Crash risk analysis for the mixed traffic flow with human-driven and connected and autonomous vehicles

Qing-Long Lu, Kui Yang* and Constantinos Antoniou

***Abstract*— In the near future, traditional or low-automation vehicles will share the roads with Connected and Autonomous Vehicles (CAVs) over many years. Yet, this complexity may impose new unknowns on the real-time crash risk evaluation. Consequently, it is important to explore crash risk analysis in such kind of mixed traffic flow environments. This paper constructed several special traffic variables in mixed traffic flow environments and proposed the kernel logistic regression (KLR) model to evaluate the crash risk in real-time. A simulated urban expressway corridor based on the North-South Elevated Road in Shanghai, China, is developed in SUMO, for the purpose of collecting the traffic safety data and traffic data (i.e., virtual detector data and Global Navigation Satellite System (GNSS) data) in mixed traffic flow environments. The prediction performance of KLR models was tested and analyzed with the simulated data, and is also compared with that of support vector machines (SVM) models. The results show that KLR has a good prediction performance like SVM. Considering KLR can provide probability estimates directly and can naturally extend to multi-class classification, priority should be given to KLR in similar problems, especially when crash risk is classified into multiple levels. The proposed KLR model is therefore recommended and has the potential to evaluate the real-time crash risk in the mixed traffic flow environment.**

Keywords–crash risk analysis, mixed traffic flow, kernel logistic regression, GNSS data

I. INTRODUCTION

Enormous efforts have been made for several years to investigate the technologies required by Connected and Autonomous Vehicles (CAVs) and cooperative driving systems (i.e., V2X or vehicle-to-all systems), which are emerging concepts to reduce traffic congestion, enhance traffic safety, improve traffic efficiencies, and etc. However, most of these studies focus on the scenarios with the CAVs' full penetration. Recently, more and more people believe that this kind of scenario will take unexpected time to achieve. Traditional or low-automation vehicles will therefore share the network space with these CAVs over many years. Besides, real-time crash risk evaluation is a research hot-spot to employ real-time traffic data to evaluate the crash risk in the road network and then identify when and where a crash is likely to occur for the further proactive traffic managements. However, most previous crash risk analyses apply the data from the monitoring devices installed fixedly in the road, such as loop detectors [1][2], automatic vehicle identification [3], and etc. Consequently, it is important to explore crash risk analysis in such kind of mixed traffic flow environments, where the data sources are different.

Logistic regression (LR) models are widely applied in previous crash risk evaluation studies, such as [1][2][4][5]. These traditional logistic regression models are built on the assumptions about the distribution of data and a well-defined function between the dependent variable and independent predictors [4]. When these basic assumptions were not met, inefficient estimations and incorrect inferences would be produced [6]. To this end, this paper proposes a kernel logistic regression model (KLR) to develop the relationship between the safety state and traffic variables.

The contents of the paper will be structured as follows. First, the overall methodology will be introduced, including the KLR model, feature construction and selection, and evaluation criterion. Then the simulation experiment, including scenario design, and network development and calibration are presented. Afterwards, the results about the feature selection, model development and performance are described and analyzed. It should be noted that the support vector machines (SVMs) and standard logistic regression models are also applied to compare with the proposed KLR models. Finally, a conclusion is given, focusing on the main findings but also limitations and future directions.

II. METHODOLOGY

A. Kernel Logistic Regression Model

A danger recognition system can be modeled as a supervised binary classification problem. A calibrated classification model for crash risk prediction needs to evaluate the safety level of a driving situation based on the vicinity traffic states, surrounding environment, etc. Training data used to estimate and evaluate the model include the data for both dangerous events and normal safe driving in the corresponding scenario. Let $\mathbf{X} \in \mathbb{R}^{n \times n_d}$ be a feature matrix where n is the number of samples and n_d is the number of features. Let $\mathbf{y}$ be a binary vector marking the class of samples. For an instance $\mathbf{x}_i \in \mathbb{R}^{n_d}$ (a row in $\mathbf{X}$), the potential label is either y_i=1 or y_i=0. In this study, y_i=1 corresponds to the occurrence of a dangerous event, while y_i=0 labels the normal driving sample. A classifier is thus to decide the instance $\mathbf{x}_i$ to be dangerous or safe based on its attributes. LR, a probabilistic statistical method, which has been proven to be a powerful classifier, solves the classification problem by optimizing

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \ln(1 + e^{-y_i f(\mathbf{x}_i)}) \qquad (1)$$

This work was supported by the European Unions Horizon 2020 research and innovation programme iDREAMS under grant agreement No 814761.

