Accident Analysis and Prevention 156 (2021) 106122

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

An integrated methodology for real-time driving risk status prediction using naturalistic driving data

Qiangqiang Shangguan [a,b], Ting Fu [a,b], Junhua Wang [a,b,*], Tianyang Luo [a,b], Shou'en Fang [a,b]

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, 201804, China
[b] College of Transportation Engineering, Tongji University, 4800 Cao'an Highway, Shanghai, 201804, China

ARTICLE INFO

Keywords:
Driving risk status prediction
Rolling time window approach
Naturalistic driving data
Car-following events
Machine learning algorithms

ABSTRACT

Real-time driving risk status prediction is critical for developing proactive traffic intervention strategies and enhance driving safety. However, the optimal observation time window length and prediction time window length, which should be the prerequisite for the timeliness and accuracy of real-time driving risk status prediction model, have been rarely explored in previous studies. In this study, a methodology which integrates driving risk status identification, rolling time window-based feature extraction, real-time driving risk status prediction and driving risk influencing factors analysis was proposed to accurately evaluate and predict real-time driving risk status. The methodology was tested based on 1,440 car-following events from Shanghai Naturalistic Driving Study. Results show that four driving risk statuses (safe, low-risk, median-risk and high-risk) are most appropriate to establish risk labelling criteria. In addition, results from driving risk status prediction show that when the observation time window length is 0.5 s, the accuracy rate of predicting medium-risk or high-risk status occurring in the next 0.7 s is higher than 85 % using multi-layer perceptron model. Meanwhile, the results from the analysis of influencing factors show that the input variables related to the risk status score higher in the ranking of feature importance. A part from that, speed difference, headway distance, speed and acceleration are still important in predicting driving risk status. The proposed methods in this paper can be applied in connected and autonomous vehicle (CAV) to reduce driver cognitive workload and hence improve driving safety fed with naturalistic driving data collected using in-vehicle systems.

1. Introduction

As shown in previous studies, the main causes of the traffic crashes are driver's incorrect perception, wrong judgement, inability to respond to emergency situations in time, underestimation of braking requirements, and other related factors such as vehicle and road conditions. (Lee and Yeo, 2016). To mitigate the traffic crash risk related to drivers, many researchers have dedicated to the driving risk status prediction (Katrakazas et al., 2019; Cai et al., 2020). In addition, due to the development of technology in the field of connected and autonomous vehicle (CAV), the investigation of real-time driving risk prediction is becoming increasingly important.

With the development of data acquisition technology, data sources for evaluate and predict real-time driving risk are becoming more and more diverse. Typical approaches include survey-based studies, macro crash data studies, driving simulation studies and naturalistic driving studies (NDS). Researchers began to use questionnaires to study drivers' risky driving behaviors to obtain the relationship between behavioural characteristics and traffic crash risks (Scott-Parker and Weston, 2017). With the development of traffic data collection methods, more and more researchers have paid attention to the traffic crash prediction. It often involves extracting and analyzing the macro statistical data of traffic crash to investigate the impact of traffic flow characteristics on traffic crash risk (Yu and Abdel-Aty, 2013b). Furthermore, due to its convenience and stability, some studies used driving simulator to investigate the driving risk in some specific risky scenarios (Shangguan et al., 2020). However, due to limited data availability, these types of data collection methods have two major drawbacks in real-time driving risk prediction. One is that the limited data sample size, the risk prediction model is usually constructed based on specific scenarios. Another shortcoming is crash data or risk-related parameters cannot be obtained in real time, and therefore cannot be used for real-time driving risk prediction. In

* Corresponding author at: The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai, 201804, China.
E-mail addresses: 1710912@tongji.edu.cn (Q. Shangguan), tingfu@tongji.edu.cn (T. Fu), benwjh@163.com (J. Wang), lty-paul@qq.com (T. Luo), fangsek@tongji.edu.cn (S. Fang).

https://doi.org/10.1016/j.aap.2021.106122
Received 30 November 2020; Received in revised form 27 January 2021; Accepted 1 April 2021
Available online 23 April 2021

contrast, in NDS, various real-time driving behavior characteristics, surrounding traffic conditions, and environment parameters can be obtained using in-vehicle devices, cameras, and sensors (Wang et al., 2020a). Due to the comprehensiveness and authenticity of the collected data, it can provide more effective real-time driving risk prediction results. Therefore, authentic and real-time naturalistic driving data collected from in-vehicle systems such as advanced driver assistance systems (ADAS) can be a practical alternative.

Despite many efforts on the topic of driving risk prediction using naturalistic driving data, some research gaps still exist. Firstly, driving risk status labelling is a valuable work but still lacking. The surrogate measures of safety (SMoS) are commonly used in previous studies to evaluate driving risk based on microscopic driving data, such as Time to Collision (TTC). However, it is more difficult to establish reasonable thresholds to determine reasonable risk level (Shi et al., 2019). Secondly, the appropriate length of observation-prediction time window significantly affects the accuracy of risk level prediction. However, the observation-prediction time windows used in existing studies are arbitrary and inconsistent, ranging from a few seconds to tens of seconds (Xiong et al., 2018; Chen et al., 2019; Panagopoulos and Pavlidis, 2019). Thirdly, naturalistic driving data contains a large number of behavior characteristics, surrounding traffic conditions, and environment parameters, so reasonable feature selection is essential to the efficiency and accuracy of the prediction model (Shi et al., 2019). Additionally, the data for different driving risk levels are usually imbalanced. The use of imbalanced data in machine learning algorithms may lead to biased prediction results and weak generalization ability (Wang et al., 2020b). These above problems make the real-time driving risk prediction more complicated.

This study aims to propose an integrated methodology for accurately evaluating and predicting real-time driving risk status, which integrates driving risk status identification, rolling time window-based feature extraction, real-time driving risk status prediction and driving risk influencing factors analysis. The remainder of this paper is organized as follows: the previous studies about driving risk prediction are introduced in Section 2. Car-following events extraction from naturalistic driving database are presented in Section 3. The methodology of this paper is described in Section 4. Section 5 elaborates on the results of driving risk prediction and the discussion of this study. The final section covers the conclusions of this study.

2. Literature review

2.1. Overview of data sources for driving risk prediction

As mentioned before, approaches to predict driving risk can be classified into four main categories based on data acquisition methods: survey-based studies, macro crash data studies, driving simulation studies and NDS.

Firstly, survey-based studies, which mainly include the questionnaire method, interview method and the online-survey method. It often involves extracting the personality, attitude and risk perception data of the driver to investigate the relationship between driving risk and driver's personality characteristics (Bıçaksız and Özkan, 2016). Although these survey-based studies can reflect the relationship between driving risk and driver's characteristics, these are two major limitations. On the one hand, the research based on the survey is relatively subjective and the sample size is limited. On the other hand, the respondent may be reluctant to provide risky driving behavior information, so the authenticity of the survey data is also limited. Secondly, macro crash data studies, including traffic crash cause analysis and traffic crash risk prediction. Previous studies have shown that the degree of dispersion of the speed and the traffic volume are considered to be the main factors causing traffic crash risk (Ahmed and Abdel-Aty, 2013; Shi and Abdel-Aty, 2015). Despite the great efforts on driving safety risk analysis based on macro crash data, there are still some limitations. As mentioned in the previous studies, historical crash data has the characteristics of small sample and mislocation, which cannot be used for driving risk prediction (Fu et al., 2018). Besides, detailed driving data, such as vehicle kinematics parameters, and other road users' motion characteristics, to reflect the real-time driving risk are still limited. Thirdly, driving simulation studies, which have been widely adopted in driving risk research due to its high safety, strong controllability and comprehensive data acquisition advantages (Wu et al., 2018). However, it is difficult to restore the real scene in the driving simulation experiment, and it is necessary to further verify the experimental results in combination with the actual vehicle test.

Alternatively, NDS, which continuously recording the driver's operations, vehicle motion status and surrounding environment characteristics, have been developed in recent years (Raju et al., 2019). The 100-Car NDS, conducted by the NHTSA, was the first naturalistic driving study in the US. The analyses presented in one technical report using the 100-Car NDS data established the relationships between driving behavior and traffic crash (Klauer et al., 2006). After the 100-Car NDS, NHSTA carried out the Strategic Highway Research Program 2 (SHRP 2), which took 3 years to record 35 million miles of driving data (Bärgman et al., 2015). Arbabzadeh and Jafari (2017) used the data from SHRP 2 NDS to establish the real-time driving safety risk prediction model using multinomial logistic regression approach. In addition, based on naturalistic driving dataset, Wang et al. (2019b) extracted 5608 cut-in events and studied the traffic characteristics and other influencing factors of cut-in behavior. NDS is expected to provide highly reliable driving data and a variety of experimental scenarios, which can improve the accuracy of driving safety risk assessment and prediction.

