Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks

Carrodano Tarantino, Cinzia

How to cite

CARRODANO TARANTINO, Cinzia. Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks. In: Scientific reports, 2024, vol. 14, p. 18948. doi: 10.1038/s41598-024-69740-6

This publication URL: https://archive-ouverte.unige.ch/unige:179319
Publication DOI: 10.1038/s41598-024-69740-6

scientific reports

Check for updates

OPEN

Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks

Cinzia Carrodano

This paper aims to demonstrate nonlinear risk factor interactions based on a data-driven approach using a Bayesian network model, providing a road safety use case. Road safety is a critical issue worldwide, with approximately 1.3 million road traffic deaths each year (WHO). Traditional road safety risk assessment methods often analyze individual factors separately; however, these assessments fail to capture the complex dynamics of real-world analysis, in which multiple factors interact through nonlinear relationships. In this study, a novel road safety risk assessment approach that uses a Bayesian network model to explore the nonlinear relationships among road safety risk factors is developed. Through the analysis of extensive crash reports from the state of Maryland, the complex interdependencies among various risk factors and their cumulative impact on road safety are investigated. Our findings show that two combined risk factors have different effects on risk level when considered individually. Two case studies related to human state risk factors and environmental risk factors, such as driving under the influence and snowy roads, as well as fatigue and snowy roads, have an amplified effect on the risk level. The findings highlight the importance of considering nonlinear interactions among risk factors when developing effective and targeted strategies for accident prevention and road safety improvement. This research contributes to the field of road safety by presenting a new methodology for understanding and mitigating road safety hazards.

The rapid expansion of urban development and increase in travel have greatly increased the potential risk of traffic accidents. According to the World Health Organization (WHO), traffic accidents cause 1.3 million preventable deaths annually worldwide and result in injuries to 20 to 50 million people. By 2030, the United Nations General Assembly aims to reduce traffic deaths and injuries by at least 50%[1]. This ambitious goal underscores the importance of enhancing road safety to mitigate the adverse outcomes of traffic accidents.

To understand and mitigate road safety risks, the various factors contributing to these risks must be explored. These complex, multifaceted risk factors include elements such as driver behavior, vehicle conditions, road infrastructure, and weather conditions. While the primary risk factors for road traffic accidents, such as speeding, driving under the influence, lack of security equipment, and distracted driving, have been well documented by the WHO[2], conventional road safety risk assessment approaches typically analyze each of these factors separately.

Although these methodologies can be used to identify key risk factors, they often neglect the complex dynamics present in real-world analysis, in which factors have complex relationships. This limitation represents a significant gap in current road safety research, highlighting the need for a more holistic understanding of how various risk factors collectively influence safety outcomes.

Previous studies have employed Bayesian methods to enhance the understanding of road safety risks. Cheng et al.[3] utilized Bayesian spatiotemporal models with mixture components to analyze crash frequencies, focusing on the space–time interactions of crash occurrences. Deublein et al.[4] applied a Bayesian hierarchical approach to predict road accidents, highlighting the non-linear relationship between exposure measures and crash types through a multivariate Poisson-lognormal regression model. Han et al.[5] investigated the varying effects of road-level factors on crash frequency across regions using Bayesian hierarchical random parameter models, accounting for regional variations and complex interactions between road-level factors. Qin et al.[6] explored the non-linear relationship between traffic volume and crash incidence using hierarchical Bayesian models, emphasizing the importance of considering non-linear exposure measures in safety performance functions.

Geneva School of Economics and Management, University of Geneva, 1205 Geneva, Switzerland. email: cinzia.carrodanotarantino@unige.ch

While these studies significantly advance the field of road safety analysis, they often do not fully capture the combined nonlinear interactions among multiple risk factors.

To address this gap, we aim to investigate the complex and nonlinear interactions among road risk factors through a comprehensive data-driven analysis approach. In this work, nonlinearity refers to an analysis in which the combined impact of multiple risk factors on safety outcomes cannot be accurately predicted by merely summing their individual effects. Such an approach is essential for road safety, as variables such as driver behavior, vehicle conditions, and environmental factors are rarely isolated but instead interact in multifaceted ways, influencing the likelihood and severity of accidents. We aim to explore the nonlinear relationships among road safety risk factors by employing a Bayesian network model. Using this model to analyze a dataset collected in the state of Maryland, which includes extensive crash reports, we investigate the complex interdependencies among various risk factors and their nonlinear impact on road safety. The results show that the model provides a robust framework for examining how combined risk factors contribute to the probability and severity of road traffic accidents.

The Maryland crash report dataset was selected based on its comprehensiveness and the depth of information included on accident circumstances, outcomes, and contributing factors. This rich dataset provides a good foundation for examining a wide array of risk factors within a diverse range of real-world analyses, making it an ideal resource. The findings of this study contribute considerably to the field of road safety analysis. By highlighting the importance of nonlinear interactions among risk factors through data-driven analyses, we offer new perspectives for developing more effective and targeted strategies for accident prevention and road safety improvement.

To the best of our knowledge, this study is the first to demonstrate the nonlinear effects on risk level when factors are combined. Although we have proven this nonlinearity using two case studies related to road safety, the developed *data-driven* approach can be applied to other fields.

This paper begins with a "Related works" section dedicated to driving risk assessment and specifically to the Bayesian network approach. We continue with the "Methods" section, where we present the methodology for building the model, and the "Results" section shows the sensitivity analysis of the model followed by two Bayesian network applications and the demonstration of nonlinear factor interactions. We conclude this study with discussion and conclusion sections.

Related works

Traditional statistical approaches may not adequately capture the complex interrelationships and outcomes of road risk factors. A probabilistic graphical model, such as a Bayesian network, considers the intertwined nature of these risk factors and provides a more nuanced and holistic understanding of road safety. Many studies have been conducted in the field of road safety. They have examined different facets of causes that can induce a bad event. Major risk factors, such as human factors, road factors, environmental factors and vehicle-related factors, have been identified[7]. Different methods have been used to assess road risks[8–11]. These studies are a small part of a wide area of research. They perform risk assessment through specific factors using mostly statistical methods. Although they provide interesting results, our study focuses on the causal relationships between risk factors and provides a more effective method for understanding how combined risk factors interweave.

The emphasis on causal analysis of accidents stems from its ability to explain the relationships between risk factors to propose appropriate actions to mitigate their consequences. Causal analysis also helps to understand the sequence of events that results in accident outcomes.

Several studies have explored the use of the Bayesian network for risk assessment in different fields; for example, Zou et al.[12] apply a Bayesian network for causal analysis of road accidents. They develop the model with an expectation–maximization (EM) algorithm to address missing data. The model shows good results in identifying and explaining road risk factors. They focus on the severity of injuries and the estimated damage. Ma et al.[13] propose a Bayesian network for hazardous materials (Hazmat). More recently, Feng et al.[14] present a Bayesian network-based risk evaluation model for railways, and Zengkai et al.[15] propose a dynamic Bayesian network risk analysis for marine oil spills. According to many other studies, using a Bayesian network for risk assessment provides good results and explains causal relationships.

To provide a model that explains the effect on the risk level of combined risk factors, we choose a causal graph, such as a Bayesian network.

Methods

Road risk is associated with uncertainty, which can lead to negative outcomes, such as accidents that may have different levels of severity. An accident could result in fatalities, injuries, or material damage affecting vehicles, nearby structures or the environment. The causal aspect is deeply intertwined with road risk[16]. Reckless behavior can cause a loss of vehicle control, potentially leading to an accident. A single event can be traced back to multiple risk sources. The weather, road conditions, and behavior of road users may interact in intricate ways, thus influencing the level of risk. The causal dimension of risk leads to a causal risk model, as described by Aven and Thekdi[16]. Based on Bayes' theorem, a Bayesian network is a graphical representation of the cause-and-effect relationships between different variables. It allows a precise assessment of uncertainty and probabilities.

Bayesian network modeling and estimation involve learning through artificial intelligence approaches. Bayesian network learning requires two steps[17].

(a) *Structure learning*: The variables (nodes) and the relationships among them are determined, and a direct acyclic graph (DAG) is constructed.
(b) *Parameter learning*: A conditional probability table (CPT) is learned for each node in the DAG.

Our methodology is shown in the Fig. 1.

The steps are explained below.

Step I We identify the risk factors associated with driving and select our study variables, considering the relevance of each factor. Once the variables have been chosen, we perform the discretization process.

In the Bayesian network construction process, risk modeling is divided into two phases: structure learning and parameter learning.

Step II Structure learning: To obtain the network structure, we first apply the K2 algorithm in MATLAB[18] to determine the DAG structure. Then, based on expert knowledge, we identify the final structure of our model.

