Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 8881545, 12 pages
https://doi.org/10.1155/2020/8881545

Research Article

Analysis of Factors Affecting the Severity of Automated Vehicle Crashes Using XGBoost Model Combining POI Data

Hengrui Chen, Hong Chen, Zhizhen Liu, Xiaoke Sun, and Ruiyu Zhou

College of Transportation Engineering, Chang'an University, Xi'an 710000, China

Correspondence should be addressed to Hong Chen; glch@chd.edu.cn

Received 6 July 2020; Revised 4 November 2020; Accepted 6 November 2020; Published 19 November 2020

Academic Editor: Kun Wang

The research and development of autonomous vehicle (AV) technology have been gaining ground globally. However, a few studies have performed an in-depth exploration of the contributing factors of crashes involving AVs. This study aims to predict the severity of crashes involving AVs and analyze the effects of the different factors on crash severity. Crash data were obtained from the AV-related crash reports presented to the California Department of Motor Vehicles in 2019 and included 75 uninjured and 18 injured accident cases. The points-of-interest (POI) data were collected from Google Map Application Programming Interface (API). Descriptive statistics analysis was applied to examine the features of crashes involving AVs in terms of collision type, crash severity, vehicle movement preceding the collision, and degree of vehicle damage. To compare the classification performance of different classifiers, we use two different classification models: eXtreme Gradient Boosting (XGBoost) and Classification and Regression Tree (CART). The result shows that the XGBoost model performs better in identifying the injured crashes involving AVs. Compared with the original XGBoost model, the recall and G-mean of the XGBoost model combining POI data improved by 100% and 11.1%, respectively. The main features that contribute to the severity of crashes include weather, degree of vehicle damage, accident location, and collision type. The results indicate that crash severity significantly increases if the AVs collided at an intersection under extreme weather conditions (e.g., fog and snow). Moreover, an accident resulting in injuries also had a higher probability of occurring in areas where land-use patterns are highly diverse. The knowledge gained from this research could ultimately contribute to assessing and improving the safety performance of the current AVs.

1. Introduction

The autonomous vehicle (AV) technique has the potential to reduce crashes significantly. More than 30 thousand people die from traffic accidents every year in the US, with 2.2 million accidents resulting in injuries [1]. Traffic crashes cost the economy $277 billion a year, twice as much as congestion [2]. Over 40% of fatal accidents involved alcohol, distraction, drug addiction, and fatigue. Drivers' error is the leading cause of 90% of the accidents. Even crashes caused mainly by vehicles, roadways, and environmental conditions are accompanied by some human factors (e.g., inattention, distraction, or speeding). With the popularization of AV technology, drivers' errors may disappear, indicating the possibility of reducing the fatal accident rate by at least 40% [3]. Therefore, clarifying how different influencing factors affect the severity of AV crashes is of considerable significance in comprehensively improving the safety of AVs.

Safety is the primary factor driving the development of AV technology. Previous literature has concentrated on the various advanced driver assistance systems (e.g., forward collision warning, vehicle collision warning system, and lane departure warning systems), traffic signal control (e.g., actuated signal control and cooperative adaptive cruise control), and accident responsibility [4–6]. However, designing a system that can operate safely in any unexpected circumstances remains a daunting challenge. The existing AV technology still has certain limitations in terms of technical indicators and driving environment requirements:

(1) Robustness of environmental perception and visual recognition needs to be improved [7].

(2) Multistrategy decision-making algorithms in AV technology lack measures against abnormal behavior [8].

(3) Although AV technology can assist the driver in completing the driving task to a certain extent, it may also affect the driver. Hence, further studies on driver behavior are necessary [9].

Road safety is a complicated issue and is influenced by a series of risk factors, such as driver, environment, and vehicle factors. AVs will play an essential role in future transportation safety. Given the uncertainty in the safety of AVs, this study applies the publicly available Traffic Collision Reports of AV crashes in California to predict the severity of crashes involving AVs and analyze the effects of the different factors on crash severity. The knowledge gained from this research could contribute to the assessment and improvement of the safety performance of current AVs.

The rest of the paper is organized as follows: Section 2 is the literature review related to our study; Section 3 describes the dataset and correlated variables; Section 4 introduces the main content of the proposed methodology in detail; Section 5 discusses the model results; and, finally, the conclusion and limitations are shown in Section 6.

2. Literature Review

Most previous studies on AV technology safety rely mainly on evaluating drivers' performance and behavior in a simulated environment and developing the performance of autonomous driving systems in a closed field environment. Some research focuses on the driving trajectory of AVs to avoid potential collisions. Hegedus et al. proposed a local trajectory optimization algorithm based on nonlinear optimization, which can provide a dynamic, feasible, comfortable, and customizable trajectory for highly automated vehicles [10]. Omidvar [11] developed an algorithm for trajectory optimization of AVs in the signalized intersections at a closed-course. This algorithm optimizes signal control and provides the best trajectory for AVs. As for the simulation studies regarding AV safety, many researchers use driving simulators as experimental tools. They focused on the driver's physiological and psychological responses in an autonomous driving environment. Winter et al. [12] found that drivers can divert their attention to secondary tasks in a highly automated driving environment without affecting the driving performance of the vehicle.

The California Department of Motor Vehicles (DMV) release massive crash data involving AVs, and many machine learning models (e.g., logistic regression models [13], Classification and Regression Tree (CART) [14], neural network [15], and random forest [16, 17]) have been utilized to identify the factors that contribute to the severity of crashes involving AVs. To investigate the factors contributing to the severity of AV involved crashes, Wang [18] developed CART models by harnessing California's Report from 2014 to 2018. The highway is recognized as the location where severe injuries are likely to happen. Crash severity significantly increases if the AV is responsible for the crash. Xu et al. [19] conducted a study based on the binary logistic regression model using California data. The driving mode of AVs, collision location, roadside parking, rear-end collision, and one-way road are the main factors that contributed to the severity level of AVs involved crashes. Boggs et al. [20] investigated factors contributing to AV involved crashes using the hierarchical Bayesian heterogeneity-based approach. According to this study, clear weather could reduce the likelihood of injury crashes involving AVs.

