Transportation Research Part C 116 (2020) 102646

Contents lists available at ScienceDirect

Transportation Research Part C

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

Predicting lane-changing risk level based on vehicles' space-series features: A pre-emptive learning approach

Tianyi Chen[a,*], Xiupeng Shi[a], Yiik Diew Wong[a], Xiaocong Yu[b]

[a] School of Civil and Environmental Engineering, Nanyang Technological University, 639798, Singapore
[b] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801-2918, USA

ARTICLE INFO

Keywords:
Pre-emptive lane-changing risk prediction
Resampling method
Feature selection
Machine learning

ABSTRACT

Vehicles' risky lane-changing (LC) maneuver has significant impact on road traffic safety. As an innovation compared with the posterior LC risk prediction methods proposed in previous studies, this study develops a pre-emptive LC risk level prediction (P-LRLP) method, which is able to estimate the crash risk level of an LC event in advance before the LC car completes the LC maneuver. The basic concept of this method is to apply a machine learning classifier to predict the LC risk level based on cars' key space-series features at the beginning of the LC event. To boost the prediction performance, an innovative resampling method, namely ENN-SMOTE-Tomek Link (EST), and an advanced machine learning classifier, namely LightGBM, are proposed and employed in the development of the P-LRLP method. Meanwhile, an algorithm which can measure the stability of the selected key features in terms of the randomness and size of training samples is developed to evaluate the feature selection methods. A digitalized vehicles' trajectory dataset, the Next Generation Simulation (NGSIM) is used for method validation. The validation results manifest that the EST can achieve satisfactory resampling performance while Random Forest (RF), as an embedded FS method, achieves remarkable performance on both stability of selected features and prediction of risk level. The results also show that the LC risk level can be most accurately predicted when the LC car moves to the position where the distance between the longitudinal center line of the LC car and the marking line separating the two lanes equals 1.5ft. As an innovative LC risk level prediction technique, the P-LRLP method could be integrated with advanced driver-assistance system (ADAS) and vehicle-to-vehicle (V2V) communication to remedy potential risky LC maneuver in the future.

1. Introduction

1.1. Background

Lane-changing (LC) is a common maneuver in on-road driving. An LC maneuver is executed whenever a vehicle changes lane(s) across multi-lane traffic stream along the road journey. Herein, LC and lane-merging (LM) crashes account for about 5% of reported road traffic accidents and 7% of crash fatalities (Hou et al., 2015). Risky LC maneuvers are associated with crashes that result in significant loss of lives and property damage, which should be remediated as much as possible. Machine learning classification is a

* Corresponding author.
E-mail addresses: TIANYI002@e.ntu.edu.sg (T. Chen), XSHI004@e.ntu.edu.sg (X. Shi), CYDWONG@ntu.edu.sg (Y.D. Wong), xy21@illinois.edu (X. Yu).

https://doi.org/10.1016/j.trc.2020.102646
Received 1 December 2019; Received in revised form 1 March 2020; Accepted 2 April 2020
Available online 24 May 2020
0968-090X/

mathematical technique which can effectively make prediction and identify patterns based on data used for training (Bishop, 2006). Machine learning classifiers have been widely used in the prediction and analysis of road traffic accident risk, for example, the prediction of crash accident (Shi and Abdel-Aty, 2015, Yang et al., 2018b), the prediction of accident severity (Iranitalab and Khattak, 2017, Jeong et al., 2018), and the analysis on risk-contributing factors (Shi et al., 2019, Wang et al., 2019).

1.2. Related work and research gaps

1.2.1. Lane-changing risk prediction

Many researchers have focused on the prediction on LC intention (i.e. whether to change lanes or not) and LC trajectory planning, which involves drivers' intention, vehicles' movement, and the states of surrounding vehicles. Sun and Elefteriadou (2010; 2012; 2014) investigated the characteristics of driver's behavior during an LC event from a microscopic perspective. Hou et al. (2013) modeled mandatory LC using Bayes classifier and decision tree to predict vehicles' decision when to change lane. Suh et al. (2018) designed an LC trajectory planning model with probabilistic and deterministic prediction techniques based on driving environment for automated driving. Yang et al. (2018a) proposed a dynamic LC trajectory planning model for automated vehicles based on the states of the surrounding vehicles. Zhang et al. (2019) used a deep learning model to model vehicles' car-following (CF) and LC behaviors and predict the behaviors based on vehicles' time-series features. Xie et al. (2019) proposed a data-driven model to predict the LC process composed of LC decision and LC implementation based on deep learning.

Nevertheless, few studies have attempted to estimate or predict LC risk. Pande and Abdel-Aty (2006) proposed a model to predict the occurrence of LC related freeway crashes based on the traffic surveillance data. They explored traffic parameters surrounding the crash location as the factors which possibly contributed to the occurrence of LC crash accident. Classification tree was used to identify key factors which were significantly associated with the binary target (i.e. crash vs. non-crash), based on which the occurrence of the LC related crashes can be predicted. Chen et al. (2019) proposed an LC risk level prediction method based on feature learning. They employed the statistical descriptions of vehicles' motion behaviors during an LC process as the features, based on which Random Forest (RF) classifier was trained to select the key features and predict the LC risk level.

Although the above-mentioned studies have provided valuable insights into LC safety, the following research gaps still remain to be resolved:

1. Many researchers have taken LC safety into consideration when planning LC trajectory or predicting LC intention. However, most studies merely considered the surrounding vehicles in the adjacent lane during an LC process. Additionally, few studies focused on the risk control during an LC process.
2. The few studies (Pande and Abdel-Aty, 2006; Chen et al., 2019) that have focused on LC risk prediction were conducted based on the features collected from upstream and downstream of an LC event. Consequently, those methods can merely conduct posterior prediction, which is inefficient to detect potential LC crashes or risky LC maneuvers in advance.

1.2.2. Feature engineering

Several researchers have studied the safety of LC maneuver and explored the impact of several factors. Lee et al. (2011) investigated the impact of traffic flow parameters on drivers' CF and LC behaviors and found that the parameters had different effects on CF and LC crash accidents. Yun et al. (2017) discussed the impact of in-vehicle navigation information on LC behavior in urban expressway diverge segments. Wang et al. (2018) studied the contribution of driver's perception characteristics to LC safety and accordingly developed an LC warning system. Yang et al. (2019) used naturalistic driving data to examine the contribution of gap acceptance and duration to the impact on the following vehicle in an LC event. Ali et al. (2019) examined the impact of connected vehicle environment (CVE) on safety during mandatory LC and found that CVE can increase the safety and efficiency of mandatory LC. Arbis and Dixit (2019) proposed a game theoretic model for LC and found that longer acceleration lanes and reduction of speed limits on on-ramps can reduce the risk of mandatory LC.

There are many features (or factors) that have potential impact on LC safety. Hence, identifying influential features and discarding noisy features is of significance to the investigation of LC safety. Feature selection (FS) is an essential data processing technique able to remove noisy, irrelevant, and redundant data from dataset (Liu and Yu, 2005). FS is proposed to reduce computational cost, improve prediction performance of the predictors, and better explain the underlying causation process of how features contribute to the targets (Guyon and Elisseeff, 2003). The evaluation of FS methods has been conducted in many studies to compare the performance of candidate FS methods. Freeman et al. (2015) tested 16 filter-based FS methods by evaluating the predictive performance with specific classifiers. Bermingham et al. (2015) investigated FS methods based on prediction accuracy and computational cost in a study of genomic prediction. Rodriguez-Galiano et al. (2018) evaluated filter, wrapper and embedded methods according to prediction accuracy in a groundwater study. However, most studies have two major limitations:

1. The evaluation places more attention on the predictive performance achieved with the selected features and ignores the stability of the selected features.
2. The size of selected features is roughly determined without much rigor.

1.3. Method development and organization

Considering the research gaps as above mentioned, we propose a pre-emptive LC risk level prediction (P-LRLP) method in this

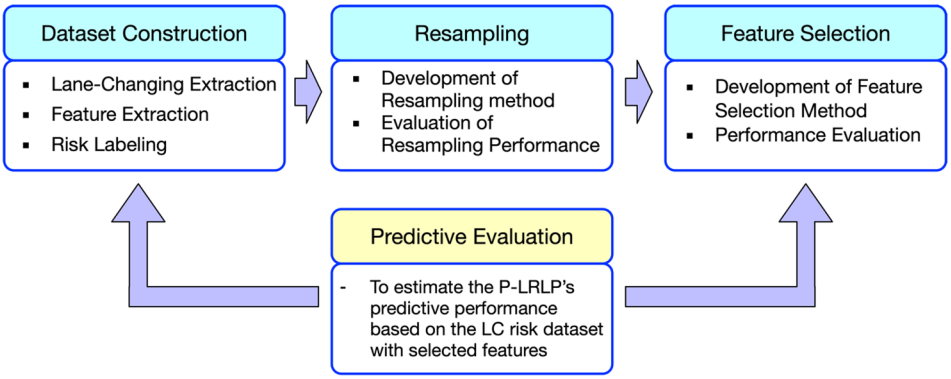


Fig. 1. Development of lane-changing risk level prediction (P-LRLP) method.

study. The method applies the machine learning classifier, which is trained on the LC risk dataset with key space-series features collected at the beginning of an LC event, to predict the risk level of an LC maneuver. Fig. 1 shows the procedure of method development, which includes the steps of dataset construction, resampling, feature selection and predictive evaluation. As innovations, we further propose a resampling method and an algorithm to measure features' stability, which are applied in the steps of resampling and feature selection, respectively. The result of predictive evaluation plays an important role in the steps of dataset construction and feature selection. The followings shall be determined in the step of predictive evaluation based on the vehicles' trajectory dataset: (a) feature selection method, (b) the number of key features, (c) the label of LC risk dataset (i.e. the number of risk levels), and (d) how much in advance the LC risk level can be predicted.

This paper is organized as follows. Section 2 introduces the step of dataset construction. Section 3 introduces the step of resampling. Section 4 introduces the steps of feature selection and predictive evaluation jointly, which aims to clearly illustrate the purpose of predictive evaluation. Section 5 uses a practical vehicles' trajectory dataset to verify the P-LRLP method and discusses the findings. Sections 6 covers the conclusions, limitations and future work.

2. Dataset construction

As shown in Fig. 2, dataset construction aims to construct LC risk dataset from vehicles' trajectory dataset. LC risk dataset comprises vehicles' candidate features (i.e. attributes) and LC risk labels (i.e. target) of each sample as an LC event. The construction of LC risk dataset is comprised of three phases, namely, LC extraction, feature extraction and risk labeling. LC extraction aims to extract LC trajectories from vehicles' trajectory dataset, which is introduced in Section 2.1. Feature extraction is proposed to extract both space-stamp features and space-interval features during an LC event, which are illustrated in Section 2.2. Risk labeling contains two steps, namely, risk quantification and risk clustering, which are explained in Section 2.3.

2.1. Lane-changing extraction

LC extraction is proposed to assemble the LC events extracted from vehicles' trajectory dataset as an LC trajectory dataset. In this study, we consider an LC event as a collocation that involves five cars. Fig. 3 illustrates right-to-left and left-to-right LC scenarios, where *Sub* is the LC car, *Pre1* and *Pre2* are respectively the preceding cars in the original lane and target lane, and *Fol1* and *Fol2* are respectively the following counterparts in those two lanes. As shown in Fig. 3, to reduce the computational expense of feature extraction, the X, Y coordinates are differently defined according to *Sub*'s LC direction.

Cars' motion behaviors (e.g. location, velocity and acceleration) are decomposed into X, Y directions in preparation for the follow-on phases. Each car is assumed to keep its longitudinal center line parallel to the lane markers during the LC to simplify the calculation of the distance between two cars. Also, cars' acceleration is supposed to be in same or opposite direction with velocity. As an example, Fig. 4 shows car's motion decomposition. The car C_1's velocity V_1 is decomposed into V_{1x} and V_{1y}, in X and Y directions, respectively. G_x indicates the lateral gap between two cars, C_1 and C_2, in X direction, while G_y indicates the longitudinal gap in Y direction. W_* and L_*describe car's width and length, respectively.