Step III Parameter learning: We determine the model parameters using Netica, a Bayesian network-based software developed by Norsys Software Corporation in Canada[19]. Netica is a widely used tool by researchers across diverse fields, including climate change[20], cybersecurity[21], health[22–24], the environment[25] and road-related studies[26,27]. We input the data into Netica and construct the structure determined in the previous step. Certain data might be unknown or missing, and we use the expectation–maximization (EM) algorithm to obtain these missing data and add them to the dataset. However, the unknown gender data were retained as unknown in our study to maintain the integrity of the original dataset. Missing and unknown data for all the other variables were handled using the EM algorithm. We then proceed to the learning stage of the process. The outcome is the Bayesian model of driving risk.

Step IV We analyze the performance of our model through a sensitivity analysis to evaluate the dependence of the nodes on the target node. In addition, we consider several applications, such as prediction, causal inference, and the most likely explanation.

These steps are detailed hereafter:

Step I: Dataset exploration and risk factor identification

Dataset exploration

We utilized a dataset from the state of Maryland[26], which was chosen for its high population density and focus on road safety. The dataset included vehicle crash reports from January to March 2018, with a total of 26,746 individual crash reports. We filtered the dataset to include only reports with a driver and excluded drivers aged less than 15 years. We identified twelve main variables. The selected risk factors, including human, environmental, and vehicle factors and behaviors, are detailed below. The target variables are the severity and collision type. This dataset is highly qualitative and contains very little unknown data. The drivers' ages (Table 1) and genders are described below. The other variables are provided in Appendix A.

These statistics are a description of the age of the drivers, considering only available information on age. This corresponds to 26,746 data points minus 4584 (=4559+25) unknown or not applicable data points, which is equal to 22,162 data points. This corresponds to 83% of the total.

The average age of the drivers is 39.92 years. The standard error indicates that the sample is representative of the population. The median age is 36 years, and the most frequent age is 28 years. The measure of the dispersion of ages is 16.79 years. Most ages are within this range above or below the average age. The difference between the highest and lowest ages of drivers is 82 years. The youngest is 15 years old, and the oldest is 97 years old.

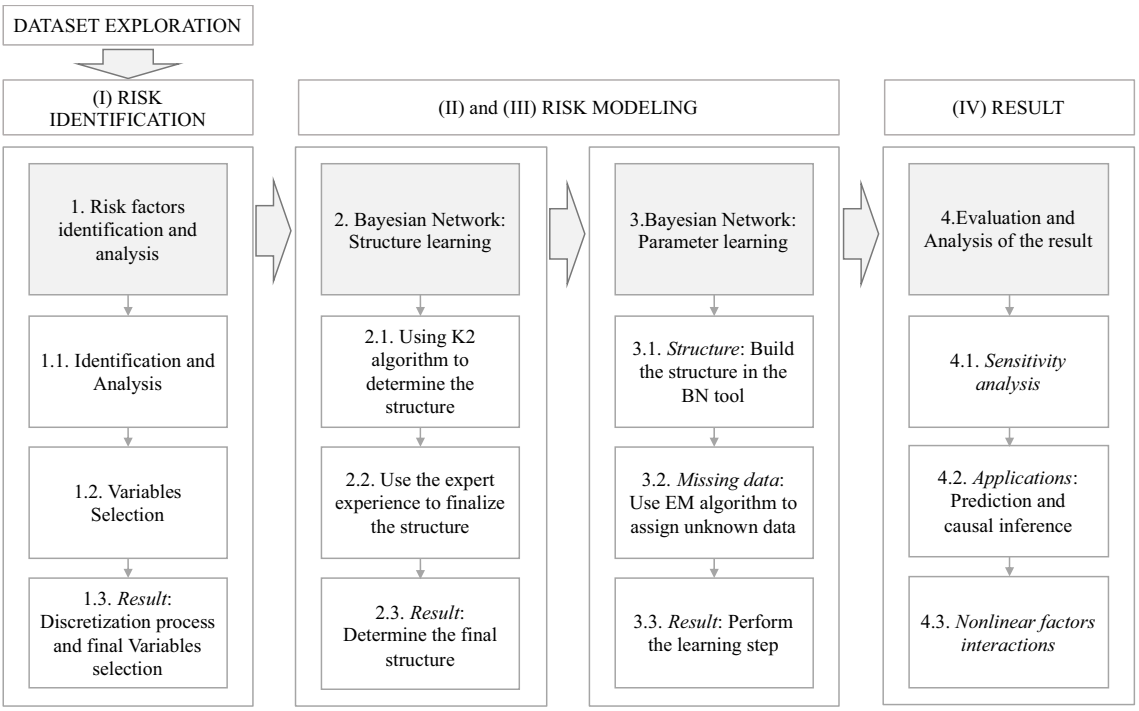

Figure 1. Methodology for building the BN model and the results.

Description	Statistics
Mean	39.92
Standard Error	0.113
Median	36
Mode	28
Standard Deviation	16.79
Sample Variance	281.80
Kurtosis	−0.237
Skewness	0.743
Range	82
Minimum	15
Maximum	97
Sample Size	22,162

Table 1. Descriptive statistics: variable AGE.

The negative kurtosis indicates a slightly flatter distribution than a normal distribution, also called a *platykurtic* distribution. The distribution of drivers' ages has fewer extreme values than expected in a normal distribution. The skewness of 0.743 reveals a slight positive skew where the right tail of the distribution is slightly longer, as also shown in Fig. 2.

The dataset is composed of 33% female, 51% male, and 16% unknown.

Risk factor identification, analysis and variable selection

This section is divided into two parts: the first part is related to risk factor identification and analysis, while the second part refers to variable selection.

We have analyzed official statistics (those of the WHO, NHTSA, European Commission, and Maryland Government) and related driving risk analysis. Three major common risk factors have been identified, in line with the literature[27]: human factors (H), vehicle factors (V) and environmental factors (E). A description of these risk factors is provided below.

Human factors (H). Human factors are the primary contributors to driving risk. The World Health Organization (WHO) identifies major risk factors such as speeding, driving under the influence, and distracted driving[28]. This conclusion aligns with the findings of the National Highway Traffic Safety Administration (NHTSA)[29] and the European Commission, which highlight speed as a critical risk factor[30]. The European Commission maintains that *"Speeding increases the likelihood of an accident. Very strong relationships have been established between speed and accident risk."* The Maryland government identifies human factors, such as distracted driving or speed and aggressive driving, and impaired driving, as the most relevant driving risks[31]. A substantial amount of academic research also refers to human behavior as a significant factor in road traffic accidents[32–35].

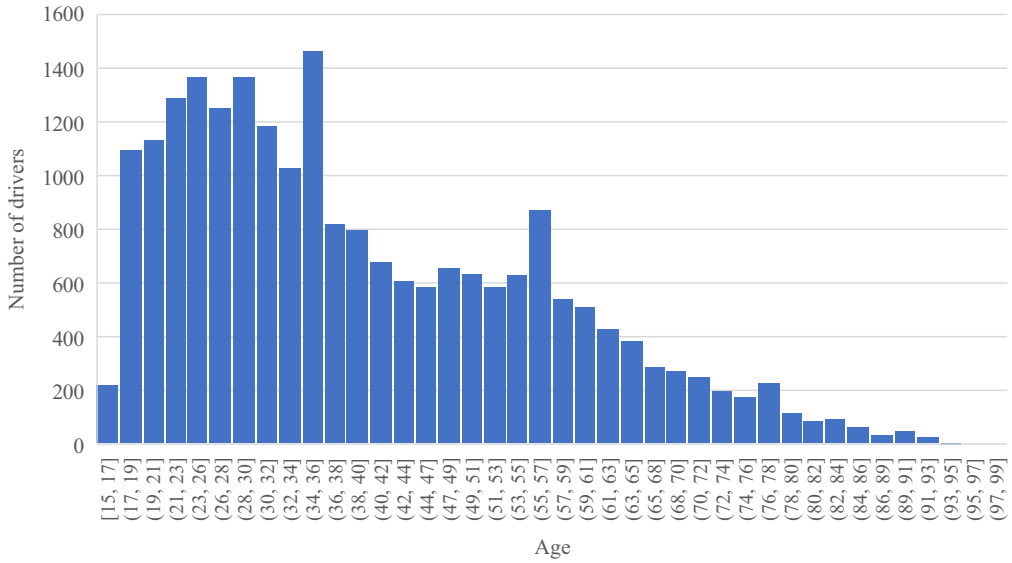

Figure 2. Histogram of the variable AGE.

Environmental factors (E). Although they might not be as prominent in accident data as human factors, elements such as road design[28], lighting conditions (daylight or night light), and inclement weather are significant risk factors[36,37].

We assume that the synergistic effect of these elements might amplify risk. For instance, the probability of an accident can increase when poor road conditions due to wet weather are combined with inadequate lighting. This amplification effect is demonstrated in Sect. 4.3.

Vehicle factors (V). Advancements in vehicle technology have significantly reduced vehicle-related risk factors, unlike the nonuse of safety equipment, such as seat belts or correct helmet use, which continue to be regarded as critical risk factors in official statistics[28].

An accident has consequences of varying severity, among which we distinguish fatalities, injuries and property damage only. Granular information comes from the collision type, which provides information about how the accident occurred. The severity and collision type are the target variables. Table 2 identifies the risk factors that have a direct and measurable impact on road risk.

