Journal of Zhejiang University-SCIENCE A in press
www.jzus.zju.edu.cn; www.springer.com/journal/11582
E-mail: jzus_a@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.A2500331

Lateral risk prediction and influencing factors analysis of container trucks based on trajectory reconstruction data

Zhihao ZHU, Hexuan LIU, Rongjun CHENG✉

Faculty of Maritime and Transportation, Ningbo University, Ningbo 315211, China

Abstract: With the continuous growth of demand for container transportation, the proportion of container trucks passing through ports and surrounding roads has increased significantly. Due to their large size and poor maneuverability, once a truck accident occurs, it is often accompanied by serious casualties and property losses. Current research on traffic conflicts for container trucks is limited by the lack of high-quality data: first, publicly available container truck trajectory datasets are extremely rare; second, although drones have the ability to collect data over a large range, their shooting data have problems such as limited accuracy and discontinuous trajectories, which make it difficult to meet the high requirements of micro-modeling for data quality. This problem directly restricts the accuracy of conflict prediction and the credibility of causal analysis. To improve the accuracy and completeness of trajectory data, we introduce a trajectory reconstruction method to repair and complete the original trajectory. The experimental results show that the reconstructed trajectory is significantly better than the original data in terms of continuity and rationality. On this basis, a two-dimensional time to collision (2D-TTC) indicator was constructed to identify side-swipe conflict events, and based on the extraction of micro-behavior features, sample sets of side-swipe and rear-end conflicts were constructed and a variety of machine learning models introduced to carry out conflict prediction analysis. The results show that the GBDT model performs best in side-swipe conflict prediction, and the XGBoost model in rear-end conflict prediction. By introducing the SHAP method to improve the interpretability of the model, our analysis shows that the key factors influencing side-swipe conflict are the lateral speed and the average longitudinal speed within 5 s. The lateral speed reflects the lateral deviation of the vehicle, and the average longitudinal speed within 5 s reflects the driving stability and acceleration trend in a short time. The two together determine the lateral controllability of the vehicle in the dynamic process. Rear-end conflict is affected mainly by the change in longitudinal acceleration, revealing the instability of the vehicle during braking and the lack of control over the distance between the vehicle and the preceding vehicle. Finally, the model performance was optimized through feature ablation experiments where, in the prediction of side-swipe conflicts, GBDT achieved an accuracy of 0.911 and an AUC of 0.953.

Key words: Real-time conflict prediction; Container Truck Dataset; Trajectory reconstruction; Explainable Machine Learning; Side-swipe conflict.

Uncorrected

1 Introduction

In the area near a port, container truck traffic flow is characterized by high density, low speed and highly overlapping paths. Not only is the traffic organization complex, but there is also a high risk of traffic conflicts. Ordinary urban roads are dominated by intelligent connected vehicles (Li and Cheng, 2025), while the traffic subjects in a port scene are mainly heavy-loaded freight vehicles. Their poor driving stability, delayed acceleration and deceleration response, and weak control flexibility lead to traffic behavior with significant nonlinearity and dynamic uncertainty (Ji et al., 2023). In this context, traditional data collection methods that rely on fixed sensors can no longer meet the needs of refined traffic modeling and safety assessment. In recent years, drones have become an important supplementary tool for obtaining road traffic data due to their high maneuverability and bird's-eye view. Many studies have used drones to construct high-precision trajectory datasets (Krajewski et al., 2018; Barmpounakis and Geroliminis, 2020; Zheng et

✉ Rongjun CHENG, chengrongjun76@126.com
Rongjun CHENG, https://orcid.org/0000-0002-5558-9364

Received July 19, 2025; Revision accepted Oct. 9, 2025; Crosschecked

© Zhejiang University Press 2025

al., 2023b). This type of aerial observation method is particularly suitable for port areas. By shooting the container truck traffic flow, it can not only obtain the complete vehicle motion trajectory but also capture the microscopic interactions that are difficult to monitor with traditional ground equipment, laying a data foundation for subsequent more comprehensive safety analysis.

High-precision vehicle trajectory data have been widely used in driving behavior analysis and traffic safety assessment. Many studies have used drones combined with related technologies (OpenCV, KLT, DCF-CSR) to extract high-resolution vehicle trajectory datasets (Wang et al., 2019; Xing et al., 2020; Shawky et al., 2023). However, the trajectory data extracted using drones combined with image recognition technology still have data accuracy problems, including vehicle type recognition errors and noise caused by some numerical differentials. These data errors not only affect the accuracy of subsequent traffic feature extraction but may also destroy the true characterization of vehicle behavior, thereby weakening the effectiveness of trajectory-based traffic safety analysis. Traditional trajectory smoothing methods mainly remove noise from trajectory data by using a moving average (Zhou et al., 2017; Gu et al., 2019; Tian et al., 2019), Savitzky–Golay filtering (Ahn et al., 2013; Zaki et al., 2014), wavelet transform (Rafati Fard et al., 2017; Hu et al., 2022a) and other technologies to make the vehicle trajectory smoother and more continuous. However, such methods usually consider only the statistical characteristics of the data, ignoring the physical constraints and dynamic laws of vehicle motion, which can easily lead to excessive smoothing of the trajectory and loss of real dynamic characteristics, or insufficient smoothing and residual noise (Thiemann et al., 2008). In addition, they are sensitive to parameter selection, resulting in limited applicability and versatility in different scenarios (Montanino and Punzo, 2015; Wu et al., 2019). In contrast, trajectory reconstruction methods are not only dedicated to smoothing and denoising trajectory data, but also can repair missing segments in trajectories, remove outliers, and ensure the continuity of dynamic variables such as speed and acceleration in the time dimension and the consistency of traffic behavior logic (Zhao et al., 2024a). In addition, the reconstruction process usually combines outlier detection and trajectory interpolation technology, which is particularly suitable for data processing needs in high-noise and complex scenarios such as ports or intersections (Barmpounakis and Geroliminis, 2020; Zhao et al., 2024b). For example, some researchers proposed a three-stage method combining wavelet transform and Savitzky-Golay filtering, which effectively improved the smoothness and coherence of intersection trajectory data (Zhao, et al., 2024b); others further introduced a smoothing framework based on Gaussian mixture and unscented RTS filtering to enhance the robustness of trajectory reconstruction and its adaptability to non-Gaussian noise (He et al., 2024). Therefore, trajectory reconstruction methods have shown increasingly important research value and application potential in improving trajectory data quality, serving micro-behavior modeling and traffic safety assessment.

Due to their long bodies and large volumes, container trucks have higher requirements for surrounding space when changing lanes and merging, and their visual blind spots are significantly enlarged, which significantly increases the potential risks during lateral interactions (Jansen and Varotto, 2022; Li et al., 2024b). In addition, trucks often travel at lower speeds in specific scenarios such as port areas, which often prompt small vehicles behind to frequently change lanes and overtake to improve traffic efficiency, further exacerbating the lateral interference of traffic flow. In such complex traffic environments, lane changes are particularly frequent, and interactions between lanes are complex. Side-swipe conflicts have become a common but underestimated high-risk event (Chen et al., 2020; Ouyang et al., 2025). At present, mainstream traffic conflict research focuses on longitudinal conflict types, especially rear-end conflicts. Identification and warning methods generally adopt longitudinal safety indicators such as time to collision (TTC) (Xing et al., 2019; Orsini et al., 2021; Hu et al., 2022b). Such methods are usually based on the following assumptions: vehicles are in the same lane, the longitudinal approach is the main source of risk, and conflict judgment depends on changes in speed and distance. However, in real traffic scenarios, especially complex sections involving frequent lane changes or

Unedited

lane merging, this one-dimensional longitudinal assumption may not accurately reflect the way in which lateral risks are generated (Hou et al., 2024). Some improved methods, such as time to collision with disturbance (TTCD), have introduced longitudinal speed fluctuation factors (Xie et al., 2019). The anticipated collision time (ACT) indicator was further expanded into a two-dimensional TTC in a study to simultaneously characterize the lateral and longitudinal relative motion characteristics. This indicator was successfully used to identify multiple types of conflicts including side collisions (Venthuruthiyil and Chunchu, 2022). In terms of risk modeling, some recent studies have begun to attempt to identify potential conflicts in two-dimensional trajectory space. For example, one study proposed a two-dimensional collision detection method based on intersection trajectory analysis to identify high-risk events in lane changes and cross paths in advance (Ward et al., 2015). Another study developed an analytical algorithm for calculating lateral TTC and theoretically modeled lateral conflicts under idealized trajectory reconstruction conditions (Hou et al., 2014). However, these methods still face some serious problems in practical applications: first, most studies rely on high-precision sensors or simulation data, which make it difficult to meet the real-time application requirements of large-scale, natural scenes (Hu, et al., 2022b); second, in truck traffic flows involving large vehicles, high response lag, and weak control flexibility, and the robustness and adaptability of existing models still need to be improved (Li et al., 2022; Tian et al., 2024).