2.2. Overview of modeling approaches for driving risk prediction

The current driving risk prediction models used in related research mainly include statistical regression models and machine learning algorithms.

Most previous studies have conducted driving risk prediction based on statistical regression models. Although these models are mainly used traditional historical statistical data for risk prediction, the model is more interpretable. Common statistical regression models proposed in previous studies for traffic crash risk prediction include logistic regression models and negative binomial regression models (Ahmed et al., 2012; Guo and Fang, 2013; Hassan and Abdel-Aty, 2013; Yu and Abdel-Aty, 2013b, a). For example, Yu and Abdel-Aty (2013a) used a random-effect logistic regression model to study the difference of traffic crashes between workday and weekend, and found that crashes on workday are more likely to occurred under congested-flow conditions while crashes on weekend often occurred under free-flow conditions. Ahmed et al. (2012) proposed a Bayesian logistic regression model to predict traffic crashes. The accuracy of prediction results reached 72.5 %, but the false alarm rate also increased to 45.83 %. Hassan and Abdel-Aty (2013) used the "case-control" logistic regression method to investigate the traffic crash risk under low visibility. The study combined traffic flow data and meteorological data, and the final crash prediction accuracy reached 69 %. Guo and Fang (2013) identified the high-risk drivers using 100-Car NDS data and negative binomial regression model. The statistical regression prediction model has good interpretability. The model can obtain the variables that significantly affect the driving risk to help traffic managers choose appropriate traffic safety management strategies. However, improving the efficiency and accuracy of these statistical regression prediction models is still challenging.

The driving risk prediction model based on machine learning algorithms has significantly improved the prediction accuracy and has been widely adopted by researchers in recent years. These models include eXtreme gradient boosting (XGBoost) (Panagopoulos and Pavlidis, 2019; Shi et al., 2019), support vector machine (SVM) (Yu and Abdel-Aty, 2013b; You et al., 2017), random forest (RF) (Lin et al.,

2015), Markov model (Xiong et al., 2018), and several neural network models (Chen et al., 2019; Wang et al., 2019a; Costela and Castro-Torres, 2020). For example, Panagopoulos and Pavlidis (2019) trained a short term prediction model to recognize risky driving behavior and predict driving risks using XGBoost. In this study, the adopted XGBoost algorithm takes the last 30 s of driving data as input and makes risk prediction for the next 10 s. Shi et al. (2019) designed a framework based on XGBoost that integrates feature selection, risk labelling and imbalanced data resampling to assess and prediction driving risk levels, and the prediction accuracy of proposed model can reach about 89 %. Besides, You et al. (2017) applied case-control method and SVM to identify the crash risk status. Lin et al. (2015) used RF method to select input variables for driving risk prediction. In addition, Xiong et al. (2018) utilized Markov chain to illustrate the transition pattern of driving risk status, and then developed multinomial logistic model to improve the performance of prediction algorithm. The results showed that the accuracy of driving risk status prediction model reaches over 85 % when the input and output time window of prediction model are 1.4 s and 0.8 s, respectively. Furthermore, Wang et al. (2019a) used back propagation neural network to predict the driving risk on expressways through simulation experiment. Chen et al. (2019) proposed a novel neural network model for crash risk prediction and found that by using 15 s as the optimal time window length of the input variable can significantly reduce the prediction error. However, this study only predicts whether a crash event will occur, but does not predict when it will occur in the future. Costela and Castro-Torres (2020) applied eye movement characteristics to driving risk prediction based on feedforward neural network.

In general, most previous studies have focused on using statistical regression methods or machine learning methods to predict traffic crashes or driving risks. However, as a key issue of real-time driving risk prediction, how to choose the appropriate length of observation-prediction time window is rarely mentioned and explored in previous studies. In this study, an integrated methodology will be proposed to address the research gap and used for real-time driving risk status prediction.

3. Data preparation

3.1. Introduction of the Shanghai Naturalistic Driving Study (SH-NDS)

This paper uses the real-world driving data provided in the SH-NDS to predict real-time driving risk. As described in previous studies, the SH-NDS has collected approximately 161,055 km of naturalistic driving data from 2012 to 2016 (Wang et al., 2019b).

The data acquisition system used in SH-NDS mainly consists of six parts: i) a high-definition camera for recording four camera views, as shown in Fig. 1; ii) a global positioning system (GPS) for collecting the latitude and longitude position information of the vehicle; iii) a 3-axis accelerometer for measuring the acceleration of the vehicle; iv) a Doppler radar for measuring the relative speed and relative distance between the naturalistic driving vehicle and the leading vehicle (LV); v) a lane deviation warning system for warning the driver when crossing a lane without turning on the turn signal; and vi) an emergency event button. The data collection frequency of the SH-NDS is 10 Hz.

3.2. Car-following events extraction

This study extracts car-following events from the SH-NDS and applies them to real-time driving risk status prediction, and based on this, verifies the effectiveness of the proposed integrated methodology. Before extracting car-following events, the original data was preprocessed, including deleting outliers, repairing data and data smoothing. Then, the car-following events were extracted based on several criteria proposed by Zhu et al. (2018):

- The detected ID of LV $>$ 0 and remained constant: ensuring the following vehicle (FV) was preceded by the same LV;
- Longitudinal distance between FV and LV $<$ 120 m: eliminating the completely unrestricted car-following situations;
- Lateral distance between FV and LV $<$ 2 m: guaranteeing that the FV and the LV are in the same lane and no overtaking behavior;
- Speed of FV $>$ 10 km/h: eliminating low-speed waiting events;
- Duration of each car-following event $>$ 15 s: ensuring that each extracted car-following event contains enough driving data for driving risk prediction.

After applied the above extraction criteria, this research extracted 1,440 car-following events and then utilized them for driving risk status prediction.

4. Methodology

The overall flowchart of the methodology proposed in this paper is illustrated in Fig. 2. In addition to data preparation mentioned in the previous part of the paper, it mainly consists of four parts: i) driving risk status identification, ii) feature extraction using rolling time window approach, iii) real-time driving risk status prediction, and iv) analysis of

Fig. 1. The face view, hands view, front view and rear view from SH-NDS.

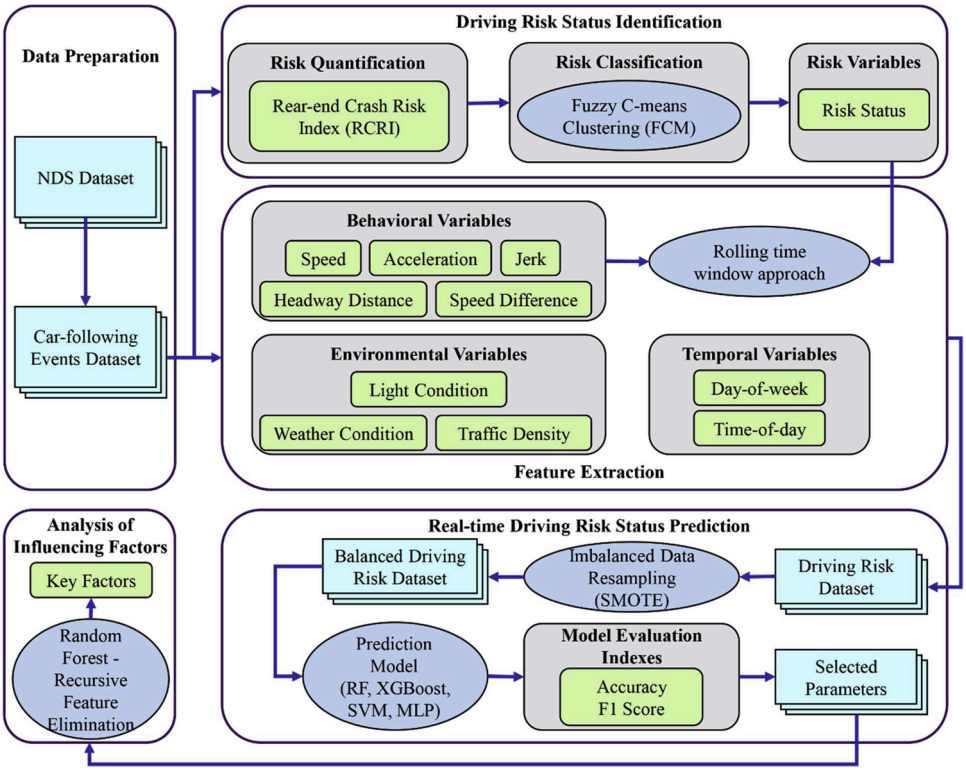

Fig. 2. Overall methodology flowchart.

influencing factors using random forest – recursive feature elimination (RF-RFE). The driving risk status identification part mainly includes driving risk quantification and classification. Feature extraction is mainly to extract input and output variables from the car-following data, and further used for real-time risk status prediction. Driving risk status prediction applies supervised algorithms to realize real-time prediction of driving risk status. Finally, the RF-RFE method is used to identify factors that have a significant impact on driving risk. The four parts of the methodology are detailed below.

