Accident Analysis and Prevention 154 (2021) 106061

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A data-driven feature learning approach based on Copula-Bayesian Network and its application in comparative investigation on risky lane-changing and car-following maneuvers

Tianyi Chen [a,*], Yiik Diew Wong [a], Xiupeng Shi [a,b], Yaoyao Yang [c]

[a] School of Civil and Environmental Engineering, Nanyang Technological University, 639798, Singapore
[b] Institute for Infocomm Research, The Agency for Science, Technology and Research (A*STAR), Singapore
[c] School of Business, Renmin University of China, 100872, Beijing, China

ARTICLE INFO

Keywords:
Crash causal inference
Feature learning
Copula-Bayesian Network
Risky driving maneuver
Lane-changing
Car-following

ABSTRACT

The era of 'Big Data' provides opportunities for researchers to have deep insights into traffic safety. By taking advantages of 'Big Data', this study proposes a data-driven method to develop a Copula-Bayesian Network (Copula-BN) using a large-scale naturalistic driving dataset with multiple features. The Copula-BN is able to explain the causality of a risky driving maneuver. As compared with conventional BNs, the Copula-BN developed in this study has the following advantages: the Copula-BN 1. Has a more rational and explainable structure; 2. Is less likely to be over-fitting and can attain more satisfactory prediction performance; and 3. Can handle not only discrete but also continuous features. In terms of technical innovations, Shapley Additive Explanation (SHAP) is used for feature selection, while Gaussian Copula function is employed to build the dependency structure of the Copula-BN. As for applications, the Copula-BNs are used to investigate the causality of risky lane-changing (LC) and car-following (CF) maneuvers, upon which the comparisons are made between the two essential but risky driving maneuvers. In this study, the Copula-BNs are developed based on the Second Highway Research Program (SHRP2) Naturalistic Driving Study (NDS) database. Upon network evaluation, the Copula-BNs for both risky LC and CF maneuvers demonstrate satisfactory structure performance and promising prediction performance. Feature inferences are conducted based on the Copula-BNs to respectively illustrate the causation of the two risky maneuvers. Several interesting findings related to features' contribution are discussed in this paper. To a certain extent, the Copula-BN developed using the data-driven method makes a trade-off between prediction and causality within the 'Big Data'. The comparison between risky LC and CF maneuvers also provides a valuable reference for crash risk evaluation, road safety policy-making, etc. In the future, the achievements of this study could be applied in Advanced Driver-Assistance System (ADAS) and accident diagnosis system to enhance road traffic safety.

1. Introduction

1.1. Research background

The era of 'Big Data (a very large amount of data)' provides opportunities for researchers to yield deep insights into traffic safety. Advanced data collection and storage techniques, such as detectors, sensors, and databases, have led to rapid development of intelligent transportation system (ITS) and resulted in significant changes in traffic safety analytics (Lian et al., 2020). However, most statistical learning methods which have been widely used for processing 'Big Data' cannot well explain the intrinsic causation of a traffic crash, even though they can attain extremely high prediction accuracy (Lian et al., 2020; Mannering et al., 2020). Consequently, with availability of 'Big Data', how to tradeoff prediction performance and causal relationship inference places a challenge to research in traffic safety.

Lane-changing (LC) and car-following (CF) are two essential driving maneuvers in road traffic stream flow. The vehicles' crashes caused by improper CF or LC maneuver can result in severe property damage and human fatality. As reported by National Highway Safety Administration

* Corresponding author.
E-mail addresses: TIANYI002@e.ntu.edu.sg (T. Chen), CYDWONG@ntu.edu.sg (Y.D. Wong), shix@i2r.a-star.edu.sg (X. Shi), yangyaoyao2017@ruc.edu.cn (Y. Yang).

https://doi.org/10.1016/j.aap.2021.106061
Received 6 December 2020; Received in revised form 20 February 2021; Accepted 22 February 2021
Available online 7 March 2021

(NHTSA), LC (including lane-merging) crashes account for 5% of police-reported crashes and 0.5 % of road traffic fatality in the United States, while CF (i.e., rear-end) crashes account for 28 % of total automobile crashes and 6% of fatal automobile crashes (NHTSA, 2015). The causation of vehicles' crashes is complex due to it involving a variety of contributing factors (Mannering et al., 2016). Although there are differences in the contributing factors between CF and LC crashes, the two maneuvers have significant effects on each other and sometimes it is hard to discuss them independently (Zhang et al., 2019). Hence, how to conduct comparative investigations into risky LC and CF maneuvers considering the inherent complexity of the two risky maneuvers also poses a challenge.

As a brief summary of this section, this study is faced with two main challenges: first, how to tradeoff between prediction and causality with 'Big Data'; and second, how to investigate and compare risky LC and CF maneuvers which have inherent complexity. To overcome the above challenges, this study develops a method incorporating data-driven feature selection techniques and causal-inference model. The method is demonstrated on a large-scale naturalistic driving dataset to explore similarity and difference between risky LC and CF maneuvers. Investigating the causation of risky LC and CF maneuvers is of paramount significance to the enhancement of road traffic safety. The proposed method and the analytical results of applying the method could provide valuable references for the development of Advanced Driver-Assistance System (ADAS) and accident diagnosis system.

1.2. Literature review and research gaps

Vehicles' crash can be regarded as a systematic failure resulting from the effect of multifold factors (i.e., features) instead of only a single mistake (Ren et al., 2008). Therefore, risk factor investigation has become a mainstream in the research arena of road traffic safety. Adanu et al. (2017) investigated the influence of drivers' residential factors on driving behaviors in order to explain the area-based difference in road traffic crash severity. Han et al. (2018) examined the effect of road factors on crash frequency across different traffic regions. Papadimitriou et al. (2019) ranked the infrastructure-related crash risk factors according to their contributions towards road traffic safety. Xing et al. (2019) discussed the lag effects of weather factors on road traffic crashes in the form of hourly association. Das et al. (2020) developed an trajectory-level LC detection method considering the effects of different weather conditions. Wang and Feng (2019) investigated the consistency between single-vehicle and multi-vehicle crash hotpots using roadway geometric features and traffic features. Wali et al. (2020) characterized the volatility in longitudinal and lateral driving decision and depicted the association between the volatility in Time-to-Collision (TTC) and injury severity. Most previous studies were conducted based on a specific category of features due to limited data sources or for the purpose of narrowing down research focus. However, the era of 'Big Data' has made large-scale traffic data with multiple categories of features becoming available. Taking advantages of 'Big Data', researchers can synthesize the features from multiple categories and better explain the inherent complexity of crash causality.

As a subset of data-driven approach, machine learning methods (especially supervised machine learning methods) have become increasingly popular in the prediction of road traffic crashes or risky driving maneuvers. With supervised learning, machine learning models can predict the label (i.e., output target) of a sample based on the performance of the sample's features (i.e., input variables). Zeng et al. (2016) used neural networks to model the nonlinear relationship between road traffic crash frequency by severity and risk factors. Bao et al. (2019) employed spatiotemporal convolutional long short-term memory network (STCL-Net) to predict citywide road traffic crash risk by leveraging multi-source dataset. Osman et al. (2019) proposed a bi-level hierarchical classification method to recognize the type of drivers' secondary tasks based on their driving behavioral features. Wang et al. (2019) applied support vector machine (SVM) to estimate road traffic crash propensity using traffic platoon features collected by floating car method. Formosa et al. (2020) developed a series of Deep Neural Network (DNN) to detect road traffic conflicts based on the traffic features collected by in-vehicle sensors. As compared to traditional methods, machine learning methods are able to attain more satisfactory predictive performance. However, machine learning methods are always seen as "black-boxes" since they cannot clearly uncover the causal relationships between features and label (Mannering et al., 2020).

Bayesian Network (BN) is a type of robust probabilistic model using a Directed Acyclic Graph (DAG) to present a set of variables and quantify their conditional dependencies (Pearl, 2009). As an effective causal-inference model, BNs have been widely used to map causal relationships between risk-related features and evaluate road traffic safety. Liang and Lee (2014) applied dynamic BN and supervised clustering to detect driver's cognitive distraction according to eye movement and driving performance. Chen et al. (2015) integrated BN with multinomial logit model to offer a good understanding of the relationship between contributing factors and driver injury severity in CF crashes. Mbakwe et al. (2016) combined BN and Delphi method to model the highway traffic crashes and estimate the crash rates in developing countries. Chen et al. (2018) proposed a probabilistic decision-making framework for CF crash avoidance systems based on a BN model with major crash-causing features. Hwang et al. (2019) developed a BN to depict the relationships between the factors that are involved in the safety rating of interstate motor carriers. However, the conventional BN applied in most previous studies has following limitations. First, the network is built either by the pre-determined structure based on a priori knowledge or by structure learning techniques. The BN with pre-determined structure might neglect the key features that have not been uncovered and result in unsatisfactory prediction performance, while the BN generated by structure learning sometimes is not easily explainable especially with multiple features as inputs. Second, the BN cannot efficiently deal with a large scale of features which especially include continuous ones.

As two essential driving maneuvers on the roads, the safety of both LC and CF has drawn much attention from academia. Many researchers have investigated the features that have effects on risky LC and CF maneuvers. Reimer et al. (2013) discussed the impact of age and cognitive effects on a regular LC maneuver based on the data from an on-road study. Yang et al. (2019) applied multilevel mixed-effects linear models to explain the relationship between distance headway and contributing factors of an LC maneuver based on the LC events extracted from a naturalistic driving dataset. Li et al. (2020b) proposed a short-term prediction model of LC impacts during the propagation of traffic oscillation and recommended the optimal speed and acceleration for a safe LC maneuver. Wang et al. (2018) developed a desired safety margin model to describe the influence of driving behaviors (e.g., reaction characteristics, acceleration habits, etc.) on the stability in CF. Ding et al. (2019) assessed drivers' crash risk variation of CF maneuver as based on the features of driver's visual perception, vehicle type, and horizontal curves. Li et al. (2020a) employed dynamic driving risk potential field in modeling the CF maneuver of automated vehicles, which considers the dynamic effects of vehicles' acceleration and steering angle. As mentioned above, the two maneuvers have significant effects on each other, which sometimes make it challenging to discuss them independently (Zhang et al., 2019). However, few studies have attempted to conduct comparative investigations into risky LC and CF maneuvers to explore the similarity and difference between the two risky maneuvers, e.g., the difference in the features contributing to the maneuvers, the difference in the effects of the common features on the maneuvers, etc.

From the above literature scan, the research gaps inherent in previous studies, which this study is proposed to address, are summarized as follows:

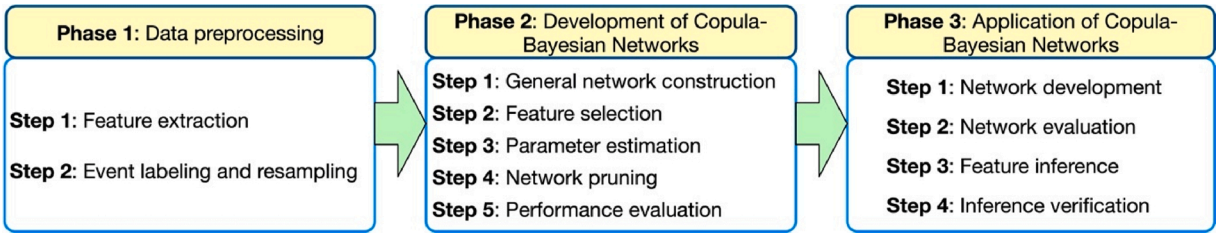


Fig. 1. Research Framework.

1 Few studies have attempted to take advantages of 'Big Data' to comprehensively consider the features belonging to multiple categories and illustrate the causal relationships between them when predicting or investigating risky driving maneuvers.
2 Conventional BN has following limitations: first, the development of BN is likely to neglect uncovered key features or lack rationality; and second, the BN is not always suitable for a large scale of features (e.g., inputs from 'Big Data'), especially continuous features.
3 The comparative investigations into the causality of risky LC and CF maneuvers have not been paid much attention. Few studies have explored the similarity and difference between the two risky maneuvers.

1.3. Research framework and paper organization

To resolve the above research challenges and research gaps, this study proposes a method to develop Copula-BN based on a large-scale naturalistic driving dataset. The Copula-BN is able to better illustrate the causality of vehicles' crash from a probabilistic perspective. The method is demonstrated on a practical naturalistic driving dataset to develop Copula-BNs respectively for risky LC and CF maneuvers, upon which the comparative investigations into the two risky maneuvers are conducted. As presented in Fig. 1, this study consists of three phases. Phase 1 is concerned with constructing the dataset that contains candidate features and event label of each sample as a preparation for the next phases. Phase 2 illustrates the procedure to develop a Copula-BN and to identify the optimal Copula-BN for a risky driving maneuver. Phase 3 uses a practical naturalistic driving database to demonstrate the method and develops the Copula-BNs respectively for risky LC and CF maneuvers, upon which feature inferences are conducted. The remaining paper is organized as follows. Section 2 introduces innovative techniques which are essential for the development of Copula-BN. Sections 3 to 5 present the procedures for Phases 1, 2 and 3 as well as the steps in each phase, respectively. Section 6 summarizes the conclusions, limitations, and future work.

2. Preliminaries

2.1. Shapley additive explanation

In this study, Shapley Additive Explanation (SHAP) is employed in the step of feature selection in Phase 2. SHAP applies the game theoretically optimal Shapley value to measure the additive importance of each feature (Lundberg and Lee, 2017), upon which the candidate key features are identified. The feature importance reflects the effect of the feature on the prediction of the target states in machine learning. The features with larger absolute SHAP importance are more important. The explanation model of SHAP is integrated with an additive feature attribution method, which is specified as:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x_j', \tag{1}$$

where $x' \in \{0,1\}^M$ is the coalition vector that indicates whether the feature is observed or not, ϕ_j refers to the contribution of the feature j as measured by the Shapley value, and M indicates the feature size. The Shapley value (i.e., SHAP value), as a weighted average of all possible subsets of input features, is computed as:

$$\phi_j = \sum_{S \subseteq \{x_1, x_2, \cdots x_M\} \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \Big[E\big[f(x)\, x_{S \cup \{i\}}\big] - E\big[f(x)\, x_S\big]. \Big] \tag{2}$$

where S denotes all feature subsets, $x_{\cdot}$ denotes a sample, and $E[f(x)\, x_i]$ is the expectation of given the subset $\{i\}$ of the features. Then, the feature importance (i.e., SHAP importance), as the average of the marginal contributions of the feature to the target states, is obtained as:

$$I_j = \frac{1}{m \cdot n} \sum_{k=1}^{m} \sum_{p=1}^{n} \left| \phi_j^{(k,p)} \right| \tag{3}$$

where m indicates the number of the target states and n refers to the sample size.

Most conventional measures of feature importance (e.g., tree-based machine learning classifier using information gain or entropy to measure feature importance) can only describe the general contribution of a feature to the target rather than the specific contribution to each state of the target. Hence, SHAP is more comprehensive than the most conventional measures of feature importance, as it considers the contributions of a feature to each state of the target based on each sample. In this study, as a tree-based machine learning classifier, Random Forest classifier is employed to be integrated with SHAP to assign importance to each candidate feature. The procedure of feature selection is introduced in Section 4.2.

2.2. Copula-Bayesian Network

Bayesian Network (BN) can represent the probabilistic causal relationships between variables (including features and label) using Bayesian inferences. BN is able to predict the conditional probabilities of an event (i.e., target) given the contributing features as possible causes. Also, it can illustrate how the features contribute to the event by estimating the posterior probability of the features given the states of the event. Hence, the capability for bi-directional inferences integrated with rigorous probabilistic computation results in the wide application of BN (Pearl, 2009). Herein, a BN is defined as $BN = G(X, E)$, where $X = \{x_1, x_2, \cdots, x_\varepsilon\}$, which indicates the set of nodes representing variables, and E refers to the set of the directed edges linking the nodes in a DAG. According to chain rule, the dependence relationships between nodes can be specified by the joint probability distribution of nodes, which can be expressed as (Russell and Norvig, 2002):

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1,\, i \in \varepsilon}^{n} P(x_i | X_i^{prt}) \tag{4}$$

where X_i^{prt} refers to a set of parent nodes of the node x_i.

