sample imbalance increases. It is evident from existing studies that the performance of the model tends to deteriorate as the proportion of sample imbalance increases. Overall, the supervised learning approaches, despite achieving good performance, are limited by their sensitivity to imbalanced samples and the large demand for sample size.

Meanwhile, information on the impact of site-specific factors and road characteristics on crash results is valuable for designing safety countermeasures (Buddhavarapu et al., 2016). Personalized modeling that takes road heterogeneity into account provides insight into the variability of empirical associations between road features and associated safety, and has recently provided better results (Behara et al., 2021). However, for roads with very low crash rates or newly paved sections, road-level personalized modeling is difficult using existing supervised learning methods due to the lack of historical crash risk samples. Therefore, it is critical to identify a new approach that is less dependent on historical crash risk samples.

In contrast to the widely adopted supervised modeling methods, modeling with unsupervised learning methods does not even require the participation of any historical crash risk samples. So these methods are generally insensitive to imbalanced samples and can be used to adapt the heterogeneity features, as there usually lacks for enough previous knowledge for supervised learning in these problems (Kriegel et al., 2008). However, most of the existing transportation related studies have used unsupervised learning methods in clustering data (Wz & Ww, 2019; Zhu et al., 2019) and conducting dimensional reductions (Boquet et al., 2020), which did not fully exploit the advantages of unsupervised learning.

In view of the above-mentioned issues, this study aims to explore the feasibility of unsupervised learning in real-time crash risk identification. The main contributions of this study are as follows:

1. Compared four classical unsupervised learning anomaly detection methods, one class support vector machine (SVM), isolation forest, autoencoder-based outlier detection and angle-based outlier detection (ABOD), to identify real-time crash risk, among which ABOD is found to have the best performance with an AUC of 0.83.
2. Improved ABOD with sliding window mechanism to consider the varying traffic flow distribution features at different times of the day, which provides an improvement of higher 6.3% sensitivity and 8.1% lower FAR.
3. Verified the adaptability of unsupervised learning methods to imbalanced samples and data heterogeneity problems.

The rest of the paper is organized as follows: In the Empirical Data section, it is explained how to build crash data. In the Methodology section, four classical unsupervised learning anomaly detection methods are introduced. In the Modeling Results and Analyses section, the model structure of ABOD with sliding windows is proposed and the experimental results are presented and analyzed. In the last section, the Conclusions and Discussion sections provide discussion and future work.

## 2. Data preparation

This study uses the coil traffic flow data and crash data detected from the Shanghai urban expressway system in May and June 2014. The Shanghai expressway system includes six urban expressways: Inner Ring Viaduct, Central Ring Viaduct, Yan'an Viaduct, Humin Viaduct, North-South Viaduct and Yixian Viaduct. The six urban expressways are divided into 237 sections according to the distribution of entrance and exit ramps, with an average section length of 949.2 m. Each section includes one or more coil detectors for collecting traffic flow data, with a data collection time interval of 20s, and the collected traffic flow data includes vehicle speed, traffic volume and occupancy rate. In the actual traffic flow data collection process, affected by equipment failure, the collected data may have some abnormalities, including data loss, data accuracy deviation, etc. Therefore, it is necessary to preprocess the original traffic coil data, including: (1) delete duplicate, unknown or invalid data due to coil failure or other reasons; (2) set a reasonable threshold range to remove data outside the range; (3) check the data consistency and remove abnormal data; (4) add the above deleted coil data using the spatial linear interpolation method; (5) correspond the coil data to the road section according to the coil number in order to facilitate the association of crash data. As the time interval of traffic flow data collection is too short, it is prone to lead to more data noise, so the original traffic flow data is merged into one segment every 5 minutes. In order to describe the traffic flow condition at a certain time, this paper extracts the traffic flow coil data of the first 20 min before that time point and divides it into 4 time-slice, named in turn as segments 1 to 4, with segment 1 being the closest to the current time. Calculate the mean, standard deviation of the speed and traffic volume of each segment in the current section, the upstream and downstream sections. Then calculate the sum of the traffic volume and the coefficient of variation of the speed respectively. Finally name a total of 72 variables according to the naming rules in Figure 1.

