



Article

Exploring Causal Factor in Highway–Railroad-Grade Crossing Crashes: A Comparative Analysis

Yubo Wang ¹, Yubo Jiao ², Liping Fu ^{1,*} and Qiangqiang Shangguan ³¹ Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada; y3878wan@uwaterloo.ca² Department of Civil Engineering, McGill University, Montreal, QC H3A0G4, Canada; yubo.jiao@mcgill.ca³ The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Department of Road and Airport Engineering, College of Transportation Engineering, Tongji University, Shanghai 201804, China; qiangshangguan@tongji.edu.cn

* Correspondence: liping.fu@uwaterloo.ca; Tel.: +51-9888-4567 (ext. 43984)

Abstract

Identification of causal factors in traffic crashes has always been a significant challenge in road safety studies. Traditional crash prediction models are limited in elucidating the underlying causal mechanisms in road crashes. This research explores the application of three graphic models, namely, the Gaussian graphical model (GGM), causal Bayesian network (CBN) and graphic extreme gradient boosting (XGBoost), through a case study using highway–railroad-grade crossing (HRGC) inventory and collision data from Canada. The three modelling approaches have generally yielded consistent findings on various risk factors such as crossing control type, track angle, and exposure, showing their potential for identifying causal relationships through the interpretation of causal graphs. With the ability to make better causal inferences from crash data, the effectiveness of safety countermeasures could be more accurately and reliably estimated.

Keywords: graphic learning methods; crash mechanism; HRGC crashes; causal inference

Academic Editors: Dimitrios Nikolaou and Panagiotis Papantoniou

Received: 10 July 2025

Revised: 5 August 2025

Accepted: 9 August 2025

Published: 18 August 2025

Citation: Wang, Y.; Jiao, Y.; Fu, L.; Shangguan, Q. Exploring Causal Factor in Highway–Railroad-Grade Crossing Crashes: A Comparative Analysis. *Infrastructures* **2025**, *10*, 216. <https://doi.org/10.3390/infrastructures10080216>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Ensuring traffic safety is pivotal for fostering economic growth and safeguarding public health. Among various safety challenges, highway–rail-grade crossings (HRGCs) represent a critical point of vulnerability, where the intersection of rail and road infrastructure creates a high risk of severe and often fatal collisions. Annually, around 23 people lose their lives and another 28 suffer serious injuries at railway crossings in Canada. According to Transport Canada, over CAD 85 million has been invested in the Rail Safety Improvement Program (RSIP) over the past four years annually to decrease the crash rates and mitigate the adverse effects of traffic crashes [1]. Numerous attempts have been made by safety professionals and traffic engineers through such safety management programs to improve the safety interventions [2–4]. One crucial motivation for traffic safety studies is to investigate the relationship between various risk factors and crash outcomes, including crash frequencies and severity [5]. Understanding the crash causal risk factors can lead to enhanced safety improvement measures, thus reducing the occurrence probability and severity of crashes.

The general framework for safety improvement studies involves constructing crash prediction models to investigate the impact of various risk factors on the crash outcomes [6].

Traditionally, researchers focus on conducting quantitative analysis and developing generalized linear regression models for the relationship between crashes and potential risk factors [6]. Although such attempts are successful in conducting collective analysis in studying the influence exerted by environmental, vehicular, human-related, and facility-relevant factors, these approaches are inherently limited in understanding causal mechanisms of crashes due to the complexity of interacting variables from massive crash datasets. Meanwhile, capturing and establishing comprehensive causal relationships among various dynamic features presented in crash data necessitates novel approaches given that traditional statistical approaches often fall short in this regard.

With the advancement of causal learning, new opportunities have been afforded to understand crash mechanisms utilizing graphic models such as the Gaussian graphical model (GGM) and Bayesian network (BN), which can reveal the causal links in crashes with representative causal graphs. However, there has been limited effort in applying these models for causal inference in road safety. The objective of this research is to conduct a comparative study of three different graphic models, including GGM, BN, and graphic extreme gradient boosting (XGBoost) for identifying the causal factors affecting highway–railroad-grade crossing (HRGC) crashes.

The rest of this paper is organized as follows. The second section presents a review on identifying relationships between factors associated with crashes, graphical models for safety studies, and challenges in causal inference of traffic crashes. The third part documents the detailed methodology for discovering the potential causal relationships between various risk factors and HRGC crashes through different graphical models. Lastly, followed by the results and discussions, this study concludes by summarizing its contributions to understanding the HRGC crash mechanisms, discussing the study's limitations, and providing recommendations for future research.

2. Literature Review

2.1. Identification of Relationships Between Risk Factors and Crashes

Traditionally, the identification of relationships between risk factors and crashes relies on quantitative analysis [7] and statistical approaches [2–4]. Regression models have been the most applied approaches to model crash frequencies from the early studies up to recent times [8,9], including linear regression [10–14], Poisson regression [10,15–19], negative binomial regression [18,20–22], logistic regression [23–25], and multivariate regression [22,26–28]. Logit-based regression models, such as mixed logit regression, have been utilized to unravel the associations between risk factors and injury severities. Meanwhile, researchers have integrated regressions with random parameter models [20,22,29,30] and limited dependent variable models [31,32] to enhance prediction accuracy and improve model performance.

With advancements in safety study approaches, numerous machine learning models have been proposed to analyze correlations between factors and crashes, thus indicating the potential causal effects. For example, tree-based ensemble models are utilized to evaluate the importance of various factors contributing to traffic crashes. Researchers adopted Shapley additive explanations (SHAP) to explain the results attained from gradient boosting methods, especially extreme gradient boosting (XGBoost) to discover the significance of contributing factors [33–37], including applications in spatial safety modelling and freight-related incidents [38,39]. Meanwhile, recent attempts have explored the application of causal forests to estimate crash modification functions (CMFunctions) while addressing confounding bias to evaluate the joint effects of work zone features, traffic volume, and weather on crash risk [40]. Apart from these efforts, association rule mining (ARM) is

applied to identify frequent patterns and correlations among binary variables to uncover the relationship between contributing factors and crash outcomes [41,42].

Although these past efforts have enhanced the current understanding of relationships between risk factors and crash outcomes, the significance of variables and correlations among them cannot necessarily demonstrate causal relationships or establish causalities, thus falling short in revealing the causal mechanism of traffic crashes.

2.2. Risk Factor Identifications in Highway–Railway–Grade Crossing (HRGC) Crashes

Safety at HRGCs continues to be of significant concern since crashes at these locations often result in fatalities or serious injuries, as discussed previously [36]. Early attempts have focused on establishing models to predict crash frequency utilizing generalized linear models [43,44], consistently showing the influence of average daily vehicle traffic, daily train traffic, warning system, nighttime through-train traffic, train maximum speed, and the number of traffic lanes on the highway. Results indicate that daily train and vehicle traffic volumes can increase the probability of crash occurrence, while active warning systems can reduce the risk.

Recently, extensive research has been conducted to elucidate the risk factors and mechanisms that contribute to the occurrence of HRGC crashes with machine learning models. Researchers have developed tree-based methods to improve the performance of models for predicting the frequency and severity of collisions at HGRCs through risk factor identifications, including the decision tree and gradient boosting algorithms [45,46]. Lasisi et al. utilized classification methods including logistic regression, support vector machine, random forest, Gaussian naïve Bayes, and multi-layer perceptron-neural network to predict crashes [47]. Results from these studies reveal that variables such as daytime train movement, nighttime train movement, daily train traffic, train speed, and highway speed are associated with the crashes. Moreover, specific influential factors such as active warning systems like Flashing Light and Bells with Gates (FLBG) have been shown to reduce crash probabilities significantly [48].

Despite the contributions of the existing literature, risk factor identification remains limited to correlations rather than causality, which may hinder the improved understanding of the causal mechanisms behind crashes and thus the more accurate and reliable estimation of the effectiveness of safety countermeasures.

2.3. Graphical Models for Causal Inference in Traffic Crash Analysis

Various approaches have been applied to make causal inferences from traffic crash data, including Granger causality [28,49], instrumental variables (IV) [50,51], structural equation modelling (SEM) [52–54], and the a priori algorithm [33,54,55]. Recently, development in the machine learning field has afforded new opportunities to identify the causal relationship in traffic crashes; examples include extreme gradient boosting (XGBoost) [34,35,37], decision trees, and random forests [34,37,56] as well as deep neural networks [11,57,58]. In comparison with other approaches, XGBoost method is known for its high predictive accuracy and capability for handling complex relationships through gradient boosting, making it an effective tool for causal predictions [59]. Apart from these methods, various graphical models have shown the potential of extracting causal information through the construction of causal graphs [60]. Researchers have attempted to generate causal graphs through Bayesian networks (BNs) [14,32] and Markov random fields [32,61] in traffic-related studies. BNs use directed acyclic graphs (DAGs) to depict causal relationships [14,62,63], while MRFs, also known as forests or trees, use undirected graphs [61]. Edges in these graphs indicate the causal relationships between variables, enabling researchers to conduct causal inference based on these established causal links.

