Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 3771640, 11 pages
https://doi.org/10.1155/2021/3771640

Research Article

Analyzing Accident Injury Severity via an Extreme Gradient Boosting (XGBoost) Model

Shubo Wu,[1] **Quan Yuan**,[2] **Zhongwei Yan**,[1] **and Qing Xu**[2]

[1]*Merchant Marine College, Shanghai Maritime University, Shanghai 201306, China*
[2]*State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing 100084, China*

Correspondence should be addressed to Quan Yuan; yuanq@tsinghua.edu.cn

Received 16 July 2021; Revised 20 August 2021; Accepted 11 September 2021; Published 27 September 2021

Academic Editor: Long Truong

Vehicle to vulnerable road user (VRU) crashes occupy a large proportion of traffic crashes in China, and crash injury severity analysis can support traffic managers to understand the implicit rules behind the crashes. Therefore, 554 VRUs-involved crashes are collected from January, 2017, to February, 2021, in a city in northern China, including 322 vehicle-pedestrian crashes and 232 vehicle-bicycle crashes. First, a descriptive statistical analysis is conducted to investigate the characteristics of VRUs-involved crashes. Second, the extreme gradient boosting (XGBoost) model is introduced to identify the importance of risk factors (i.e., time of day, day of week, rushing hour, crash position, weather, and crash involvements) of VRUs-involved crashes. The statistical analysis demonstrates that the risk factors are closely related to VRUs-involved crash injury severity. Moreover, the results of XGBoost reveal that time of day has the greatest impact on VRUs-involved crashes, and crash position shows the minimum importance among these risk factors.

1. Introduction

Crash injury severity analysis plays a crucial role in traffic crash analysis, which can assist traffic management [1–4]. Crash injury severity is defined as the degree of injury and property damage caused by a crash event. The crash injury severity analysis aims to explore the correlation between crash injury severity and various contributing factors, such as road-users-related factors, temporal-related factors, environmental conditions, and crash types. The universal rules support traffic managers to better understand the contributions of factors on crash injury severity and further reduce the crash severity and improve traffic safety by developing countermeasures [5–7].

Currently, the research approaches on crash injury severity can be divided into two categories, which are statistical models and machine learning-based models. The statistical models assume that the contributing factors affecting crash injury severity follow a particular distribution, which needs to be defined carefully for better capturing the relationship between crash injury severity and explanatory variables. The commonly used models contain multivariate Poisson regression model [8, 9], ordered probit model [10, 11], bivariate binary/ordered probit model [12, 13], random parameter probit model [14], etc. Wang et al. focused on mountainous expressways and proposed a partial proportional odds model to determine the determinants of truck-involved crash injury severity [15]. Xu et al. attempted to investigate pedestrian-involved crash injury severity by using geographically and temporally weighted regression model taking into account spatial-temporal correlation [16]. The statistical models could demonstrate and explain clearly the correlation between crash severity and related variables with the help of explainable and logical theoretical deductions. However, due to the nonlinear relationship between crash injury severity and contributing factors, these statistical models difficultly capture the inner and intrinsic correlations [17–19].

Machine-learning-based models have a powerful internal inferential capability, which makes them more flexible by learning without or little prior assumptions of related factors to describe the complex characteristics of crash events.

Previous researches employed logistic models (e.g., random parameter logit model, and mixed/ordered logit model) [20, 21], support vector machine (SVM) [22, 23], random forest (RF) [24, 25], Bayesian-related models [26, 27], etc., to explore the complex relationship between crash injury severity and contributing factors and further identify the risk factors on crash injury severity. For comprehensive accounting of the observed heterogeneity, Behnood et al. introduced a random parameter multinomial logit model for comparing the contribution of risk factors to crash injury severity under bicycle-vehicle crashes [28]. Liu et al. introduced an ordinal logistic regression model to examine the risk factors on pedestrian-motor vehicle collisions, taking into account the spatial-temporal correlation [29]. Li et al. introduced SVM model to investigate the potential correlation between external factors and crash injury severity, but the performance was suppressed due to multiclass classification problems [30]. Li et al. analyzed the key factors affecting electric bicycle-related crash injury severity with the help of random forest model [31].

Beyond that, Bayesian approaches, as a classical machine learning model, have been widely used in crash injury severity modelling, which were regarded as Bayesian-related models. For instance, Bayesian binomial logistic model [32, 33], Bayesian multivariate regression model [34, 35], Bayesian spatial model [36, 37], and Bayesian mixed logit model [38] have successfully demonstrated their applicability in crash injury severity-involved correlation issues. Yuan et al. divided crash severity into two categories (property damage only and injury/fatality) and integrated bivariate probit model and Bayesian approach to identify the contributing factors associated with crash injury severity [39]. Haq et al. developed binary logistic model with Bayesian inference approach to investigate the effects on truck-involved crashes, especially on occupant injury severity considering comprehensive factors [40]. Guo et al. proposed a novel random parameter, that is, multivariate Tobit model, to identify risk factors on crash severity under different crash types [41]. Zhang et al. utilized a Bayesian multinomial logit model with conditional autoregression prior to examining the hazardous factors that contributed to freeway crash injury severity [42].

In sum, previous researches focused on identifying risk factors towards traffic injury severity by various statistical models and machine learning-based models, and satisfying results were obtained. However, decision tree ensemble-based models, also a type of machine leaning model, including adaptive boosting (AdaBoost), gradient boost decision tree (GBDT), light gradient boosting machine (LightGBM), and extreme gradient boosting (XGBoost), are less utilized for crash injury severity analysis. Additionally, vulnerable road users (VRUs), as the vulnerable groups of the traffic participants, are prone to fatal injuries in crashes [43]. As the most common VRUs, the pedestrians and bicyclists have been paid much attention to, but decision tree ensembles-based models are rarely utilized to investigate the potential universal rules behind the VRUs-involved crashes. Hence, this paper attempts to describe the characteristics of VRUs-involved crashes and identify the contributing factors associated with crash injury severity. Based on this purpose, 554 VRU-vehicle crashes (contains 232 bicycle-vehicle crashes and 322 pedestrian-vehicle crashes) were collected. Moreover, XGBoost is introduced for crash injury severity modelling and ranking the importance of risk factors on crash injury severity. The contribution of this paper is twofold:

(1) Conduct a descriptive statistical analysis to investigate the characteristics of VRUs-involved crashes from the perspective of six risk factors (i.e., time of day, day of week, rush hour, crash position, weather, and crash involvements), and further transform into universal rules to support traffic management

(2) XGBoost is adopted to identify the risk factors contributing to VRUs-involved crash injury severity with the help of VRUs-involved crashes dataset from policy records, which further determine the real causes to enhance traffic safety

The rest of this paper is organized as follows. Section 2 introduces the data details and candidate variables analyzed in this paper. Section 3 describes the details of XGBoost model. Section 4 provides the experimental results, which consist of crash severity characteristics and identified risk factors. Section 5 briefly concludes the study.