Traffic conflict prediction generally adopts data-driven methods, and the mainstream methods mostly adopt a binary classification framework, that is, to identify whether a conflict occurs. Among them, machine learning has been widely used in conflict detection and risk assessment tasks due to its excellent prediction performance (Katrakazas et al., 2018; Li et al., 2020; Mohammadian et al., 2021; Yuan et al., 2022). Compared with traditional statistical methods, machine learning and deep learning have significant advantages, especially in dealing with complex nonlinear relationships and high-dimensional features (Yao et al., 2021; Islam and Abdel-Aty, 2023). In recent years, researchers have also proposed a variety of methods to further improve model performance and adaptability. Some scholars developed a non-parametric model based on machine learning for conflict prediction at signalized intersections, and improved the accuracy and transferability of the model through Bayesian optimization (Zheng et al., 2023a); others have proposed a unified probabilistic modeling method that regards traffic conflicts as extreme events in interactions, thereby achieving effective detection of multiple environments and types of conflicts (Jiao et al., 2025). However, despite the continuous improvement in model performance, most of these methods are still "black box models" and it is difficult to explain how the features specifically affect the prediction results. This limitation restricts their application in traffic safety analysis and policy making. To enhance the interpretability of the model, local interpretation methods (such as LIME and SHAP) have been gradually introduced in recent years. LIME approximates the behavior of complex models near a single point by constructing a local linear model, and the calculation is relatively simple (Ribeiro et al., 2016), while SHAP is based on the Shapley value principle and measures the contribution of each feature to the model output from a game theory perspective, which is suitable for local and global interpretation (Lundberg, 2017). Studies have shown that SHAP can provide stable and consistent interpretation results in traffic risk identification. Some studies have used CatBoost and SHAP frameworks to identify key factors affecting highway safety (Li et al., 2024a). Others have compared the differences in risk characteristics of different types of drivers (Peng et al., 2024). These studies show that interpretability methods are of great significance in improving model transparency and policy application value.

This paper focuses on the traffic flow around a port, which is composed mainly of container trucks, and conducts real-time traffic safety analysis around side-swipe and rear-end conflicts. The original trajectory data are reconstructed to improve data accuracy and be more in line with physical laws. On this basis, a two-dimensional TTC indicator is developed based on the coordinate format output by YOLOv8 to realize the identification of side-swipe conflicts. At the same time, rear-end conflict data are collected to construct a sample set containing two

Unedited

types of conflict events. Subsequently, a variety of machine learning models are constructed to predict and model side-swipe conflicts and rear-end conflicts, and the performance of the models under different feature combinations is compared. Then, combined with the SHAP interpretability method, the role mechanism of each influencing factor in different conflict types is deeply analyzed. The best performing model based on preliminary analysis is further improved through feature superposition experiments. The main contributions of this paper include:

(1) The trajectory reconstruction process was performed on a traffic flow trajectory dataset consisting mainly of container trucks, and a high-precision dataset that complies with physical laws was obtained.

(2) A TTC index based on two-dimensional space is proposed to identify and extract side-swipe conflict samples, which solves the problem of insufficient side-swipe conflict recognition of traditional TTC methods.

(3) A variety of machine learning models were used to model and predict the two types of typical conflicts. Through different feature combinations and feature ablation experiments, a model with better prediction performance was obtained, which will play an important role in future intelligent truck loading conflict prediction modules.

(4) The SHAP theory was used to conduct interpretable analysis of the model results, and the specific impact of various characteristics on the two types of conflicts was deeply explored.

This paper is organized as follows: Section 2 introduces the methods involved in this study; Section 3 presents the results of trajectory reconstruction and conflict prediction modeling and conducts a preliminary analysis; Section 4 provides a discussion, analyzing in detail the specific impact of features on these two conflict types and pointing out the shortcomings of this study; Section 5 summarizes the entire paper.

2 Methodology

First, the original trajectory data are reconstructed to ensure the quality and rationality of the trajectory data. Then, 2D-TTC and TTC are used to identify the side-swipe and rear-end conflicts in the reconstructed data. On this basis, feature extraction is performed to complete the construction of the conflict sample, thus completing the sample production process. The samples are divided into training sets and test sets. The training set is used to train the machine learning model. The model with the best prediction performance is obtained through evaluation indicators to predict whether a conflict will occur. Then, the SHAP theory is used to analyze how each factor specifically affects the occurrence of conflict. Finally, according to the order of feature importance, the features are superimposed in sequence to perform feature ablation experiments to further improve the prediction performance of the model. The overall method framework is shown in Fig. 1.

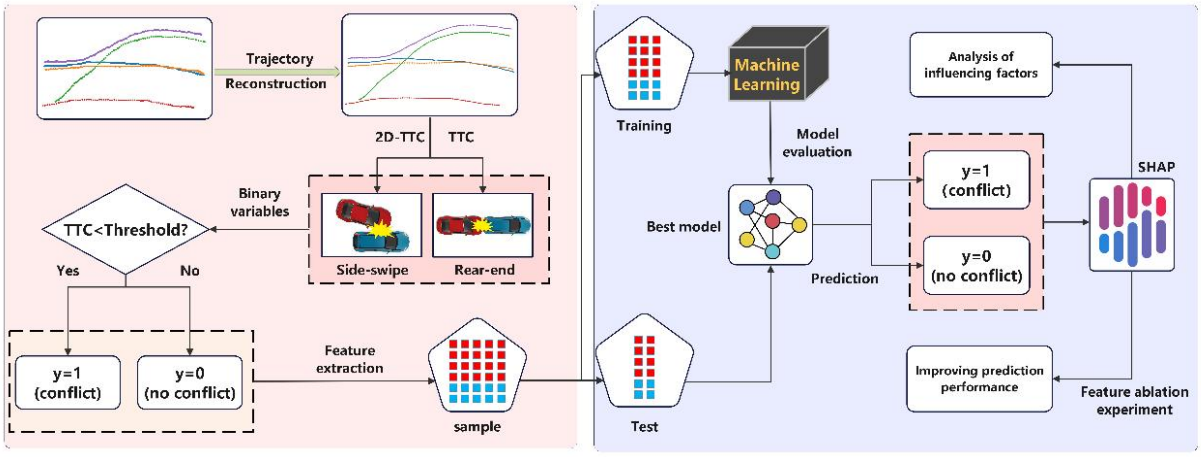


Fig. 1. Method framework diagram

Unedited

2.1 Trajectory Reconstruction

The video dataset was shot in windless and sunny weather, and abnormal videos were removed. All videos were anti-shake processed, but may still have had data quality issues. Therefore, it was necessary to reconstruct the trajectory data extracted by YOLOv8. The method of trajectory reconstruction is based on a previous study (Zhao, et al., 2024a), but due to the inclusion of different scenarios, the method was changed.

2.1.1 Objective Function

The Euclidean distance is used to measure the deviation between the reconstructed trajectory and the original trajectory (Fig. 2). The trajectory reconstruction problem is transformed into a nonlinear optimization problem, and the optimization goal is to minimize the RMSE value of the reconstructed trajectory and the original trajectory. The objective function can be expressed by the following equation:

$$\min C = \min \sqrt{\frac{1}{N}\sum_{i=1}^{N} D(L_i)} \quad (1)$$

where C is the cost function of reconstructing the trajectory, L_i is the coordinate information of each sampling point, i is the index of the sampling point, N is the number of sampling points in each trajectory, and $D(L_i)$ is the cost of reconstructing the i-th sampling point of a single trajectory, which can be calculated by Eq. 2.

$$D(L_i) = \left(x_i - x_{orignal,i}\right)^2 + \left(y_i - y_{orignal,i}\right)^2 \quad (2)$$

where x_i and y_i are the position coordinates of the i-th reconstructed trajectory point, $x_{orignal,i}$ and $y_{orignal,i}$ are the position coordinates of the i-th original trajectory point, and the unit of the coordinates is m.

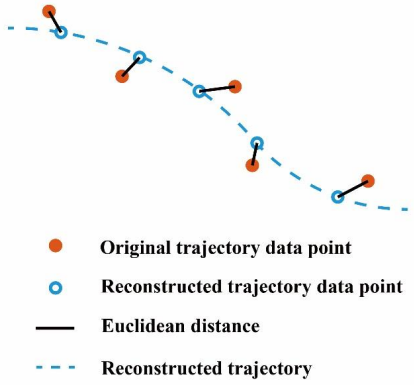


Fig. 2. Trajectory reconstruction principle.

2.1.2 Constraints

The reconstructed trajectory needs to conform to objective laws, so the speed, acceleration, and rate of change of acceleration should be limited within a reasonable range according to the traffic rules and vehicle characteristics of the study area. At the same time, different vehicles have different characteristics, and the range of restrictions should also be different. The given constraints are as follows:

(1) Speed constraint: vehicles in the study area travel in one direction, so the speed must be greater than or equal to 0, and the maximum speed normally does not exceed the speed limit of the road section.

$$0 \le v \le v_{\max} \quad (3)$$

where $v_{\max}$ is set to 23 m/s.

(2) Acceleration constraints. The vehicles in the study area can be divided into two main types: container trucks and cars. The characteristics of these two vehicles are different. Container trucks usually have slower acceleration than cars, so there are different constraints on these two types of vehicles.

$$a_{T\min} \le a_T \le a_{T\max}, a_{C\min} \le a_C \le a_{C\max} \quad (4)$$

where a_T represents the acceleration of the truck and a_C is the acceleration of the car. The acceleration range of the truck is limited to between -3 and $3\ m/s^2$, and the acceleration range of the car is limited to between -5 and $5\ m/s^2$, (Liu et al., 2022).

(3) Acceleration rate constraint (jerk constraint). Jerk is the first-order derivative of acceleration with respect to time, which measures the change in acceleration per unit time. In human-vehicle interaction applications, it is an important variable

Unedited

used to evaluate comfort.

$$-j_{\max} \le \frac{a(i+1)-a(i)}{\Delta t} \le j_{\max} \tag{5}$$

where $j_{\max}$ is the maximum value of the rate of change of acceleration, which is set to $10\ m/s^3$ (Martinez and Canudas-De-Wit, 2007), i represents the time index, and Δt represents the time interval, which is $0.1\ s$.