Despite the significant past efforts in risk factor analysis, researchers still face challenges when attempting to discover the causal relationships and draw causation assertions. One major obstacle is to establish illustrative and straightforward causal relationships, e.g., using causal graphs, of the crash causality between various risk factors and crash outcomes, such as occurrence frequency, from crash datasets [4]. In addition, the number of candidate variables has been increasing dramatically in recent years due to a significant increase in available data related to traffic crashes. Although the enrichment of variables allows a more comprehensive analysis of various confounding effects and interactions, understanding the crash causality requires a modelling methodology that can accurately capture the complexity of causal relationships in traffic crashes [4,27,32].

Another challenge is the reliability of observational studies. Generally, controlled experiments are utilized to draw causation conclusions in medical and environmental fields [64,65]. However, it is impossible to follow such a process in traffic safety studies due to ethical considerations. Thus, researchers have attempted to conduct analyses based on observations from historical crash data. And the reliability of the observational studies needs to be thoroughly explored given the uncontrolled nature of latent variables and missing data.

This paper aims to address the above-mentioned challenges through the comparison of different graphic causal discovery models, exploring the potential of applying these models for understanding causality in traffic crashes.

3. Methodology

In this section, the causal inference problem is stated by introducing the framework of the study. Following this, the methods utilized in this paper are demonstrated.

3.1. Identification of Causal Relationships

In this study, the causal discovery process is framed around establishing and optimizing causal relationships through constructing and refining causal graphs. The most widely recognized approach to represent causal relationships is directed acyclic graph (DAG), which consists of a set of nodes representing the variables in the study such as risk factors (e.g., driver age, road conditions, and traffic control type), environment conditions (e.g., weather and environment), and crash outcomes (e.g., crashes and fatalities) and a set of directed edges between nodes indicating the conditional dependence between different variables or causal influence of one variable on another [66,67], as shown in Figure 1.

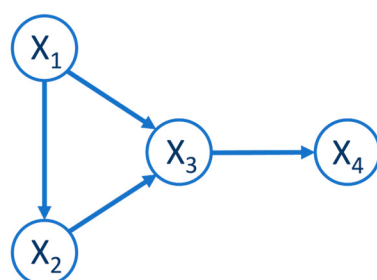


Figure 1. Example Directed Acyclic Graph (DAG).

Consider a specific DAG with a set of variables $X = \{X_i\}_{i=1}^M$, where X_i represents a potential factor corresponding to a node i . Each X_i is linked to other nodes—its parent nodes and its value could be estimated from the functional relationship of its parent nodes within the graph, augmented by an independent additive noise term ε_i , as defined in Equation (1).

$$X_i = f_i(X_{parent(i)}) + \varepsilon_i \quad (1)$$

where $X_{parent(i)}$ represents the parent nodes of node i ; ε_i is the noise component; $fi(.)$ is the link function to be calibrated using data from real-world observations. Let $D = [x_k]$, where x_k is a vector of k th observations of all variables in the graph; that is, $\varphi_k := [x_{k,1}, x_{k,2}, \dots, x_{k,i}, \dots, x_{k,M}]^T \in \mathbb{R}^d$.

In this paper, three different models are integrated to identify causal relationships within a directed acyclic graph (DAG) framework, as shown in Figure 2.

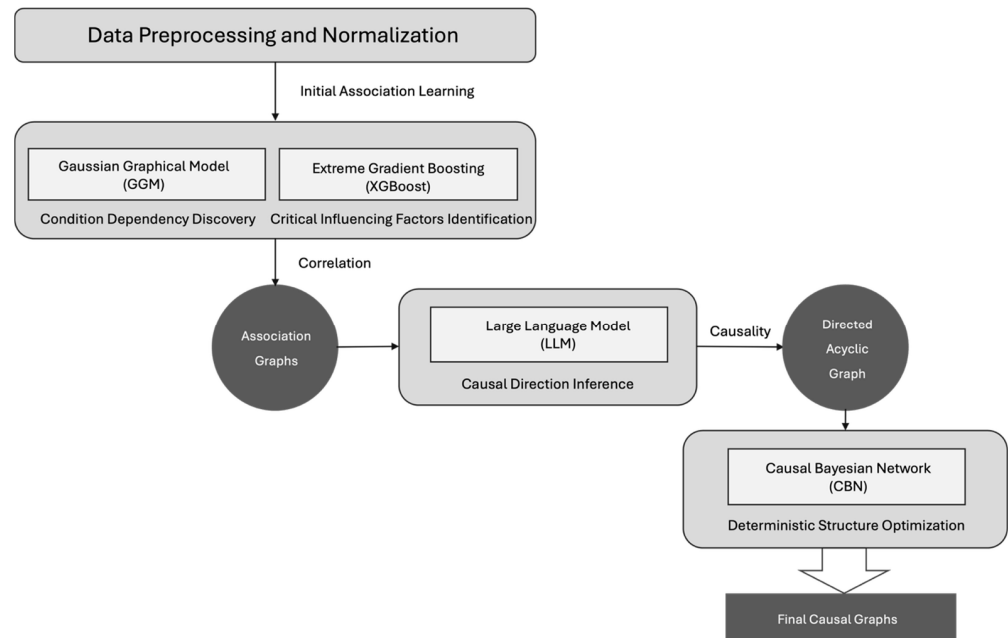


Figure 2. Integrated Workflow of the Causal Discovery Framework.

The Gaussian graphical model (GGM) is employed for efficiently capturing conditional dependencies among variables through sparse inverse covariance estimation, making it suitable for high-dimensional accident data. Extreme gradient boosting (XGBoost) is used to identify non-linear associations and feature importance, offering interpretable predictors via SHAP values. These two models collectively provide the structural foundation for causal Bayesian network (CBN) learning, which formalizes the directionality of relationships and quantifies causal effects through probabilistic reasoning. This multi-model approach ensures a balance between computational efficiency, interpretability, and causal rigor. The directionality among variables is primarily informed by domain knowledge and empirical patterns observed in the dataset. The following subsections provide detailed explanations of each method.

3.2. Gaussian Graphical Model (GGM)

In this paper, the Gaussian graphical model (GGM) is applied to handle the collinearity among variables [68] through Markov random field (MRF). MRF is an undirected graphical model that represents the conditional dependencies between random variables using an undirected graph [69], where each node corresponds to a variable, and edges denote direct probabilistic interactions between the variables, thereby capturing the joint probability distribution of the entire system. And GGM constructs a sparse graph by estimating the inverse of the covariance matrix of a data matrix X with n multivariate normal observations and p dimensions. The approach involves regularizing the estimation process by imposing an L_1 penalty on the elements of the precision matrix (the inverse of the covariance matrix). Denoting the empirical covariance matrix as S , the GLE can be obtained from the steps below (Equations (2)–(6)).

1. Empirical Covariance Matrix Calculation

$$S = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T \quad (2)$$

where x_i is the i -th observation in the data matrix $X = (x_1, x_2, \dots, x_n)$, and μ is the mean vector of the observations.

2. Likelihood Function

Assuming that the data X follow a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, the likelihood function can be written as

$$L = \prod_{i=1}^n \mathcal{N}(x_i, \Sigma) = \prod_{i=1}^n \frac{1}{2\pi^{\frac{p}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)} \quad (3)$$

The log-likelihood function is

$$\log(L) = -\frac{n}{2} (p \log(2\pi) + \log(\det(\Sigma))) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \quad (4)$$

3. Simplifying the Log-likelihood Function

Removing constant terms and simplifying the log-likelihood function can be attained as

$$\log(L) \propto -\log(\det(\Sigma)) - \text{tr}(S \Sigma^{-1}) \quad (5)$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix.

4. Adding the L_1 Penalty

To enforce sparsity, an L_1 penalty on the elements of the inverse covariance matrix $\Theta = \Sigma^{-1}$ is introduced. The optimization problem becomes

$$\hat{\Theta} = \underset{\Theta}{\text{argmin}} (-\log(\det(\Theta)) + \text{tr}(S\Theta) + \lambda \|\Theta\|_1) \quad (6)$$

where λ is the tuning parameter controlling the sparsity, and $\|\Theta\|_1$ is the L1 norm of Θ .

5. Convex Optimization

The above optimization problem is a convex problem and can be solved firstly through the Gaussianization of the data with no prior latent clustering or constraints, relaxing the normality assumption required for GLE [70]. The resulting graph is identified through this optimization procedure by applying the rotation information criterion (RIC) approach to find the optimal λ value regarding the relationship between variables [70]. In the proposed research, the scikit-learn package in Python 3.11.5 is applied to estimate the GGM model [71].