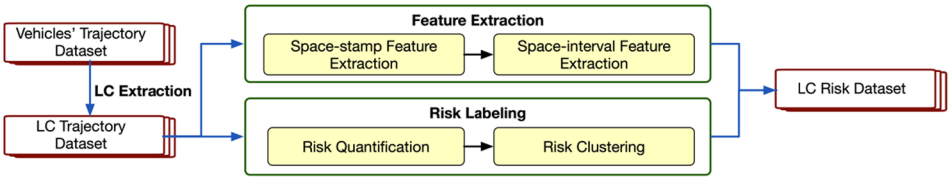
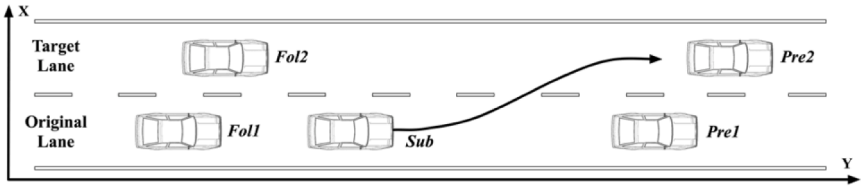
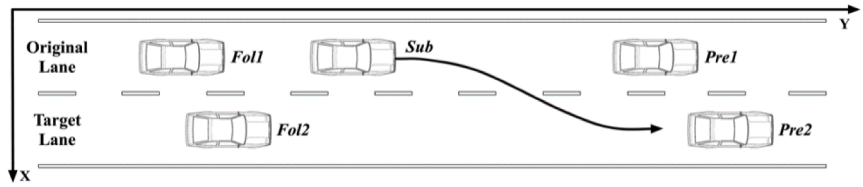


Fig. 2. General procedure of dataset construction.



(a) Right-to-left lane-changing



(b) Left-to-right lane-changing

Fig. 3. Illustration of lane-changing scenarios.

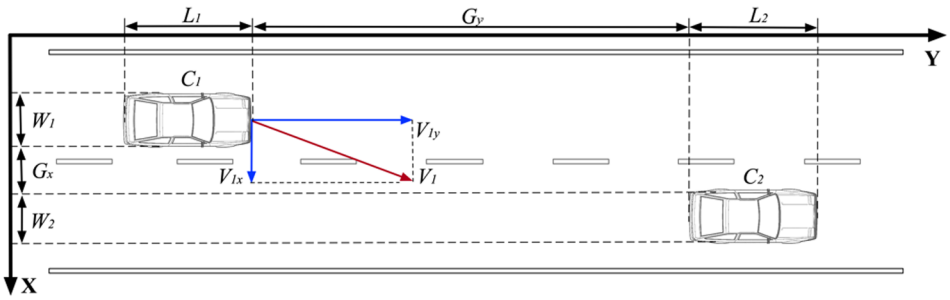
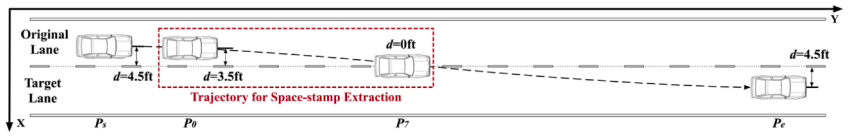


Fig. 4. Illustration of car's motion decomposition.

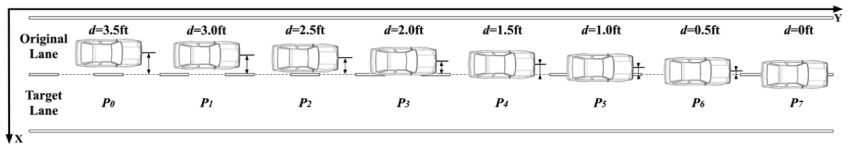
2.2. Feature extraction

The P-LRLP method is proposed to predict LC risk level mainly based on the motion behaviors of the five cars as *Sub* moves in the original lane during an LC event, which can reflect the car's LC intention. As shown in Fig. 5(a), we define that an LC event starts and ends at the positions P_s and P_e where the distance, d, between *Sub*'s longitudinal center line and the dashed marking line, which separates the original lane and target lane, equals 4.5 ft. Cars' motion behavioral features (of the five cars) are collected as *Sub* moves along the trajectory from P_0, where d equals 3.5 ft, to P_7, where d equals 0 ft. On one hand, less information of car's LC intention can be obtained if d is larger than 3.5 ft. On the other hand, the remediation actions (e.g. stop changing lane and keep following car in the original lane) in the case of high-risk LC event might not be effective when the car's longitudinal canter line has already moved to the target lane.

Fig. 5(b) illustrates the seven positions (i.e. space-stamps), the d of which decreases from 3.5 ft to 0 ft with decrement of 0.5 ft, on the trajectory for feature extraction. Candidate features contain cars' basic information (i.e. car length, car width, *Sub*'s LC direction and *Sub*'s LC duration) and cars' behavioral features. Cars' motion behavioral features can be categorized into space-stamp features



(a) *Sub*'s lane-changing trajectory



(b) Space-stamps for feature extraction

Fig. 5. Trajectory for space-stamp extraction.

Table 1
Categorizations and descriptions of candidate features.

Categories		Candidate features	Number of features
Basic information	Individual features	Car length	5
		Car width	5
		Sub's LC direction	1
		Sub's LC duration	1
Space-stampfeatures	Individual features	Cars' X coordinate	5
		Cars' Y coordinate	5
		Cars' velocity in X direction	5
		Cars' velocity in Y direction	5
		Cars' acceleration in X direction	5
		Cars' acceleration in X direction	5
	Interaction features	Gap distance between *Sub* and its surrounding cars in X direction	4
		Gap distance between *Sub* and its surrounding cars in Y direction	4
		Velocity difference between *Sub* and its surrounding cars in X direction	4
		Velocity difference between *Sub* and its surrounding cars in Y direction	4
		Acceleration difference between *Sub* and its surrounding cars in X direction	4
		Acceleration difference between *Sub* and its surrounding cars in Y direction	4
Space-interval features	Individual features	Cars' X coordinate*	50
		Cars' Y coordinate*	50
		Cars' velocity in X direction*	50
		Cars' velocity in Y direction*	50
		Cars' acceleration in X direction*	50
		Cars' acceleration in X direction*	50
	Interaction features	Gap distance between *Sub* and its surrounding cars in X direction*	40
		Gap distance between *Sub* and its surrounding cars in Y direction*	40
		Velocity difference between *Sub* and its surrounding cars in X direction*	40
		Velocity difference between *Sub* and its surrounding cars in Y direction*	40
		Acceleration difference between *Sub* and its surrounding cars in X direction*	40
		Acceleration difference between *Sub* and its surrounding cars in Y direction*	40

*Indicates the features referring to the statistical descriptions, such as mean, standard deviation, range, maximum, minimum, 0.25 quantile, 0.5 quantile, 0.75 quantile, kurtosis and skewness, of a motion behavior.
For each space-stamp, a total of 54 space-stamp features are extracted. For each space-interval, a total of 540 space-interval features are extracted.

and space-interval features. Space-stamp features refer to the features of cars' motion behaviors at the space-stamps shown in Fig. 5(b). Space-interval features refer to the statistical descriptions of cars' motion behaviors in the space-interval between the start position P_s and a space-stamp P_y. Table 1 summarizes the candidate features.

One purpose of this study is to determine which position (i.e. space-stamps) *Sub* moves to such that the P-LRLP method can most accurately predict the LC risk level. The measurements of prediction performance are discussed in detail in latter Section 4.3.2. Consequently, seven LC risk datasets with different spatial conditions are constructed in preparation for validation and evaluation. Table 2 summarizes the space-stamps, space-interval, $d_{min.}$ and number of features involved in the feature extraction for each dataset. Herein, we take the No.4 dataset as an example to calculate the number of features. Since the No.4 dataset involves 12 features of basic information, five space-stamps with 54 features at each stamp, and 540 space-interval features, the number of features can be obtained as $12 + 5 \times 54 + 540 = 822$.

2.3. Risk labeling

Risk labeling is proposed to determine the risk level (i.e. class) of each sample as an LC event. As illustrated in Fig. 2, the labeling

Table 2
Candidate feature extraction.

LC risk dataset No.	Feature extraction			
	Space-stamps	Space-interval	$d_{min.}$*	Number of features
1	P_0, P_1	$[P_s, P_1]$	3.0 ft	660
2	P_0, P_1, P_2	$[P_s, P_2]$	2.5 ft	714
3	P_0, P_1, P_2, P_3	$[P_s, P_3]$	2.0 ft	768
4	P_0, $\cdots$, P_4	$[P_s, P_4]$	1.5 ft	822
5	P_0, $\cdots$, P_5	$[P_s, P_5]$	1.0 ft	876
6	P_0, $\cdots$, P_6	$[P_s, P_6]$	0.5 ft	930
7	P_0, $\cdots$, P_7	$[P_s, P_7]$	0 ft	984

*Refers to the distance d of the last involved space-stamp. For example, the $d_{min.}$ of the No.4 dataset indicates the distance d of the space-stamp P_4, which equals 1.5 ft.

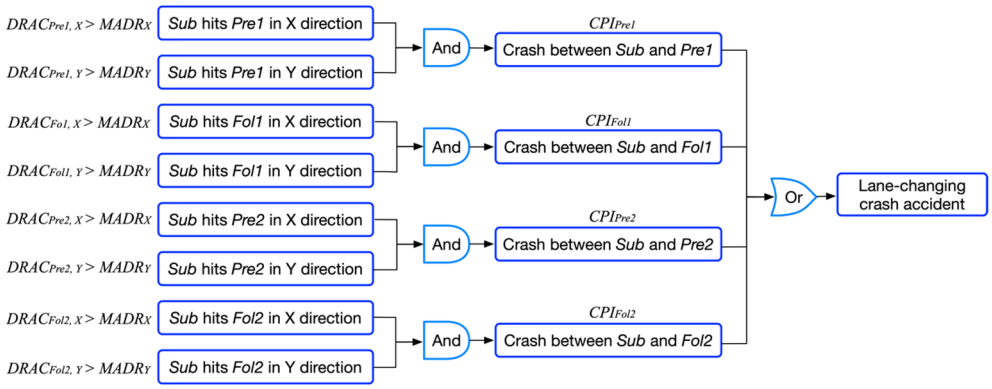


Fig. 6. Illustration of fault tree analysis.

process contains two phases, namely risk quantification and risk clustering. The first phase aims to quantify the risk of each LC event by integrating fault tree analysis method with surrogate measures. The second phase applies unsupervised learning method to cluster the quantified risk into several risk levels, based on which the LC events are classified.

Fault tree analysis method is a useful analytic technique to estimate the probability of a system failure based on the propagation of the root causes. In this study, the LC event that involves a cluster of five cars, as shown in Fig. 3, is treated as a system. The system failure, namely the LC crash accident, occurs when the LC car crashes into any one of its surrounding cars. Fig. 6 illustrates the fault tree of an LC event. The surrogate measures such as Deceleration Rate to Avoid a Crash (DRAC) and Crash Potential Index (CPI) are involved in the fault tree analysis to quantify the LC risk.

