Rough Set Based Method for Vehicle Collision Risk Assessment Through Inferring Driver's Braking Actions in Near-Crash Situations

Liqun Peng
The School of Transportation and Logistics at East China Jiaotong University, Nanchang, 330013, China. E-mail: lq.peng@ecjtu.jx.cn

Miguel Angel Sotelo
Professor at the Department of Computer Engineering. University of Alcalá, Alcalá de Henares (Madrid), Spain. E-mail: miguel.sotelo@uah.es

Yi He
California PATH, University of California, Berkeley, CA, USA. E-mail: yihe@berkeley.edu

Yunfei Ai
National Engineering Laboratory for Transportation Safety and Emergency Informatics, Beijing 100011, China. E-mail: aiyunfei@cttic.cn

Zhixiong Li
School of Mechatronic Engineering, Ocean University of China, Qingdao 266110, China; Also with School of Mechanical, Materials, Mechatronic and Biomedical Engineering, University of Wollongong, Wollongong, NSW 2522, Australia. E-mail: zhixiong.li@ieee.org

©ISTOCKPHOTO.COM/MICROVONE

Digital Object Identifier 10.1109/MITS.2019.2903539
Date of publication: 18 March 2019

1939-1390/19©2019IEEE

Abstract—Driving information and data under potential vehicle crashes create opportunities for extensive real-world observations of driver behaviors and relevant factors that significantly influence the driving safety in emergency scenarios. Furthermore, the availability of such data also enhances the collision avoidance systems (CASs) by evaluating driver's actions in near-crash scenarios and providing timely warnings. These applications motivate the need for heuristic tools capable of interpreting the correlations of driving risk with driver/vehicle characteristics and incidental traffic factors. In this paper, we acquired amount of real-world field data and built a comprehensive "driver-vehicle-road" dataset for actual driver behavior evaluation. The proposed method works in two steps. In the first step, a variable precision rough set (VPRS) based classifier derives a simplified decision rules from field driving dataset, which presents the essential attributes relevant to driving safety. In the second step, we quantify the mutual information entropy of each attribute to evaluate the significance of different factors on happening a vehicle crash, then an accumulation of weighted "driver-vehicle-road" is calculated to achieve an index reflecting the driving safety level. The performance of the proposed method is demonstrated in an offline analysis of the driving data collected from field trials, where the goal is to infer the emergency braking actions in next short term. The results indicate that our proposed model is a good alternative for providing drivers immediate warnings with high prediction accuracy and stability.

I. Introduction

Driving safety is a high priority issue for governmental agencies, the majority of vehicle manufacturers and other stakeholders. In order to enhance the safety for both the drivers and pedestrian, a number of improvements have been proceeded, ranging from enhancement of infrastructure to vehicle-based safety systems. Advanced driver assistance systems (ADAS) techniques have made a large-scale sensor embedded vehicle study to collect amounts of driving data on the actual road to investigate the relationship between the emergency driving safety and driver maneuvers (e.g., Acceleration/deceleration, and steering) [1]. This approach was described as most helpful in revealing driver behavior and vehicle crash causation mechanism. With access to field driving data, the safety related events could be observed and measured more precisely.

Problem motivation: Effective ADAS requires awareness of actual driving situation, a reliable assessment of the vehicle crash risks, and making rapid decisions on assisting actions [2, 3]. On the one hand, understanding of the multi-factors on road, especially the driver behavior, will remarkably improve the vehicle crash risk assessment. For example, if the speed of the vehicle under study is 90 km/h and the relative distance from the vehicle ahead is 50 m, the acceleration volition would be considered as dangerous/risky, conversely, if the driving volition is slow-down, the risk level is low and therefore the action should be considered as not dangerous. On the other hand, although a variety of roadside and vehicular onboard sensors are capable of collecting a large-scale information, it is still needed to be considered whether all these collected data are appropriate for traffic safety applications. The inclusion of abundant factors for crash detection may lead to overfitting actual driving safety and making false warnings to drivers.

Approaches for vehicle collision risk assessment: Driving safety problems involve with complex interactions between the driver's perceptual and decisional contribution, vehicle motion and incidental effects under varying traffic environment. However, the nature of instant driver behavior under emergency situations has hardly been estimated in previous studies, especially when considering the entire complexity of scenarios in the context of driving. Furthermore, the less significant characteristics of driver behavior, vehicle or traffic information take negative efforts for driving safety analysis. From this perspective, it appears that existing methods integrating multiple factors to judge safety is not satisfied enough to realistically model real-world driving safety issues.

In this work, we investigate the actual driving behaviors in near-crash events as well as the involved interactions among "driver-vehicle-road" multi factors. The field driving data was collected under potential threats in dynamic traffic, then a comprehensive dataset is built to record all related factors as a whole. The extent of near-crashes are profiled into different risk levels, namely high deceleration, medium deceleration, and low deceleration respectively, based on intensity of braking process features. Therefore, given the set of aforementioned braking process characteristics drivers may take is pre-established, the problem of evaluating the driving safety at the current instant can be seen as a classification task to infer the most suitable driver behavior in the next short term. Once the upcoming dangerous driving events are detected, we are able to assist driver to adapt more safety maneuver. This latter step has been widely implemented following [1]–[3], which are out of the scope of this work.

Effective ADAS requires awareness of actual driving situation, a reliable assessment of the vehicle crash risks, and making rapid decisions on assisting actions.

In the present paper, we propose a rough set based model to assess the vehicle collision risk involved in the emergency events. The emergency near-crash events are identified by learning the actual driving data from the field experiments. We advocate the use of rough set theory for addressing the noisy and imprecise issue of these data commonly collected by test vehicle equipped sensors, such as accelerometers and gyroscopes. Rough set theory has been leveraged to perform data sorting with this kind of information, which has been fully described in Section "IV. Modeling process". Rough set models are mainly used for interpretation of data dependencies, the estimation of the significance of attributes, and the reduction of all redundant samples and attributes to a minimal representative set of attributes. Therefore, given an inputs-set composed of data collected from the previously mentioned sensors, the function of the rough set is to simplify the induced decision rules without reducing the classification accuracy.

The main contribution of this work is the elaboration of applying rough set model for driving risk assessment. More accurately, the set of driving safety related attributes is reduced to a minimal set, which represents the most critical attributes that should be taken into consideration for driving safety assessment. Such minimal subset is then used during the training and classification phases to generate a set of decision rules for the prediction of the emergency driver actions in near-crash scenarios. As it will be seen, in the light of the comparison with other algorithms, this accuracy aim has been achieved. Besides, instead of using expert knowledge, rough set processes the problem by investigating the sample data and estimating the conditional probabilities related to a special many-valued logic. While other inference models, such as fuzzy set, which is also widely applied for modeling vagueness and uncertainty issues, classify whether driver behavior into safe or risky status mainly based on subjectively assigning membership function value [4]. In this case, no additional operators are predefined and expert knowledge are referred to define rough set operators [5].

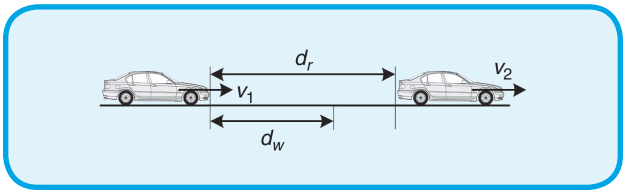

FIG 1 Safety distance analysis scheme.

The reminder of this paper is organized as follows. Section II presents a brief overview of the related research on vehicle collision avoidance system. Section III introduces a field test and data collection, including experiment design, data processing and driving risk definition. Section IV describes the modeling process for vehicle crash risk assessment. The model evaluation and test results are illustrated in Section V. Finally, the main conclusions and the future work are discussed in Section VI.

II. Related Work

A variety of driving safety assessment have been explored. Some have evaluated the safety issue based on real-time vehicle kinematics. Others have comprehensively tried to monitor the D-V-E (driver, vehicle, and environment) statues. In this study, we account for vehicle crash risk with more complex scenarios and factors.

A. Crash Risk Assessment Based on Vehicle Kinematics

Safety distance (SD) model is one of the most important methods in identification of longitudinal crash risk [2]. As presented in Fig. 1, where v_1 and v_2 are the longitudinal velocity of the following and preceding vehicles respectively, d_r is the gap between them, it is intended to calculate the critical warning distance d_w, which could be expressed as the general function form as follows:

$$d_w = f(v_{\mathrm{rel}}, v_1, \alpha_1, \alpha_2, \tau) + d_0 \tag{1}$$

Where v_{rel} is the relative velocity between the following and preceding vehicles, τ is the delay, and α_1 and α_2 are the maximum deceleration of the following and leading vehicles respectively, d_0 is headway offset, the variables represented above are comprehensively taken into consideration when evaluating the safety distance d_w. Then, by comparing the measured headway d_r with the safety distance d_w, the vehicle will be in the safe driving situation when $d_r > d_w$, while the following vehicle should be warned or decelerated to avoid a crash if $d_r < d_w$. The safety distance model could be transformed into time to collision (TTC) model [6, 7]. Both SD and TTC models have been extensively applied in many modern developed in-vehicle safety systems based on Information and Communication Technology [8, 9]. Such systems have been expected to support the driver to maintain safe speed and headway in all driving situations by providing timely warning to driver when a critical safety situation emerges.

Other researchers developed two dimensions algorithms based on dynamic equations of vehicle motion. Considering the vehicles points in space, as present in Fig. 2, implies that collision risk exist when

$$\sqrt{(x_1(t_1)-x_2(t_2))^2+(y_1(t_1)-y_2(t_2))^2} \le \sum_{i=1}^{2} R_i \qquad (2)$$

where

$$R_i = g(x_i,\, y_i,\, \vec{v_i},\, L_i,\, W_i)$$

L_i and W_i is respectively the length and width of the subject vehicle and the obstacle. The safety zones algorithm creates a safety virtual zone around vehicles and detects the overlap areas between the subject vehicle and each approaching obstacle to indicate collision danger [10, 11]. However, the algorithms based on vehicle kinematics are very susceptible to generate false warnings, especially when driver behavior is ignored in analyzing complex traffic scenarios.