Q.L. Lu, K. Yang and C. Antoniou are with the Chair of Transportation Systems Engineering, Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Germany.

*corresponding author: kui.yang@tum.de

where C is a regularization parameter, $\mathbf{w}$ is the vector of parameters to be determined. $f(\mathbf{x})$ is given by

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \tag{2}$$

According to the representer theorem, the optimal $\mathbf{w}$ can be written as

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{3}$$

LR, however, tends to be biased towards the majority class [14], so its kernel version will be used in this study, which has shown performance as good as SVM. In kernel-based classification methods, an input vector $\mathbf{x}$ will be mapped into the Hilbert spaces generated by a positive definite kernel K, i.e., $\mathbf{x} \mapsto \varphi(\mathbf{x})$ where $\varphi : \mathbb{R}^{n_d} \mapsto \mathbb{R}^{n_d^{(\varphi)}}$. However, it is not required to do the actual mapping $\varphi(\cdot)$ as it can be done implicitly by K when $\varphi(\mathbf{x})^T\varphi(\hat{\mathbf{x}})$ can be computed as $K(\mathbf{x}, \hat{\mathbf{x}})$. The kernel version of Equation (3) is given by

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \varphi(\mathbf{x}_i) \tag{4}$$

Combine Equation (2) and (4) gives the optimal f(x):

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

Hence, fitting a KLR is equivalent to a finite-dimensional convex programming problem descried as

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j \tilde{K}(\mathbf{x}_i, \mathbf{x}_j) + C \sum_{i=1}^{n} \ln(1+e^{-\sum_j \alpha_j \tilde{K}(\mathbf{x}_i, \mathbf{x}_j)}) \tag{6}$$

where $\tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$. This problem can be solved efficiently by the Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm presented in [16].

Note that, $\ln(1 + e^{-\sum_j \alpha_j \tilde{K}(\mathbf{x}_i, \mathbf{x}_j)})$ is the negative log-likelihood (NLL) associated with the probabilistic model

$$P(y|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + e^{-\sum_j \alpha_j \tilde{K}(\mathbf{x}_i, \mathbf{x}_j)}} \tag{7}$$

In this study, we apply the Radial Basis Function (RBF) kernel given as below.

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2} \tag{8}$$

where γ is a positive parameter reflecting the influence of a single training sample. Generally, γ is set to be $1/n$.

B. Event Definition Based on Safety Surrogate Measures

Due to the lack of field data of mixed traffic, most studies investigating the safety effect of CAVs used surrogate safety measures (SSM) to evaluate the crash risk in the mixed traffic context (e.g., [10], [11]). Time-to-collision (TTC) is one of the most popular indicators for rear-end crash risk assessment. TTC reflects the time for a faster follower to crash into a leader if their relative speed stays unchanged in follow-lead situations, which is calculated as below.

$$TTC = \begin{cases} \frac{g_l - g_f - L}{v_f - v_l} & \text{if } v_f > v_l \\ \infty & \text{otherwise} \end{cases} \tag{9}$$

where g_l, g_f are the longitudinal location of the leader and follower, respectively, while v_l and v_f are the corresponding speeds. L is the length of the following vehicle.