DRAC refers to the minimum deceleration rate required by the following vehicle to match the velocity of the preceding vehicle to avoid a crash accident (Cooper and Ferguson, 1976). In this study, DRAC is decomposed into X and Y directions. The DRAC between *Sub* and one of its surrounding cars at timestamp t can be obtained as:

$$DRAC_{C^*,D^*}(t) = \begin{cases} \frac{(V_{Sub,D^*}(t) - V_{C^*,D^*}(t))^2}{G_{D^*}(t)}, \ if \quad V_{Sub,D^*}(t) > V_{C^*,D^*}(t) \\ 0, \ otherwise \end{cases} \tag{1}$$

where C^* denotes one of *Sub*'s surrounding cars, namely *Pre1*, *Fol1*, *Pre2* or *Fol2*, and D^* indicates the direction, namely X or Y direction. $V_*(t)$ and $G_*(t)$ indicates car's velocity and the gap between two cars at timestamp t, which are illustrated in Fig. 4. Maximum Available Deceleration Rate (MADR) refers to the maximum threshold of DRAC (Cooper and Ferguson, 1976). The following vehicle is deemed to 'hit' the preceding vehicle if the DRAC exceeds MADR. In this study, MADR is defined as $1.4m/s^2$, as recommended in our previous study (Chen et al., 2019). The probability that *Sub* 'hits' one of its surrounding cars in D^* direction at timestamp t can be obtained as:

$$P_{C^*,D^*}(t) = \begin{cases} 1, \ if DRAC_{C^*,D^*}(t) > MADR_{D^*}(t) \\ 0, \ otherwise \end{cases} \tag{2}$$

Crash Potential Index (CPI) refers to the probability that the DRAC exceeds MADR for every Δt (Δt is extremely short) of the observation time (Cunto and Saccomanno, 2008). The crash between *Sub* and one of its surrounding cars is deemed to occur if *Sub* 'hits' the surrounding car in X and Y directions simultaneously at timestamp t, i.e., $P_{C^*,X}(t)=P_{C^*,Y}(t) = 1$. CPI can be seen as the general probability that the crash accident occurs between *Sub* and one of its surrounding cars in the observation time, T (i.e., the duration of an LC event), which can be obtained as:

$$CPI_{C^*} = \frac{\sum_t P_{C^*}(t) \bullet \Delta t}{T} \tag{3}$$

where $P_{C^*}(t)$ refers to the probability of the crash between *Sub* and one of its surrounding cars at timestamp t, which can be obtained as:

$$P_{C^*}(t) = \begin{cases} 1, \ if P_{C^*,X}(t)=P_{C^*,Y}(t) = 1 \\ 0, \ otherwise \end{cases} \tag{4}$$

Accordingly, the probability of the system failure, namely the occurrence of crash accident for an LC event, can be obtained as (Ruijters and Stoelinga, 2015):

$$P = 1 - \prod_{C^*} (1 - CPI_{C^*}) \tag{5}$$

The probability of the system failure, P, is regarded as the quantified risk of an LC event, based on which k-Means algorithm (Cover and Hart, 1967) is used to classify the samples into several risk levels (Shi et al., 2019; Chen et al., 2019). Determining the

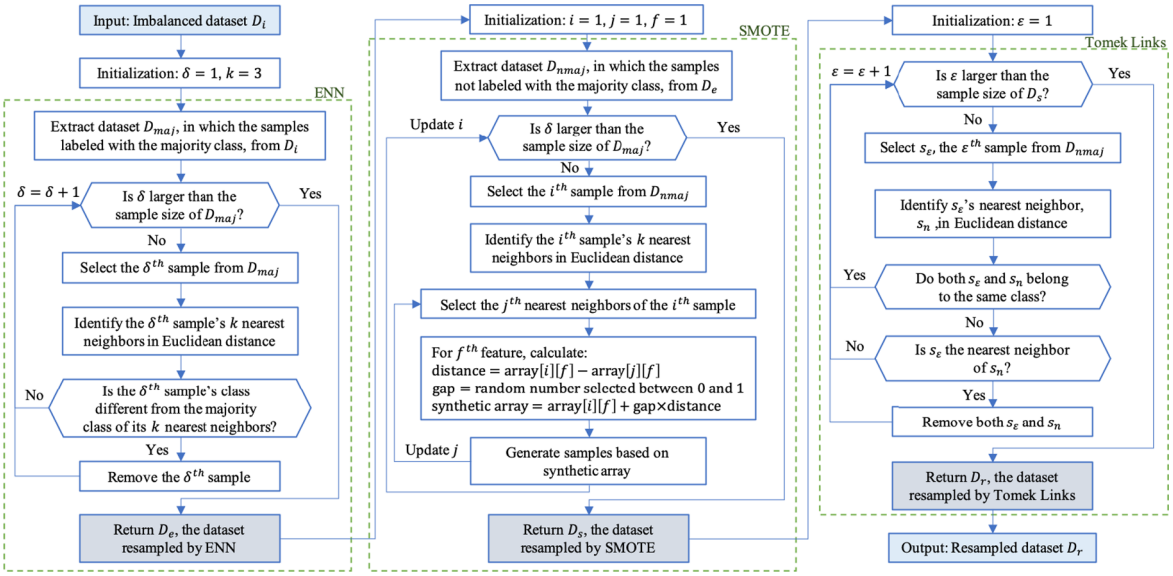


Fig. 7. General flowchart of EST resampling method.

number of risk levels, *n*, is also a purpose of this study. Herein, 3, 4, 5, 6 and 7 are considered as the candidate values of *n*, from which the optimal number of risk levels is selected according to the performance of LC risk level prediction. The details of prediction performance are discussed latter in Section 4.3.2. The risk levels are roughly-defined and less convincing if *n* is smaller than 3. However, the risk levels with *n* larger than 7 would place more challenges on the prediction of LC risk level, which might reduce the prediction accuracy. Therefore, we only consider the above-mentioned five values as the candidate values of *n* in this study.

3. Resampling

3.1. Resampling method

Resampling methods have been widely used to address class imbalance problem (Estabrooks et al., 2004), which occurs when the ratio of the majority class size to the minority class size is high in a dataset. Resampling methods can be categorized into over-sampling method, under-sampling method, and the combination of those two methods. To balance the class distribution of a dataset, over-sampling methods increase the sample size of minority class while under-sampling methods decrease the sample size of majority class in the training set (Drummond and Holte, 2003). LC risk dataset is a multi-class dataset suffering from class imbalance problem. The class distribution of LC risk dataset is skewed since the risky LC events, which refer to the samples of the high-risk classes, occur infrequently in real-life. To obtain better-defined class distribution for LC risk dataset, we propose an innovative resampling method combining Edited Nearest Neighbors (ENN) (Wilson, 1972), Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) and Tomek Link (Tomek, 1976). The flowchart of ENN-SMOTE-Tomek Link (EST) is shown in Fig. 7.

3.2. Performance measurements

The performance of resampling methods is measured from two perspectives. First, measuring the imbalance ratios of the dataset before and after resampling can facilitate judgement as to whether the class distribution of the dataset is properly balanced after resampling. Hence, we propose a measure of imbalance ratio for *n*-class dataset, which can be obtained as:

$$IR = \frac{\sum_{i,j} max(s_i/s_j,\ s_j/s_i)}{n(n-1)} \tag{6}$$

where $max(s_j/s_i,\ s_i/s_j)$ indicates the imbalance ratio between the i and j. s_i and s_j refer to the sample sizes of the classes i and j, respectively. Second, over-sampling can potentially lead to overfitting on the samples of minority classes (Chawla, 2009) and increase computational cost (Liang and Zhang, 2012) in training scenario, while under-sampling might cause the loss of important information by discarding or removing samples (Liang and Zhang, 2012). Hence, to minimize such potential problems caused by resampling, we propose the measures of over-sampling ratio and under-sampling ratio to assess the resampling results, which can be obtained as:

$$OR = \frac{\sum \text{Oversamplingsize}}{\text{Originalsamplesize}} \tag{7}$$

$$UR = \frac{\sum \text{Undersamplingsize}}{\text{Originalsamplesize}} \tag{8}$$

where 'oversampling size' refers to the increased sample size of the over-sampled class, while 'under-sampling size' refers to the decreased sample size of the under-sampled class. 'Original sample size' refers to the sample size of the dataset before being re-sampled.

4. Feature selection

4.1. Feature selection method

The purpose of FS is to select key features from the candidate features based on the weights (i.e. importance or coefficients) or ranks assigned to the features. Feature weight is calculated based on the mutual relationship of features or the relationship between features and the label, which can explain features' contributions to the target results (label) (Razmjoo et al., 2017). In this study, we consider the supervised FS methods, which can be further categorized into filter, wrapper and embedded approaches (Tang et al., 2014).

Filter approach relies on the measures of the statistical characteristics of training data such as correlation, dependency, information, etc. (Kohavi and John, 1997). In this study, we take two typical filter FS methods, namely, Removing Features with Low Variance (RFLV) (He et al., 2006) and Univariate Feature Selection (UFS) (Guyon and Elisseeff, 2003), into consideration. Wrapper approach selects features based on the prediction performance of the predetermined learning algorithm integrated with feature search method (Kohavi and John, 1997). Two typical feature search methods, namely, Sequential Forward Floating Selection (SFFS) (Pudil et al., 1994, Jain and Zongker, 1997) and Recursive Feature Elimination (RFE) (Guyon and Elisseeff, 2003). Additionally, the feature search methods are respectively wrapped with four efficient learning algorithms (i.e. classifiers) to rank features, namely, Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Decision Tree (DT) (Yan et al., 2016; Breiman, 2017), Gradient Boosting Decision Tree (GBDT) (Friedman, 2002), and Random Forest (RF) (Breiman, 2001). Embedded approach incorporates the statistical criterion of filter approach and the ranking criterion of the classifier in wrapper approach to select features (Chandrashekar et al., 2014). The above-mentioned four learning algorithms are employed as the classifiers for being embedded. The FS methods are summarized in Table 3.

4.2. Performance evaluation

Performance evaluation is proposed to evaluate FS methods and select the most suitable FS method for LC risk dataset. Fig. 8 presents the general procedure of performance evaluation, which comprises two phases, stability evaluation and predictive evaluation. The two phases of performance evaluation are introduced in detail in Sections 4.2.1 and 4.2.2, respectively.

4.2.1. Stability evaluation

Stability is defined as the sensitivity of an FS method to the variations in the training set and quantified robustness of feature preferences (Kalousis et al., 2007). Stability can be measured by weight-based similarity and rank-based similarity. Weight-based similarity refers to the similarity of the weights assigned by multiple training scenarios of an FS method to the features, while rank-based stability indicates the similarity of the feature ranks generated in different training scenarios. In this study, we employ rank-based similarity to measure the stability of an FS method. On one hand, for wrapper FS methods, feature preferences are obtained

Table 3
Summary of feature selection methods.

No.	Approaches	Feature selection		
		Feature search	Evaluation criteria	Classifiers
1	Filter	Exhaustive search	RFLV (data variance)	
2			UFS (Pearson's correlation)	
3	Wrapper	SFFS	Feature rank based on prediction performance	SVM
4				DT
5				GBDT
6				RF
7		RFE	Feature rank based on features' importance	SVM
8				DT
9				GBDT
10				RF
11	Embedded	Exhaustive search*	Coefficients*	SVM
12		Stochastic search*	Information entropy*	DT
13				GBDT
14				RF

*Refers to the method or criterion included in the classifiers of feature selection (Pal and Foody, 2010; Hastie et al., 2009).

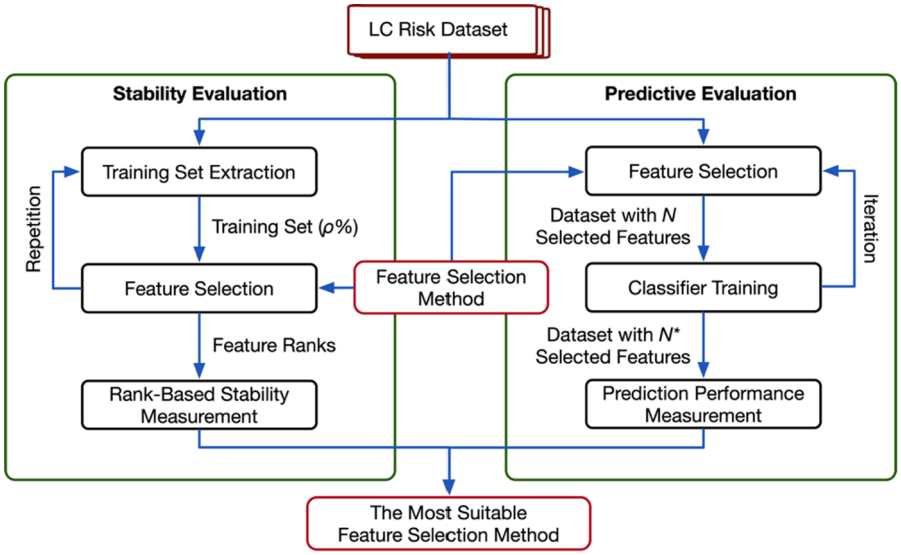


Fig. 8. General procedure of performance evaluation.

based on feature rank. Weighting schema can be easily cast as ranking schema, but not vice versa. On the other hand, rank-based similarity is more sensitive to the difference in feature preferences, especially when the weights assigned to features are fairly close to each other.