However, aforementioned studies were typically done by setting absolute thresholds on the vehicle kinematics measurements, without taking account of the relationship between the crash risk severity and detailed driving maneuver and (e.g., constant speed, acceleration, braking, and steering).

B. Crash Risk Assessment Based on D-V-E Arrangement

It is well known that driving involves complex interactions between the driver, the vehicle and the environment under varying conditions (road characteristics and properties, weather, incidental effects, etc). The D-V-E (driver, vehicle, traffic environment) factors have been generally considered to be the most important factors in crash occurrence [12, 13]. Hence, It is necessary to develop reasoning models, as shown in Fig. 3, which integrate the main constituents of driving situation with generic phases of completing driving, i.e. perception, analysis, decision making and action. Such a reasoning model will improve the current development of driving safety analysis.

Naturalistic driving studies provide an opportunity to more precisely observe and measure safety-related events [14]–[16]. In these studies, the driver's factor was fully considered as one of the precipitating and contributing factors of crashes and provided the critical exposure of pre-crash data. Naturalistic driving studies have recorded a large-scale field data, which in turn, could provide a useful supplement to effectively control laboratory and field studies to further enhance the understanding of the effects of driver characteristics on traffic safety [17]–[19].

Other studies presented a wider survey of the D-V-E arrangement, taking into consideration driving safety related factors, such as obstacle detection, driver intention, real-time weather and roadway geometry [20]–[22]. For instance, Hassan et al., [23] used the structural equation modeling approach to explore significant factors associated with young drivers' involvement in at-fault crashes. It was revealed in the study that, aggressive violations, in-vehicle distractions and demographic characteristics were significant factors affecting young drivers' involvement in at-fault crashes. Ahmed et al., [24] also assessed the effectiveness of the weather on real-time road crash risk in locations with recurrent fog problems. Although some studies have made effort to address these issues by combining multiple elements (e.g., detecting the driving context, analysis of conditions, and proposing actions), their number and genuine contribution were relatively low.

C. Crash Risk Assessment Inference Reasoning Model

Machine learning methods have been widely studied to evaluate vehicle crash risk in dynamic traffic. [25] illustrated

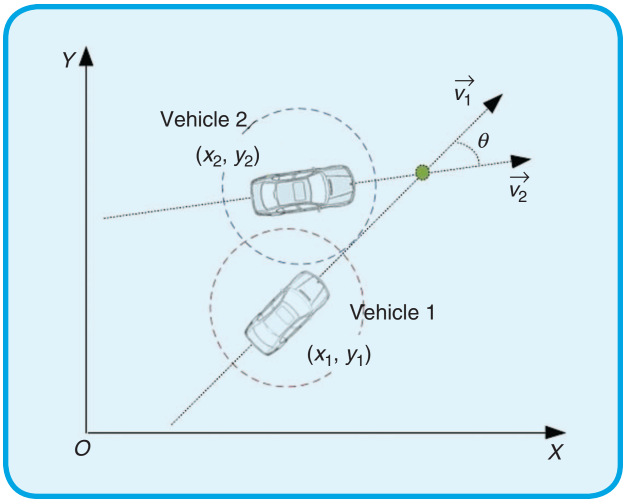

FIG 2 Two dimensions analysis scheme.

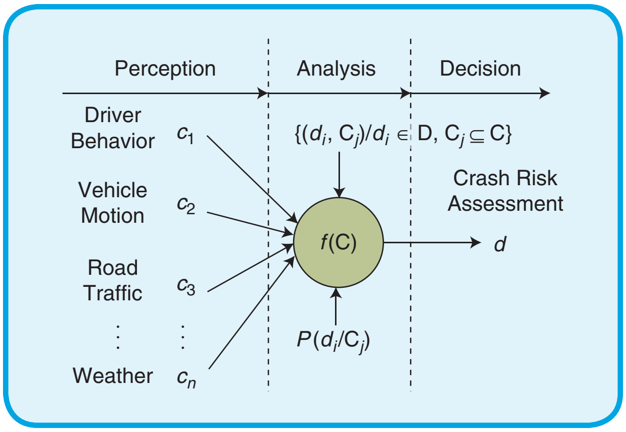

FIG 3 Crash risk assessment with consideration of multi-factors.

that using multi-roles as inputs of machine learning predictor significantly reduced the number of false collision warnings to drivers compared to the analytically derived formula based on the minimum safety gap. It also explicates the advantage of machine learning models capable of training a large volumes of collected data to form a concise and meaningful understanding of actual driving situation. Forming a high-level view of the world is a necessary requirement for intelligent vehicles to interact safely with both human drivers as well as other intelligent vehicles [26].

Artificial neural network model is one of the most practical tools used for fitting the relationships between risk driving behavior and traffic environmental factors, which model vehicle collision risk as a complicated nonlinear function of "driver-vehicle-road" attributes. The nonlinear function is defined by a multilayer network, including one or two hidden layers, with "driver-vehicle-road" attributes as inputs and collision risk prediction as output [27]. It is believed that a three layer neural network with a sufficiently large number of hidden neurons can model any nonlinear relationships between inputs and outputs [28]. There is variety of evolving neural network algorithms illustrated for improving application, [29] processed with a large number of inputs from accelerometer and gyro measurements based on a self-organized neural network model. The proposed approach is capable of recognizing dangerous conditions though heuristically tuning thresholds from simulated training crash tests, which outperforms the benchmark method by setting absolute thresholds on the inertial measurements. [30] proposed a probabilistic neural network is trained to predict prospective steering angles based on collected video data and the vehicles CAN bus data during human driving, thus imitating human behavior. The integration of end-to-end learning into a modularized architecture allows for additional safety constraints and complementary sensor information to be combined with intuitive steering.

Logit-based model is another popular methodology for analyzing crash the factors associated with accident severity. For example, [31] developed an unsupervised and Bayesian model that generates local multivariate linear models describing how the risky driver behavior is associated with the input data (independent variables), which are segmented into blocks of linear data sequences based on local statistical patterns. The advantage of this model formulation is using matrix-variate distribution theory, providing a general, intuitive and flexible parameterization. [32] employed a tree-based rules to analyze accidents involving powered two-wheelers, and demonstrated that the curve alignment, rural areas, run-off-the-road crashes, nighttime, and rainy weather were significantly associated with accident severity. These studies provided some insights into the factors that affect the likelihood of a vehicle crash. [33] obtained new insights into driving risk by using classification and regression tree (CART) model to analyze the relationship of driver characteristics, road conditions, and vehicle characteristics in near-crash database. The results indicate that the velocity when braking, triggering factors, potential object type, and potential crash type had the greatest influence on the driving-risk level involved in near-crashes. It also evaluated the application of CART model for predicting motor vehicle crashes, and showed that CART model performed better than traditional decision tree models.

In this work, we investigate the actual driving behaviors in near-crash events as well as the involved interactions among "driver-vehicle-road" multi factors.

In process of machine learning algorithms, the data may include easily hundreds of variables, however, it is obviously that not all these variables produce significant information gain for evaluating driver's risk behavior. Since there are a lot of redundant variables may decrease the performance of the learning dataset and actual validation dataset, it is necessary to keep an eye on the overfitting issues. Many machine learning techniques such as neural networks, tree-based models, and support vector machines perform worse when extra irrelevant predictors are added, and therefore a variable selection technique should always precede the modeling [32]. Rough set based models are some of the most practical tools, which are highly resistant to the inclusion of irrelevant variables through automatic variable subset selection. One of the main advantages of rough set based models is their simple interpretability. In spite of artificial neural network models having strong nonlinear fitting capabilities, their input-output relationships cannot be interpreted or verified explicitly, whereas the rough set based fitting model explicitly indicate the input-output relationships characterized by reduced rules, which are interpretable and easy to understand [38]. The transparent input-output relationships are very important for retro designing ADAS, especially evaluate the significant factors impacting the driving safety in emergency cases.

II. Data Collection and Processing

In this section, a field test was carefully designed and performed to collect real driving data in naturalistic and low-intervention environment, which was used to analyze the driving safety in near-crash scenarios under complex road traffic environment.

A. Experiment Design (Test Vehicle/Sensors/Drivers/Route)

The field driving experiments were conducted using a YUEXIANG sedan, which was provided by CHANGAN Auto company. The vehicle was equipped with instruments to detect driver behavior, vehicle motion state, and dynamic traffic in real time situation. The on-board units equipped in the experimental vehicle includes two cameras, Mobileye, two radars and on-board computer, as shown in Fig. 4. The two cameras were used to record vehicle forward view and driver's facial expression. Mobileye was used to identify the road traffic environment (lane lines and obstacles information preceding vehicle) and judge vehicle crash risk by detecting TTC (time to collision). The two radars were used to measure the headway between vehicle and approaching vehicles in front and behind respectively. On-board computer was used to record data obtained by sensors, including GPS, brake signal, steering signal, three-axis acceleration information.

Field trip was completed by our experiment vehicle equipped with above mentioned on-board units. Testing route was designed surrounding over the area in the central part of the Wuhan city, China, as shown in Fig. 5. The Google Map image records the test route (solid black) where the data was collected by vehicle with on-board GPS equipment. However, the GPS raw data described in longitudinal and lateral degree could not be directly used for evaluating the vehicles' trajectory in the travelled distance. More so, the corresponding RTK (real time kinematic) positions in Fig. 6 represents the vehicle trajectory in plane-coordinate. The points of origin and destination have been remarked. The RTK positions represent the vehicle trajectory in the test area.