On the other hand, for lateral movement, we consider the critical moment when the vehicle finishing lane-changing and check if its distance to the proceeding vehicle and the following vehicle (on the target lane) is safe. Here we define a novel indicator named lane-changing distance difference ratio (DDR) to measure the risk of lateral movements. DDR is calculated as

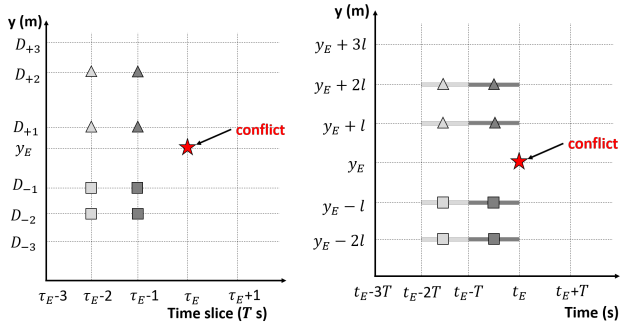
$$DDR = \min\left(\frac{d_f - d_f^*}{d_f}, \frac{d_l - d_l^*}{d_l}\right) \tag{10}$$

where d_l, d_f are the distance of the ego vehicle to its leader and follower right after it finishes lane-changing. d_l^* (d_f^*) is the corresponding required secured longitudinal gap to the nearest leader (follower) to fulfill deceleration constraints.

A smaller TTC or DDR indicates a more hazardous situation. Referring to [10] and [11], the TTC threshold for recording a dangerous situation is set to be 2 s in this study. And the DDR threshold is set to be -0.12.

C. Feature Construction

In previous studies on crash risk analysis (e.g., [12],[13]), loop detector data (e.g., speed, volume) are mostly used when estimating crash predicting models and distinguishing precursor crash conditions. In general, the detector data are aggregated in a T-second interval to eliminate the randomness. It has been validated that the closest two data points in either temporal dimension or spatial dimension on the upstream and downstream of the crash are most significant. Therefore, 8 data slices are matched with a dangerous event, as specifically shown in Figure 1a. For instance, for a crash happened at 07:33:21, when T is 30, data of the closest two detectors on the upstream/downstream during [07:32:00, 07:32:30) and [07:32:30, 07:33:00) are collected as candidate explanatory variables for the corresponding event. Note that, weather conditions and geometric features are not explicitly considered here. However, one can easily extend the model with these variables in any cases if they are available.



(a) Data points of detector data (b) Data points of GPS data

Fig. 1. Conceptual definition of data points

In the era of mixed traffic with human driving vehicles (HVs) and CAVs, however, data sources are not limited to detectors installed on the network. Vehicle-to-infrastructure (V2I), vehicle-to-vehicle (V2V) and GNSS devices (without

loss of generality, we take Global Positioning System (GPS) as example hereafter) assembled to CAVs provide the possibility to catch the exact speed and location of CAVs and the relative distance between them. Consequently, from the GPS data of CAVs, we can extract similar variables as those refined from loop detector data. Clearly, assuming the control center has the access to the GPS data in real-time, we can construct finer and more accurate variables to represent the traffic states of upstream/downstream relevant to the sample. Figure 1b shows the definition of GPS-based variables. For the same event arisen at 07:33:21, the data point in any sub-segment in $\{[y_E-2l, y_e-l), [y_E-l, y_E), (y_E, y_E+l], (y_E+l, y_E+2l]\}$ during any time slice in $\{[07:32:21, 07:32:51), [07:32:51, 07:33:21)\}$ is extracted, resulting in 8 data points as well.

Furthermore, except the statistics (i.e., mean and standard deviation) of speed and volume of CAVs, each data point also contains the statistics of the distance between every two CAVs. One should know that the variables constructed on speed and volume can only capture the temporal characteristics of the traffic, while the variables developed on distance can explain the spatial characteristics that are ignored in the literature. Thus, GPS-based variables can capture both spatial and temporal factors that may lead to a dangerous event.

D. Feature Selection

Overall, we extract 32 detector-based and 32 GPS-based variables from the detector dataset and CAV GPS dataset, respectively. However, some of them are highly correlated and thus may lead to biased estimates. Besides, not all variables are significant in predicting dangerous events. Random Forest (RF) is an ensemble of randomized decision trees that has been widely used to evaluate the variable importance in preliminary before applying any learning methods to solve certain regression/classification problem. In this study, variables are selected based on Gini importance by leveraging RF. Algorithm 1 presents the procedure for selecting the most significant and uncorrelated variables.