We propose an innovative algorithm to measure rank-based stability considering the effect of both randomness and size of the training set on the feature preferences. The pseudocode of the algorithm is shown as follows:

Start
input: dataset, D_t; FS method, M_f; Number of features, N.
$T \leftarrow$ [75%, 80%, 85%, 90%, 95%], which is the size of training set
M_f is trained on D_t
return R_r, which is the reference feature rank with top N features
for training size, ρ, **in** T:
  $i \leftarrow 1$
  while $i <= 20$:
    Randomly select ρ of D_t as the training set, $D_{\rho,i}$
    M_f is trained on $D_{\rho,i}$
    return R_i^{ρ}, which is sorted by R_r
    $i = i + 1$
  end while
  $i', i'' \in [1, 20]$
  compute $Corr_{\rho}^{i',i''}$, which is Spearman's *rho* of any two R_i^{ρ}
  compute $Corr_{\rho} = \frac{\sum_{i''}(\sum_{i'} Corr_{\rho}^{i',i''} - 1)}{20 \times 19}$
end for
compute $RS = \frac{\sum_{\rho} Corr_{\rho}}{5}$
return RS
output: rank-based stability, RS.
End

ρ (ρ = 75%, 80%, 85%, 90%, 95%) of the samples (i.e. LC events) in the LC risk dataset are randomly allocated as the training set. For each value of ρ, we repeat the allocation 20 times and obtain 20 randomly shuffled training sets, upon which a given FS method is trained to generate feature ranks respectively. For each FS method, the feature rank generated by the LC risk dataset (i.e. when $\rho = 1$) is treated as reference feature rank. As the results of different training scenarios, the features with various ranks shall be sorted by reference feature rank before measuring rank-based stability. Rank-based stability is measured by Spearman's rank correlation coefficient (Spearman's *rho*) (Kalousis et al., 2007). The Spearman's *rho* between two feature ranks, R_i^{ρ} and $R_{i''}^{\rho}$ can be obtained as:

$$Corr_{\rho}^{i',i''} = 1 - 6\sum_{m=1}^{N} \frac{\left(R_{i'}^{\rho}(m) - R_{i''}^{\rho}(m)\right)}{N(N^2-1)} \tag{4}$$

where $R_i^{\rho}(m)$ and $R_{i''}^{\rho}(m)$ respectively refer to the rankings of m^{th} feature in R_i^{ρ} and $R_{i''}^{\rho}$. As shown in the pseudocode, $Corr_{\rho}$ refers to the

average similarity of the feature ranks over the 20 training scenarios with a certain value of ρ. *RS* refers to the average similarity of the feature ranks over the training scenarios with the five possible values of ρ.

4.2.2. Predictive evaluation

The prediction performance of predictor, which is trained on the dataset with selected features, is of significance to not only the selection of FS method but also the determination of *Sub*'s position (i.e. space-stamp) for prediction and the number of risk levels n, as mentioned above. The predictive evaluation involves two main steps. Firstly, the predictor is trained on the LC risk dataset with the first N selected features to find out the least number of features N^* that achieves the highest prediction accuracy. In this study, prediction accuracy is measured by 5-fold cross-validation score (Rodriguez et al., 2010). We use Light Gradient Boosting Machine (LightGBM) classifier as the predictor trained on the dataset with selected features. LightGBM is an advanced GBDT algorithm developed by Microsoft Research, which can deal with a large number of data distances and achieve remarkable prediction performance (Ke et al., 2017). Secondly, the prediction performance of the predictor (i.e. LightGBM) trained on the dataset with N^* features is measured by precision (P), recall (R), F1 score (F1), and area under the curve of receiver operating characteristics (AUC) (Powers, 2011; Fawcett, 2006). Considering the multi-classes in LC risk dataset, the measures are integrated with one-against-all (OAA) scheme (Rifkin and Klautau, 2004).

5. Method validation

5.1. Data source

The vehicles' trajectory data used for method validation was streamed from a digitalized video-based dataset, the Next Generation Simulation (NGSIM) US-101 dataset, which was collected on a road segment of US Highway 101 in Los Angeles, California, USA from 7:50 am to 8:35 am on June 15, 2005 (FHWA, 2005). The vehicles' motion information was collected for every 0.1 sec ($\Delta t = 0.1s$). The lane width of standard highway in the US is 12 ft (3.6 m) (FHWA, 2014). In this study, we only consider the LC events in five mainline lanes and one auxiliary lane, excluding on-ramp and off-ramp. Accordingly, a total of 428 LC events involving a collocation of five cars in each event are obtained as samples. Savitzky-Golay filter (Savitzky and Golay, 1964) is employed to preprocess the NGSIM dataset as data cleaning to remove the potential noise and improve the quality of the dataset.

5.2. Resampling

This section aims to evaluate the resampling performance of EST, the method proposed in this study. Herein, we take the No.4 dataset in Table 2 as an example to evaluate the performance. SMOTETomek (Batista et al., 2003) and SMOTEENN (Batista et al., 2004), the two commonly used resampling methods that combine both over-sampling and under-sampling strategies, are employed to resample the LC risk dataset. Fig. 9 shows the resampling results of the three methods with respect to the number of risk levels. In

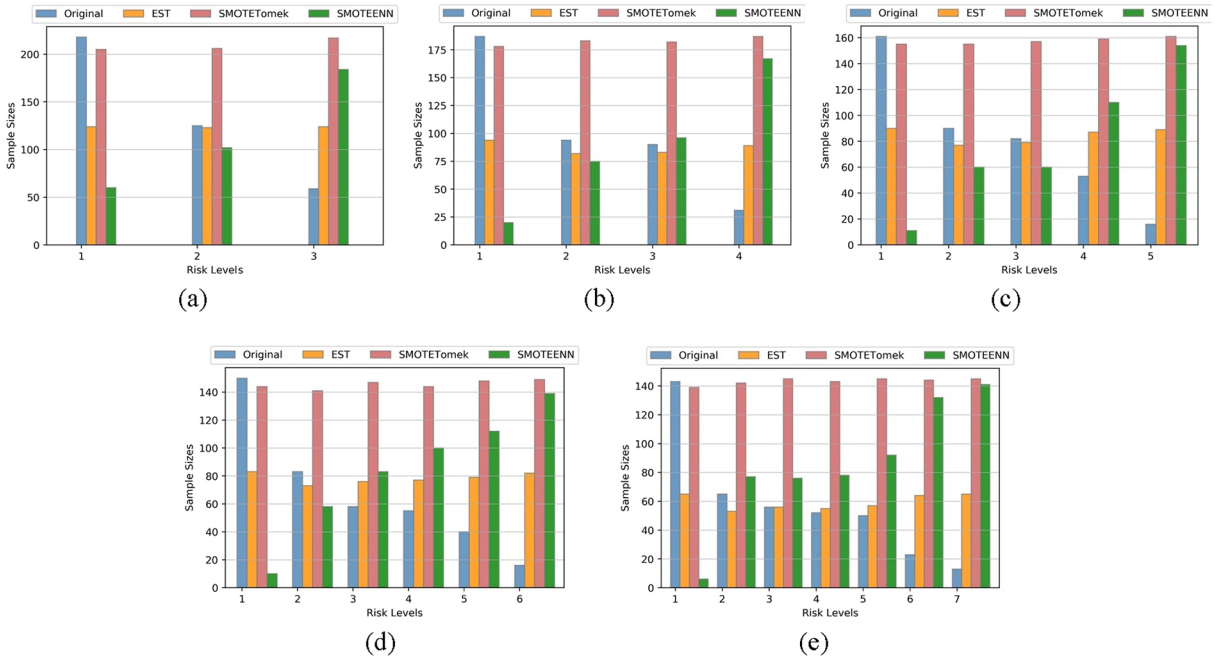


Fig. 9. Resampling results.

Table 4
Performance of resampling methods.

Resampling methods	Measures	Number of risk levels (n)				
		3	4	5	6	7
EST	*IR*	1.01	1.08	1.09	1.02	1.03
	OR	0.16	0.14	0.27	0.36	0.25
	UR	0.24	0.28	0.21	0.19	0.22
SMOTETomek	*IR*	1.02	1.02	1.02	1.01	1.01
	OR	0.59	1.07	0.97	1.19	1.50
	UR	0.03	0.02	0.01	0.01	0.00
SMOTEENN	*IR*	2.19	3.69	3.86	4.01	5.44
	OR	0.31	0.35	0.49	0.66	0.84
	UR	0.45	0.46	0.51	0.41	0.34
Original	*IR*	2.48	2.68	3.29	2.68	2.79

Fig. 9, 'Original' refers to the LC risk dataset before being resampled. Table 4 presents the resampling performance of each method based on the measurements introduced in Section 3.2.

As shown in Fig. 9 and Table 4, the LC risk dataset before being resampled has a serious class imbalance problem. The dataset resampled by SMOTETomek can effectively solve the class imbalance problem. However, the over-sampling ratio (*OR*) of SMOTETomek is higher than those of the other two methods. The large sample size over-resampled by SMOTETomek is more likely to result in overfitting on the samples of minority classes. SMOTEENN cannot effectively alleviate the class imbalance problem. The dataset resampled by SMOTEENN still has a high imbalance ratio (*IR*). EST achieves better resampling performance, compared with the other two methods, in terms of resampled sample size and class distribution of resampled dataset, which means that EST is less likely to cause overfitting problem and mass information loss of the majority class. Consequently, EST is selected as the resampling method in this study.

5.3. Stability of feature selection

In this study, the FS methods are respectively trained on the seven LC risk datasets introduced in Table 2 to rank the features, based on which the rank-based stability is measured using the algorithm introduced in Section 4.2.1. Herein, as examples to illustrate the stability of each FS method, the FS methods are trained on the LC risk datasets labeled with three risk levels. As shown in Figs. 10–13, RFLV, as a filter FS method, and RF, as an embedded FS method, attain high performance on stability, whereas the other FS methods fail to achieve acceptable stability.

As a filter FS method, UFS achieves unacceptable stability since Pearson's correlation cannot effectively detect the noisy features and the non-linear association between features (Wilcox, 2001). The wrapper FS methods, including SFFS-based and RFE-based methods, cannot achieve satisfactory performance on stability due to the lack of robustness across classifiers (Dash et al., 2002) and generality (Zhang et al., 2014), especially for a dataset with a large number of candidate features. As an embedded FS method, RF employs bagging (i.e. bootstrap aggregation), for which the method can decrease both variance and bias of prediction results (Friedman and Hall, 2007) and improve the stability of classifier (Breiman, 1996), which enables RF to achieve higher stability than the other embedded FS methods.