2. Data Description

2.1. Data Source. For exploring the characteristic of VRUs-involved crashes, 554 crash samples were collected from police records on crashes, which have occurred in a city in northern China within about four years. The dataset contains various information, such as crash time, position, involvements, and injury severity, and six factors are extracted to explain the characteristics of VRUs-involved crashes. Vehicles and bicycles or pedestrians were involved in one crash, and bicyclists and pedestrians were defined as VRUs. The crashes dataset contains 323 vehicle-bicycle crashes and 322 vehicle-pedestrian crashes. The property-damage-only crashes are excluded because the vehicle-bicycle or vehicle-pedestrian crashes are prone to injury or death, which belong to injury or fatal accidents. Additionally, the crashes dataset consists of 385 injury crashes and 169 fatal crashes, which caused 517 injuries and 173 deaths.

2.2. Candidate Variables. Generally, if fatal or injured occupants are involved in a crash, it can be regarded as a severe accident. Considering that the dataset only contains fatal accidents and injury accidents, but without property-damage-only accidents, the crash injury severity is divided into two categories: injury accident (only injured occupant involved in the crash), which is coded as 0, and fatal accident (at least one fatality occupant involved in the crash), which is coded as 1. Figure 1 describes the extracted factors related to crashes from the dataset, which are time of day, day of week, rush hour, weather, crash position, and crash involvements. These six factors are extracted to investigate the

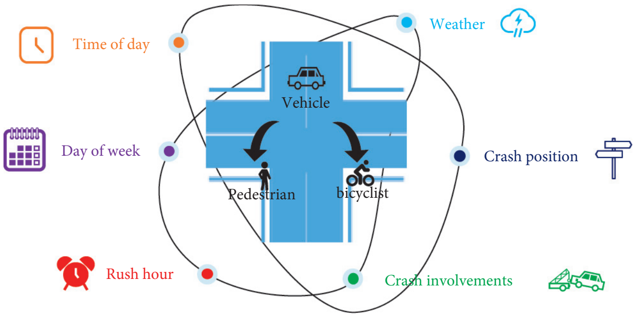


Figure 1: Selected factors related to crashes.

characteristics of VRUs-involved crashes, divided into two typical injury categories (see Table 1).

To some extent, time of day reflects the lighting conditions laterally, which is a crucial factor for traveling. Considering that, the crash position is complex, which mainly contains road section and intersection, but less sidewalk, roundabout. For better modelling, the crashes happened on sidewalk were regarded as road section. The weather information is collected from the related website (see http://www.tianqihoubao.com/lishi) based on the date and time of crashes [26]. Noting that this website provides the weather information only in two periods, that is daytime and night, it is not detailed enough to the specific hours. Additionally, due to the various types of weather, some of them have similar impact on traveling environment, for instance, sunny and cloudy, rainy and snowy. Therefore, the weather was divided into two categories: good and adverse.

3. Methodology

Extreme gradient boosting (XGBoost), as a typical decision tree ensemble-based model, was proposed by Chen in 2016 [44]. XGBoost is optimized from GBDT, which introduced second-order derivatives into optimization process. It outperforms with advantages of parallel learning, high flexibility, built-in cross-validation, etc. Previous studies have proved the successful use in traffic crash severity analysis and risk prediction [45, 46].

3.1. Objective Function. Given training data $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the objective function is described as

$$\mathrm{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad (1)$$

where $L(y_i, \hat{y}_i)$ is the training loss, which measures the fitting performance of the model to training data, and $\Omega(f_k)$ denotes the regularization term, which controls the model complexity for preventing overfitting. n is the volume of training dataset, and K denotes the number of trees. Generally, for classification problems, the logistic loss is adopted as loss function, and the expression is given in

$$L(y_i, \hat{y}_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i)\ln\left(1 + e^{-\hat{y}_i}\right), \quad (2)$$

where y_i is the truth value, and $\hat{y}_i$ is the predictive value. In XGBoost model, the predictive value is the sum of the score for each tree and the $\hat{y}_i$ can be defined as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}, \quad (3)$$

where $\mathcal{F}$ denotes the functional space and f_k is the function of k-th tree.

3.2. Additive Training. In the training process, it is intractable to learn all trees simultaneously. Instead, XGBoost introduces an additive strategy, which corrects what we have learned and adds one new tree at a time. The details of additive strategy are provided as

$$\begin{aligned} \hat{y}_i^{(0)} &= 0, \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i), \\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i), \\ &\ldots, \\ \hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), \end{aligned} \quad (4)$$

where $\hat{y}_i^{(t)}$ and $f_t(x_i)$ denote the predictive value and the added predictive function (i.e., new tree) at step t, respectively. For obtaining the best tree at each step, the objective function at step t is defined as

$$\begin{aligned} \mathrm{Obj}^{(t)} &= \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^t\right) + \sum_{i=1}^{t} \Omega(f_i), \\ &= \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + C, \end{aligned} \quad (5)$$

where $C = \sum_{i=1}^{t-1} \Omega(f_i)$ is a constant. For the training loss function, Taylor formula is introduced in loss function and its second-order expansion is expressed as

Table 1: Data list of main variables related to crashes.