Agarwal et al. [21] proposed a relatively novel technology in 2016: eXtreme Gradient Boosting (XGBoost). It has high precision and fast processing speed as well as lower cost and complexity. Two studies [22, 23] have shown that XGBoost is more accurate than other machine learning techniques (logistic regression, SVM, deep neural network, etc.) in predicting the likelihood of an accident. Meng et al. [24] use XGBoost to combine multiple data sources to predict the occurrence and duration of accidents, including geometric road design, historical accident data, and weather data. Fan et al. [25] also used artificial neural networks to integrate multiple XGBoost models to predict the duration of the accident. Finally, as an integrated algorithm, it is not affected by the multicollinearity of data.

However, the lack of reliable data and insufficient data sources have limited studies on accident analysis, especially for the accident mechanisms of AVs. Fortunately, reliable points-of-interest (POI) data can be collected from anywhere globally, providing a broad space for detailed accident detection [26]. Although these POI data may not be the typical factors used in traditional traffic accident analysis, they are specific data on land-use factors with precise location information [27]. Additionally, they are expected to be highly correlated with traffic accidents in the macro- and microaspects. The current study employs POI data to describe the built environment to replace traditional land-use data. It specifies the city's infrastructure distribution and has much better statistical granularity [28]. Simpson's diversity index is selected as the POI diversity evaluation index to quantify the diversity of land-use patterns in the buffer zone.

The primary purpose of this study is to use the XGBoost model incorporating POI data to predict the severity of crashes involving AVs and investigate the effects of the different factors on crash severity. This study employed 94 crash reports involving AV in California received in 2019. Synthetic Minority Oversampling Technique (SMOTE) was applied to address the imbalanced data. Ultimately, the knowledge gained from this study could contribute to the assessment and improvement of the safety performance of the current AVs.

3. Data Preparation

3.1. Data Sources. With the implementation of California Senate Bill 1298, the Department of Motor Vehicles (DMV) demanded that crash reports involving AV be provided within ten business days of the crash occurrence [19]. This study employed 94 crash reports involving AVs in California received in 2019. Information was manually extracted from crash reports submitted by various manufacturers for a comprehensive understanding of AV-related information (e.g., type of collision, manufacturer's name, crash severity, vehicle information, and weather). A vast number of reports did not count vehicle speed before the crash. Thus, it was not adopted for model development.

This paper analyzes the diversity of land-use patterns based on POI data because of the lack of traffic volume and land-use data. The integration of traffic accident data and POI data can enable a more accurate identification of land-use intensity on traffic safety [29]. The POI data were obtained from Google Map Application Programming Interface (API). The buffer analysis and cross summary toolbox in ArcGIS was used to match the POI data according to the latitude and longitude of the accident site. Different types of POI may have different effects on traffic status, but some types have similar functions. Therefore, the POIs are divided into four major categories, as shown in Table 1.

Simpson's diversity index was selected as the POI diversity evaluation index to quantify land-use development intensity in the buffer zone, as shown in the following equation:

$$\text{POI\_}D = 1 - \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2, \tag{1}$$

where N_i and N represent the number of POIs of a specific type and the total amount of POIs, respectively. The larger the value of D, the higher the diversity of POIs.

3.2. Statistical Analysis. Figure 1 illustrates the distribution of collision type. Rear-end collisions are the primary type of crashes, accounting for 64%. The road environment's perceptual system might cause the AVs emergency brakes, although the report does not provide clear instructions. According to the statistical data, most crashes are conventional vehicles hitting the rear of AVs [30]. Furthermore, rear-end collisions usually occur at intersections because the trajectory of intersections is more complicated than that of the road segment [31]. The other common collision types are siding swipe (15%), broadside (12%), and head-on (9%). Crashes involving AVs are caused primarily by the complicated interaction between AVs and conventional vehicles [32]. Therefore, specific attention should be given to the adverse effects of mixed traffic flow composed of AVs and conventional vehicles on the autonomous driving system during the low penetration rate of AVs [33]. AVs–pedestrian collisions or hit objects are not reported, which indicates the benefit of road environment perception and motion control systems for AVs.

Figure 2 describes the proportion of crashes for each company. Cruise has the most crash reports in 2019, accounting for 58%, followed by Waymo (25%). Cruise is a representative company because it has launched many test vehicles in congested San Francisco. By contrast, Waymo's test site is in Arizona. The traffic environment in San Francisco is much more complicated than Arizona's, with its lots of intersections, steep hills roads, and aggressive driving. Therefore, the probability of an emergency occurring is higher. Moreover, because of the insufficient sample size, which company's test vehicles are more prone to accidents cannot be proved.

Figure 3 indicates the vehicle movement preceding the collision. The most common states of AVs and conventional vehicles before collision are stopped and proceeding straight, respectively. Unexpected situations in front (e.g., a pedestrian crossing the road) may cause the AVs to emergency brake, while a conventional vehicle behind cannot evade in time, resulting in a rear-end collision. Consistent with previous studies [30], most crashes are conventional vehicles hitting the rear of AVs. The second-largest percentage of AVs and conventional vehicle movements are proceeding straight and changing lanes. Taking effective emergency avoidance measures immediately when a conventional vehicle makes unsafe lane changes is challenging for the automatic driving system. Ultimately, researchers have pointed out that AV technology still needs to overcome many barriers to respond accurately in complex traffic environments.

Figure 4 shows that most collisions involving AVs are significantly less severe than regular accidents, especially for severe injuries and fatal collisions. Specifically, 81% of crashes are property-damage-only (PDO) crashes, and 19% have minor injuries. Similarly, 72% of AVs are only minor damage, thereby suggesting that collisions occurred at low-speed conditions. Speed and speed variations have been frequently regarded as critical factors closely connected with the injured crash [34, 35]. AVs would not fall prey to personal faults. Drivers' error is the leading cause of 90% of accidents. AV technology reduces crash severity by overcoming driver error (e.g., speeding, aggressive driving, inexperience, slow reaction times, inattention, and various other driver shortcomings).

The specific location of the collision can be collected from the crash report, while the approximate latitude and longitude of each accident can be obtained through OpenStreetMap. We use ArcGIS software to draw the heat map of AV crashes (shown in Figure 5). It provides the visualization and distribution of accident locations among counties. The accidents mainly occurred in San Francisco and Palo Alto because they were the main test sites for AVs. In the future, the use of AVs will be extended to any corner of any city in the United States, making it necessary to analyze further the effects of land-use intensity around the accident site on the crash.