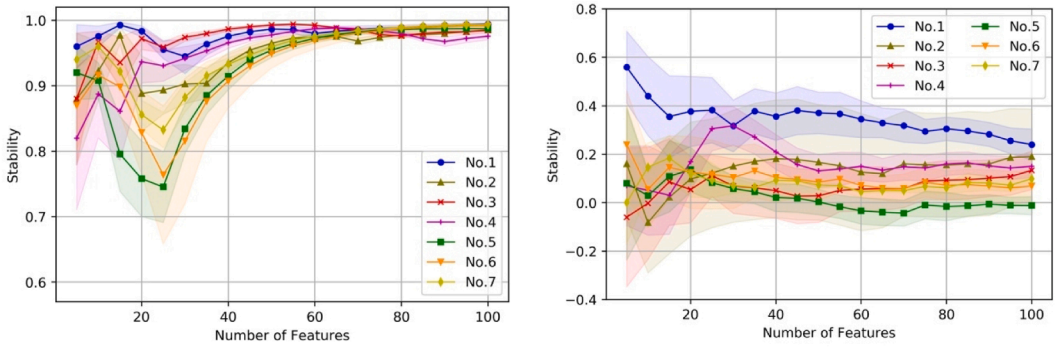
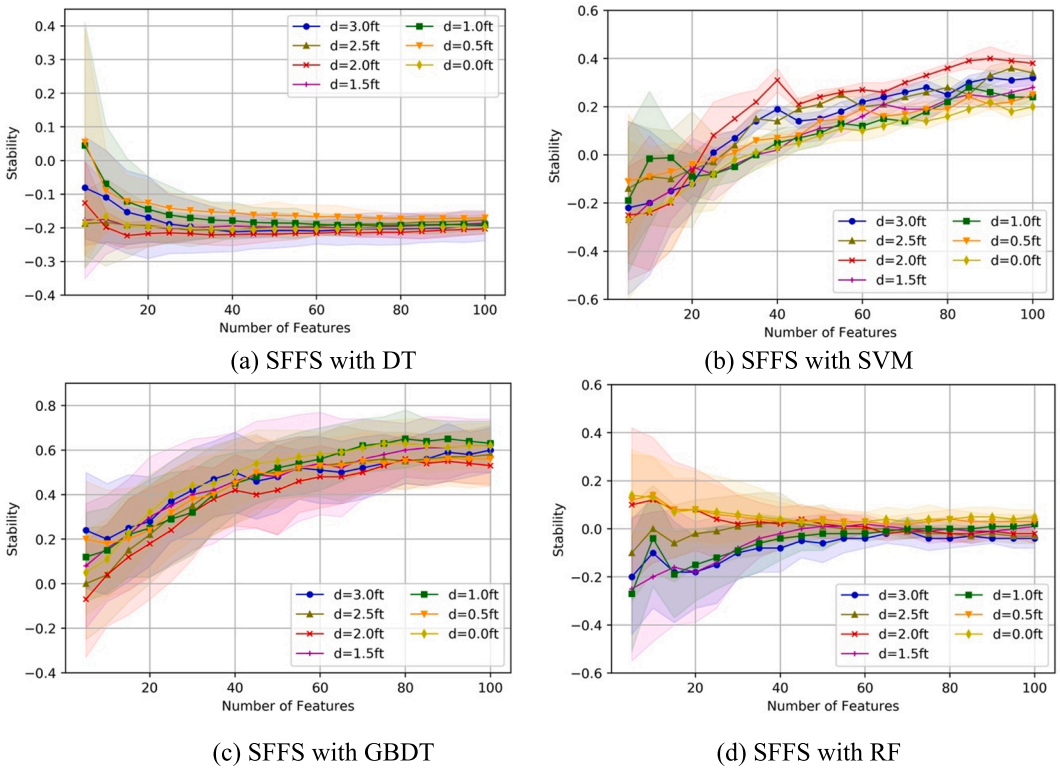


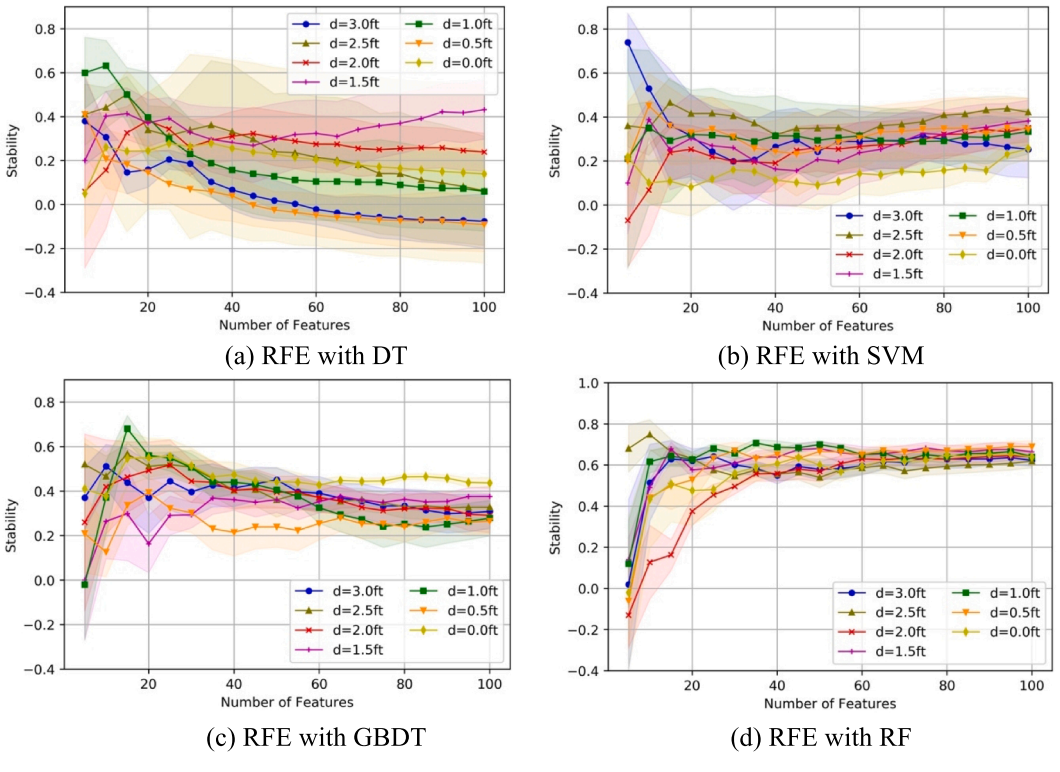
Fig. 10. Rank-based stability of filter feature selection methods.



(a) SFFS with DT (b) SFFS with SVM

(c) SFFS with GBDT (d) SFFS with RF

Fig. 11. Rank-based stability of wrapper feature selection methods (SFFS).



(a) RFE with DT (b) RFE with SVM

(c) RFE with GBDT (d) RFE with RF

Fig. 12. Rank-based stability of wrapper feature selection methods (RFE).

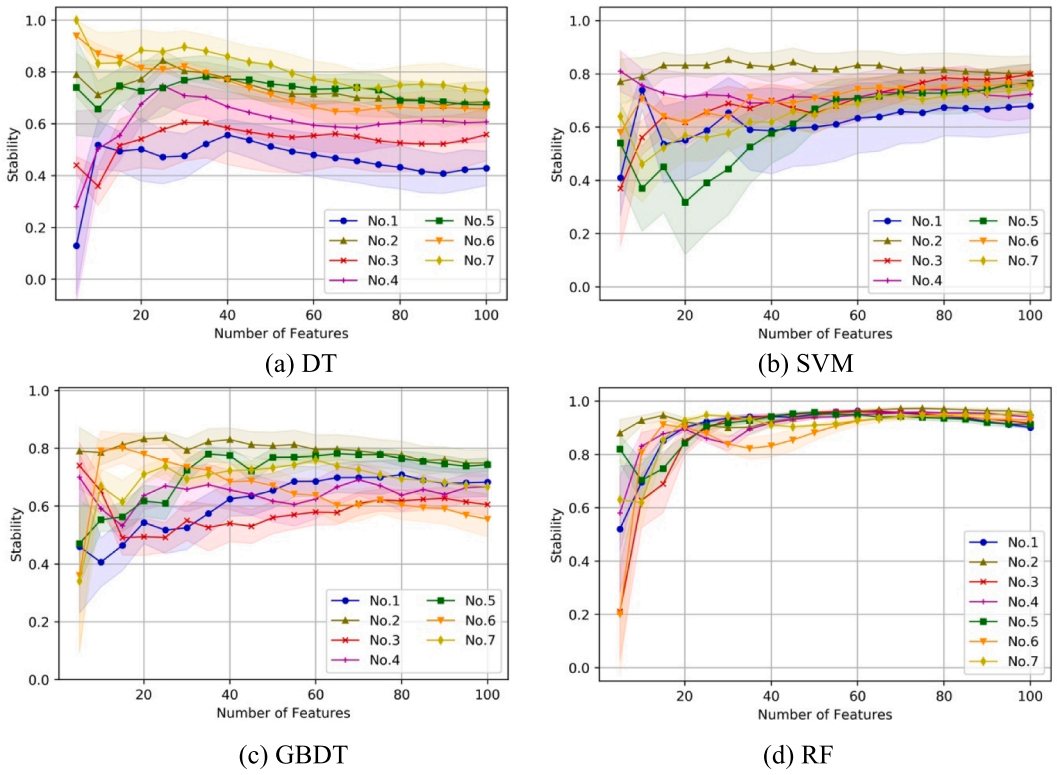


Fig. 13. Rank-based stability of embedded feature selection methods.

5.4. Prediction performance

As mentioned above, the evaluation of prediction performance has four purposes: a) to select the most suitable FS method, b) to select the key features, c) to determine the number of risk levels (i.e. n), and d) to determine which position *Sub* moves to such that LC risk can be accurately predicted. The FS methods excluding RFLV and RF are discarded because of the unacceptable stability, as shown in Section 5.3. Hence, we only consider the prediction performance of the predictor trained on the dataset with the features selected by RFLV and RF, respectively, for the selection of FS method. Herein, as examples, RFLV and RF are respectively trained on the No.4 LC risk datasets labeled with three risk levels to selected key features. Then, the predictor (i.e. LightGBM) is trained on the LC risk dataset with the selected key features. The results of prediction performance are shown in Fig. 14 and Table 5.

As shown in Fig. 14 and Table 5, the predition performance achieved by the predictor trained on the N^* features selected by RF is higher. Therefore, RF is selected as the FS method in this study because of the high performance it achieves on both stability and LC risk level prediction. The prediction accuracy with respect to the number of the features selected by RF, which is trained on the LC

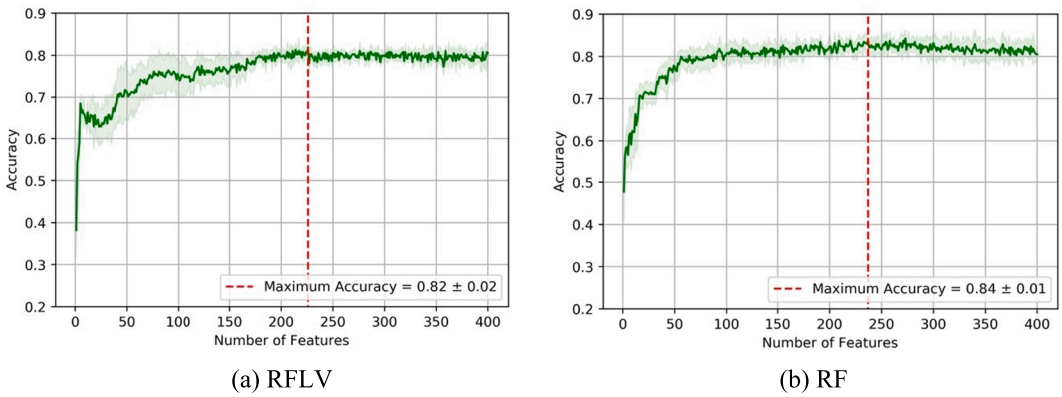


Fig. 14. Prediction accuracy based on features selected by RFLV and RF. 1. The horizontal-axis value of the dash line refers to the least number of features, N^*, with which the predictor is trained to achieve the maximum prediction accuracy.

Table 5
Prediction performance based on N^* features selected by RFLV and RF.

FS methods	RFLV				RF			
Measures	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.87	0.94	0.90	0.97	0.87	0.95	0.91	0.98
Risk level 2	0.72	0.63	9.67	9,86	0.79	0.71	0.75	0.87
Risk level 3	0.72	0.63	0.67	0.93	0.83	0.84	0.83	0.95
Micro avg.	0.79	0.79	0.79	0.91	0.83	0.83	0.83	0.90
Macro avg.	0.79	0.79	0.79	0.92	0.83	0.83	0.83	0.93

risk datasets with different values of n (i.e. number of risk levels), is shown in Figs. A1–A7 in Appendix A. Accordingly, the maximum prediction accuracy based on the LC risk datasets with each value of n is shown in Fig. 15. The prediction performance achieved by the predictor trained on the LC risk dataset with N^* features is shown in Tables B1–B7 in Appendix B.

As shown in Fig. 15 and Table B4, the predictor trained on the No.4 LC risk dataset with n equal to 3 can achieve the most satisfactory prediction accuracy (0.85 ± 0.02). Consequently, the LC risk is recommended to be classified into three risk levels. As introduced in Table 2, cars' motion behavioral features of the No.4 LC risk dataset are collected till the LC car, *Sub*, moves to the position P_4 during an LC event as illustrated in Fig. 5(b). It means that, based on the P-LRLP method, the LC risk can be most accurately predicted when *Sub* moves to the position where the distance between *Sub*'s longitudinal center line and the marking line seperating original lane and target lane equals 1.5 ft.

5.5. Discussion

5.5.1. Analysis of feature relevance

As mentioned above, the key features are selected by RF as an embedded FS method in this study. Herein, we take the No.4 dataset with three risk levels as an example, on which FS methods are trained, to illustrate the relevance between the selected features. The heatmaps in Fig. 16 present the correlation coefficients of the first 60 features selected by the embedded FS methods as mentioned above.