The selected route for driving test is representative of most urban city traffic conditions in China, i.e., city ring road and expressway (usually low traffic volume and may have congestion). The experiments were implemented from 7:30 am to 9:30 am and 17:00 pm to 19:00 pm. Within these time frames, the traffic flow is denser and traffic crash is more frequent. In this study, the driver's high deceleration behavior was considered to be a crash risk related event. A totally of 51 drivers, who signed the consent form, participated in the designed driving experiments. The experiment lasted 60 days on average 4 hours per day, during which, the driving time and range was approximately 265.8 hours and over 5101.79 km respectively. Among the 51 drivers, 6 were female and 45 were male, all the participants held a valid driving license. The average age was 37 years (ranging from 25 to 56). They had 12 years (ranging from 3 to 16) mean period of driving experience.

B. Dataset Processing

This research focuses on the driving safety analysis and assessment in near crash scenarios. Driving risk is identified as a potential threat that could cause vehicle collision accidents. Usually, the consequence of driving risk for a driver in his/her normal state is mainly reflected by rapid evasive maneuvers (i.e. emergency braking and/or steering operation), which have been employed by many studies on naturalistic driving to identify near-crashes situations [12]–[14]. Near crash implies that the driver performs a rapid evasive maneuver (i.e. emergency braking and/or steering operation) that did not result in real crash. In the experiments, near crash events in naturalistic driving were identified by detecting unusual vehicle kinematic. When the vehicle deceleration reached a threshold value (longitudinal $-1.5\ m/s^2$, lateral: $-1\ m/s^2$) or TTC (time to collision) between the test vehicle and preceding vehicle is less than 3 s, the data collection system recorded the vehicle state (i.e., speed, brake signal, steering signal, and three-axis acceleration), the TTC with approaching vehicles in the longitudinal direction, and video sequence of the events happening at the time. Note that, it is very necessary to review the recorded video data to decide whether an event triggered by kinematic thresholds was actually safety critical. If not, such an event was not defined as near-crash and was deleted from the dataset. The recorded cases were checked manually.

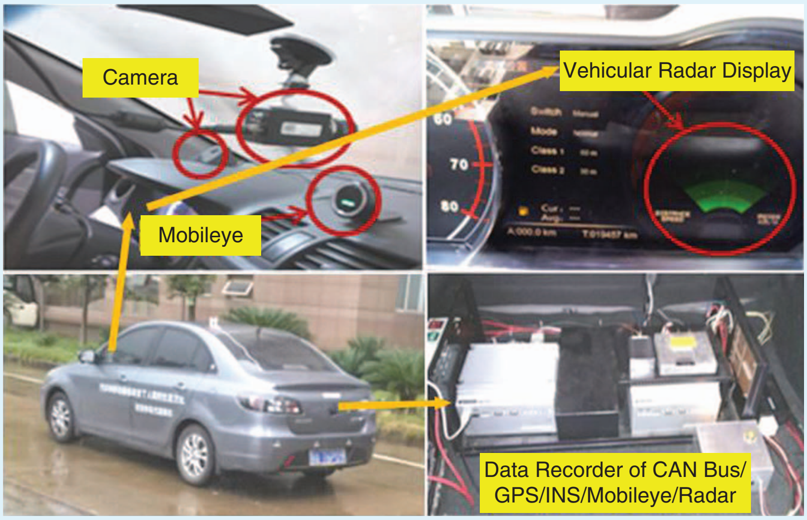

FIG 4 Experimental vehicle and apparatus.

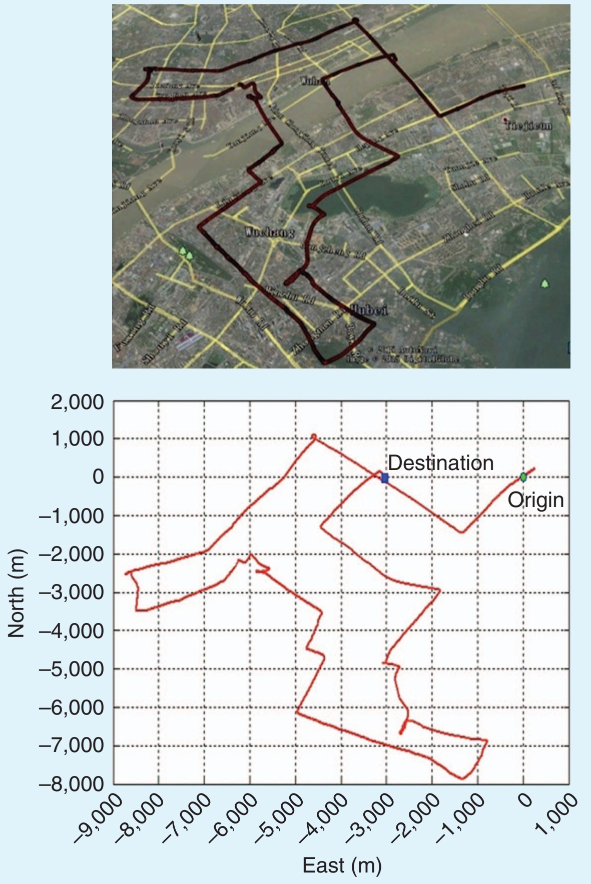

FIG 5 The route of the area of one on-road test.

Totally, 3374 near-crash events (only in the longitudinal direction, with 1 real crash accident) were recorded throughout the 30 days real driving test. Nearly all the near-crashes had large longitudinal deceleration, implying that the drivers tended to adopt the rapid braking maneuver to avoid potential crash. Hence, the driving-risk level was represented by the braking process characteristics. Intuitively, the driving risk is higher if the braking maneuver is performed with greater urgency in a near-crash. The clustering braking process characteristics data were investigated to evaluate the involvement of driving risk in a near-crash event [22]. The distribution of these near-crashes by deceleration is summarized in Table I. The driving-risk level in each near-crash case will be placed in one of the following three groups: low-risk, moderate-risk and high risk.

As outlined in previous studies, driver behavior, vehicle motion and traffic environment have been largely investigated and testified as among the major factors influencing driving safety in varying degrees. In this study, we conduct a reasoning model to predict driver's response and action in near-crash situation. It incorporates procedures that (1) detect the driving environment and to extract safety related conditional information about it, the status of the vehicle, and the conditions of the driver. It usually performs the sub-processes for monitoring, detection and classification of the information, for recognition of driver behavior and environmental factors and vehicle's actual states (position, orientation, conditions), (2) analyze the driving situation and conditions, which is to achieve the comprehension of

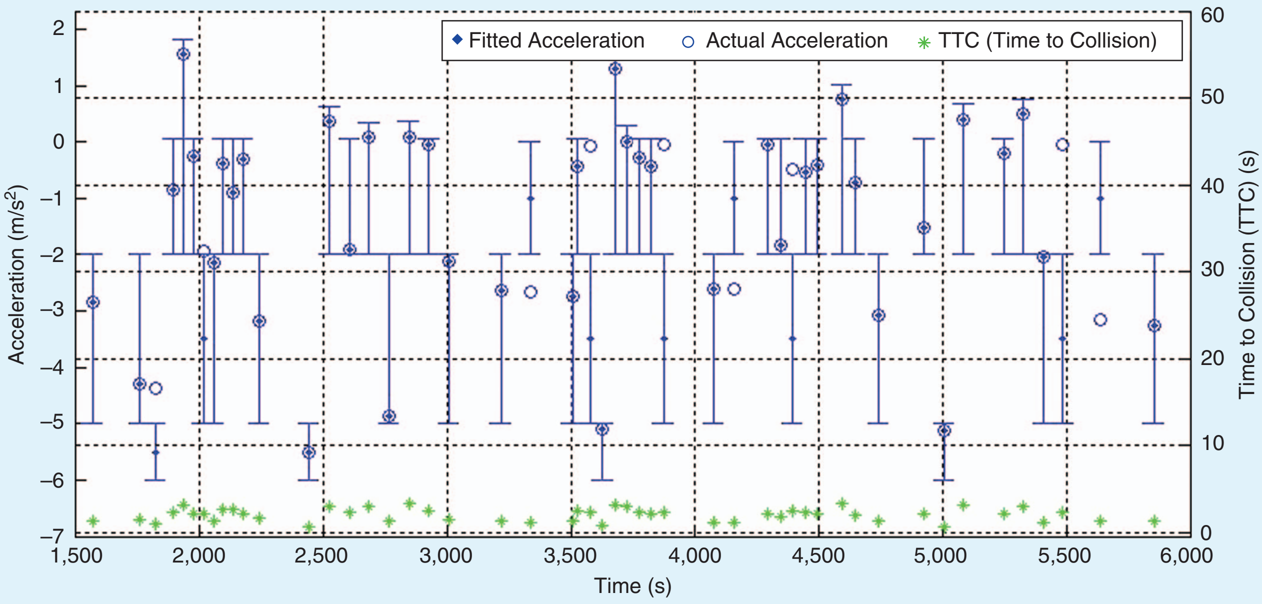

FIG 6 Crash risk evaluation of longitudinal near-crash scenarios.