2.1.3 Solution

The trajectory reconstruction problem is transformed into a nonlinear optimization problem using the above formula. Python language is used for programming and the IPOPT algorithm (interior point optimizer) in the numerical optimization toolbox CasADi is used for solving (Wächter and Biegler, 2006; Andersson et al., 2019).

This experiment is carried out on multiple servers equipped with a 12 vCPU Intel(R) Xeon(R) Silver 4214R 2.40 GHz processor, 1 NVIDIA RTX 3080 Ti (12 GB) GPU, and 90 GB memory. The convergence condition is that the change in the objective function value is less than a certain threshold, which is set to 1×10^{-6}.

2.2 Conflict Identification

This study used characteristic thresholds to distinguish between conflict and non-conflict scenarios. Not all interactions were considered conflicts. A more detailed distinction was made between rear-end conflicts and side-swipe conflicts. A rear-end conflict was considered a conflict if its magnitude was below a certain threshold, while a side-swipe conflict was considered a conflict only if it satisfied Eq. 11 and was below a certain threshold.

2.2.1 Rear-end conflict

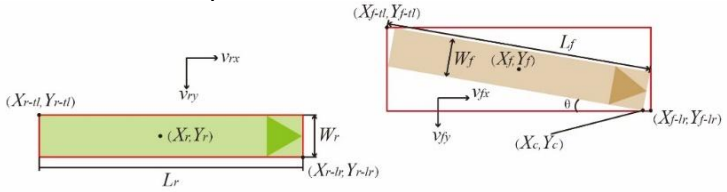
This study used TTC (Hayward, 1972) to identify rear-end conflicts. TTC can be calculated by the following equation:

$$TTC=\begin{cases}\dfrac{g_x}{v_r-v_f}, v_r \ne v_f \\ \infty \quad , v_r = v_f\end{cases} \tag{6}$$

where g_x is the distance between the front and rear vehicles in the x direction, v_r is the speed of the rear vehicle, and v_f is the speed of the front vehicle.

2.2.2 Side-swipe conflict

For side-swipe conflicts, a two-dimensional TTC was developed based on TTC and combined with the coordinate characteristics of the YOLOv8 detection box to identify side-swipe conflicts. The specific principle is shown in Fig. 3, First, some implicit conditions for side-slip collisions must be clarified. The longitudinal speed of the rear vehicle must be greater than that of the front vehicle, otherwise there will be no collision. Then, the interaction between the front and rear vehicles in the two-dimensional direction is considered, and finally two extreme situations where collisions may occur are obtained (Fig. 3 (b) and (c)).



(a) Precursor of side-swipe conflict

(b) Extreme case 1 of sidd-swipe conflict (c) Extreme case 2 of sidd-swipe conflict

Fig. 3. Side-swipe conflict identification principle.

Unedited

The coordinates (X_c , Y_c) of the collision point can be calculated by Eq. 7.The TTC in the longitudinal direction of the two extreme cases is different and can be calculated by Eqs 8-9.The lateral TTC can be calculated according to the Eq. 10. If the lateral TTC is between the longitudinal TTC of these two cases, it is considered that conflict is likely to occur, which can be expressed by formula 11.

$$\begin{cases} X_c = X_{f-lr} - W_f \sin\theta \\ Y_c = Y_{f-lr} \end{cases} \tag{7}$$

where (X_{f-lr}, Y_{f-lr}) is the coordinate of the lower right corner of the front vehicle detection box, W_f is the width of the front vehicle, and θ is the heading angle of the vehicle.

$$T_{x1} = \frac{X_c - X_{r-lr}}{v_{rx} - v_{fx}} \tag{8}$$

$$T_{x2} = \frac{X_c - X_{r-lr} + L_r}{v_{rx} - v_{fx}} \tag{9}$$

where X_{r-lr} is the X coordinate of the lower right corner of the rear vehicle detection box, L_r is the length of the rear vehicle, v_{rx} and v_{fx} are the longitudinal speeds of the rear vehicle and the front vehicle, respectively.

$$T_y = \frac{Y_{r-tl} - Y_e}{v_{fy} - v_{ry}} \tag{10}$$

$$T_{x1} \le T_y \le T_{x2} \tag{11}$$

where Y_{r-tl} is the Y coordinate of the top left corner of the rear vehicle detection box, v_{fy} and v_{ry} are the lateral speeds of the front and rear vehicles, respectively.

2.3 Conflict Analysis

2.3.1 Conflict prediction

Traffic conflict prediction is essentially a binary classification problem. Machine learning models are well suited for solving this problem. In this study, we trained multiple machine learning models for traffic conflict prediction. Logistic regression (Safavian and Landgrebe, 1991) (LR) maps input features to probability values through a logical function and is suitable for simple binary classification problems. Support vector machines (Vapnik, 1995) (SVM) maximize class intervals by finding the optimal hyperplane and are particularly suitable for high-dimensional data. XGBoost (Chen and Guestrin, 2016) is an efficient gradient boosting framework that gradually improves model performance by iteratively optimizing weak learners. Gradient boosted decision tree (Friedman, 2001) (GBDT) gradually improves the overall prediction performance by iteratively training multiple decision trees to fit the residuals and is suitable for dealing with nonlinear relationships and feature interaction problems.

2.3.2 Model Evaluation

We used five criteria to evaluate model performance. The first three criteria were accuracy (ACC), false negative rate (FNR) and false positive rate (FPR). These indicators were based on the confusion matrix and calculated using Eqs (S1)-(S3). The higher the ACC value, the better the model performance, and the lower the FNR and FPR values, the better the model performance. The other two indicators were the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC), which are also used to evaluate model performance (Bradley, 1997). ROC is used to show the performance of a classification model at different decision thresholds. It depicts the relationship between the true positive rate (TPR) and the FPR. TPR can be calculated using Eq. (S4). The AUC value ranges from 0 to 1: the closer to 1, the better the model performance.

Note: If the section, figure, table, or formula in the main text below has an S at the beginning of its number, it means that the content is in the appendix.

2.3.3 SHapley Additive explanation

SHAP has been used to explain machine learning models (Islam and Abdel-Aty, 2023). It provides a consistent and fair way to assign the contribution of features to predictions based on the Shapley value in game theory. The SHAP value measures the importance of each feature by calculating its marginal contribution in different combinations, thereby explaining the output of the model. A positive or negative SHAP value indicates that the feature has a positive or negative impact on the prediction. In this study, we used SHAP to analyze how features affect conflicts. The detailed principle of SHAP can be found in the electronic version of the attached

Unedited

material, Eqs (S5)-(S7).

3 Data and results

3.1 Data preparation

This dataset is a collection of high-definition images of vehicles on the road section of Ningbo Meishan Island taken by drones. The total video length is about 40 h, the resolution is 1920*1080, and the frame rate is 30 frames per s. The dataset contains vehicle images at different time periods, all taken in windless and clear or light wind conditions, with high resolution and real scene characteristics. Since Ningbo Meishan Island is an important port and logistics center, almost 80-90% of the vehicles on the road are container trucks, and the traffic flow properties are special.

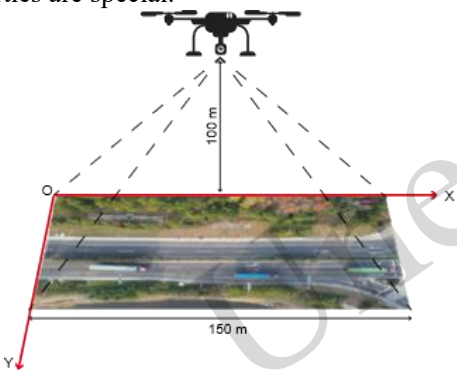
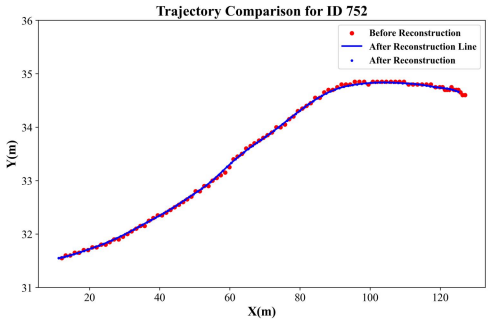


Fig. 4. Road section recorded by drone.

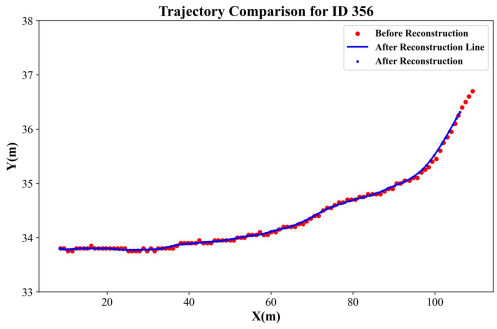
First, the video dataset was subjected to abnormal video removal and anti-shake processing, and then the vehicle trajectory was extracted using YOLOv8 (Fig. 4), The vehicle position was recorded using the global coordinate system. The origin of the coordinate was set at the upper left corner of the road section. The x coordinate increased as it moved to the right, and the y coordinate increased as it moved down the road. The initially extracted trajectory data included the frame number, ID, type of target vehicle, center coordinates of the target vehicle and their respective lengths and widths, coordinates of the upper left and lower right corners of the target vehicle detection frame, and other calculated data (Table S1 of the electronic supplementary materials). Details of the dataset construction process can be found published in a previous article (Zhu et al., 2025).