4.1. Driving risk status identification

4.1.1. Description of the crash risk index

As mentioned in previous studies, the crash risk can be determined by evaluating the initial conditions of the LV and the FV, the disturbance of the LV, the reaction time and the evasive action taken by follower

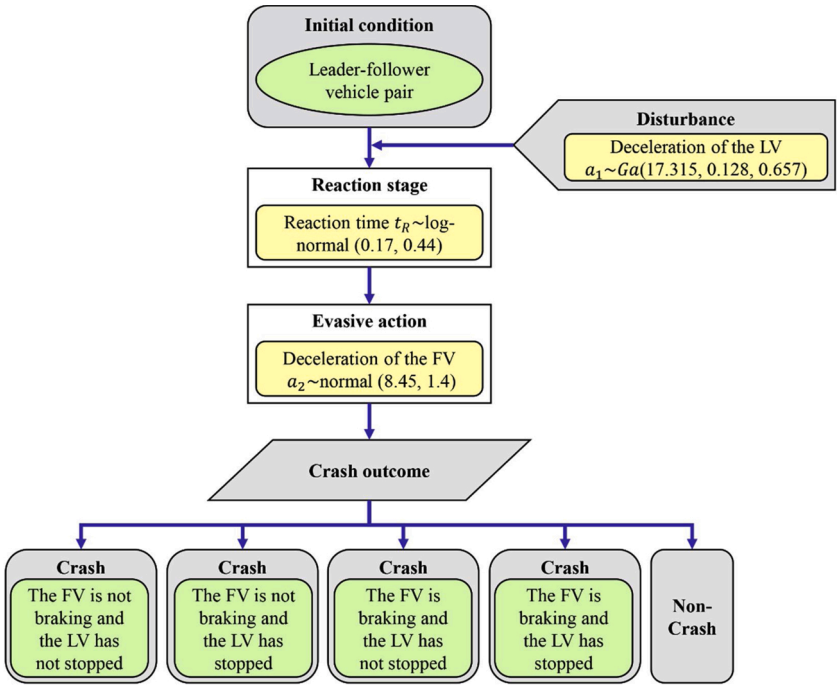

Fig. 3. The concept diagram of the RCRI.

(Johnsson et al., 2018). In addition, the disturbance, reaction time and the maximum available deceleration rate all follow a specific distribution. As described in Shangguan et al. (2021), the Monte Carlo simulation method was carried out to randomly select the deceleration rate taken by the LV and the response time and deceleration rate of the FV. Then, according to the initial conditions of the leader-follower vehicle pair and the above three selected parameters, the crash outcome which includes the probability and consequence of the crash can be obtained. The rear-end crash risk index (RCRI) is finally defined as the product of the probability and the consequence of the crash event.

The concept diagram of the RCRI is shown in Fig. 3. According the previous studies, the deceleration rate taken by LV follow a shifted gamma distribution (17.315, 0.128, 0.657) (Kuang et al., 2015). The reaction time of the FV follows a log-normal distribution (0.17, 0.44), and the braking coordination time is 0.175 s (Wang, 2002). The maximum available deceleration rate of cars is specified as a truncated normal distribution. The maximum deceleration rate of cars is distributed between 4.23 m/s^2 and 12.68 m/s^2, with an average value of 8.45 m/s^2 and a variance of 1.40 m/s^2 (Cunto and Saccomanno, 2008). Then, based on the five crash outcomes, the RCRI, which considers the crash probability and severity simultaneously, can be calculated and shown as below. Detailed information about the calculation of RCRI was described in Shangguan et al. (2021).

$$RCRI_i = \frac{\sum_{j=1}^{N} crash_{ij} \times SASD_{ij}}{N} \quad (1)$$

where $RCRI_i$ indicates the risk of crash for i^{th} vehicle interaction; N = 10,000 is the total number of samples generated by the Monte Carlo simulation method. A crash occurs, $crash_{ij} = 1$; otherwise, $crash_{ij} = 0$. The square of the absolute speed difference ($SASD_{ij}$) between LV and FV at the time of the crash is defined as the severity of the crash risk.

4.1.2. Fuzzy C-means (FCM) clustering algorithm for driving risk status labelling

In order to apply RCRI to driving risk status prediction, the evaluation criteria of RCRI was developed based on FCM clustering algorithm. The FCM clustering algorithm adopts the concept of fuzzy theory. Compared with K-Means clustering, FCM provides more flexible clustering results. In most cases, the samples in the data set cannot be divided into clearly separated clusters. Therefore, it is more reasonable to assign a weight to each sample and each cluster to indicate that the sample belongs to a certain cluster. In this paper, the FCM cluster algorithm was employed to group the RCRI into different levels based on the minimization of the loss function. The detailed calculation process of the loss function was described in Oh et al. (2006).

4.2. Feature extraction using rolling time window approach

To make the form of the input variables more suitable for real-time driving risk status prediction, a rolling time window approach was proposed to reconstruct driving features and determine the optimal observation-prediction time window length. As shown in Fig. 4, t_1 and t_m represent the beginning and the end of the risk prediction respectively, and $x_{1,\ldots,n}$ represents the extracted input features. The time window length of the observation sample is ω, and it is continuously rolling forward along the rolling direction with a rolling time step δ (equals to the data collection frequency 0.1 s). Besides, the time window length of the prediction label is φ. To achieve the optimal driving risk prediction accuracy, the length of observation time window ω and prediction time window φ for feature extraction will be selected based on the performance of driving risk status prediction model by grid searching.

In this study, the highest driving risk status in the prediction time window was selected as the output variable, which is determined by the risk status characteristics, behavioral characteristics and several traffic characteristics within observation time window. The mean RCRI within observation time window (R_M), the last observed driving risk status in the observation time window (R_L), and a trend statistic value of the driving risk status in the observation time window (R_T) were selected as the risk status characteristics and input into the driving risk prediction model. R_T was defined as follows:

$$R_T = \sum_{a,b} (b-a)|b-a|N_{ab} \quad (2)$$

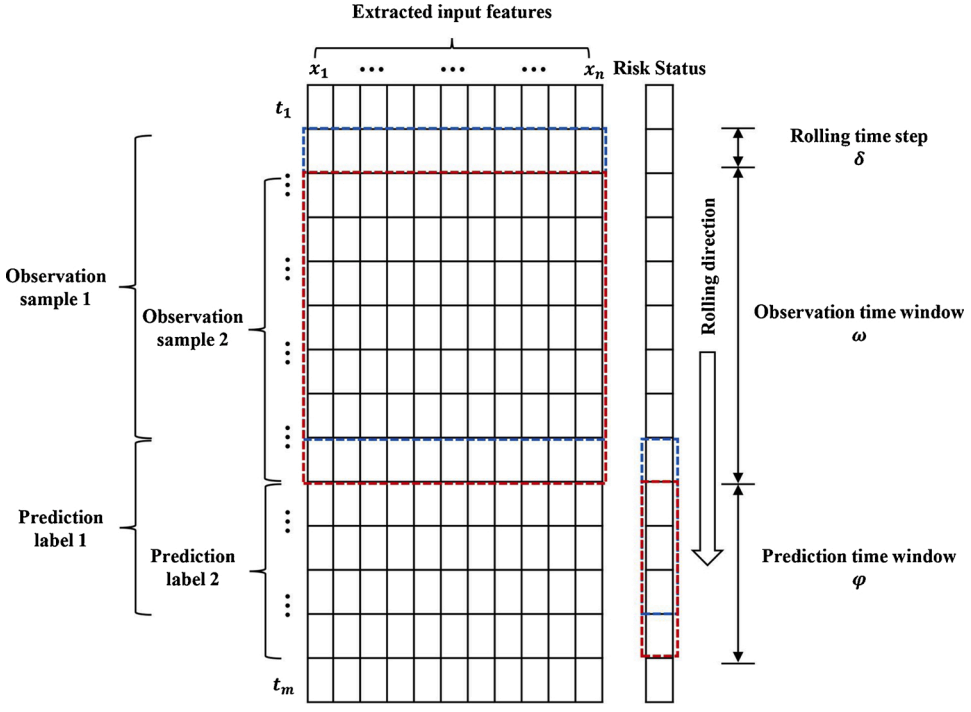

Fig. 4. Schematic diagram of the rolling time window approach.

$$N_{ab} = \frac{number\ of\ driving\ risk\ status\ pairs\ (a,b)}{total\ number\ of\ possible\ driving\ risk\ status\ pairs} \quad (3)$$

where a and b represent the instantaneous driving risk status.