the driving safety and to project it to the whole of the driving arrangement and process. Mathematically, the reasoning model can be defined as: IF $C_1^k \wedge C_2^k \wedge \ldots \wedge C_T^k$, THEN $\{(D_1, \beta_{k1}), (D_2, \beta_{k2}), \ldots, (D_N, \beta_{kN})\}$ with $\{\beta_{kj} \leq 1 \,|\, j = 1,2,\ldots,N\}$, a rule weight θ_k and attributes weight $\{\omega_i \,|\, i = 1,2,\ldots,T\}$, where $\{(D_1, \beta_{k1}), (D_2, \beta_{k2}), \ldots, (D_N, \beta_{kN})\}$ is referred to as a reliability β_{kj} of the inferred results for each D_j. The belief rule can be understood as if $C_1 = C_1^k$, $C_2 = C_2^k, \ldots, C_T = C_T^k$, then the consequence is THEN $\{(D_1, \beta_{k1}), (D_2, \beta_{k2}), \ldots, (D_N, \beta_{kN})\}$, where $C_1, C_2, \ldots, C_T$ are conditional attributes of the inference rule, $D_1, D_2, \ldots, D_N$ are assessment grade used in the consequence, and β_{kj} is the belief degree to which $\{D_j \,|\, j = 1,2,\ldots,N\}$ is to be believed to be the consequence. (3) predict the threats and consequences of that threat based on the comprehensive similarity of current conditional attributes $C_1^{k+1} \wedge C_2^{k+1} \wedge \ldots \wedge C_T^{k+1}$ with attributes in each inference rule. The prediction can be expressed as $D_{k+1} = \{D_j, \max_{N \geq j \geq 1} \beta_{kj} \,|\, \beta_{kj} = \theta_k * \Sigma_{i=1}^{T} \omega_i S(C_i^k, C_i^{k+1})\}$. Based on the weighted similarity results for each attribute, the prediction of the sample $C_1^{k+1} \wedge C_2^{k+1} \wedge \ldots \wedge C_T^{k+1}$ can be assessed using decision in D_j that maximizes the β_{kj}. From the above, we can apply the inference model in particular for assessing driving safety status with comprehensive consideration to driver behavior, vehicle motion and road environment. Then the prediction result is responsible for generating warnings for the driver and for the execution of the corrective actions, depending on the level of risk.

Altogether, the experiment dataset included the following five major categories: Participants information (age, gender, driving experience); Driver behavior and decision (acceleration, deceleration, steering); Road obstacles (time to collision in longitudinal direction); Vehicle kinematic status (velocity); Road traffic (traffic flow, road segment, road slipperiness). The dataset presented above is comprehensive and contains important attributes that describe the conditions affecting vehicle crash risk. It also provides potential information for analyzing the relationship among driving risk, driver/vehicle characteristics, and road environment. However, each attribute in the dataset has been defined and described by a specific performance measure (i.e. vehicle velocity is expressed by km/h, vehicular approaching status is expressed by Time to collision (s). Driver age is expressed by years and driver braking action is expressed by Boolean signal (1 or 0)). These could not be used directly for comprehensive evaluation by integrating other items with different property unit. In order to unify the property of the above attributes, a quantitation protocol is proposed in Table II, with explicitly considering the factors distribution in Chinese traffic situations and the distribution statistics of crash accidents [15]. Then, the heterogeneity among the attributes presented above can be eliminated, which in turn, can be applied for comprehensive analysis. It should be noted that, the quantitation range of attribute quantification need to be covered according to the real application.

III. Modeling Process

In this section, To achieve the accurate prediction, a hybrid VPRS (variable precision rough set) and Information entropy model is investigated for evaluating the field test data and categorizing the near-crash driving situation in

Table I. Distributed category of near-crash risk.

Driving Risk Level	Low	Moderate	High
Deceleration when braking m/s^2	(−2, 0]	(−5, −2]	(−8, −5]

Table II. Quantitation of attributes.

Attributes	Type	Description	
Participants information			
Gender	Boolean	1: Male; 2: Female	
Age	Continuous	Driver age (years) is categorized into four groups, 1: 18–30; 2: 31–45; 3: 46–60; 4: >60	
Driver behavior			
Acc pedal	Boolean	0: No; 1: Yes	Further categorized into:

Brake switch	Boolean	0: No; 1: Yes	1: Keep constant;
			2: Acceleration;

Turn indicator	Boolean	0: No; 1: Yes	3: Deceleration; 4: Steering
Road obstacles			
Vehicular distance with approached obstacle in longitudinal direction	Continuous	Evaluated by TTC (time to collision, seconds) and quantified into three levels: 1: >5; 2: 2.1–5; 3: 0–2	
Vehicle kinematic status			
Velocity	Continuous	Evaluated by km/h, quantified into four levels: 1: 0–40; 2: 41–50; 3: 51–60; 4: >60	
Road Traffic			
Road segment type	Qualitative	1: Corridor link; 2: Intersection; 3: Viaduct; 4: Tunnel	
Traffic flow	Qualitative	1: Congested; 2: Moderate flow; 3: Free flow	
Road slipperiness	Continuous	Evaluated by coefficient of friction between tyre and road surface, quantified into three levels: 1: 0.7–1; 2: 0.4–0.69; 3: 0–0.39	

corresponding safety level. The correlation of driving safety with all types of attributes is revealed and analyzed, and significance of relevant attributes, such as driver decision, vehicle motion and traffic environment, on the influence of driving risk is evaluated.

A. Rough Set Models for Database Classification

Rough Set (RS) is an effective approach for addressing problems of data classification, based on the conception of upper and lower approximation in a Decision Table (DT), which are constructed from empirical data and can represent the correlation of condition factors with decision factors [5]. The DT is characterized by four tuple set $S = \{U,\ A = C \cup D, V,\ f\}$, where $U = \{x_1, x_2, \ldots, x_{|U|}\}$ denotes a non-empty finite set called universe, A is a non-empty finite set of attributes that contains condition attribute set $C = \{a_1,\ a_2 \ldots,\ a_m\}$ and decision attribute set $D = \{d_1, d_2, \ldots, d_n\}$. $V = \quad V_a$ is the value domain of the attribute a, which represent the properties of either condition attributes or decision attributes. $f : U \times A \to V$ is a total function such that $f(x_i, a_j) \in V_a$ for every $\forall x_i \in U$ and $\forall a_j \in A$, e.g., $f(x_i, a_j) = v$, which means for element x_i, its attribute a_j has the value of v. For an arbitrary nonempty subset $B \subseteq A$, an indiscernibility relation is defined as:

$$IND(B) = \{\{x_i, x_j\} \in U * U / f(x_i, a) = f(x_j, a),\ \forall a \in B\}$$

$IND(B)$ partial U into a family of disjoint subsets $U/IND(B)$ called a quotient set of U:

$$U/IND(B) = \{[x]_B / x \in U\}$$

where $[x]_B$ denotes equivalence class determined by B. Then for a decision table (DT), the indiscernibility class of U with regards to condition attribute set C can be expressed as $[x]_C = \{c_1, c_2, \ldots, c_m\}$, and with regards to decision attribute set D can be expressed as $[x]_D = \{d_1, d_2, \ldots, d_n\}$. The relationship between condition attribute set C with decision attribute d_j can be evaluated by lower approximation and upper approximation, which are defined as:

$$\begin{cases} \underline{\mathrm{apr}_C}(d_j) = \cup\{x \in U |\ [x]_C \subseteq [x]_{d_j}\} \\ \overline{\mathrm{apr}_C}(d_j) = \cup\left\{x \in U |\ [x]_C \cap [x]_{d_j} \neq \phi\right\} \end{cases} \tag{3}$$

The positive region of d_j to C is defined as $\mathrm{POS}_C(d_j) = \underline{\mathrm{apr}_C}(d_j)$.

Variable precision rough set (VPRS) is proposed as an important extension of classical RS. The VPRS gives a less rigorous definition of the inclusion relation compared with Eq. (1), which makes the classical RS more fault tolerant. By introducing a precision parameter value $\beta \in (0.5, 1]$, the lower and upper approximations of d_j can be defined as follows:

$$\begin{cases} \underline{\mathrm{apr}_C^\beta}(d_j) = \cup\{x \in U | P(d_j | C) \geq \beta\} \\ \overline{\mathrm{apr}_C^\beta}(d_j) = \cup\{x \in U | P(d_j | C) > 1 - \beta\} \end{cases} \tag{4}$$

where $P(d_j | C)$ is the inclusion degree function:

$$P(d_j | C) = \frac{\left|[x]_C \cap [x]_{d_j}\right|}{|[x]_C|} \tag{5}$$

The improved positive region of d_j to C is defined as

$$\mathrm{POS}_C^\beta(d_j) = \underline{\mathrm{apr}_C^\beta}(d_j). \tag{6}$$

The classification quality of VPRS is evaluated by classification quality degree. If the decision attribute set divides the U into n classes, the classification quality degree of a certain attribute set $P\ (P \subseteq C)$ can be defined as follow:

$$\gamma^\beta(P,\ D) = \frac{\sum_{i=1}^{n} |\ \underline{\mathrm{apr}_P^\beta}(d_j)|}{|U|} \tag{7}$$

$\gamma^\beta(P,\ D)$ presents the percentage of effective sorting decision information D based on P in certain knowledge set, e.g., $\gamma^\beta(P,\ D) = 0$, it denotes that condition attribute set P includes no significant factors related to decision attributes in D. Noted that $0 \leq \gamma^\beta(P,\ D) \leq 1$.

B. β-Reducts for VPRS Attribute Reduction

For certain DT, not all of the condition attributes included is effective for information system category, which means that, some of the condition attributes are redundant. The condition attributes reduction in DT is one of the core problems for both VPRS. The process of finding the reduct is to identify the important attributes and remove the redundant attributes from condition attribute set in a certain DT. Formally in VPRS, β -reducts of condition attributes is expressed as $\mathrm{red}^\beta(C, D)$, which should be satisfied with the following two properties:

1) $\gamma^\beta(C, D) = \gamma^\beta(\mathrm{red}^\beta(C, D), D)$,
2) No proper subset of $\mathrm{red}^\beta(C,D)$, subject to the same β value can also give the same quality of classification.

The parameter β can be interpreted as confidence value, on which, the largest proportion of condition equivalence classes $[x]_C$ can be allocated correctly to different decision equivalence classes $[x]_D$. β-reducts derived subsets of the attributes, which are capable, through construction of decision rules, of explaining allocations given by the whole set of condition attributes subject to the majority inclusion aspect. The β-reducts should ensure the reduct-derived decision rules compatible with those from the original DT. Two propositions should be clearly considered to investigate the allowable ranges of β for attributes reduction, which allow the subsets of condition attributes to remain β-reducts [37].