Algorithm 1 Random Forest for variable selection

Candidate explanatory variable set: $\mathbb{X}$
Initialize selected variable set: $\mathbb{S} = \varnothing$
Define the minimum acceptable correlation: ρ_{min}
$i = 0$
for $i < n_v$ **do**
    $\tilde{x}_i = \text{argmax}_{x \in \mathbb{X}} \text{Gini}(\text{RandomForest}(\mathbb{X}))$
    $\mathbb{C} = \{x_j | corr(\tilde{x}_i, x_j) \geq \rho_{min}, x_j \in \mathbb{X}\}$
    Update $\mathbb{S} : \mathbb{S} = \mathbb{S} \cup \{\tilde{x}_i\}$
    Update $\mathbb{X} : \mathbb{X} = \mathbb{X}/\mathbb{C}$
    $i = i + 1$
end for

E. Evaluation Criterion

We apply the following metrics to evaluate the performance of the proposed model and make a comparison with other advanced models.

1) Balanced accuracy: The balanced accuracy (B-accuracy) of a binary classification problem is given by

$$B\text{-}accuracy = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{FP+TN}) \tag{11}$$

where TP, FN, TN and FP are true positive, false negative, true negative and false positive, respectively. It is worth mentioning that, B-accuracy ([0,1]) is essentially the arithmetic mean of the recall of each class, which implies two classes have the same weight and importance. Accuracy instead, depends on the performance achieved in the majority class, while the performance in one another is less important. Thus, B-accuracy is more preferable than accuracy in crash risk analysis.

2) F1 score:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{12}$$

F1 score ([0,1]) is calculated as the weighted average between precision and recall. It has been found that this metric imposes more importance on the smaller class and it mostly rewards models coming out with similar precision and recall.

3) AUC: As the most popular metric in classifier evaluation, the area under the Receiver Operating Characteristic curve (AUC) is also used to evaluate the models here. The Receiver Operating Characteristic (ROC) curve is a trade-off between the true positive rate (TPR) and false positive rate (FPR) with FPR on the x-axis and TPR on the y-axis. AUC ([0,1]) summarizes the information of the ROC curve in one number to represent the predictive performance of the model.

III. EXPERIMENT SETUP

A. Scenario and Simulation

To implement and evaluate the proposed modeling framework, we conduct the simulation on a 17 km segment of the North-South Elevated Road (from the connection with the Inner Ring Road to the connection with the Outer Ring Expressway) in Shanghai, China, with 3 lanes or 4 lanes at most sub-segments. To simplify the implementation, we only simulate the northbound corridor. Apart from the upstream and downstream of the trunk link, this segment also connects to 7 on-ramps and 12 off-ramps. Figure 2 shows the abstract layout of the simulated segment marking the relative location of on-ramps and off-ramps. It has a speed limit of 80 km/h (trucks are not considered). The North-South Elevated Road passing through the city center of Shanghai is one of the main trunk links crossing the Huangpu River and plays a significant role in daily commuting. Due to its function in distributing the high demand spatially, North-South Elevated Road has become a typical high-risk trunk link. High demand renders traffic congestion/jams, especially during peak hours, and thus increases the possibility of rear-end collisions. On the other hand, distributing the demand via the linked on-ramps and off-ramps induces more lane-changing actions, which also increases the crash risk. Hence, implementing experiments on this segment provides an opportunity to analyze the crash mechanism under complex traffic situations,

especially at the early stage of mixed traffic with HVs and CAVs.

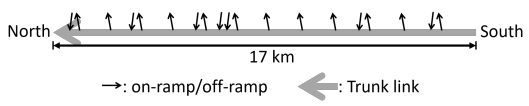


Fig. 2. The Silhouette of the simulation road segment.

The open-source simulator, Simulation of Urban MObility (SUMO [7]), is used to simulate the traffic of a day on the target road segment and collect the data as input to the proposed model. Corresponding to the deliverable 3.1 [8] of the Transition Areas for Infrastructure-Assisted Driving (TransAID) project, where the modeling and implementation of AVs and CAVs in SUMO are explicitly discussed, we distinguish HVs and CAVs in car-following behavior and lane-changing behavior[1]. It is worth mentioning that, we focus on the early stage of the operation of mixed traffic, where neither CAV exclusive lanes are available nor platooning driving is developed due to the low penetration rate. As a result, no external control algorithm is required to manipulate the driving of CAVs at each simulation step. We randomly select 1000 dangerous conflicts from the recorded event dataset and select 4 relevant normal driving cases for each dangerous driving event. After dropping the samples without CAV data being recorded and thus cannot construct the complete candidate explanatory variable pool, the final dataset contains 572 dangerous events and 2080 normal driving reference cases. Consider the mean distance (377 m) between every two detectors and the speed limit, we use T=30 and l=300 m when constructing the candidate explanatory dataset.