3.2 Trajectory reconstruction effect

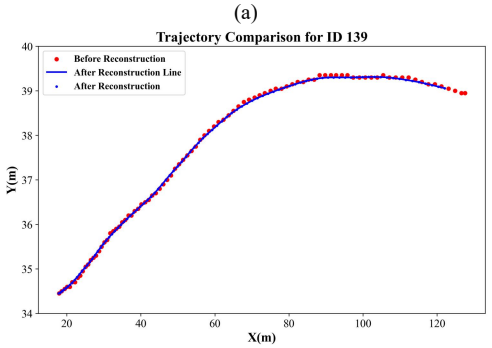
After eliminating abnormal trajectories, a total of 24038 trajectories were reconstructed. The reconstructed trajectories of some vehicles are shown in Fig. 5. Fig. 5(a, c) are the complete trajectories of a vehicle changing lanes to the right. The vehicle in Fig. 5a changes one lane and the vehicle in Fig. 5c changes two lanes. Fig. 5(b) shows the trajectory of a vehicle driving to the right into the right-turn lane.

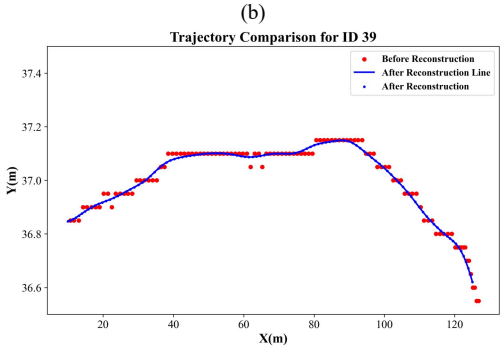


(a)

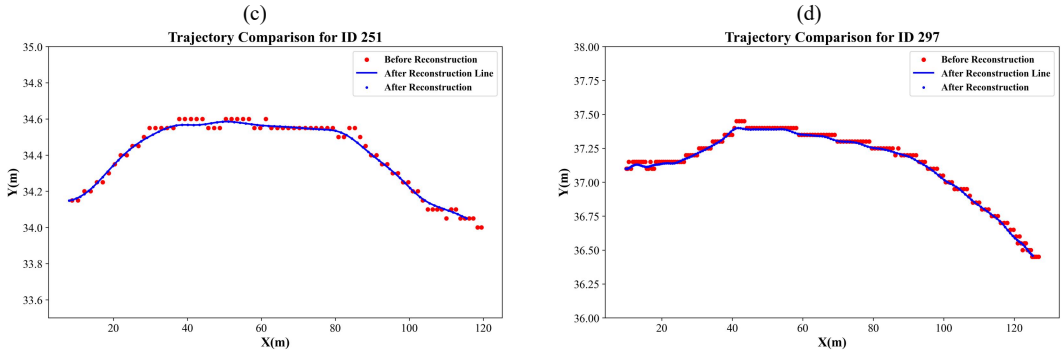


(b)





Unedited

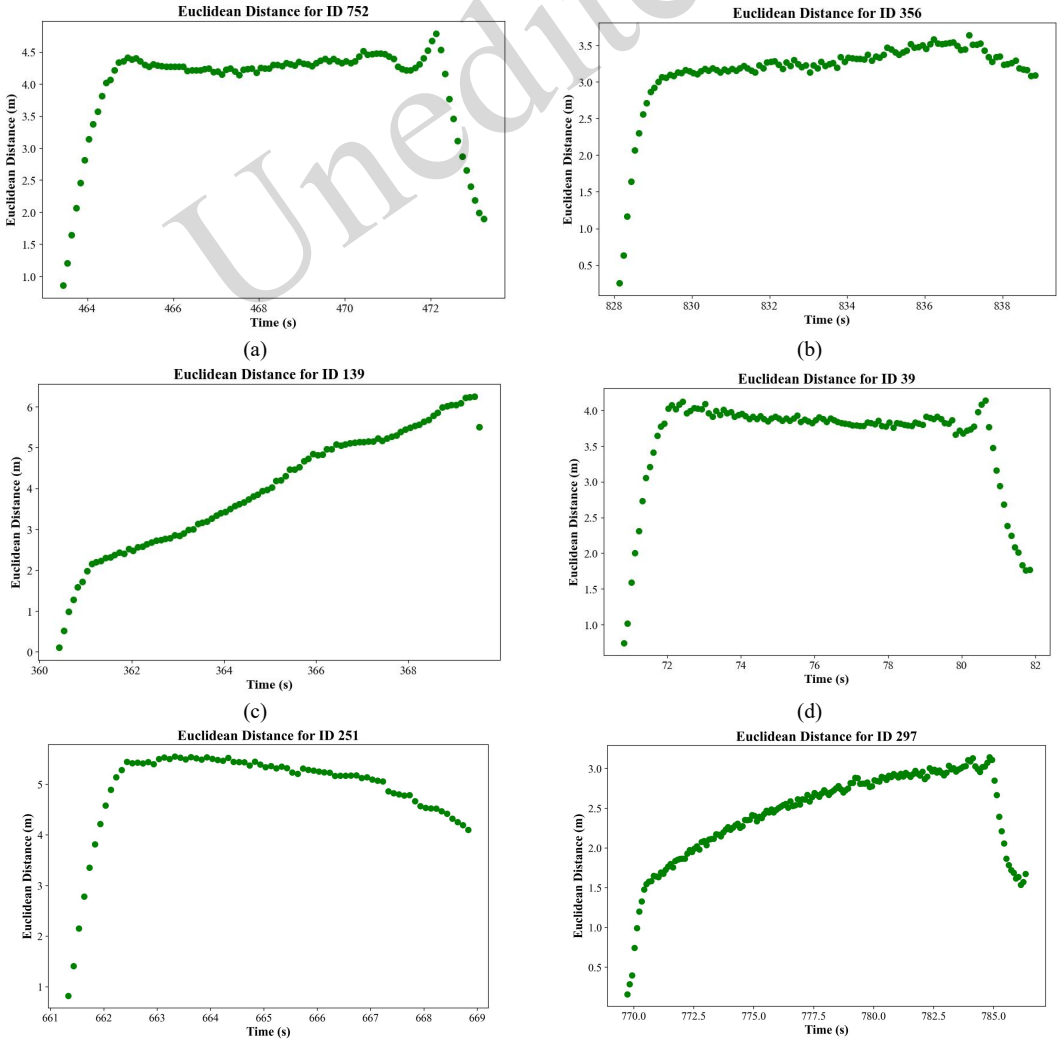


(e) (f)

Fig. 5. Comparison of trajectories before and after reconstruction.

Although the trajectories of Fig. 5(d, e, f) seem to fluctuate greatly, the ordinate indicates that they were in a straight-line state, and the overall fluctuation is consistent with the actual driving situation. Although the reconstructed trajectory is relatively close to the original trajectory, the zoom data show that the data points of the original trajectory are jagged and there is some obvious noise, while the reconstructed trajectory is a relatively smooth path.

Unedited



(e) (f)

Fig. 6. Euclidean distance of reconstructed trajectory data points.

Fig. 6 shows the Euclidean distances between the reconstructed trajectory data points and the original data points. The Euclidean distances for most data points were relatively large, indicating that the reconstructed trajectory differs significantly from the original trajectory. However, Fig. 5 shows that the differences between the data points in the Y direction were not very large, and theoretically, a large Euclidean distance should not result. Looking at the X direction, the reconstruction reveals that the data point at the end of the original trajectory has shifted to the left, which is likely a reasonable phenomenon. Drones are affected by wind when collecting video data. Although shooting in clear, calm weather is recommended, they can still be affected. Drones are easily affected in the X direction, resulting in some unreasonable data points in this direction, such as excessive longitudinal displacement between data points. This explains the large Euclidean distances between the data points in Fig. 6.

Fig. S1-S5 shows the comparison of speed and acceleration in two-dimensional space before and after reconstruction, as well as the comparison of Jerk. The data of the original trajectory are largely inconsistent with the actual situation, while the data of the reconstructed trajectory have been greatly improved and are more consistent with the actual situation.

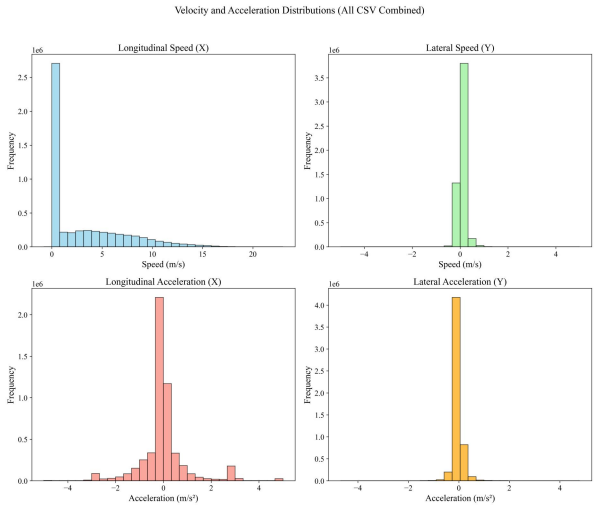


Fig. 7. Histograms of velocity and acceleration in the longitudinal and lateral directions.

To ensure the rationality of the reconstruction results, we merged all trajectory dataset files and plotted longitudinal and lateral velocity and acceleration histograms to verify the accuracy of the trajectory reconstruction (Fig. 7). The longitudinal velocity was concentrated between 0 and 5 m/s, peaking at 0, reflecting frequent vehicle stops and starts. The longitudinal acceleration was centered around 0, with slightly more negative values, consistent with deceleration, yielding, and starting. Both the lateral velocity and acceleration were concentrated near 0, indicating no significant lane changes or turns, and minimal lateral movement, consistent with the vehicle approaching an intersection but not entering it. The overall distribution was reasonable and consistent with the dataset's collection scenario (i.e., a road section near an intersection).