Traffic characteristics consist of the following variables: i) behavioral variables, including vehicle speed, acceleration, jerk (the change rate of vehicle acceleration), headway distance and speed difference between LV and FV; ii) Temporal variables, including time-of-day and day-of-week; iii) Environmental variables, including light conditions, weather conditions and traffic density. Among these variables, jerk is applied because it can represent the change rate of acceleration. The larger the jerk, the faster the acceleration changes. This variable is also used in the study of identifying aggressive drivers and predicting traffic crashes (Bagdadi and Várhelyi, 2011; Feng et al., 2017). The mean, standard deviation, and slope of regression fit of the behavioral variables within observation time window were extracted as the input variables. Meanwhile, the temporal variables and environmental variables were also extracted as the input variables for driving risk prediction. Table 1 shows the summary of the extracted features.

To ensure the performance of the prediction model, the min-max standardization method was applied to the extracted continuous variables, so as to normalize the ranges of different continuous input variables to the same scale.

4.3. Real-time driving risk status prediction

In real traffic data, driving risk at different statuses always imbalanced. The proportion of high driving risk situation is much smaller than the proportion of low driving risk situation. The most common approach to address this problem is Synthetic Minority Over-sampling Technique (SMOTE) (Cai et al., 2020). The main concept of the SMOTE is to use the existing samples to artificially generate new samples of minority groups using nearest neighbor algorithm (Chawla et al., 2002). In this study, we used SMOTE to create an equal number of samples between different driving risk levels. In addition, to ensure the usability of the risk prediction model, SMOTE was only applied to training dataset to obtain the parameters of the prediction model, while the test dataset were still original unbalanced data.

Through the above data preparation and resampling approaches, the driving data at different risk statuses could be balanced. In order to obtain the optimal combination of observation-prediction time window and the best prediction model, four prediction models including RF, XGBoost, SVM, and multi-layer perceptron (MLP) and the selected evaluation indexes of prediction model are briefly introduced in the following subsections.

Table 1
Summary of the Extracted Features.

Categories	Extracted Features	Conditions
Risk Variables	R_M	Continuous variable
	R_L	Dummy variable (Different risk levels)
	R_T	Continuous variable
Behavioral Variables	Speed-mean	Continuous variable (km/h)
	Speed-std	Continuous variable (km/h)
	Speed-srf	Continuous variable
	Acceleration-mean	Continuous variable (m/s^2)
	Acceleration-std	Continuous variable (m/s^2)
	Acceleration-srf	Continuous variable
	Jerk-mean	Continuous variable (m/s^3)
	Jerk-std	Continuous variable (m/s^3)
	Jerk-srf	Continuous variable
	Headway distance-mean	Continuous variable (m)
	Headway distance-std	Continuous variable (m)
	Headway distance-srf	Continuous variable
	Speed difference-mean	Continuous variable (m/s)
	Speed difference-std	Continuous variable (m/s)
	Speed difference-srf	Continuous variable
Temporal Variables	Day-of-week	Dummy variable (Holiday, workday)
	Time-of-day	Dummy variable (Off peak, morning peak, evening peak)
Environmental Variables	Light condition	Dummy variable (Daytime, nighttime)
	Weather condition	Dummy variable (Sunny, rainy)
	Traffic density	Dummy variable (High, median, low)

Note: std - standard deviation; srf - slope of regression fit.

4.3.1. Random Forest (RF)

RF is a machine learning algorithm that uses multiple decision trees for training, classification and prediction, and is mainly used in regression and classification scenarios. RF is a well-known ensemble learning method, which belongs to the part of the ensemble learning algorithm where there is no dependence between weak learners, and because of this advantage, the RF algorithm can run in parallel. RF reduces the correlation between decision trees by randomly selecting samples and features. The model randomly selects the same amount of data from the original training data as training samples, and when building a decision tree, randomly selects some of the features to build a decision tree. Through these two kinds of random selection, the degree of similarity between the decision trees in the RF is smaller, which further improves the accuracy of the model (Svetnik et al., 2003). RF overcomes the problem of decision tree overfitting and is widely used in supervised learning. Therefore, this study uses random forest as an alternative algorithm for driving risk prediction.

4.3.2. eXtreme gradient boosting (XGBoost)

XGBoost provides a parallel tree boosting which can be conducted to deal with the binary and multi-classification problems. The general idea of XGBoost is to continuously perform feature splitting to grow a tree, and continuously learn to fit the residuals of the last prediction until the training is completed to obtain k trees. The prediction process of XGBoost is based on the characteristics of the sample, adding up the scores of the corresponding leaf nodes of each tree to obtain the optimal predicted value of the samples (Chen and Guestrin, 2016).

4.3.3. Support vector machine (SVM)

Standard SVM model is a two-class classification model. The basic concept of SVM is to construct a separating hyperplane to solve the two-class classification problem (Yu and Abdel-Aty, 2013b). Using the standard SVM calculation process to construct multiple decision boundaries in an orderly manner can achieve multi-classification of samples. In this study, a Gaussian kernel function is chosen and applied to decision function. Then, the classification decision function based on radial basis function can be represented as:

$$f(x) = sgn\left(\sum_{i=1}^{N} a_i^* y_i e^{-\gamma\|x_i - x\|^2} + b^*\right) \quad (4)$$

where $a_i^* y_i$, γ, x_i, b^* are the parameters obtained through model optimization.

4.3.4. Multi-layer perceptron (MLP)

MLP is also called artificial neural network (ANN), which mainly includes input layer, multiple hidden layers, output layer and the weights of each layer. The layers of the MLP are fully connected, that is, any neuron in the upper layer is connected to all the neurons in the next layer. The MLP model mainly uses the gradient descent method to optimize the loss function, that is, randomly initialize all parameters, and then iteratively train, continuously calculate the gradient and update the parameters, and stop training when the loss function reaches a certain threshold (Hu et al., 2020).

4.3.5. Evaluation Indexes of driving risk status prediction model

The selected evaluation indexes of driving risk status prediction model provide a basis for determining the optimal parameters and prediction model. Considering the difference in sample size under different risk statuses, the accuracy rate and the F1 score that can integrate precision and recall were both calculated to evaluate the performance of the driving risk status prediction model (Chen et al., 2019).

4.4. Analysis of influencing factors using random forest – recursive feature elimination (RF-RFE)

Finally, the RF-RFE approach was proposed to identify the key factors that significantly affect driving risks. This two-step hybrid algorithm integrates the advantages of feature ranking and feature elimination. This algorithm firstly sorts the importance of features based on RF, and filters a set of relatively key features. Then, the RFE method was applied to find the best subset from the filtered features. The RFE feature selection method is essentially a recursive process, grading features according to certain importance (Zvarevashe and Olugbara, 2018).

5. Results and discussion

5.1. Driving risk status labelling using FCM

This study uses Wilk's lambda (Λ), which is defined as the ratio of within-group variance to the sum of within-group variance and between-group variance, to determine the appropriate number of risk clusters (Johnson and Wichern, 2002). Actually, the appropriate cluster size can be determined when the marginal Wilk's lambda ratio change is minimal (Oh et al., 2006).

$$\Lambda = \frac{|W|}{|T|} = \frac{|W|}{|B+W|} \tag{5}$$

where W is the within-group variance, T is the total variance, and B is the between-group variance.

It can be seen from Fig. 5 that no significant marginal Wilk's lambda ratio change reduction is observed after four clusters, which means the most optimal cluster size is four. Based on the cluster results and the value of RCRI, four driving risk status labelling criteria can be obtained and shown in Table 2. The safe status is considered with low RCRI up to 0.0034, accounting for 72.7 %. The RCRI of low-risk and medium-risk ranged from 0.0034 to 0.0114 and 0.0114 to 0.036, accounting for 22.1 % and 4.8 %, respectively. On the other hand, high-risk status with RCRI in excess of 0.036, accounting for 0.4 %.

In order to better demonstrate the driving risk status labelling results in this study, taking an actual car-following event as an example, the change process of driving risk status can be observed and shown in Fig. 6.