B. Network Calibration

To simulate the real traffic accurately, we calibrate the demand of different origin-destination pairs by using the traffic measurements recorded by loop detectors on the trunk link (measurements on on/off-ramps are not available) on June 3rd, 2016. The southernmost cross-section of the trunk link and 7 on-ramps are defined as origins, while the northernmost cross-section of the trunk link and 12 off-ramps are defined as destinations, which results in 56 OD pairs. There are 45 detecting locations. Traffic measurements, including speed and volume, are aggregated into a one-hour interval for each detector. The popular Simultaneous Perturbation Stochastic Approximation (SPSA [9]) algorithm, which has been widely applied to solve the dynamic demand estimation problem in the transportation community attributed to its efficiency on large-scale problems, is used to address the demand calibration task in this study with the following objective function.

$$\min w_1 f(\hat{\boldsymbol{s}}^t, \boldsymbol{s}^t) + w_2 f(\hat{\boldsymbol{q}}^t, \boldsymbol{q}^t) \tag{13}$$

where $\boldsymbol{s}^t$ and $\boldsymbol{q}^t$ are the vector of observed mean speed and total volume in interval t, respectively, while $\hat{\boldsymbol{s}}^t$ and $\hat{\boldsymbol{q}}^t$ are the corresponding simulated vectors. $f(\hat{\boldsymbol{z}}, \boldsymbol{z})$ is used to evaluate the goodness-of-fit. Here we apply the Root Mean Square Normalized error (RMSN), which is defined as below.

$$f(\hat{\boldsymbol{z}}, \boldsymbol{z}) = \frac{\sqrt{n \sum_{i=1}^{n} (\hat{z}_i - z_i)^2}}{\sum_{i=1}^{n} z_i} \tag{14}$$

By calibrating the demand for different routes (ODs), we expect to generate similar car-following and lane-changing maneuvers as in real traffic. With the fact that lane-changing behavior is a main trigger for traffic accidents, it is critical to generate realistic lane-changing maneuvers in crash risk analysis.

We set w_1=w_2=0.5. Figure 3 compares calibrated values with true measurements. The calibrated network ultimately leads to an average RMSN in 22 hours (the first and last hour are used for traffic warm-up and dissipation process) of 29%, which is capable of recreating the real traffic properly.

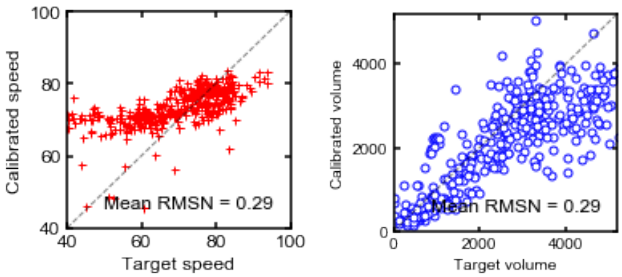


Fig. 3. Comparison of target values and calibrated values

IV. RESULTS

A. Feature Evaluation

As described in Section II-D, by applying Algorithm 1, we want to select the variables with most Gini importance but slightly correlated to compose the explanatory dataset for the binary classification model. In this section, first, we deploy the first 10 variables ranked by Gini importance in Figure 4. The superscript of the variable indicates the data source, and the subscript indicates the spatial-temporal sub-segment and the feature (only appropriate to standard deviations). For example, $\sigma^{(gps)}_{d,(-2,-2)}$ is the standard deviation of distance between CAVs in the second sub-segment on the upstream during the time slice before the previous slice; $\sigma^{(det)}_{s,(+1,-2)}$ is the standard deviation of vehicle speeds recorded by the detector in the first sub-segment on the downstream during the time slice before the previous time slice. Surprisingly, all of them are extracted from the CAV GPS dataset, which implies the necessity and benefit of introducing GPS information into consideration when constructing variables to model the upstream and downstream traffic states ahead of the happening of events.