Proposition 1

If condition attribute set C is discernible to d_j with a certain β value between $(0.5,\ 1]$, then the C is also discernible at any level between $(0.5,\ \beta)$.

Proposition 2

If condition attribute set C is not discernible to d_j with a certain β value between $(0.5, 1]$, then the C is also not discernible at any level between $(\beta, 1]$.

Thus, the confidence level associated with a set of attributes is defined by the least upper bound value on β such that all the condition classes satisfy the majority inclusion relation at this value. The least of these upper bounds of the β is defined as:

$$\beta = \min(m_1, m_2) \tag{8}$$

where

$$\begin{cases} m_1 = 1 - \max\{\Pr(d_j \mid c_i) \mid P(d_j \mid c_i) < 0.5\} \\ m_2 = \min\{\Pr(d_j \mid c_i) \mid P(d_j \mid c_i) > 0.5\} \end{cases}$$

The definition operates in terms of the quality of classification, which is used to define and extract β-reducts. The requirement is that a β-reduct should permit the use of subsets of attributes without loss of classification quality. β-reducts extract significant attributes from decision table and build rules for classification of unseen samples by matching the description of the sample to the condition part of each rule.

C. Attributes Weighted Similarity for Decision Making

Decision making rules extracted by VPRS reduct cannot cover the complete cases, where the sample does not match any of the rules. That is, if the matched rule is certain, it is clear that the class of the sample can be evaluated using the decision of the matched rule. However, if the matched rule has not been included, the classification is ambiguous. In this section, we propose weighted attributes to evaluate extracted rules, and design a decision algorithm based on attributes weights similarity to classify an unseen sample.

Suppose $U/X = \{X_1, X_2, \dots, X_l\}$ is an equivalence class produced by a set of condition attributes X, $X \subseteq C$. $U/D = \{Y_1, Y_2, \dots, Y_{|D|}\}$ is an equivalence class produced by decision attribute set D. The information entropy of subset X can be defined as [36, 37]:

$$H(X) \;= -\sum_{i=1}^{l} \frac{|X_i|}{|U|} \log_2 \frac{|X_i|}{|U|} \Bigg) \tag{9}$$

The conditional entropy of D given X is defined as:

$$H(D/X) \;= -\sum_{i=1}^{l} \frac{|X_i|}{|U|} \sum_{j=1}^{|D|} \frac{|Y_j \cap X_i|}{|X_i|} \log 2 \left(\frac{|Y_j \cap X_i|}{|X_i|} \right) \tag{10}$$

The mutual information entropy of D to X is defined as:

$$I(X, D) = \; H(D) - H(D/X) \tag{11}$$

If $X_i \in X$, the significance of X_i to the classification results can be evaluated by:

$$\mathrm{SIG}(X_i, X, D) = \mathrm{abs}(\triangle I) \tag{12}$$

Where, the $\mathrm{abs}(\triangle I)$ is the absolute value of the $\triangle I$, which is the mutual information degree. It is defined as:

$$\triangle I = I(X, D) - I(X - \{X_i\}, D) = H(D/X - \{X_i\}) - H(D/X) \tag{13}$$

If $X_p, X_q \in X$, the relative significance of X_p to X_q can be evaluated as:

$$\mathrm{SIG}_{p,q} = \mathrm{SIG}(X_p, X, D)/\mathrm{SIG}(X_q, X, D) \tag{14}$$

Suppose $B = \{b_1, b_2, \dots, b_{|B|}\}$ is condition attributes set in extracted rules, the corresponding weights set for each condition attribute is $\omega_b = \{\omega_{b_1}, \omega_{b_2}, \dots, \omega_{b_{|B|}}\}$, which is calculated by the geometric average method as follows:

$$\omega_{b_i} = \left(\prod_{q=1}^{|B|} \mathrm{SIG}_{i,\,q} \right)^{1/|B|} \tag{15}$$

After normalizing, the weight of each condition attribute is presented as follows:

$$\varepsilon_i = \frac{\omega_{b_i}}{\omega_B}, \;\; \omega_B = \sum_{i=1}^{|B|} \omega_{b_i} \tag{16}$$

If the sample does not match any of the rules, the decision algorithm based on attributes weighted similarity as shown below could be used to deal with the remaining cases.

Suppose that $u_i \in B$, $u_j \in U$, the similarity of the u_i and u_j on attribute b_i is defined as:

$$S_{b_i}(u_i, u_j) = 1 - \frac{|v_i - v_j|}{|b_{\max} - b_{\min}|} \tag{17}$$

In Eq. (14), v_i and v_j denotes the value of attribute b_i in u_i and u_j respectively. $b_{\max}$ and $b_{\min}$ respectively denotes the maximum and minimum of attribute b_i in B. Using the weighted similarity measurement to evaluate the similarity of u_i and u_j as follows:

$$S(u_i, u_j) = \frac{1}{n} \sum_{i=1}^{n} \omega_i S_{b_i}(u_i, u_j) \tag{18}$$

In Eq. (18), n is the number of rules in B. Based on the similarity result, the class of the sample u_j can be assessed using decision in u_i that maximizes $S(u_i, u_j)$.

From the above, the improved rough set model examines the driving safety with comprehensively evaluating the state of driver behavior, vehicle motion and road traffic, and extracting rule sets for matching the sample conditions to each classification of driving safety. Then, the significance of each factor will be evaluated by mutual information entropy and the weights are calculated based on the unseen sample in the real field situation that would be

A field test was carefully designed and performed to collect real driving data in naturalistic and low-intervention environment, which was used to analyze the driving safety in near-crash scenarios under complex road traffic environment.

finally classified into decision by matching the similarity of the sample to the condition part of each rule.

IV. EVALUATION AND DISCUSSION

A. Illustrative Examples

1) Extreme Driving Event Detection

In this section, the details of driver behavior and vehicle motion are visualized. Then, extreme braking events will be identified in accordance with special rules. One of the rules this study used is TTC-based thresholds [2], since driving behaviors vary at different TTC situation, implying different driving safety contexts. Further, these extreme braking events were linked to instantaneous vehicle control statuses and driving contexts to understand why they occur. Note that this study uses driver's extreme acceleration as key measures to capture driver's instantaneous collision avoiding decisions, i.e., how a vehicle is maneuvered instantaneously. Driving behaviors can also be captured by other measures, such as steering angles and the position of the accelerator or brake in a vehicle. Given that accelerations are the outcomes of maneuvering by drivers, the authors prefer to use them for analysis.

In our work, we extracted 678 groups of samples from field test, as described in Section 3.1, and take vehicle longitudinal emergency cases as example to explicitly evaluate the crash risk of the test vehicle and preceding vehicle in near-crash scenarios. This subset is a representative sample, in which the experimental vehicle recorded all the parameters in Table II when the vehicle deceleration reached a threshold of -1.5 m/s^2 or the TTC less than 3 s, the immediate data and previous sampling points were both recorded. Then, an offline behavior analysis is performed by randomly divided these samples into two subsets: 628 groups of these data were used for searching a β-reduct decision table (DT) of condition attributes which provide the same information for classification purposes as the full set of available attributes, then the significance of potential risk factors on driving safety can be evaluated and quantified based on the β-reduct DT by taking advantage of mutual information entropy for assigning attributes weights. The other 50 groups of data were extracted for inferring the extreme brake maneuvers happened in next 0.5 s by integrating the weighted β-reduct attributes as inputs, the results of predicted braking extent, actual driver deceleration and real time headway (TTC) for this case are summarized in Fig. 6.

2) General Results

In Fig. 6, the actual driver acceleration/deceleration are represented with circle point and classified into three scopes (m/s^2): $(-2, 1.8]$, $(-5, -2]$ and $[-6, -5]$, which respectively indicate three crash risk levels. The inferred driver's acceleration/deceleration in next short term are represented with solid dot. The longitudinal headway between vehicle on changing lane and approaching vehicles in neighbor lane is also evaluated by TTC, which is usually widely accepted as binomial judgement for assessing vehicle crash risk by setting a threshold [2, 3]. For example, it was identified that, at t = 1571s, 2768s, 3628s, etc., the TTC between the test vehicle and preceding vehicle was less than 2 s, these scenarios are viewed as risk situation.

Our proposed method examines driver intension, vehicle motion state as well as the approaching vehicles in dynamic traffic, and infer the driver's compulsory reaction for safety driving according to the current vehicle crash risk. Fig. 6 shows a comparison of our predicted results and actual driver deceleration, the results confirm that the combination of "driver-vehicle-road" based classifier produces more accurate predictions and few errors outperform the classifier using only TTC. It is observed that, at t = 1827s, the TTC between the test vehicle and preceding vehicle was less than 0.6 s, and the driver was predicted to apply a hard breaking, which represent high risk situation. Actually, that was the only real collision accident during the whole test period. However, at t = 3218s, the longitudinal headway (TTC) is 1.2s, which indicate the short distance to the preceding vehicle, and this near-crash is correctly identified as moderate risk (the consequent deceleration was –2.64 m/s^2). In this case, the driver was detected taking a timely deceleration action in advance. Consequently, the driver could only take a moderate deceleration to avoid collision accident. It was also noticed that, at t = 3628s, where the TTC is 0.92s, but we predict that the high risk will come to the next short term (the actual deceleration was –5.09 m/s^2), since the driver took an acceleration action before recognizing the potential crash risk and taking a harsh brake to avoid the accident, which was evaluated to be a high risk situation. It illustrates the influence of driver behavior and decision on the driving safety.