3.3. Extreme Gradient Boosting (XGBoost)

XGBoost has recently been applied to risk factor identifications through tree-based feature classification methods [60]. In this research, XGBoost is employed as another alternative for identifying the potential causal relationship due to its capability to capture complex patterns and interactions effectively [35,36]. With its tree-based construct for representing the relationships between the outcome of interest (e.g., collisions) and influencing factors, as shown in Figure 3, the XGBoost algorithm quantifies the effects of specific variables and their degree of importance, thereby offering an interpretable approach to understanding complex interactions within the data.

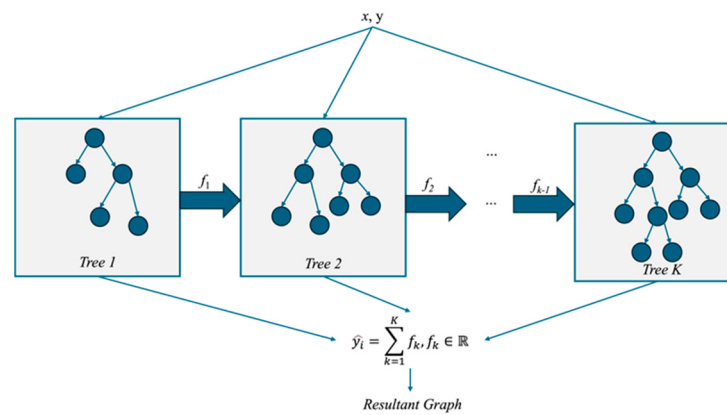


Figure 3. The General Architecture of Graphic XGBoost.

Despite its strength, the results from the XGBoost method, such as feature importance scores, do not allow direct causal interpretations, which would require domain knowledge with additional investigation. While these scores indicate predictive relevance, translating them into potential causal impacts and applying them for quantifying the intervention effectiveness requires careful domain-specific understanding and adjustments [72].

In this paper, the “importance” of each predictor variable is evaluated based on its total “gain” contribution to the resulting XGBoost model. Variables with a gain value exceeding 0.02 are considered important [73] and are subsequently used to construct graphical representations for further analysis. By integrating these results, XGBoost contributes to the overall framework by identifying critical variables and supporting the development of robust causal structures.

3.4. Causal Bayesian Network (CBN) Learning

A causal Bayesian network (CBN) is a graphical model that represents a set of variables and their conditional dependencies through the Bayesian network modelling. This model is encapsulated within a directed acyclic graph (DAG), providing a structured representation of causal relationships between candidate variables. In a CBN, nodes represent variables, and directed edges represent causal influences from parent nodes (as nodes A and B) to child nodes (as node C and D). Parent nodes are the direct causes of the child nodes, indicating that the state of a parent node can influence the state of its child nodes. By using directed edges to connect each variable to its children’s variables, a graph $\mathcal{G}$ can be constructed to represent the topology of the network. In a CBN model, the joint distributions of a set of variables X are specified by a decomposition as in Equation (7).

$$P(X) = \prod_{i=1}^n P(X_i | \prod_i G) \quad (7)$$

where $\prod_i G$ is a subset of $\{X_1, X_2, \dots, X_n\}$ and called the parent set of X_i . The causal relationship represented by $\mathcal{G}$ is calculated by Equation (8).

$$P(X) = \sum_{i=1}^n P(X_i)P(X|X_i) \quad (8)$$

In context of the Bayesian network, both the prior probability and the posterior probability should be calculated as the conditional independence between X_i and its parent nodes in the graph $\mathcal{G}$. Assume that a set of non-overlapping events exist in X , denoted as

$A_1, A_2, \dots, A_n$ and outcome E . Then, the posterior probability of any random event A with independence of outcome E is calculated as Equation (9).

$$P(A_i|E) = \frac{P(A_i)P(E|A_i)}{\sum_{j=1}^n P(A_j)P(E|A_j)} = \frac{P(A_i)P(E|A_i)}{P(A)} \quad (9)$$

To any random variable set X and its parent nodes $\text{Parent}(X_i)$ in the CBN, the joint distribution can be calculated through conditional probabilities as in Equation (10).

$$P(X) = \prod_{i=1, i \in n}^n P(X_i = x_i | X_{i+1} = x_{i+1}, \dots, X_n = x_n) = \prod_{i=1, i \in n}^n P(X_i = x_i | \text{Parent}(X_i)) \quad (10)$$

Therefore, the variable distributions can be understood through the identification of the probabilistic relationships. For the implementation of BN learning, the hill-climbing algorithm from the ‘pgmpy’ library in Python 3 was used to find the optimal DAG [73]. The final BN graph represents the statistically significant relationships between variables with a 95% confidence level.

4. Case Study Dataset

In this paper, the case study will focus on identifying the critical risk factors causing collisions at grade crossings in Canada and evaluating the effectiveness of some common safety countermeasures such as upgrading control devices and improving sight distance.

The dataset is presented as highway–railroad-grade crossings (HRGCs) inventory and collision databases provided by Transport Canada [74]. The inventory database contains data from 24,967 HRGCs across Canada, including features on intersecting railways and highways, control devices, and traffic exposure attributes such as train and vehicle traffic flow. The collision dataset includes all collisions that have happened at the HRGCs over the past eight years, including a total of 614 collisions and 75 fatalities. In the 2024 database, 614 collisions occurred at 560 HRGC, of which 91.6% had one collision, 7.0% had two collisions, and the remaining had three collisions.

Variables related to railway features, highway features, crossing features, traffic control features, and pedestrians are considered as the influential factors of collision occurrence based on the existing literature [36,75]. Given that the focus of this paper is on understanding the risk factors contributing to collision occurrences, the dependent variable will represent the number of collisions for each HRGC as a count variable (e.g., 0, 1, 2, ...). The inventory dataset has many missing values, which are addressed with some commonly applied imputation techniques, such as K-nearest neighbours [76].

Table 1 presents a summary of the explanatory variables utilized in the analysis, along with detailed descriptions and statistics.

Table 1. Variable descriptions and statistics.

Variable	Description	Possible Values	Mean	STD	Min	Max
		Railroad				
Track Number	Number of railway tracks	≥ 0	1.17	0.46	0.00	10.00
Track Angle	Angle degree of track	≥ 0	78.86	22.33	0.00	173.00
Daily Trains	Total number of trains per day	≥ 0	7.91	11.02	0.00	289.00
Railway Speed Limit	Overall maximum speed for rail train approaching from both sides	≥ 0	38.240	21.45	0.00	250.00

Table 1. *Cont.*

Variable	Description	Possible Values	Mean	STD	Min	Max
Highway						
Highway Lanes	Number of highway lanes	≥ 0	1.840	0.63	0.00	7.000
Daily Vehicles	Number of vehicles per day	≥ 0	1110.59	3545.88	0.00	65,104.00
Daily Pedestrians	Number of pedestrians per day	≥ 0	1.55	6.98	0.00	284.00
Road Speed Limit	Road posted/unposted maximum speed for road approach from both sides	≥ 0	55.19	25.94	0.000	110.00
Environment						
Area Type	Whether the HRGC is in rural areas or urban areas	0—rural; 1—urban	-	-	-	-
Lightings	Lighting conditions along highway	0—unlighted; 1—single side; 2—both sides	-	-	-	-
Crossing						
Gradient Stopping Sight Distance	Percent of gradient stopping sight distance for left side	$[-100, 100]$	1.24	3.95	36.40	64.99
Distance to Intersection	Distance between the crossing and intersection for left side	≥ 0	7.83	19.32	0.00	500.00
Whistle	Train is required to whistle or not	0—no; 1—yes	-	-	-	-
Sign	Presence of signs at passive crossings	0—no; 1—yes	-	-	-	-
Stop	Presence of stop signs at passive crossings	0—no; 1—yes	-	-	-	-
Gate	Active warning devices for vehicles	0—no; 1—yes	-	-	-	-
Flashing lights & Bells (FLB)	Traffic control signal or device is interconnected to a crossing warning system or not	0—no; 1—yes	-	-	-	-
Crossing Type	Type of the crossing	0—passive; 1—active warning system	-	-	-	-
Collision	Whether a collision occurred or not	0—no collisions; 1—at least one collision	-	-	-	-

5. Results and Discussions

The discussion is organized as follows: first, the results obtained from the Gaussian graphical model (GGM) and extreme gradient boosting (XGBoost) are presented. These results highlight their respective contributions to identifying correlated factors and covariates that can provide the most significant predictive power. Following this, the causal links are identified using causal Bayesian network (CBN) learning, paving the way to the comparison among the results attained from different methods.

5.1. Gaussian Graphical Model (GGM) Results

In this section, potential causal relationships indicated by the coefficient among various variables are visualized through the correlation graphs, representing the potential causal mechanisms in the highway–railroad-grade crossings (HRGCs) as shown in Figures 3 and 4.