The dataset comprises 26,746 crash reports and 79 variables. It includes both major variables and numerous detailed variables. For example, a primary variable is the "physical condition: code 102" (Appendix A), while detailed variables include the "alcohol test code: test refused, positive preliminary test, evidence test", as well as "alcohol test type: breath, blood, urine, others." By concentrating on these major variables, as detailed in Appendix A, we ensure a relevant and impactful selection of variables that significantly influence risk factors while aligning with the study's objectives and the literature review analysis (Table 2).

Bayesian networks can handle both continuous and discrete variables[58]. The majority of our dataset variables are discrete. For continuous variables, such as the driver's age, we performed a discretization process. Categorization is applied to several variables: physical condition, surface condition, weather, vehicle type, and safety equipment.

The variables with a consequent influence on risk used in our study are enumerated in Appendix A and distinguished into human factors, called "driver", environmental factors (mentioned as "road" and "environment"), and vehicle factors. "Driver" includes gender, age, and physical condition. "Road" refers to road geometry and surface conditions. The "environment" is related to the weather and light conditions. Vehicle type, safety equipment, and vehicle movement are variables of the "vehicle" factor.

Step II: risk modeling: structure learning with the K2 algorithm

Constructing a Bayesian network (BN) structure using an expert-based approach can be challenging due to the varying assumptions of different experts. Moreover, it may be difficult to obtain the necessary information in the absence of an expert, which could result in an inaccurate structure. A *data-driven* approach may be more efficient, especially in complex domains, as it provides an initial structure that can subsequently be analyzed and fine-tuned based on expert knowledge[58]. This method was used for structure learning, which is divided into three steps, as in[59]: (1) a preliminary structure is built based on expert knowledge, (2) the structure is built using a data-driven analysis approach, and (3) the structure is adjusted based on expert knowledge and data-driven analyses.

Several algorithms for learning structures based on data have been developed. These methods can be categorized into three main categories: constraint-based, score-based, and hybrid approaches[17,58,60]. For our study, we selected the score-based K2 algorithm implemented in MATLAB[14] because score-based algorithms are commonly used to build Bayesian network structures[60]. Additionally, the K2 algorithm is efficient for

Factors	Variables	Reason	References
Human	Age	Young drivers are more likely to engage in risky driving habits and perceive less danger in various driving situations, while older drivers may have more medical conditions associated with higher crash risk	38–41
	Gender	Studies have shown that gender influences risky driving behaviors, with differences in risk perception and sensation seeking between males and females. Male drivers often exhibit more aggressive and risky behaviors	42,43
	Physical condition	The use of alcohol and drugs among drivers is linked to impaired driving performance, increased risk of accidents, and risky behaviors such as speeding and texting while driving. Fatigue also contributes to a higher risk	44–47
Environment	Road geometry	Road geometry design implies vehicle stability, speed selection, and driver perception of risk, contributing to the likelihood of accidents. Specific geometric features, such as horizontal curves and lane widths, influence accident frequency. For instance, accident frequency is diminished by narrow lanes and sharp horizontal curves, as these features encourage drivers to reduce speed and drive more cautiously. Improved road design can therefore play a critical role in enhancing road safety	48
	Surface condition	Surface conditions such as wetness or ice play a critical role in driving risk by affecting vehicle handling, braking distances, and the likelihood of hydroplaning	49,50
	Weather	Weather conditions such as rain, fog, and snow impact driving risk by altering visibility, road friction, and vehicle handling	51
	Light condition	The risk of driving varies significantly under different lighting conditions, such as daylight, dusk, dawn, and nighttime. Reduced visibility during nighttime and the glare from oncoming vehicles can increase the likelihood of accidents	52,53
Vehicle	Vehicle type	Different vehicle types, including cars, trucks, buses, and motorcycles, have varied risk profiles due to differences in size, weight, maneuverability, and visibility	54
	Safety equipment use	Safety equipment use impacts the severity of risk accident. Seat belt use significantly reduces the risk of fatal injury to front-seat passenger car occupants	55
	Vehicle movement	Different vehicle movements, including lane changing, accelerating, and navigating through occluded scenes, significantly impact driving risk by influencing driver's ability to predict trajectories, react to potential hazards, and maneuver safely	56,57

Table 2. Risk identification and analysis of risk factors.

large datasets and involves a heuristic approach in which the node order is assumed to be known, which are characteristics that align with our driving risk dataset. Then, causal relationships are defined. The computational complexity of the algorithm is polynomial based on the number of variables[61]. Compared to other algorithms, the K2 algorithm is less computationally intensive.

The K2 algorithm, developed by Cooper and Herskovits[62], is based on the concept of maximizing a score function to determine the best network structure and uses a heuristic method called greedy search. The algorithm makes a locally optimal choice at each stage according to the following steps. First, a predefined node order is provided by an expert. We assume that a node has no parents and then add the preceding nodes according to their parent nodes. Then, we calculate the score of the network as each parent is added[58]. As a greedy algorithm that computes the local optimal choice, the K2 algorithm might not guarantee a global optimal solution; thus, expert knowledge is necessary to finalize the Bayesian network structure. After the initial structure is learned with the K2 algorithm, the structure is refined according to expert knowledge to ensure that the model accurately reflects the situation.

In this study, experts in the field validated the structure determined based on the K2 algorithm. The twelve selected variables are represented as nodes, while the connections between nodes designate the causal relationships among variables.

Step III: Risk modeling: Parameter learning with the EM algorithm

Once the topology of the Bayesian network has been determined, we perform parameter learning. The data in traffic road accident datasets are often unknown or missing. From a statistical perspective, missing or unknown data may lead to bias[63], and this issue can be addressed using the expectation–maximization (EM) algorithm[64–66]. Given that our dataset contains unknown data, we use the EM algorithm to address this issue. The EM algorithm is a very popular parameter estimation method[65]. This algorithm was introduced by Dempster et al.[67]. The algorithm operates in two steps: the expectation step (E-step) and the maximization step (M-step). The algorithm alternates between the E-step and the M-step in each iteration to find the maximum likelihood solutions when dealing with latent variables. In the E-step, missing or unknown data are estimated according to the observed data and current parameter estimates. In the M-step, these estimates are used to update the parameters by maximizing the likelihood of the observed data. The EM algorithm is explained hereafter based on[27]. The EM algorithm is usually formulated in the context of a statistical model with observed variables, Z, and missing variables, Z^m, depending on the parameters θ. The complete dataset is denoted as $T = (Z, Z^m)$, with the log-likelihood $l_0(\theta; T)$.

(a) First, the parameters are estimated for $\widehat{\theta}^{(0)}$; the set accuracy is ε, and the correction value is $\widehat{\theta}\prime$.

$$\left|\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}^{'}\right| > \boldsymbol{\varepsilon}, \textbf{then}\widehat{\boldsymbol{\theta}} \leftarrow \widehat{\boldsymbol{\theta}}'$$

(b) E-step: The expectation of the log-likelihood is computed using the current parameter estimate. The posterior distribution of the variables Z^m is calculated given Z and the current parameter estimates $\theta^{(t)}$, where t is the current iteration.

$$Q(\theta\prime, \widehat{\theta}^{(t)}) = E[l_0(\theta\prime; T)|Z, \widehat{\theta}^{(t)}] \tag{1}$$

where $\boldsymbol{l}_0(\boldsymbol{\theta}'; \boldsymbol{T})$ is the log-likelihood and $\mathrm{E}[\boldsymbol{l}_0(\boldsymbol{\theta}'; \boldsymbol{T})|\boldsymbol{Z}, \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})}]$ denotes the expectation.

(c) M-step: The expectation computed in the *E-step* is maximized to update the parameter estimate. The parameter $\boldsymbol{\theta}$ maximizes the Q-function $\boldsymbol{Q}(\boldsymbol{\theta}', \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})})$ computed in the E-step.

(d) The *E-step* and *M-step* are iterated until the algorithm converges.

After the model structure was created in Netica software, the "unknown" data were addressed using the EM algorithm. Figure 3 shows the refined version of the model after the EM algorithm was applied, where unknowns were assigned values and the nodes' conditional probability distributions were computed.

Results

Our approach to building the model includes four main steps: (I) identifying the relevant risk factors, (II) modeling the risk factors via structure learning, (III) modeling the risk factors via parameter learning, and (IV) analyzing the results. This section describes Step IV of the methodology and comprises the sensitivity analysis, the applications of the model and a demonstration of the nonlinear effect on the risk. Our model is built based on the real-world dataset of the State of Maryland, which includes 26,746 crash reports from January to March 2018. The winter season was chosen for the snowy period; additionally, this is a large qualitative dataset.

Sensitivity analysis and applications

We evaluated the performance of our model with sensitivity analyses to determine the degree to which the nodes were dependent on the target node. This analysis included examining the performance of our model in applications such as prediction, causal inference, and the most probable explanation, with the results providing valuable insights into the applicability of the proposed model.