Copula-Bayesian Network (Copula-BN) is a BN which implements Copula as the backbone to specify the dependency between variables (Elidan, 2010). Herein, a Copula is a multivariate joint distribution of random variables which can have various marginal distributions

Table 1
Descriptions of feature categories.

Feature categories	Abbr.	Descriptions
Driver's physiological features	DP	Driver's physiology-related features, i.e., physical strength, demographic features, sleep habit, attention performance, visual and cognitive ability, and risk perception ability, collected from interview and relevant tests.
Driving experience features	DE	The features including driver's driving history (e.g., average annual mileage, years of driving, etc.) and driving characteristics (e.g., distance headway, brake activations, etc.) on an entire trip, from which the event was identified.
Driving behavioral features	DB	The features of driver behaviors, which occurred within seconds prior to the event including the ones that contribute to the crash or near crash, e.g., risky behaviors, head position, attention, etc.
Surrounding environmental features	SE	The features relevant to vehicle's surrounding environment when the event occurred, e.g., traffic condition, road condition, weather, etc.
Vehicle's performance features	VP	The features of vehicle's performance and configuration, e.g., tire pressure, malfunction, driving assistance, etc.
Surrounding vehicles' features	SV	The features of the surrounding vehicles, e.g., vehicle types, kinetic features, interactions with the subject vehicle, etc. The features are collected during the epoch starting 20 s before the event started and ending 10 s after the event started.
Vehicle's kinetic features	VK	The features that describe vehicles' movement, e.g., velocity, acceleration, etc. The features are collected during the epoch starting 20 s before the event started and ending 10 s after the event started.

(Nelsen, 2007). The multivariate cumulative distribution function is derived as a corresponding Copula function, which is obtained as:

$$F(x_1,\ x_2,\ \cdots,\ x_n) = C(F_1(x_1),\ F_2(x_2),\ \cdots, F_n(x_n)) \tag{5}$$

where $C(\cdot)$ indicates a Copula function which represents a multivariate dependence structure, and $F_i(x_i)$ denotes the marginal distribution of the random variable x_i. In this study, Gaussian Copula is applied as the Copula function, which can effectively handle the dataset with a mix of continuous and discrete variables (Zilko and Kurowicka, 2016). The multivariate cumulative distribution in a form of the Gaussian Copula function can be expressed as:

$$F(x_1,\ x_2,\ \cdots,\ x_n) = C_\phi(F_1(x_1),\ F_2(x_2),\ \cdots, F_n(x_n); \theta) \tag{6}$$

where $C_\phi(\cdot)$ indicates the Gaussian Copula function, and θ is the set of the Copula parameters corresponding to the edges between the variables (i.e., nodes) in the BN. The Copula parameter describes the correlation between two variables, which is measured by Kendall rank correlation coefficients in this study.

3. Data preprocessing

3.1. Feature extraction

Feature extraction aims to identify the candidate features from dataset, which are likely to have contributions to the vehicles' on-road crashes. Feature extraction involves the sub-steps of data cleaning, feature categorization, and feature transformation. Data cleaning aims at improving the quality of data by removing the inconsistencies and errors from the dataset (Rahm and Do, 2000). The features with missing data and duplicate features are eliminated from the dataset. Savitzky-Golay filter (Savitzky and Golay, 1964) is employed to remove the noises from the time-series features (e.g. velocity, acceleration, etc.). Feature categorization is an effective data processing method that handles the dataset with a large scale of features from multiple aspects. The candidate features are classified into seven categories, which are summarized in Table 1. Herein, all the features are collected from, or relevant to, the subject vehicle (which is a regular car in Section 5).

Feature transformation aims to transfer the existing features to the forms that can enhance the quality of knowledge extracted from data (Kusiak, 2001). In this study, the candidate features are classified into categorical features and numerical features, and the categorical features are further classified into nominal features and ordinal features. The nominal features refer to the features without inherent order, e.g., road surface condition ('dry', 'wet', 'icy', etc.), while the ordinal features indicate the features with an implied order, e.g., sleeper type ('light', 'normal', or 'heavy'). The numerical features are comprised of discrete features and continuous features. The discrete features indicate the features categorized into a classification, e.g., annual driving miles ('<5, 000', '5000–10,000', etc.), while the continuous features are the features measured on a continuum or scale, e.g., traffic density, vehicle's velocity, etc.

To improve computational efficiency, one-hot encoding is used to treat nominal features, which represents each state of a feature as a binary vector (Harris and Harris, 2010). Also, integer encoding is used to handle with ordinal features and discrete features, which assigns an integer value to each state of a feature. Meanwhile, the time-series features (e.g., velocity, acceleration, etc.), which belong to continuous features, shall be statistically described in order to reduce the feature's dimensionality. In this study, the time-series features are represented by the statistical descriptions (i.e. mean, standard deviation, maximum, minimum, range, kurtosis, skewness, 0.25 quantile, 0.5 quantile, and 0.75 quantile) of a temporal sequence (Shi et al., 2019; Chen et al., 2019).

3.2. Event labeling and resampling

This study only considers vehicles' on-road crashes, excluding the crashes that occur at the junction areas (i.e., intersection, entrance/exit ramp, interchange area, rail grade crossing, array access, driveway, etc.). The events (i.e., target) are labeled with non-crash, near-crash, and crash. Herein, non-crash denotes any incident or maneuver within the bound of normal driving scenarios, which also includes the baseline events selected as comparison with crash and near-crash events. Near-crash refers to any circumstance that requires a rapid evasive maneuver by the subject vehicle to avoid a crash. Near-crash has the following four criteria: 1. The subject vehicle must not contact with any surrounding vehicles; 2. An evasive maneuver is required by the subject vehicle; 3. The evasive maneuver must be conducted rapidly; and 4. The maneuver conducted by the subject vehicle must not be pre-meditated. Crash denotes any contact between the two vehicles at any velocity in which kinetic energy is measurably dissipated or transferred (see (Hankey et al., 2016) for the details of event labeling).

Resampling is applied to resolve class imbalance problem, which occurs when the sample size of any one class is much lower than that of the other classes (Galar et al., 2011). A dataset with class imbalance problem would lead to the decline in the prediction performance of a classifier (Japkowicz and Stephen, 2002). The classifier tends to be more biased towards the class with majority of samples than the others, which results in the unsatisfactory prediction performance on the class with few samples. The dataset in this study has a severe class imbalance problem since the sample size of non-crash event is much larger than the other two types of event, as vehicles' crashes rarely occur in real-life. Consequently, as a promising resampling method, SMOTETomek, which integrates Synthetic Minority Over-sampling Technique (SMOTE) and Tomek-Link, is applied to resolve the class imbalance problem (Chen et al., 2019). SMOTE is an over-sampling method synthetically propagating the samples of the minority class (Chawla et al., 2002). Tomek-Link is an under-sampling method that eliminates the noisy and borderline samples of the majority class (Tomek, 1976). The resampling

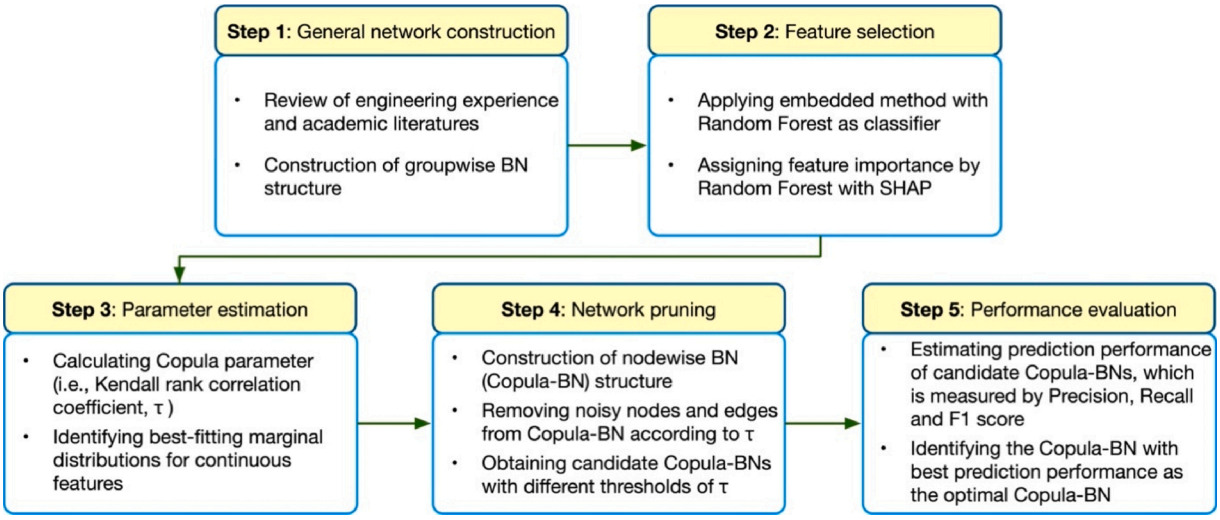


Fig. 2. Flowchart of developing Copula-Bayesian Network.

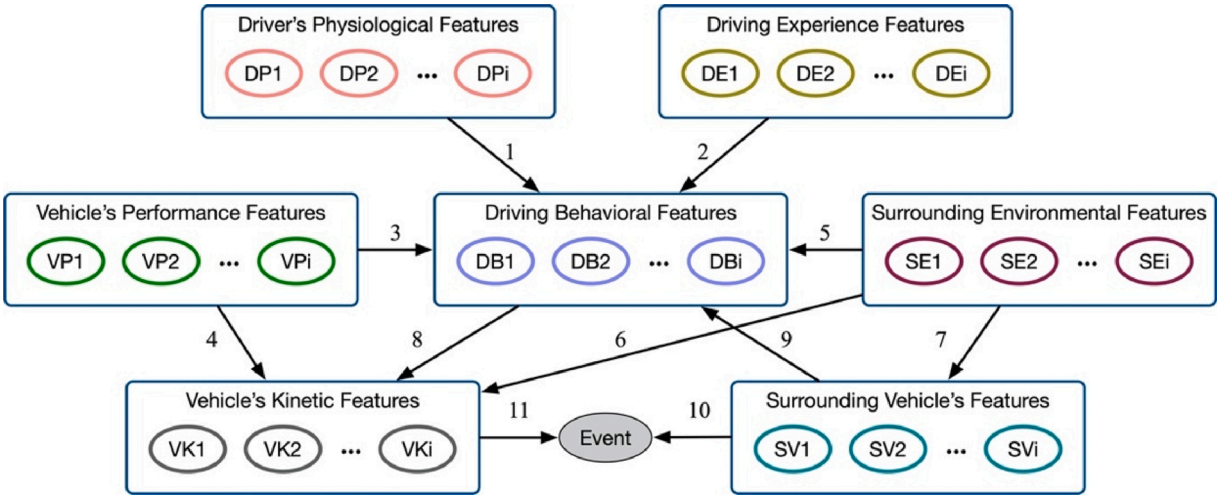


Fig. 3. Structure of general Bayesian Network[1].
[1] Nodes named DP_i, DE_i, VP_i, DB_i, SE_i, VK_i, SV_i refer to the features of each category, respectively.

method applies one-against-all (OAA) scheme (Rifkin and Klautau, 2004) considering the multi-class scenario of the dataset. The resampling dataset shall be randomly split into training set and test set. The BN is trained and validated on the training set containing 80 % of the samples in the resampling dataset, while the test set consists of the remaining samples, which offers unbiased evaluation for the trained BN.

4. Development of Copula-Bayesian Network

As illustrated in Fig. 1, the development of Copula-BN contains five major steps, namely, general network construction, feature selection, parameter estimation, network pruning, and performance evaluation. Herein, Fig. 2 generally presents the methodology and techniques applied in each step of developing Copula-BN. In accordance with the flowchart in Fig. 2, this section introduces the development of Copula-BN in detail.

4.1. General network construction

The general BN construction is proposed to build a groupwise dependency structure of feature categories, with each category as a group of nodes. The edges between groups of a general BN are built based on engineering experience and academic literatures, as based on which the edges between nodes are built within groups. Herein, no nodewise edge is built inside a group as it is assumed that there is no obvious intra-group causal relationship. The details of nodewise BN construction are introduced in Sections 4.2-4.4. The structure of the general BN is illustrated in Fig. 3, in which the rectangles indicate the node groups (i.e., feature categories) and the ovals within the rectangles denote the nodes (i.e., features). The dependent relations between feature categories represented by the groupwise edges are described in Table 2, as referring to a priori knowledge in previous studies. As a systematic failure, a crash event is usually caused by a series of errors (Ren et al., 2008). Such a general BN with multiple categories of features is able to effectively handle complex information, comprehensively explain the causal chain of a risky maneuver, and make causal inference of an event being more explainable and practical.

4.2. Feature selection

Feature selection provides a foundation for the construction of BN based on a large number of candidate features. By filtering out redundant and irrelevant features, feature selection reduces the computational cost, improves the prediction performance, and facilitates the

Table 2
Descriptions of relations between feature categories.

Edges[1]	Descriptions of relations	References
1	Driver's physiological features, e.g., gender, sleep habit, etc., are significantly correlated with hazardous driving behaviors, e.g., fatigue driving, distraction, excessive speed, etc.	(Rhodes and Pivik, 2011) (Hallvig et al., 2013) (Poó and Ledesma, 2013)
2	Driving experience in terms of driving skills and styles could have effects on driver's driving behaviors, especially decision making when encountering an unexpected incident on the road.	(Fischer et al., 2007) (Simons-Morton et al., 2019)
3, 4	Vehicle's unsatisfactory performance might lead to possible errors made by the driver, e.g., distraction, improper operation, etc. Vehicle's performance, e.g., power, tire condition, etc., also has effects on vehicle's kinetic features	(Krahé and Fenske, 2002) (Weng and Meng, 2012) (Rahimi et al., 2020)
5, 6, 7	Adverse environment, e.g., slippery road condition, adverse weather, heavy traffic density, etc., are highly associated with hazardous driving behaviors. Vehicle's kinetic characteristics are more or less dependent on surrounding environmental factors, e.g., traffic flow, traffic density, etc.	(Liu and Wu, 2009) (Weng and Meng, 2012) (Chong et al., 2013) (Hamdar et al., 2016) (Das et al., 2019)
8	Driver's driving behaviors affect vehicle's kinetic features, such as vehicle's trajectory and movement, via driver's control of vehicle.	(Zhu et al., 2017) (Yang et al., 2018)
9	Kinetic features of surrounding vehicles, e.g., velocity, acceleration, etc., and the interactions between subject vehicle and its surrounding vehicles, e.g., following gaps, time-to-collision, etc., could have impact on driver's driving behaviors.	(Farah et al., 2012) (Ali et al., 2020)
10, 11	Occurrence and severity of a vehicles' crash are directly determined by the kinetic features, e.g., velocity, acceleration, etc., of the vehicles involved in the crash.	(Shi et al., 2019) (Chen et al., 2019) (Shi et al., 2020) (Chen et al., 2021)

[1] Indices of groupwise edges are shown in Fig. 3.

explanation of the causal relationship between features and target (Molina et al., 2002; Guyon and Elisseeff, 2003). Herein, feature selection is proposed to select candidate key features from the candidate features. The candidate key features are selected according to the importance assigned to each feature, which is measured by the correlation between the feature and target (i.e., event's label) (Razmjoo et al., 2017). The steps of general network construction (as introduced in Section 4.1) and feature selection effectively handle a large scale of features and make the BN become more explainable than the BNs developed by conventional structure learning methods.