Variable	Categories	Value	Fatal ($n = 169$)	Injury ($n = 385$)
Time of day	Day (7:00–19:00)	1	92 (54.4%)	263 (68.3%)
	Night (19:00–7:00)	0	77 (45.6%)	122 (31.7%)
Day of week	Weekday	1	118 (69.8%)	294 (76.4%)
	Weekend/holiday	0	51 (30.2%)	91 (23.6%)
Rush hour	Rush hour (7:00–9:00, 17:00–20:00)	1	49 (29.0%)	126 (32.7%)
	Off-peak hour	0	120 (71.0%)	259 (67.3%)
Weather	Good (sunny, cloudy)	1	138 (81.7%)	320 (83.1%)
	Adverse (rainy, snowy, etc.)	0	31 (18.3%)	65 (16.9%)
Crash position	Road section	1	110 (65.1%)	266 (69.1%)
	Intersection	0	59 (34.9%)	119 (30.9%)
Crash involvements	Vehicle-bicycle	1	53 (31.4%)	179 (46.5%)
	Vehicle-pedestrian	0	116 (68.6%)	206 (53.5%)

$$\mathrm{Obj}^{(t)} \approx \sum_{i=1}^{n}\left(L\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2\right) + \Omega(f_t) + C, \tag{6}$$

where g_i and h_i are the first-order and second-order partial derivative on loss function and can be expressed as

$$g_i = \partial_{\hat{y}^{(t-1)}} L\left(y_i, \hat{y}_i^{(t-1)}\right), h_i = \partial^2_{\hat{y}^{(t-1)}} L\left(y_i, \hat{y}_i^{(t-1)}\right). \tag{7}$$

Generally, the constant terms in the objective function are ignored, and the simplified objective function can be obtained as

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2\right] + \Omega(f_t). \tag{8}$$

3.3. Model Complexity. For defining the complexity of tree, we refine the tree $f(x)$ as (9). It contains the vector of weight on leaves ω and mapping function q, which maps each data sample to the corresponding leaf.

$$f_t(x) = \omega_{q(x)}, \omega \in \mathscr{R}^T, q\colon \mathscr{R}^d \longrightarrow \{1, 2, \ldots, T\}, \tag{9}$$

Here, T is the number of leaves. Then, the complexity can be defined as (10). γ is the coefficient of leaf, and λ denotes the coefficient of L2 regularization term.

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2. \tag{10}$$

We define $I_j = \{i | q(x_i) = j\}$, which denotes the set of data samples assigned to j-th leaf. Then, we introduce (10) into (9), and the objective function of t-th tree can be rewritten as

$$\begin{aligned}\mathrm{Obj}^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2\right] + \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T}\left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j{}^2\right] + \gamma \mathrm{T}.\end{aligned} \tag{11}$$

$G_j = \sum_{i \in I_j} g_i$ represents the sum of first-order partial derivative on jth leaf. $H_j = \sum_{i \in I_j} h_i$ denotes the sum of second-order partial derivative on jth leaf. We take partial derivation with respect to $\sum_{j=1}^{T} (G_j w_j + (1/2)(H_j + \lambda) w_j^2)$, and the best ω_j and objective value can be obtained as

$$\omega_j^* = -\frac{G_j}{H_j + \lambda}, \tag{12}$$

$$\mathrm{Obj}^* = -\frac{1}{2}\sum_{j=1}^{T}\left(\frac{G_j^2}{H_j + \lambda}\right) + \gamma \mathrm{T}. \tag{13}$$

To measure how good a tree is and find the best tree structure, a greedy algorithm is introduced. It starts from a single leaf and attempts to split each leaf into two leaves and then calculates the information gain in this split process (see equation (14)).

$$\mathrm{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{\left(G_L + G_R\right)^2}{H_L + H_R + \lambda}\right) - \gamma, \tag{14}$$

where G_L, H_L and G_R, H_R are the derivative values of left and right after split. Gain denotes the gain for loss function in the split. If $\mathrm{Gain} > 0$, the result of split will be considered.

4. Results

Based on the 544 crashes data, the time-related information, crash position, weather, and crash involvements are investigated. In the section, six risk factors are extracted to explore the characteristics of VRUs-involved crashes and further determine the risk factors contributing to crash injury severity.

4.1. Descriptive Statistical Analysis

4.1.1. Temporal Characteristics. Figure 2 illustrates the proportion of different accident types under three time-related factors. From the perspective of the time of day, the VRUs-involved crashes are probable to occur in the daytime, while the proportion of fatal crashes at night is relatively higher than those in the daytime, with values of 38.7% and

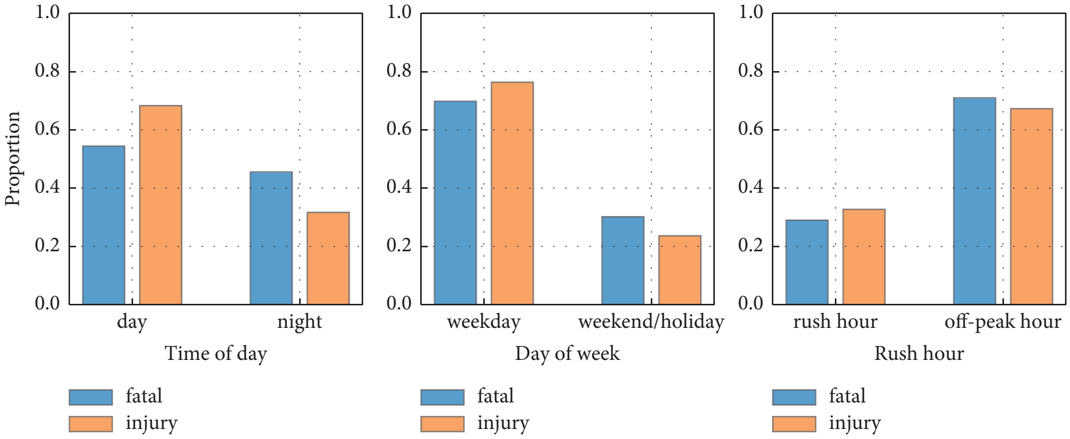


Figure 2: The proportion of different accident types under time-related factors.