B. Scoring Comparison

This section explains how the model performance are evaluated. We analyze the driving risk of near-crash scenarios in the naturalistic driving experiments. Near-crash implies that the driver performs a rapid evasive maneuver (i.e., emergency braking and/or steering operation), failing which a real crash may occur. Previous studies indicated that these evasive driving events are associated with comprehensive "driver-vehicle-road" conditions, the main component of vehicle crash assessment models are interpreting the factors importance and understanding their relationship with driving safety. The attributes subset $C = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9\}$ extracted from field trial have been investigated for this study, where the c_1 denotes driver behavior or intention, c_2 denotes the gender of test driver, c_3 denotes the age of test driver, c_4 denotes vehicle velocity, c_5 denotes TTC in occupied lane, c_6 denotes TTC in neighbor lane, c_7 denotes road segment type, c_8 denotes traffic congestion, c_9 denotes road slipperiness. However, in process of machine learning algorithms, the data may include easily hundreds of variables, a key question therefore whether or not all these variables actually lead to true information gain? The answer is apparently not, since redundant variables may increase the performance of the learning dataset but they do not necessarily increase the performance on the actual validation dataset which can be easily controlled for by keeping an eye on the over-fitting. A scoring process is conducted to examine the accuracy and reliability of our proposed method for extreme driving events detection, it is also compared with the results derived by PNN algorithm, CART algorithm and TTC algorithm respectively.

1) Model Accuracy Evaluation

The scoring comparison were conducted for four different model. Model-1 was calibrated using rough set reduct attributes vector $\{c_1, c_4, c_5, c_6, c_9\}$ as input according the results of our proposed model. In order to examine the prediction accuracy that can be achieved depending only on one dataset at a time and to account for significance of reduct element from the collected data source, another three models were calibrated and compared; Model-2 based on PNN algorithm using all available factors gathered from field trial as model inputs; Model-3 based on CART algorithm consider experiment gathered factors as inputs, of which, the relative importance of all variables are normalized to characterize their ability to influence the model; Model-4 based on considering the threshold of TTC $\{a_5\}$ as criteria for determining the risk level.

The Receiver Operating Characteristics (ROC) curve is capable of examining the classification problem with positive and negative class values [37]. Through plotting a True Positive Rate (TPR) vs False Positive Rate (FPR) graph, it shows how well the model is at discriminating between the crash and non-crash cases in the target variable. In our study, the low vehicle crash risk is defined as the negative class and the high & moderate crash risk is defined as the positive class. We use ROC curve indexes as the main criteria to examine the performance of models for vehicle crash risk detection. The TPR is the ability to predict a crash case correctly and True Negative Rate (TNR = 1-FPR) is the ability to predict a non-crash case correctly. The overall accuracy indicates the proportion of correctly identified positive and negative cases, and the area under the ROC curve (AUC) represents the expected performance as a single scalar. The exact criteria for all models validation datasets are listed in Table III.

Consequently, Model-1 is consistently superior in term of classification accuracy and area under the ROC curve. Model-2 is ranked second after the full model (Model-1),

Table III. Validation: Classification rates and index.

Algorithms	Description of Inputs	TPR (%)	FPR (%)	TNR (%)	OCR (%)	AUC
M-1: VPRS+En	Reduct attributes	91.7	3.3	96.7	94.2	0.94
M-2: PNN	All collected attributes	88.3	6.7	93.3	90.8	0.88
M-3: CART	Normalized attributes	83.3	10.0	90.0	86.7	0.84
M-4: TTC	Vehicular TTC	71.7	13.3	86.7	79.2	0.71

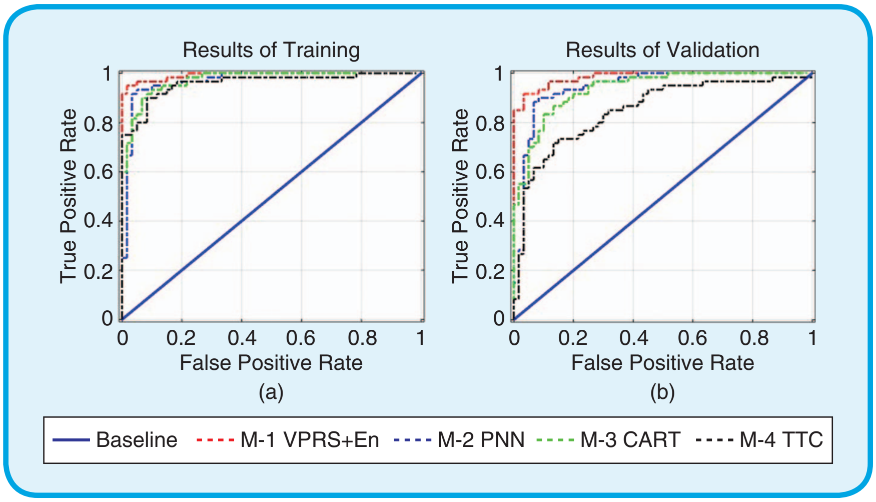

FIG 7 Overall performance of crash risk assessment models.

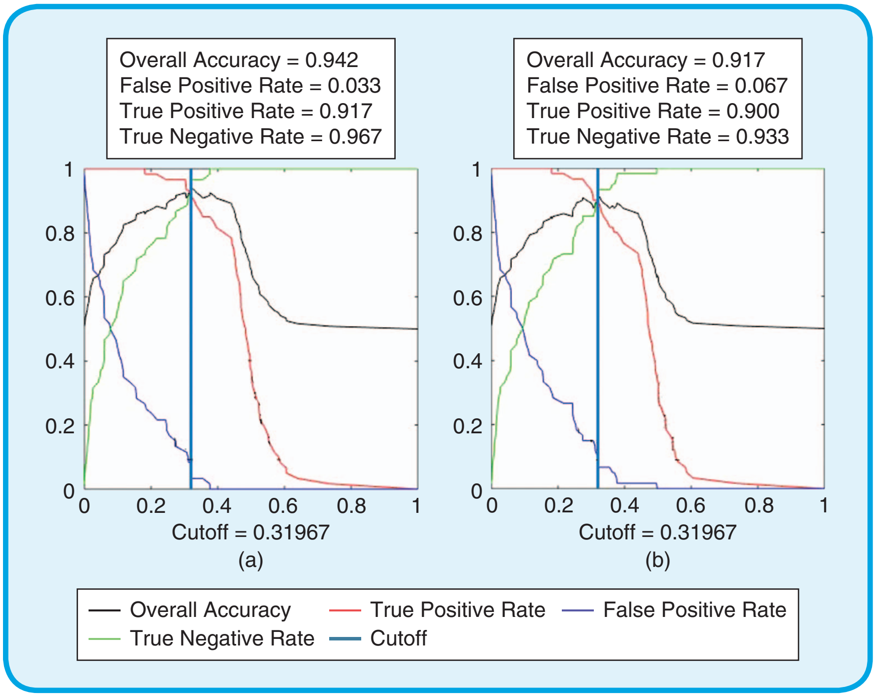

FIG 8 Model-1 vehicle crash risk pre-detection.

while Model-3 is relatively ranked lower than Model-1 and Model-2 but still providing satisfactory performance. Model-4 is ranked the lowest on these measures. Area under the ROC curves as shown in Fig. 7 and listed in Table III was found to be 0.94 for Model-1 validation dataset, 0.88 and 0.84 for Model-2 and Model-3, respectively while Model-4 achieved ROC of 0.71 all for the validation dataset. It may also be observed that Model-1 achieves 91.7% correct prediction of driver harsh deceleration in consequent short term by using the β-reduct attributes as input, while only 71.7% of vehicle crash risk has been predicted by using Model-4 (TTC model). It indicates the significance of driver volition a_4 and weather condition a_9 on the impact of safety driving. Although the attribute a_4 and a_9 have no direct relativity with vehicle crash risk, when comprehensively consider all attributes, the prediction has been improved, which testify the over speeding behavior and road snippiness effectively characterize the potential vehicle crash risk. We further conduct the prediction based on Model-2 and Model-3 respectively, and examine the prediction performance involved all the attributes C as inputs, then we achieve the less accurate results of 88.3% and 83.3% driver deceleration maneuver when compared to the performance of using β-reduct attributes, which account for the redundant attributes a_2, a_3 and a_8 having insignificant impact on driving safety.

It should be noted that the overall accuracy and error rate are particularly suspicious performance measures when the class distribution of a data set strongly biases to the majority class. Highly imbalanced problems generally have highly non-uniform error costs that often favor the minority class of primary interest. For instance, identifying a dangerous driver behavior as safe may be a fatal error, while identifying a safe driving behavior as risk is usually considered a much less serious error since this mistake can be corrected in later detections. ROC graphs are consistent for a given problem even if the distribution of positive and negative samples is highly skewed. The comparisons of model predictions between the observed and predicted risk levels for the learning and testing data are also presented in Fig. 7. The overall Model-1 prediction accuracy for the learning data is approximately 95.9% and that for the testing data is approximately 94.2%, which is a more optimal range compared with the other training models. For instance, in Fig. 7, we investigate the PNN model for vehicle crash prediction and achieved accuracies of 93.9% and 90.8% in the training and testing phases, and achieved accuracies of 89.6% and 86.7% for training and testing procedure based on CART model. The prediction performance of our proposed model demonstrates that the VPRS model structure can reflect the pattern hidden behind naturalistic data to some extent.

2) Efficiency Measurement of Pre-detection

To further understand the reliability of our proposed model, we apply the ROC curve indexes to evaluate the algorithm on drawing meaningful knowledge from the observed data and inferring the risk driving level for a given prediction time horizon. As shown in Figs. 8–11 that different overall accuracy, true positive rate, false positive rate and true negative rate are calculated by changing the cutoff value. In each figure, the vehicle crash risk predicted for specific prediction time from 0.5 second to 1 second are shown in left graph and right graph respectively.