3.3 Analysis of factors affecting conflict

The conflict samples were extracted from the reconstructed trajectory data. A TTC value of 3 s was finally selected as the threshold for distinguishing conflicts, based on references (Shahana and Vedagiri, 2024). Finally, 1506 rear-end conflicts and 264 side-swipe conflicts were extracted (Table 1). The selection of relevant features retained the speed and acceleration features, as well as the calculated heading angle, and the mean and standard deviation of these features. The calculation range of the mean and standard deviation was selected to be within 5 s. The statistical description of the relevant variables is shown in Table S2 and Table S3.

The two types of conflict samples were trained separately: 80% of the sample data were used as the training set, 20% as the test set, and five-fold cross validation was used to make full use of the data. The selected classification models were SVM, LR, XGBoost, and GBDT. To select the model with the best performance, a cross grid was used to find the best parameters, and the model with the best preliminary performance was selected by combining the features.

Table1 Sample composition

Samples	Rear-end	Side-swipe
0	2259	573
1	1506	264

Unedited

Note: 0 means no conflict, 1 means conflict.

Table 2 Comparison of model performance with different feature combinations.

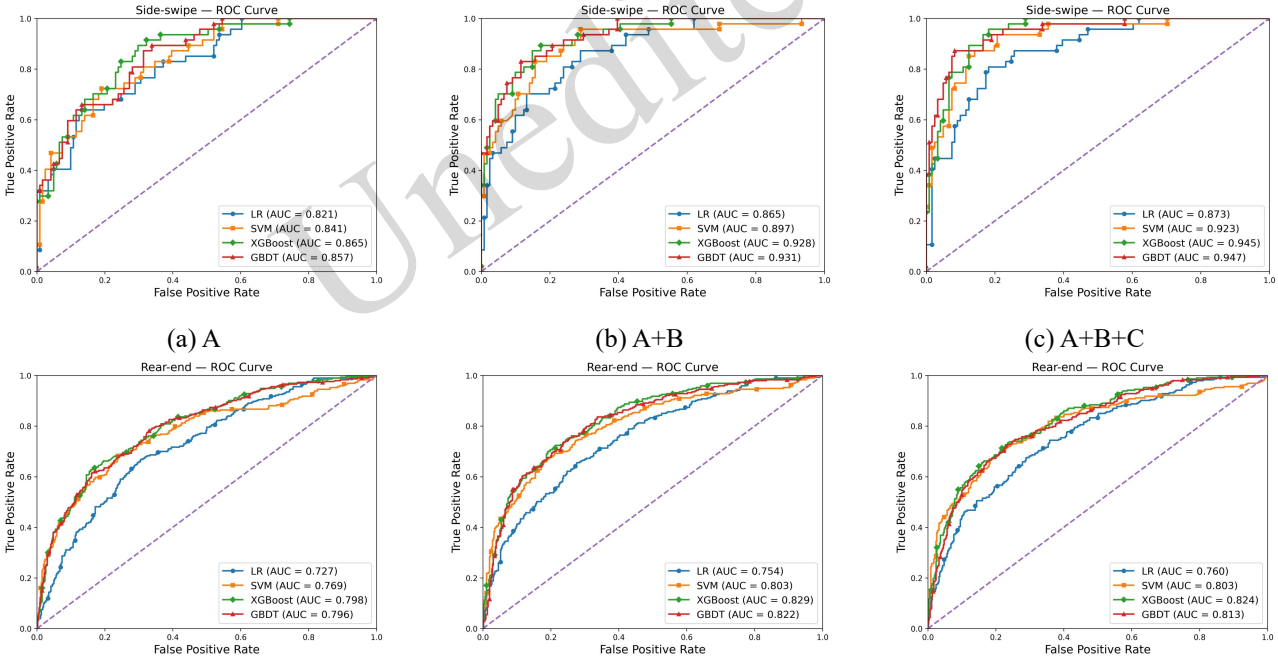
Feature	Model	ACC	FPR	FNR
A	LR	0.768	0.099	0.574
	SVM	0.804	0.091	0.468
	XGBoost	0.810	0.132	0.340
	GBDT	0.780	0.165	0.362
A+B	LR	0.810	**0.066**	0.511
	SVM	0.833	0.107	0.319
	XGBoost	0.857	0.099	0.255
	GBDT	0.863	0.083	0.277
A+B+C	LR	0.798	0.074	0.532
	SVM	0.845	0.074	0.362
	XGBoost	0.869	0.083	0.255
	GBDT	**0.905**	0.074	**0.149**

Note: The higher the ACC value and the lower the FNR and FPR values, the better the model performance.

3.3.1 Side-swipe conflict

The model evaluation results of side-swipe conflict are shown in Table 2. As the number of features increased, the performance of the models improved significantly, except for the LR model whose performance decreased, perhaps because LR is not flexible enough in dealing with nonlinear problems. When all features were used for training, the GBDT model had the best performance, with an accuracy of 0.905, an FPR of 0.074, and an FNR of 0.149. Fig. 8 (a) (b) (c) shows the ROC curves and AUC values of each model under different feature combinations. The AUC of the GBDT model when using all features reached 0.948. Overall, the model performance of GBDT was the best.

Unedited



(a) A (b) A+B (c) A+B+C

(d) A (e) A+B (f) A+B+C

Fig. 8. ROC curves under different feature combinations

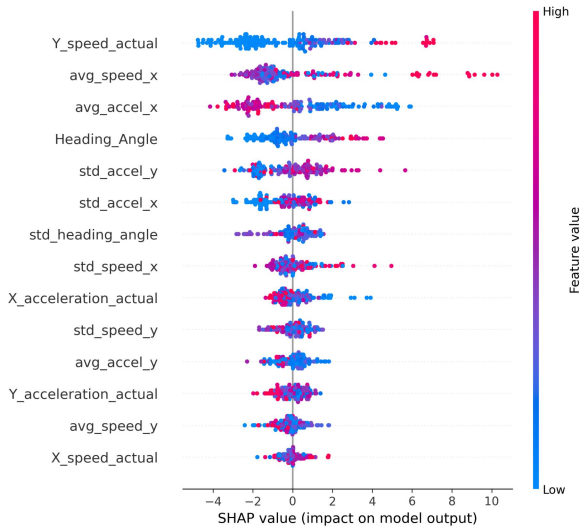


Fig. 9. Summary plot of model with SHAP values.

Then, the results of the GBDT model were analyzed for interpretability. Fig. 9 shows the SHAP value summary of the GBDT model. The features on the left are sorted by importance: those with higher speed are more important. The most important feature is the lateral speed. The higher the speed, the more likely it is to cause a side-swipe conflict. The second is the average longitudinal speed within 5 s. Faster speeds are more likely to cause conflicts. The average longitudinal acceleration within 5 s is the opposite: higher acceleration is less likely to cause conflicts. This may be because the driver judges that the road conditions are safer and will maintain a higher acceleration. The standard deviation of acceleration within 5 s measures the driver's driving style. A higher value indicates that the driver's operation is relatively unstable, which will also increase the possibility of a conflict. Overall, the impact of the first six features on conflicts is relatively important

3.3.2 Rear-end conflict

The model evaluation results of rear-end conflict are shown in Table 3. The performance of the model using ABC features and AB features did not change much, but the accuracy and FNR were better when using AB features. When using AB features, the best performing model was XGBoost, with an accuracy of 0.772, an FPR of 0.128, and an FNR of 0.386. Fig. 8 (d, e, f) shows the ROC curves and AUC values of each model under different feature combinations. The AUC of the XGBoost model reached 0.828 when using AB features. Overall, the model performance of XGBoost was relatively good.

Table 3 Comparison of model performance with different feature combinations.

Feature	Model	ACC	FPR	FNR
A	LR	0.681	0.165	0.560
	SVM	0.732	0.109	0.519
	XGBoost	0.741	**0.104**	0.502
	GBDT	0.742	0.107	0.495
A+B	LR	0.696	0.215	0.444
	SVM	0.756	0.122	0.437
	XGBoost	**0.772**	0.128	0.386
	GBDT	0.765	0.146	0.375
A+B+C	LR	0.697	0.224	0.427
	SVM	0.756	0.130	0.423
	XGBoost	0.764	0.126	0.410
	GBDT	0.750	0.178	**0.362**

Note: The evaluation indicators of the model results in Tables 2 and 3 correspond to the test part of the dataset.

Then, the interpretability of the XGBoost model results was analysed. Fig. 10 shows the SHAP value summary of the XGBoost model. Longitudinal acceleration was the most critical influencing factor in rear-end conflict, followed by factors such as the average longitudinal acceleration and longitudinal speed within 5 s. When the longitudinal acceleration was lower, conflicts were more likely to occur. This may be due to the large number of vehicles on the road, frequent vehicle interactions, and insufficient vehicle traffic, which easily leads to conflicts. When the longitudinal speed of the vehicle was low, its impact on the conflict was both positive and negative. This corresponds to two situations. When the traffic flow is smooth overall, conflicts are not likely to occur at a low speed. When the traffic flow is slow, even at a low speed, conflicts may occur easily due to increased vehicle interactions and shortened vehicle distances. This shows that it is necessary to consider multiple factors comprehensively, and confirms that when predicting conflicts, when the number of features is large, the model often predicts conflicts more accurately.