For Shanghai urban expressways, taking the road section as the basic analysis unit, assuming a real-time traffic crash risk prediction with frequency of 1 min, around 341.28 million traffic flow data records will be generated per day, and $1.02384\times10^8$ records per month. When a crash occurs, the record will be marked as 1. If there is no crash, it will be marked as 0. Since a crash will have an impact on the traffic conditions of the road section where the crash occurred, all records of the road section 1 hour after the crash were deleted to eliminate the data noise caused by the crash. The full sample data of Shanghai expressway in May 2014 used for

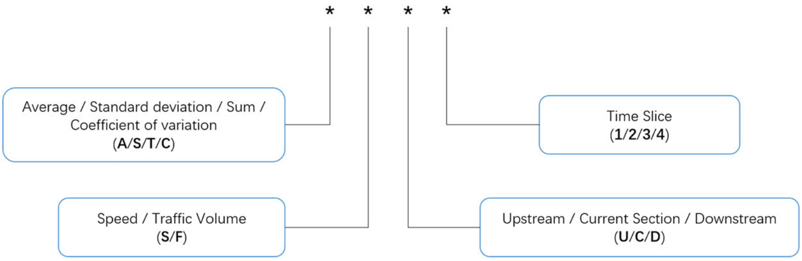
* * * *
Average / Standard deviation / Sum / Coefficient of variation
(A/S/T/C)
Time Slice
(1/2/3/4)
Speed / Traffic Volume
(S/F)
Upstream / Current Section / Downstream
(U/C/D)

Figure 1. Nomenclature rule for traffic variables.

the experiment consisted of 7559352 records, including 1,210 crash records and 7,558,142 non-crash records.

To investigate the effects of the imbalanced data, datasets with different crash to non-crash ratios were created, namely 1:1, 1:5, 1:10, 1:20, 1:100. In addition, for real-time crash risk analysis using our dynamic model, several sets of samples in continuous 60 minutes were randomly selected to simulate the real traffic environment. And it's ensured that there was at least one crash event in each set.

## 3. Methodology

In this paper, anomaly detection methods in unsupervised learning are used to perform real-time crash risk identification. In the experiments, the non-crash traffic flow data are considered as normal samples while the crash risk samples are regarded as abnormal samples. The 72-dimensional features extracted in the previous section are used as the input to the model. After screening, a total of four classical unsupervised learning anomaly detection methods are selected as the models to be tested, namely Angle-Based Outlier Detection (ABOD), one class SVM, Isolation Forest and Autoencoder-Based Outlier Detection. These four algorithms and the model performance evaluation methods are described below.

### 3.1. Angle-based outlier detection (ABOD)

Angle-Based Outlier Detection (Kriegel et al., 2008) is used to detect outliers in a large set of data objects. ABOD mitigates the effects of the "curse of dimensionality" by assessing the angle variance between difference vectors of different points instead of distances in the full-dimensional Euclidean data space.

Imagine a simple dataset, as shown in Figure 2. In the case of a point in a cluster, the angle between the difference vectors differs considerably, as shown by the angle α. The angle variance becomes smaller for points that

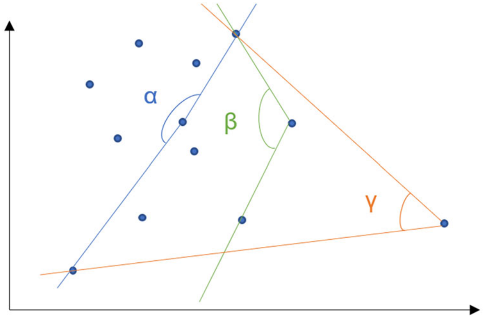

Figure 2. Intuition of angle-based outlier detection.

are at the boundary of the cluster, as indicated by the angle β. However, the variance is still relatively high compared to the angle variance of the true outliers, as shown by the angle γ. As most points are concentrated in certain directions, the angle to most point pairs will be small.

So an angle-based outlier factor (ABOF) is used to describe the divergence of objects in directions relative to each other. Given a database $\mathcal{D} \subseteq \mathbb{R}^d$, a point $\vec{A} \in \mathcal{D}$, and a norm $\| . \|: \mathbb{R}^d \to \mathbb{R}_0^+$. The scalar product is denoted by $\langle ., .\rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For two points $\vec{B}, \vec{C} \in \mathcal{D}$, $\overrightarrow{BC}$ denotes the difference vector $\vec{C} - \vec{B}$. The angle-based outlier factor $ABOF(\rightharpoonup A)$ is the angle variance between the difference vectors of $\vec{A}$ to all pairs of points in $\mathcal{D}$ :

$$ABOF\left(\vec{A}\right) = VAR_{\vec{B}, \vec{C} \in \mathcal{D}}\left(\frac{\langle \overrightarrow{AB}, \ \overrightarrow{AC}\rangle}{\|\overrightarrow{AB}\| \cdot \|\overrightarrow{AC}\|}\right) \tag{1}$$

In our model, the statistical information of traffic flow will be used as the input of the model in the form of a vector of length 72, which presents a point in high-dimensional space. By applying ABOD, we will tell the outliers from the full sample, which is considered as crash risk identification.