In Fig. 8(a), 91.7% of risky driving behavior involved in near crash situation is correctly predicted 0.5 s before the driver taking the harsh deceleration in those scenarios, while 96.7% of safety driving status are accurately identified, which means that the false warning rate is only 3.3%, the overall driving risk prediction accuracy is 94.2%. The risky driving behavior also has been predicted before 1 s in Fig. 8(b), we target the 90.0% harsh deceleration behavior. It illustrates that prediction accuracy of Model-1 presents lightly reduction at longer time interval prediction.

We also examine the prediction accuracy of Model-2 and Model-3 in Fig. 9 and Fig. 10 respectively. The results show that the performance of Model-2 fluctuates when predicting the driving safety in longer time intervals. In Fig. 9(a), 93.3% of safety driving status are accurately identified, however, in Fig. 9(b), only 83.3% of true negative cases are effectively achieved, which means the increase of false warning for risk driving. In Fig. 10, although the overall accuracy of 86.7% and 85% respectively in 0.5 s and 1 s prediction time with very low false positive rate is considered reasonable, Model-3 performs not as good as the Model-1 and Model-2, the inclusion of abundant information may cause the overfitting in crash risk assessment. In Fig. 11, we achieved the lowest prediction accuracy by applying TTC model, only 61.7% of harsh braking and 80.0% of safety status is predicted before 1 s. The results further indicate that the vehicle near-crash events can diversify into different driving safety level when having same headway (TTC) before driver making the effort, since the driver maneuver will influence the driving safety in most emergency cases. These results show that the VPRS model framework seems to be quite robust with respect to realistic vehicle near crash risk assessment.

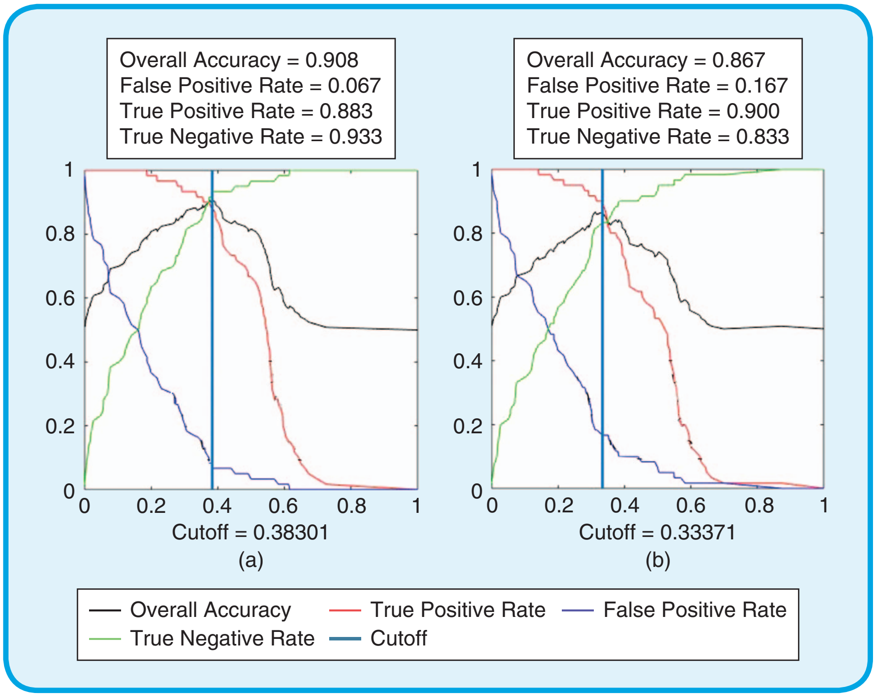

FIG 9 Model-2 vehicle crash risk pre-detection.

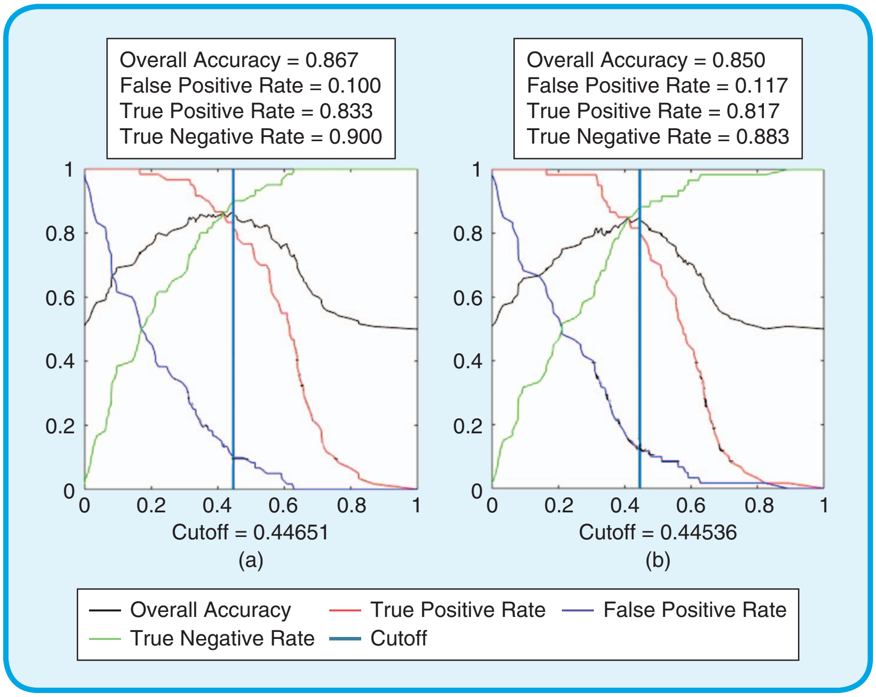

FIG 10 Model-3 vehicle crash risk pre-detection.

VI. Conclusions

In this paper, we explore driving safety problems based on a systematic "driver-vehicle-road" arrangement, which has been illustrated to outperform the TTC based method when pre-detecting potential vehicle crash risk. The involvement in near-crash situation is linked with the related attributes (driver behavior, vehicle motion, etc.) through an improved rough set model, which can be trained and validated using driving data. The rules bases are also used to make judgement for identifying whether a new case in near-crash scenario is involved with crash risk. Furthermore, the proposed rough set model reveals the input-output relationships by extracted rules in an easily interpretable way, while other models are black boxes in nature and their input-output relationships cannot be straightforwardly verified. This transparent input-output relationships are very important for retro designing ADAS. The proposed method can further accommodate

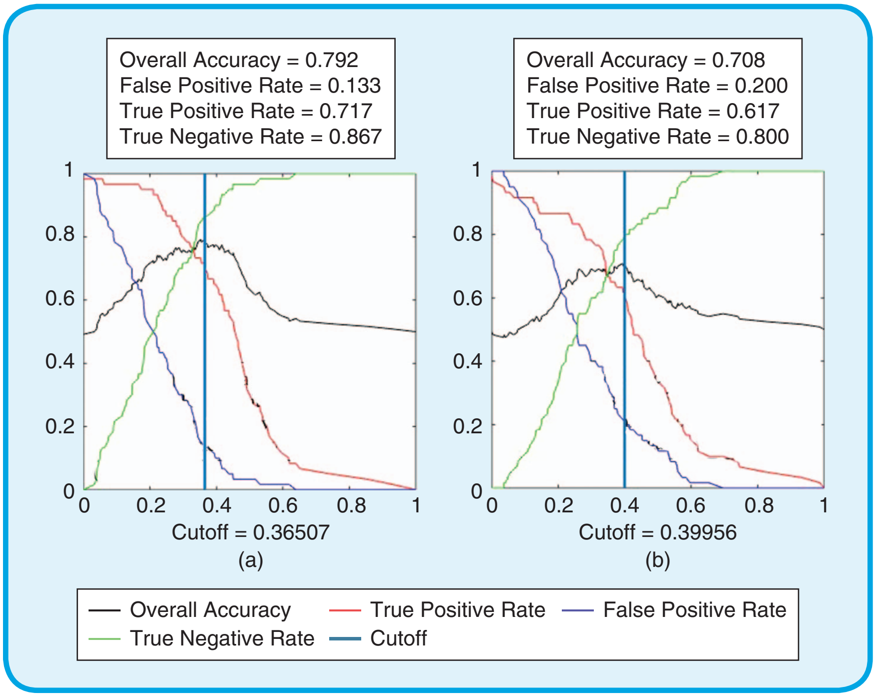

FIG 11 Model-4 vehicle crash risk pre-detection.

driver's experience, although it has not been discussed in this paper. Expert knowledge can be incorporated in rough set based models as constraints, or the initial value of training parameters in the proposed model can be set by experts intuitively whenever possible, leading to an expert-data driven system.

It also should be noted that, there are some limitations in our conducted field driving test. In our current database, the influence of multi factors on the driving risk was not fully addressed. Only longitudinal driving safety situation assessment has been processed and evaluated. The time-duration of the current experiment was not very long enough to collect data under all conditions. Despite such limitations, the proposed VPRS quantify the driving risk in near-crash event and to analyze the associated risk-factors, this can be extrapolated to specific studies on other datasets.

As drivers with different personality may weigh safety, comfort, driving efficiency and other factors very differently, further research will consider the influence of driver's personality on driver behavior in near-crash situations. In this case, driver's personality (e.g., personal weights to above aspects) can be modeled by choosing different parameters for the related cost function, and the challenge is to design a suitable cost function reflecting different driving styles. Furthermore, other driving intension should be considered to capture more complex scenarios, such as lane change, overtaking and turn round etc. These scenarios can be separated into sequent scenarios of lane changing and car following behavior. Finally, it would be interesting to apply the proposed method for retro design of vehicle collision avoidance system to helps driver to take effective action before vehicle involves in high risk situation in near crash scenarios.