However, considering the potential multicollinearity between variables, whenever we pick one variable, we have to drop the highly correlated variables. In this study, 0.3 is used as the correlation threshold ρ_{min}. This selecting principle, ultimately, recommends the variables listed in Table I. Four of them are extracted from CAV GPS data. And two of them are distance-based variables, while the rest three are speed-based. This validates the aforementioned expectation,

[1]Refers to Table 26 and Table 28 in [8] for the details of the modeling of HVs and CAVs in SUMO.

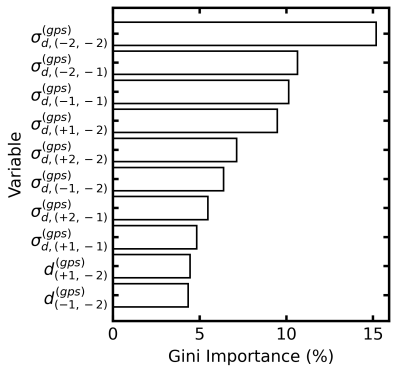


Fig. 4. Feature Gini importance

i.e., distance-based variables can also provide very useful auxiliary information when model traffic states as it can capture the spatial correlation of traffic. Furthermore, standard deviations are reserved, evidencing that incongruous traffic is more likely to produce an accident. Table I summarizes the statistics of the selected important variables, where the minimum value equaling zero means only one vehicle in the relevant spatial-temporal sub-segment. And the correlation matrix is given in Table II, which shows that there are no significantly statistical correlations between these variables.

TABLE I
STATISTICS OF SELECTED VARIABLES

Variable	Mean	Min	50%	Max	Gini importance
$\sigma^{(gps)}_{d,(-2,-2)}$	49.44	0.00	48.80	129.07	15.89%
$\sigma^{(gps)}_{d,(-1,-1)}$	48.01	0.00	48.56	90.07	8.54%
$\sigma^{(gps)}_{s,(-1,-1)}$	2.32	0.00	1.97	11.12	0.23%
$\sigma^{(gps)}_{s,(+2,-2)}$	1.84	0.00	1.62	15.98	0.12%
$\sigma^{(det)}_{s,(+1,-2)}$	2.02	0.02	1.79	10.39	0.06%

TABLE II
CORRELATION MATRIX OF SELECTED VARIABLES

Variable	①	②	③	④	⑤
$\sigma^{(gps)}_{d,(-2,-2)}$ ①	1				
$\sigma^{(gps)}_{d,(-1,-1)}$ ②	0.29	1			
$\sigma^{(gps)}_{s,(-1,-1)}$ ③	0.16	0.23	1		
$\sigma^{(gps)}_{s,(+2,-2)}$ ④	0.05	0.04	-0.02	1	
$\sigma^{(det)}_{s,(+1,-2)}$ ⑤	-0.01	0.02	0.05	0.08	1

B. Model Comparison

In this section, we develop and test the proposed KLR model, support vector machine (SVM) models and the conventional logistic regression (LR) model, and then compare their prediction performances. SVM with two typical kernels, linear (SVM-Linear) and radial basis function (SVM-RBF), are considered. First, from Figure 5, we can see that SVM-RBF and the proposed KLR results in almost the same figures. While SVM-RBF has been validated to be an effective model in crash risk analysis, we can claim that the KLR model is also an appropriate alternative in similar tasks. On the other hand, SVM-Linear and LR show relatively lower precision, reflected as more FP.