### 3.2. One class SVM

One class SVM (Schölkopf et al., 2001) was proposed to solve the One-class classification problem using the SVM approach. After the input data set X is mapped to the feature space H by the kernel function, the coordinate origin is first considered as the unique member of the outlier set, and then the distance between the member points of the sample set and the coordinate origin is calculated using the relaxation variable $\xi_i$. Those points that are close enough will be considered as members of the outlier set, i.e., abnormal samples. In this paper, we use the linear kernel function:

$$K\left(\vec{x}_i, \ \vec{x}_j\right) = \ \vec{x}_i \cdot \ \vec{x}_j \tag{2}$$

According to the formula:

$$\begin{aligned} \min W(a) = \ & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i y_i a_j y_j K\left(\vec{x}_i, \ \vec{x}_j\right) \\ \text{s.t.} \ \sum_{i=1}^{l} a_i = 1, \ & 0 \le a_i \le \frac{1}{\nu l}, \ i = 1, \ldots, l \end{aligned} \tag{3}$$

where $\nu \in (0, \ 1)$ is a balance parameter to maximize the distance from the coordinate origin to the normal sample dataset and to ensure that as many normal samples as possible are correctly classified. The resulting judgment function is:

$$f(x) = sgn\left[\sum_{i=1}^{1} a_i y_i K(\vec{x}_i,\ \vec{x}) - \rho\right] \tag{4}$$

where ρ can be calculated from any edge vector $\vec{x}_j$.

$$\rho = \sum_{i=1}^{l}\sum_{j=1}^{l} a_i y_i K(\vec{x}_i,\ \vec{x}_j) \tag{5}$$

For any given feature vector $\vec{x}$, if $f(x) \geq 0$, this feature vector belongs to a normal sample. Otherwise, it is an abnormal sample.

### 3.3. Isolation forest

The core of the Isolation Forest algorithm (Fei et al., 2008) is the construction of a forest (iForest) consisting of iTrees. iTree is a random binary tree, where each node is either a leaf or an internal node containing two children. The iTree is first constructed by randomly selecting an attribute A and a splitting value p from the data set $\mathcal{D}$. Each data object $d_i$ is then divided by the value of its attribute A, denoted as $d_i(\mathrm{A})$. It is placed in the left subtree if $d_i(\mathrm{A}) < \mathrm{p}$, and vice versa in the right subtree. The procedure is repeated until one of the following conditions is met. (1) there is only one data or multiple identical data left in D; (2) there is a maximum height of the tree.

In total, the Isolation Forest algorithm generates a specified number of iTrees and forms an iForest. Specifically, each iTree is constructed by randomly extracting a subset of $\mathcal{D}$ to ensure the diversity of iTrees. For the query object x, the leaf node of x is determined by traversing the set of iTrees in the iForest. Then, the anomaly score of x is calculated based on its path length, and the anomaly evaluation of x is performed.

Since the iTree is structurally equivalent to a binary search tree, the path length of the leaf node containing x is equal to the path length of the failed query in the binary search tree. Given the dataset $\mathcal{D}$, the path length of the failed query in the binomial search tree is:

$$c(n) = 2H(n-1) - 2\frac{n-1}{n} \tag{6}$$

where $H(k) = \ln(k) + \gamma$, $\gamma$ is the Euler constant, $c(n)$ is the average value of $P(x)$ for a given $n$, which is used to normalize $P(x)$. The anomaly score $s$ of the query object $x$ is shown in Equation (2):

$$s(x,\ n) = 2^{-\frac{E(P(x))}{c(n)}} \tag{7}$$

Where $E(P(x))$ is the average value of $P(x)$ in the iTree set.

### 3.4. Autoencoder-based outlier detection

Self-encoder is an unsupervised deep learning method in which the input to the model is also the optimization target of the model. In the training and optimization process, a higher-order abstract representation of the data information is obtained by compressing the intermediate layer nodes, and the consistency of the output and the input characterizes the ability of the features to restore the original signal (Lange & Riedmiller, 2010). The basic structure of the model is shown in Figure 3.

As can be seen from Figure 3, the self-encoder maps the input $X$ through f to obtain the new feature output $H$ by enoder; then $H$ is mapped through g to obtain the output $X'$ by decoder; the reconstruction error is minimized by iterative training to ensure that $X'$ approximates $X$ as much as possible.

The self-encoder can obtain the abstract representation of the input data and input the test features into the training model for identification, and use the mean square error as an index to measure the similarity of the input and output layers and compare it with the abnormality threshold. Since the self-encoder learns the internal laws of the features under normal operation, and the performance of the model deteriorates significantly during the reduction process of abnormal features, and it is necessary to issue a warning if the similarity between the input and output is low. The overall process of abnormal state recognition is shown in Figure 4.