Unedited

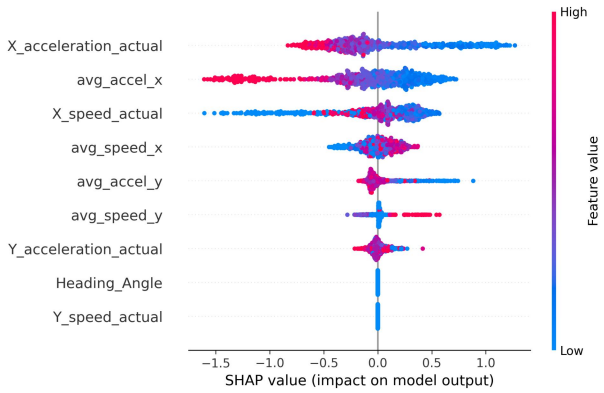
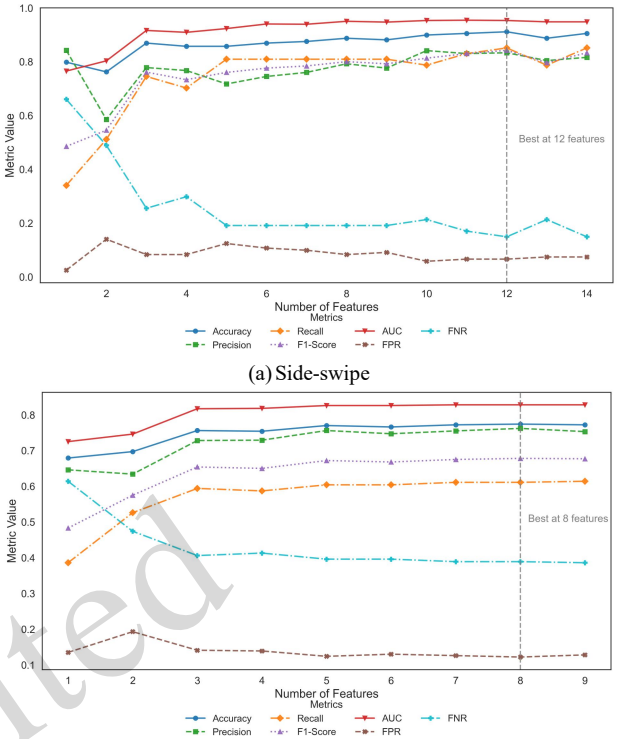


Fig. 10. Summary plot of model with SHAP values.

3.3.3 Conflict prediction model development

The next step was to improve the model. As shown in Table S4 in the Appendix, we have clarified the importance of each feature. Now, we will present a more detailed exploration of the performance of the prediction models for side-swipe conflicts and rear-end conflicts. Starting from the most important feature, we will add features in turn to compare the changes in the model's prediction performance. The model prediction performance uses more comprehensive indicators, namely accuracy, precision, recall, F1-Score, AUC value, FPR and FNR.

As the best model for predicting side-swipe conflict was GBDT when using 14 features, we added features to the GBDT model to explore the changes in model performance. The model performance became more stable when 3 features were added (Fig. 11 (a)). After the 12th feature was added, the overall performance of the model peaked, with Accuracy, Precision, Recall, F1-Score, AUC, FPR, and FNR being 0.911, 0.833, 0.851, 0.842, 0.953, 0.066, and 0.149, respectively. As shown in Fig. 11 (b), the above steps were performed on the model with the best performance in predicting rear-end conflicts. There was a common feature in the changes in their performance indicators. After adding the third feature, the model performance indicators changed relatively smoothly, which shows that the first three features are crucial for correctly predicting collisions. The best model performance appeared when the eighth feature was added, and the indicators were 0.774, 0.762, 0.611, 0.678, 0.828, 0.122, and 0.389, respectively. Table S5 in the Appendix lists the parameter settings of the best model.



(a) Side-swipe

(b) Rear-end

Fig. 11. Evaluation index change chart.

4 Discussion

This study was based on a vehicle trajectory dataset composed mainly of container trucks near a port area. The data were first reconstructed to improve the data quality. Then, prediction modeling was performed for two typical conflict types: rear-end conflict and side-swipe conflict. The SHAP theory was used to analyze the interpretability of the model results and explore how each factor affects the occurrence of conflicts. Finally, by evaluating the performance of multiple machine learning models with different numbers of features, the key variables that affect the occurrence of conflicts were determined, and a prediction model with better performance was constructed.

4.1 Analysis of factors affecting conflict

Figs 12 and 13 show the SHAP values of the top six important features for identifying the two types of collisions. For side-swipe conflicts, the most important feature is the lateral speed. This is reasonable because side-swipe conflicts usually occur in lane change scenarios. Fig. 12 (a) shows that

Unedited

side-swipe conflicts are likely to occur when the lateral speed is above 0.3 m/s. This is consistent with previous research findings (Shahana and Vedagiri, 2024). Fig. 12 (b) shows that vehicles are more likely to have conflicts when the average longitudinal speed within 5 s is low or high. Fig. 12 (c) shows that vehicles are more likely to have conflicts when they are decelerating. They are less likely to have conflicts when they are accelerating for a long time. This corresponds to the scenarios of sudden braking and smooth traffic flow, respectively. Fig. 12 (d) shows that higher heading angles are more likely to cause side-swipe conflicts. This finding complements previous studies that showed that higher heading angles are less likely to cause rear-end conflicts (Islam and Abdel-Aty, 2023).

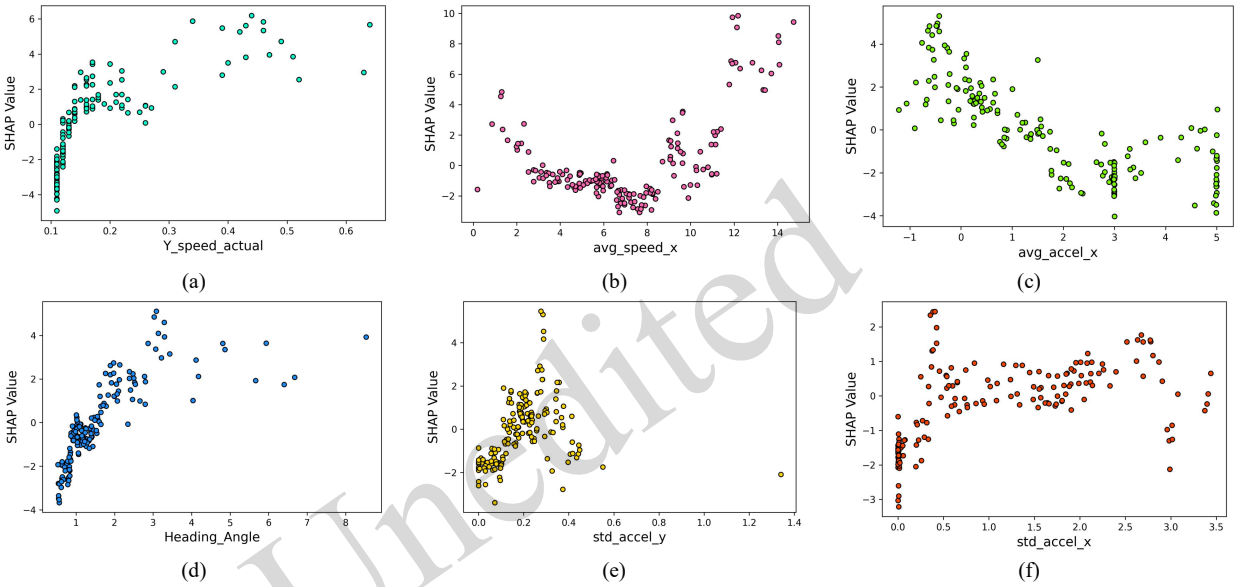
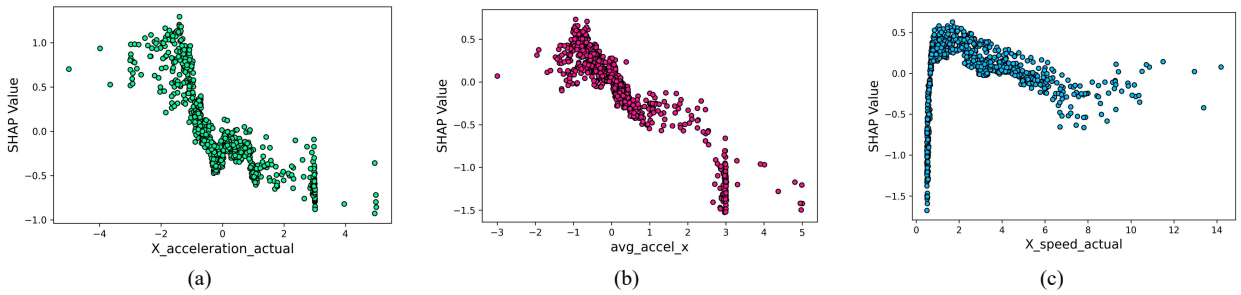
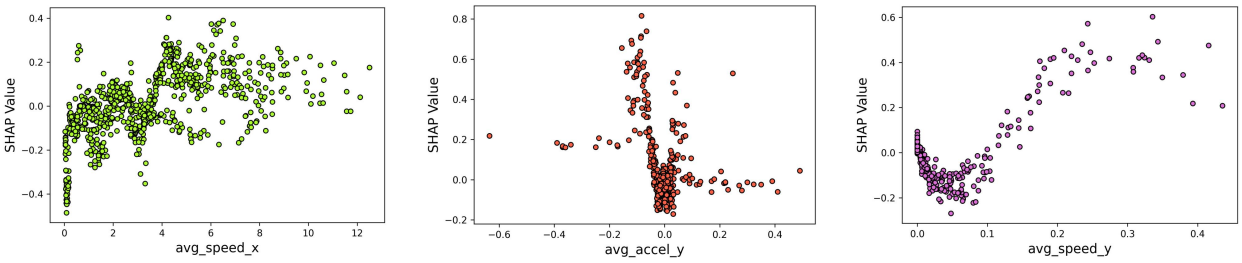


Fig. 12. SHAP values of the side-swipe feature.