25.9% (see Table 2). Maybe most people intend to travel during the daytime, which is prone to cause crashes. But at night, due to the terrible travel environment (i.e., poor light visible condition), the crashes are easy to cause deaths. Additionally, the proportion of crashes on weekdays is larger than that on weekends/holidays, but the fatality rate is the opposite and the values of weekday and weekend/holiday are 28.6% and 35.9%, respectively. The reason may be that people keep a relatively low safety alert when traveling on weekends/holidays than on weekdays. Moreover, the VRUs-involved crashes are prone to happen in off-peak hours than in rush hours due to the longer period of off-peak hours. Similarly, the fatality rate of off-peak hours is higher than those of rush hours (the values are 31.7% and 28.0%, respectively).

The variation tendency of VRUs-involved crashes counted by different days of the week is shown in Figure 3, and Table 3 provides the crash injury severity information under each day of the week. It illustrates that the largest number of crashes appears on Thursday, while Sunday occupies the least number. The main reason possible is that Thursday is the day near the weekend, the busiest day for most people as well as for the traffic, and yet Sunday is the final of a weekend when people are more likely to take a rest at home. However, the fatality rate is higher on Sunday (the value is 41.0%) because of the low safety awareness of people during leisure travel. Additionally, Monday takes up the minimum fatality rate with a value of 19.7%. The reason may be that Monday is the first day of weekday, and people will maintain a relatively high-security alert while commuting to work.

The statistical information of VRUs-involved crashes for injury severity is shown in Table 4, and Figure 4 illustrates the variation tendency of crashes counted by hours of the day. It indicates that crashes are prone to appear in rush hour (i.e., 7:00–9:00, 17:00–20:00), especially in the rush hours of the morning, with the highest peak existing in 7:00–8:00 (the total number of crashes is 45). It is because that this period is the time to go to work when the traffic is busy, likely to cause crashes. Moreover, most of the crashes happened at 6:00–23:00, which is the time for human activities, while few crashes occur within 23:00–6:00, which is the sleeping time. Overall, we found that the mortality at night is relatively higher than that in the daytime.

4.1.2. Spatial Characteristics. In the raw crash dataset, the crash position is complex, which makes the spatial characteristics hard to be described. Hence, we reorganized the complicated crashes environment into two types: road section and intersection. Table 5 provides the statistical information of crashes under two types of positions. There are 169 fatalities involved in crashes, including 110 on road sections and 59 at intersections. Moreover, the crashes that occurred on road sections take a higher proportion than intersections, and the mortality of crashes on road sections and at intersections are 0.293 and 0.331, respectively. Additionally, the proportion of fatal crashes that happened at intersections is higher than that of injury crashes, with values of 34.9% and 30.9%. Therefore, we can obtain that the crashes are more likely to happen on road sections, but the crashes happening at intersections have higher fatalities.

4.1.3. Weather Characteristics. There are various types of weather, so that it is hard to describe the weather characteristics associated with crash injury severity. Hence, the weather is divided into good (including sunny and cloudy) and adverse weather (including rainy, snowy, etc.). Table 6 shows the statistical information of injury severity in all weathers. Most VRUs-involved crashes happened in good weather, taking up 82.7%. That is because people prefer to travel in good weather compared to adverse weather. However, the mortality of crashes in adverse weather is higher than that in good weather, with values of 0.323 and 0.301, respectively. Similarly, the crashes that happened in adverse of fatal accidents account for a high proportion than injury accidents; the values are 18.3% and 16.9%, respectively. The results illustrate that VRUs-involved crash rarely happens in adverse weather. But once it happens, it may cause fatality.

4.1.4. Crash Involvements' Characteristics. In the crash dataset, the simultaneous participants in the crashes are vehicle and bicycle or vehicle and pedestrian; thus, the crash

Table 2: Statistical information of time-related factors for injury severity.

	Time of day		Day of week		Rush hour	
	Day	Night	Weekday	Weekend/Holiday	Rush hour	Off-peak hour
Fatal	92 (25.9%)	77 (38.7%)	118 (28.6%)	51 (35.9%)	49 (28.0%)	120 (31.7%)
Injury	263 (74.1%)	122 (61.3%)	294 (71.4%)	91 (64.1%)	126 (72.0%)	259 (68.3%)
Total	355	199	412	142	175	379

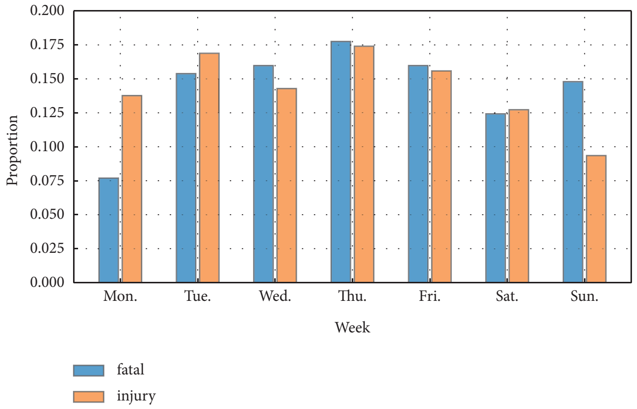


Figure 3: The variation tendency of crashes counted by different days of the week.

Table 3: Statistical information of different days of the week for injury severity.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Fatal	13 (19.7%)	26 (28.6%)	27 (32.9%)	30 (30.9%)	27 (31.0%)	21 (30.0%)	25 (41.0%)
Injury	53 (80.3%)	65 (71.4%)	55 (67.1%)	67 (69.1%)	60 (69.0%)	49 (70.0%)	36 (59.0%)
Total	66	91	82	97	87	70	61

Table 4: Statistical information of hours of the day for injury severity.