Compared with the other three embedded FS methods, the features selected by RF are more correlated to each other, especially for those features ranked ahead, which can be explained by the feature weighting criterion of the RF classifier (Guyon et al., 2002; Ratanamahatana and Gunopulos, 2003; Strobl et al., 2008). Despite the increasing relevance among the selected features improves the interpretation of those features (Altmann et al., 2010), selected features with high relevance might cause certain negative effects on the prediction, such as increasing computational time (Yu and Liu, 2003). Nevertheless, in this study, LightGBM, as a predictor with high performance (Ke et al., 2017), is able to effectively attenuate the potential negative effects caused by the features selected by RF.

5.5.2. Evaluation of predictor's performance

LightGBM algorithm is implemented with Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to improve prediction accuracy and computational speed (Ke et al., 2017). Herein, to verify LightGBM's performance on LC risk level prediction, four conventional machine learning classifiers, namely DT, SVM classifier (SVC), RF, and Multi-Layer Perceptron classifier (MLP) (Hastie et al., 2009), are employed to be trained on the No.4 LC risk dataset with n equal to 3 as comparisons. The results of prediction accuracy achieved by each predictor are presented in Fig. 17. LightGBM achieves higher accuracy than the conventional machine learning classifiers, which is superior in LC risk level prediction.

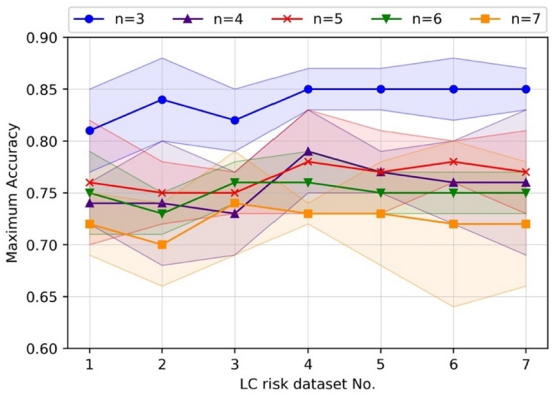


Fig. 15. Maximum prediction accuracy.

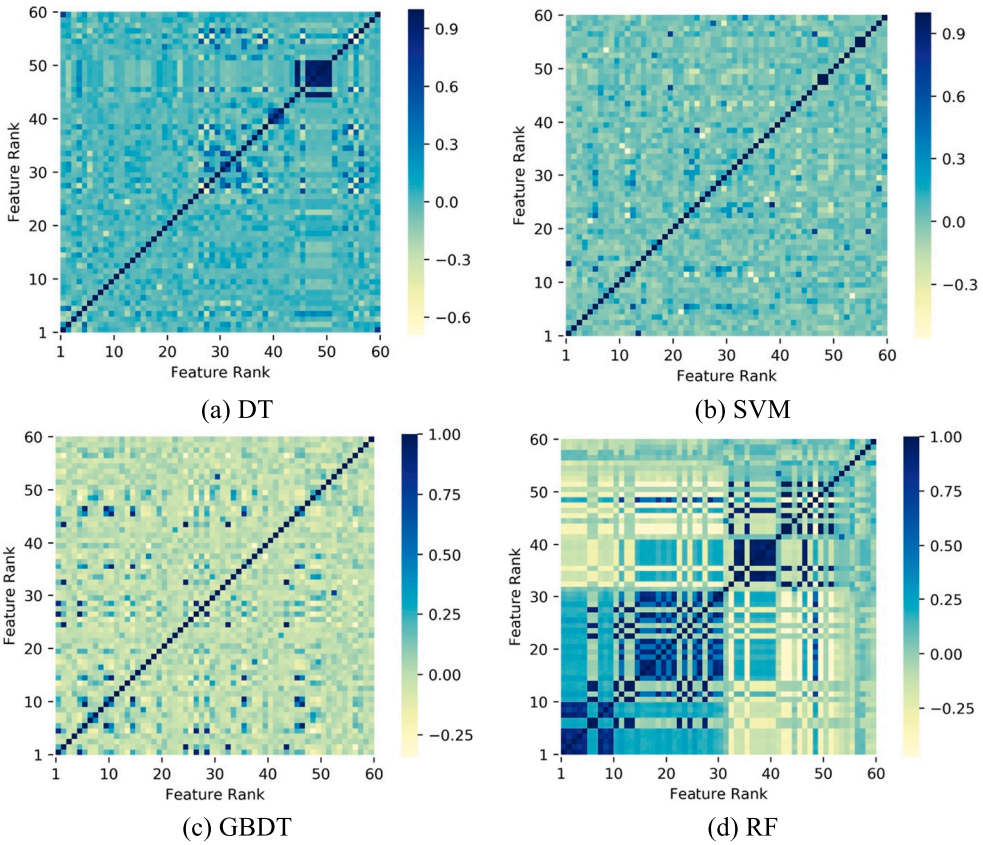


Fig. 16. Heatmaps of features' correlation coefficients based on embedded feature selection method.

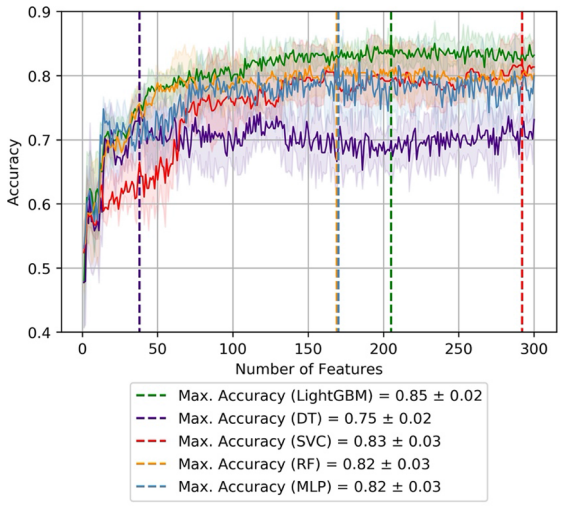


Fig. 17. Prediction accuracy achieved by LightGBM and conventional predictors.

5.5.3. Analysis of key features

As shown in Fig. A4(a), the first 205 features selected by RF based on the No.4 LC risk dataset with three risk levels are regarded as key features. Fig. 18 illustrates the importance of the key features catagorized into individual features and interaction features as introduced in Table 1. The key features can also be catagorized into the groups of acceleration (e.g. cars' acceleration, acceleration difference), velocity (e.g. cars' velocity, velocity difference), and distance (e.g. cars' size, cars' coordinates, gap distance). Three main findings are summarized as follows:

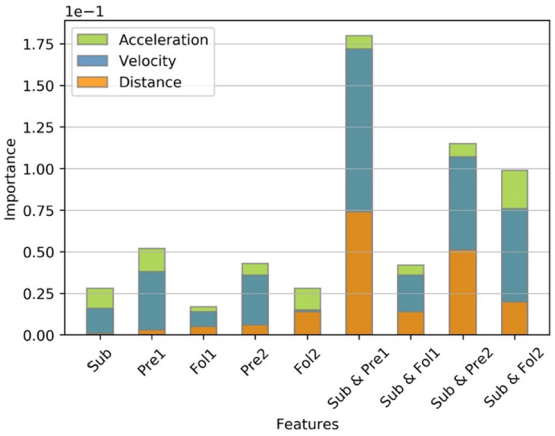


Fig. 18. Histogram of feature importance.

(a) *Sub*'s LC maneuver can be treated as an interruption to the car following maneuver (i.e *Fol2* follows *Pre2*) in the target lane. Hence, features with respect to the interaction between *Sub* and the cars in the target lane show great importance.
(b) The interaction features between *Sub* and *Pre1* are also accorded greater importance. On one hand, the features are extracted when *Sub*'s longitudinal center line is still placed within the original lane, which basically can be treated as LC maneuver. On the other hand, *Sub*'s driver would easily draw more attention on the cars in the target lane (i.e. *Pre2* and *Fol2*) than *Pre1* when changing lanes (Salvucci and Liu, 2002).
(c) The features relevant to the preceding cars (i.e. *Pre1* and *Pre2*) contribute more to the LC risk, which can be ascribed to the lesser attention those two cars draw on the situation behind. The verification of this hypothesis can be proposed in future work.

6. Conclusions

6.1. Summary and conclusions

In this study, we propose a pre-emptive LC risk level prediction (P-LRLP) method, which is able to predict the LC risk of an LC event that involves a cluster of five cars before the LC car completes the LC maneuver. The P-LRLP method includes the steps of dataset construction, resampling, feature selection and predictive evaluation. The space-series motion behavioral features of each car involved in an LC event are included as the candidate features of the LC risk dataset with LC risk levels as label. We also propose an innovative resampling method, namely EST, and an algorithm which can comprehensively evaluate the rank-based stability of selected features considering the randomness and size of training samples. As the predictor, LightGBM is trained on the resampled LC risk dataset with the selected features to predict the LC risk level of an LC event.

The NGSIM dataset is employed for the validation of the P-LRLP method. The evaluation of resampling methods shows that EST is able to achieve satisfactory resampling performance. RF, as an embedded FS method, is selected as the FS method in this study due to its high performance on the stability of selected features and risk prediction. Based on the results of predictive evaluation, the LC risk is recommended to be clustered into three levels and the first 205 features selected by RF are regarded as the key features. As the results of key feature analysis, RF is an acceptable FS method and the features selected by RF can provide several useful interpretations

Using the P-LRLP method, the LC risk can be most accurately predicted when the LC car moves to the position where the distance between the longitudinal center line of the LC car and the marking line separating original lane and target lane equals 1.5 ft in original lane. The method comprehensively considers the LC scenario which includes four surrounding cars in both original and target lanes. The P-LRLP method is able to predict the LC risk in advance before the LC car completes the LC maneuver, which is different from the posterior prediction methods proposed in previous research. Furthermore, this study could provide a basis for future research on risk control during an LC process.

6.2. Limitations and future work

From the perspective of data source, we only consider the LC scenario that involves a cluster of five cars on the highway and employ the NGSIM dataet from the US for validation. Also, the quality of the NGSIM dataset has drawn concern from research community (Coifman and Li, 2017). The calibration accuracy of the trajectory dataset is of importance to the P-LRLP method. Consequently, in future, the high-quality vehicles' trajectory dataset which involves various LC scenarios could be employed to improve the generalization of the P-LRLP method and gain deeper understanding of LC safety. Additionally, the cases of LC crash accident could be involved to further investigate the accident causation into how high-risk LC maneuver could evolve into an LC crash accident.

From the technical perspective, as a method akin to a “black-box”, machine learning techniques should be further transparentized for better interpretability. The “black-box” can be resolved from two angles in future. Firstly, quantitative risk analysis methods (e.g. advanced Bayesian-Network) can be employed to investigate the causal relationship and risk mechanism of an LC crash accident. A large scale dataset of vehicles’ trajectory and LC crash accident record could provide a solid basis for the quantitative analysis. Secondly, “in-car” experiments could be conducted to qualitatively investigate the contributions of human factors to risky LC maneuvers. Upon transparentizing the “black-box”, an LC warning system can be developed by integrating the P-LRLP method with advanced driver-assistance system (ADAS) and vehicle-to-vehicle (V2V) communication. The LC warning system could detect risky LC behavior and provide corresponding assistance as feedback to the different predicted LC risk levels. Last but not least, more efforts could be devoted to improve the prediction performance of the P-LRLP method.

CRediT authorship contribution statement

Tianyi Chen: Conceptualization, Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing. **Xiupeng Shi:** Conceptualization, Formal analysis. **Yiik Diew Wong:** Supervision, Conceptualization, Writing - review & editing. **Xiaocong Yu:** Methodology, Software.

Acknowledgements

This paper presents a part of the first author’s PhD research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Figures of prediction accuracy based on candidate LC risk dataset

Figs. A1–A7.