In this study, embedded feature selection method (i.e., embedded method) is employed to select candidate key features. Embedded method involves feature selection into the training process of a classifier, upon which the features are sorted based on the feature importance (Lal et al., 2006). Embedded method is computational efficient since the method uses a reliable measure of feature importance and avoid retraining the classifier for feature subsets (Peralta and Soto, 2014). Besides, the prediction accuracy of the classifier provides a solid basis for determining the number of candidate key features, which is superior to a rough estimation of the number as used by the conventional methods.

Herein, Random Forest classifier is employed as the classifier incorporated with the embedded method, which constructs an ensemble of decision trees using bagging as bootstrap aggregation method (Breiman, 2001). The embedded method with Random Forest classifier is able to attain satisfactory stability performance and excellent prediction performance (Chen et al., 2020). The stability performance denotes the sensitivity of a feature selection method to the robustness of feature preferences and variations of training dataset (Kalousis et al., 2007). In this study, the feature importance assigned by Random Forest classifier is measured by SHAP importance, which can illustrate the contributions

Table 3
Candidate marginal distributions.

Distributions	Probability density function	Parameter sets[1]
Beta[2]	$f(y,\ a,\ b) = \frac{\Gamma(a+b)y^{a-1}(1-y)^{b-1}}{\Gamma(a)\Gamma(b)}$	$(a,\ b,\ loc,\ scale)$
Exponential	$f(y) = \exp(-y)$	$(loc,\ scale)$
Gamma[2]	$f(y,\ a) = \frac{y^{a-1}\exp(-y)}{\Gamma(a)}$	$(a,\ loc,\ scale)$
Log-normal	$f(y,s) = \frac{1}{sy\sqrt{2\pi}}\exp(-\frac{log^2(y)}{2s^2})$	$(s,\ loc,\ scale)$
Normal	$f(y) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2})$	$(loc,\ scale)$

[1] $y = \frac{x - loc}{scale}$.
[2] $\Gamma(\cdot)$ refers to Gamma function.

of each feature to target comprehensively, as introduced earlier in Section 2.1. Herein, the embedded method is respectively applied on the training sets with each category of features and selects the candidate key features for each category.

4.3. Parameter estimation

This step aims to estimate Copula parameter and identify the best-fitting marginal distributions for the continuous features based on the samples in the training set. Copula parameter measures the dependency between variables and provides a basis for the dependency structure of Copula-BN, as mentioned in Section 2.2. The Copula enables the BN to efficiently process not only discrete features but also continuous features. In this study, Copula parameter is determined by Kendall rank correlation coefficient (τ) (Abdi, 2007; Pan et al., 2019), which is obtained as:

$$\tau = \frac{2}{n(n-1)}\sum_{i<j} sgn(x_i - x_j)sgn(y_i - y_j) \tag{7}$$

where n refers to sample size, and $x_.$ and $y_.$ indicate the observations of the variables X and Y, respectively. The best-fitting marginal distribution of a continuous feature is selected from five candidate distributions, which are summarized in Table 3. Herein, Sum of Square Error (SSE) is employed as the fitting criterion, and the candidate distribution with the lowest SSE would be identified as the best-fitting marginal distribution for the feature. Kolmogorov-Smirnov test (K–S test) (Massey, 1951) is applied to validate the goodness-of-fit achieved by the best-fitting marginal distribution. The K–S statistic measures the distance between the cumulative distribution function of the best-fitting distribution and the empirical distribution function of samples. Herein, the hypothesis that the two distributions are identical is rejected when the K–S statistic is larger than its corresponding P-value.

4.4. Network pruning

A nodewise Copula-BN is obtained by incorporating the candidate key features into the general BN, in which the probabilistic dependency between the variables (including features and event) is presented in a form of Copula function. However, the embedded method assigns importance to features only considering the relationships between features and target (i.e., event) instead of the correlation between features. Hence, such a nodewise BN with the candidate key features still has a complex structure with noises. The step of network pruning, which is inspired by the pruning technique used by decision trees (Rokach and Maimon, 2008), is proposed to remove the noisy nodes and edges in a nodewise BN. Network pruning is able to alleviate overfitting problem, increase computational efficiency, and improve prediction accuracy, by reducing the complexity of BN. In this study, Kendall rank correlation coefficient (τ) (see Eq. 7) is used to quantify the correlations between the variables. Accordingly, candidate Copula-BNs with different thresholds

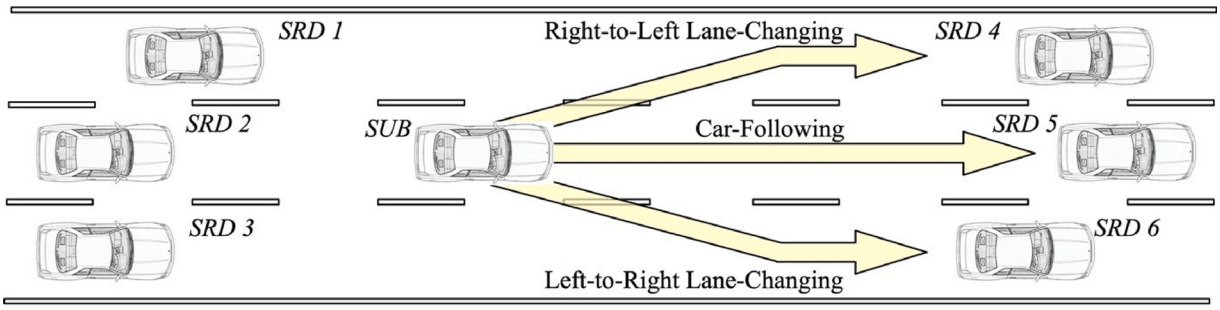


Fig. 4. Illustration of lane-changing and car-following maneuvers.

of, where noisy nodes and edges with τ lower than the threshold, are pruned. Herein, the candidate thresholds of τ are identified from 0.20 to 0.45 with increment of 0.05.

4.5. Performance evaluation

The candidate Copula-BNs are evaluated from the perspective of prediction performance. The Copula-BN that has better prediction performance than the other candidate BNs is identified as the optimal Copula-BN. The features of the optimal Copula-BN are regarded as key features. Those candidate Copula-BNs with higher threshold of τ are more likely to misidentify some key variables or dependency relationships between variables as noises, which might lead to the loss of critical information and have a negative impact on network's prediction performance. Consequently, satisfactory prediction performance can be seen as a trade-off between network complexity and information loss of a BN. The prediction performance of the candidate BNs is evaluated using the samples of the test set. Herein, prediction performance is measured by Precision (P), Recall (R), and F1 score (F1), which can be obtained as follows.

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2(P \cdot R)}{P + R} \tag{10}$$

where True Positives (*TP*) refers to the samples correctly predicted as positives, False Positives (*FP*) refers to the negative samples incorrectly predicted as positives, and False Negatives (*FN*) refers to the positive samples incorrectly predicted as negatives (Powers, 2011). Accordingly, the Copula-BN illustrating the causation of risky maneuver can be obtained, based on which simulations are conducted for the casual inference of the risky maneuver.

5. Application of Copula-Bayesian Network

5.1. Data source

The dataset from the Second Highway Research Program (SHRP2) Naturalistic Driving Study (NDS) database is used for validation in this study. The SHRP2 collected over a total of 4300 years of naturalistic driving data from nearly 3400 participant drivers around the United States between 2010 and 2013. The database information including video views (i.e., driver behaviors, surrounding environment, etc.), vehicle behaviors (e.g., speed, acceleration, etc.), and sensory information (e.g., radar, GPS, etc.) are collected by a data acquisition system (DAS) installed in the participant vehicles. The features are extracted and digitalized by the research team from Virginia Tech Transportation Institute (VTTI) (Hankey et al., 2016; Sears et al., 2016, 2019). In this study, a total of 475 features are extracted from the SHRP2 NDS database as the candidate features, which are then categorized into the seven categories as mentioned in Section 3.1.

Table 4
Resampling results.

Maneuvers		Events			Total
		Non-crash	Near-crash	Crash	
LC	Before resampling	562	317	13	892
	After resampling	553	326	150	1029
CF	Before resampling	12,560	1778	72	14,440
	After resampling	3465	1780	400	5645

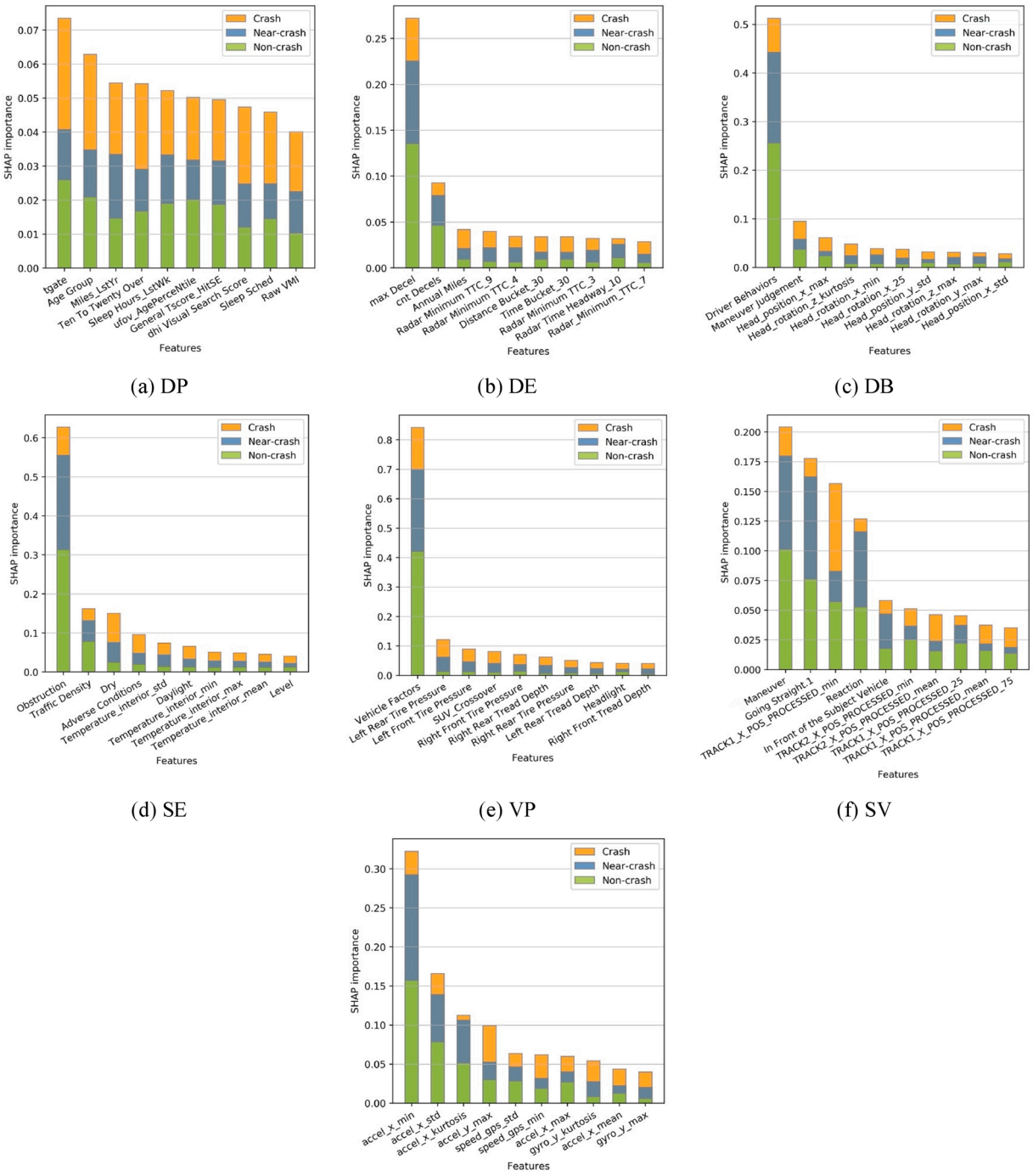
Both LC and CF maneuvers are extracted in accordance with the variable named '*pre-incident maneuver*' in the SHRP2 NDS dataset provided by VTTI, which denotes the last type of driving maneuver that the subject vehicle engaged in just prior to or at the time of the event. Also, the extracted events are manually checked according to the corresponding video footages in case of misclassification. The events with missing data or abnormal data are eliminated. Fig. 4 shows both LC and CF maneuvers, where *SUB* denotes the subject vehicle and *SRD·* indicates the surrounding vehicle if it exists. LC refers to the maneuver that the subject vehicle traverses to an adjacent lane from its present lane, while CF denotes the maneuver that the subject vehicle keeps following the preceding vehicle in a lane of traffic stream. The case studies in Section 5 only take the scenarios in which the vehicles (including the subject vehicle and its surrounding vehicles) involved are regular cars into consideration. Both LC and CF datasets with the 475 features in each dataset are extracted from the SHRP2 NDS database. As presented in Table 4, before resampling, both LC and CF datasets have a severe class imbalance problem since the size of non-crash events is much larger than that of crash event (herein, each event is seen as a sample). Therefore, a total of 1029 LC samples and 5645 CF samples are retrieved by using SMOTETomek, which effectively resolves the class imbalance problem.

5.2. Network development

To make this paper concise, this section only uses LC maneuver to demonstrate the procedure of developing a Copula-BN.

5.2.1. Feature selection

First of all, the candidate features in each feature category are sorted by the decreasing SHAP importance as calculated based on Random Forest classifier. The features assigned with higher SHAP importance value are deemed to be more influential in the outcome of event. Fig. 4 shows the first ten important features with corresponding SHAP importance belonging to each feature category based on the LC dataset. As presented in Fig. 5, SHAP comprehensively takes the contributions of a feature to the three classes of label, namely, non-crash, near-crash, and crash, into account, which is superior to the conventional feature



(a) DP (b) DE (c) DB

(d) SE (e) VP (f) SV

(g) VK

Fig. 5. SHAP importance of candidate features for lane-changing maneuver.

importance measures that spans across the entire label but not on a specific class.

Then, the Random Forest classifier is trained by the samples with important features, and the candidate key features are identified based on the prediction accuracy of the classifier. Fig. 6 illustrates the prediction accuracy with respect to the ranking of important features in each category using the LC dataset, upon which the candidate key features belonging to a category can be determined. For example, as presented in Fig. 6, the classifier trained on the samples with VK features in the LC dataset attains the maximum prediction accuracy when the feature ranking equals to five. Therefore, the first five important features in VK category, as shown in Fig. 5(g), are selected as the candidate key features for risky LC maneuver.