For rear-end conflicts, the most important feature is the longitudinal acceleration. Fig. 13 (a) shows that when the driver decelerates, the driver's judgment is lower than his/her safety standard. When accelerating, the driver may feel safe. The impact of the longitudinal average acceleration within 5 s on rear-end conflicts is basically the same as for the longitudinal acceleration. Fig. 13 (c) shows that the longitudinal speed of vehicles in the range of 1-5 m/s is prone to conflict. Such a speed range shows that the overall traffic flow speed is not fast, indicating that there are many vehicles, the distance between vehicles may be small, and a certain speed is maintained, so it is more likely to cause conflicts.



Unedited



(d) (e) (f)

Fig. 13. SHAP value of the rear-end feature.

The feature correlation analysis part is presented in Section S4.2 of the Appendix.

4.2 Limitations

While we systematically analyzed influencing factors and constructed a rear-end conflict and side-swipe conflict prediction model based on reconstructed container truck trajectory data in a port area, several limitations remain. First, the data source was relatively limited, with strong regional and vehicle type restrictions, which may affect the model's generalization. Furthermore, we focused on container truck traffic flow, a scenario that is relatively underrepresented in current research. Second, the model input features are based mainly on kinematic metrics and fail to incorporate multiple sources of information, such as road structure and environmental factors, making it difficult to fully characterize the conflict formation mechanism. Third, the performance of the rear-end conflict prediction model remains suboptimal, possibly due to the complex formation mechanism of rear-end conflicts and incomplete consideration of influencing factors. Furthermore, the model training and validation in this study were based on offline data and have not yet been tested in real traffic environments or real-time systems. Therefore, the applicability and stability of our model requires further verification. Finally, while methods such as SHAP were introduced for feature interpretation, SHAP analysis results may be affected by inter-feature correlations and model structure, resulting in limitations in interpreting causal relationships. Subsequent research could include the following aspects: at the data level, trajectory data of different regions, vehicle types and time periods could be expanded to improve the generalization of the model; at the feature level, multi-source information such as road geometry, traffic flow density and meteorological conditions could be integrated to characterize the conflict mechanism more comprehensively; at the method level, multi-scale time series modeling and model integration could be explored and verified in real traffic or simulation environments; at the same time, explainable artificial intelligence and traffic safety mechanisms could be combined to improve the stability, credibility and generalizability of the model.

Unedited

5 Conclusions

This study reconstructed the trajectory data of vehicles, mainly container trucks, near a port area to obtain a high-resolution and high-precision vehicle trajectory dataset. Then, based on the coordinate format output by YOLOv8, a two-dimensional TTC was developed to identify side-swipe conflicts. Samples of two types of conflicts, side-swipe and rear-end conflicts, were collected for container truck traffic flow. A machine learning model and SHAP theory were used to analyze the influencing factors and a model for predicting these two conflicts was developed. The main conclusions of this study are as follows:

Compared with the original data, after trajectory reconstruction, the trajectory and its speed, acceleration and jerk become more reasonable, which is in line with the actual vehicle driving conditions. The video dataset shot by drones has the problem of data accuracy, and the trajectory reconstruction method solves this problem well.

Considering the interaction of vehicles in two-dimensional space, the developed 2D-TTC can well identify side-swipe conflicts. Higher lateral speed and average longitudinal speed within 5 s are more likely to cause side-swipe conflicts, and lower

average acceleration within 5 s is also more likely to cause conflicts. The GBDT model with 12 features was finally selected to achieve the best performance in identifying side-swipe conflicts, with an accuracy of 0.911 and AUC of 0.953.

The most important features affecting rear-end conflicts are longitudinal acceleration and average acceleration within 5 s. When these two values are low, rear-end conflicts are more likely to occur. Finally, the XGBoost model with 8 features was selected, which achieved the best performance in identifying rear-end conflicts, with an accuracy of 0.774 and AUC of 0.828.

Acknowledgments

This work is supported by the Program of Humanities and Social Science of the Education Ministry of China (No. 24YJA630013) and the Ningbo Natural Science Foundation of China (Grant No. 2024J125) and "Innovation Yongjiang 2035" Key R&D Programme (Grant No.2024H032).

Author contributions

Zhihao Zhu: Conceptualization, Methodology, Writing-original draft. Hexuan Liu: Investigation, Validation, Data curation, Formal analysis. Rongjun Cheng: Writing - review & editing, Supervision.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahn S, Vadlamani S, Laval J, 2013. A method to account for non-steady state conditions in measuring traffic hysteresis. *Transportation Research Part C: Emerging Technologies*, 34:138-147. https://doi.org/10.1016/j.trc.2011.05.020

Andersson JaE, Gillis J, Horn G, et al., 2019. Casadi: A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1):1-36. https://doi.org/10.1007/s12532-018-0139-4

Barmpounakis E, Geroliminis N, 2020. On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment. *Transportation Research Part C: Emerging Technologies*, 111:50-71. https://doi.org/10.1016/j.trc.2019.11.023

Bradley AP, 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Chen AY, Chiu Y-L, Hsieh M-H, et al., 2020. Conflict analytics through the vehicle safety space in mixed traffic flows using uav image sequences. *Transportation Research Part C: Emerging Technologies*, 119:102744. https://doi.org/10.1016/j.trc.2020.102744

Chen T, Guestrin C, 2016. Xgboost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p.785-794. https://doi.org/10.1145/2939672.2939785

Friedman JH, 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, :1189-1232.

Gu X, Abdel-Aty M, Xiang Q, et al., 2019. Utilizing uav video data for in-depth analysis of drivers' crash risk at interchange merging areas. *Accident Analysis & Prevention*, 123:159-169. https://doi.org/10.1016/j.aap.2018.11.010

Hayward JC, 1972. Near miss determination through use of a scale of danger.

He J, Peng B, Feng Z, et al., 2024. A gaussian mixture unscented rauch–tung–striebel smoothing framework for trajectory reconstruction. *IEEE Transactions on Industrial Informatics*, 20(5):7481-7491. https://doi.org/10.1109/TII.2024.3360478

Hou J, List GF, Guo X, 2014. New algorithms for computing the time - to - collision in freeway traffic simulation models. *Computational intelligence and neuroscience*, 2014(1):761047.

Hou K, Zou J, Zheng F, et al., 2024. On the precise quantification of the impact of a single discretionary lane change on surrounding traffic. *arXiv preprint arXiv:2407.18557*,

Hu X, Zheng Z, Chen D, et al., 2022a. Processing, assessing, and enhancing the waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research Part C: Emerging Technologies*, 134:103490. https://doi.org/10.1016/j.trc.2021.103490

Hu Y, Li Y, Huang H, et al., 2022b. A high-resolution trajectory data driven method for real-time evaluation of traffic safety. *Accident Analysis & Prevention*, 165:106503. https://doi.org/10.1016/j.aap.2021.106503

Islam Z, Abdel-Aty M, 2023. Traffic conflict prediction using connected vehicle data. *Analytic Methods in Accident Research*, 39 https://doi.org/10.1016/j.amar.2023.100275

Jansen RJ, Varotto SF, 2022. Caught in the blind spot of a truck: A choice model on driver glance behavior towards cyclists at intersections. *Accident Analysis & Prevention*, 174:106759. /https://doi.org/10.1016/j.aap.2022.106759

Ji Q, Lyu H, Yang H, et al., 2023. Bifurcation control of solid angle car-following model through a time-delay feedback method. *Journal of Zhejiang University-SCIENCE A*, 24(9):828-840. https://doi.org/10.1631/jzus.A2300026

Jiao Y, Calvert SC, Van Cranenburgh S, et al., 2025. A unified probabilistic approach to traffic conflict detection. *Analytic Methods in Accident Research*, 45:100369. https://doi.org/10.1016/j.amar.2024.100369

Katrakazas C, Quddus M, Chen WH, 2018. A simulation study of predicting real-time conflict-prone traffic conditions.

Unedited

IEEE Transactions on Intelligent Transportation Systems, 19(10):3196-3207. https://doi.org/10.1109/TITS.2017.2769158

Li G, Yang Y, Li S, et al., 2022. Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness. *Transportation Research Part C: Emerging Technologies*, 134:103452. https://doi.org/10.1016/j.trc.2021.103452

Li J, Wang X, Yang X, et al., 2024a. Analyzing freeway safety influencing factors using the catboost model and interpretable machine-learning framework, shap. *Transportation Research Record*, 2678(7):563-574. https://doi.org/10.1177/03611981231208903

Li J, Cheng R, 2025. A real-time adaptive signal control method for multi-intersections in mixed connected vehicle environments. *J. Zhejiang Univ.-Sci. A Appl. Phys. Eng*, 1:189.

Li L, Lyu H, Wang T, et al., 2024b. Stdi4dmpc: Distributed model predictive control for connected and automated truck platoon with mixed traffic flow based on spatiotemporal trajectory prediction. *IEEE Transactions on Vehicular Technology*,

Li P, Abdel-Aty M, Cai Q, et al., 2020. The application of novel connected vehicles emulated data on real-time crash potential prediction for arterials. *Accident Analysis & Prevention*, 144

Liu M, Zhao J, Hoogendoorn S, et al., 2022. A single-layer approach for joint optimization of traffic signals and cooperative vehicle trajectories at isolated intersections. *Transportation Research Part C: Emerging Technologies*, 134:103459. https://doi.org/10.1016/j.trc.2021.103459

Lundberg S, 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*,

Martinez J-J, Canudas-De-Wit C, 2007. A safe longitudinal control for adaptive cruise control and stop-and-go scenarios. *IEEE Transactions on control systems technology*, 15(2):246-258.