In Figure 4, blue links represent positive covariances, while red links indicate negative covariances. The thickness of each link corresponds to the strength of these relationships, with thicker links denoting stronger covariances. In this figure, the alpha parameter is set to 0.01 to retain as many correlations as possible, thereby maximizing the potential to identify underlying causal relationships. Figure 5 displays the coefficients between various variables as derived from the graphic GGM method, providing a detailed view of the interactions and their respective strengths. The heatmap provides valuable correlation insights into the dynamics of HRGC collisions, particularly in relation to traffic control features, paving the

way to causal discovery. A notable cluster of variables includes stop signs, flashing lights and bells, and collision. This cluster suggests the presence of these traffic control features may significantly influence the frequency of collisions. Among these features, flashing lights and bells appears to exhibit a stronger link to collision occurrence than stop signs, indicating that active warning devices might play a more crucial role in mitigating collision risks. Further analysis of the coefficient strengths could refine this interpretation.

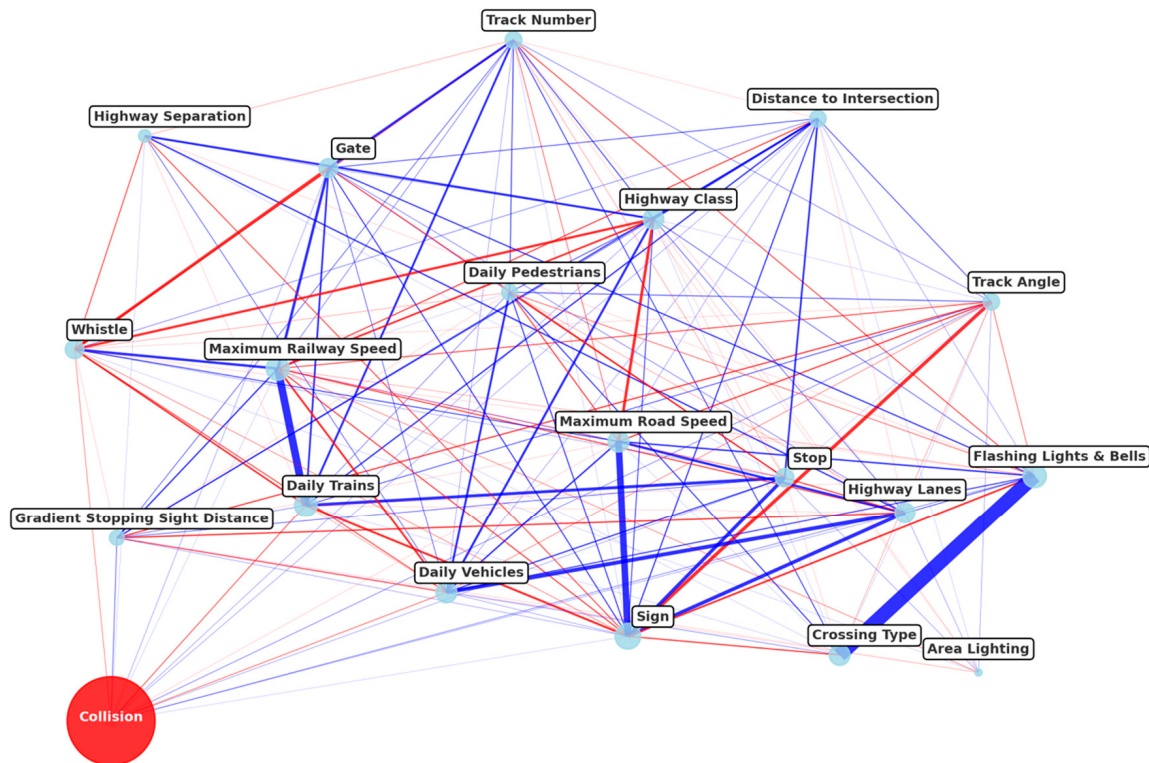


Figure 4. Resultant Causal Graph from GGM.

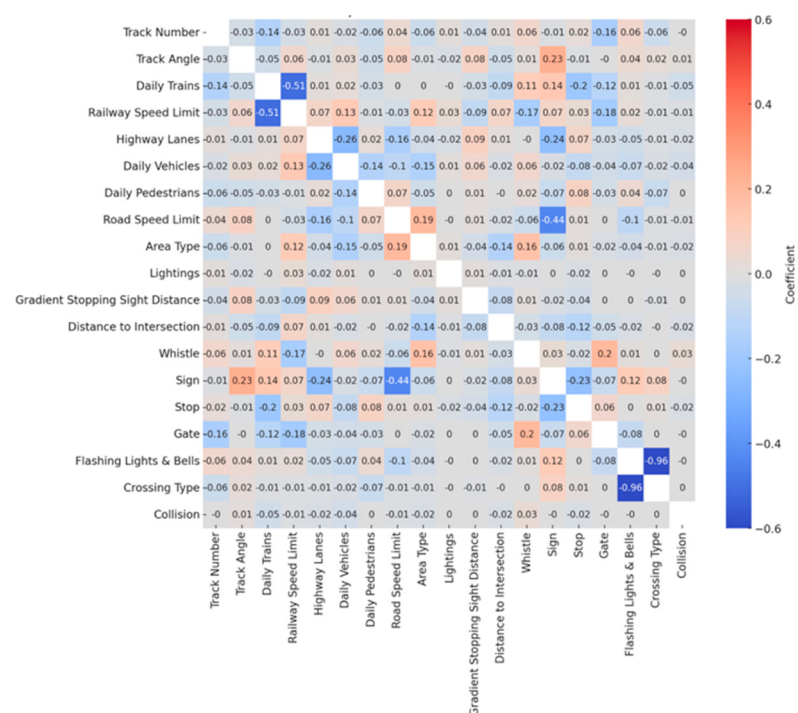


Figure 5. Heatmap of Coefficients between Variables (GGM).

The gradient stopping sight distance is found to have a negative coefficient with collision frequency (-0.04 ± 0.03), suggesting that longer sight distances may lower collision risk. Although a coefficient of -0.04 represents a moderate effect size, it is meaningful in the context of traffic safety, where small adjustments in sight distance can potentially reduce crash rates. Additionally, the relationship between track angle and collision (-0.03 ± 0.08) implies that as track angles diverge from straight alignments, there may be a slight increase in the likelihood of collisions.

To evaluate the impact of the regularization parameter α on the sparsity structure of the estimated correlation graph, a sensitivity analysis was conducted (Figure 6). It illustrates the relationship between varying α values and the resulting number of edges in the learned graph. As expected, the number of edges decreases monotonically with increasing α , reflecting the stronger penalization of partial correlations. At low α (e.g., 0.1–0.3), the model admits a dense network with over 200 edges, potentially including spurious or weak associations. In contrast, higher α values (≥ 0.7) yield a much sparser graph, highlighting only the strongest dependencies among variables. Notably, when $\alpha = 1.0$, the graph reduces to just 11 edges, suggesting a highly conservative structure. This trend underscores the necessity of selecting an optimal α value that balances model complexity with interpretability. Based on the sensitivity analysis, an α value between 0.5 and 0.7 offers a balanced trade-off—sufficiently reducing noise without discarding meaningful correlations—thus avoiding both overfitting and under-representation of potential causal links.

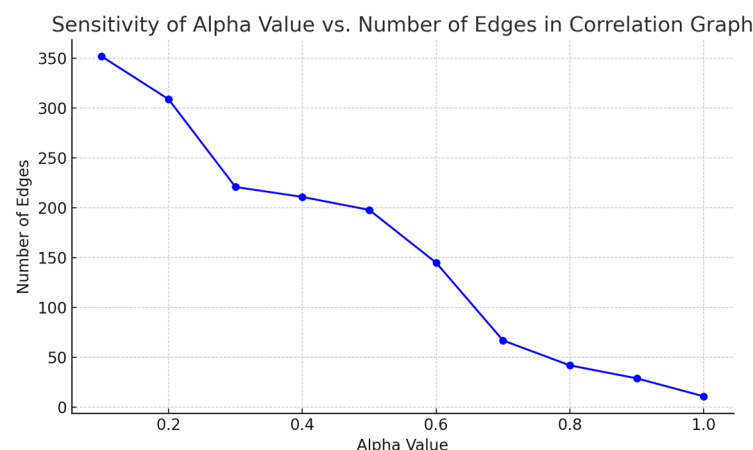


Figure 6. Sensitivity of Alpha Value vs. Number of Edges in Correlation Graph.

5.2. Graphic Extreme Gradient Boosting (XGBoost) Results

Using the graphic XGBoost method, the relationships among various risk factors are illustrated through their feature importance in predicting collisions, as shown in Figures 7 and 8. These figures present the results derived from the graphic XGBoost approach, indicating the critical variables influencing the crash outcomes. This method provides a broader perspective on the factors influencing collision frequency at highway–railway-grade crossings (HRGCs).

In this method, each feature importance value is normalized by dividing it by the maximum score in the matrix, yielding a relative measure of importance across variables. These normalized scores are presented using two complementary visualizations: a correlation graph (Figure 7) and a heatmap (Figure 8). In both figures, a consistent color scheme is applied—values greater than or equal to 0.5 are highlighted in red, while those below 0.5 appear in blue—clearly illustrating the comparative significance of each factor across the dataset.