SVM-Linear		Predicted	
		N	P
Actual	N	335	3
	P	8	110

SVM-RBF		Predicted	
		N	P
Actual	N	337	1
	P	3	115

LR		Predicted	
		N	P
Actual	N	335	3
	P	14	104

KLR-RBF		Predicted	
		N	P
Actual	N	337	1
	P	4	114

Fig. 5. Confusion matrix of different models

In addition, Table III (five-fold cross validation) includes the metrics for evaluating models in a more systematic way, within which B-accuracy and F1 score are directly computed from the confusion matrix, while AUC is calculated based on the ROC curve. SVM-RBF performs best among the four models in B-accuracy and F1 score, reaching 97.99% and 96.32%, respectively, while LR performs worst (94.21% and 93.14%, respectively). KLR-RBF has a similar B-accuracy (97.26%) as that of SVM-RBF, but its F1 score shows an apparent deficiency compared to the SVMs. As we explain in Section II-E, B-accuracy gives the same weight to two classes, while F1 score emphasizes more on the smaller class. This means, KLR can perform as good as the best SVM if the problem imposes the same weights on two classes, but SVMs work better if the smaller class was emphasized. Regarding the AUC, similarly, SVM-RBF and KLR get close results, while SVM-Linear and LR are behind them. Figure 6 illustrates the ROC curves of these models. Furthermore, to measure the effectiveness of the feature selecting procedure, we also apply KLR on the first five principal components (PCs) constructed by the principal component analysis (PCA) algorithm. The result shows that, KLR-RBF and KLR-RBF(PCA) are almost the same in B-accuracy and AUC, though the latter outperforms the former in F1 score. That is to say, the selected variables can capture as much variance of the entire candidate explanatory dataset as the first five orthogonal (uncorrelated) PCs.

TABLE III
MODEL COMPARISON

Model	B-accuracy	F1 score	AUC
SVM-Linear	0.9565	0.9454	0.9614
SVM-RBF	0.9799	0.9632	0.9926
LR	0.9421	0.9314	0.9696
KLR-RBF	0.9726	0.9395	0.9927
KLR-RBF(PCA)	0.9735	0.9684	0.9978

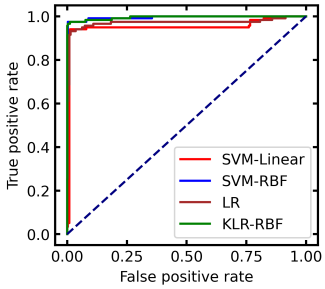


Fig. 6. The ROC curves of different models

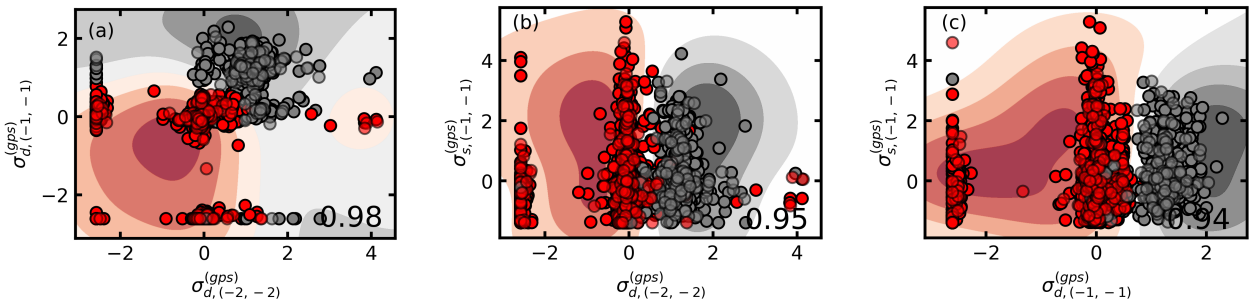


Fig. 7. KLR-RBF classifier performance

C. KLR Performance Analysis

Figure 7 demonstrates scatter plots of two of the first three variables listed in Table I and the decision boundaries of KLR. Besides, the B-accuracy is also provided at the right-bottom corner. As we can see from the figure, (a) shows an apparent non-linear boundary that separates two classes, while (b) and (c) have nearly linear boundaries (this also explains why SVM-Linear and LR can also perform well in this task). Clearly, KLR can distinguish two classes effectively with either non-linear or linear boundaries, which is also supported by high B-accuracy values. Thus, one can expect a good performance of KLR in various classification problems.