Mohammadian S, Haque MM, Zheng Z, et al., 2021. Integrating safety into the fundamental relations of freeway traffic flows: A conflict-based safety assessment framework. *Analytic Methods in Accident Research*, 32:100187. https://doi.org/10.1016/j.amar.2021.100187

Montanino M, Punzo V, 2015. Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. *Transportation Research Part B: Methodological*, 80:82-106. /https://doi.org/10.1016/j.trb.2015.06.010

Orsini F, Gecchele G, Rossi R, et al., 2021. A conflict-based approach for real-time road safety analysis: Comparative evaluation with crash-based models. *Accid Anal Prev*, 161:106382. https://doi.org/10.1016/j.aap.2021.106382

Ouyang P, Yanyong G, Pan L, et al., 2025. An approach for evaluating traffic safety of expressway weaving segments: Investigating risk patterns of lane-changing conflicts. *Journal of Transportation Safety & Security*, 17(2):125-157. https://doi.org/10.1080/19439962.2024.2372292

Peng Z, Jingping Z, Hao J, et al., 2024. A comparative analysis of risk factors in taxi-related crashes using xgboost and shap. *International Journal of Injury Control and Safety Promotion*, 31(3):508-520. https://doi.org/10.1080/17457300.2024.2349555

Rafati Fard M, Shariat Mohaymany A, Shahri M, 2017. A new methodology for vehicle trajectory reconstruction based on wavelet analysis. *Transportation Research Part C: Emerging Technologies*, 74:150-167. https://doi.org/10.1016/j.trc.2016.11.010

Ribeiro MT, Singh S, Guestrin C, 2016. "Why should i trust you?". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p.1135-1144. https://doi.org/10.1145/2939672.2939778

Safavian SR, Landgrebe D, 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660-674. https://doi.org/10.1109/21.97458

Shahana A, Vedagiri P, 2024. Developing rear-end and side-swipe conflict prediction models for urban signalized intersections under disordered traffic conditions. *IATSS Research*, 48(1):1-13. https://doi.org/10.1016/j.iatssr.2023.12.004

Shawky M, Alsobky A, Al Sobky A, et al., 2023. Traffic safety assessment for roundabout intersections using drone photography and conflict technique. *Ain Shams Engineering Journal*, 14(6):102115. https://doi.org/10.1016/j.asej.2023.102115

Thiemann C, Treiber M, Kesting A, 2008. Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data. *Transportation Research Record*, 2088(1):90-101. https://doi.org/10.3141/2088-10

Tian J, Zhang HM, Treiber M, et al., 2019. On the role of speed adaptation and spacing indifference in traffic instability: Evidence from car-following experiments and its stochastic model. *Transportation Research Part B: Methodological*, 129:334-350. https://doi.org/10.1016/j.trb.2019.09.014

Tian Z, Chen F, Ma S, et al., 2024. Analysis of the severity of heavy truck traffic accidents under different road conditions. *Applied Sciences*, 14(22):10751.

Vapnik VN, 1995. The nature of statistical learning theory. translators,

Venthuruthiyil SP, Chunchu M, 2022. Anticipated collision time (act): A two-dimensional surrogate safety indicator for trajectory-based proactive safety assessment. *Transportation Research Part C: Emerging Technologies*, 139:103655. https://doi.org/10.1016/j.trc.2022.103655

Wächter A, Biegler LT, 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25-57. https://doi.org/10.1007/s10107-004-0559-y

Unedited

Wang C, Xu C, Dai Y, 2019. A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. *Accident Analysis & Prevention*, 123:365-373. https://doi.org/10.1016/j.aap.2018.12.013

Ward JR, Agamennoni G, Worrall S, et al., 2015. Extending time to collision for probabilistic reasoning in general traffic scenarios. *Transportation Research Part C: Emerging Technologies*, 51:66-82. https://doi.org/10.1016/j.trc.2014.11.002

Wu F, Stern RE, Cui S, et al., 2019. Tracking vehicle trajectories and fuel rates in phantom traffic jams: Methodology and data. *Transportation Research Part C: Emerging Technologies*, 99:82-109. https://doi.org/10.1016/j.trc.2018.12.012

Xie K, Yang D, Ozbay K, et al., 2019. Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure. *Accident Analysis & Prevention*, 125:311-319. https://doi.org/10.1016/j.aap.2018.07.002

Xing L, He J, Abdel-Aty M, et al., 2019. Examining traffic conflicts of up stream toll plaza area using vehicles' trajectory data. *Accident Analysis & Prevention*, 125:174-187. https://doi.org/10.1016/j.aap.2019.01.034

Xing L, He J, Li Y, et al., 2020. Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza. *Accident Analysis & Prevention*, 135:105343. https://doi.org/10.1016/j.aap.2019.105343

Yao R, Zeng W, Chen Y, et al., 2021. A deep learning framework for modelling left-turning vehicle behaviour considering diagonal-crossing motorcycle conflicts at mixed-flow intersections. *Transportation Research Part C: Emerging Technologies*, 132:103415. https://doi.org/10.1016/j.trc.2021.103415

Yuan C, Li Y, Huang H, et al., 2022. Using traffic flow characteristics to predict real-time conflict risk: A novel method for trajectory data analysis. *Analytic Methods in Accident Research*, 35 https://doi.org/10.1016/j.amar.2022.100217

Zaki MH, Sayed T, Shaaban K, 2014. Use of drivers' jerk profiles in computer vision–based traffic safety evaluations. *Transportation Research Record*, 2434(1):103-112. https://doi.org/10.3141/2434-13

Zhao J, Ma R, Wang M, 2024a. A behaviourally underpinned approach for two-dimensional vehicular trajectory reconstruction with constrained optimal control. *Transportation Research Part C: Emerging Technologies*, 159 https://doi.org/10.1016/j.trc.2024.104489

Zhao J, Xiaoliang Y, And Zhang C, 2024b. Vehicle trajectory reconstruction for intersections: An integrated wavelet transform and savitzky-golay filter approach. *Transportmetrica A: Transport Science*, 20(2):2163207. https://doi.org/10.1080/23249935.2022.2163207

Zheng L, Hu Z, Sayed T, 2023a. Traffic conflict prediction at signal cycle level using bayesian optimized machine learning approaches. *Transportation Research Record*, 2677(5):183-195. https://doi.org/10.1177/03611981221128812

Zheng O, Abdel-Aty M, Yue L, et al., 2023b. Citysim: A drone-based vehicle trajectory dataset for safety-oriented research and digital twins. *Transportation Research Record: Journal of the Transportation Research Board*, 2678(4):606-621. https://doi.org/10.1177/03611981231185768

Zhou M, Qu X, Li X, 2017. A recurrent neural network based microscopic car following model to predict traffic oscillation. *Transportation Research Part C: Emerging Technologies*, 84:245-264. https://doi.org/10.1016/j.trc.2017.08.027

Zhu Z, Meng Y, Cheng R, 2025. Container truck high-risk events prediction and its influencing factors analyses based on trajectory data. *Systems*, 13(5):326.

Electronic supplementary materials

Section S1, Table S1, Figs. S1–S3

中文概要

题　目：基于轨迹重构数据的集装箱卡车横向风险预测及影响因素分析

作　者：祝志豪，刘赫煊，程荣军

机　构：宁波大学，海运学院，中国宁波，315211

目　的：港口集装箱卡车在运行过程中，由于车辆微观行为特征复杂多变，易诱发侧擦和追尾等交通冲突事件，进而对道路安全造成潜在威胁。本文旨在研究集装箱卡车运行中关键微观行为特征对交通冲突的影响机理，构建基于可解释性机器学习的冲突预测方法，以提升冲突事件的提前识别能力与分析可信度，从而为主动交通安全管理提供支持。

创新点：1. 提出基于二维空间的 TTC 指标，实现了侧擦冲突事件的有效识别，弥补了传统单维 TTC 方法在侧滑冲突识别中的不足；2. 构建面向集装箱卡车的可解释冲突预测模型，通过特征组合与消融实验优化预测性能，并利用 SHAP 分析揭示关键微观行为特征的影响机理，为主动安全决策提供支持。

方　法：1. 通过轨迹重构方法对无人机采集的容器卡车交通流数据进行处理，提升轨迹的连续性和物理一致性，为后续特征提取和冲突分析提供可靠数据基础；2. 基于二维空间构建 TTC 指标，实现侧滑冲突事件的识别与提取，并结合微观行为特征构建侧擦和追尾冲突样本集，为冲突预测模型的建立提供数据支持；

Unedited

3. 通过多种机器学习模型对典型冲突事件进行短时预测，并采用 SHAP 方法从可解释性角度分析关键特征对两类冲突的影响机理，验证所提出方法的有效性。

结　论： 1. 轨迹重构能够有效提升无人机采集数据在速度、加速度和加加速度层面的合理性，使其更符合车辆真实运动特性；2. 考虑车辆在二维空间中的交互关系，构建的 2D-TTC 指标可以有效识别侧滑冲突事件，并揭示横向速度与短时纵向速度特征对冲突发生的显著影响；3. 基于微观行为特征构建的冲突预测模型中，GBDT 与 XGBoost 分别在侧滑和追尾冲突识别中表现最佳，模型精度与 AUC 指标均达到较高水平，为主动交通安全管理提供可靠技术支撑。

关键词： 实时冲突预测；集装箱卡车数据集；轨迹重构；可解释机器学习；侧滑冲突。

Unedited