5.2.2. *Parameter estimation*

The Copula parameter is calculated as the Kendall rank correlation coefficient (τ) between variables (including features and event), if two variables have an edge to connect each other in a BN. Fig. 7 presents the heatmap of the Kendall rank correlation coefficients between the variables for LC maneuver. The correlation coefficients between the variables also provide a reference for the step of network pruning. The best-fitting marginal distribution is identified from the candidate

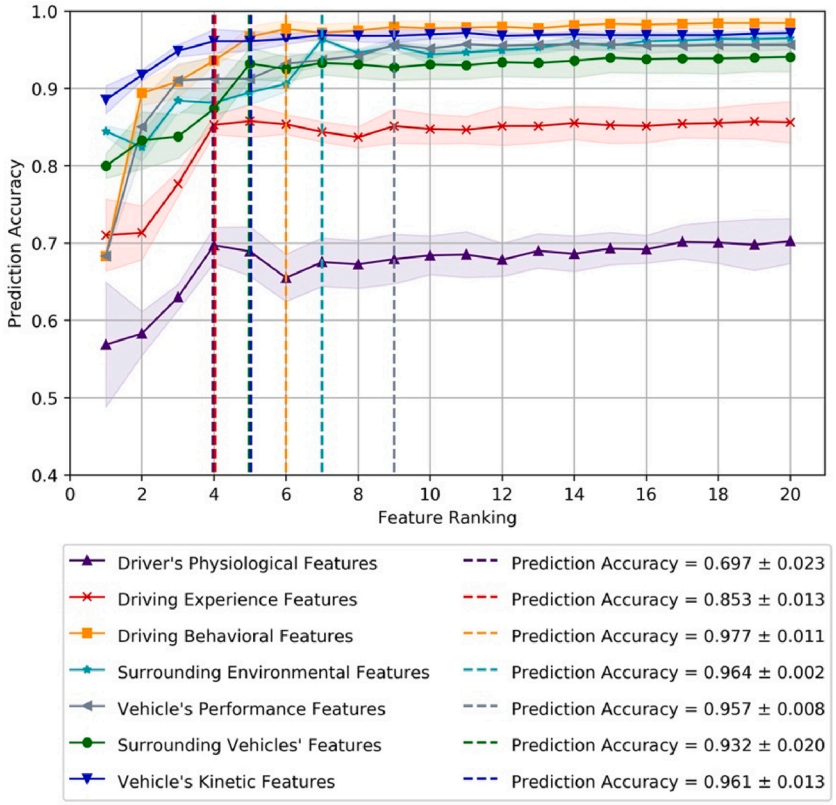


Fig. 6. Prediction accuracy with respect to feature ranking for lane-changing maneuver.

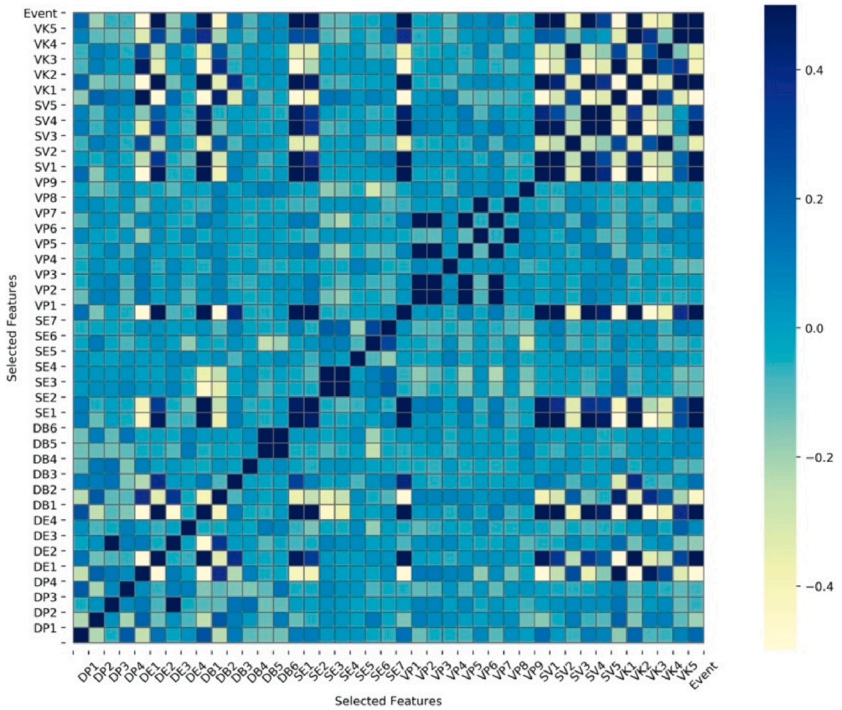


Fig. 7. Heatmaps of Kendall ranking correlation coefficients between variables for lane-changing maneuver.

distributions for each continuous feature. Herein, SSE is employed as the criterion to identify the best-fitting marginal distribution. Then, K–S test is used to validate the marginal distributions. Table 5 lists the best-fitting marginal distribution of each continuous feature with its parameters, as well as the validation results. Table 5 only records the continuous features denoted by the nodes in Fig. 8(a) when the threshold of Kendall rank correlation coefficient (τ_t) equals 0.20. As presented in Table 5, the K–S statistic of each feature is smaller than its corresponding P-value, which demonstrates the satisfactory goodness-of-fit achieved by the marginal distributions and provides a solid

Table 5
Marginal distribution identification and validation of continuous features for lane-changing maneuver.

Features	Distributions	Parameters[1]	SSE	KS stat.	P-value
DE1	Beta	(9.70×10^7, 1.64, -2.22 $\times 10^7$, 2.22×10^7)	3.762	0.109	0.171
DE3	Log-normal	(0.351, -1.736, 3.719)	0.024	0.205	0.325
DB3	Beta	(1.014×10^9, 4.083, -1.043×10^{11}, 1.043×10^{11})	7.8×10^{-5}	0.329	0.460
SV3	Beta	(0.439, 45.426, -6.055, 984.542)	1.7×10^{-3}	0.160	0.350
SE2	Normal	(2.960, 1.188)	0.022	0.287	0.321
VK1	Beta	(31.698, 0.924, -16.773, 16.790)	2.679	0.114	0.137
VK2	Exponential	(0.008, 0.083)	5.156	0.086	0.428
VK3	Normal	(-0.151, 2.043)	0.089	0.197	0.375
VK4	Log-normal	(0.860, -0.009, 0.148)	1.794	0.086	0.430
VK5	Gamma	(1.347, -0.080, 8.163)	0.002	0.083	0.373

[1] Please refer to Table 3 for the parameter set of each distribution.

foundation for the construction of Copula-BN.

5.2.3. Network pruning

Fig. 8 shows the structures of the candidate BNs with different thresholds of Kendall rank correlation coefficient (τ_t) between the features for risky LC maneuver. The heatmaps of the correlation coefficients are presented in Fig. 7. Then, each candidate BN shall be validated by the test set, from which the BN with the best prediction performance is selected as the optimal BN. Meanwhile, the features constituting the optimal BN are identified as the key features to the risky maneuver. The descriptions of the candidate key features for both risky LC and CF maneuvers when τ_t equals 0.20 are listed in Appendix A.

5.2.4. Performance evaluation

The optimal BN is supposed to have a capacity of identifying risky events (i.e., near-crash and crash events) precisely given a dataset. Consequently, two additional criteria are set up for the performance evaluation. First, as compared to non-crash events, more attention shall be put on the prediction performance of both near-crash and crash events, which reflects the ability of a BN identifying the samples of non-majority classes. Second, Recall (R), which indicates the ratio of the total number of relevant samples that are retrieved, shall be given more weight than the other two measures, especially of both near-crash and crash events.

Table 6 presents the prediction performance of each candidate BN with respect to risky LC maneuver. According to the abovementioned criteria of prediction performance, the BN with τ_t equal to 0.35 is selected as the optimal BN for risky LC maneuver. As compared to the other candidate BNs, the optimal BN not only achieves higher general prediction accuracy, but also higher prediction performance on both near-crash and crash events. It means that the optimal BNs could detect a potential LC crash more precisely and provide a better explanation for the causation of a crash accident.

5.3. Discussion

5.3.1. Network evaluation

Network evaluation is proposed to verify the superiority of the optimal Copula-BNs. Herein, the evaluation is conducted from two perspectives, namely, structure performance evaluation, and comparison to conventional BNs with discretized features. Using the proposed method of developing Copula-BN, the optimal Copula-BNs for risky LC and CF maneuvers are generated as Figs. 9(a) ($\tau_t = 0.35$) and 9(b) ($\tau_t = 0.40$), respectively. The descriptions of the key features for both two maneuvers can be found in Appendix A.

5.3.1.1. Structure performance. In this sub-section, the structure performance of the Copula-BNs is evaluated from qualitative and quantitative perspectives, respectively. From a qualitative perspective, as a result from the step of feature selection (see Sections 4.2 and 5.2.1), the prediction accuracy can generally reflect the contribution of feature category to the outcome label of event. For examples, as presented in Fig. 6, it can be found that the prediction accuracy achieved by the classifier trained with Driver's Physiological (DP) features and Driving Experience (DE) features is less satisfactory as compared with the other feature categories for LC maneuver, which means a lower contribution made by those two categories to the event. In a DAG like Copula-BN, the node of a feature with less contribution to the label is less likely to have a direct connection to the target node. Hence, the prediction accuracy resulting from the step of feature selection (e.g., Figs. 6) manifests the rationality of the pre-determined general network (i.e., Fig. 3) from a general perspective.

From a quantitative perspective, structure performance reflects the complexity of a BN, which can be measured by Bayesian model selection criteria. The BN with a lower complexity would have less chance to have overfitting problem during training process (Cawley and Talbot, 2007). In this sub-section, three Bayesian model selection criteria, namely, K2 score (Cooper and Herskovits, 1992), Bayesian Dirichlet equivalent uniform (BDeu) (Buntine, 1991), and Bayesian Information Criterion (BIC) score (Schwarz, 1978) are used to measure the structure performance of a Copula-BN. The Copula-BN with a lower complexity would attain a higher criterion score. Table 7 summarizes the structure performance of the candidate Copula-BNs with respect to LC and CF maneuvers. As listed in Table 7, the criterion scores of the candidate Copula-BNs with respect to LC and CF maneuvers respectively remain at a high level when $0.30 \leq \tau_t \leq 0.45$ and $0.35 \leq \tau_t \leq 0.45$. Consequently, it can be concluded that the structure performance of the optimal Copula-BNs for both of the two maneuvers is satisfactory.

5.3.1.2. Network comparison. In this sub-section, the comparisons on prediction performance are made between Copula-BNs and the conventional BNs of which continuous features are discretized before training. Herein, as an efficient discretization method, *k*-means clustering algorithm (Cover and Hart, 1967) is used to classify the continuous features by partitioning the continuous value into *k* cluster. The number of clusters, *k*, is determined by Silhouette score (Rousseeuw, 1987), which achieves the global maximum at the optimal *k* where there is a trade-off between samples' intra-cluster distance and within-cluster distance. As comparisons, the structures of the conventional BNs with discretized features are identical to those of the optimal Copula-BNs, which are presented as Figs. 9(a) and (b) with respect to LC and CF maneuvers. The prediction performance of the two conventional BNs respectively for the two maneuvers are summarized in Table 8. As presented in Table 8, for both two maneuvers, the conventional BNs achieve less satisfactory prediction performance, especially on the near-crash and crash events, than the Copula-BNs. Thus, it can be concluded that Copula-BNs are superior to the conventional BNs in prediction performance, especially in predicting or identifying risky events (i.e., near-crash and crash events).

5.3.2. Feature inferences

5.3.2.1. General inferences. Based on the structure of the optimal Copula-BN and the marginal distribution of each variable (including feature and label), 500 samples are respectively generated for LC and CF maneuvers to simulate the causal relationships. As a graphical interpretation tool, cobweb plots are used to provide visual exploratory analysis. In cobweb plots, each sample is represented as a jagged line intersecting the vertical lines which denote the variables. The intersecting point on the vertical line refers to the percentile value of the variable assigned by the sample. Figs. 10 and 11 respectively show the

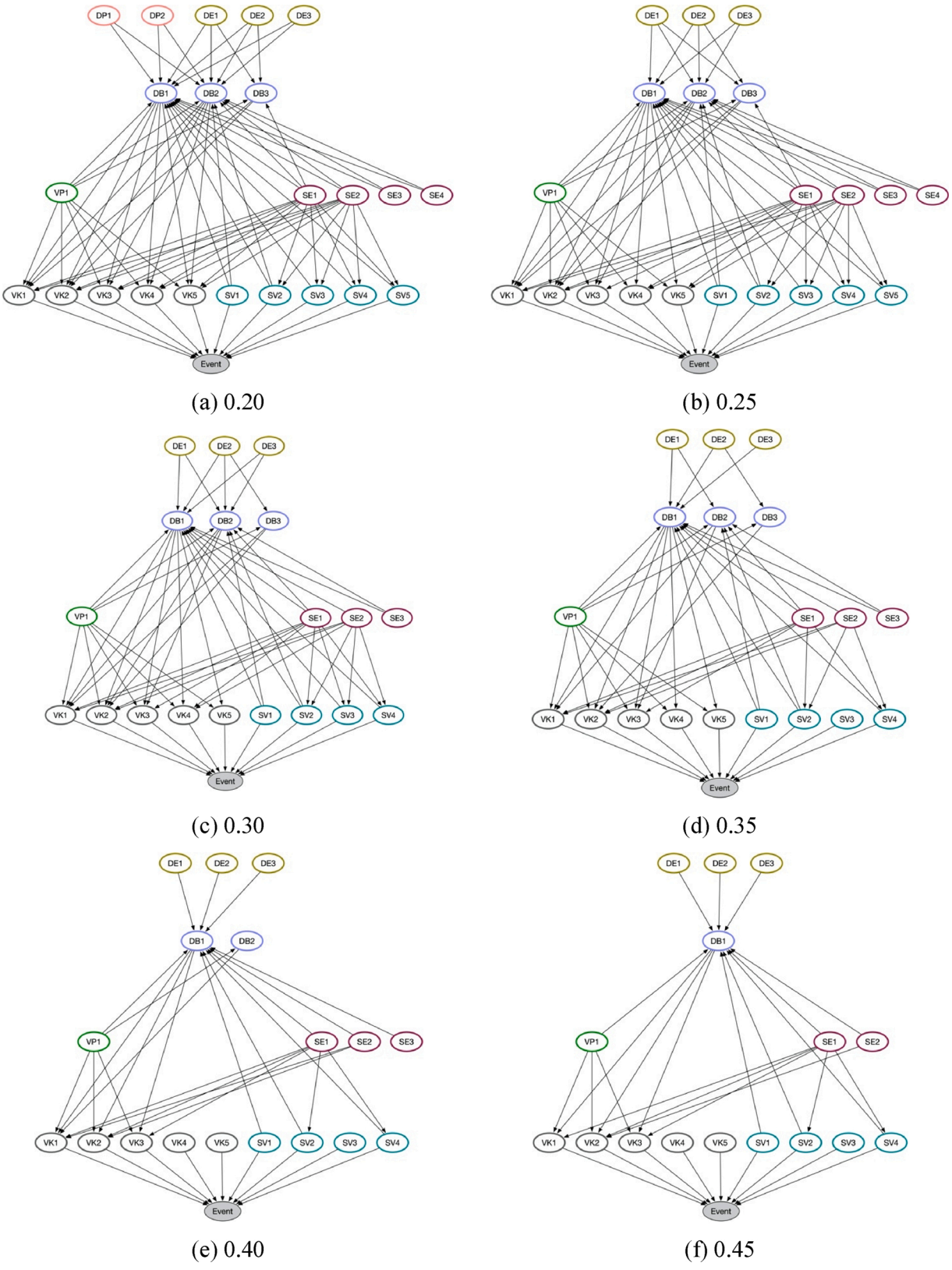


Fig. 8. Candidate Copula-Bayesian Networks of lane-changing maneuver.

cobwebs of non-crash, near crash, and crash events, as well as the probability density distribution of each variable, in terms of LC and CF maneuvers.

The cobwebs describe the joint distribution and dependence relations between the features from a general perspective, which provides an overview of the contributions made by the features, especially the continuous features, to the events. Interestingly, taking the vehicle's kinetic (VK) features of both LC and CF maneuvers as examples, it can be found that VK1 and VK3 with lower value, as well as VK2 with higher value, are more likely to result in a risky event (i.e., near-crash or crash).

Table 6
Prediction performance of candidate Copula-Bayesian Networks for lane-changing maneuver.