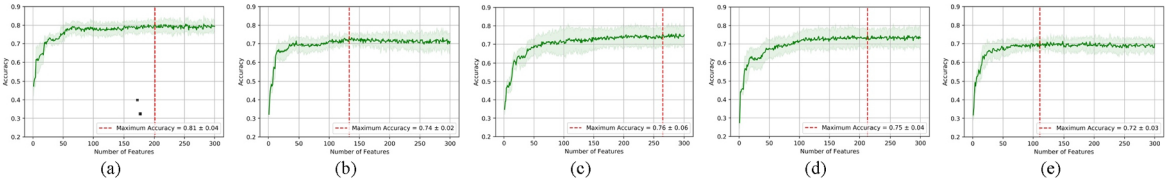


Fig. A1. Prediction accuracy based on No.1 LC risk dataset.

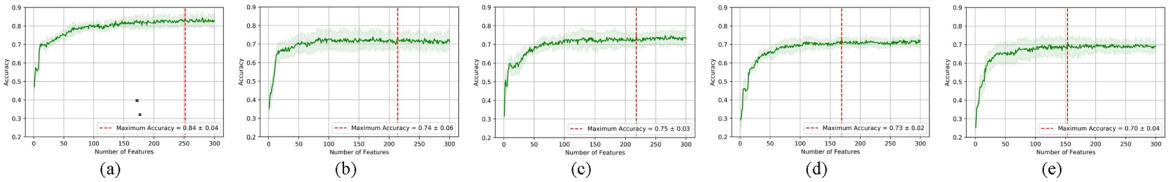


Fig. A2. Prediction accuracy based on No.2 LC risk dataset.

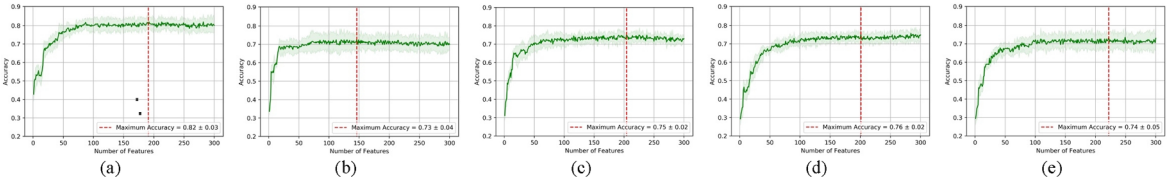


Fig. A3. Prediction accuracy based on No.3 LC risk dataset.

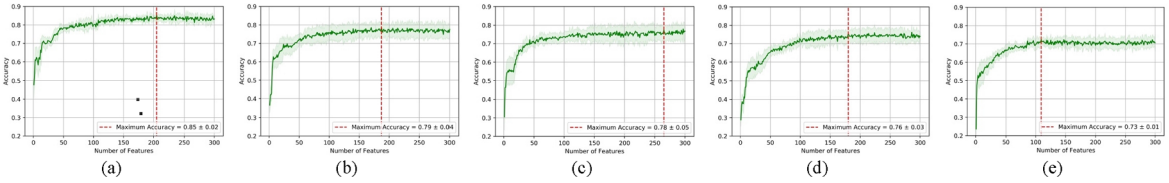


Fig. A4. Prediction accuracy based on No.4 LC risk dataset.

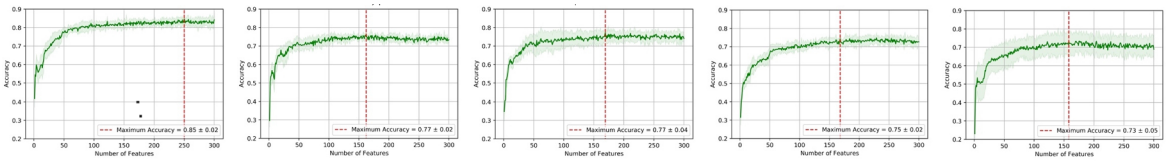


Fig. A5. Prediction accuracy based on No.5 LC risk dataset.

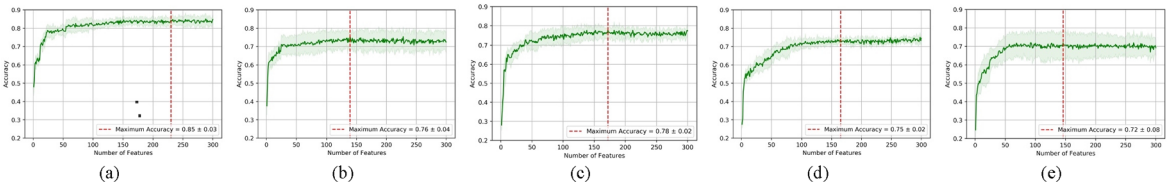


Fig. A6. Prediction accuracy based on No.6 LC risk dataset.

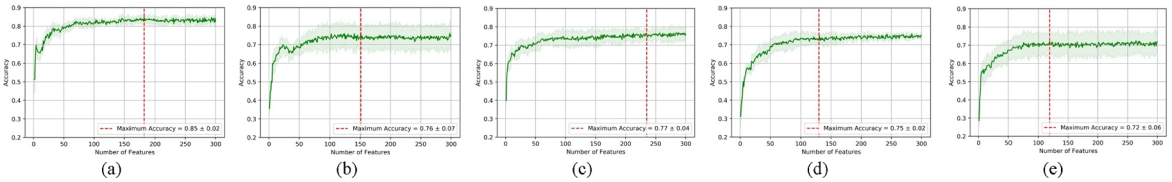


Fig. A7. Prediction accuracy based on No.7 LC risk dataset.

Appendix B. Tables of prediction performance based on candidate LC risk dataset

Tables B1–B7.

Table B1
Prediction performance based on No.1 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.85	0.90	0.87	0.96	0.80	0.96	0.87	0.98	0.86	0.99	0.92	0.99	0.80	0.96	0.87	0.99	0.79	0.92	0.85	0.98
Risk level 2	0.74	0.70	0.72	0.85	0.66	0.45	0.54	0.79	0.62	0.56	0.59	0.85	0.67	0.43	0.52	0.87	0.57	0.45	0.51	0.84
Risk level 3	0.83	0.83	0.83	0.94	0.62	0.60	0.61	0.87	0.57	0.44	0.50	0.83	0.62	0.64	0.63	0.89	0.51	0.39	0.44	0.81
Risk level 4					0.80	0.90	0.85	0.96	0.74	0.74	0.74	0.84	0.71	0.57	0.63	0.85	0.67	0.69	0.68	0.92
Risk level 5									0.86	0.99	0.92	0.99	0.75	0.81	0.77	0.97	0.65	0.58	0.61	0.89
Risk level 6													0.86	0.99	0.92	0.99	0.77	0.88	0.82	0.99
Risk level 7																	0.88	0.98	0.93	0.99
Micro avg.	0.81	0.81	0.81	0,89	0.74	0.74	0.74	0.93	0.76	0.76	0.76	0.92	0.75	0.75	0.75	0.93	0.72	0.72	0.72	0.89
Macro avg.	0.81	0.81	0.81	0.91	0.72	0.73	0.72	0.90	0.73	0.74	0.73	0.92	0.73	0.73	0.73	0.92	0.69	0.70	0.69	0.92

Table B2
Prediction performance based on No.2 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.88	0.91	0.89	0.97	0.85	0.99	0.92	0.99	0.86	0.98	0.92	0.99	0.83	0.96	0.89	0.98	0.79	0.98	0.88	0.99
Risk level 2	0.79	0.77	0.78	0.90	0.63	0.45	0.52	0.82	0.57	0.50	0.53	0.89	0.58	0.53	0.55	0.89	0.48	0.30	0.68	0.84
Risk level 3	0.86	0.84	0.85	0.95	0.59	0.54	0.57	0.84	0.52	0.43	0.47	0.84	0.61	0.65	0.63	0.87	0.52	0.42	0.47	0.82
Risk level 4					0.79	0.93	0.85	0.97	0.73	0.79	0.76	0.95	0.68	0.51	0.58	0.83	0.60	0.49	0.54	0.87
Risk level 5									0.92	0.97	0.94	0.99	0.76	0.73	0.74	0.96	0.66	0.70	0.68	0.89
Risk level 6													0.84	0.95	0.89	0.99	0.75	0.92	0.83	0.99
Risk level 7																	0.89	0.98	0.93	0.99
Micro avg.	0.84	0.84	0.84	0.95	0.74	0.74	0.74	0.95	0.75	0.75	0.75	0.90	0.73	0.73	0.73	0.96	0.71	0.71	0.71	0.85
Macro avg.	0.84	0.84	0.84	0.94	0.72	0.73	0.71	0.91	0.72	0.73	0.72	0.93	0.72	0.72	0.72	0.92	0.67	0.69	0.67	0.91

Table B3
Prediction performance based on No.3 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.86	0.90	0.88	0.97	0.86	0.89	0.88	0.98	0.85	0.98	0.91	0.99	0.80	0.97	0.88	0.99	0.81	0.92	0.86	0.99
Risk level 2	0.73	0.73	0.73	0.89	0.55	0.53	0.54	0.83	0.63	0.61	0.62	0.88	0.58	0.56	0.57	0.86	0.54	0.52	0.53	0.85
Risk level 3	0.82	0.78	0.80	0.93	0.59	0.54	0.57	0.84	0.57	0.41	0.48	0.82	0.67	0.58	0.62	0.87	0.58	0.43	0.50	0.85
Risk level 4					0.86	0.91	0.88	0.98	0.72	0.72	0.72	0.95	0.73	0.61	0.67	0.88	0.72	0.63	0.67	0.89
Risk level 5									0.88	0.99	0.93	0.99	0.80	0.76	0.78	0.95	0.67	0.61	0.64	0.90
Risk level 6													0.86	0.97	0.91	0.99	0.81	0.97	0.88	0.99
Risk level 7																	0.90	1.00	0.95	0.99
Micro avg.	0.80	0.80	0.80	0.88	0.73	0.73	0.73	0.92	0.75	0.75	0.75	0.95	0.75	0.75	0.75	0.91	0.74	0.74	0.74	0.85
Macro avg.	0.80	0.80	0.80	0.93	0.71	0.72	0.72	0.91	0.73	0.74	0.73	0.93	0.74	0.74	0.74	0.92	0.72	0.73	0.72	0.92

Table B4
Prediction performance based on No.4 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.89	0.94	0.91	0.98	0.85	0.97	0.91	0.99	0.86	0.99	0.92	0.99	0.88	0.97	0.92	0.99	0.77	0.98	0.86	0.99
Risk level 2	0.80	0.76	0.78	0.88	0.73	0.58	0.64	0.89	0.67	0.59	0.63	0.91	0.56	0.60	0.58	0.89	0.52	0.28	0.36	0.82
Risk level 3	0.87	0.85	0.86	0.94	0.68	0.65	0.67	0.86	0.57	0.55	0.56	0.87	0.72	0.64	0.68	0.89	0.46	0.37	0.41	0.81
Risk level 4					0.84	0.91	0.88	0.98	0.78	0.71	0.74	0.94	0.72	0.48	0.58	0.85	0.52	0.47	0.50	0.86
Risk level 5									0.92	0.99	0.95	0.99	0.72	0.83	0.77	0.97	0.63	0.65	0.64	0.89
Risk level 6													0.89	0.97	0.93	0.99	0.76	0.94	0.84	0.99
Risk level 7																	0.90	0.98	0.94	0.99
Micro avg.	0.85	0.85	0.85	0.91	0.79	0.79	0.79	0.95	0.77	0.77	0.77	0.94	0.76	0.76	0.76	0.95	0.69	0.69	0.69	0.84
Macro avg.	0.85	0.85	0.85	0.93	0.78	0.78	0.77	0.93	0.76	0.77	0.76	0.94	0.75	0.75	0.74	0.93	0.65	0.67	0.65	0.91

Table B5
Prediction performance based on No.5 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.90	0.94	0.92	0.98	0.84	0.95	0.89	0.98	0.93	0.97	0.95	0.99	0.86	0.96	0.91	0.99	0.82	0.98	0.90	0.99
Risk level 2	0.77	0.81	0.79	0.91	0.67	0.52	0.59	0.84	0.62	0.66	0.64	0.89	0.64	0.55	0.59	0.87	0.51	0.43	0.47	0.85
Risk level 3	0.88	0.80	0.84	0.94	0.65	0.61	0.63	0.87	0.62	0.49	0.54	0.88	0.57	0.66	0.61	0.89	0.63	0.55	0.59	0.87
Risk level 4					0.84	0.95	0.89	0.98	0.73	0.74	0.73	0.93	0.66	0.51	0.57	0.85	0.60	0.52	0.56	0.88
Risk level 5									0.91	0.98	0.94	0.99	0.77	0.78	0.77	0.95	0.67	0.65	0.66	0.88
Risk level 6													0.92	0.97	0.94	0.99	0.81	0.94	0.87	0.98
Risk level 7																	0.94	1.00	0.97	0.99
Micro avg.	0.85	0.85	0.85	0.94	0.77	0.77	0.77	0.91	0.77	0.77	0.77	0.93	0.74	0.74	0.74	0.94	0.73	0.73	0.73	0.85
Macro avg.	0.85	0.85	0.85	0.94	0.75	0.76	0.75	0.92	0.76	0.77	0.76	0.94	0.74	0.74	0.73	0.92	0.71	0.72	0.72	0.92