3.3. Variable Collinearity Analysis. Multicollinearity refers to the situation in which several explanatory variables in a regression model are highly linearly related. As an integrated algorithm, XGBoost is not influenced by the multicollinearity of the data; however, introducing excessive variables may cause overfitting of the model. Moreover, the interpretability of the model may be significantly affected, thereby increasing the complexity of the model. Variance inflation factor (VIF) was calculated using SPSS 26.0, which is a common indicator of multicollinearity [36–38].

Generally, independent variables with VIF values higher than 10 indicate severe collinearity between two variables, which suggests that one of them should be eliminated [39]. Finally, nine categorical variables were determined. The descriptive statistics of the variables are shown in Table 2.

4. Methodology

In the current study, two classification models were used to train the data. This section provides the relevant concepts of these models. We applied the Scikit-learn (sklearn) library in Python 3.6. Overall, the proposed models consist of these steps:

Step 1. We employed the SMOTE algorithm to deal with imbalanced datasets.

Step 2. We randomly selected 70% of the data as the training set, and the remaining 30% were employed to test the model.

Step 3. We inputted the divided training set into the XGBoost and CART models, respectively, and used the grid search to determine the best combination of parameters to prevent the model from overfitting. The cross-validation method was used to measure the stability of the model.

Step 4. By comparing the performance of two models, choosing the well-performing model to predict the severity of crashes involving AVs and analyze the effects of the different factors on crash severity.

4.1. XGBoost Model.

The core of XGBoost is an integrated algorithm based on gradient boosted decision trees. It utilizes a series of decision trees, where every tree studies from the prior tree and influences the following tree to promote model performance [40]. In this section, we explain the formulas and evaluation indicators behind XGBoost. Interested readers can refer to the study published by Chen [21] for further detailed information. Chen and Guestrin made some improvements based on the Gradient Boosting [41] and presented the XGBoost in 2016. One of the unprecedented progress is the regularization of the loss function. The regularized objective L_k for the k^{th} iteration can be expressed, as shown in the following equation:

$$L_k = \sum_{i=1}^{n} l\left(y^{(i)}, \hat{y}_k^{(i)} \right) + \sum_{j=1}^{k} \Omega\left(f_j\right), \tag{2}$$

where n is the number of samples, $\widehat{y_k}^{(i)}$ is the prediction value of the sample i at iteration k, and l is the original loss function. Ω represents the regularization term, as shown in the following equation:

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_J^2. \tag{3}$$

Here, T is the number of leaf nodes and γ and λ are two constants employed to constrain the degree of regularization.

Another development of XGBoost is the application of an additive learning approach [42] that combines the most reliable tree model $f_k(x^i)$ into the current classification model to provide the m^{th} iteration prediction result [43]. Therefore, equation (3) can be expressed further as follows:

$$L_k = \sum_{i=1}^{n} l\left(y^{(i)}, \hat{y}_{k-1}^{(i)} + f_k\left(x^{(i)}\right) \right) + \Omega\left(f_k\right) + \sum_{j=1}^{k-1} \Omega\left(f_j\right). \tag{4}$$

Additionally, XGBoost utilizes the second-order Taylor expansion to the objective function and equation (4) can be expressed further as the following equation:

$$L_k = \sum_{i=1}^{n} \left[l\left(y^{(i)}, \hat{y}_{k-1}^{(i)} + g_i * f_k\left(x^{(i)}\right) + \frac{1}{2} h_i * f_k\left(x^{(i)}\right) \right] + \Omega\left(f_k\right) + C. \tag{5}$$

Here, $g_i = \eth_{\widehat{y}_{k-1}} l(y^i, \widehat{y}_{k-1})$ and $h_i = \delta^2_{\widehat{y}_{k-1}} l(y^i, \widehat{y}_{k-1})$ are the first and second derivatives of the loss function, respectively, and C represents the constant.

Finally, as an integrated algorithm, XGBoost is not affected by the multicollinearity of the data. This advantage makes XGBoost possibly gain more reliable results even if the variables have a strong linear correlation.

4.2. Classification and Regression Tree.

Classification and Regression Tree (CART) is a nonparametric decision tree learning method [15]. It can summarize decision rules from a series of data with features and labels and present them in a tree structure to solve classification and regression problems. The CART method usually consists of two main steps: tree growing and pruning. The tree extends from the root node, which includes all the data in the dataset. Divide the root node into two child nodes through a splitter (independent variable) to improve the purity of the two child nodes. The Gini index is used as the splitting criterion in the current study. If the root node m is divided into two child nodes (child nodes n_1 and n_2) by the variable θ, the Gini coefficient of any child node is calculated as follows:

$$H(n(\theta)) = 1 - \sum_k p\left(\frac{k}{n}\right)^2, \quad n \in (n_1, n_2). \tag{6}$$

Here, $H(n(\theta))$ represents the Gini index of the child node n , and $p(k/n)$ is the proportion of class k records in node n. The impurity at node m is calculated as follows:

$$G(\theta) = \frac{o_1}{N_m} H\left(n_1(\theta)\right) + \frac{o_2}{N_m} H\left(n_2(\theta)\right). \tag{7}$$

Here, N_m is the total number of observations at node mm and o_1 and o_2 are numbers of observations in child nodes n_1 and n_2. The method tries to divide the root node m by selecting the variable θ^*:

$$\theta^* = \arg\min_\theta G(\theta). \tag{8}$$

Table 1: Categories of POIs.

Categories	Types of POIs
Commercial buildings	Foods, hotels, shopping areas, living services, beauty services, leisure and entertainments, exercise and fitness
Residential buildings	Apartments and houses, dormitories
Office building	Governmental agencies, businesses
Transportation facilities	Traffic facilities

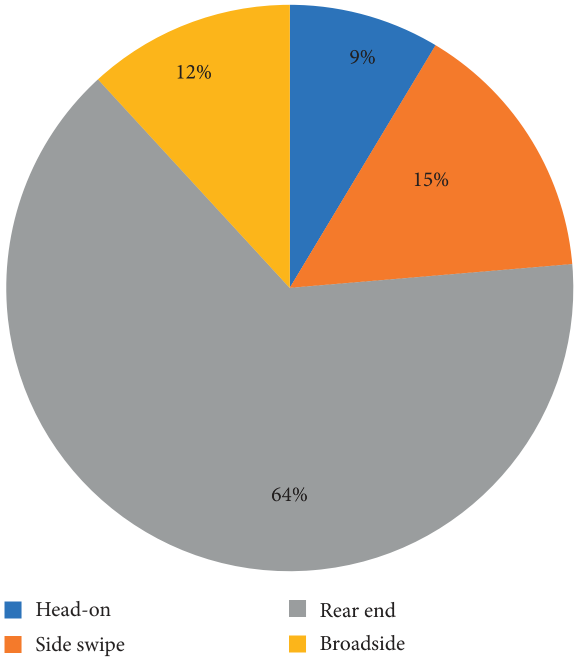


Figure 1: The distribution of collision type.