τ_t		Non-crash	Near-crash	Crash	Macro avg.	Micro avg.	Accuracy
	P	0.973	0.841	0.765	0.860	0.902	0.898
0.20	R	0.938	0.841	0.867	0.882	0.898	
	F1	0.955	0.841	0.813	0.870	0.899	
	P	0.991	0.877	0.849	0.905	0.935	0.932
0.25	R	0.947	0.905	0.944	0.928	0.930	
	F1	0.968	0.891	0.889	0.916	0.933	
	P	0.991	0.887	0.794	0.891	0.931	0.927
0.30	R	0.965	0.873	0.900	0.913	0.927	
	F1	0.978	0.880	0.844	0.900	0.928	
	P	**0.991**	**0.905**	**0.853**	**0.916**	**0.944**	**0.942**
0.35	R	**0.956**	**0.905**	**0.967**	**0.942**	**0.942**	
	F1	**0.973**	**0.905**	**0.906**	**0.928**	**0.942**	
	P	0.972	0.841	0.800	0.871	0.907	0.903
0.40	R	0.929	0.841	0.933	0.901	0.903	
	F1	0.950	0.841	0.862	0.884	0.904	
	P	0.995	0.831	0.750	0.860	0.911	0.903
0.45	R	0.929	0.857	0.900	0.895	0.983	
	F1	0.963	0.844	0.818	0.875	0.906	

As listed in Appendix A, the first three key VK features (i.e., VK1, VK2, and VK3) of LC maneuver are identical to those of CF maneuver, which denote the minimum, standard deviation, and kurtosis of vehicle acceleration in the longitudinal direction. However, the values of the three VK features for the two maneuvers respectively show the same trend but different amplitude as event changes from non-crash to crash. Besides, as compared to LC maneuver, the samples of CF maneuver fall in a much narrower value range of each VK feature. In other words, more noises can be seen in the distributions of the VK features for LC maneuver.

Such noises are inferred to result from the inherent complexity of LC maneuver. Accordingly, it is not reasonable to conduct risk classification for LC maneuver by simply clustering the VK features without considering additional features.

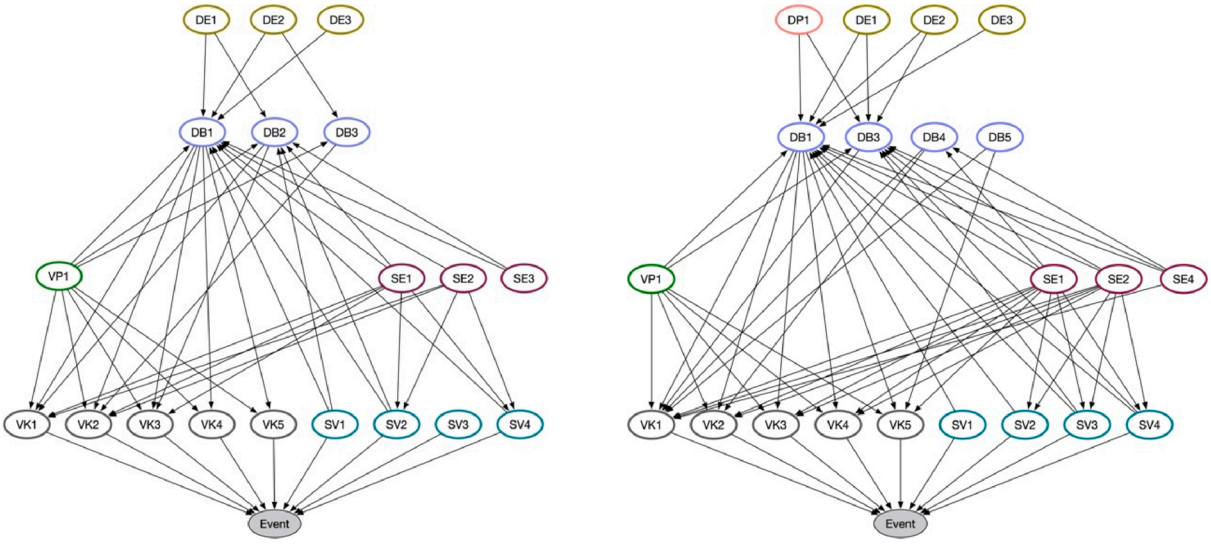
In summary, the cobwebs provide a general description of how the key features corelate with each other and make contributions to the event. As the examples mentioned above, the cobwebs illustrate the sensitivity of the VK features to the event, upon which the comparisons are made from both event-wise and maneuver-wise. Using the cobwebs, this sub-section discusses the roles that the acceleration-related features play respectively in risky LC and CF maneuvers and the impact of the noises in the distribution of the VK features for LC maneuver. However, the cobwebs have following limitations: first, the contributions made by the discrete features to the event cannot be well-described; second, the minor difference in the value distribution of some variables under different conditions is hard to be observed; and third, the cobwebs only provide empirical results based on generated samples rather than a quantitative measure of the relationships between variables. Therefore, in the following sub-section, forward and backward inference methods are proposed to address the above limitations.

5.3.2.2. Quantitative contributions. Forward inference, also known as predictive inference, reasons from given information about one or more variables to new beliefs of the others following the directions of the edges, while backward inference, also known as diagnostic inference, reasons in the opposite directions of the edges in a Copula-BN (Ding, 2010). Therefore, the forward and backward inferences could illustrate how the states of a variable are affected by its parent variables and how the variable contributes to its offspring variables via the causal structure of the Copula-BN. The dependence relationships between variables can be measured in a form of conditional probability given the predefined states of one or more variables.

Herein, contribution of a feature to the event is measured by the conditional expectation of the event given the states of the feature based on above-mentioned forward inference. Tables 9 and 10 respectively

Table 7
Structure performance of candidate Copula-Bayesian Networks.

τ_t	LC			CF		
	K2	BDeu	BIC	K2	BDeu	BIC
0.20	-2.55×10^5	-3.57×10^5	-3.58×10^6	-2.23×10^6	-1.52×10^6	-1.98×10^8
0.25	-2.15×10^5	-3.43×10^4	-3.22×10^6	-1.15×10^6	-9.33×10^5	-1.33×10^8
0.30	-1.41×10^4	-1.43×10^4	-2.16×10^5	-2.35×10^5	-1.13×10^5	-2.15×10^7
0.35	$\mathbf{-1.39 \times 10^4}$	$\mathbf{-1.38 \times 10^4}$	$\mathbf{-2.02 \times 10^5}$	-7.15×10^4	-9.29×10^4	-3.20×10^7
0.40	-1.35×10^4	-1.32×10^4	-2.00×10^5	$\mathbf{-7.93 \times 10^4}$	$\mathbf{-8.01 \times 10^4}$	$\mathbf{-2.19 \times 10^6}$
0.45	-1.20×10^4	-1.17×10^4	-1.31×10^5	-7.89×10^4	-7.84×10^4	-2.11×10^6

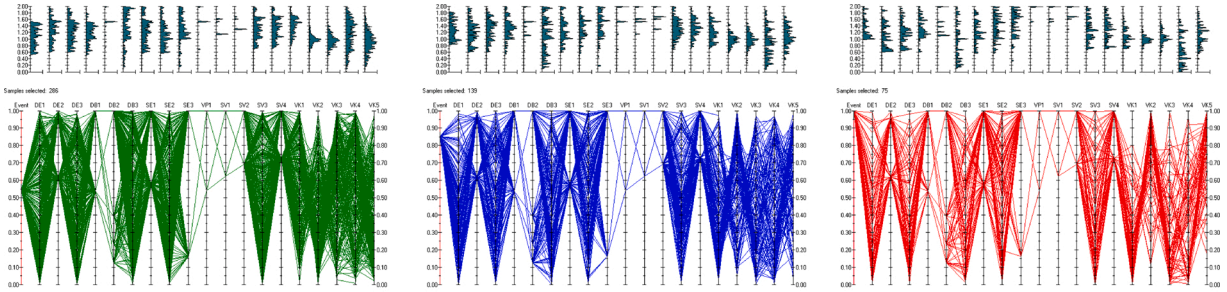


(a) Lane-changing maneuver ($\tau_t = 0.35$) (b) Car-following maneuver ($\tau_t = 0.40$)

Fig. 9. Optimal Copula-Bayesian Networks.

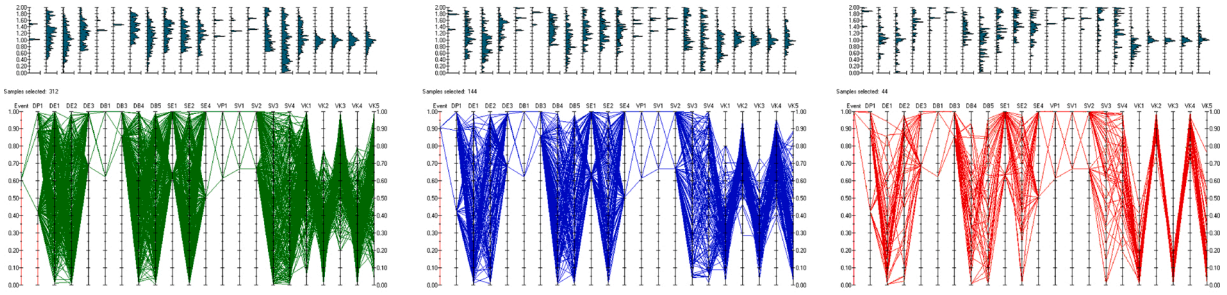
Table 8
Comparisons between prediction performance of Bayesian Networks.

BNs			Non-crash	Near-crash	Crash	Macro avg.	Micro avg.	Accuracy
LC ($\tau_t = 0.35$)	Conventional	P	0.890	0.963	0.920	0.924	0.917	0.913
		R	0.996	0.813	0.793	0.869	0.913	
		F1	0.942	0.881	0.852	0.892	0.910	
	Copula	P	**0.991**	**0.905**	**0.853**	**0.916**	**0.944**	**0.942**
		R	**0.956**	**0.905**	**0.967**	**0.942**	**0.942**	
		F1	**0.973**	**0.905**	**0.906**	**0.928**	**0.942**	
CF ($\tau_t = 0.40$)	Conventional	P	0.967	0.992	0.983	0.981	0.977	0.976
		R	0.998	0.954	0.866	0.940	0.976	
		F1	0.983	0.972	0.920	0.959	0.976	
	Copula	P	**0.994**	**0.985**	**0.925**	**0.968**	**0.987**	**0.987**
		R	**0.993**	**0.987**	**0.925**	**0.968**	**0.988**	
		F1	**0.993**	**0.986**	**0.925**	**0.968**	**0.985**	



(a) Non-crash (b) Near-crash (c) Crash

Fig. 10. Cobwebs of lane-changing maneuver.



(a) Non-crash (b) Near-crash (c) Crash

Fig. 11. Cobwebs of car-following maneuver.

Table 9
Conditional expectations of event given discrete features for lane-changing maneuver.

Features	Conditional expectations of event[1]			
	0	1	≥2	Dif.[2]
DE2	0.558 ± 0.712	0.672 ± 0.751	0.766 ± 0.773	0.208
DB1	0.375 ± 0.600	0.879 ± 0.711		0.504
DB2	1.26 ± 0.733	0.948 ± 0.754	0.377 ± 0597	0.883
VP1	0.376 ± 0.601	0.885 ± 0.775		0.509
SE1	0.470 ± 0.663	0.793 ± 0.771		0.323
SE3	0.620 ± 0.736	0.601 ± 0.729		0.019
SV1	0.337 ± 0.561	1.050 ± 0.754		0.713
SV2	0.507 ± 0.685	0.834 ± 0.778		0.327
SV4	0.500 ± 0.681	0.893 ± 0.781		0.393

[1] Original expectation:0.606 ± 0.731
[2] Dif. $= Max(\text{mean}) - Min(\text{mean})$

Table 10
Conditional expectations of event given discrete features for car-following maneuver.

Features	Conditional expectations of event[1]			
	0	1	≥2	Dif.[2]
DP1	0.346 ± 0.574	0.546 ± 0.681	0.761 ± 0.744	0.415
DE2	0.448 ± 0.641	0.557 ± 0.691	0.659 ± 0.727	0.211
DB1	0.182 ± 0.406	0.990 ± 0.692		0.808
DB3	0.417 ± 0.617	0.889 ± 0.753		0.472
VP1	0.214 ± 0.443	0.931 ± 0.712		0.717
SE1	0.360 ± 0.584	0.708 ± 0.727		0.348
SE4	0.493 ± 0.664	0.478 ± 0.657		0.015
SV1	0.447 ± 0.640	0.560 ± 0.692		0.113
SV2	0.399 ± 0.610	0.667 ± 0.724		0.268

[1] Original expectation:0.486 ± 0.659
[2] Dif. $= Max(\text{mean}) - Min(\text{mean})$

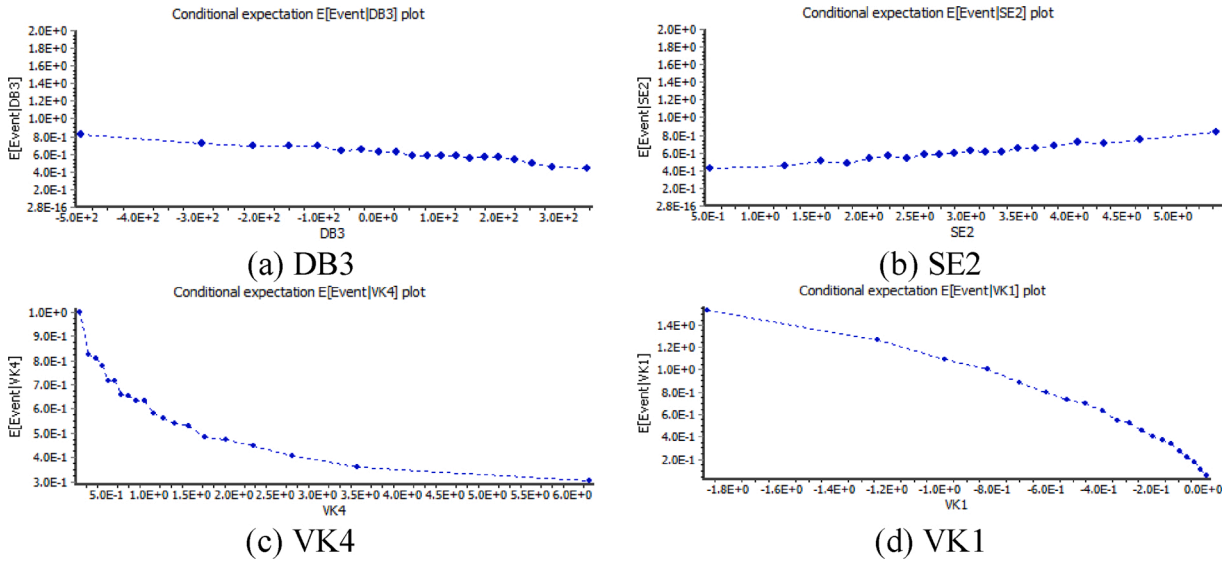


Fig. 12. Conditional expectations of event given continuous features for lane-changing maneuver.