Table B6
Prediction performance based on No.6 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.89	0.96	0.92	0.98	0.81	0.96	0.88	0.97	0.85	0.96	0.90	0.99	0.83	0.95	0.89	0.99	0.80	0.92	0.86	0.98
Risk level 2	0.82	0.70	0.76	0.91	0.65	0.43	0.51	0.84	0.64	0.53	0.58	0.90	0.57	0.52	0.54	0.88	0.53	0.50	0.51	0.87
Risk level 3	0.82	0.87	0.84	0.95	0.63	0.65	0.64	0.89	0.63	0.53	0.58	0.87	0.66	0.68	0.67	0.89	0.63	0.50	0.51	0.87
Risk level 4					0.86	0.95	0.90	0.98	0.77	0.83	0.80	0.94	0.63	0.57	0.60	0.87	0.51	0.43	0.47	0.88
Risk level 5									0.85	0.96	0.90	0.99	0.78	0.73	0.76	0.94	0.66	0.68	0.67	0.90
Risk level 6													0.90	0.97	0.93	0.99	0.85	0.94	0.89	0.99
Risk level 7																	0.92	1.00	0.96	0.99
Micro avg.	0.85	0.85	0.85	0.91	0.76	0.76	0.76	0.94	0.78	0.78	0.78	0.95	0.74	0.74	0.74	0.93	0.72	0.72	0.72	0.83
Macro avg.	0.84	0.84	0.84	0.95	0.74	0.74	0.74	0.92	0.76	0.77	0.76	0.94	0.73	0.74	0.73	0.93	0.71	0.72	0.71	0.93

Table B7
Prediction performance based on No.7 LC risk dataset.

	3				4				5				6				7			
Measures	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
Risk level 1	0.89	0.90	0.90	0.98	0.81	0.94	0.87	0.98	0.81	0.98	0.89	0.99	0.80	0.94	0.87	0.98	0.74	0.92	0.82	0.98
Risk level 2	0.79	0.75	0.77	0.92	0.65	0.51	0.57	0.86	0.62	0.49	0.55	0.88	0.55	0.53	0.54	0.87	0.48	0.47	0.47	0.87
Risk level 3	0.86	0.89	0.87	0.95	0.67	0.67	0.67	0.87	0.57	0.46	0.51	0.88	0.66	0.68	0.67	0.86	0.69	0.47	0.56	0.83
Risk level 4					0.88	0.91	0.90	0.98	0.73	0.80	0.77	0.95	0.65	0.48	0.55	0.86	0.62	0.52	0.57	0.88
Risk level 5									0.94	1.00	0.97	0.99	0.76	0.79	0.77	0.96	0.63	0.63	0.63	0.87
Risk level 6													0.92	0.96	0.94	0.99	0.83	0.97	0.89	0.99
Risk level 7																	0.93	1.00	0.96	0.99
Micro avg.	0.85	0.85	0.85	0.92	0.77	0.77	0.77	0.94	0.76	0.76	0.76	0.96	0.73	0.73	0.73	0.94	0.72	0.72	0.72	0.83
Macro avg.	0.85	0.85	0.85	0.95	0.75	0.76	0.75	0.92	0.73	0.75	0.74	0.94	0.72	0.73	0.72	0.92	0.70	0.71	0.70	0.92

References

Ali, Y., Haque, M.M., Zheng, Z., Washington, S., Yildirimoglu, M., 2019. A hazard-based duration model to quantify the impact of connected driving environment on safety during mandatory lane-changing. Transportation Res. Part C: Emerg. Technol. 106, 113–131.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.

Arbis, D., Dixit, V.V., 2019. Game theoretic model for lane changing: Incorporating conflict risks. Accid. Anal. Prev. 125, 158–164.

Batista, G.E., Bazzan, A.L., Monard, M.C., 2003. Balancing training data for automated annotation of keywords: a case study. WOB 10–18.

Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter 6, 20–29.

Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., Agakov, F., Navarro, P., 2015. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. 5, 10312.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Breiman, L., 1996. Bagging predictors. Mach. Learning 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learning 45, 5–32.

Breiman, L., 2017. Classification and Regression Trees. Routledge.

Chandrashekar, G., Sahin, F.J.C., Engineering, E., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28.

Chawla, N.V., 2009. Data mining for imbalanced datasets: An overview. Data Mining and Knowledge Discovery Handbook. Springer.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357.

Chen, T., Shi, X., Wong, Y.D., 2019. Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data. Accid. Anal. Prev. 156–169.

Coifman, B., Li, L., 2017. A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset. Transportation Res. Part B: Methodol. 105, 362–377.

Cooper, D.F., Ferguson, N., 1976. Traffic studies at T-Junctions. 2. A conflict simulation record. Traffic Eng. Control 17.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learning 20, 273–297.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27.

Cunto, F., Saccomanno, F.F., 2008. Calibration and validation of simulated vehicle safety performance at signalized intersections. Accid. Anal. Prev. 40, 1171–1179.

Dash, M., Choi, K., Scheuermann, P., Liu, H., 2002. Feature selection for clustering-a filter solution. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, pp. 115–122.

Drummond, C., Holte, R.C., 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Workshop on Learning From Imbalanced Datasets II. Citeseer, pp. 1–8.

Estabrooks, A., Jo, T., Japkowicz, N., 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. 20, 18–36.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 861–874.

FHWA, 2005. Next Generation Simulation (NGSIM) [Online]. Available: https:// ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm [Accessed September 2018].

FHWA, 2014. Safety [Online]. Available: https://safety.fhwa.dot.gov/geometric/pubs/mitigationstrategies/chapter3/3_lanewidth.cfm [Accessed September 2018].

Freeman, C., Kulić, D., Basir, O., 2015. An evaluation of classifier-specific filter measure performance for feature selection. Pattern Recogn. 48, 1812–1826.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.

Friedman, J.H., Hall, P., 2007. On bagging and nonlinear estimation. J. Stat. Planning Inference 137, 669–683.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learning Res. 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learning 46, 389–422.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, Springer, New York.

He, X., Cai, D., Niyogi, P., 2006. Laplacian score for feature selection. Adv. Neural Inf. Process. Syst. 507–514.

Hou, Y., Edara, P., Sun, C., 2013. Modeling mandatory lane changing using Bayes classifier and decision trees. IEEE Trans. Intell. Transp. Syst. 15, 647–655.

Hou, Y., Edara, P., Sun, C., 2015. Situation assessment and decision making for lane change assistance using ensemble learning methods. Expert Syst. Appl. 42, 3875–3882.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.

Jain, A., Zongker, D., 1997. Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19, 153–158.

Jeong, H., Jang, Y., Bowman, P.J., Masoud, N., 2018. Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. Accid. Anal. Prev. 120, 250–261.

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. 12, 95–116.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 3146–3154.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97, 273–324.

Lee, C., Park, P.Y., Abdel-aty, M., 2011. Lane-by-lane analysis of crash occurrence based on driver's lane-changing and car-following behavior. J. Transportation Saf. Sec. 3, 108–122.

Liang, G., Zhang, C., 2012. An efficient and simple under-sampling technique for imbalanced time series classification. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. ACM, pp. 2339–2342.

Liu, H., Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. 491–502.

Pal, M., Foody, G., 2010. Feature selection for classification of hyperspectral data by SVM. IEEE Trans. Geosci. Remote Sens. 48, 2297–2307.
Pande, A., Abdel-aty, M., 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. Accid. Anal. Prev. 38, 936–948.
Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learning Technol. 2 (1), 37–63.
Pudil, P., Novovičová, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recogn. Lett. 15, 1119–1125.
Ratanamahatana, C.A., Gunopulos, D., 2003. Feature selection for the naive bayesian classifier using decision trees. Appl. Artificial Intelligence 17, 475–487.
Razmjoo, A., Xanthopoulos, P., Zheng, Q.P., 2017. Online feature importance ranking based on sensitivity analysis. Expert Syst. Appl. 85, 397–406.
Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learning Res. 5, 101–141.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575.
Rodriguez-Galiano, V., Luque-Espinar, J., Chica-Olmo, M., Mendes, M., 2018. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Sci. Total Environ. 624, 661–672.
Ruijters, E., Stoelinga, M., 2015. Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools. Comput. Sci. Rev. 15, 29–62.
Salvucci, D.D., Liu, A., 2002. The time course of a lane change: Driver control and eye-movement behavior. Transportation Res. Part F: Traffic Psychol. Behav. 5, 123–132.
Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.
Shi, Q., Abdel-aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transportation Res. Part C: Emerging 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.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinf. 9, 307.
Suh, J., Chae, H., Yi, K., 2018. Stochastic model-predictive control for lane change decision of automated driving vehicles. IEEE Trans. Veh. Technol. 67, 4771–4782.
Sun, D., Elefteriadou, L., 2010. Research and implementation of lane-changing model based on driver behavior. Transp. Res. Rec. 2161 (1), 1–10.
Sun, D., Elefteriadou, L., 2012. Lane-changing behavior on urban streets: An “in-vehicle” field experiment-based study. Comput.-Aided Civ. Infrastruct. Eng. 27 (7), 525–542.
Sun, D., Elefteriadou, L., 2014. A driver behavior-based lane-changing model for urban arterial streets. Transportation Sci. 48 (2), 184–205.
Tang, J., Alelyani, S., Liu, H., 2014. Feature selection for classification: A review. Data Classification: Algorithms and Applications 37.
Tomek, I., 1976. Two modifications of CNN. IEEE Trans. Syst., Man Cybernetics 6, 769–772.
Wang, C., Sun, Q., Fu, R., Li, Z., Zhang, Q., 2018. Lane change warning threshold based on driver perception characteristics. Accid. Anal. Prev. 117, 164–174.
Wang, L., Abdel-aty, M., Lee, J., Shi, Q., 2019. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. Accid. Anal. Prev. 122, 378–384.
Wilcox, R.R., 2001. Modern insights about Pearson’s correlation and least squares regression. Int. J. Selection Assessment 9, 195–205.
Wilson, D.L., 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst., Man, Cybernetics 408–421.
Xie, D.-F., Fang, Z.-Z., Jia, B., He, Z., 2019. A data-driven lane-changing model based on deep learning. Transportation Res. Part C: Emerging Technol. 106, 41–60.
Yan, F., Eilers, M., Baumann, M., Luedtke, A., 2016. Development of a lane change assistance system adapting to driver's uncertainty during decision-making. In: Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 93–98.
Yang, D., Zheng, S., Wen, C., Jin, P.J., Ran, B., 2018a. A dynamic lane-changing trajectory planning model for automated vehicles. Transportation Res. Part C: Emerging Technol. 95, 228–247.
Yang, K., Wang, X., Yu, R., 2018b. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. Transportation Res. Part C: Emerging Technol. 96, 192–207.
Yang, M., Wang, X., Quddus, M., 2019. Examining lane change gap acceptance, duration and impact using naturalistic driving data. Transportation Res. Part C: Emerging Technol. 104, 317–331.
Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 856–863.
Yun, M., Zhao, J., Zhao, J., Weng, X., Yang, X., 2017. Impact of in-vehicle navigation information on lane-change behavior in urban expressway diverge segments. Accid. Anal. Prev. 106, 53–66.
Zhang, X., Sun, J., Qi, X., Sun, J., 2019. Simultaneous modeling of car-following and lane-changing behaviors using deep learning. Transportation Res. Part C: Emerging Technol. 104, 287–304.
Zhang, Y., Wang, S., Phillips, P., Ji, G., 2014. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowl.-Based Syst. 64, 22–31.