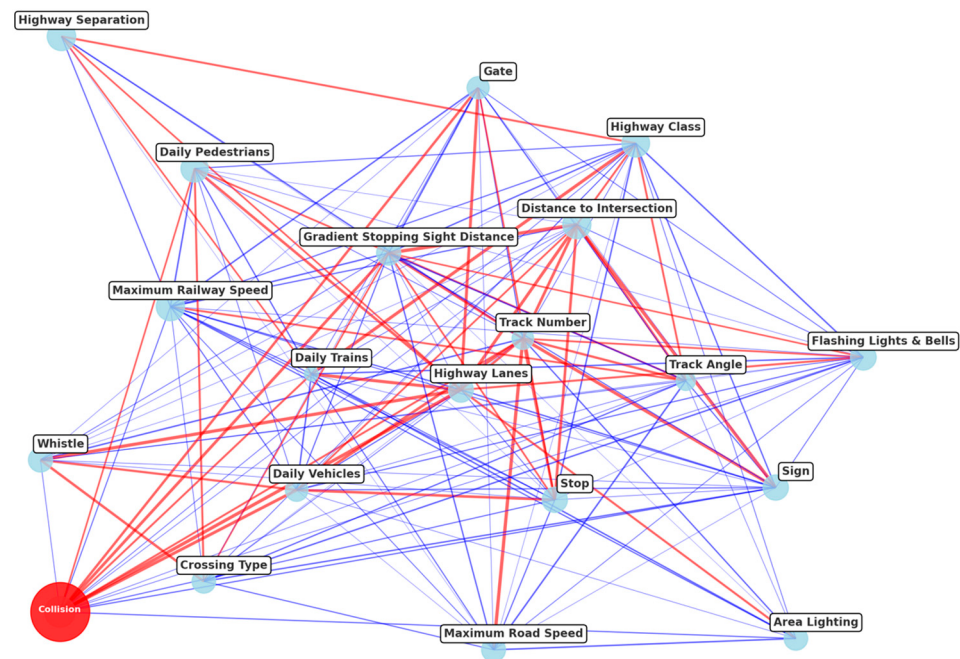


Figure 7. Resultant Causal Graph from Graphic XGBoost.

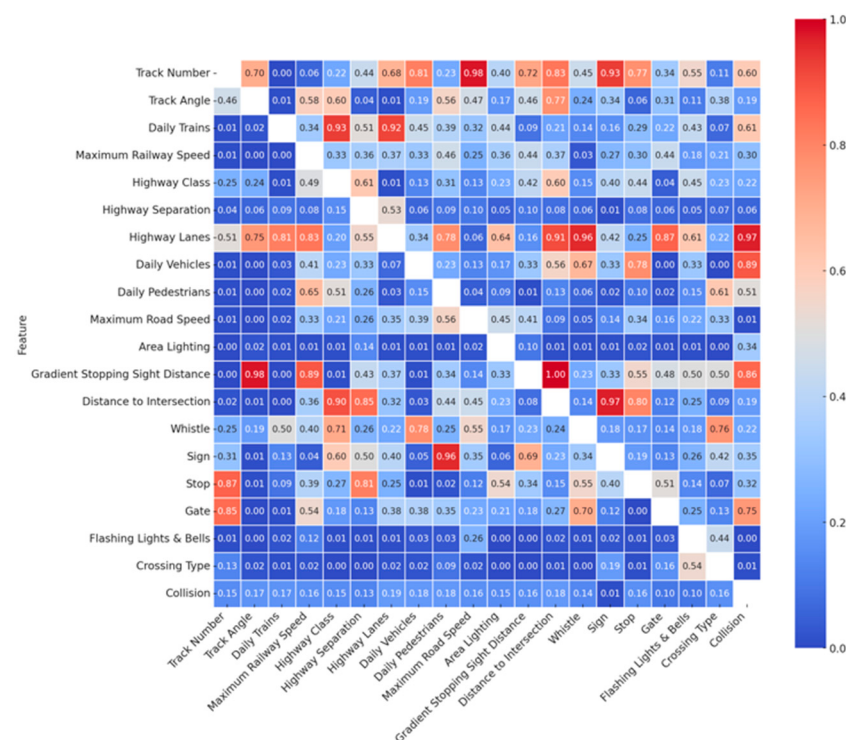


Figure 8. Heatmap of Importance between Variables (Graphic XGBoost, Normalized).

In terms of other significant variables, the analysis suggests that higher roadway classes, typically with more lanes, are associated with an increased likelihood of collisions. The variable importance of highway lanes in XGBoost (0.97 ± 0.17) and the positive link observed in the other methods imply that enhanced infrastructure, like dedicated lanes or additional traffic control measures, is necessary to improve safety in these areas. Moreover, the result reinforces the significance of features such as signs, stop signs, and flashing lights and bells in mitigating collision risks, with an importance score consistently around 0.06.

More importantly, the results indicate that distance to intersections and gradient stopping sight distance show significance in collisions. These features can influence drivers'

visibility and stopping distances, thereby contributing to crashes. The analysis also reveals that the maximum railway speed (0.30) has a notable impact on collision occurrences. Higher railway speeds likely increase the risk of collisions due to shorter reaction times and longer stopping distances required for trains. Additionally, highway lanes (0.97) and daily pedestrian volume (0.51) significantly affect collision rates, indicating that areas with more lanes and high pedestrian traffic near grade crossings are more susceptible to crashes. Furthermore, the presence of gates (0.75) appears to have an association with collisions.

5.3. Causal Bayesian Network (CBN) Learning Results

In this paper, the CBN method requires an initial tree structure as a starting point to optimize and generate the final directed acyclic graph (DAG). To construct this initial structure, the relationships derived from both coefficient values and feature importance in the previous two methods are utilized as input for the CBN approach. As a result, the generated DAG and the causal directionality between variables are depicted in Figure 8. In the CBN models, causal relationships are represented by the directionality of the edges in the DAG, illustrating how one variable influences another within the network.

Overall, the directionality of relationships derived from the CBN method provides deeper insights into the causal dynamics of traffic crashes. Consistent with the findings from the GGM analysis, the presence of signs, stop signs, and flashing lights and bells shows a relationship with collision occurrences at HRGCs, highlighting the importance of these features in influencing safety outcomes. Meanwhile, track angle appears to have a potential causal relationship with collision occurrences; however, further validation from the other two methods is needed to strengthen this finding. Furthermore, rather than recommending specific treatments, these findings indicate that these features could be part of a broader strategy for risk mitigation, with field personnel assessing which measures are most effective based on site-specific conditions. Separating the effects of each feature could provide further clarity on their individual contributions to safety. For instance, flashing lights and bells, as active warning devices, may offer more substantial protection in high-traffic areas compared to passive signs, which might be more suitable in less complex environments. Figure 9 illustrates a conceptual example of a causal graph that could be generated using the CBN method.

Additionally, there is a notable association between the frequency of daily trains and collision occurrences, suggesting that higher train frequency may increase exposure to potential collisions. The link between maximum road speed and collision occurrence also indicates that higher speeds may contribute to elevated collision risks, underscoring the need for tailored speed management strategies in areas with significant pedestrian or vehicle activity near crossings.

Meanwhile, the potential causal relationships between daily trains and the presence of gates as well as between track number and gates suggest that areas with higher train frequency and multiple tracks might benefit from enhanced protective measures such as gates to reduce collision risks. The impact of gates on collision frequency underscores their role as a critical safety feature in busy areas. Furthermore, the associations between area type, whistle, and flashing lights and bells indicate that certain environmental contexts (such as urban versus rural settings) influence the types of safety features implemented. Specifically, urban areas may incorporate more active warning devices, like flashing lights and bells, to address increased traffic and pedestrian presence.

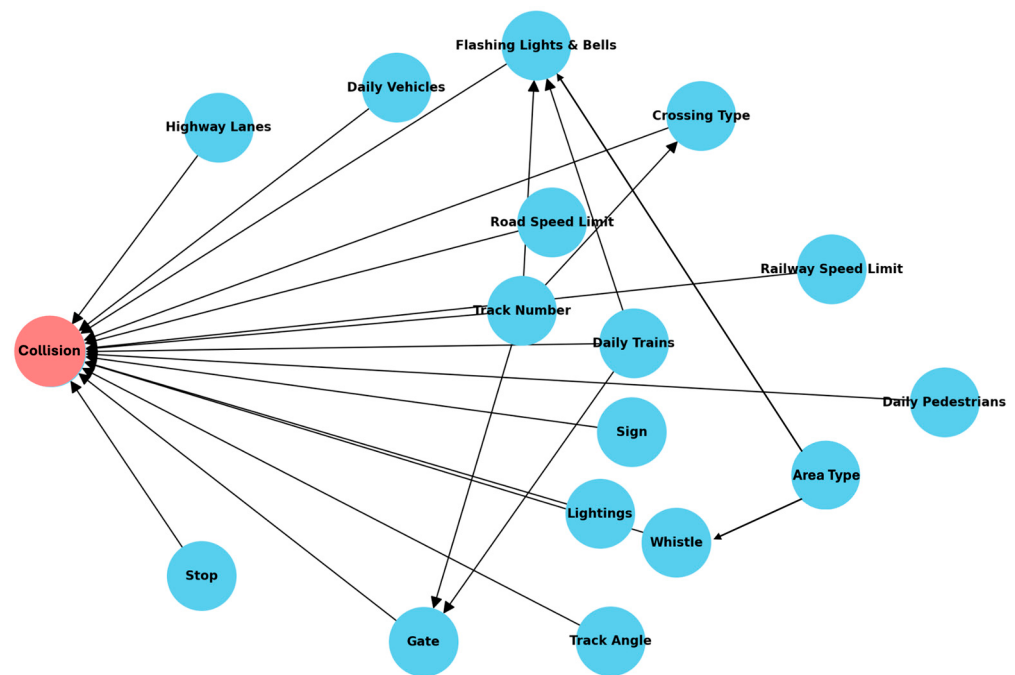


Figure 9. Resultant Causal Graph from CBN.