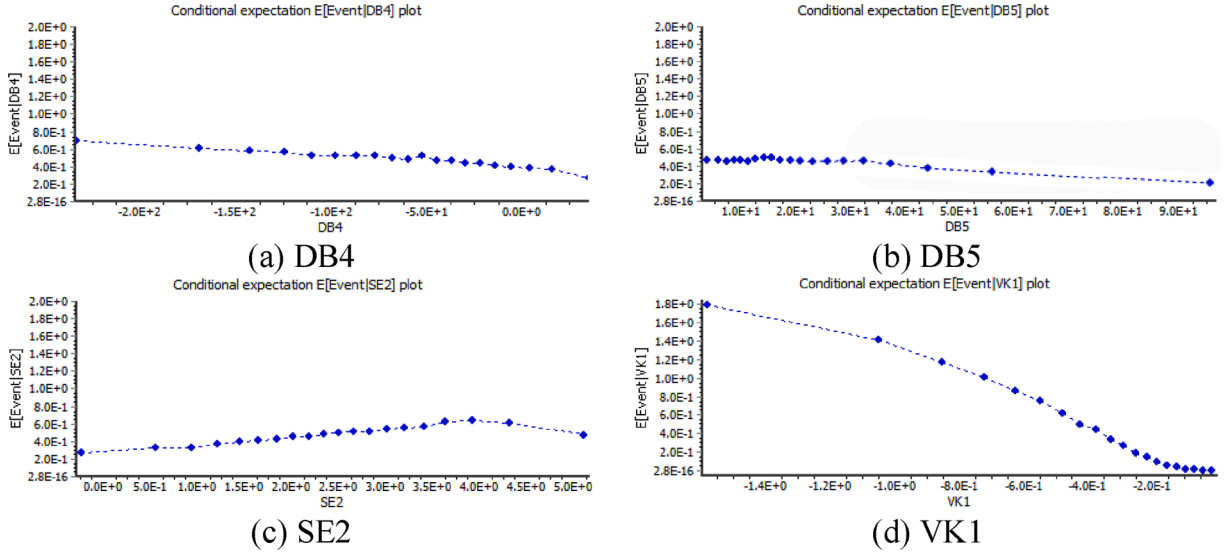


Fig. 13. Conditional expectations of event given continuous features for car-following maneuver.

summarize the contributions of discrete features to the event with respect to LC and CF maneuvers. Followings are interesting findings:

1 Despite that there is no key DP feature selected for LC maneuver, the feature DP1, i.e., driving sleepiness, shows a fairly high impact on CF safety, as the conditional expectation of the event increases strongly with increasing DP1 for CF maneuver. As compared to LC maneuver, CF maneuver is less attention-demanding since driver does not need to observe the movement of surrounding vehicles frequently. Driver is likely to feel fatigued or bored, especially for the one who has accumulated drowsiness, when performing a less attention-demanding (or less complex) task, such as following a car. Hence, driving sleepiness can have a significant contribution to risky CF maneuver.

2 Although the impacts of SE features are subtle as compared to the most other features, the difference between the impacts on LC and CF maneuvers raises concerns. It can be found that road surface condition (e.g., SE3 for LC maneuver) exerts more impact on LC safety, while traffic infrastructure (e.g., SE4 for CF maneuver) has more impact on CF safety. When a car is changing lane, adverse road surface condition can have negative effects on tire-road friction, especially when the car is not in a good condition, which might have a further impact on the car's kinetic features and lead to hazardous maneuvers, e.g., tire slip. Since CF is a basic maneuver easier than LC, driver is more likely to draw attention on the traffic infrastructure away from surrounding vehicles when following a car. Therefore, the SE features related to traffic infrastructure can have effects on DB features, especially DB3 (attention), and then further contribute to a risky CF maneuver along the edges in the BN.

3 As compared to CF maneuver, the SV features of LC maneuver show greater impacts on the event, especially the feature (i.e., SV1 for both two maneuvers) that judge whether the closest car conducted a risky maneuver. LC maneuver is more inherently complex than CF maneuver since LC maneuver involves vehicles in multiple lanes. Moreover, LC crash can be seen as a systematic failure, in which the movement of surrounding vehicles also plays an indispensable role. Hence, LC vehicle should not always take responsible for LC crashes, and the behaviors of the surrounding vehicles are supposed to be drawn much attention in the research on LC safety.

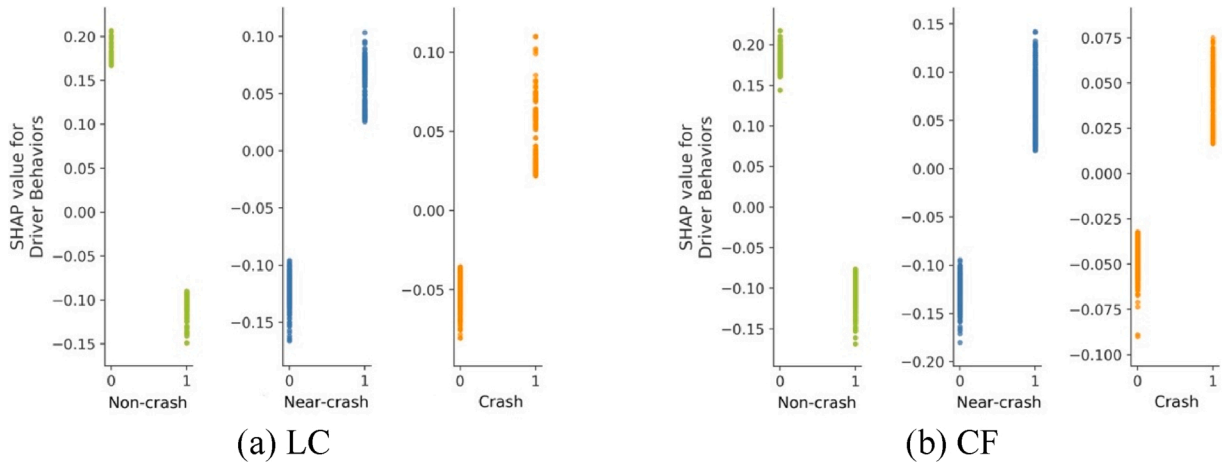


Fig. 14. SHAP values for DB1.

Also, several interesting findings from the conditional expectation plots of continuous features with respect to LC and CF maneuvers (see Figs. 12 and 13) are discussed as follows:

1 As shown in Figs. 12(a), 13 (a), and (b), the continuous DB features of both LC and CF maneuvers describe the rotation of driver's head. The figures illustrate that the vertical rotation of head (i.e., visual attention to upside or downside) when changing lane and the lateral rotation of head (i.e., visual attention to left or right side) when following a car are likely to increase crash risk. To ensure driving safety, driver is supposed to aptly put visual attention on left or right side to observe the surrounding vehicle when changing lane, while to always put the visual attention forward in longitudinal direction to keep a safe distance headway when following a car. However, despite that the above illustrations of the head rotation features seem reasonable, the contributions of those features to the risky maneuvers are limited as shown by the figures. This can be explained by driver sometimes offsetting visual attention from the supposed direction with rational purposes, e.g., observing traffic signs. Hence, it is not highly recommended to predict or detect potential risky maneuvers merely based on driver's visual attention which is collected from head rotation or eye-ball movement.
2 In this study, SE2 indicates traffic density for both LC and CF maneuvers. As presented in Fig. 12(b), the LC crash potential increases with the increasing traffic density all along. However, as illustrated in Fig. 13(c), for CF maneuver, there is an initial increase of the crash risk with increasing traffic density, and then a decrease when traffic density reaches a high level. Using forward inferences with the Copula-BNs, the causal relations between SE2 and VK features and between SE2 and SV features are investigated for both LC and CF maneuvers. It is found that, when traffic density reaches a high level (e.g., SE2 $>$ 4.0 in Figs. 12(b) and 13 (c)), the subject car is more likely to keep a low velocity, and the gaps between the subject car and its surrounding cars become narrower. For LC maneuver, even though the LC car moves in a low velocity, such an intricate maneuver that involve multiple surrounding vehicles is more likely to be more challenging and riskier when the gaps with the surrounding cars become narrower. Nevertheless, for CF maneuver, the low velocity as kept by the CF car can counteract the crash risk resulting from the narrower gaps with the surrounding cars.
3 As shown in Fig. 12(c), LC crash potential decreases with the LC car's increasing lateral acceleration (VK4) during an LC event. Nevertheless, in most previous studies on vehicle's LC trajectory planning for ADAS or automated vehicles (AVs), researchers would like to set vehicle's lateral acceleration as low as possible in order to make the LC trajectory a smooth profile and ensure the comfort of driver and passengers. It can be inferred that the lower lateral acceleration means a longer duration of an LC maneuver, which increases the chance of risky interactions with the surrounding vehicles in both present and adjacent lanes. Hence, it is of significance for researchers to carefully determine lateral acceleration and make a better trade-off between comfort and safety when planning trajectory for an LC vehicle.

5.3.3. *Inference verification*

The above inferences are conducted based on the structures of the optimal Copula-BNs. Herein, the SHAP value assigned by Random Forest classifiers, as introduced in Section 2.1, is used to quantify the contribution of each key feature to the event, which is proposed to verify the above feature inferences using the field samples in the training set. The

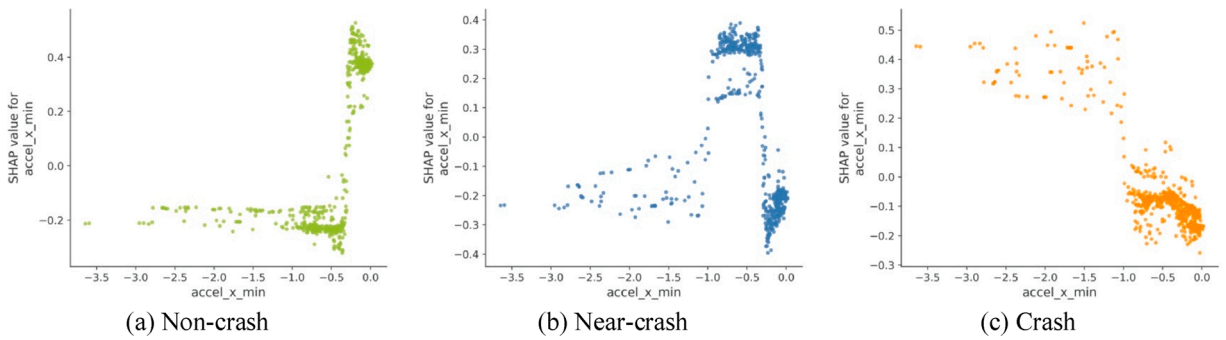


Fig. 15. SHAP values of VK1 with respect to lane-changing maneuver.

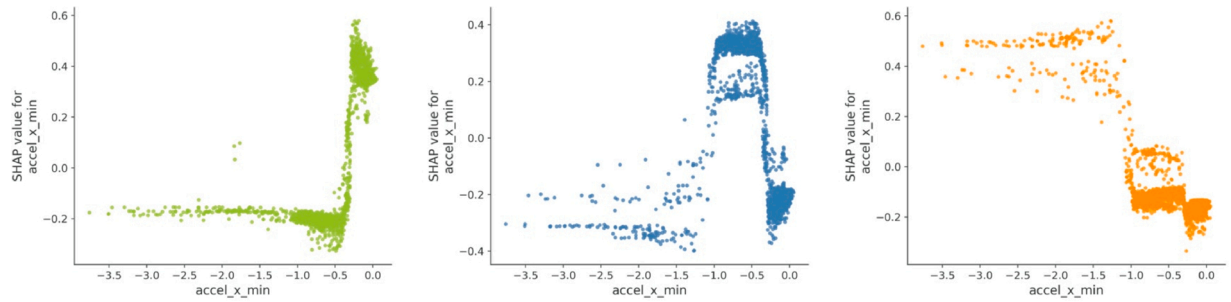


Fig. 16. SHAP values of VK1 with respect to car-following maneuver.

positive and negative SHAP values respectively indicate the positive and negative contributions of the features to the event. This section takes the contributions of DB1 (driver behaviors) and VK1 (accel_x_min) as the examples to demonstrate the verification of the inferences in terms of discrete and continuous features, respectively.

Fig. 14 shows the contributions of DB1 to each state of the event in a form of SHAP values with respect to LC and CF maneuvers. For both of the two maneuvers, as the state of DB1 changes from 0 to 1, the SHAP value on non-crash event turns negative, while the SHAP values on both near-crash and crash events turn positive. It means risky driver behaviors contribute positively to both near-crash and crash events, which have negative impacts on both LC and CF safety. Figs. 15 and 16 respectively present the contributions of VK1 to each state of the event with respect to LC and CF maneuvers. For both of the two maneuvers, with the increase of VK1 value, the SHAP values on non-crash, near-crash, and crash events turn positive in succession. The VK1 with low value contributes positively to crash event, while the VK1 with high value contributes positively to non-crash event. It can be inferred that the cars with higher deceleration recorded when conducting the maneuvers are more likely to be involved into a crash.

Therefore, from a qualitative perspective, similar results with regard to the contributions of the two features (DB1 and VK1) can also be obtained from the feature inference conducted based on the optimal Copula-BNs, as shown in Tables 9 and 10 and Figs. 12(d) and 13 (d). Also, the SHAP values of the key features (including DB1 and VK1) on the event for both LC and CF maneuvers are summarized in Appendix B, with which the feature inferences conducted on Copula-BN can be verified. However, as a limitation of using SHAP value to describe feature's contribution, SHAP cannot clearly and concisely describe the causal relations between features. Hence, the forward and backward inferences cannot be conducted on the basis of SHAP value. Despite that the contributions measured by SHAP values enable the verification of the feature inference using Copula-BN from a qualitative view, the above limitation makes SHAP inferior to the feature inference. Consequently, further verification of the probabilistic feature inference based on Copula-BN could be a direction of the future work.

6. Conclusions

6.1. Summary and conclusions

This study proposes a data-driven method to develop Copula-BN for investigating the causality of vehicles' crash on road. Besides, using a large-scale naturalistic driving dataset, the Copula-BNs are employed to provide insights into the causation of risky LC and CF maneuvers. The main steps of Copula-BN development as well as the brief description and purpose of each step are summarized as follows: 1. general network construction: building a groupwise dependency structure using feature categories to ensure the BN structure to be reasonable and explainable; 2. feature selection: identifying key features from candidate features for each category, which can reduce computational cost, improve prediction performance, and facilitate the explanation of causal relationships; 3. parameter estimation: estimating Copula parameter and identifying the best-fitting marginal distributions for features, which forms the basis for the dependency structure of Copula-BN; 4. network pruning: removing the noisy nodes and edges from BN to alleviate overfitting problem, reduce computational cost, and improve prediction accuracy; and 5. performance evaluation: evaluating the prediction performance of the candidate Copula-BNs as results from network pruning, upon which the optimal Copula-BN can be identified.

As two innovative techniques applied in the proposed method, SHAP is employed to measure the feature importance in the step of feature selection, while Copula function is used to describe the probabilistic causal relationship between the variables in a BN. Besides, the contributions of this study are summarized as follows:

1 This study offers a new approach to investigating the causality of road traffic crash in the era of 'Big Data'. The proposed method can effectively deal with a large-scale dataset with multiple features. To a certain extent, the Copula-BN as developed by the proposed method makes a trade-off between prediction and causality with 'Big Data'.
2 As compared with the conventional BNs, the Copula-BNs as developed in this study have the following advantages: 1) the step of general network construction forms a solid basis for the rationality of the BN's structure; 2) the steps of feature selection and network pruning can reduce the noises in the BN and increase prediction accuracy; and 3) the Copula embedded in the BN can efficiently handle not only discrete but also continuous variables.
3 The Copula-BNs are used to uncover the underlying process of how key features contribute to risky LC and CF maneuvers, based on which the similarity and difference between the causality of both two risky maneuvers are explored. The comparison between risky LC and CF maneuvers provides a valuable reference for crash risk evaluation, road safety policy-making, driving assistance system development, etc.