### 3.5. Model performance evaluation

Finally, to evaluate the classification performance of outlier detection models, sensitivity and false alarm rate (FAR) are often utilized in crash risk analysis (Li et al., 2020). Sensitivity is defined as the percentage of the total number of crashes that are correctly predicted across all real crash cases. The FAR is the ratio of the total number of mispredictions of non-crash

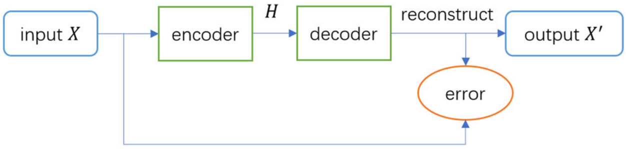

Figure 3. The training process of auto-encoder.

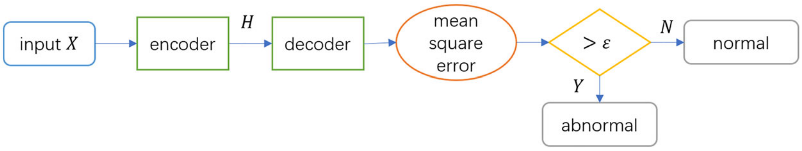

Figure 4. The testing process of auto-encoder.

**Table 1.** 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 |

across all predicted non-crash cases. The evaluation matrix is given in Table 1 for explanation.

$$Sensitivity = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Negative} \tag{8}$$

$$False\ \ Alarm\ \ Rate = \frac{Flase\ \ Positive}{False\ \ Positive + True\ \ Negative} \tag{9}$$

## 4. Modeling results and analyses

### 4.1. Results of four classical unsupervised learning methods

In this section, a total of four widely adopted anomaly detection methods, including one-class SVM, Isolation Forest and Autoencoder, are developed. Different from setting a threshold in supervised learning, anomaly detection in unsupervised learning requires setting the contamination rate of the data to determine the threshold to distinguish normal samples from abnormal samples. In our experiment, the contamination rate is fixed at 0.20. For example, in the ABOD method, it means a sample will be recognized as a crash event when the ABOF value is smaller than 80% of the training samples. The contamination rate was tuned by a grid search at 0.01 intervals from 0.1 to 0.5. It was found that when the contamination rate was set to 0.20, the sensitivity of the predictions and the FAR were best balanced.

The comparison results are shown in Table 2. It can be seen that ABOD provides the best AUC with 0.830 and one-class SVM holds the worst performance. This verifies the original ABOD paper that angles are more stable than distances in high-dimensional space (Kriegel et al., 2008). Therefore, ABOD was selected as the benchmark model for the following analyses.

### 4.2. Results of dynamic ABOD with sliding windows

In fact, the distribution of traffic low on a specific road segment varies with different times of one day (Wang et al., 2018). For instance, the distribution of traffic flow during the peak period is significantly different from the off-peak period. Therefore, a dynamic model based on Sliding Window Mechanism was further proposed, which was named dynamic ABOD.

Table 2. AUC Comparison between different unsupervised learning methods.

| Method | AUC |
|---|---|
| OneClassSVM | 0.676 |
| IsolationForest | 0.772 |
| AutoEncoder | 0.742 |
| ABOD | **0.830** |

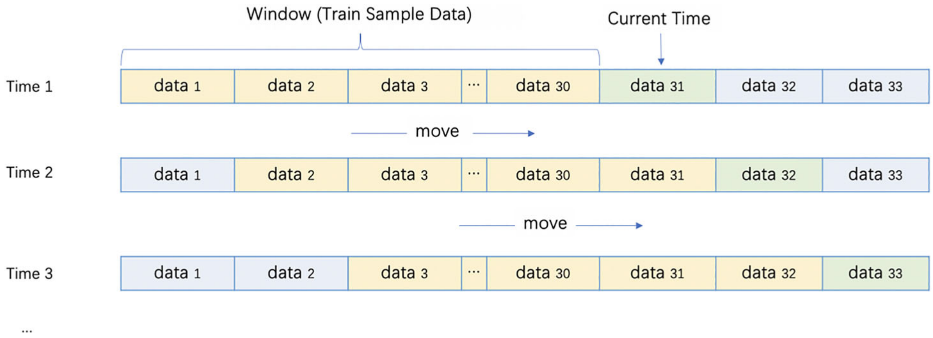

Figure 5. Dynamic modeling based on the sliding window mechanism.

In this study, taking into account the differences in traffic flow under different road sections and different time periods, we propose a dynamic modeling method based on ABOD using the sliding window mechanism.

In the dynamic model, the train sample set will be updated at a frequency of one minute. All training data is processed in a window. when updating, the window slides from left to right as shown in Figure 5. The data in the first window will be discarded, and new data will be added from the new last of the window.