For Bayesian networks, sensitivity analyses elucidate how various nodes affect particular target nodes, thereby enabling us to identify important parameters.

We performed sensitivity analyses for the target nodes, namely, severity, safety equipment and collision type, in Netica. Tables 3, 4 and 5 show the results of the sensitivity analysis, including the mutual information and the variance of beliefs, which are explained as follows.

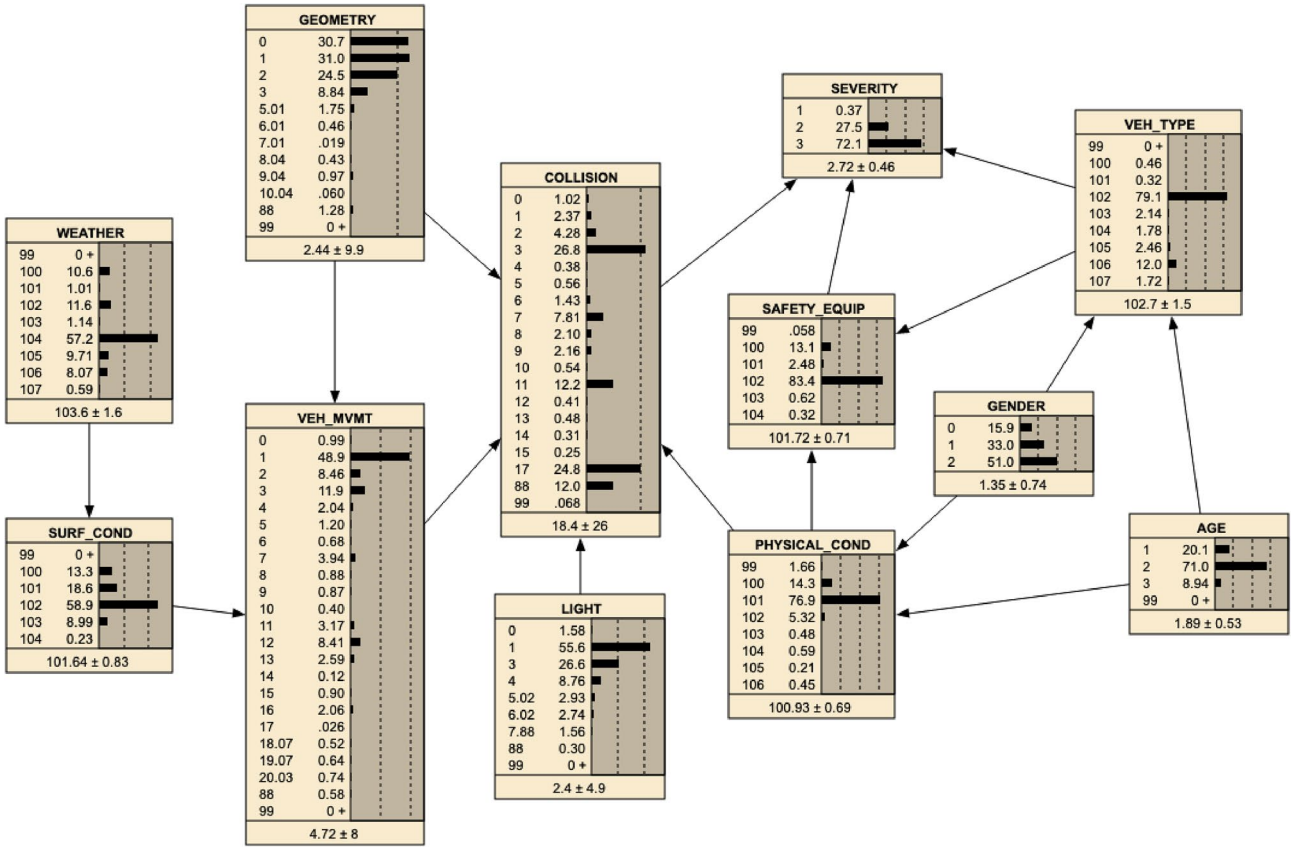

Figure 3. Parameter adjustment according to the EM algorithm.

Node	Mutual information	Percent	Variance of beliefs
SEVERITY	0.88196	100.00000	0.2042617
COLLISION	0.02073	2.35000	0.0054494
SAFETY_EQUIP	0.01995	2.26000	0.0038936
PHYSICAL_COND	0.00799	0.90600	0.0018384
VEH_TYPE	0.00532	0.60300	0.0011410
GENDER	0.00349	0.39500	0.0009062
GEOMETRY	0.00276	0.31300	0.0007697
VEH_MVMT	0.00265	0.30100	0.0007220
LIGHT	0.00008	0.00870	0.0000184
AGE	0.00005	0.00603	0.0000071
SURF_COND	0.00001	0.00085	0.0000017
WEATHER	0.00000	0.00024	0.0000004

Table 3. Sensitivity analysis of the severity node.

Node	Mutual information	Percent	Variance of beliefs
SAFETY_EQUIP	0.81343	100.00000	0.1555180
PHYSICAL_COND	0.25779	31.70000	0.0433205
GENDER	0.12149	14.90000	0.0231785
VEH_TYPE	0.03332	4.10000	0.0021846
SEVERITY	0.01995	2.45000	0.0016127
COLLISION	0.01357	1.67000	0.0019271

Table 4. Sensitivity analysis of the safety equipment node.

Node	Mutual information	Percent	Variance of beliefs
COLLISION	2.98481	100.00000	0.6936152
VEH_MVMT	0.35030	11.70000	0.0270559
GEOMETRY	0.18700	6.27000	0.0073266
PHYSICAL_COND	0.04496	1.51000	0.0013272
SURF_COND	0.00643	0.21500	0.0002478
SEVERITY	0.02073	0.69500	0.0003949
LIGHT	0.02342	0.78500	0.0020717
GENDER	0.01360	0.45600	0.0003603
SAFETY_EQUIP	0.01357	0.45500	0.0003427
WEATHER	0.00138	0.04630	0.0000537
VEH_TYPE	0.00013	0.00423	0.0000040
AGE	0.00009	0.00315	0.0000013

Table 5. Sensitivity analysis of the collision type.

Mutual information is a concept from information theory that provides a measure of the dependence between variables to assess the strength of their relationship. When two nodes are dependent, understanding the value of one node can provide information about the value of the other node. Mutual information measures the information gain among nodes[68]. A high mutual information value implies a high degree of dependency between variables[69], and the mutual information, namely, the entropy reduction, is calculated using Eq. 2. The mutual information between Q, the query variable, and F, the varying variable, is measured in bits.

$$I = H(Q) - H(Q|F) = \sum_{q}\sum_{f} \frac{P(q,f)log_2\left[P(q,f)\right]}{P(q)P(f)} \tag{2}$$

where H(Q) is the initial entropy of Q and H(Q|F) is the entropy of Q after new findings are obtained according to variable F[70,71].

In a Bayesian network, the **variance of beliefs**, **s2**, is the expected squared change in the posterior probability of a node, taken over all its states, due to the information of another node. This metric can be used to evaluate the possible impact or sensitivity of each element in a network[72]. The variance of beliefs, s2, is expressed in Eq. 3 [73].

$$s2 = \sum_{f}\sum_{q} \frac{P(q,f)}{\left[P(q|f) - P(q)\right]^2} \tag{3}$$

Higher variance values indicate that our understanding of this factor can change significantly based on new information. Therefore, a higher variance value indicates that a factor is highly sensitive. Lower variance values indicate that the factor is less sensitive to new information.

Model validation was performed using sensitivity analyses for the severity, safety equipment, and collision nodes, as shown in Tables 3, 4 and 5, respectively. Both of them are the target variables, severity and collision types, and we have added the safety equipment, as it has shown a strong dependency on the severity.

The sensitivity analysis of the severity node showed mutual information values of 0.02073 for the collision node and 0.01995 for the safety equipment node (Table 3). The collision node had a higher mutual information value, indicating that this node had a stronger influence on severity. The variable with the second highest mutual information value is safety equipment. Both factors directly influence accident severity. These variables are followed by physical condition (0.00799), which has an indirect influence on severity and a direct influence on safety equipment and collisions. This analysis revealed the critical factors that contribute to the severity of accidents, emphasizing the importance of focusing on collision prevention and the use of safety equipment as primary interventions. Additionally, the role of physical condition suggests that strategies aimed at improving the DUI or fatigue of individuals could further enhance safety measures and reduce the severity of accidents.

Next, the mutual information of the safety equipment node was assessed, revealing only the first six variable dependencies (Table 4). The results showed a strong dependency between safety equipment use and physical condition (0.25779). This implies that if a driver is fatigued or under the influence of alcohol or drugs, he or she may be less likely to use safety equipment, such as a seat belt. This relationship underscores the need for educational and enforcement strategies that address not only the importance of using safety equipment but also the impact of physical condition on safety practices. By highlighting the influence of physical condition, gender, and vehicle type on the use of safety equipment, this analysis provides valuable insights for developing targeted interventions aimed at improving road safety and reducing the risk of severe accidents.