This study focuses on the maneuvers conducted on the road but excluding the ones at junction areas. The candidate features are extracted and classified into seven categories, while the events are labeled with non-crash, near-crash, and crash. The datasets for LC and CF maneuvers are respectively resampled by SMOTETomek technique to alleviate class imbalance problem. The SHRP NDS database is employed for application, from which 1029 LC samples and 5645 CF samples are obtained after resampling. Accordingly, the Copula-BNs are developed

for LC and CF maneuvers, respectively. Upon network evaluation, the optimal Copula-BNs for both two maneuvers demonstrate satisfactory structure performance and prediction performance. A total of 500 samples are respectively generated for LC and CF maneuvers using the optimal Copula-BNs, upon which the feature inferences are conducted to explain the causation of risky LC or CF maneuver from a probabilistic perspective. Interesting inference results are summarized as follows (see detailed inferences in Section 5.3.2):

1 As compared with CF maneuver, more noises are found in the distributions of vehicle's kinetic (VK) features for LC maneuver, which can be inferred as results of the inherent complexity of LC maneuver. Hence, it is not recommended to conduct risk classification for LC maneuver by merely clustering the VK features.
2 Although there is no key driver's physiological (DP) feature for LC maneuver, the feature of driving sleepiness has a fairly high impact on CF maneuver. This can be explained by driver being more likely to feel fatigued or bored when following a car, which can be seen as a less attention-demanding maneuver.
3 As surrounding environmental (SE) features, road surface condition exerts more impact on LC safety while traffic infrastructure has more impact on CF safety. Adverse road surface condition can have negative effects on the VK features of a LC car that can further result in a risky maneuver. Driver's attention might be easily distracted by traffic infrastructure when following a car.
4 As compared with CF maneuver, surrounding vehicle (SV) features show greater impacts on LC maneuver. Since LC is a complex maneuver that involves surrounding vehicles in multiple lanes, the behaviors of the surrounding vehicles shall be paid much attention in the research on LC safety.
5 As driving behavioral (DB) features, the features related to rotation direction of driver's head show a difference between the causation of risky LC and CF maneuvers. However, the contributions of those features to both two risky maneuvers are limited because driver sometimes offsets visual attention with legitimate purpose. Consequently, it is not highly recommended to predict risky driving maneuver merely based on driver's visual attention.
6 LC crash risk increases with increasing traffic density, while the CF crash risk decreases with increasing traffic density when the density reaches a high level. The narrow gaps with surrounding vehicles resulting from increasing traffic density make LC more challenging and riskier. Nevertheless, for CF maneuver, the low velocity can counteract the crash risk caused by the narrow gaps when traffic density is at high level.
7 LC crash risk decreases with the LC car's increasing lateral acceleration during an LC event, while lower lateral acceleration can ensure the comfort of passengers. Consequently, it is significant for the researcher to carefully determine lateral acceleration and trade off safety and comfort when planning trajectory for an LC maneuver.

6.2. Limitations and future work

The limitations and relevant future work of this study are summarized as follows:

1 Using BN to investigate the causality has following two limitations: first, BN structure is built upon the associations or correlations between variables, which might not adequately represent the causal relationships; second, Bayesian inference is conducted based on the statistical characteristics of a population, which might over-represent an individual sample.

Therefore, in the future, much effort can be made to disentangle causation from correlation and enhance generalization capability of BN.

2 This study merely discusses the contribution of an individual feature using the probabilistic inference conducted via the Copula-BNs. In future, the Copula-BNs could also be applied to investigate how a cluster of multiple features contributes to the event. Besides, with the development of data collection techniques, the categorization of features can be optimized and the 'hidden' nodes in the structure of Copula-BNs can be exposed in the future, which could provide a better insight into the causation of a crash.
3 As mentioned in Section 5.3.3, in this study, the feature inferences conducted via Copula-BNs are verified by SHAP from a qualitative perspective. However, SHAP cannot provide sufficient verification (e.g., quantitative verification) due to its limitations. Hence, in future, more efforts could be put on the further validation of feature inferences. Moreover, relevant field experiments could be designed to validate and solidify the feature inference.
4 Although SHRP NDS database is comprehensive as it collects multiple aspects of features, the dataset has a lot of noisy data and missing data, which have impacts on the development of Copula-BNs. Hence, in future, advanced data cleaning methods could be developed to make a trade-off between information loss and quality improvement for the dataset. Also, high-quality datasets with more detailed features and various scenarios (e.g., the events collected from different countries) could be employed to improve the Copula-BNs.
5 Based on the forward inference, the Copula-BN could be integrated with Advanced Driver-Assistance System (ADAS). Using the Copula-BN, the conditional expectation of the event can be calculated based on the states of the key features. The conditional expectation of the event can be seen as a risk indicator, upon which the ADAS could provide warning notices to the driver if the crash risk exceeds certain threshold.
6 Based on the backward inference, a crash diagnosis system can be developed to investigate the causation of a crash accident, which is significant to road traffic safety. Given a crash (Event = 1) and the states of some key features, the conditional expectation of the other key features can be obtained via the Copula-BN. The features with high conditional expectation are more likely to be the main causes of the crash, which shall be of concern to investigators.

CRediT authorship contribution statement

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

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

This paper presents a part of the first author's PhD research. The authors have been granted approval by the Virginia Tech Transportation Institute (VTTI) to use the SHRP2 NDS database for academic research. The authors would like to sincerely thank the VTTI for providing the SHRP2 NDS database. This study did not receive any specific grant from funding agencies in the commercial, public, or not-for-profit sectors.

Appendix A

Feature descriptions
Table A1

Table A1
Feature descriptions.

Cat.	Features	Abbr.		Descriptions
		LC	CF	
DP	tgate	DP1		How often the participant reported "tailgating" in the risk-taking investigation
	Age Group	DP2		The age group corresponding to the driver's birthdate
	Driving Sleepy		DP1	How often the participant reported driving sleepy in the past 12 months
DE	max Decel	DE1	DE1	Maximum deceleration value recorded during the trip in lateral (x) direction
	cnt Decels	DE2	DE2	Number of the accelerations larger than 0.4 gravitational acceleration sustained for at least 1200 milliseconds in x direction
	Annual Miles	DE3		The participant's estimated average annual mileage over the past five years
	Radar Time Headway_10		DE3	Cumulative time where vehicle is moving greater than 5 kph while there was a lead vehicle present with a headway between 0.5 and 1.0 s
DB*	Driver Behaviors	DB1	DB1	Whether the driver behaviors that (possibly) contribute to the crash or near-crash occur or not
	Maneuver Judgement	DB2		Judgement of the safety and legality of the pre-incident maneuver
	Head_position_x_max	DB3		Maximum head position relative to the face camera in x direction
	Attention		DB3	Whether the driver was distracted or not prior to the event
	Head_position_y_min		DB4	Minimum head position relative to the face camera in y direction
	Head_position_y_std		DB5	Standard deviation of head position relative to the face camera in y direction
SE	Obstruction	SE1	SE1	Whether there were visual obstructions, which were relating to sight distance or blind spots in the roadway infrastructure that (possibly) have impacted the ability of the driver to recognize potential risks or respond effectively to the event, or not
	Traffic Density	SE2	SE2	The level of traffic density at the time of the start of the event
	Dry	SE3		Whether the road surface condition was dry or not
	Adverse Conditions	SE4		Whether there was an adverse weather condition at the time of the start of the event
	Divided (median strip or barrier)		SE4	Whether the opposing traffic of lane was divided by median strip/barrier or not
VP	Vehicle Factors	VP1	VP1	Whether there were factors, which were relating to the mechanical malfunction in vehicle that (might) have contributed to the event or to the ability of the driver to respond effectively to the event, or not
	Maneuver	SV1	SV1	Whether the closest surrounding vehicle conducted risky maneuver that (possibly) contribute to the crash or near-crash occur or not.
	Going Straight.1	SV2		Whether the ongoing actions of the driver of the closest surrounding vehicle prior to the start of the event was going straight or not
SV*	Track1_X_POS_PROCESSED_min	SV3	SV4	Minimum distance between the closest surrounding vehicle and subject vehicle front bumper in x direction
	Reaction	SV4	SV2	Whether the driver of the closest surrounding vehicle had reaction or avoidance maneuver in response to the event or not
	In front of the Subject Vehicle	SV5		Whether the position of the closest surrounding vehicle that is involved in the event was in front of the subject vehicle or not
SV*	Track1_X_VEL_PROCESSED_std		SV3	Standard deviation of the relative velocity between the closest surrounding vehicle and subject vehicle in x direction
	accel_x_min	VK1	VK1	Minimum vehicle acceleration in the x direction versus time
	accel_x_std	VK2	VK2	Standard deviation of vehicle acceleration in the x direction versus time
	accel_x_kurtosis	VK3	VK3	Kurtosis of vehicle acceleration in the x direction versus time
VK*	accel_y_max	VK4		Maximum vehicle acceleration in the y direction versus time
	speed_gps_std	VK5		Standard deviation of vehicle speed from GPS
	speed_gps_min		VK4	Minimum vehicle speed from GPS
	accel_x_max		VK5	Maximum vehicle acceleration in the x direction versus time

* The coordinate systems of vehicle and head position are respectively illustrated in Figs. A1 and. A2

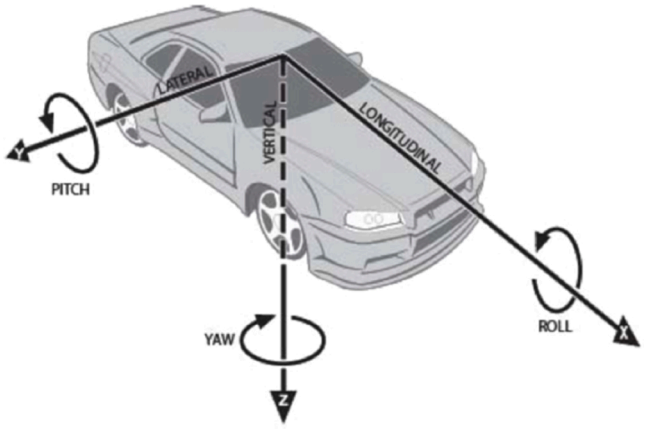


Fig. A1. Coordinate system of vehicle.

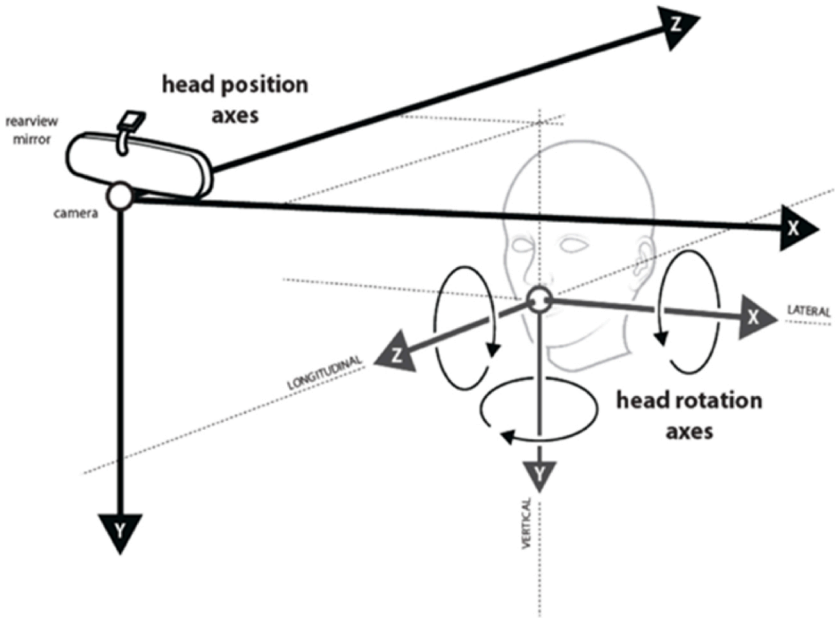


Fig. A2. Machine-vision based coordinate system of head position.

Appendix B

Inference verification using SHAP
Figs. B1 and B2

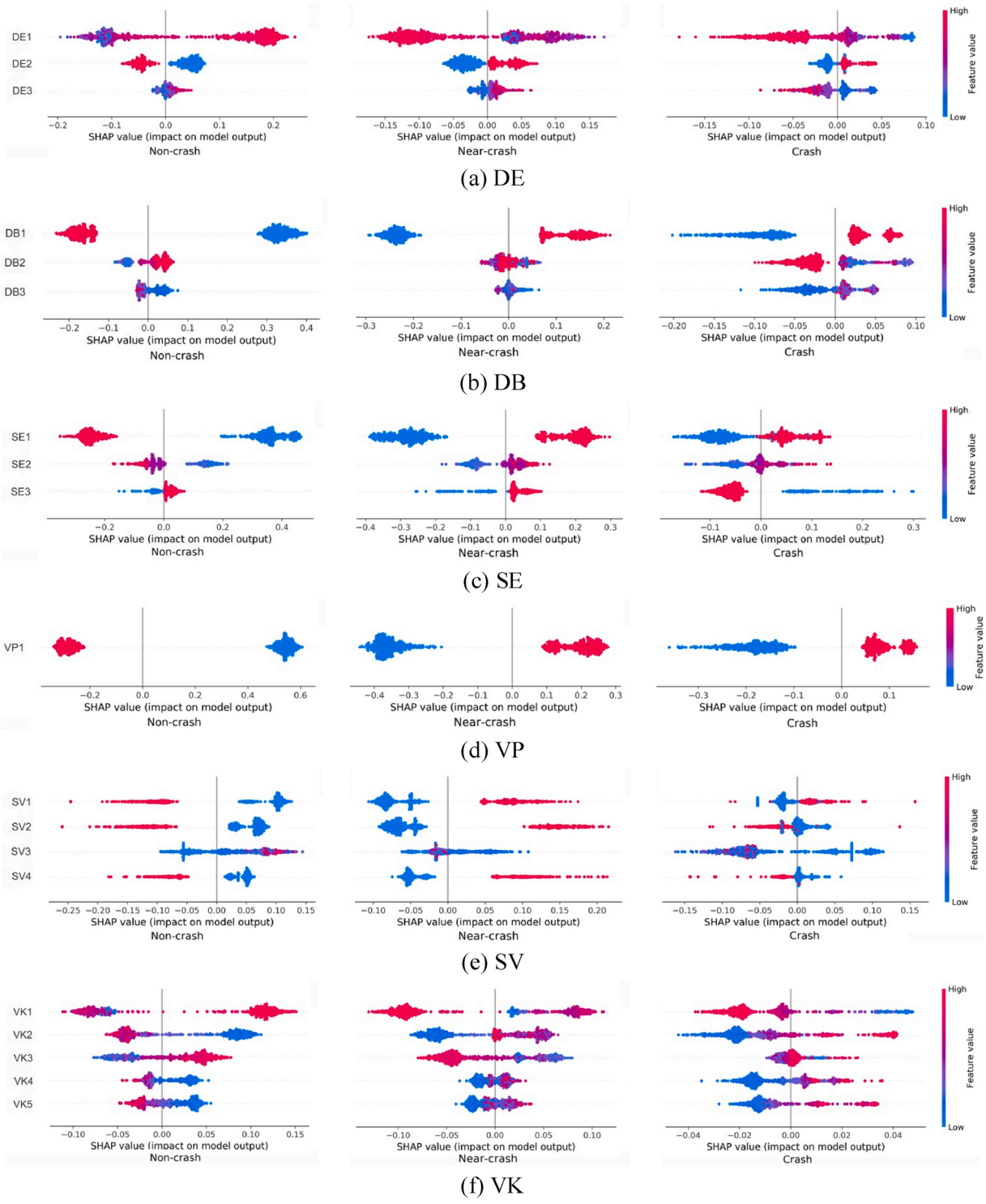


Fig. B1. SHAP values of key features with respect to lane-changing maneuver.