To predict crash events at a certain time, samples from the past time should be selected as the training set. That is, each time the system makes a prediction, a new and different sample is selected to train the model.

The window size is critical to the dynamic model, which determines the number of training samples. For example, when the window size is set to 30, the traffic flow statistics for the past 30 minutes will be used as train samples. In the dynamic model, the contamination rate is set to 0, which means a sample will be recognized as a crash event when its ABOF value is less than all training samples. And several sets of continuous 60-minute data are selected to simulate a real-time analysis environment.

Experiments with different window sizes such as 10, 20, 30, 40, 50 minutes were performed and the results are shown in Table 3. As the window size increases, it showed substantial negative influences on sensitivity and positive influences on FAR. When the window size is too small, the dynamic model is very sensitive to the change in traffic flow and may tend to incorrectly identify more samples as crash events. It also should be noted that a larger window size allows the model to learn the long-term trend of

**Table 3.** Results in dynamic models under different window size settings.

| Window size | FAR | Sensitivity |
|---|---|---|
| 10 | 0.312 | 0.940 |
| 20 | 0.251 | 0.883 |
| 30 | **0.173** | **0.867** |
| 40 | 0.127 | 0.730 |
| 50 | 0.096 | 0.711 |

traffic flow in the past. But when the window size is set too large, our model will lose temporal information and degrades to the above city-level model or even worse. From the experimental results, it can be seen that 30 minutes is a relatively good value for window size which is highlighted in red, and will be used in the following experiments.

## 5. Discussions

### *5.1. Model sensitivity on different crash and non-crash ratios*

To verify the advantage of unsupervised learning being insensitive to imbalanced samples, we built models based on various ratios of crash and non-crash cases, namely 1:1, 1:5, 1:10, 1:20 and 1:100. Table 4 presents the sensitivity, FAR and AUC values for the different ratios.

It can be seen that the strong advantage of unsupervised learning being insensitive to imbalanced data is demonstrated when the model performance does not show large fluctuations under different crash and non-crash ratios. For instance, the AUC values under different ratios are all close to 0.829 and the FAR values vary from 0.252 to 0.283. It implies that the model tends to learn the statistical distribution information of the training samples, for which the ratio has minimal effect on the results.

### *5.2. Personalized modeling at road level*

As mentioned above, insensitivity to the ratio of the crash and non-crash cases is an important advantage of unsupervised learning. In addition, the essence of anomaly detection by ABOD is to learn the structural statistical distribution of normal data. So it doesn't even require any crash cases to train models, which means the method can be applied in a road segment where no crashes occurred. In contrast, supervised learning often requires a large number of high-quality crash risk samples. Therefore, the city-level model is preferred for supervised learning methods. It will be difficult for them to build an elegant road-level model and much less likely to develop a dynamic model like our paper.

And in fact, only a small number of normal samples is needed to complete the modeling. To explore how many train samples are sufficient for modeling, we performed an experiment and the results are shown in

**Table 4.** The summary of the modeling results on different ratios of crash and non-crash.

| Crash/Non-Crash Ratio | Sensitivity | FAR | AUC |
|---|---|---|---|
| 1:1 | 79.1% | 28.3% | 0.823 |
| 1:5 | 80.4% | 25.4% | 0.830 |
| 1:10 | 77.5% | 25.2% | 0.826 |
| 1:20 | 82.0% | 26.9% | 0.829 |
| 1:100 | 81.2% | 26.5% | 0.829 |

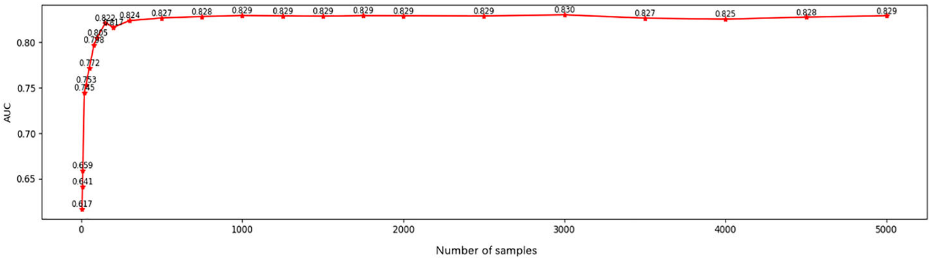

**Figure 6.** The AUC value under different numbers of train samples.

**Table 5.** Examples of classification results in roadway segment models.

| Road ID | FAR | Sensitivity |
|---|---|---|
| 21113 | 0.170 | 0.865 |
| 22104 | 0.166 | 0.880 |
| 22110 | 0.121 | 0.907 |
| 22218 | 0.258 | 0.707 |
| 23105 | 0.281 | 0.712 |
| 23209 | 0.182 | 0.891 |
| 24106 | 0.129 | 0.959 |
| 24125 | 0.160 | 0.932 |
| 24133 | 0.167 | 0.868 |
| 24215 | 0.139 | 0.922 |
| 24230 | 0.127 | 0.889 |
| Unified model | 0.254 | 0.804 |

Figure 6. In general, the performance of the ABOD model reaches the peak when the number of train samples exceeds around 500. So it means 500 samples are enough for our model to learn the structural statistical distribution of normal traffic flow.