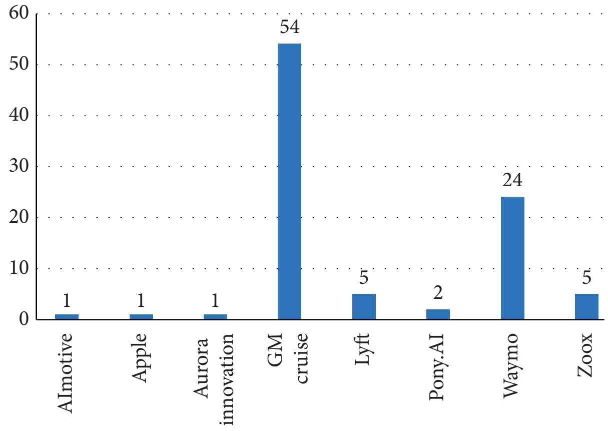


Figure 2: AVs crashes of different companies.

When CART detects that no further gains can be made by further growing the tree deeply or when specific predetermined criteria that are stopping rules are met, the segmentation will stop. Given the defined branches and nodes of the tree, each corresponding variable falls into a terminal node.

4.3. Model Evaluation. The confusion matrix is a multidimension measurement index system of binary classification problems that has been used widely in evaluating model performance (see Table 3) [44].

The overall accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}. \tag{9}$$

However, this index could not be suitable for unbalanced data. Because the number of injury accidents in the current study is significantly less than the uninjured accidents, even if all minority instances are misclassified, the overall

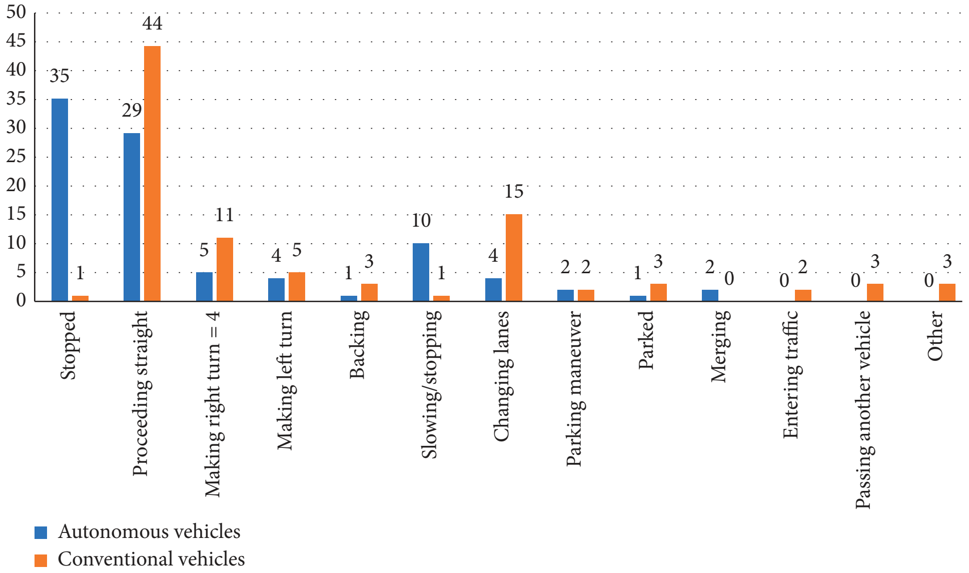
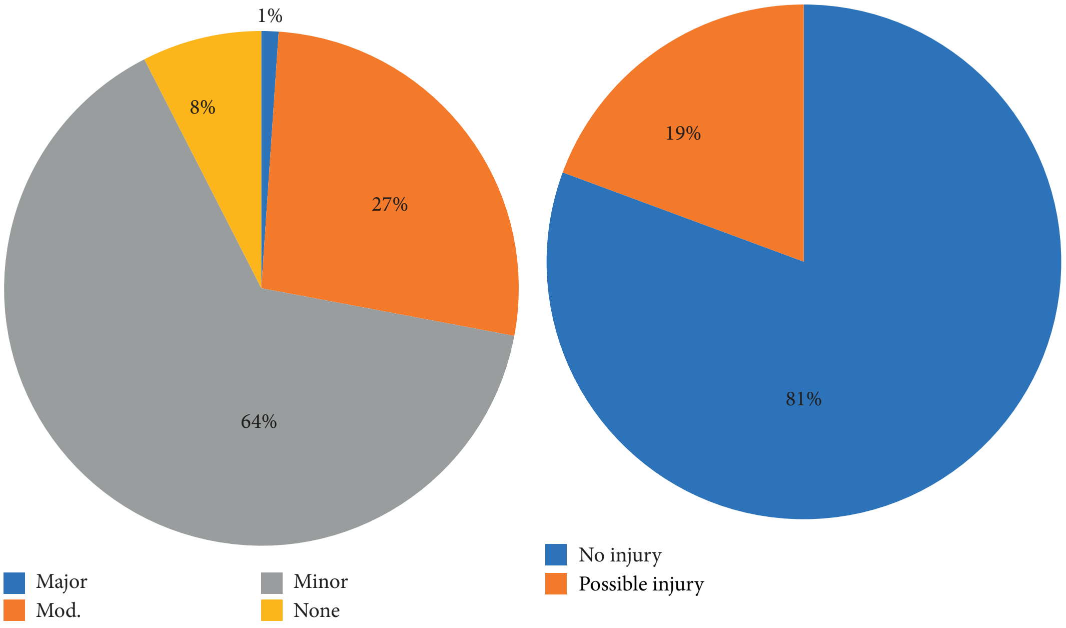


Figure 3: Movement preceding collision.



(a)

(b)

Figure 4: Crash severity and vehicle damage. (a) Vehicle damage description. (b) Crash severity.

accuracy might still be very high. To address the limitations of the overall classification accuracy, the *G*-mean (geometric mean) is considered a reasonable index to evaluate imbalanced data. It has a high value by balancing the classification accuracy of the minority and majority instances [45]. The *G*-mean is calculated as follows:

$$G-\text{mean} = \sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}} * \frac{\text{TN}}{\text{TN}+\text{FP}}}. \qquad (10)$$

The recall rate indicates the classification accuracy of minority instances, as shown in the following equation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}. \qquad (11)$$

Finally, *G*-mean and recall are employed as indexes to measure model performance.