V. CONCLUSIONS

Since V2I, V2V and GNSS devices assembled to CAVs provide more diverse data, such as real-time speed, location and the relative distance between them, it is necessary to explore new methodologies to evaluate the real-time crash risk in the mixed traffic flow environment. This paper developed a simulated urban expressway corridor in SUMO and collected the traffic safety data and traffic data in mixed traffic flow environments. After constructing several special traffic feature variables, the important variables are identified by the random forest. It found that the variables from GPS are more significant than those from loop detectors. With the help of the important variables, the kernal logistic regression model based on radial basis function (KLR-RBF), support vector machines (SVM) models and standard logistic regression models are introduced to tackle the real-time crash risk evaluation tasks. After validation and comparison, the results show that KLR-RBF has a good prediction performance like SVM. And KLR-RBF can distinguish safe traffic flow and dangerous traffic flow effectively with either non-linear or linear boundaries. Still, this research does not come without limitations. The dataset did not include the existing variables on human drivers demographics, attitudes and perceptions. Future work should also focus on active traffic managements based on the real-time crash risk evaluation results.

REFERENCES

[1] Abdel-Aty, M., Uddin, N. and Pande, A., 2005. Split models for predicting multivehicle crashes during high-speed and low-speed operating conditions on freeways. Transportation research record, 1908(1), pp.51-58.

[2] Yang, K., Wang, X. and Yu, R., 2018. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. Transportation research part C: emerging technologies, 96, pp.192-207.

[3] Ahmed, M.M., Abdel-Aty, M. and Yu, R., 2012. Bayesian updating approach for real-time safety evaluation with automatic vehicle identification data. Transportation research record, 2280(1), pp.60-67.

[4] Xu, C., Wang, W. and Liu, P., 2012. A genetic programming model for real-time crash prediction on freeways. IEEE Transactions on Intelligent Transportation Systems, 14(2), pp.574-586.

[5] Xu, C., Wang, W. and Liu, P., 2013. Identifying crash-prone traffic conditions under different weather on freeways. Journal of safety research, 46, pp.135-144.

[6] Delen, D., Sharda, R. and Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis & Prevention, 38(3), pp.434-444.

[7] Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Fltterd, Y.P., Hilbrich, R., Lcken, L., Rummel, J., Wagner, P. and Wiener, E., 2018, November. Microscopic traffic simulation using sumo. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 2575-2582). IEEE.

[8] Mintsis, E., Koutras, D., Porfyri, K., Mitsakis, E., Lcken, L., Erdmann, J., Fltterd, Y.P., Alms, R., Rondinone, M., Maerivoet, S. and Carlier, K., 2019. TransAID Deliverable 3.1-Modelling, simulation and assessment of vehicle automations and automated vehicles' driver behaviour in mixed traffic-iteration 2.

[9] Spall, J.C., 1998. Implementation of the simultaneous perturbation algorithm for stochastic optimization. IEEE Transactions on aerospace and electronic systems, 34(3), pp.817-823.

[10] Papadoulis, A., Quddus, M. and Imprialou, M., 2019. Evaluating the safety impact of connected and autonomous vehicles on motorways. Accident Analysis & Prevention, 124, pp.12-22.

[11] Zhang, J., Wu, K., Cheng, M., Yang, M., Cheng, Y. and Li, S., 2020. Safety Evaluation for Connected and Autonomous Vehicles Exclusive Lanes considering Penetrate Ratios and Impact of Trucks Using Surrogate Safety Measures. Journal of advanced transportation, 2020.

[12] Sun, J. and Sun, J., 2015. A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. Transportation Research Part C: Emerging Technologies, 54, pp.176-186.

[13] Yu, R., Wang, X., Yang, K. and Abdel-Aty, M., 2016. Crash risk analysis for Shanghai urban expressways: a Bayesian semi-parametric modeling approach. Accident Analysis & Prevention, 95, pp.495-502.

[14] Maalouf, M. and Trafalis, T.B., 2011. Robust weighted kernel logistic regression in imbalanced and rare events data. Computational Statistics & Data Analysis, 55(1), pp.168-183.

[15] Zhu, J. and Hastie, T., 2005. Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics, 14(1), pp.185-205.

[16] Liu, D.C. and Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Mathematical programming, 45(1), pp.503-528.