Hour	Fatal	Injury	Total
0	6 (42.9%)	8 (57.1%)	14
1	4 (66.7%)	2 (33.3%)	6
2	3 (33.3%)	6 (66.7%)	9
3	2 (33.3%)	4 (66.7%)	6
4	3 (42.9%)	4 (57.1%)	7
5	6 (60.0%)	4 (40.0%)	10
6	12 (35.3%)	22 (64.7%)	34
7	4 (8.9%)	41 (91.1%)	45
8	8 (21.6%)	29 (78.4%)	37
9	9 (24.3%)	28 (75.7%)	37
10	9 (29.0%)	22 (71.0%)	31
11	5 (21.7%)	18 (78.3%)	23
12	7 (36.8%)	12 (63.2%)	19
13	6 (22.2%)	21 (77.8%)	27
14	5 (23.8%)	16 (76.2%)	21
15	10 (35.7%)	18 (64.3%)	28
16	5 (19.2%)	21 (80.8%)	26
17	11 (45.8%)	13 (54.2%)	24
18	12 (33.3%)	24 (66.7%)	36
19	12 (48.0%)	13 (52.0%)	25
20	9 (42.9%)	12 (57.1%)	21
21	9 (30.0%)	21 (70.0%)	30
22	5 (20.0%)	20 (80.0%)	25
23	7 (53.8%)	6 (46.2%)	13

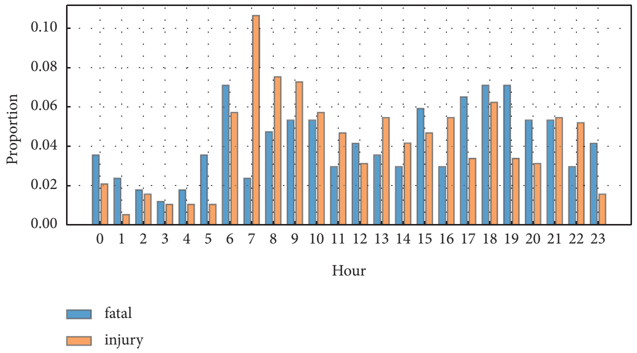


Figure 4: The variation tendency of crashes counted by the hours of the day (note that 0 in abscissa denotes 0:00–1:00, and others follow the rule).

Table 5: Statistical information of crash position for injury severity.

Position	Fatal accident	Injury accident	Total	Mortality
Road section	110 (65.1%)	266 (69.1%)	376 (67.9%)	0.293
Intersection	59 (34.9%)	119 (30.9%)	178 (32.1%)	0.331
Total	169	385	554	0.305

Table 6: Statistical information of weather for injury severity.

Weather	Fatal accident	Injury accident	Total	Mortality
Good	138 (81.7%)	320 (83.1%)	458 (82.7%)	0.301
Adverse	31 (18.3%)	65 (16.9%)	96 (17.3%)	0.323
Total	169	385	554	0.305

involvements are divided into vehicle-bicycle and vehicle-pedestrian. It can be seen that vehicle-pedestrian crashes take up a relatively high proportion not only in fatal accidents but also in injury accidents (see Table 7), and the proportion of fatal crashes is higher than that of injury crashes, with values of 68.6% and 53.5%, respectively. Additionally, the mortality of vehicle-pedestrian crashes is higher than vehicle-bicycle crashes, with values of 0.360 and 0.228. In sum, we can infer that vehicle-pedestrian crashes more easily result in death compared to vehicle-bicycle crashes, and most of these crashes may happen in intersections and crosswalks. It is probably because that the targets of bicycles are larger than pedestrians, more likely to attract the attention of vehicle drivers. And the reaction distance of cyclists is longer than pedestrians, which can reduce the injury severity in crashes.

4.2. Importance Identification for Risk Factors

4.2.1. Parameters Optimization. In this section, XGBoost is utilized to identify the contributing factors influencing crash injury severity. It is noted that the parameters of XGBoost are crucial for the model performance, and the grid search algorithm is introduced to obtain the optimal parameters. For binary classification problem in this study, the logistic loss and area under receiver operating characteristic curve are defined as objective loss function and evaluation metric, respectively. Moreover, four parameters, including number of estimators (n_estimators), learning rate, maximum depth, and coefficient of regularization (λ), are selected to optimize by grid search algorithm, and the candidate values are given in Table 8. The number of estimators refers to the number of iterations (i.e., the number of decision tree), learning rate controls the step size in weight updating, and maximum depth denotes the maximum depth of a tree. All these parameters contribute to preventing overfitting.

Based on the grid search results, we found that the optimal parameters model can be obtained, when the number of estimators is set as 10, learning rate as 0.05, maximum depth as 4, and λ as 3, and the scores of AUC and accuracy are 0.675 and 0.706, respectively. Figure 5 provides the AUC variation trends under different parameter settings. From Figure 5(a), the AUC scores show a up and down trend, and the maximum scores is 0.675 when number of estimators is set as 10, which indicates the optimal value of number of estimators is 10. The optimal values of learning rate, max depth, and λ are 0.05, 4, and 3, respectively. It is noted that the other three parameters are set as optimal values (i.e., learning rate is set as 0.05, max depth as 4, and λ as 3) in Figure 5(a), and other cases follow this rule.