5. Results and Discussion

5.1. Model Results. We use the grid search to determine the best combination of parameters to prevent the model from overfitting. The optimal parameter values are shown

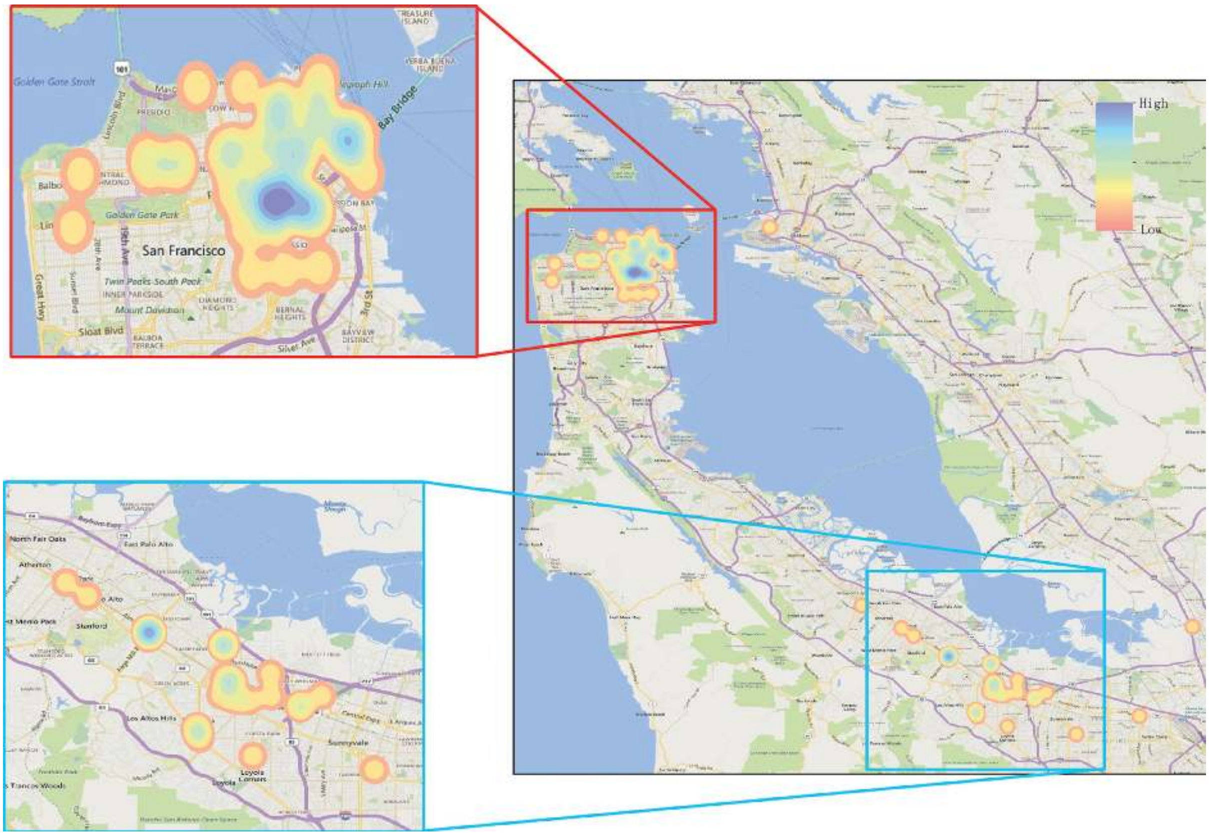


Figure 5: Heat map analysis of AV crashes.

in Table 4. The number of "n_estimators" is the number of trees that are fitted in the model. The parameter "gamma" is the minimum loss reduction required to make a further partition on a leaf node of the tree. The learning rate is used to shrink the weights in an update to prevent overfitting. The maximum depth of the tree represents the maximum number of splits; increasing the maximum depth can cause overfitting. The parameters "criterion" has the function of measuring the quality of a split. The parameter "min_samples_leaf" represents the minimum number of samples required to be at a leaf node.

The optimal CART and XGBoost models are established after parameter tuning. Then, the two models are tested on the same testing data to compare the predicted results. Table 5 shows the estimation results of the crash severity model.

The accuracy, recall, and G-mean results for the two modes are shown in Table 5, in which we can see that the XGBoost model performed better than the CART model, thereby reflecting the stability of XGBoost. Besides, the XGBoost model's accuracy is reduced by 26.1% after incorporating POI data, but the recall and *G*-mean have increased by 100% and 11.1%. It indicates that highly mixed land-use areas have a positive effect on identifying injury accidents. Additionally, as mentioned in Section 4.3, the *G*-mean and recall metrics are appropriate for imbalanced data because we need to identify injury accidents as much as possible. The recall and *G*-mean results of the calibration dataset in the XGBoost model are 84.6% and 69.9%, respectively; the recall and *G*-mean results of the validation dataset are 80% and 68.8%, respectively. The results between the calibration and validation dataset are relatively balanced, indicating that the model is of good fitting performance and prediction ability. In summary, the XGBoost model with POI data performs well in identifying the injured crashes.

5.2. Feature Analysis. Figure 6 illustrates the relationship between collision severity and potential contributing factors. Variables include the type of collision, the AVs movement preceding the crash, vehicle damage, accident location, driving mode, and weather.

We can observe that the weather is the most critical feature in the model. Specifically, injured accidents are more likely to occur in extreme weather conditions (e.g., fog and snow) [46] because sensors have poor perception performance in extreme weather. Rain and fog are composed of small water droplets that block the reflector and produce false alarms during obstacle detection [47]. According to Hasirlioglu et al. [48], in foggy weather, the relationship between temperature and visibility is inversely proportional, and visibility represents the distance that the detector can detect in this case.

Table 2: Variables description and distribution.