Table 5 shows the sensitivity analysis results for the collision type. According to the results, the collision type strongly depends on the vehicle movement (0.35030) and road geometry (0.18700), followed by the physical condition (0.04496).

This analysis highlights the critical influence of vehicle movement and road geometry on the occurrence and type of collisions, underscoring the importance of these factors in road safety and accident prevention strategies. The significant dependency on vehicle movement suggests that the way vehicles are driven, including speed and

directional changes, plays a key role in the risk and nature of collisions. Similarly, road geometry, which includes the design of roads, is shown to be a crucial determinant of the likelihood and severity of accidents.

Applications

In this section, we introduce the concept of posterior probabilistic reasoning and present two applications of the Bayesian network model: prediction and causal inference.

Posterior probabilistic reasoning: Prediction and causal inference

According to[17], probabilistic reasoning, also called belief updating, was introduced by Pearl[69]. The use of posterior probabilities enables the incorporation of new information and the updating of beliefs, which is a core principle of Bayesian statistics[17]. Bayesian networks can estimate outcomes based on causes; this process is called prediction. Moreover, Bayesian networks can perform causal inference, which involves determining a cause according to an observed outcome. The BN models of the following examples are shown in Appendix B.

Prediction

A "prediction" can be performed with the Bayesian network model by setting the predictive variables, namely, evidence variables, at 100%. As previously mentioned, driving under the influence (DUI) is a primary crash risk factor. DUI was identified as "influenced by medications, drugs, alcohol" in the "physical condition" node in our Bayesian network model. The variables were set to a maximum value of 100% in Netica. The DUI analysis results show a difference for the single vehicle factor in the collision type: the probability of a collision with a single vehicle notably increased from 24.8% to 38.7%. Moreover, the percentage of severe and, more specifically, fatal crashes increased from 0.37% to 1.23%. These results suggest that when people are driving under the influence, most crashes occur with a single vehicle and tend to be more severe.

For example, considering the influence of alcohol, in 2021, in the U.S., drunk driving accidents were responsible for the deaths of approximately 37 people each day, with a total of 13,384 deaths in one year[74]. As explained by the National Highway Traffic Safety Administration (NHTSA)[74], alcohol affects driving ability and how people operate vehicles. The results of the model show that drunk driving increases the probability of single-vehicle crashes and the severity of such accidents. This result is consistent with those of previous studies, such as those by Morland[75], Behnood and Mannering[76], and Maistros et al.[77]. In addition, in this study, we evaluated the use of safety equipment, such as safety belts, with a lack of safety equipment increasing the severity of most accidents. When driving under the influence, the lack of safety equipment usage increased notably, from 2.48 to 10.3%. The NHTSA[74] mentions that driving under the influence causes physiological impairments, such as reduced visual function, reduced ability to perform tasks, reduced coordination, and decreased ability to brake appropriately, with these impairments increasing reaction times and affecting vehicle control. By slowing brain function, alcohol alters judgment, reasoning and muscle coordination[78]. Individuals under the influence of alcohol might neglect safety protocols, e.g., safety belts. Our model provides results consistent with official statistics and literature in the field.

Causal inference

Bayesian networks can be used to perform simulations of causal inference by intervening on variables. We propose hereafter an example of a prediction statement compared to official analysis. This will provide another way to validate the model. The target variable in this analysis is the collision node, which represents "the angle meets the left turn head on". The results revealed two obvious changes during collisions, namely, an increase in the intersection factor (from 24.5 to 55.6%) in the geometry category and an increase in the "making a left turn" factor (from 8.41 to 41.7%) in the vehicle movement category. In addition, the dark/no lights factor in the light category increased from 8.76 to 15%. The results suggest that these kinds of crashes tend to occur at night, and the physical condition of the driver appears to be normal (changing from 76.9 to 70.7%), which implies that light conditions are the principal risk factor for this kind of road geometry.

Intersections and left turns are critical points in traffic flow. Moreover, such driving maneuvers require high vigilance. During the night, visibility is reduced, making it more difficult for drivers to see other vehicles, other drivers, pedestrians and road signs. Moreover, making left turns requires drivers to assess oncoming traffic and judge speed and distance, which is more challenging under low light conditions. The glare caused by headlights may also impact the driver's vision, making it more difficult to determine the position and speed of other vehicles. Additionally, driver fatigue typically increases during nighttime driving, which may increase the probability of accidents.

Demonstration of nonlinear factor interactions

We investigated the application of Bayesian networks (BNs) for road safety analysis, focusing on analyses derived from the model. These analyses provide insights into the factors affecting fatal accident severity and the interplay among various risk factors. Then, we consider the nonlinear effect of these combined risk factors. In our first analysis, we aim to broadly understand the influence of different factors on accident severity. The values of the analysis presented hereafter are obtained from the BN model (Fig. 3) by setting evidence variables indicated on a scale from 0 to 100%. The "Base" column refers to the original BN model in Fig. 3. The obtained results are shown in Appendices C and D. In the table, we include only variables that have a notable impact. We focus primarily on fatal accidents to determine the most important risk factors (Table 6). Furthermore, we consider the obtained specific factors in combination and evaluate the probability of a fatal accident occurring in each case (Table 7). This analysis deepens the understanding of the previous risk factors. Finally, we determine whether two combined risk factors have a greater impact on the overall risk than the sum of the same factors considered

Analysis #1					
Node	**Code**	**Description**	**Base**	**Values**	**Relative change**
Severity	1	Fatal	0.366	**100.000**	
	2	Injury	27.515	0.000	
	3	Damage only	72.119	0.000	
Collision	3	Same direction rear end	26.842	26.819	0%
	11	Same movement angle	12.192	13.227	8%
	17	Single vehicle	24.751	41.484	68%
Physical cond	101	Normal	76.942	70.363	−9%
	102	DUI	5.325	17.953	237%
Safety Equip	101	None	2.475	29.521	1093%
	102	Belt	83.394	59.006	−29%

Table 6. Analysis #1: The evidence variable is set for severity-fatal of the BN model (base = Fig. 3).

Analysis #2											
Node	**Code**	**Description**	**Base**	**Values**	**Relative change**	**Values**	**Relative change**	**Values**	**Relative change**	**Values**	**Relative change**
Severity	1	Fatal	0.366	**1.235**	237%	**1.179**	222%	**0.438**	20%	**7.940**	2068%
	2	Injury	27.515	28.009	2%	26.281	−4%	22.980	−16%	58.610	113%
	3	Damage only	72.119	70.756	−2%	72.540	1%	76.582	6%	33.450	−54%
Collision	3	Same direction rear end	26.842	24.556	−9%	0.000		0.000		0.000	
	11	Same movement angle	12.192	3.971	−67%	0.000		0.000		0.000	
	17	**Single vehicle**	24.751	38.735	56%	**100.000**		100.000		100.000	
Physical cond	101	Normal	76.942	0.000	−100%	0.000		0.000		0.000	
	102	**DUI**	5.325	**100.000**		100.000		100.000		100.000	
Safety Equip	101	**None**	2.475	10.260	315%	10.260	315%	0.000		**100.000**	
	102	**Belt**	83.394	83.033	0%	83.033	0%	**100.000**			

Table 7. Analysis #2: Several evidence variables are included in the model (Fig. 3). Primary influences with different variables: DUI, single vehicle, and safety equipment use.

separately (Tables 8 and 9). The goal of the first two analyses is to define interesting risk factors for use in the demonstration.

Analysis #1: fatal accident analysis

This first analysis is designed to investigate cases with fatal outcomes, revealing marked changes in several variables (Table 6). The analysis is based on the BN model (Fig. 3) mentioned in the "Base" column in Table 6. The next column is the result with the evidence variable "severity, fatal" set at 100%. The output might be interpreted

Analysis #3 : DUI / Snow			A	B	C	D	E *(computed)*
Node	**Code**	**Description**	**Normal situation**	**Physical condition**	**Surface condition**	**Combined factors**	**Added factors E = (C-A) + B)**
Severity	1	Fatal	**0.332**	**1.228**	**0.350**	**1.329**	
	2	Injury	29.517	27.924	28.951	28.655	
	3	Damage only	70.151	70.848	70.699	70.015	
Physical cond	101	Normal	100		100		
	102	DUI		100		100	
	104	Fatigue					
Surface cond	102	Dry	100	100			
	103	Snow	100	100			
Physical condition Surface condition			Normal Dry	DUI Dry	Normal Snow	DUI Snow	DUI Snow
Severity level: fatal			**0.332**	**1.228**	**0.350**	**1.329**	**1.245**
delta vs. normal sit.(A)				*0.895*	*0.018*	***0.997***	***0.913***

Table 8. Results for Analysis #3 (DUI/Snow), nonlinear risk factor demonstration.