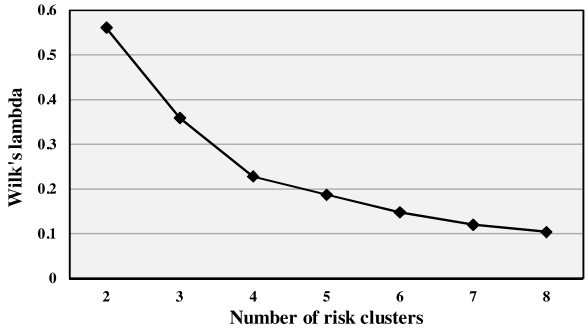

Fig. 5. FCM clustering results.

Table 2
Driving risk status labelling criteria.

Category	Real-time driving risk criteria
Safe	RCRI $\leq$ 0.0034
Low-risk	RCRI $>$ 0.0034 and RCRI $\leq$ 0.0114
Median-risk	RCRI $>$ 0.0114 and RCRI $\leq$ 0.036
High-risk	RCRI $>$ 0.036

5.2. Real-time driving risk status prediction model comparison and influencing factors analysis

5.2.1. Parameter selection and comparison of evaluation indexes

As mentioned above, the selection of optimal observation-prediction time window length is crucial for driving risk status prediction. If the length of the observation-prediction time window is not appropriate, the effective information of the input variables in the observation time window will be missing, and the driving risk status cannot be accurately predicted. In addition, as shown in previous studies, at least 60 % to 90 % of rear-end crashes can be avoided by warning the driver 0.5–1.0 s prior to the potential collisions (National Transportation Safety Board, 2001). So, considering the requirement of timeliness and prediction accuracy, the length of observation time window $\omega \in [0.1, 5]$ with a rolling time step 0.1 s, and the length of prediction time window $\varphi \in [0.5, 3]$ (gridded at 0.1 s) were selected through grid searching to obtain the optimal combination of ω and φ and then utilized in driving risk status prediction model.

In this study, we used the sklearn Python library, an open source and user-friendly package suitable for various machine learning algorithms, as the platform for applying driving risk status prediction algorithms (RF, XGBoost, SVM, and MLP). The hyperparameter settings of the four algorithms are as follows: estimators '100', depth 'unlimited,' and criterion 'gini' were set in RF. Similarly, the estimators '160', maximum depth of a tree '5', and learning rate '0.1' were set in XGBoost. The gamma value '0.1', C value '0.8', kernel function 'radial basis function' were set in SVM. The three hidden layer size '30, 30, 20', and the max number of iterations '500' were set in MLP.

To evaluate these driving risk status prediction models, the whole dataset was divided into training dataset (80 %) and test dataset (20 %). The resampling method was only applied to the training dataset for developing driving risk prediction model, while the test dataset were still original unbalanced driving data. To compare the performance of these prediction models with different time window length combinations, two evaluation indexes of accuracy rates and F1 score were calculated. The results of prediction accuracy and F1 score are presented in Figs. 7 and 8. Obviously, as the width of the prediction time window increases, the prediction accuracy and F1 value of the four models all show a downward trend. Besides, the results also show that under the same observation time window and prediction time window, the prediction accuracy and F1 score of RF and XGBoost are relatively close. In addition, the prediction accuracy and F1 value of SVM are generally lower than the other three models.

Among all driving risk status prediction results from different observation-prediction time window length, the highest accuracy rate and F1 score of each prediction model under a given observation or prediction time window length was selected as its optimal prediction results. Fig. 9 illustrates the accuracy rates of four different prediction models with different observation/prediction time window lengths. Similarly, the F1 score of four different prediction models with different observation/prediction time window lengths are shown in Fig. 10.

As shown in Figs. 9 and 10, the prediction accuracy and the F1 score did not decrease significantly with the increase of the observation time window length. On the contrary, as the length of the prediction time window increases, the prediction accuracy and the F1 score decrease significantly. The reason may be that as the length of prediction time

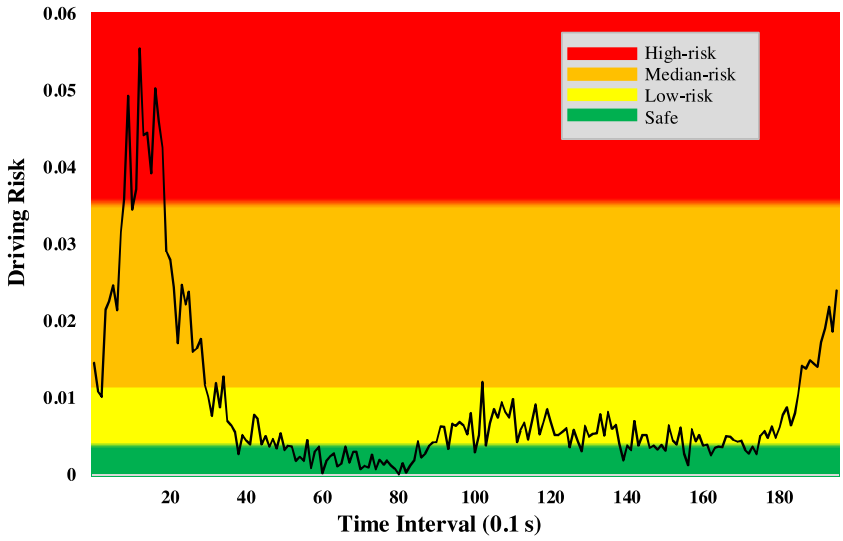

Fig. 6. Illustration of changes in driving risk.

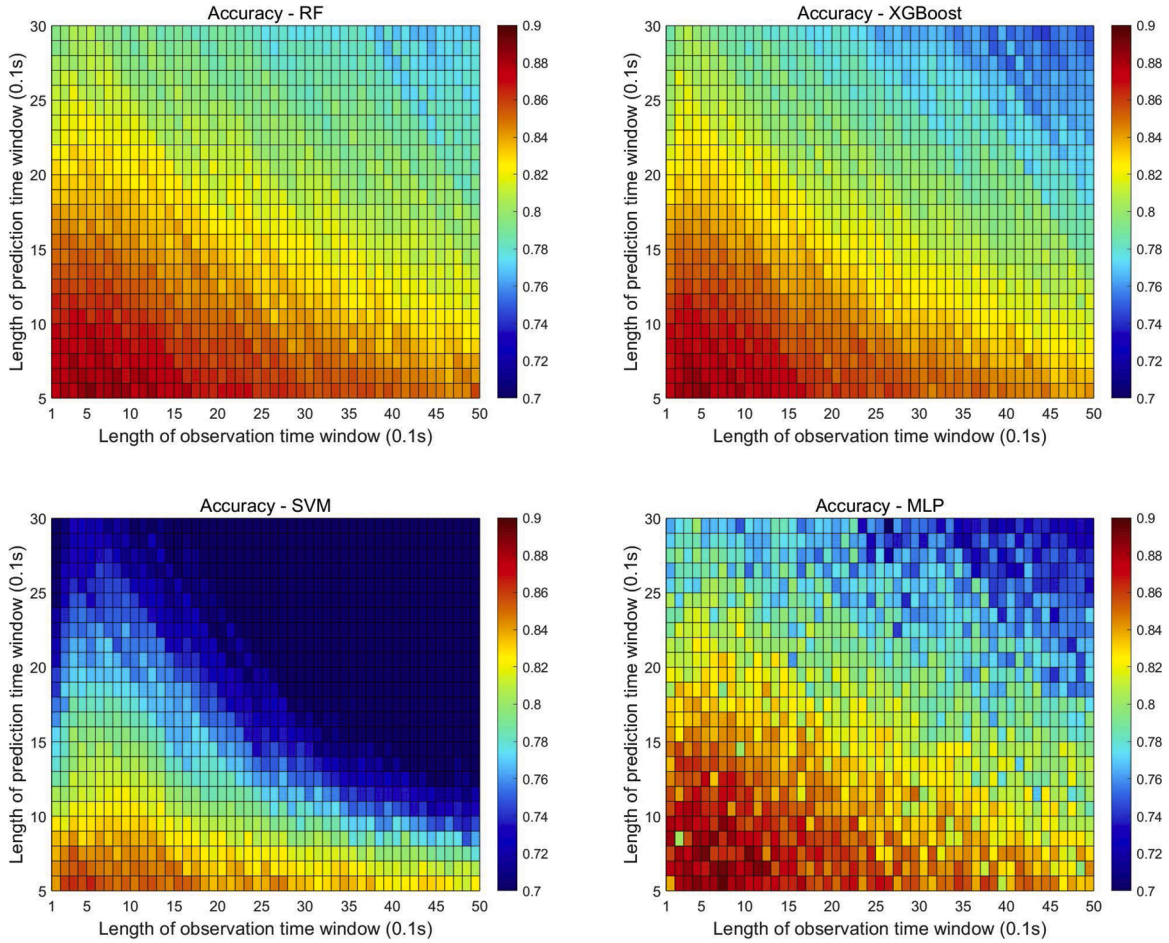

Fig. 7. Accuracy rates of prediction models with different time window length combinations.

window increases, the driving risk status needs to be predicted for a longer time in the future, which makes it more challenging to accurately predict driving risk status under the current information. In addition, results from Figs. 9 and 10 also show that the driving risk prediction performance of the RF, XGBoost and MLP models is much better than SVM.