Variable	Total ($N = 94$) Distribution (%)	Uninjured ($N = 76$) Distribution (%)	Injured ($N = 18$) Distribution (%)
Type of collision			
Head-on = 0	8.60	10.67	0.00
Side swipe = 1	15.05	16.00	11.11
Rear end = 2	64.52	62.67	72.22
Broadside = 3	11.83	10.67	16.67
AV state (the AV movement preceding the collision)			
Stopped = 0	38.71	38.67	38.89
Moving = 1	61.29	61.33	61.11
Other vehicle state (other vehicle movement preceding the collision)			
Stopped = 0	1.08	1.33	0.00%
Moving = 1	98.92	98.67	100.00
Vehicle damage (describe vehicle damage)			
None = 0	7.52	9.33	0.00
Minor = 1	64.52	74.67	22.22
Mod = 2	26.89	14.675	77.78
Major = 3	1.08	1.33	0.00
Driving mode (driving mode preceding the collision)			
Conventional = 0	52.69	54.67	44.44
Autonomous = 1	47.31	45.33	55.56
Accident location			
Intersection = 1	47.31	45.33	55.56
Street = 2	35.48	44.00	0.00
Highway = 3	13.98	6.67	44.44
Parking lot = 4	3.23	4.00	0.00
Weather (weather conditions at the time of the accident)			
Clear = 1	77.42	90.67	22.22
Cloudy = 2	19.35	5.33	77.78
Raining = 3	2.15	2.67	0.00
Fog/visibility = 4	1.08	1.33	0.00
Lighting (lighting conditions at the time of the accident)			
Daylight = 1	64.52	69.33	44.44
Dusk-dawn = 2	2.15	1.33	5.56
Dark-street lights = 3	33.33	29.33	50.00
POI_D (the POI diversity evaluation index)			
POI_D ≤ 0.5 = 0	55.91	50.67	55.56
0.5 < POI_D ≤ 0.7 = 1	38.71	33.33	33.33
0.7 < POI_D = 2	5.38	16.00	11.11

Table 3: Confusion matrix.

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

Table 4: Parameter tuning results.

Optimal parameters	XGBoost model	CART model
n_estimators	670	/
gamma	0.1	/
learning_rate	0.2	/
max_depth	5	4
criterion	/	Gini
min_samples_leaf	/	1

The next two most crucial features are vehicle damage and accident location. The higher degree of vehicle damage corresponds to the higher severity of the accident. Crashes have a higher probability of occurring at intersections regardless of signalized or nonsignalized intersections [49]. The traffic environment at intersections is complex and changeable because vehicles, nonmotor vehicles, and pedestrians are highly mixed [50]. The crash reports do not cover the number of crossings under autonomous mode. Therefore, it is necessary to study the stability and safety of AVs when crossing intersections in autonomous driving mode. Existing AVs do not take advantage of the convenience brought by infrastructure-to-vehicle (I2V) communication. These facilities can decrease the spatial and temporal instabilities and promote the safety of drivers, cyclists, and pedestrians [51].

Table 5: Estimation results of the crash severity model.

	Calibration			Validation		
	Accuracy (%)	Recall (%)	*G*-mean (%)	Accuracy (%)	Recall (%)	*G*-mean (%)
CART model	81.5	23.1	47.1	82.1	20.0	43.7
XGBoost model	83.3	46.2	65.9	85.2	40.0	61.9
XGBoost model + POI	63.1	84.6	69.9	63.0	80.0	68.8

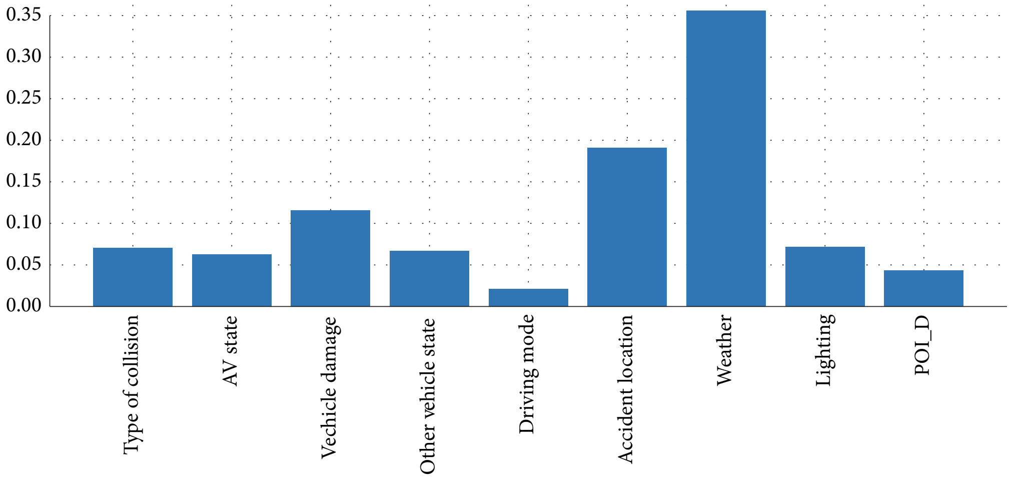


Figure 6: Feature importance.

Injury crashes are also more likely to occur in areas with high mixed land-use areas. Chen [52] supposed that areas with a high degree of mixed land-use have various functions, which significantly increased the conflict points between vehicles, bicycles, and pedestrians. Diverse land-use regions are prone to diversified traffic behaviors and increased regional traffic flow, affecting traffic safety substantially. This finding seems intuitive because mixed land-use patterns typically exhibit diverse land-use types leading to complex roadway layouts. The results suggest that areas with mixed land-use require additional local-level research to develop effective target-oriented treatments to improve the safety performance of AVs.

Many rear-end accidents occur in the data set (i.e., 65%), but AVs are usually not the responsible party. Conventional vehicles are not accustomed to the AVs' characteristics, and, thus, crashes are often caused by conventional vehicles hitting the rear of AVs. Wang et al. [18] find that crash severity significantly increases if the AV is responsible for the crash. From the perspective of driving mode and vehicle movement before the accident, 56% of AVs' injury accidents were in autonomous driving mode. In February 2016, the AV produced by Google had its first accident when it changed lanes and collided with a bus at a low speed, causing no casualties. Three months later, Tesla in the US suffered a more serious fatal crash while driving in autonomous mode. It was the first known fatal crash in the history of AV technology. These two crashes were regarded as important events since the advent of AV, indicating a new type of traffic crash. The main reason for these two crashes was that the driver ignored the warning to take over from the AVs, which meant that the two drivers did not take over the driving in time to ensure driving safety. This also shows that AVs are responsible for these crashes. In conclusion, severe injuries can happen if the vehicle is on automated driving mode and is the crash's primary responsible party.