Acknowledgment

The authors would also like to thank those who participated in the driving experiments.

About the Authors

Dr. Liqun Peng is currently an associate professor with the School of Transportation and Logistics at the East China Jiaotong University. He received his Ph.D. degree in Intelligent Transportation System Engineering from Wuhan University of Technology, in China, in 2015. He had worked at University of Alberta as a Postdoctoral Fellow (2016–2017). His research interests are in the area of advanced driving assistant systems and connected vehicle, specifically for improving roadway mobility and safety for both arterial and freeway.

Professor Miguel Angel Sotelo **(Fellow IEEE)** received the Ph.D. degree in Electrical Engineering in 2001 from the University of Alcalá, Spain. He is Head of the INVETT Research Group and Vice-President for International Relations at the University of Alcalá. He had served as Editor-in-Chief of IEEE Intelligent Transportation Systems Magazine (2014–2016) and Associate Editor of IEEE Transactions on Intelligent Transportation systems (2008–2015). He is currently the President of the IEEE Intelligent Transportation Systems Society.

Dr. Yi He is currently a research fellow at California PATH, University of California, Berkeley, CA, USA. He received his Ph.D degree in Intelligent Transportation System Engineering from Wuhan University of Technology, China in 2015. His research interests include driving behavior analysis, connected vehicle and advanced driving assistance system for addressing traffic safety problem.

Dr. Yunfei Ai currently works as Postdoctoral Research Associate with the National Engineering Laboratory for Transportation Safety and Emergency Informatics, China. He received his Ph.D degree in Transportation Planning and Management from Dalian Maritime University, China, in 2016. His research interests include transportation safety, emergency management and Emergency Informatics.

Dr. Zhixiong Li (M'16) received his Ph.D. degree in Transportation Engineering from Wuhan University of Technology, China. Currently he is a ARC DECRA Fellow at School of Mechanical, Materials, Mechatronic and Biomedical Engineering, University of Wollongong, Australia. He is also an adjunct professor at Faculty of Engineering, Ocean University of China. His research interests include Intelligent Vehicles and Control, Loop Closure Detection, and Mechanical System Modeling and Control. He is an Associate Editor of Measurement (Elsevier) and a Column Editor of IEEE Intelligent Transportation Systems Magazine.

References

[1] J. Liu and A. J. Khattak, "Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles," *Transp. Res. Part C, Emer. Technol.*, vol. 68, pp. 83–100, 2016.

[2] H. S. Tan and J. Huang, "DGPS-based vehicle-to-vehicle cooperative collision warning: engineering feasibility viewpoints," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 4, pp. 415–428, 2006.

[3] S. Moona, I. Moon, and K. Yi, "Design, tuning, and evaluation of a full-range adaptive cruise control system with collision avoidance," *Control Eng. Pract.*, vol. 17, no. 4, pp. 442–455, 2009.

[4] J. N. K. Liu, Y. Hu, and Y. He, "A set covering based approach to find the reduct of variable precision rough set," *Inf. Sci.*, vol. 275, pp. 83–100, 2014.

[5] I. Park and G. Choi, "A variable-precision information-entropy rough set approach for job searching," *Inform. Syst.*, vol. 48, pp. 279–288, 2015.

[6] J. Weng, Q. Meng, and X. Yan, "Analysis of work zone rear-end crash risk for different vehicle-following patterns," *Accid. Anal. Prev.*, vol. 72, pp. 449–457, 2014.

[7] J. Weng, S. Xue, Y. Yang, X. Yan, and X. Qu, "In-depth analysis of drivers' merging behavior and rear-end crash risks in work zone merging areas," *Accid. Anal. Prev.*, vol. 77, pp. 51–61, 2015.

[8] M. Montemerlo, J. Becker, and B. Suhrid, "Junior: the Stanford entry in the urban challenge," *J. Field Robot.*, vol. 25, no. 9, pp. 569–597, 2008.

[9] R. Schubert, K. Schulze, and G. Wanielik, "Situation assessment for automatic lane-change arbitrary maneuvers," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 607–616, 2010.

[10] N. Kaempchen, B. Schiele, and K. Dietmayer, "Situation assessment of an autonomous emergency brake for arbitrary vehicle-to-vehicle collision scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 678–687, 2009.

[11] C. Oh, E. Jung, and H. Rim, "Intervehicle safety warning information system for unsafe driving events: methodology and prototypical implementation," *Transp. Res. Rec. J. Transp. Res. Board.*, vol. 2324, no. 1, pp. 1–10, 2011.

[12] S. B. McLaughlin, J. M. Hankey, and T. A. Dingus, "A method for evaluating collision avoidance systems using naturalistic driving data," *Accid. Anal. Prev.*, vol. 40, no. 1, pp. 8–16, 2008.

[13] Z. Bareket, P. S. Fancher, H. Peng, K. Lee, and C. A. Assaf, "Methodology for assessing adaptive cruise control behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 4, no. 3, pp. 123–131, 2003.

[14] T. A. Dingus et al., "The 100-Car naturalistic driving study, phase II-results of the 100-Car field experiment," *National Highway Traffic Safety Admin (DOT HS 810593)*, 2005.

[15] J. Stutts et al., "Driver's exposure to distractions in their natural driving environment," *Accid. Anal. Prev.*, vol. 37, no. 6, pp. 1093–1101, 2005.

[16] R. J. Hanowski, M. A. Perez, and T. A. Dingus, "Driver distraction in long-haul truck drivers," *Transp. Res. Part F, Traffic Psychol. Behav.*, vol. 8, no. 6, pp. 441–458, 2005.

[17] F. Guo, K. G. Klauer, J. M. Hankey, and T. A. Dingus, "Near crashes as crash surrogate for naturalistic driving studies," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2147, no. 1, pp. 66–74, 2010.

[18] X. D. Yan, E. Radwan, and D. H. Guo, "Effects of major-road vehicle speed and driver age and gender on left-turn gap acceptance," *Accid. Anal. Prev.*, vol. 39, no. 4, pp. 843–852, 2007.

[19] E. Rendon-Velez, I. Horváth, and E. Z. Opiyo, "Progress with situation assessment and risk prediction in advanced driver assistance systems: a survey," in *Proc. 16th ITS World Congress*, 2009, pp.21–25.

[20] H. L. Huang and M. Abdel-Aty, "Multilevel data and Bayesian analysis in traffic safety," *Accid. Anal. Prev.*, vol. 42, no. 6, pp. 1556–1565, 2010.

[21] T. Kim and H. Y. Jeong, "Crash probability and error rates for head-on collisions based on stochastic analyses," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 896–904, 2010.

[22] H. M. Hassan and M. Abdel-Aty, "Exploring the safety implications of young drivers' behavior, attitudes and perceptions," *Accid. Anal. Prev.*, vol. 50, pp. 361–370, 2013.

[23] M. Ahmed, M. Abdel-Aty, J. Y. Lee, and R. J. Yu, "Real-time assessment of fog-related crashes using airport weather data: a feasibility analysis," *Accid. Anal. Prev.*, vol. 72, pp. 309–317, 2014.

[24] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 63–77, Mar. 2006.

[25] P. Szczurek, B. Xu, O. Wolfson, and J. Lin, "Estimating relevance for the emergency electronic brake light application," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1638–1656, 2012.

[26] M. Müller, D. Böhmländer, and W. Utschick, " Machine learning based prediction of crash severity distributions for mitigation strategies," *J. Adv. Inform. Technol.*, vol. 9, no. 1, pp. 15–24, 2018.

[27] H. Abdelwahab and M. Abdel-Aty, "Artificial neural networks and logit models for traffic safety analysis of toll plazas," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1784, no. 1, pp. 115–125, 2002.

[28] Q. Zeng and H. Huang, "A stable and optimized neural network model for crash injury severity prediction," *Accid. Anal. Prev.*, vol. 73, pp. 351–358, 2014.

[29] D. Selmanaj, M. Corno, and S. M. Savaresi, "Hazard detection for motorcycles via accelerometers: a self-organizing map approach," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3609–3620, 2017.

[30] C. Hubschneider, A. Bauer, J. Doll, M. Weber, S. Klemm, F. Kuhnt, "Integrating end-to-end learned steering into probabilistic autonomous driving," in *Proc. 20th Int. Conf. Intelligent Transportation Systems*, 2017, pp. 1–7.

[31] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot, "An unsupervised approach for inferring driver behavior from naturalistic driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3325–3336, 2015.

[32] A. I. Montella and L. Liana, "Safety performance functions incorporating design consistency variables," *Accid. Anal. Prev.*, vol. 74, pp. 133–144, 2015.

[33] J. Wang, Y. Zheng, X. Li, C. Yu, K. Kodaka, and K. Li, "Driving risk assessment using near-crash database through data mining of tree-based model," *Accid. Anal. Prev.*, vol. 84, pp. 54–64, 2015.

[34] M. Ahmed and M. Abdel-Aty, "A data fusion framework for real-time risk assessment on freeways," *Transp. Res. Part C, Emer. Technol.*, vol. 26, pp. 203–213, 2013.

[35] W. Ziarko, "Variable precision rough set model," *J. Comput. Syst. Sci.*, vol. 46, no. 1, pp. 39–59, 1993.

[36] H. S. Own and Y. Hamdi, "Rough set based classification of real world Web services. Rough set based classification of real world Web services," *Inform. Syst. Front.*, vol. 17, no. 6, pp. 1301–1311, 2015.

[37] J. Liu, Q. Hu, and D. Yu, "A weighted rough set based method developed for class imbalance learning," *Inf. Sci.*, vol. 178, no. 4, pp. 1235–1256, 2008.

[38] M. Beynon, "Reducts within the variable precision rough sets model: a further investigation," *Eur. J. Oper. Res.*, vol. 134, no. 3, pp. 592–605, 2001.

ITS