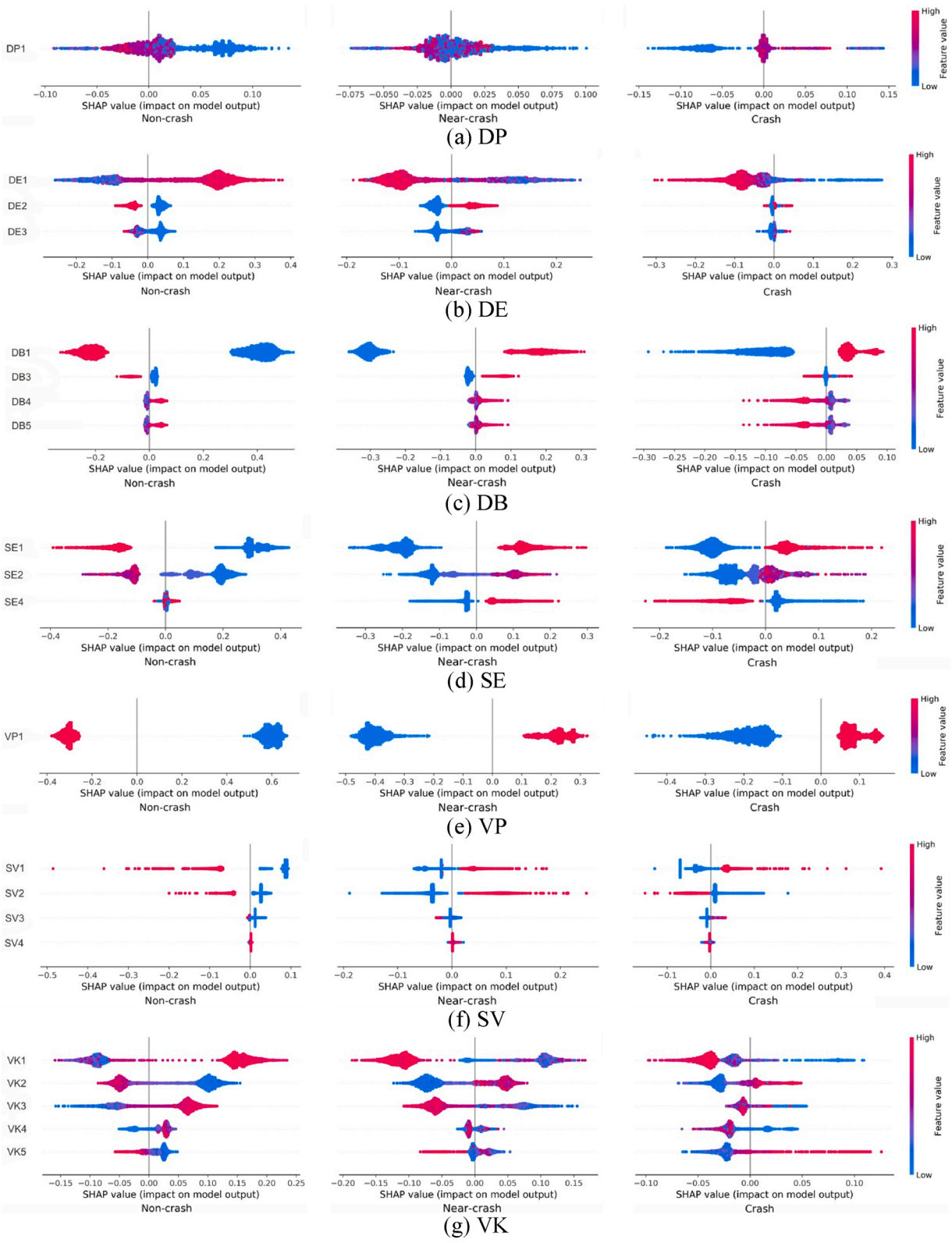


Fig. B2. SHAP values of key features with respect to car-following maneuver.

References

Abdi, H., 2007. The Kendall rank correlation coefficient. Encycl. Measure. Stat. 508–510.

Adanu, E.K., Smith, R., Powell, L., Jones, S., 2017. Multilevel analysis of the role of human factors in regional disparities in crash outcomes. Accid. Anal. Prev. 109, 10–17.

Ali, Y., Sharma, A., Haque, M.M., Zheng, Z., Saifuzzaman, M., 2020. The impact of the connected environment on driving behavior and safety: a driving simulator study. Accid. Anal. Prev. 144, 105643.

Bao, J., Liu, P., Ukkusuri, S.V., 2019. A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accid. Anal. Prev. 122, 239–254.

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

Buntine, W., 1991. Theory Refinement on Bayesian Networks. Uncertainty Proceedings 1991. Elsevier.

Cawley, G.C., Talbot, N.L., 2007. Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. J. Mach. Learn. Res. 8, 841–861.

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

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., Guan, H., 2015. A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. 80, 76–88.

Chen, C., Liu, X., Chen, H.-H., Li, M., Zhao, L., 2018. A rear-end collision risk evaluation and control scheme using a Bayesian network model. Ieee Trans. Intell. Transp. Syst. 20, 264–284.

Chen, T., Shi, X., Wong, Y.D., 2019. Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data. Accid. Anal. Prev. 129, 156–169.

Chen, T., Shi, X., Wong, Y.D., Yu, X., 2020. Predicting lane-changing risk level based on vehicles' space-series features: a pre-emptive learning approach. Transp. Res. Part C Emerg. Technol. 116, 102646.

Chen, T., Shi, X., Wong, Y.D., 2021. A lane-changing risk profile analysis method based on time-series clustering. Phys. A Stat. Mech. Appl. 565, 125567.

Chong, L., Abbas, M.M., Flintsch, A.M., Higgs, B., 2013. A rule-based neural network approach to model driver naturalistic behavior in traffic. Transp. Res. Part C Emerg. Technol. 32, 207–223.

Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. 9, 309–347.

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

Das, A., Ghasemzadeh, A., Ahmed, M.M., 2019. Analyzing the effect of fog weather conditions on driver lane-keeping performance using the SHRP2 naturalistic driving study data. J. Safety Res. 68, 71–80.

Das, A., Khan, M.N., Ahmed, M.M., 2020. Detecting lane change maneuvers using SHRP2 naturalistic driving data: a comparative study machine learning technique. Accid. Anal. Prev. 142, 105578.

Ding, J., 2010. In: Rebai, A. (Ed.), Probabilistic Inferences in Bayesian Networks. Bayesian Network, pp. 39–53.

Ding, N., Jiao, N., Zhu, S., Liu, B., 2019. Structural equations modeling of real-time crash risk variation in car-following incorporating visual perceptual, vehicular, and roadway factors. Accid. Anal. Prev. 133, 105298.

Elidan, G., 2010. Copula bayesian networks. Adv. Neural Inf. Process. Syst. 559–567.

Farah, H., Koutsopoulos, H.N., Saifuzzaman, M., Kölbl, R., Fuchs, S., Bankosegger, D., 2012. Evaluation of the effect of cooperative infrastructure-to-vehicle systems on driver behavior. Transp. Res. Part C Emerg. Technol. 21, 42–56.

Fischer, M., Barkley, R.A., Smallish, L., Fletcher, K., 2007. Hyperactive children as young adults: driving abilities, safe driving behavior, and adverse driving outcomes. Accid. Anal. Prev. 39, 94–105.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., Yuan, J., 2020. Predicting real-time traffic conflicts using deep learning. Accid. Anal. Prev. 136, 105429.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. Ieee Trans. Syst. Man Cybern. Part C 42, 463–484.

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

Hallvig, D., Anund, A., Fors, C., Kecklund, G., Karlsson, J.G., Wahde, M., Åkerstedt, T., 2013. Sleepy driving on the real road and in the simulator—a comparison. Accid. Anal. Prev. 50, 44–50.

Hamdar, S.H., Qin, L., Talebpour, A., 2016. Weather and road geometry impact on longitudinal driving behavior: exploratory analysis using an empirically supported acceleration modeling framework. Transp. Res. Part C Emerg. Technol. 67, 193–213.

Han, C., Huang, H., Lee, J., Wang, J., 2018. Investigating varying effect of road-level factors on crash frequency across regions: a Bayesian hierarchical random parameter modeling approach. Anal. Methods Accid. Res. 20, 81–91.

Hankey, J.M., Perez, M.A., McClafferty, J.A., 2016. Description of the SHRP 2 Naturalistic Database and the Crash, Near-crash, and Baseline Data Sets. Virginia Tech Transportation Institute.

Harris, D., Harris, S., 2010. Digital Design and Computer Architecture. Morgan Kaufmann.

Hwang, S., Boyle, L.N., Banerjee, A.G., 2019. Identifying characteristics that impact motor carrier safety using Bayesian networks. Accid. Anal. Prev. 128, 40–45.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.

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

Krahé, B., Fenske, I., 2002. Predicting aggressive driving behavior: the role of macho personality, age, and power of car. Aggress. Behav. 28, 21–29.

Kusiak, A., 2001. Feature transformation methods in data mining. Ieee Trans. Electron. Packag. Manuf. 24, 214–221.

Lal, T.N., Chapelle, O., Weston, J., Elisseeff, A., 2006. Embedded methods. Feature Extraction. Springer.

Li, L., Gan, J., Ji, X., Qu, X., Ran, B., 2020a. Dynamic driving risk potential field model under the connected and automated vehicles environment and its application in car-following modeling. Ieee Trans. Intell. Transp. Syst.

Li, M., Li, Z., Xu, C., Liu, T., 2020b. Short-term prediction of safety and operation impacts of lane changes in oscillations with empirical vehicle trajectories. Accid. Anal. Prev. 135, 105345.

Lian, Y., Zhang, G., Lee, J., Huang, H., 2020. Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles. Accid. Anal. Prev. 146, 105711.

Liang, Y., Lee, J.D., 2014. A hybrid Bayesian Network approach to detect driver cognitive distraction. Transp. Res. Part C Emerg. Technol. 38, 146–155.

Liu, Y.-C., Wu, T.-J., 2009. Fatigued driver's driving behavior and cognitive task performance: effects of road environments and road environment changes. Saf. Sci. 47, 1083–1089.

Lundberg, S.M., Lee, S.-I.A., 2017. Unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 4765–4774.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Mannering, F., Bhat, C.R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Anal. Methods Accid. Res. 25, 100113.

Massey Jr., F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46, 68–78.

Mbakwe, A.C., Saka, A.A., Choi, K., Lee, Y.-J., 2016. Alternative method of highway traffic safety analysis for developing countries using delphi technique and Bayesian network. Accid. Anal. Prev. 93, 135–146.

Molina, L.C., Belanche, L., Nebot, À., 2002. Feature selection algorithms: a survey and experimental evaluation. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, pp. 306–313.

Nelsen, R.B., 2007. An Introduction to Copulas. Springer Science & Business Media.

NHTSA, 2015. Traffic Safety Facts 2015: a Compilation of Motor Vehicle Crash Data From the Fatality Analysis Reporting System and the General Estimates System. National Highway Traffic Safety Administration, Washington, DC.

Osman, O.A., Hajij, M., Karbalaieali, S., Ishak, S., 2019. A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. Accid. Anal. Prev. 123, 274–281.

Pan, Y., Ou, S., Zhang, L., Zhang, W., Wu, X., Li, H., 2019. Modeling risks in dependent systems: a Copula-Bayesian approach. Reliab. Eng. Syst. Saf. 188, 416–431.

Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., Yannis, G., 2019. Review and ranking of crash risk factors related to the road infrastructure. Accid. Anal. Prev. 125, 85–97.

Pearl, J., 2009. Causality. Cambridge university press.

Peralta, B., Soto, A., 2014. Embedded local feature selection within mixture of experts. Inf. Sci. (Ny) 269, 176–187.

Poó, F.M., Ledesma, R.D., 2013. A study on the relationship between personality and driving styles. Traffic Inj. Prev. 14, 346–352.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. 2, 37–63.

Rahimi, E., Shamshiripour, A., Samimi, A., Mohammadian, A.K., 2020. Investigating the injury severity of single-vehicle truck crashes in a developing country. Accid. Anal. Prev. 137, 105444.

Rahm, E., Do, H.H., 2000. Data cleaning: problems and current approaches. IEEE Data Eng. Bull. 23, 3–13.

Razmjoo, A., Xanthopoulos, P., Zheng, Q.P., 2017. Online feature importance ranking based on sensitivity analysis. Expert Syst. Appl. 85, 397–406.

Reimer, B., Donmez, B., Lavallière, M., Mehler, B., Coughlin, J.F., Teasdale, N., 2013. Impact of age and cognitive demand on lane choice and changing under actual highway conditions. Accid. Anal. Prev. 52, 125–132.

Ren, J., Jenkinson, I., Wang, J., Xu, D., Yang, J., 2008. A methodology to model causal relationships on offshore safety assessment focusing on human and organizational factors. J. Safety Res. 39, 87–100.

Rhodes, N., Pivik, K., 2011. Age and gender differences in risky driving: the roles of positive affect and risk perception. Accid. Anal. Prev. 43, 923–931.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141.

Rokach, L., Maimon, O.Z., 2008. Data Mining With Decision Trees: Theory and Applications. World Scientific.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Russell, S., Norvig, P., 2002. Artificial Intelligence: A Modern Approach. Prentice Hall.

Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Sears, E.S., Perez, M.A., Brown, M., Vickers, D., Van Horn, A., Holladay, K., Barde, J., 2016. SwRI: Investigation of Techniques for Detecting Distracted Drivers Using SHRP2 Vehicle Data, v2 ed. VTTI.

Sears, E., Perez, M.A., Dan, K., Shimamiya, T., Hashimoto, T., Kimura, M., Yamada, S., Seo, T., 2019. A Study on the Factors That Affect the Occurrence of Crashes and Near-Crashes. DRAFT VERSION ed. VTTI.

Shi, X., Wong, Y.D., Li, M.Z.-F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accid. Anal. Prev. 129, 170–179.
Shi, X., Wong, Y.D., Chai, C., Li, M.Z.F., 2020. An automated machine learning (AutoML) method of risk prediction for decision-making of autonomous vehicles. Ieee Trans. Intell. Transp. Syst.
Simons-Morton, B.G., Gershon, P., Gensler, G., Klauer, S., Ehsani, J., Zhu, C., O'Brien, F., Gore-Langton, R., Dingus, T., 2019. Kinematic risky driving behavior among younger and older drivers: differences over time by age group and sex. Traffic Inj. Prev. 20, 708–712.
Tomek, I., 1976. Two modifications of CNN. Transact. Syst. Man Cybernet. 6, 769–772.
Wali, B., Khattak, A.J., Karnowski, T., 2020. The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Anal. Methods Accid. Res. 28, 100136.
Wang, X., Feng, M., 2019. Freeway single and multi-vehicle crash safety analysis: influencing factors and hotspots. Accid. Anal. Prev. 132, 105268.
Wang, Y., Zhang, J., Lu, G., 2018. Influence of driving behaviors on the stability in car following. Ieee Trans. Intell. Transp. Syst. 20, 1081–1098.
Wang, J., Luo, T., Fu, T., 2019. Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach. Accid. Anal. Prev. 133, 105320.
Weng, J., Meng, Q., 2012. Effects of environment, vehicle and driver characteristics on risky driving behavior at work zones. Saf. Sci. 50, 1034–1042.
Xing, F., Huang, H., Zhan, Z., Zhai, X., Ou, C., Sze, N., Hon, K., 2019. Hourly associations between weather factors and traffic crashes: non-linear and lag effects. Anal. Methods Accid. Res. 24, 100109.
Yang, L., Ma, R., Zhang, H.M., Guan, W., Jiang, S., 2018. Driving behavior recognition using EEG data from a simulated car-following experiment. Accid. Anal. Prev. 116, 30–40.
Yang, M., Wang, X., Quddus, M., 2019. Examining lane change gap acceptance, duration and impact using naturalistic driving data. Transp. Res. Part C Emerg. Technol. 104, 317–331.
Zeng, Q., Huang, H., Pei, X., Wong, S., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accid. Res. 10, 12–25.
Zhang, X., Sun, J., Qi, X., Sun, J., 2019. Simultaneous modeling of car-following and lane-changing behaviors using deep learning. Transp. Res. Part C Emerg. Technol. 104, 287–304.
Zhu, X., Yuan, Y., Hu, X., Chiu, Y.-C., Ma, Y.-L., 2017. A Bayesian Network model for contextual versus non-contextual driving behavior assessment. Transp. Res. Part C Emerg. Technol. 81, 172–187.
Zilko, A.A., Kurowicka, D., 2016. Copula in a multivariate mixed discrete–continuous model. Comput. Stat. Data Anal. 103, 28–55.