The city-level model takes into account the traffic flow distribution of all roads, and thus lacks personalized consideration of the roads. For example, a set of traffic flow data that is considered normal on one road may be abnormal on another. Considering that the traffic flow distribution of different road segments could be quite different, we calculated the AUC, sensitivity, and specificity of the dynamic ABOD model in road segments. The results of our experiments are shown in Table 5.

Table 5 lists the crash risk classification results of our roadway segment models. It shows that for some road segments, for example, road 22104, 23209 and 24106, the sensitivity value reaches 0.850, 0.891 and 0.959 respectively with a lower FAR. Nevertheless, the result of some roads, such

Table 6. The comparation between some city-level models and road-level models.

| | Sensitivity (city-level model) | FAR (city-level model) | Sensitivity (road-level model) | FAR (road-level model) |
|---|---|---|---|---|
| Average | 0.867 | 0.173 | 0.925 | 0.058 |
| Std | 0.0826 | 0.0521 | 0.0737 | 0.0074 |

as 22218 and 23105, is even worse than the average performance. It can be seen that the model performance of different road segments varies considerably, so it is necessary to consider the road section heterogeneity and construct a roadway segment model.

Thanks to the advantages of unsupervised learning, we can build models on roads in the absence of a large number of samples, even on roads that do not contain any historical crash risk sample. So to predict whether a crash event will occur on a certain road segment at any given moment, the dynamic model should be rebuilt each time only using the traffic flow data on the road segment. After applying the dynamic models, both sensitivity and FAR are greatly improved, as shown in Table 6. Overall, the average sensitivity improves by 6.7% and the average FAR decreases by 66.5%. There is also a significant decrease in the variance of sensitivity and specificity metrics after road-level modeling on the road segments, which indicates the improved stability of the model. It proves that the performance of the road-level model considering road heterogeneity is better than the city-level model.

### 5.3. Model performance comparisons

Compared to the literature using other modeling approaches, the proposed dynamic ABOD model shows state-of-the-art performance (summarized in Table 7). The proposed dynamic ABOD model reaches a high sensitivity at 92.5% but a low FAR at 5.8%.

## 6. Summary

The benefits of real-time crash risk analysis are to provide insight into crash precursors and to implement proactive traffic safety management strategies. Significant efforts have been made in terms of various operational sensing data and advanced modeling techniques to obtain better crash risk prediction performance. However, most of the existing studies build models based on supervised learning methods, while supervised learning methods are highly dependent on positive crash samples and extremely sensitive to imbalanced data. Generally, supervised learning methods cannot be personalized to model the different traffic flow characteristics of specific road segments due to the lack of enough crash risk samples for each

**Table 7.** Crash risk classification results based on test data in other literature.

| Authors | Modeling Algorithm | Sensitivity | FAR | Ratio of Crash and Non-crash |
|---|---|---|---|---|
| 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 |
| 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 |
| Hossain and Muromachi (2011) | Classification and regression trees | 63.3% | 20.0% | 1:92 |
| Gao *et al.* (2018) | Cost sensitive MLP | 75.79% | 15.69% | full-sample |
| Cai et al. (2019) | DCGAN and CNN model | 88.8% | 9.3% | full-sample |
| This study | **Dynamic ABOD** | **92.5%** | **5.8%** | **real-time data** |

road section. Therefore, a modeling method with low requirements for historical crash risk samples, or even does not have historical crash risk samples, is more competitive. On the one hand, it can fundamentally solve the problem of imbalanced data, and on the other hand, it can more fully consider the spatial-temporal characteristics for personalized modeling.

The developed modeling scheme has the characteristics of being independent of historical crash risk samples and insensitive to imbalanced data, thus enabling the construction of a dynamic ABOD model that fully takes spatial and temporal information into account. A crash analysis dataset with different percentages of crash and non-crash samples was created and modeled using ABOD. The sensitivity of the city-level ABOD model under balanced 1:1 ratio is 79.1% with the FAR of 28.3%. Compared to the modeling results of the city-level ABOD, there is a conclusion that the accuracy of crash risk identification has been improved due to the addition of temporal information. And finally, the road-level dynamic ABOD model takes both temporal and spatial information fully into account, thus it has the best results with high sensitivity at 92.5% and low FAR at 5.8%, which is tested in a simulated real-time environment.