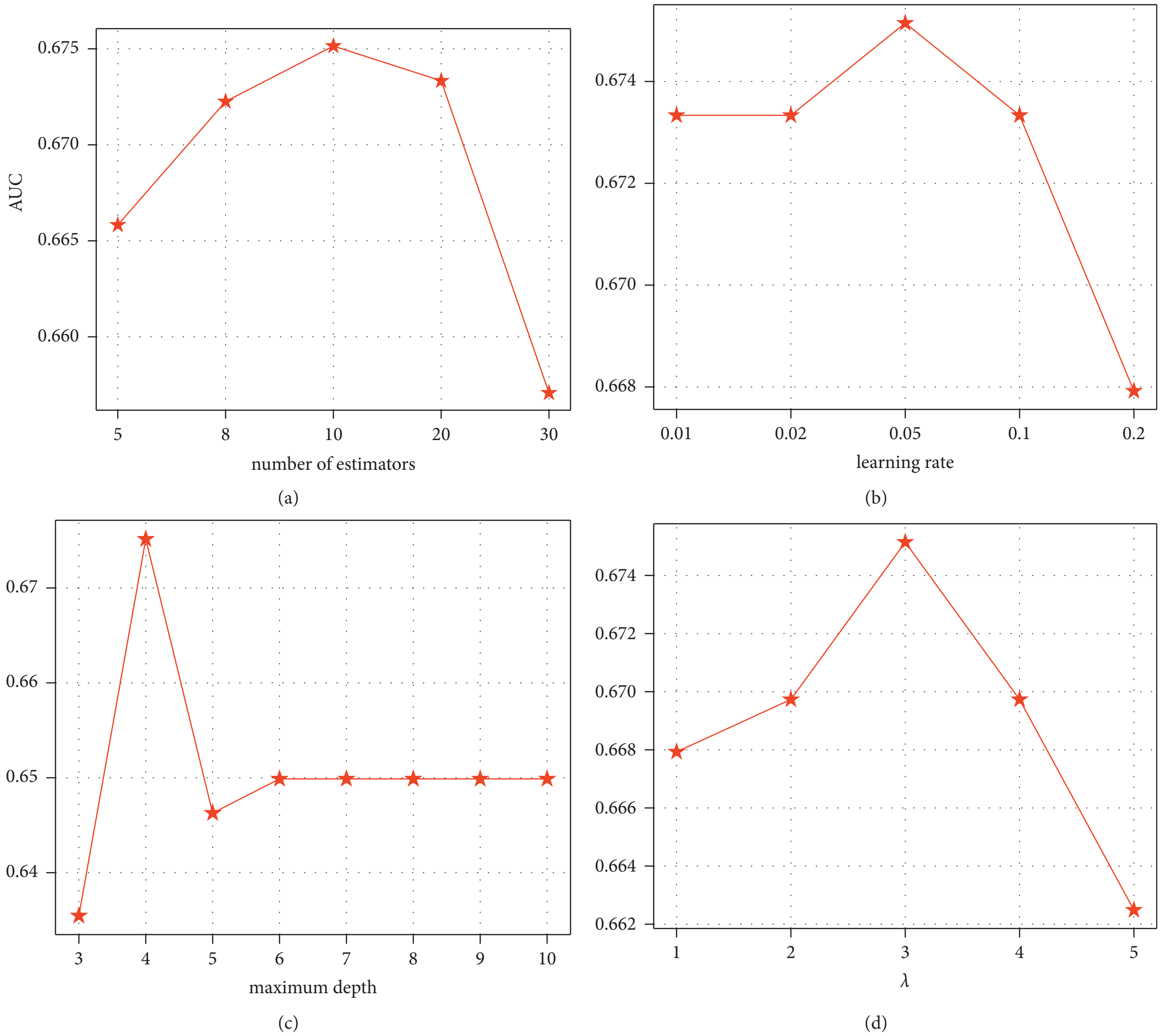
4.2.2. Risk Factors' Analysis. The XGBoost model with optimal parameters can be obtained after the parameters optimization procedure by using grid search algorithm. Then, the contributing factors were identified such that which factors show greater impact on VRUs-involved crashes injury severity. Figure 6 shows the importance of various risk factors from XGBoost model based on information gain, which is defined as the average gain for objective function optimization across all splits the feature (i.e., factor) is used in. The time of day occupies the most important role in VURs-involved crash injury severity, with the information gain score as 4.56. It reveals that time of day

Table 7: Statistical information of crash involvements for injury severity.

Crash involvements	Fatal accident	Injury accident	Total	Mortality
Vehicle-bicycle	53 (31.4%)	179 (46.5%)	232 (41.9%)	0.228
Vehicle-pedestrian	116 (68.6%)	206 (53.5%)	322 (58.1%)	0.360
Total	169	385	554	0.305

Table 8: The candidate values of four parameters for grid search algorithm.

Parameter	Number of estimators	Learning rate	Maximum depth	λ
Value	5, 8, 10, 20, 30	0.01, 0.02, 0.05, 0.1, 0.2	3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5



(a)

(b)

(c)

(d)

Figure 5: The AUC variation trends under different parameters. (a) Number of estimators. (b) Learning rate. (c) Maximum depth. (d) λ.

(day/night), which can use lighting conditions (good/adverse) instead, has a greater impact on VRUs-involved crashes, maybe because that the crashes are prone to happen in the daytime (or good lighting condition), while the crashes that occurred at night (or adverse conditions) are more likely to cause deaths.

Moreover, rush hour, day of week, and crash involvements show relatively similar importance, with the information gain

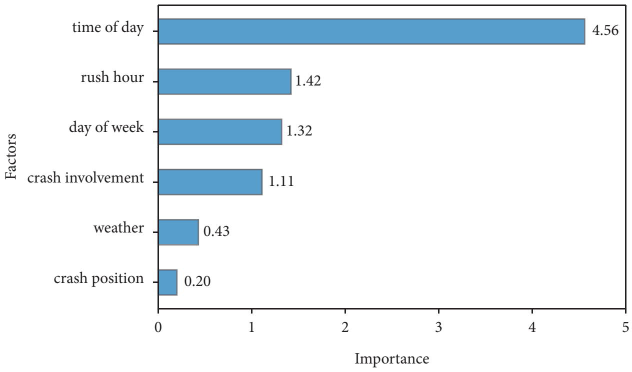


Figure 6: The importance of various risk factors.

scores as 1.42, 1.32, and 1.11, respectively. The reason may be that the categories of rush hour (i.e., rush hour and off-peak hour) show a minor difference of influence on VRUs-involved crashes, and day of week (i.e., weekday and weekend/holiday) and crash involvements (i.e., vehicle-bicycle and vehicle-pedestrian) are similar. The weather and crash position represent the least importance to VRUs-involved crash injury severity, and the information gain values are 0.43 and 0.20. Therefore, we infer that the people who travel in good or adverse weather show a similar impact on crash injury severity, which is consistent with the result of Section 4.1.3 (the mortalities are close in Table 6). This may be because people do not like to travel in adverse weather and they keep a relatively high safety awareness when traveling. Additionally, the VRUs-involved crashes that happened in different position (i.e., road section and intersection) show semblable result.