6. Conclusions and Limitations

6.1. Conclusions

In this paper, the analysis explores different causal discovery models, including the Gaussian graphical model (GGM), casual Bayesian network, and graphic extreme gradient boosting, based on the case study of highway–railroad-grade crossing (HRGC) crash datasets and consistently identifies key factors influencing collision occurrences and their interrelationships. These consistent findings provide a robust foundation for potential safety improvements at HRGCs. The details are as follows.

- Stop Signs and Flashing Lights and Bells

Across all three methods, the presence of stop signs and flashing lights and bells is shown to significantly reduce collision occurrences, which aligns with the findings in the existing literature [48]. These traffic control features provide crucial visual and auditory warnings, enhancing reaction times for drivers and pedestrians. Results show that the HRGC with higher pedestrian and traffic volumes require more comprehensive warning facilities to effectively reduce the crash risk.

- Track Angle

Track angle is identified as a contributing factor with potential causal links to both collision occurrence and train frequency. Specifically, crossings with acute (sharper) track angles are associated with a higher probability of collisions. This is likely due to reduced sight distance and more complex vehicle maneuvering paths, which can impair drivers' ability to detect and respond to approaching trains. Moreover, track angle shows a modest association with train frequency, which may further influence overall crash risk. These findings suggest the need for additional safety measures—such as improved signage, enhanced lighting, or advanced warning systems—at crossings with sharper angles to mitigate visibility challenges and increase driver and pedestrian awareness.

- Daily Trains

A significant correlation exists between the frequency of daily trains and collision occurrences, consistent with existing findings [77]. Higher train frequencies increase the

exposure to potential conflicts between trains and vehicles or pedestrians. This emphasizes the need for robust warning systems such as flashing lights and bells and physical barriers like crossing gates to manage these interactions effectively and prevent collisions.

- **Maximum Road Speed**

The analysis consistently shows that areas with higher pedestrian traffic tend to have lower maximum road speeds, likely due to safety measures that reduce speed limits in high pedestrian areas to minimize crash risks. Additionally, roads with higher traffic volumes often have higher speed limits, as indicated by the positive coefficient between daily vehicles and maximum road speed. This relationship further increases collision risks, as higher traffic volumes and speeds contribute to more dangerous crossing environments. This finding underscores the importance of implementing speed control measures, such as speed limits and traffic calming devices near GCs.

- **Daily Pedestrians**

The coefficient between daily pedestrians and collisions varies slightly among the methods but remains positive, indicating that higher pedestrian traffic increases the risk of collisions. This highlights the necessity for enhanced pedestrian safety measures, such as well-marked crosswalks, pedestrian signals, and audible warnings, at crossings with high foot traffic. These measures help ensure that pedestrians are aware of approaching trains and can cross safely.

In summary, the consistent findings across different analytical methods reinforce the reliability of the identified risk factors and their impact on HRGC safety. Implementing comprehensive safety measures that address these key factors—such as enhanced traffic control features, improved signage and lighting, robust warning systems, speed control measures, and pedestrian safety enhancements—can significantly reduce the risk of collisions and improve overall safety at grade crossings.

Beyond the specific causal insights uncovered in this study, a comparative reflection on the three graphical modeling approaches—GGM, XGBoost, and CBN—provides further understanding of their respective methodological characteristics and practical applicability. Each method offers unique advantages and trade-offs in terms of computational efficiency, scalability, interpretability, and suitability for causal inference in traffic safety analysis. GGM demonstrated high computational efficiency, leveraging convex optimization with regularized covariance estimation. Its rapid estimation of the precision matrix, even with a large number of variables, is consistent with findings from prior work on graphical lasso techniques (e.g., [67,68]), making it well-suited for high-dimensional exploratory analysis.

XGBoost also showed strong efficiency through its parallelized, tree-based architecture and yielded robust predictive performance. The use of SHAP values enabled interpretation of variable importance; however, its outputs are inherently predictive, requiring post hoc analyses to approximate causal meaning—thus limiting direct interpretability of causal relationships.

CBN, in contrast, offered explicit modeling of causal directionality through its directed acyclic graph (DAG) structure, providing valuable insights into the underlying mechanisms of observed relationships. Nonetheless, structure learning and parameter estimation proved computationally intensive, particularly with increasing data dimensionality. In our case study, the CBN model required considerable tuning and domain-informed constraints to remain tractable. Although its interpretability and alignment with causal reasoning are advantageous, its reliance on expert input and significant computational cost pose limitations for real-time or large-scale deployment.

These findings suggest a complementary role for these methods: GGM and XGBoost are well-suited for preliminary analyses and variable screening in high-dimensional

settings, whereas CBN is more appropriate for detailed causal investigations where interpretability and directional insight are essential.

6.2. Limitations and Future Work

This paper attempts to use graphical models to identify the causal relationships between various factors associated with highway–railroad-grade crossing (HRGC) crashes. The application of GLE, CBN, and XGBoost methods provides a robust analysis of the key risk factors influencing collision occurrences. The consistent findings across these methods validate the reliability of the identified factors and their causal relationships with collisions.

However, this study has several limitations that should be acknowledged. First, while graphical models can reveal associations among variables, there remains a risk of misinterpreting these links as direct causal relationships when they may instead reflect statistical dependencies inherent to the data-generating process. Although the Bayesian network (BN) approach offers improved interpretability by incorporating directionality through directed acyclic graphs (DAGs), the accuracy of inferred causal links still depends heavily on the quality and completeness of the data.

Second, the analysis focuses solely on collision occurrence without incorporating injury severity levels, which are critical for fully understanding the risk and consequences of HRGC incidents. Furthermore, the study is limited by the scope of available variables. Important contextual factors—such as weather conditions, lighting, time of day, and pedestrian infrastructure (e.g., pedestrian signals)—were not included due to data limitations in the HRGC inventory and collision databases. The absence of these variables may introduce residual confounding and reduce the precision of causal interpretations.

To address these issues, future research should aim to integrate additional data sources, such as historical weather archives, time-stamped incident logs, and pedestrian facility inventories.

Author Contributions: Conceptualization, L.F., Y.W., Y.J. and Q.S.; methodology, Y.W.; software, Y.W. and Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Transport Canada through its RSIP (Railway Safety Improvement Program) and Natural Science and Engineering Research Council of Canada Discovery Grant (NSERC DG).

Data Availability Statement: The data used in this study are openly available from Transport Canada’s Grade Crossings Inventory and Collision Database at the Government of Canada Open Data Portal, accessible at: <https://open.canada.ca/data/en/dataset/d0f54727-6c0b-4e5a-aa04-ea1463cf9f4c> (accessed on 9 July 2025).