Analysis #4: Fatigue / Snow			A	B	C	D	E *(computed)*
Node	**Code**	**Description**	**Normal situation**	**Physical condition**	**Surface condition**	**Combined factors**	**Added factors E = (C-A) + B)**
Severity	1	Fatal	**0.332**	**0.497**	**0.350**	**0.534**	
	2	Injury	29.517	27.329	28.951	27.775	
	3	Damage only	70.151	72.174	70.699	71.691	
Physical cond	101	Normal	100		100		
	102	DUI					
	104	Fatigue		100		100	
Surface Cond	102	Dry	100	100			
	103	Snow			100	100	
Physical condition Surface condition			Normal Dry	Fatigue Dry	Normal Snow	Fatigue Snow	Fatigue Snow
Severity level: fatal			**0.332**	**0.497**	**0.350**	**0.534**	**0.515**
delta vs. normal sit.(A)				*0.165*	*0.018*	***0.201***	***0.183***

Table 9. Results for Analysis #4 (Fatigue/Snow), nonlinear risk factors demonstration.

as changes in the probabilities (relative change column) of accidents being fatal compared to the base rate (base column). Table 6 shows only the most important impact factors. Details are shown in Appendix C. The results show that single vehicle collisions were associated with a marked increase (+68%) in fatalities. Moreover, driving under the influence (DUI) was strongly associated with fatalities (+237%), while fatalities notably decreased when safety equipment, particularly seat belts, was used (-29%). These results demonstrate the critical impact of driving under the influence and the lack of safety measures on fatal accidents. We investigate these findings further in the next analysis.

Analysis #2: interactions among DUI, safety equipment use, and single vehicle collisions

This section aims to deepen the previous analysis. In this analysis, we build upon the findings of the previous analysis by adding the variables single vehicle, DUI, and safety equipment usage. These predictive variables, set at 100%, are determined as evidence variables in our BN model to investigate their effect on accident severity. Similarly, as in the previous analysis, we only mention the key factors in Table 7. The details are shown in Appendix D.

The analysis results indicate that when driving under the influence (DUI) in a single-vehicle collision, the risks of fatal outcomes and injury are lower when using a seat belt (fatal outcomes, 0.44% with a seat belt vs. 7.94% without seat belt; injury, 22.98% with a seat belt vs. 58.61% without seat belt). Under the same conditions, if 100% of drivers used seat belts instead of 83%, the number of fatalities would decrease from 1.18% to 0.44%, and the number of injuries would decrease from 26.28% to 22.98%. This analysis highlights the increased risk associated with neglecting seat belts and the exacerbating effect of DUI on accident severity, which were also identified by the NHTSA and the WHO as major driving risk factors.

Analyses #3 and #4: effect of risk factors associated with physical conditions and surface conditions

In this section, we demonstrate the nonlinear effect of risk factors. We have chosen the previous key factor, the DUI for physical conditions, for the first demonstration. We have focused this study on two risk factors, human and environment. Therefore, we have provided a normal situation considering a low risk level (human normal state for drivers combined with dry road surface conditions). We have analyzed it with a higher level of risk (DUI human state combined with snowy road surface conditions). The DUI was chosen because we previously identified it as a key risk factor. As we wanted to provide a second demonstration, we chose another variable for the human state, such as fatigue, instead of DUI. In the following analyses, we assess the effects of combining two risk factors: DUI and snowy conditions (Analysis #3) and fatigue and snowy conditions (Analysis #4). We provide two different situations to show that nonlinearity is present in different cases. Analysis #3 revealed that the combined impact of driving under the influence and snowy driving conditions on fatalities (1.329–0.332 = 0.997) was greater than the sum of their individual effects ((DUI 1.228–0.332) + (SNOW 0.35–0.332) = 0.913). This effect is much lower for Analysis #4, in which the combined impact of fatigue and snowy conditions was more similar to the sum of both factors. Next, we provide the details of the analysis (Tables 8 and 9) followed by a more concise explanation of the nonlinearity demonstration. The details of Tables 8 and 9 are shown in Appendix E.

This finding reveals the nonlinear relationships among risk factors, which differ from the effects of individual risk factors (Tables 8, 9). We describe the nonlinearity effect hereafter for Analyses #3 and #4. We have demonstrated this effect for two cases, considering that other situations should be studied in further analysis.

We start with fatalities in a normal situation (normal human state and dry road conditions), which represent a lower risk situation. We then consider fatalities for two other higher risk situations, one for the human state (human state = DUI) and the other for the road condition (road = snow). We compute the variation, representing the effect of these factors. We sum both risk variations for both situations *(a)*. We consider the risk for fatalities with two risk factors (human state = DUI and road = snow) *(b)*. The subtraction of both situations *(a)* – *(b)* shows that the combined factors have a greater effect than the added factors.

We performed a similar calculation for another human state, such as fatigue, instead of DUI (Analysis #4). This case shows a weaker effect than the previous one.

Analysis #3	*Fatal*	*Variation*	*Description*
Fatalities in normal situation	0.33	–	Human state = normal; road = dry
Fatalities with DUI	1.23	0.895	Human state = DUI; road = dry
Fatalities with Snow	0.35	0.018	Human state = normal; road = snow
Subtotal		0.913 *(a)*	
Fatalities with DUI and Snow	1.33	0.997 *(b)*	Human state = DUI; road = snow
Difference (a)—(b)		−0.084	

Analysis #4	*Fatal*	*Variation*	*Description*
Fatalities in normal situation	0.33	–	Human state = normal; road = dry
Fatalities with Fatigue	0.50	0.165	Human state = fatigue; road = dry
Fatalities with Snow	0.35	0.018	Human state = normal; road = snow
Subtotal		0.183*(a)*	
Fatalities with Fatigue and Snow	0.534	0.201 *(b)*	Human state = fatigue; road = snow
Difference (a)—(b)		−0.018	

These analyses provide a comprehensive view of how different risk factors, both individually and in combination, influence the likelihood and severity of road accidents. The analysis highlights the complex interactions among these factors and their nonlinear impact on road safety.

Discussion and conclusion

Our study provides a comprehensive analysis of the nonlinear interactions among various risk factors contributing to road safety to elucidate the complex dependencies that traditional methods may overlook. We propose a novel holistic approach to risk analysis considering the multifaceted nature of risk for a specific study case, such as road safety. In this study, the concept of nonlinearity means that the combined influence of multiple risk factors on safety outcomes is not simply the summation of their separate effects. Nonlinear risk factor interactions are demonstrated with a *data-driven* approach, such as a Bayesian network.

In this study, various analyses using a data-driven model demonstrated the nonlinear interactions among road safety risk factors. These nonlinear effects are particularly evident in the previous analysis involving combined risk factors, such as DUI and adverse weather conditions. The other analysis, combining fatigue and adverse weather conditions results in a less pronounced nonlinear effect. While fatigue does impair a driver's alertness and reaction time, its impact is generally less severe than that of DUI. Fatigued drivers may experience slower reflexes and reduced attention, but they retain more control over their cognitive functions compared to drivers under the influence of alcohol or drugs. Therefore, the combined effect of fatigue and snowy conditions, is not as much amplified as the effect of DUI and adverse weather conditions.

The increased risk when these factors co-occur suggests a complex, multiplicative effect rather than a simple additive effect. This underscores the importance of considering interactions among risk factors to understand and mitigate road safety hazards. By emphasizing these interactions, the study provides a more comprehensive framework for policymakers to design targeted strategies.

Our results are consistent with previous findings in the literature, such as those of the WHO, which identified individual factors such as DUI and poor weather conditions as significant contributors to road accidents. However, by quantifying the compounded impact of these factors, our research provides a novel perspective, illustrating how their interactions increase the probability of severe outcomes. Unlike traditional statistical methods, a BN captures the complex interrelationships of risk factors and is more suitable for this kind of analysis. Although many studies propose risk assessment using Bayesian networks[12–15], none of them has analyzed the nonlinear effects of combined risk factors compared to individual effects.

Our findings have important implications for road safety policies and intervention strategies. Identifying the increased risk due to the combination of factors such as DUI and inclement weather can inform the development of targeted enforcement strategies. For instance, law enforcement agencies might prioritize DUI checks during bad weather conditions, or public safety campaigns could specifically address the heightened risks of driving under the influence during adverse weather conditions. Additionally, our study highlights the critical role of safety equipment, suggesting that interventions promoting its use could substantially mitigate risks, especially in high-risk analysis.

While our study offers important insights, some limitations should be considered. Because we focused on road data collected in Maryland, the results should be considered in this specific context, and these findings might vary in different geographic locations or cultural environments.

Future research should aim to replicate this study for other road risk locations. The method can be generalized to other fields, such as health or leisure activities, like cycling and skiing, to evaluate the universality of these nonlinear interactions. Although we performed the demonstration using two risk factors, the analysis may also include more risk factor interactions.

Our study utilized only three months of data, which may exclude natural seasonal variability in crash data and its potential influence on risk analysis. Seasonal variations can significantly impact factors such as weather

conditions, traffic volume, and driving behavior. Future research should aim to include data spanning all seasons of the year to capture these variations and provide a more comprehensive analysis of risk factors. This approach will ensure that the findings are applicable across different times of the year.