5. Conclusions

VRUs-involved crash injury severity analysis transforms the relationship behind the crashes into universal rules and further supports traffic management. This paper demonstrates a descriptive statistical analysis of the characteristics of VRUs-involved crashes based on 554 crashes data collected in a city of northern China and further utilizes XGBoost to identify the risk factors affecting crash injury severity. The important conclusions are summarized as follows. (1) The risk factors (i.e., time of day, day of week, rush hour, crash position, weather, and crash involvements) are closely related to VRUs-involved crash injury severity. More specifically, vehicle-bicycle and vehicle-pedestrian crashes are prone to involve fatalities at intersections on the weekend night in adverse weather. (2) The time of day plays a more important role in VRUs-involved crash injury severity compared with other factors, which reveals that VRUs-involved crashes that happened at night are prone to cause deaths. Additionally, the weather has little effect on VRUs-involved crash injury severity. (3) Compared to vehicle-bicycle crashes, vehicle-pedestrian crashes are prone to happen at intersections (especially at the crosswalk near the intersection), and these crashes readily cause deaths.

Although few factors were analyzed, the AUC and accuracy of XGBoost are 0.675 and 0.706, respectively, and the results still can be accepted and meet the current study. To obtain more accurate and detailed characteristics of VRU-vehicle crash injury severity, several research directions are proposed. (1) More risk factors (e.g., lighting condition, drivers' age, gender, crash pattern, and crash location related factors) can be considered to better explain the characteristics of VRU-vehicle crash injury severity and further identify the crucial risk factors. The characteristics of VRU-vehicle crash injury severity are not perfectly and accurately exploited due to the limitation of the risk factors. However, abundant risk factors may cause unfaithful characteristics to be described. To this topic, how to extract an appropriate number and precise risk factors is a crucial challenge. (2) Risk factors identified mechanism can be developed with high accuracy and robustness on crash injury severity analysis, such as random forest (RF) and nonparametric Bayesian approach, to better explain the characteristics and determine the real causes of crashes. The XGBoost model facilitates the investigation of crash injury severity issues, but the accuracy is limited due to the small sample size. Therefore, how to develop risk factors identified approach with a small sample size is a hot point. In addition, how to consider the spatial-temporal correlations in modelling process is a crucial challenge.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (nos. 52102397 and 52072214) and the National Key R&D Program of China (no. 2019YFB1600605).

References

[1] Y. Yuan, M. Yang, Y. Guo, S. Rasouli, Z. Gan, and Y. Ren, "Risk factors associated with truck-involved fatal crash severity: analyzing their impact for different groups of truck drivers," *Journal of Safety Research*, vol. 76, pp. 154–165, 2021.

[2] P. Liu and W. Fan, "Analysis of head-on crash injury severity using a partial proportional odds model," *Journal of Transportation Safety & Security*, vol. 13, no. 7, pp. 714–734, 2019.

[3] D. Li, P. Ranjitkar, Y. Zhao, H. Yi, and S. Rashidi, "Analyzing pedestrian crash injury severity under different weather conditions," *Traffic Injury Prevention*, vol. 18, no. 4, pp. 427–430, 2017.

[4] X. Chen, Z. Li, Y. Yang, L. Qi, and R. Ke, "High-resolution vehicle trajectory extraction and denoising from aerial videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3190–3202, 2020.

[5] C. Lee and M. Abdel-Aty, "Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida," *Accident Analysis & Prevention*, vol. 37, no. 4, pp. 775–786, 2005.

[6] M. G. Mohamed, N. Saunier, L. F. Miranda-Moreno, and S. V. Ukkusuri, "A clustering regression approach: a comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada," *Safety Science*, vol. 54, pp. 27–37, 2013.

[7] X. Chen, S. Wang, C. Shi, H. Wu, J. Zhao, and J. Fu, "Robust ship tracking via multi-view learning and sparse representation," *Journal of Navigation*, vol. 72, no. 1, pp. 176–192, 2019.

[8] J. Ma, K. M. Kockelman, and P. Damien, "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accident Analysis & Prevention*, vol. 40, no. 3, pp. 964–975, 2008.

[9] K. El-Basyouny, S. Barua, and M. T. Islam, "Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models," *Accident Analysis & Prevention*, vol. 73, pp. 91–99, 2014.

[10] A. J. Anarkooli, M. Hosseinpour, and A. Kardar, "Investigation of factors affecting the injury severity of single-vehicle rollover crashes: a random-effects generalized ordered probit model," *Accident Analysis & Prevention*, vol. 106, pp. 399–410, 2017.

[11] G. Fountas and P. C. Anastasopoulos, "Analysis of accident injury-severity outcomes: the zero-inflated hierarchical ordered probit model with correlated disturbances," *Analytic Methods in Accident Research*, vol. 20, pp. 30–45, 2018.

[12] Y. Guo, J. Zhou, Y. Wu, and J. Chen, "Evaluation of factors affecting e-bike involved crash and e-bike license plate use in China using a bivariate probit model," *Journal of Advanced Transportation*, vol. 2017, Article ID 2142659, 12 pages, 2017.

[13] F. Chen, M. Song, and X. Ma, "Investigation on the injury severity of drivers in rear-end collisions between cars using a random parameters bivariate ordered probit model," *International Journal of Environmental Research and Public Health*, vol. 16, no. 14, p. 2632, 2019.

[14] S. Darban Khales, M. M. Kunt, and B. Dimitrijevic, "Analysis of the impacts of risk factors on teenage and older driver injury severity using random-parameter ordered probit," *Canadian Journal of Civil Engineering*, vol. 47, no. 11, pp. 1249–1257, 2020.

[15] Y. Wang and C. G. Prato, "Determinants of injury severity for truck crashes on mountain expressways in China: a case-study with a partial proportional odds model," *Safety Science*, vol. 117, no. April, pp. 100–107, 2019.

[16] X. Xu, X. Luo, C. Ma, and D. Xiao, "Spatial-temporal analysis of pedestrian injury severity with geographically and temporally weighted regression model in Hong Kong," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 69, pp. 286–300, 2020.

[17] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: methodological frontier and future directions," *Analytic Methods in Accident Research*, vol. 1, pp. 1–22, 2014.

[18] X. Chen, J. Lu, J. Zhao, Z. Qu, Y. Yang, and J. Xian, "Traffic flow prediction at varied time scales via ensemble empirical mode decomposition and artificial neural network," *Sustainability*, vol. 12, no. 9, p. 3678, 2020.