Besides, the proposed model was developed and implemented on the basis of NumPy framework. The experiments were conducted using Python 3.7 under Linux 18.04 operation system with 32 GB RAM, and CPU. In the real-time data environment testing, the CPU-based training and predicting took around 135 ms for dynamic ABOD model with the window size of 30 minutes. Therefore, dynamic ABOD model is well suited for real-time analysis due to its high efficiency.

Furthermore, in the field of traffic safety analysis, although some studies have attempted to apply unsupervised learning methods such as clustering for crash risk analysis, many of them have stayed in classifying and analyzing the entropy of crashes without unsupervised learning-based methods for crash identification. This study pioneered the application of anomaly detection methods in unsupervised learning to the field of crash risk identification to deal with the problem of data imbalance and refined modeling at the same time. The experiments prove that the modeling cost of unsupervised learning is extremely low, firstly, it does not require any historical crash data, and secondly, in our experiments, only a small number of historical non-crash samples of around 500 are needed to build the model. Also, the imbalance of the samples and their ratios have little effect on unsupervised learning, and the proposed ABOD-based model achieves approximately the same performance for data sets with different ratios of 1:1, 1:5, 1:10, 1:20, and 1:100. Furthermore, Road-level modeling that takes road heterogeneity into account provides insight into the variability of empirical associations between road characteristics and associated safety, and has proven to be better than city-level modeling in our experiments.

Besides ABOD, other unsupervised learning methods can be explored for crash risk analysis, and ABOD can also be optimized for practical problems instead of being used directly. Since most of the existing models are one-way, methods with feedback such as reinforcement learning can be explored to meet the requirements of active traffic safety management applications. Moreover, road characteristics and weather conditions data are also important factors influencing crash risk (Geedipally et al., 2019; Llopis-Castelló et al., 2021), and the introduction of such factors into models will also be a focus of future work. In addition, the application of the proposed modeling methods and the portability of the models are issues that need to be investigated in the future.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This study was sponsored by the Science and Technology Project of Zhejiang Province, No. 2021C01011 and the National Natural Science Foundation of China (NSFC), No.71771174.

## References

Abdel-Aty, M., Uddin, N., & Pande, A. (2005). Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. *Transportation Research Record: Journal of the Transportation Research Board*, *1908*(1), 51–58. doi:10.1177/0361198105190800107

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., & Hsia, L. (2004). Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board*, *1897*(1), 88–95. doi:10.3141/1897-12

Ahmed, M. M., & Abdel-Aty, M. A. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on Intelligent Transportation Systems*, *13*(2), 459–468. doi:10.1109/TITS.2011.2171052

Ahmed, M. M., Abdel-Aty, M., & Yu, R. (2012). Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. *Transportation Research Record: Journal of the Transportation Research Board*, *2280*(1), 60–67. doi:10.3141/2280-07

Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. *Transportation research part C: emerging technologies*, *86*, 202–219.

Behara, K. N., Paz, A., Arndt, O., & Baker, D. (2021). A random parameters with heterogeneity in means and Lindley approach to analyze crash data with excessive zeros: A case study of head-on heavy vehicle crashes in Queensland. *Accident Analysis & Prevention*, *160*, 106308. doi:10.1016/j.aap.2021.106308

Boquet, G., Morell, A., Serrano, J., & Vicario, J. L. (2020). A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transportation Research Part C: Emerging Technologies*, *115*, 102622. doi:10.1016/j.trc.2020.102622

Bottou, L. (2010). *Large-scale machine learning with stochastic gradient descent, Proceedings of COMPSTAT'2010* (pp. 177–186). Springer.

Buddhavarapu, P., Scott, J. G., & Prozzi, J. A. (2016). Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. *Transportation Research Part B Methodological*, *91*, 492–510. doi:10.1016/j.trb.2016.06.005

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. *Transportation Research Part A: Policy and Practice*, *127*, 71–85. doi:10.1016/j.tra.2019.07.010

Fei, T. L., Kai, M. T., & Zhou, Z. H. (2008). Isolation forest. IEEE International Conference on Data Mining IEEE.

Gao, Y., Gao, Z., Yu, R., Huang, Z., & Feng, J. (2018). Utilizing multilayer perceptron neural network for crash risk prediction based on a full set of data. In *CICTP 2018: Intelligence, Connectivity, and Mobility* (pp. 1947–1956). Reston, VA: American Society of Civil Engineers.