In order to comprehensively consider the prediction accuracy and the F1 score of the driving risk status prediction model, the sum of prediction accuracy and F1 score under different time window lengths were sorted. The optimal observation time window length and prediction model under different prediction time window lengths are listed in Table 3.

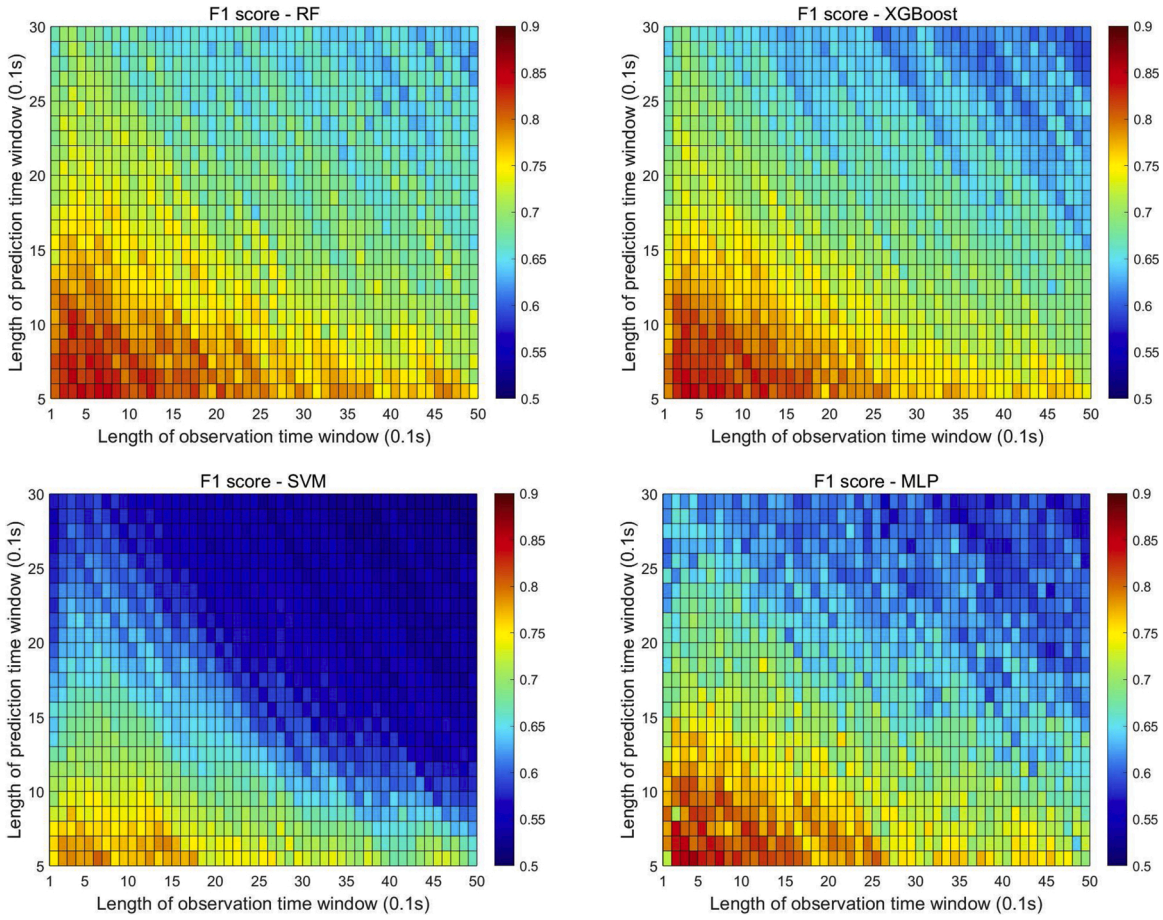

Fig. 8. F1 score of prediction models with different time window length combinations.

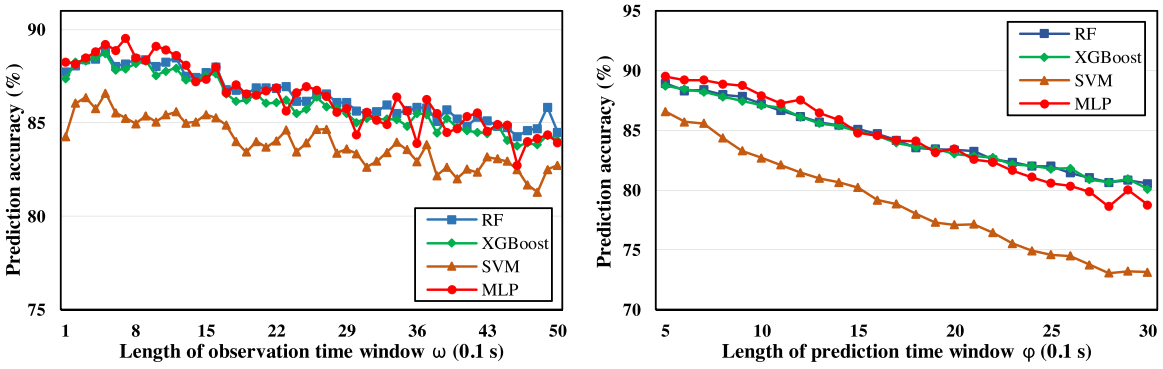

Fig. 9. Accuracy rates of prediction models with different observation/prediction time window lengths.

As shown in Table 3, all the prediction accuracy rates and F1 scores exceed 80 % and 69 %, respectively. Obviously, the prediction accuracy and F1 score gradually decrease as the length of prediction time window increases. In addition, the prediction results from Table 3 show that when the length of prediction time window is less than 1.5 s, all the accuracy rates of the prediction model are greater than 85 %. Among them, the highest prediction accuracy is 89.2 % when the prediction time window length is 0.7 s, and the F1 score of the prediction model is 83.7 %. Considering the prediction accuracy and timeliness requirements in real-time driving assistant applications, the length of observation time window $\omega = 0.5$ s and the length of prediction time window $\varphi = 0.7$ s are selected as the optimal parameters, and the MLP model is selected for driving risk status prediction. Fig. 11 presents the confusion matrix of driving risk status prediction results with selected parameters and prediction model. Accurate recognition of high-risk driving conditions is critical to improving driving safety. In previous studies, 61 % of crashes and 76 % of near-crashes can be accurately identified using multinomial logistic regression model (Arbabzadeh and Jafari, 2017). In addition, based on the approaches proposed by Shi et al. (2019), the prediction accuracy of low-risk, medium-risk and high-risk are 85.5 %, 79.8 %, and 82.4 % respectively. As the prediction results shown in Fig. 11, the selected model has a prediction accuracy rate of higher than 85 % for both the median-risk status and high-risk status.

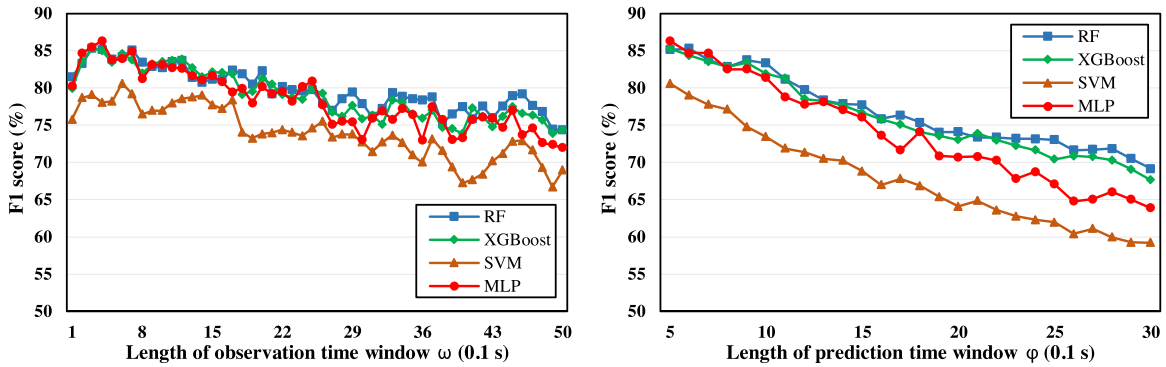

Fig. 10. F1 score of prediction models with different observation/prediction time window lengths.