In conclusion, our study contributes to road safety research by elucidating the nonlinear interactions among risk factors through a data-driven approach. The findings suggest that new paradigms for assessing and managing road safety risks should be developed. Approaches that consider the complex relationships among multiple risk factors can lead to more effective and targeted road safety policies and interventions, ultimately contributing to a reduction in road traffic accidents and fatalities. Considering the generalizability of this risk analysis method to other fields, this may also contribute to enhancing preventive measures across various domains.

Data availability

The dataset is open source and is published by the Maryland Government: (https://catalog.data.gov/dataset/maryland-statewide-vehicle-crashes-cy2018-quarter-1).

Received: 21 March 2024; Accepted: 8 August 2024
Published online: 15 August 2024

References

1. "WHO kicks off a decade of action for road safety", World Health Organization, Oct. 28, 2021. [Online] Available: https://www.who.int/news/item/28-10-2021-who-kicks-off-a-decade-of-action-for-road-safety. [Accessed: 18.12.2023].
2. "World Health Organization, Road Safety", World Health Organization, [Online]. Available: https://www.who.int/health-topics/road-safety#tab=tab_2. [Accessed: 18.12.2023].
3. Cheng, W., Gill, G. S., Zhang, Y. & Cao, Z. Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions. *Accid. Anal. Prev.* **112**, 84–93. https://doi.org/10.1016/j.aap.2017.12.020 (2018).
4. Deublein, M., Schubert, M., Adey, B. T., Köhler, J. & Faber, M. H. Prediction of road accidents: A Bayesian hierarchical approach. *Accid. Anal. Prev.* **51**, 274–291. https://doi.org/10.1016/j.aap.2012.11.019 (2013).
5. Han, C., Huang, H., Lee, J. & Wang, J. 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. https://doi.org/10.1016/j.amar.2018.10.002 (2018).
6. Qin, X., Ivan, J. N., Ravishanker, N. & Liu, J. Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov chain Monte Carlo modeling. *J. Transp. Eng.* **131**(5), 345–351. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:5(345) (2005).
7. Razzaghi, A. *et al.* Risk factors of deaths related to road traffic crashes in World Health Organization regions: A systematic review. *Arch. Trauma Res.* **8**(2), 57–86 (2019).
8. Chen, F., Wang, J. & Deng, Y. Road safety risk evaluation by means of improved entropy TOPSIS–RSR. *Saf. Sci.* **79**, 39–54 (2015).
9. Shah, S. A. R. *et al.* Road safety risk assessment: An analysis of transport policy and management for low-, middle-, and high-income Asian countries. *Sustainability* **10**(2), 389. https://doi.org/10.3390/su10020389 (2018).
10. Almoshaogeh, M. *et al.* Traffic accident risk assessment framework for Qassim, Saudi Arabia: Evaluating the impact of speed cameras. *Appl. Sci.* **11**(15), 6682. https://doi.org/10.3390/app11156682 (2021).
11. Chen, Y., Wang, K., Zhang, Y. & Shi, Q. Identification of black spots on highways using fault tree analysis and vehicle safety boundaries. *J. Transp. Saf. Secur.* **13**(1), 46–68 (2021).
12. Zou, X. & Yue, W. L. A Bayesian network approach to causation analysis of road accidents using Netica. *J. Adv. Transp.* https://doi.org/10.1007/s10994-006-6889-7 (2017).
13. Ma, X., Xing, Y. & Lu, J. Causation analysis of hazardous material road transportation accidents by Bayesian network using genie. *J. Adv. Transp.* **2018**, 1–12 (2018).
14. Feng, F., Jia, J., Liang, A. & Liu, C. Bayesian network-based risk evaluation model for the operational requirements of the China Railway Express under the Belt and Road initiative. *Transp. Saf. Environ.* **4**(3), tdac019. https://doi.org/10.1093/tse/tdac019 (2022).
15. Liu, Z. *et al.* Risk assessment of marine oil spills using dynamic Bayesian network analyses. *Environ. Pollut.* **317**, 120716. https://doi.org/10.1016/j.envpol.2022.120716 (2023).
16. Aven, T. & Thekdi, S. *Risk science: An introduction* (Routledge, 2021).
17. Scutari, M. & Denis, J. B. *Bayesian networks: with examples in R* (CRC Press, 2021).
18. Guangdi Li (2023). K2 algorithm for learning DAG structure in Bayesian network (https://www.mathworks.com/matlabcentral/fileexchange/23273-k2-algorithm-for-learning-dag-structure-in-bayesian-network), MATLAB Central File Exchange.
19. "Norsys Software Corp., Netica Application", Norsys Software Corp., [Online]. Available: https://www.norsys.com/netica.html. [Accessed: 18.12.2023].
20. Lee, S. H., Kang, J. E., Park, C. S., Yoon, D. K. & Yoon, S. Multi-risk assessment of heat waves under intensifying climate change using Bayesian Networks. *Int. J. Disaster Risk Reduct.* **50**, 101704. https://doi.org/10.1016/j.ijdrr.2020.101704 (2020).
21. Massel, A. & Daria, G. Scenario approach for analyzing extreme situations in energy from a cybersecurity perspective. *Industry 4.0* **3**(5), 266–269 (2018).
22. Shojaei Estabragh, Z. *et al.* Bayesian network modeling for diagnosis of social anxiety using some cognitive-behavioral factors. *Netw. Model. Anal. Health Inf. Bioinf.* **2**, 257–265. https://doi.org/10.1007/s13721-013-0042-x (2013).
23. Lixandru-Petre, I. O. Modeling a Bayesian Network for a Diabetes Case Study. *2020 International Conference on e-Health and Bioengineering (EHB)*, Iasi, Romania, 2020, (pp. 1–4). IEEE. (2020) https://doi.org/10.1109/EHB50910.2020.9280179.
24. Jeon, B. J., & Ko, I. Y. Ontology-based semi-automatic construction of Bayesian network models for diagnosing diseases in e-health applications. In 2007 Frontiers in the Convergence of Bioscience and Information Technologies (pp. 595–602). IEEE. (2007) https://doi.org/10.1109/FBIT.2007.63
25. Sahin, O. *et al.* Spatial Bayesian Network for predicting sea level rise induced coastal erosion in a small Pacific Island. *J. Environ. Manag.* **238**, 341–351. https://doi.org/10.1016/j.jenvman.2019.03.008 (2019).
26. "Maryland Statewide Vehicle Crashes CY2018 Quarter 1", Data Gov. [Online]. Available: https://catalog.data.gov/dataset/maryland-statewide-vehicle-crashes-cy2018-quarter-1. [Accessed: 18.12.2023].
27. Haimes, Y. Y. *Risk modeling, assessment, and management* (John Wiley & Sons, 2005).
28. "Road Traffic Injuries," World Health Organization, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: 18.12.2023].
29. "National Highway Traffic Safety Administration", [Online]. Available: https://www.nhtsa.gov/. [Accessed: 18.12.2023].
30. "Speed and Accident Risk", European Commission, [Online]. Available: https://road-safety.transport.ec.europa.eu/eu-road-safety-policy/priorities/safe-road-use/safe-speed/archive/speed-and-accident-risk_en. [Accessed: 18.12.2023].