Geedipally, S. R., Pratt, M. P., & Lord, D. (2019). Effects of geometry and pavement friction on horizontal curve crash frequency. *Journal of Transportation Safety & Security*, *11*(2), 167–188. doi:10.1080/19439962.2017.1365317

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

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., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident; Analysis and Prevention*, *124*, 66–84. doi:10.1016/j.aap.2018.12.022

Hossain, M., & Muromachi, Y. (2011). Understanding crash mechanisms and selecting interventions to mitigate real-time hazards on urban expressways. *Transportation Research Record: Journal of the Transportation Research Board*, *2213*(1), 53–62. doi:10.3141/2213-08

Karimpour, A., Kluger, R., & Wu, Y.-J. (2021). Traffic sensor data-based assessment of speed feedback signs. *Journal of Transportation Safety & Security*, *13*(12), 1302–1325. doi:10.1080/19439962.2020.1731038

Katrakazas, C., Quddus, M., Chen, W.-H., & Deka, L. (2015). Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C: Emerging Technologies*, *60*, 416–442. doi:10.1016/j.trc.2015.09.011

Khoda Bakhshi, A., & Ahmed, M. M. (2022). Real-time crash prediction for a long low-traffic volume corridor using corrected-impurity importance and semi-parametric generalized additive model. *Journal of Transportation Safety & Security*, *14*(7), 1165–1200. doi:10.1080/19439962.2021.1898069

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. doi:10.1007/s13748-016-0094-0

Kriegel, H.-P., Schubert, M., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (444–452). doi:10.1145/1401890.1401946

Lange, S., & Riedmiller, M. (2010). Deep auto-encoder neural networks in reinforcement learning. International Joint Conference on Neural Networks IEEE.

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident; Analysis and Prevention*, *135*, 105371. doi:10.1016/j.aap.2019.105371

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.

Llopis-Castelló, D., Findley, D. J., & Garcia, A. (2021). Comparison of the highway safety manual predictive method with safety performance functions based on geometric design consistency. *Journal of Transportation Safety & Security*, *13*(12), 1365–1386. doi:10.1080/19439962.2020.1738612

Ma, X., Lu, J., Liu, X., & Qu, W. (2022). A genetic programming approach for real-time crash prediction to solve trade-off between interpretability and accuracy. *Journal of Transportation Safety & Security*, 1–23. doi:10.1080/19439962.2022.2076756

Oh, C., Oh, J.-S., Ritchie, S., & Chang, M. (2001). Real-time estimation of freeway accident likelihood. In 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Roshandel, S., Zheng, Z., & Washington, S. (2015). Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident; Analysis and Prevention*, *79*, 198–211. doi:10.1016/j.aap.2015.03.013

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471. doi:10.1162/089976601750264965

Shi, Q., & Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, *58*, 380–394. doi:10.1016/j.trc.2015.02.022

Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, *54*, 176–186. doi:10.1016/j.trc.2015.03.006

Wang, L., Abdel-Aty, M., Shi, Q., & Park, J. (2015). Real-time crash prediction for expressway weaving segments. *Transportation Research Part C: Emerging Technologies*, *61*, 1–10. doi:10.1016/j.trc.2015.10.008

Wang, X., Zhou, Q., Quddus, M., Fan, T., & Fang, S. (2018). Speed, speed variation and crash relationships for urban arterials[J]. *Accident; Analysis and Prevention*, *113*(APR), 236–243. doi:10.1016/j.aap.2018.01.032

World Health Organization. (2018). *Global status report on road safety 2018*. World Health Organization.

Wz, A., & Ww, B. (2019). Learning V2V interactive driving patterns at signalized intersections. *Transportation Research Part C: Emerging Technologies*, *108*, 151–166.

Xu, C., Liu, P., Yang, B., & Wang, W. (2016). Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. *Transportation Research Part C: Emerging Technologies*, *71*, 406–418. doi:10.1016/j.trc.2016.08.015

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. *Transportation Research Part C: Emerging Technologies*, *38*, 167–176. doi:10.1016/j.trc.2013.11.020

Yang, K., Wang, X., & Yu, R. (2018). A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. *Transportation Research Part C: Emerging Technologies*, *96*, 192–207. doi:10.1016/j.trc.2018.09.020

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259.

Yuan, J., Abdel-Aty, M., Gong, Y., & Cai, Q. (2019). Real-time crash risk prediction using long short-term memory recurrent neural network. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(4), 314–326. doi:10.1177/0361198119840611

Zhu, B., Jiang, Y., Zhao, J., He, R., Bian, N., & Deng, W. (2019). Typical-driving-style-oriented Personalized Adaptive Cruise Control design based on human driving data. *Transportation Research Part C: Emerging Technologies*, *100*, 274–288. doi:10.1016/j.trc.2019.01.025