6. Summary and Conclusion

In the current study, the XGBoost model incorporating POIs data is adopted to investigate the factors contributing to the severity of AV involved crashes using reported crashes from California. The descriptive statistics analysis was employed to investigate the characteristics of AV involved crashes in terms of the crash location, collision type, crash severity, and vehicle movement before crash occurrence. A total of 94 accident cases were employed to train the model that reached the G-mean and recall of 68.8% and 80%, respectively. The recall and G-mean have increased by 100% and 11.1% after incorporating POI data. It indicates that highly mixed land-use areas have a positive effect on identifying injury accidents.

We find that the degree of vehicle damage, accident location, and type of collision significantly affect the severity of the crash, which is consistent with previous research. The difference is that this study finds weather conditions to be the most critical factor. Extreme weather and intersection accidents have a significant effect on the severity of an accident. Regardless of signalized or nonsignalized intersections, intersections are the most likely places for rear-end collisions. Mainly because a crash occurred when the vehicle was waiting at the intersection or driving slowly. We recommend using the vehicle sensor with strong stability and high sensitivity. Besides, switching to manual driving is also a solution to avoid severe accidents caused by automatic driving.

The results from the model can also provide a basis for policy decisions. For example, results reveal a significantly higher likelihood of injured crashes in mixed land-use settings. Hence, specific recommendations are made to promote the mixed-use of land and reduce traffic accidents. First, urban planning should focus on developing small-scale and high-intensity diverse land-use at the microlevel and constructing a multicenter urban pattern at the macrolevel to achieve a balanced population and transportation. The mixed land-use development needs to be based on the construction of rapid rail transit facilities, encourage walking and nonmotorized travel by improving road traffic conditions, and use public transportation as the core to reduce traffic accidents caused by the rapid development of vehicles. Areas with highly mixed land-use require additional local-level studies to develop effective treatments to improve the safety performance of AVs.

As a limiting factor, this study did not collect data on vehicle speed and driver characteristics before the accident to evaluate the safety performance of AVs. Despite the limited sample size, the collision database employed in this study includes all issued crash reports involving AVs in 2019. However, to gain a deeper understanding of the mechanism of AV crash, future research should continue to collect AV crash data and apply multisource data fusion to enhance prediction accuracy.

The model adopted in this study supplies an accepted method for investigating and understanding AV safety issues. Moreover, because the sample size increases in the future, this advantage can continue to increase. The knowledge gained from this research could contribute to the assessment and improvement of the safety performance of the current AVs.

Data Availability

The data used to support the findings of this study are publicly available at https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/autonomousveh_ol316.

Conflicts of Interest

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

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant no. 52002282) and Natural Science Foundation of Zhejiang Province (LQ19E080003).

References

[1] D. J. Fagnant and K. Kockelman, “Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations,” *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, 2015.

[2] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Kouros Mohammadian, “Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis,” *Accident Analysis and Prevention*, vol. 136, Article ID 105405, 2020.

[3] J. Stilgoe, “Machine learning, social learning and the governance of self-driving cars,” *Social Studies of Science*, vol. 48, no. 1, pp. 25–56, 2018.

[4] A. Hevelke and J. Nida-Rümelin, “Responsibility for crashes of autonomous vehicles: an ethical analysis,” *Science and Engineering Ethics*, vol. 21, no. 3, pp. 619–630, 2015.

[5] K. Massow and I. Radusch, “A rapid prototyping environment for cooperative advanced driver assistance systems,” *Journal of Advanced Transportation*, vol. 2018, Article ID 2586520, 32 pages, 2018.

[6] J. Khoury, K. Amine, and R. A. Saad, “An initial investigation of the effects of a fully automated vehicle fleet on geometric design,” *Journal of Advanced Transportation*, vol. 2019, Article ID 6126408, 10 pages, 2019.

[7] D. Yang, K. Jiang, D. Zhao et al., “Intelligent and connected vehicles: current status and future perspectives,” *Science China Technological Sciences*, vol. 61, no. 10, pp. 1446–1471, 2018.

[8] A. D. McDonald, H. Alambeigi, J. Engström et al., “Toward computational simulations of behavior during automated driving takeovers: a review of the empirical and modeling literatures,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 61, no. 4, pp. 642–688, 2019.

[9] S. Deb, M. M. Rahman, L. J. Strawderman, and T. M. Garrison, “Pedestrians’ receptivity toward fully automated vehicles: research review and roadmap for future research,” *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 3, pp. 279–290, 2018.

[10] F. Hegedüs, T. Bécsi, S. Aradi, and P. Gápár, “Model based trajectory planning for highly automated road vehicles,” *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6958–6964, 2017.

[11] A. Omidvar, M. Pourmehrab, P. Emami et al., “Deployment and testing of optimized autonomous and connected vehicle trajectories at a closed-course signalized intersection,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 19, pp. 45–54, 2018.

[12] J. C. F. De Winter, R. Happee, M. H. Martens, and N. A. Stanton, “Effects of adaptive cruise control and highly automated driving on workload and situation awareness: a review of the empirical evidence,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 27, pp. 196–217, 2014.

[13] C. Ma, J. Zhou, and D. Yang, “Causation analysis of hazardous material road transportation accidents based on the ordered logit regression model,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 4, 1259 pages, 2020.

[14] S. Dong and J. Zhou, “A comparative study on drivers ’ stop/go behavior at signalized intersections based on decision tree classification model,” *Journal of Advanced Transportation*, pp. 1–13, 2020.

[15] H. Jeong, Y. Jang, P. J. Bowman, and N. Masoud, “Classification of motor vehicle crash injury severity: a hybrid

approach for imbalanced data," *Accident Analysis and Prevention*, vol. 120, pp. 250–261, 2018.

[16] Z. Liu, H. Chen, Y. Li, and Q. Zhang, "Taxi demand prediction based on a combination forecasting model in hotspots," *Journal of Advanced Transportation*, vol. 2020, pp. 1–13, Article ID 1302586, 2020.

[17] Z. Liu, H. Chen, X. Sun, and H. Chen, "Data-driven real-time online taxi-hailing demand forecasting based on machine learning method," *Applied Science*, vol. 10, no. 19, 2020.

[18] S. Wang and Z. Li, "Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches," *PLoS ONE*, vol. 14, no. 3, pp. 1–16, 2019.

[19] C. Xu, Z. Ding, C. Wang, and Z. Li, "Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes," *Journal of Safety Research*, vol. 71, pp. 41–47, 2019.