Table 3
Optimal observation time window length and prediction model under different prediction time window lengths.

φ (0.1 s)	5	6	7	8	9	10	11	12	13	14	15	16	17
ω (0.1 s)	4	3	5	4	3	3	2	5	4	7	6	5	7
Prediction model	MLP	RF	MLP	MLP	RF	RF	XGB	RF	MLP	RF	RF	RF	RF
Accuracy	88.8	88.1	89.2	88.4	87.6	87.2	86.6	86.1	86.4	85.4	85.0	84.7	84.1
F1 score	86.3	85.3	83.7	82.5	83.8	83.3	81.3	79.8	78.1	77.8	77.7	75.9	76.3
φ (0.1 s)	18	19	20	21	22	23	24	25	26	27	28	29	30
ω (0.1 s)	7	6	5	3	9	2	4	3	2	4	3	3	1
Prediction model	RF	RF	RF	XGB	RF	RF	RF	RF	XGB	RF	RF	RF	RF
Accuracy	83.5	83.4	83.4	82.8	82.5	82.1	82.0	82.0	81.8	80.5	80.4	80.7	80.0
F1 score	75.4	74.0	74.1	73.9	73.4	73.2	73.1	73.0	70.9	71.7	71.8	70.5	69.2

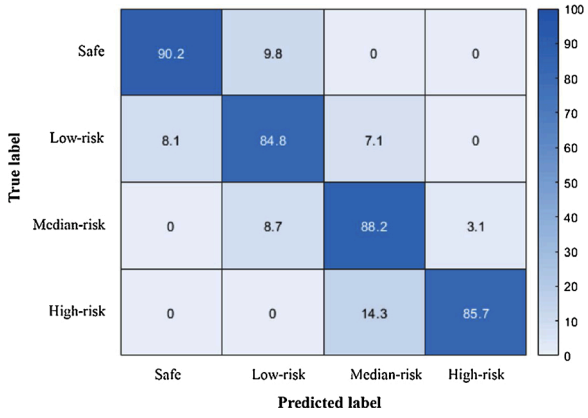

Fig. 11. Confusion matrix of driving risk status prediction results with selected parameters (%).

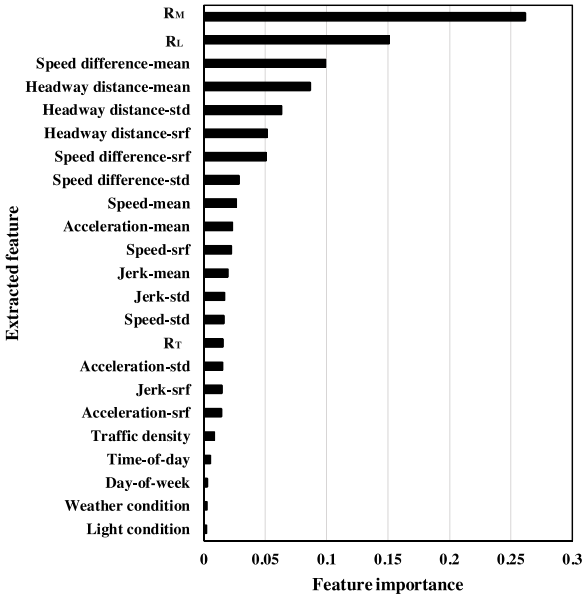

Fig. 12. Feature importance.

5.2.2. Analysis of influencing factors based on selected parameters and RF-RFE

The severity of driving risk is related to various traffic factors, including behavioural characteristics, vehicle characteristics and driving environment characteristics (Wang et al., 2020a). Therefore, it is meaningful to explore the factors that have a significant impact on driving risk. Based on the above selected time window parameters (ω = 0.5 s, φ = 0.7 s), a total 399,842 observation-prediction samples were obtained. The feature importance of 23 extracted features are presented in Fig. 12 based on RF model. Fig. 13 shows the cross-validation score of each iteration in the feature selection process. The results of RF-RFE approach show that the optimal combination has 9 features, including R_M, R_L, speed difference-mean, headway distance-mean, headway distance-std, headway distance-srf, speed difference-srf, speed-mean and acceleration-mean. As can been see from Fig. 12, the importance scores of the 9 filtered features are all the top 10 features.

Obviously, the two driving risk features of R_M and R_L dominate the driving risk status prediction. It is not surprising since the future driving risk status is based on the former driving risk status. As a supplement, the features related to speed difference and headway distance also significantly affect driving risk status prediction. In addition, the mean of speed and the mean of acceleration are still important in predicting driving risk status. Similarly, the relevant conclusions about the influencing factors on driving risk obtained in this study have also been mentioned in previous studies. For instance, Son et al. (2011) proposed a new SMoS, namely unsafe following condition (UFC), and applied it as

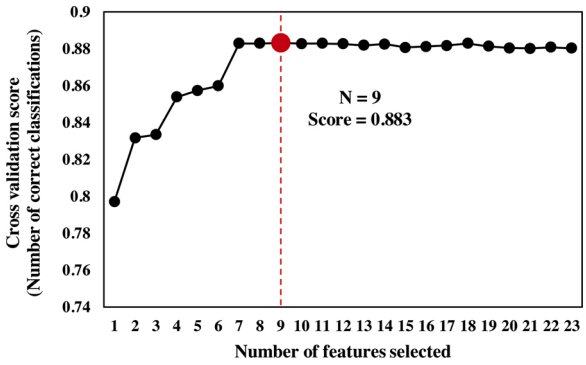

Fig. 13. Feature combination selection by RFE.

the input variable to predict the driving risk. The results showed that the UFC can effectively predict the occurrence of traffic crashes. Furthermore, some previous studies have also found that the speed differences are significantly related to the traffic crash risk, and have a positive impact on the possibility and severity of traffic crash (Cai et al., 2020). Besides, several features about velocity, acceleration and headway distance are selected as the key features used in driving assessment and risk prediction, which is consistent with Shi et al. (2019).

6. Conclusions

To sum up, this paper proposed an integrated methodology to assess and predict real-time driving risk status. In the methodology, the FCM clustering algorithm was applied to identify driving risk status. Different features were then extracted from driving data using rolling time window approach. SMOTE method was used to balance the samples under different driving risk statuses. Then, four modeling methods including RF, XGBoost, SVM and MLP were introduced and applied to predict the real-time driving risk status. Finally, the RF-RFE approach was conducted to analyze driving risk influencing factors. The methodology was tested based on 1,440 car-following events data extracted from SH-NDS database. Some key conclusions can be drawn:

- To assess and predict real-time driving risk, a new SMoS named RCRI was constructed. Then, from the results of FCM unsupervised clustering algorithm, four driving risk statuses, including safe, low-risk, median-risk and high-risk, were obtained to assess and predict real-time driving risk.
- To mine effective information from original driving behavior data, and extract effective features for driving risk prediction and obtain optimal observation-prediction time window length, the rolling time window approach was applied, which produce flexible and in-depth measures on driving behavior and feature extraction.
- Results from real-time driving risk status prediction model comparison and optimal observation and prediction time window length selection show that it is crucial to select the appropriate observation-prediction time window length. The unreasonable length of the time window will cause the effective information of the input variables to be lost, thereby affecting the accuracy of the driving risk status prediction model.
- The features related to the former driving risk status, speed difference, headway distance, speed and acceleration significantly affect driving risk. This illustrates the importance of surrogate measures of safety as input variables and applied to driving risk status prediction models. In addition, maintaining a stable driving speed and safe headway distance can help reduce crash risks.

Early warning of high-risk conditions can more effectively reduce the occurrence of traffic accidents (Lee and Yeo, 2016; Cai et al., 2020). The main contribution of this study is proposed an integrated methodology that can be used to quantify driving risks, extract driving behavior features, obtain the optimal observation-prediction time window length, identify key influencing factors, and accurately predict real-time driving risk based on machine learning algorithms. This methodology could better determine key features and optimal time window length to develop a more accurate and efficient real-time driving risk status prediction system. The RF-RFE model identified the key factors for real-time driving risk status prediction, including former driving risk status, speed difference, headway distance, speed, and acceleration. These results can guide traffic managers to implement effective management strategies such as safe headway distance maintenance, variable speed control and dynamic message release.

However, this study still has some limitations. The driving risk prediction method adopted in this paper only focuses on the car-following process, and it is not enough to explore the driving risk during lane-changing or overtaking process. For future work, high-risk lane-changing events and overtaking events will be collected through NDS or actual vehicle test to further improve and validate the accuracy of the proposed driving risk prediction model. In addition, some deep learning algorithms, such as recurrent neural network, can be applied and compared with the prediction models proposed in this research. Meanwhile, other driving risk influencing factors including vehicle characteristics and road geometry characteristics can be obtained and added to the input variables to further improve the performance of the prediction model. For practical applications, the model will be further applied in the smart vehicle industry fed with real-time naturalistic driving data collected by, for example, ADAS.