Acknowledgments: During the preparation of this manuscript/study, the author(s) used ChatGPT 4o for the purposes of grammar checking and certain visualization generations. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Transport Canada. 2024–2025 *Departmental Plan*; Government of Canada: Ottawa, ON, Canada, 2024; Available online: https://tc.canada.ca/sites/default/files/2024-03/2024-25_DEPARTMENTAL_PLAN_ENGLISH.pdf (accessed on 9 July 2024).
2. Cerwick, D.M.; Gkritza, K.; Shaheed, M.S.; Hans, Z. A Comparison of the Mixed Logit and Latent Class Methods for Crash Severity Analysis. *Anal. Methods Accid. Res.* **2014**, 3–4, 11–27. [CrossRef]
3. Hauer, E. Cause, Effect and Regression in Road Safety: A Case Study. *Accid. Anal. Prev.* **2010**, 42, 1128–1135. [CrossRef] [PubMed]
4. Hauer, E. Crash Causation and Prevention. *Accid. Anal. Prev.* **2020**, 143, 105528. [CrossRef] [PubMed]
5. Elvik, R. Problems in Determining the Optimal Use of Road Safety Measures. *Res. Transp. Econ.* **2014**, 47, 27–36. [CrossRef]
6. Hauer, E. Statistical Road Safety Modeling. *Transp. Res. Rec.* **2004**, 1897, 81–87. [CrossRef]

7. Khan, K.; Zaidi, S.B.; Ali, A. Evaluating the Nature of Distractive Driving Factors towards Road Traffic Accident. *Civ. Eng. J.* **2020**, *6*, 1555–1580. [\[CrossRef\]](#)
8. Lord, D.; Mannering, F. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transp. Res. Part. A Policy Pract.* **2010**, *44*, 291–305. [\[CrossRef\]](#)
9. Elvik, R. Risk Factors as Causes of Accidents: Criterion of Causality, Logical Structure of Relationship to Accidents and Completeness of Explanations. *Accid. Anal. Prev.* **2024**, *197*, 107469. [\[CrossRef\]](#)
10. Joshua, S.C.; Garber, N.J. Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models. *Transp. Plan. Technol.* **1990**, *15*, 41–58. [\[CrossRef\]](#)
11. Zhang, J.; Wang, J.; Fang, S. Prediction of Urban Expressway Total Traffic Accident Duration Based on Multiple Linear Regression and Artificial Neural Network. In Proceedings of the 2019 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, UK, 14–17 July 2019; IEEE: New York, NY, USA, 2019; pp. 503–510.
12. Cai, H.; Zhu, D.; Yan, L. Using Multi-Regression to Analyze and Predict Road Traffic Safety Level in China. In Proceedings of the 2015 International Conference on Transportation Information and Safety (ICTIS), Wuhan, China, 25–28 June 2015; IEEE: New York, NY, USA, 2015; pp. 363–369.
13. Goldstein, H. Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika* **1986**, *73*, 43–56. [\[CrossRef\]](#)
14. Huang, H.; Abdel-Aty, M. Multilevel Data and Bayesian Analysis in Traffic Safety. *Accid. Anal. Prev.* **2010**, *42*, 1556–1565. [\[CrossRef\]](#)
15. El-Basyouny, K.; Barua, S.; Islam, M.T. Investigation of Time and Weather Effects on Crash Types Using Full Bayesian Multivariate Poisson Lognormal Models. *Accid. Anal. Prev.* **2014**, *73*, 91–99. [\[CrossRef\]](#)
16. Aguero-Valverde, J.; Jovanis, P.P. Analysis of Road Crash Frequency with Spatial Models. *Transp. Res. Rec.* **2008**, *2061*, 55–63. [\[CrossRef\]](#)
17. Abdel-Aty, M.A.; Radwan, A.E. Modeling Traffic Accident Occurrence and Involvement. *Accid. Anal. Prev.* **2000**, *32*, 633–642. [\[CrossRef\]](#)
18. Akin, D. Darçın Akin Analysis of Highway Crash Data by Negative Binomial and Poisson Regression Models. In Proceedings of the 2nd International Symposium on Computing in Science and Engineering, Izmir, Turkey, 1–4 June 2011. [\[CrossRef\]](#)
19. Wu, K.-F.; Aguero-Valverde, J.; Jovanis, P.P. Using Naturalistic Driving Data to Explore the Association between Traffic Safety-Related Events and Crash Risk at Driver Level. *Accid. Anal. Prev.* **2014**, *72*, 210–218. [\[CrossRef\]](#)
20. Coruh, E.; Bilgic, A.; Tortum, A. Accident Analysis with Aggregated Data: The Random Parameters Negative Binomial Panel Count Data Model. *Anal. Methods Accid. Res.* **2015**, *7*, 37–49. [\[CrossRef\]](#)
21. Usman, T.; Fu, L.; Miranda-Moreno, L.F. Accident Prediction Models for Winter Road Safety: Does Temporal Aggregation of Data Matter? *Transp. Res. Rec.* **2011**, *2237*, 144–151. [\[CrossRef\]](#)
22. Dong, C.; Clarke, D.B.; Yan, X.; Khattak, A.; Huang, B. Multivariate Random-Parameters Zero-Inflated Negative Binomial Regression Model: An Application to Estimate Crash Frequencies at Intersections. *Accid. Anal. Prev.* **2014**, *70*, 320–329. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Chen, P.; Jou, R.; Saleh, W.; Pai, C. Accidents Involving Pedestrians with Their Backs to Traffic or Facing Traffic: An Evaluation of Crash Characteristics and Injuries. *J. Advcd Transp.* **2016**, *50*, 736–751. [\[CrossRef\]](#)
24. Al-Ghamdi, A.S. Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity. *Accid. Anal. Prev.* **2002**, *34*, 729–741. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Yau, K.K.W. Risk Factors Affecting the Severity of Single Vehicle Traffic Accidents in Hong Kong. *Accid. Anal. Prev.* **2004**, *36*, 333–340. [\[CrossRef\]](#)
26. Elvik, R. Evaluating the Effectiveness of Norway’s “Speak Out!” Road Safety Campaign: The Logic of Causal Inference in Road Safety Evaluation Studies. *Transp. Res. Rec.* **2000**, *1717*, 66–75. [\[CrossRef\]](#)
27. Elvik, R. Assessing Causality in Multivariate Accident Models. *Accid. Anal. Prev.* **2011**, *43*, 253–264. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Ageli, M.M.; Zaidan, A.M. Road Traffic Accidents in Saudi Arabia: An ADRL Approach and Multivariate Granger Causality. *IJEF* **2013**, *5*, 26. [\[CrossRef\]](#)
29. Yu, R.; Xiong, Y.; Abdel-Aty, M. A Correlated Random Parameter Approach to Investigate the Effects of Weather Conditions on Crash Risk for a Mountainous Freeway. *Transp. Res. Part. C Emerg. Technol.* **2015**, *50*, 68–77. [\[CrossRef\]](#)
30. Anastasopoulos, P.C.; Mannering, F.L. A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accid. Anal. Prev.* **2009**, *41*, 153–159. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Kocatepe, A.; Ulak, M.B.; Ozguven, E.E.; Horner, M.W. Who Might Be Affected by Crashes? Identifying Areas Susceptible to Crash Injury Risk and Their Major Contributing Factors. *Transp. A Transp. Sci.* **2019**, *15*, 1278–1305. [\[CrossRef\]](#)
32. Ulak, M.B.; Ozguven, E.E. Identifying the Latent Relationships between Factors Associated with Traffic Crashes through Graphical Models. *Accid. Anal. Prev.* **2024**, *197*, 107470. [\[CrossRef\]](#)