[19] X. Chen, L. Qi, Y. Yang et al., "Video-based detection infrastructure enhancement for automated ship recognition and behavior analysis," *Journal of Advanced Transportation*, vol. 2020, Article ID 7194342, 12 pages, 2020.

[20] Q. Yuan, H. Yang, J. Huang, S. Kou, Y. Li, and A. Theofilatos, "What factors impact injury severity of vehicle to electric bike crashes in China?" *Advances in Mechanical Engineering*, vol. 9, no. 8, 2017.

[21] K. Haleem and A. Gan, "Effect of driver's age and side of impact on crash severity along urban freeways: a mixed logit approach," *Journal of Safety Research*, vol. 46, pp. 67–76, 2013.

[22] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, vol. 108, no. August, pp. 27–36, 2017.

[23] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, "Crash injury severity analysis using a two-layer Stacking framework," *Accident Analysis & Prevention*, vol. 122, no. May, pp. 226–238, 2019.

[24] X. Zhou, P. Lu, Z. Zheng, D. Tolliver, and A. Keramati, "Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree," *Reliability Engineering & System Safety*, vol. 200, no. July, Article ID 106931, 2020.

[25] L. Li, C. G. Prato, and Y. Wang, "Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: a sequential approach of multivariate imputation by chained equations and random forest classifier," *Accident Analysis & Prevention*, vol. 146, no. January, Article ID 105744, 2020.

[26] Q. Yuan, Z. Liang, W. Hao, and Y. Li, "Investigating severity and main weather factors in tricycle crashes in Beijing," in *Proceedings of the 18th COTA International Conference of Transportation*, pp. 1710–1719, Beijing, China, July 2018.

[27] Y. Guo, Z. Li, Y. Wu, and C. Xu, "Exploring unobserved heterogeneity in bicyclists' red-light running behaviors at different crossing facilities," *Accident Analysis & Prevention*, vol. 115, no. November, pp. 118–127, 2018.

[28] A. Behnood and F. Mannering, "Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances," *Analytic Methods in Accident Research*, vol. 16, pp. 35–47, 2017.

[29] J. Liu, A. Hainen, X. Li, Q. Nie, and S. Nambisan, "Pedestrian injury severity in motor vehicle crashes: an integrated spatio-temporal modeling approach," *Accident Analysis & Prevention*, vol. 132, no. August, p. 105272, 2019.

[30] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Analysis & Prevention*, vol. 45, pp. 478–486, 2012.

[31] Y. S. Li, X. Zhang, W. J. Wang, and X. F. Ju, "Factors affecting electric bicycle rider injury in accident based on random forest model," *Journal of Transportation Systems Engineering and Information Technology*, vol. 21, no. 1, pp. 196–200, 2021.

[32] Y. Guo, Z. Li, Y. Wu, and C. Xu, "Evaluating factors affecting electric bike users' registration of license plate in China using Bayesian approach," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 59, pp. 212–221, 2018.

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

[34] S. Heydari, L. Fu, L. Joseph, and L. F. Miranda-Moreno, "Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings," *Analytic Methods in Accident Research*, vol. 12, no. September, pp. 18–34, 2016.

[35] Q. Zeng, W. Gu, X. Zhang, H. Wen, J. Lee, and W. Hao, "Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors," *Accident Analysis & Prevention*, vol. 127, no. February, pp. 87–95, 2019.

[36] X. Xu, S. Xie, S. C. Wong, P. Xu, H. Huang, and X. Pei, “Severity of pedestrian injuries due to traffic crashes at signalized intersections in Hong Kong: a Bayesian spatial logit model,” *Journal of Advanced Transportation*, vol. 50, no. 8, pp. 2015–2028, 2016.

[37] Q. Zeng, H. Wen, H. Huang, and M. Abdel-Aty, “A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments,” *Accident Analysis & Prevention*, vol. 100, pp. 37–43, 2017.

[38] Z. H. Khattak and M. D. Fontaine, “A Bayesian modeling framework for crash severity effects of active traffic management systems,” *Accident Analysis & Prevention*, vol. 145, Article ID 105544, 2020.

[39] Q. Yuan, X. Xu, M. Xu, J. Zhao, and Y. Li, “The role of striking and struck vehicles in side crashes between vehicles: bayesian bivariate probit analysis in China,” *Accident Analysis & Prevention*, vol. 134, no. October, Article ID 105324, 2020.

[40] M. T. Haq, M. Zlatkovic, and K. Ksaibati, “Investigating occupant injury severity of truck-involved crashes based on vehicle types on a mountainous freeway: a hierarchical Bayesian random intercept approach,” *Accident Analysis & Prevention*, vol. 144, Article ID 105654, 2020.

[41] Y. Guo, Z. Li, P. Liu, and Y. Wu, “Modeling correlation and heterogeneity in crash rates by collision types using full bayesian random parameters multivariate Tobit model,” *Accident Analysis & Prevention*, vol. 128, pp. 164–174, 2019.

[42] X. Zhang, H. Wen, T. Yamamoto, and Q. Zeng, “Investigating hazardous factors affecting freeway crash injury severity incorporating real-time weather data: using a Bayesian multinomial logit model with conditional autoregressive priors,” *Journal of Safety Research*, vol. 76, pp. 248–255, 2021.

[43] Q. Yuan and H. Chen, “Factor comparison of passenger-vehicle to vulnerable road user crashes in Beijing, China,” *International Journal of Crashworthiness*, vol. 22, no. 3, pp. 260–270, 2017.

[44] T. Chen and C. Guestrin, “XGBoost,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17, pp. 785–794, 2016.

[45] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, “A feature learning approach based on XGBoost for driving assessment and risk prediction,” *Accident Analysis & Prevention*, vol. 129, no. March, pp. 170–179, 2019.

[46] M. Guo, Z. Yuan, B. Janson, Y. Peng, Y. Yang, and W. Wang, “Older pedestrian traffic crashes severity analysis based on an emerging machine learning xgboost,” *Sustainability*, vol. 13, no. 2, pp. 926–26, 2021.