[20] A. M. Boggs, B. Wali, and A. J. Khattak, "Exploratory analysis of automated vehicle crashes in California: a text analytics and hierarchical Bayesian heterogeneity-based approach," *Accident Analysis and Prevention*, vol. 135, Article ID 105354, 2020.

[21] A. K. Agarwal, S. Wadhwa, and S. Chandra, "Diagnosis of tuberculosis--newer tests," *Journal Association Physicians India*, vol. 42, no. 8, p. 665, 1994.

[22] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher, "A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset," *Accident Analysis and Prevention*, vol. 127, pp. 134–149, 2019.

[23] S. R. Mousa, P. R. Bakhit, and S. Ishak, "An extreme gradient boosting method for identifying the factors contributing to crash/near-crash events: a naturalistic driving study," *Canadian Journal of Civil Engineering*, vol. 46, no. 8, pp. 712–721, 2019.

[24] H. Meng, X. Wang, and X. Wang, "Expressway crash prediction based on traffic big data," in *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning - SPML '18*, Taiwan, China, November 2018.

[25] L. Kuang, H. Yan, Y. Zhu, S. Tu, and X. Fan, "Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor," *Journal of Intelligent Transportation Systems*, vol. 23, pp. 1–14, 2019.

[26] Y. Pan, S. Chen, S. Niu, Y. Ma, and K. Tang, "Investigating the impacts of built environment on traffic states incorporating spatial heterogeneity," *Journal of Transport Geography*, vol. 83, Article ID 102663, 2020.

[27] R. Jia, A. Khadka, and I. Kim, "Traffic crash analysis with point-of-interest spatial clustering," *Accident Analysis and Prevention*, vol. 121, pp. 223–230, 2018.

[28] E. Chen, Z. Ye, C. Wang, and W. Zhang, "Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data," *Cities*, vol. 95, Article ID 102359, 2019.

[29] J. Bao, X. Shi, and H. Zhang, "Spatial analysis of bikeshare ridership with smart card and POI data using geographically weighted regression method," *IEEE Access*, vol. 6, pp. 76049–76059, 2018.

[30] S. Wang and Z. Li, "Exploring causes and effects of automated vehicle disengagement using statistical modeling and classification tree based on field test data," *Accident Analysis and Prevention*, vol. 129, pp. 44–54, 2019.

[31] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 143–163, 2016.

[32] J. Rios-Torres and A. A. Malikopoulos, "A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 5, pp. 1066–1077, 2017.

[33] X. Li, W. Wang, C. Xu, Z. Li, and B. Wang, "Multi-objective optimization of urban bus network using cumulative prospect theory," *Journal of Systems Science and Complexity*, vol. 28, no. 3, pp. 661–678, 2015.

[34] C. Xu, X. Wang, H. Yang, K. Xie, and X. Chen, "Exploring the impacts of speed variances on safety performance of urban elevated expressways using GPS data," *Accidental Analysis and Preview*, vol. 123, pp. 29–38, 2019.

[35] X. Wang, Q. Zhou, M. Quddus, T. Fan, and S. Fang, "Speed, speed variation and crash relationships for urban arterials," *Accidental Analysis and Preview*, vol. 113, pp. 236–243, 2018.

[36] D. A. Belsley, "A Guide to using the collinearity diagnostics," *Computer Science in Economics and Management*, vol. 4, no. 1, pp. 33–50, 1991.

[37] C. H. Mason and W. D. Perreault, "Collinearity, power, and interpretation of multiple regression analysis," *Journal of Marketing Research*, vol. 28, no. 3, p. 268, 1991.

[38] T. Naes and B. H. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," *Journal of Chemometrics*, vol. 15, no. 4, pp. 413–426, 2001.

[39] Y. Pan, S. Chen, T. Li, S. Niu, and K. Tang, "Exploring spatial variation of the bus stop influence zone with multi-source data: a case study in Zhenjiang, China," *Journal of Transport Geography*, vol. 76, pp. 166–177, 2019.

[40] M. Petrere, "Pesque-solte," *Ciência Hoje*, vol. 53, no. 5, pp. 1189–1232, 2014.

[41] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[42] H. Zhang, D. Qiu, R. Wu, Y. Deng, D. Ji, and T. Li, "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Applied Soft Computing*, vol. 80, pp. 57–79, 2019.

[43] J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan, and J. Zhang, "Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: a city management perspective," *IEEE Access*, vol. 7, pp. 148059–148072, 2019.

[44] X. Li, J. Tang, X. Hu, and W. Wang, "Assessing intercity multimodal choice behavior in a Touristy City: a factor analysis," *Journal of Transport Geography*, vol. 86, Article ID 102776, 2020.

[45] A. P. Bradley, "The use OF the area under the roc curve IN the evaluation OF machine learning algorithms," *Transportation Research Part A Policy Practice*, vol. 44, no. 5, pp. 291–305, 1997.

[46] X. Yu and M. Marinov, "A study on recent developments and issues with obstacle detection systems for automated vehicles," *Sustain.*vol. 12, no. 8, 2020.

[47] R. H. Rasshofer, M. Spies, and H. Spies, "Influences of weather phenomena on automotive laser radar systems," *Advances in Radio Science*, vol. 9, pp. 49–60, 2011.

[48] S. Hasirlioglu, A. Riener, W. Ruber, and P. Wintersberger, "Effects of exhaust gases on laser scanner data quality at low ambient temperatures," *IEEE Intelligent Vehicle Symposium*, no. Iv, pp. 1708–1713, 2017.

[49] V. Gadepally, A. Krishnamurthy, and Ü. Ozguner, "A framework for estimating long term driver behavior," *Journal*

of Advances Transportation, vol. 2017, Article ID 3080859, 11 pages, 2017.

[50] Y. Hu, J. Ou, and L. Hu, "A review of research on traffic conflicts based on intelligent vehicles perception technology," in *Proceedings of the 2019 International Conference on Advances in Construction Machinery and Vehicle Engineering (ICACMVE)*, Changshu, China, May 2019.

[51] O. Grembek, A. Kurzhanskiy, A. Medury, P. Varaiya, and M. Yu, "Making intersections safer with I2V communication," *Transportation Research Part C: Emerging Technologies*, vol. 102, pp. 396–410, 2019.

[52] P. Chen, "Built environment factors in explaining the automobile-involved bicycle crash frequencies: a spatial statistic approach," *Safety Science*, vol. 79, pp. 336–343, 2015.