33. Chen, H.; Chen, H.; Zhou, R.; Liu, Z.; Sun, X. Exploring the Mechanism of Crashes with Autonomous Vehicles Using Machine Learning. *Math. Probl. Eng.* **2021**, 2021, 1–10. [[CrossRef](#)]
34. Elyassami, S.; Hamid, Y.; Habuza, T. Road Crashes Analysis and Prediction Using Gradient Boosted and Random Forest Trees. In Proceedings of the 2020 6th IEEE Congress on Information Science and Technology (CiSt), Agadir-Essaouira, Morocco, 5–12 June 2021; IEEE: New York, NY, USA, 2020; pp. 520–525.
35. Parsa, A.B.; Movahedi, A.; Taghipour, H.; Derrible, S.; Mohammadian, A. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accid. Anal. Prev.* **2020**, 136, 105405. [[CrossRef](#)]
36. Rana, P.; Sattari, F.; Lefsrud, L.; Hendry, M. Machine Learning Approach to Enhance Highway Railroad Grade Crossing Safety by Analyzing Crash Data and Identifying Hotspot Crash Locations. *Transp. Res. Rec. J. Transp. Res. Board.* **2023**, 2678, 1055–1071. [[CrossRef](#)]
37. Cheng, Z.; Liu, B.; Huang, J. Causal Analysis of Road Safety Accidents in Britain Based on a Univariate Decision Tree Method. In Proceedings of the 2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI), Zakopane, Poland, 15–16 August 2022; IEEE: New York, NY, USA, 2022; pp. 436–441.
38. Yang, C.; Chen, M.; Yuan, Q. The Application of XGBoost and SHAP to Examining the Factors in Freight Truck-Related Crashes: An Exploratory Analysis. *Accid. Anal. Prev.* **2021**, 158, 106153. [[CrossRef](#)] [[PubMed](#)]
39. Ziakopoulos, A.; Vlahogianni, E.; Antoniou, C.; Yannis, G. Spatial Predictions of Harsh Driving Events Using Statistical and Machine Learning Methods. *Saf. Sci.* **2022**, 150, 105722. [[CrossRef](#)]
40. Zhang, Z.; Akinci, B.; Qian, S. Inferring Heterogeneous Treatment Effects of Work Zones on Crashes. *Accid. Anal. Prev.* **2022**, 177, 106811. [[CrossRef](#)] [[PubMed](#)]
41. Liu, P.; Guo, Y.; Liu, P.; Ding, H.; Cao, J.; Zhou, J.; Feng, Z. What Can We Learn from the AV Crashes?—An Association Rule Analysis for Identifying the Contributing Risky Factors. *Accid. Anal. Prev.* **2024**, 199, 107492. [[CrossRef](#)]
42. Hong, J.; Tamakloe, R.; Park, D. Application of Association Rules Mining Algorithm for Hazardous Materials Transportation Crashes on Expressway. *Accid. Anal. Prev.* **2020**, 142, 105497. [[CrossRef](#)] [[PubMed](#)]
43. Lu, D.P. Highway-Rail Grade Crossing Traffic Hazard Forecasting Model. In Proceedings of the Transportation Research Board Annual Meeting, Washington, DC, USA, 13–17 January 2019.
44. Saccomanno, F.F.; Fu, L.; Miranda-Moreno, L.F. Risk-Based Model for Identifying Highway-Rail Grade Crossing Blackspots. *Transp. Res. Rec.* **2004**, 1862, 127–135. [[CrossRef](#)]
45. Zheng, Z.; Lu, P.; Tolliver, D. Decision Tree Approach to Accident Prediction for Highway–Rail Grade Crossings: Empirical Analysis. *Transp. Res. Rec.* **2016**, 2545, 115–122. [[CrossRef](#)]
46. Soleimani, S.; Leitner, M.; Codjoe, J. Applying Machine Learning, Text Mining, and Spatial Analysis Techniques to Develop a Highway-Railroad Grade Crossing Consolidation Model. *Accid. Anal. Prev.* **2021**, 152, 105985. [[CrossRef](#)]
47. Lasisi, A.; Li, P.; Chen, J. Hybrid Machine Learning and Geographic Information Systems Approach—A Case for Grade Crossing Crash Data Analysis. *Adv. Data Sci. Adapt. Data Anal.* **2020**, 12, 2050003. [[CrossRef](#)]
48. Shangguan, Q.; Wang, Y.; Fu, L. Quantifying the Effectiveness of an Active Treatment in Improving Highway-Railway Grade Crossing Safety in Canada: An Empirical Bayes Observational before–after Study. *Can. J. Civ. Eng.* **2024**, 52, 17–28. [[CrossRef](#)]
49. Zhang, L.; Fu, K.; Ji, T.; Lu, C.-T. Granger Causal Inference for Interpretable Traffic Prediction. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; IEEE: New York, NY, USA, 2022; pp. 1645–1651.
50. Yu, Q.; Zhou, Y.; Ayele Atumo, E.; Qu, L.; Zhang, N.; Jiang, X. Addressing Endogeneity between Hazardous Actions and Motorcyclist Injury Severity by Integrating Generalized Propensity Score Approach and Instrumental Variable Model. *Accid. Anal. Prev.* **2023**, 192, 107297. [[CrossRef](#)] [[PubMed](#)]
51. Afghari, A.P.; Papadimitriou, E.; Pilkington-Cheney, F.; Filtness, A.; Brijs, T.; Brijs, K.; Cuenen, A.; De Vos, B.; Dirix, H.; Ross, V.; et al. Investigating the Effects of Sleepiness in Truck Drivers on Their Headway: An Instrumental Variable Model with Grouped Random Parameters and Heterogeneity in Their Means. *Anal. Methods Accid. Res.* **2022**, 36, 100241. [[CrossRef](#)]
52. Najaf, P.; Thill, J.-C.; Zhang, W.; Fields, M.G. City-Level Urban Form and Traffic Safety: A Structural Equation Modeling Analysis of Direct and Indirect Effects. *J. Transp. Geogr.* **2018**, 69, 257–270. [[CrossRef](#)]
53. Zheng, L.; Sayed, T.; Essa, M. Validating the Bivariate Extreme Value Modeling Approach for Road Safety Estimation with Different Traffic Conflict Indicators. *Accid. Anal. Prev.* **2019**, 123, 314–323. [[CrossRef](#)]
54. Shaaban, K.; Gaweesh, S.; Ahmed, M.M. Investigating In-Vehicle Distracting Activities and Crash Risks for Young Drivers Using Structural Equation Modeling. *PLoS ONE* **2020**, 15, e0235325. [[CrossRef](#)]
55. Xi, J.; Zhao, Z.; Li, W.; Wang, Q. A Traffic Accident Causation Analysis Method Based on AHP-Apriori. *Procedia Eng.* **2016**, 137, 680–687. [[CrossRef](#)]
56. John, M.; Shaiba, H. Apriori-Based Algorithm for Dubai Road Accident Analysis. *Procedia Comput. Sci.* **2019**, 163, 218–227. [[CrossRef](#)]

57. Delen, D.; Sharda, R.; Bessonov, M. Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks. *Accid. Anal. Prev.* **2006**, *38*, 434–444. [\[CrossRef\]](#)
58. Jamal, A.; Umer, W. Exploring the Injury Severity Risk Factors in Fatal Crashes with Neural Network. *IJERPH* **2020**, *17*, 7466. [\[CrossRef\]](#)
59. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: San Francisco, CA, USA, 2016; pp. 785–794.
60. Cunningham, S. *Causal Inference: The Mixtape*; Yale University Press: New Haven, Connecticut, 2021; ISBN 978-0-300-25588-1.
61. Kataoka, S.; Yasuda, M.; Furtlehner, C.; Tanaka, K. Traffic Data Reconstruction Based on Markov Random Field Modeling. *Inverse Probl.* **2014**, *30*, 025003. [\[CrossRef\]](#)
62. Ellis, B.; Wong, W.H. Learning Causal Bayesian Network Structures From Experimental Data. *J. Am. Stat. Assoc.* **2008**, *103*, 778–789. [\[CrossRef\]](#)
63. Needham, C.J.; Bradford, J.R.; Bulpitt, A.J.; Westhead, D.R. Inference in Bayesian Networks. *Nat. Biotechnol.* **2006**, *24*, 51–53. [\[CrossRef\]](#)
64. Aguilera, P.A.; Fernández, A.; Fernández, R.; Rumí, R.; Salmerón, A. Bayesian Networks in Environmental Modelling. *Environ. Model. Softw.* **2011**, *26*, 1376–1388. [\[CrossRef\]](#)
65. Fortin, S.P.; Johnston, S.S.; Schuemie, M.J. Applied Comparison of Large-scale Propensity Score Matching and Cardinality Matching for Causal Inference in Observational Research. *BMC Med. Res. Methodol.* **2021**, *21*, 109. [\[CrossRef\]](#)
66. Pearl, J. Comment: Understanding Simpson’s Paradox. *Am. Stat.* **2014**, *68*, 8–13. [\[CrossRef\]](#)
67. Friedman, J.; Hastie, T.; Tibshirani, R. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics* **2008**, *9*, 432–441. [\[CrossRef\]](#)
68. Aubert, G.; Kornprobst, P. Image Processing: Mathematics. In *Encyclopedia of Mathematical Physics*; Elsevier: Amsterdam, The Netherlands, 2006; pp. 1–10. ISBN 978-0-12-512666-3.
69. Liu, H.; Lafferty, J.; Wasserman, L. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *J. Mach. Learn. Res.* **2009**, *10*, 2295–2328.
70. Kuusimäki, M.; Sillanpää, M.J. MCPeSe: Monte Carlo Penalty Selection for Graphical Lasso. *Bioinformatics* **2021**, *37*, 726–727. [\[CrossRef\]](#) [\[PubMed\]](#)
71. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *Mach. Learn. Python* **2011**, *12*, 2825–2830.
72. Ankan, A.; Panda, A. Pgmpy: Probabilistic Graphical Models Using Python. In Proceedings of the 14th Python in Science Conference (Scipy 2015), Austin, Texas, 6–12 July 2015; pp. 6–11.
73. Ma, J.; Ding, Y.; Cheng, J.C.P.; Tan, Y.; Gan, V.J.L.; Zhang, J. Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective. *IEEE Access* **2019**, *7*, 148059–148072. [\[CrossRef\]](#)
74. Transport Canada Grade Crossings Inventory—Open Government Portal. Available online: <https://open.canada.ca/data/en/dataset/d0f54727-6c0b-4e5a-aa04-ea1463cf9f4c> (accessed on 6 November 2024).
75. Hong, W.-T.; Clifton, G.; Nelson, J.D. Railway Accident Causation Analysis: Current Approaches, Challenges and Potential Solutions. *Accid. Anal. Prev.* **2023**, *186*, 107049. [\[CrossRef\]](#) [\[PubMed\]](#)
76. Keerin, P.; Boongoen, T. Improved KNN Imputation for Missing Values in Gene Expression Data. *Comput. Mater. Contin.* **2022**, *70*, 4009–4025. [\[CrossRef\]](#)
77. Heydari, S.; Fu, L. Developing Safety Performance Functions for Railway Grade Crossings: A Case Study of Canada. In Proceedings of the 2015 Joint Rail Conference, American Society of Mechanical Engineers, San Jose, CA, USA, 23 March 2015; p. V001T06A017.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.