Author statement

Qiangqiang Shangguan: Conceptualization, Methodology, Data curation, Investigation, Software, Writing - original draft. **Ting Fu:** Investigation, Software, Writing - review & editing, Funding acquisition. **Junhua Wang:** Conceptualization, Data curation, Writing - review & editing, Funding acquisition. **Tianyang Luo:** Data curation, Software, Writing - review & editing. **Shou'en Fang:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

This study was jointly supported by the Chinese National Natural Science Foundation (71871161), the National Key R&D Program of China (2019YFB1600703), and the Shanghai Sailing Program (20YF1451800).

References

Ahmed, M.M., Abdel-Aty, M., 2013. A data fusion framework for real-time risk assessment on freeways. Transp. Res. Part C Emerg. Technol. 26, 203–213.
Ahmed, M.M., Abdel-Aty, M., Yu, R., 2012. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transp. Res. Rec.: J. Transp. Res. Board 2280 (1), 60–67.
Arbabzadeh, N., Jafari, M., 2017. A data-driven approach for driving safety risk prediction using driver behavior and roadway information data. Ieee Trans. Intell. Transp. Syst. 19 (2), 446–460.
Bagdadi, O., Várhelyi, A., 2011. Jerky driving—an indicator of accident proneness? Accid. Anal. Prev. 43 (4), 1359–1363.
Bärgman, J., Lisovskaja, V., Victor, T., Flannagan, C., Dozza, M., 2015. How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from shrp2. Transp. Res. Part F: Traffic Psychol. Behav. 35, 152–169.
Bıçaksız, P., Özkan, T., 2016. Impulsivity and driver behaviors, offences and accident involvement: a systematic review. Transp. Res. Part F: Traffic Psychol. Behav. 38, 194–223.
Board, N.T.S, 2001. Special Investigation Report-Highway Vehicle and Infrastructure-Based Technology for the Prevention of Rear-End Collisions. NTSB Number SIR-01.

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., Wu, Y., 2020. Real-time crash prediction on expressways using deep generative models. Transp. Res. Part C: Emerg. Technol. 117, 102697.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794.
Chen, J., Wu, Z., Zhang, J., 2019. Driving safety risk prediction using cost-sensitive with nonnegativity-constrained autoencoders based on imbalanced naturalistic driving data. IEEE Trans. Intell. Transp. Syst. 20 (12), 4450–4465.
Costela, F.M., Castro-Torres, J.J., 2020. Risk prediction model using eye movements during simulated driving with logistic regressions and neural networks. Transp. Res. Part F Traffic Psychol. Behav. 74, 511–521.
Cunto, F., Saccomanno, F.F., 2008. Calibration and validation of simulated vehicle safety performance at signalized intersections. Accid. Anal. Prev. 40 (3), 1171–1179.
Feng, F., Bao, S., Sayer, J.R., Flannagan, C., Manser, M., Wunderlich, R., 2017. Can vehicle longitudinal jerk be used to identify aggressive drivers? An examination using naturalistic driving data. Accid. Anal. Prev. 104, 125–136.
Fu, T., Miranda-Moreno, L., Saunier, N., 2018. A novel framework to evaluate pedestrian safety at non-signalized locations. Accid. Anal. Prev. 111, 23–33.
Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3–9.
Hassan, H.M., Abdel-Aty, M.A., 2013. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. J. Saf. Res. 45, 29–36.
Hu, J., Huang, M.-C., Yu, X., 2020. Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. Accid. Anal. Prev. 144, 105665.
Johnson, R.A., Wichern, D.W., 2002. Applied Multivariate Statistical Analysis. Prentice hall, Upper Saddle River, NJ.
Johnsson, C., Laureshyn, A., De Ceunynck, T., 2018. In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators. Transp. Rev. 38 (6), 765–785.
Katrakazas, C., Quddus, M., Chen, W.H., 2019. A simulation study of predicting real-time conflict-prone traffic conditions. Ieee Trans. Intell. Transp. Syst. 19 (10), 3196–3207.
Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., 2006. The Impact of Driver Inattention on Near-Crash/Crash Risk: an Analysis using the 100-Car Naturalistic Driving Study Data. National Highway Traffic Safety Administration, United States.
Kuang, Y., Qu, X., Wang, S., 2015. A tree-structured crash surrogate measure for freeways. Accid. Anal. Prev. 77, 137–148.
Lee, D., Yeo, H., 2016. Real-time rear-end collision-warning system using a multilayer perceptron neural network. Ieee Trans. Intell. Transp. Syst. 17 (11), 3087–3097.
Lin, L., Wang, Q., Sadek, A.W., 2015. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transp. Res. Part C: Emerg. Technol. 55, 444–459.
Oh, C., Park, S., Ritchie, S.G., 2006. A method for identifying rear-end collision risks using inductive loop detectors. Accid. Anal. Prev. 38 (2), 295–301.
Panagopoulos, G., Pavlidis, I., 2019. Forecasting markers of habitual driving behaviors associated with crash risk. IEEE Trans. Intell. Transp. Syst. 21 (2), 841–851.
Raju, N., Kumar, P., Arkatkar, S., Joshi, G., 2019. Determining risk-based safety thresholds through naturalistic driving patterns using trajectory data on expressways. Saf. Sci. 119, 117–125.
Scott-Parker, B., Weston, L., 2017. Sensitivity to reward and risky driving, risky decision making, and risky health behaviour: a literature review. Transp. Res. Part F: Traffic Psychol. Behav. 49, 93–109.
Shangguan, Q., Fu, T., Liu, S., 2020. Investigating rear-end collision avoidance behavior under varied foggy weather conditions: a study using advanced driving simulator and survival analysis. Accid. Anal. Prev. 139, 105499.
Shangguan, Q., Fu, T., Jiang, R., Fang, S.E., 2021. Use of Naturalistic Driving Data to Quantify Influencing Factors of Driving Risk Based on a New Surrogate Measure of Safety. Transportation Research Board, Washington DC.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C: Emerg. Technol. 58, 380–394.
Shi, X., Wong, Y.D., Li, M.Z.-F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on xgboost for driving assessment and risk prediction. Accid. Anal. Prev. 129, 170–179.
Son, H., Kweon, Y.-J., Park, B.B., 2011. Development of crash prediction models with individual vehicular data. Transp. Res. Part C: Emerg. Technol. 19 (6), 1353–1363.
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and qsar modeling. J. Chem. Inf. Comput. Sci. 43 (6), 1947–1958.
Wang, D., 2002. Traffic Flow Theory.
Wang, J., Kong, Y., Fu, T., 2019a. Expressway crash risk prediction using back propagation neural network: a brief investigation on safety resilience. Accid. Anal. Prev. 124, 180–192.
Wang, X., Yang, M., Hurwitz, D., 2019b. Analysis of cut-in behavior based on naturalistic driving data. Accid. Anal. Prev. 124, 127–137.
Wang, J., Huang, H., Li, Y., Zhou, H., Liu, J., Xu, Q., 2020a. Driving risk assessment based on naturalistic driving study and driver attitude questionnaire analysis. Accid. Anal. Prev. 145, 105680.
Wang, K., Xue, Q., Xing, Y., Li, C., 2020b. Improve aggressive driver recognition using collision surrogate measurement and imbalanced class boosting. Int. J. Environ. Res. Public Health 17 (7), 2375.
Wu, Y., Abdel-Aty, M., Park, J., Zhu, J., 2018. Effects of crash warning systems on rear-end crash avoidance behavior under fog conditions. Transp. Res. Part C: Emerg. Technol. 95, 481–492.
Xiong, X., Chen, L., Liang, J., 2018. Vehicle driving risk prediction based on markov chain model. Discrete Dyn. Nat. Soc. 2018.
You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. J. Mod. Transp. 25 (2), 116–123.
Yu, R., Abdel-Aty, M., 2013a. Investigating the different characteristics of weekday and weekend crashes. J. Saf. Res. 46, 91–97.
Yu, R., Abdel-Aty, M., 2013b. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
Zhu, M., Wang, X., Tarko, A., 2018. Modeling car-following behavior on urban expressways in shanghai: a naturalistic driving study. Transp. Res. Part C: Emerg. Technol. 93, 425–445.
Zvarevashe, K., Olugbara, O.O., 2018. Gender voice recognition using random forest recursive feature elimination with gradient boosting machines. Proceedings of the 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD) 1–6.