31. "Zero Deaths MD, 'Road Safety & Driving Behaviors Resulting in Crashes," Maryland HighwaySafety Office, [Online]. Available: https://zerodeathsmd.gov/road-safety/. [Accessed: 18.12.2023].
32. Petridou, E. & Moustaki, M. Human factors in the causation of road traffic crashes. *Eur. J. Epidemiol.* **16**, 819–826. https://doi.org/10.1023/A:1007649804201 (2000).
33. Lakhan, R., Pal, R., Baluja, A., Moscote-Salazar, L. R. & Agrawal, A. Important aspects of human behavior in road traffic accidents. *Indian J. Neurotrauma* **17**(02), 085–089 (2020).
34. Bucsuházy, K. *et al.* Human factors contributing to the road traffic accident occurrence. *Transp. Res. Proc.* **45**, 555–561. https://doi.org/10.1016/j.trpro.2020.03.057 (2020).
35. Behnood, A. & Mannering, F. L. The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. *Traffic Injury Prevent.* **18**(5), 456–462 (2017).
36. Bakhshi, V., Aghabayk, K., Parishad, N. & Shiwakoti, N. Evaluating rainy weather effects on driving behaviour dimensions of driving behaviour questionnaire. *J. Adv. Transp.* **2022**, 1–10. https://doi.org/10.1155/2022/6000715 (2022).
37. Jima, D. & Sipos, T. The impact of road geometric formation on traffic crash and its severity level. *Sustainability* **14**(14), 8475. https://doi.org/10.3390/su14148475 (2022).
38. Cai, A. *et al.* Younger drivers are more impaired by sleep loss than older drivers, with blink duration and eye closures increased for younger drivers only. *Sci. Rep.* **11**, 19644. https://doi.org/10.1038/s41598-021-99133-y (2021).
39. Ulleberg, P. & Rundmo, T. Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Saf. Sci.* **41**(5), 427–443. https://doi.org/10.1016/S0925-7535(01)00077-7 (2003).
40. Hatfield, J. & Fernandes, R. The role of risk-propensity in the risky driving of younger drivers. *Accid. Anal. Prevent.* **41**(1), 25–35. https://doi.org/10.1016/j.aap.2008.08.023 (2009).
41. Payyanadan, R. P., Sanchez, F. A. & Lee, J. D. Route familiarity breeds risk taking in older adult drivers. *IEEE Trans. Human-Mach. Syst.* **49**(1), 20–31. https://doi.org/10.1109/THMS.2018.2874180 (2019).
42. Rhodes, N. & Pivik, K. Age and gender differences in risky driving: The roles of positive affect and risk perception. *Accid. Anal. Prevent.* **43**(3), 923–931. https://doi.org/10.1016/j.aap.2010.11.015 (2011).
43. Song, X. *et al.* The mediating effect of driver characteristics on risky driving behaviors moderated by gender, and the classification model of driver's driving risk. *Accid. Anal. Prevent.* **153**, 106038. https://doi.org/10.1109/ICTIS60134.2023.10243731 (2021).
44. Shyhalla, K. Alcohol involvement and other risky driver behaviors: effects on crash initiation and crash severity. *Traffic Injury Prevent.* **15**(4), 325–334. https://doi.org/10.1080/15389588.2013.822491 (2014).
45. Jørgenrud, B. *et al.* Association between speeding and use of alcohol and medicinal and illegal drugs and involvement in road traffic crashes among motor vehicle drivers. *Traffic Injury Prevent.* **19**(8), 779–785. https://doi.org/10.1080/15389588.2018.1518577 (2018).
46. Zhang, G., Yau, K. K., Zhang, X. & Li, Y. Traffic accidents involving fatigue driving and their extent of casualties. *Accid. Anal. Prevent.* **87**, 34–42. https://doi.org/10.1016/j.aap.2015.10.033 (2016).
47. Williamson, A. *et al.* The link between fatigue and safety. *Accid. Anal. Prevent.* **43**(2), 498–515. https://doi.org/10.1016/j.aap.2009.11.011 (2011).
48. Wang, C., Quddus, M. A. & Ison, S. G. The effect of traffic and road characteristics on road safety: A review and future research direction. *Saf. Sci.* **57**, 264–275. https://doi.org/10.1016/j.ssci.2013.02.012 (2013).
49. Carlson, A., & Vieira, T. (2021). The effect of water and snow on the road surface on rolling resistance. Statens väg-och transportforskningsinstitut.
50. Chen, X. & Wang, H. Analysis and mitigation of hydroplaning risk considering spatial-temporal water condition on the pavement surface. *Int. J. Pave. Eng.* **24**(2), 2036988. https://doi.org/10.1080/10298436.2022.2036988 (2023).
51. Fu, L., Thakali, L., Kwon, T. J. & Usman, T. A risk-based approach to winter road surface condition classification. *Canad. J. Civ. Eng.* **44**(3), 182–191. https://doi.org/10.1139/cjce-2016-0215 (2017).
52. Evans, T., Stuckey, R. & Macdonald, W. Young drivers' perceptions of risk and difficulty: Day versus night. *Accid. Anal. Prevent.* **147**, 105753. https://doi.org/10.1016/j.aap.2020.105753 (2020).
53. Mikoski, P., Zlupko, G. & Owens, D. A. Drivers' assessments of the risks of distraction, poor visibility at night, and safety-related behaviors of themselves and other drivers. *Transp. Res. Part F Traffic Psychol. Behav.* **62**, 416–434. https://doi.org/10.1016/j.trf.2019.01.011 (2019).
54. Blackman, R. A. & Haworth, N. L. Comparison of moped, scooter and motorcycle crash risk and crash severity. *Accid. Anal. Prevent.* **57**, 1–9. https://doi.org/10.1016/j.aap.2013.03.026 (2013).
55. Høye, A. How would increasing seat belt use affect the number of killed or seriously injured light vehicle occupants?. *Accid. Anal. Prevent.* **88**, 175–186. https://doi.org/10.1016/j.aap.2015.12.022 (2016).
56. Miyajima, C., Ukai, H., Naito, A., Amata, H., Kitaoka, N., & Takeda, K. Driver risk evaluation based on acceleration, deceleration, and steering behavior. *In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1829–1832). IEEE. (2011) https://doi.org/10.1109/ICASSP.2011.5946860
57. von Hoermann, C., Pagany, R., Kirchner, K., Dorner, W., Heurich, M., & Storch, I. Predicting the risk of deer-vehicle collisions by inferring rules learnt from deer experience and movement patterns in the vicinity of roads. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 368–373). IEEE. (2020) https://doi.org/10.1109/ACIT49673.2020.9208843
58. LernerMalka, B. R. Investigation of the K2 algorithm in learning Bayesian network classifiers. *Appl. Artif. Intell.* **25**(1), 74–96. https://doi.org/10.1080/08839514.2011.529265 (2011).
59. Ma, X., Xing, Y. & Lu, J. Causation analysis of hazardous material road transportation accidents by Bayesian network using genie. *J. Adv. Transp.* https://doi.org/10.1155/2018/6248105 (2018).
60. Scanagatta, M., Salmerón, A. & Stella, F. A survey on Bayesian network structure learning from data. *Progr. Artif. Intell.* **8**, 425–439. https://doi.org/10.1007/s13748-019-00194-y (2019).
61. Behjati, S., & Beigy, H. An order-based algorithm for learning structure of bayesian networks. In International Conference on Probabilistic Graphical Models (pp. 25–36). PMLR. (2018)
62. Cooper, G. F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9**, 309–347. https://doi.org/10.1007/BF00994110 (1992).
63. Demissie, S., LaValley, M. P., Horton, N. J., Glynn, R. J. & Cupples, L. A. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat. Med.* **22**(4), 545–557. https://doi.org/10.1002/sim.1340 (2003).
64. Fenton, N. & Neil, M. *Risk assessment and decision analysis with Bayesian networks* (CRC Press, 2018).
65. Jensen, F. V. & Nielsen, T. D. *Bayesian networks and decision graphs* Vol. 2 (Springer, 2007).
66. Lauritzen, S. L. The EM algorithm for graphical association models with missing data. *Comput. Stat. Data Anal.* **19**(2), 191–201. https://doi.org/10.1016/0167-9473(93)E0056-A (1995).
67. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodological)* **39**(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x (1977).
68. Cheng, J., Bell, D., & Liu, W. (1998). Learning Bayesian networks from data: An efficient approach based on information theory. *Retrieved from:* http://www.cs.ualberta. *ca/~ jcheng/bnpc. htm.*
69. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan kaufmann, 1988).

70. Marcot, B. G. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecol. Model.* **230**, 50–62. https://doi.org/10.1016/j.ecolmodel.2012.01.013 (2012).
71. Marcot, B. G. (2006). Characterizing species at risk I: modeling rare species under the Northwest Forest Plan. Ecology and Society, 11(2). Available: http://www.jstor.org/stable/26266002. [Accessed: 18.12.2023]
72. Neapolitan, R. E. *Probabilistic reasoning in expert systems: theory and algorithms* (John Wiley & Sons Inc, 1990).
73. "Sensitivity Equations in Netica Documentation", Norsys Software Corp. [Online]. Available: https://www.norsys.com/WebHelp/NETICA/X_Sensitivity_Equations.htm. [Accessed: 18.12.2023].
74. "Drunk Driving", National Highway Traffic Safety Administration, [Online]. Available: https://www.nhtsa.gov/risky-driving/drunk-driving. [Accessed: 18.12.2023].
75. Mørland, J. *et al.* Drugs related to motor vehicle crashes in northern European countries: a study of fatally injured drivers. *Accid. Anal. Prevent.* **43**(6), 1920–1926. https://doi.org/10.1016/j.aap.2011.05.002 (2011).
76. Behnood, A. & Mannering, F. L. The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. *Traffic Injury Prevent.* **18**(5), 456–462. https://doi.org/10.1080/15389588.2016.1262540 (2017).
77. Maistros, A., Schneider, W. H. IV. & Savolainen, P. T. A comparison of contributing factors between alcohol related single vehicle motorcycle and car crashes. *J. Saf. Res.* **49**, 129-e1. https://doi.org/10.1016/j.jsr.2014.03.002 (2014).
78. National Institute on Alcohol Abuse and Alcoholism, "Health Topics: Alcohol and the Brain," National Institutes of Health, [Online]. Available: https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics/health-topics-alcohol-and-brain. [Accessed: 24.12.2023]

Author contributions
C.C. wrote the main manuscript text and prepared all the figures and tables.

Competing interests
The author declares no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-69740-6.

Correspondence and requests for materials should be addressed to C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.