IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2019, accepted October 2, 2019, date of publication October 9, 2019, date of current version October 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946401

Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective

JUN MA[1,2], YUEXIONG DING[2], JACK C. P. CHENG[1], YI TAN[3], VINCENT J. L. GAN[1], AND JINGCHENG ZHANG[4]

[1]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong
[2]Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong
[3]College of Civil Engineering, Shenzhen University, Shenzhen 518060, China
[4]Shenzhen Mariocode Science and Technology Company Ltd., Shenzhen 518060, China

Corresponding author: Jingcheng Zhang (zhangjingcheng0306@outlook.com)

ABSTRACT Traffic accidents have been one of the most important global public problems. It has caused a severe loss of human lives and property every year. Studying the influential factors of accidents can help find the reasons behind. This can facilitate the design of effective measures and policies to reduce the traffic fatality rate and improve road safety. However, most of the existing research either adopted methods based on linear assumption or neglected to further evaluate the spatial relationships. In this paper, we proposed a methodology framework based on XGBoost and grid analysis to spatially analyze the leading factors on traffic fatality in Los Angeles County. Characteristics of the collision, time and location, and environmental factors are considered. Results show that the proposed method has the best modeling performance compared with other commonly seen machine learning algorithms. Eight factors are found to have the leading impact on traffic fatality. Spatial relationships between the eight factors and the fatality rates within the Los Angeles County are further studied using the grid-based analysis in GIS. Specific suggestions on how to reduce the fatality rate and improve road safety are provided accordingly.

INDEX TERMS XGBoost, factors analysis, GIS, grid-based analysis, non-linear machine learning, traffic fatality.

I. INTRODUCTION

Traffic accidents are considered as one of the most critical and dangerous problems all around the world. According to World Health Organization (WHO), about 1.3 million people die each year due to traffic accidents. An additional 20-50 million are injured or disabled [1]. In the United States, traffic fatalities increased by 15.3 percent from 2011 to 2016 (29,867 to 37,461) [2]. How to reduce the occurrence of fatal traffic accidents and improve road safety has been a significant problem for both governments and research institutions. To this end, many scholars have conducted different kinds of research to study traffic accidents. Some focused on the road safety management and education [3], [4], some paid attention to the improvement of vehicles [5], [6], some studied the emergency medical service [7], [8], and some focused on the factors that influence the severity of traffic accidents [9]–[12].

Knowing what the influential factors are and how they affect the accidents can help us better understand the cause-effect behind. This is beneficial to improve the estimation of the accident severity and the preparation of countermeasures. Many factors have been studied by previous literature. For example, Adanu *et al.* [10] studied the factors that influence the severity of single-vehicle accidents that happen on weekdays and weekends. They found that factors such as driver unemployment, driving with invalid license, no seatbelt, fatigue, have a high correlation with the severity of the accidents. Mohamed *et al.* [11] studied the influential factors of rear-end crashes. They found that factors like tailgating, driving too fast, years of experience, number of lanes, are significantly affecting the severity of rear-end crashes. Lee *et al.* [13] analyzed the impact of rainfall intensity and water depth on traffic accident severity.

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

Schneider *et al.* [14] examined the statewide motorcycle crash data in a five year range in Ohio and found that female, the presence of alcohol and high degree curve would increase the likelihood of severe traffic accidents. Hashimoto *et al.* [15] studied the correlations between traffic accidents and environmental factors like population and road condition. Aziz *et al.* [16] concluded that low-lighted-roads and pedestrian crossing intersections would increase the likelihood of severe pedestrian-vehicle accidents.

When analyzing the impact factors, these literature had adopted different kinds of statistical methods. For example, Russo and Savolainen [17] applied three regression models namely negative binomial, order logit and multinomial logit to ascertain the relationship between different median barrier types and freeway median crash frequency and severity. Dapilah *et al.* [18] examined how motorcyclist characteristics influence the road traffic accidents using chi-square analysis. Karimi and Kashi [19] investigated the effect of geometric parameters on accident reduction using sensitivity analysis. Onozuka *et al.* [20] studied whether a full moon contributes to the road traffic accidents using conditional Poisson regression. However, most of these methods are based on linear assumptions. Their performances are limited since the real cases are very complicated, and the relationships between the factors and accidents are non-linear. Therefore, non-linear machine learning algorithms were applied in some recent research to address this problem. For example, Mussone *et al.* [21] predicted the severity of the accidents in urban road intersections using Artificial Neural Network (ANN). Li *et al.* [22] used Support Vector Machine (SVM) models for crash injury severity analysis. However, algorithms like ANN and SVM are referred to as black box processes because they do not provide a direct explanation of the variable importance [23]. This does not help when analyzing the cause-effect behind traffic accidents. Therefore, a non-linear methodology framework with the ability to calculate the variable importance needs to be proposed.

Furthermore, geographical information system (GIS) is gradually becoming more popular due to its ability to capture, store, manipulate, analyze, manage and present spatial or geographical data [24]–[26]. Many previous studies have used GIS to analyze the data spatially and temporally. For example, Kazmi and Zubair [27] estimated the vehicle damage cost involved in road traffic accidents based on GIS. Din *et al.* [28] assessed the level of service of public transportation using GIS. Zangeneh *et al.* [29] conducted a spatial-temporal cluster analysis of mortality from road traffic injuries using GIS. However, few studies have combined GIS and machine learning algorithms to analyze the influential factors on traffic accident fatality.

The objective of this paper is to propose a Grid-based non-linear machine learning framework to analyze the critical factors that influence traffic accidents fatality. A non-linear machine learning algorithm namely XGBoost is implemented to process the data and distinguish the variable with higher importance. Its performacne is compared with five other machine learning algorithms, including multiple linear regression (MLR), logistic regression (LR), multi-layer perceptron (MLP), support vector machine (SVM) and random forest (RF). Factors including crash characteristics, time and locations, and environmental characteristics are considered in the model. Whether a traffic accident involves fatality is the classification target. Results show that XGBoost outperforms other algorithms with higher classification accuracy. Eight factors such as alcohol use and lighting conditions are found to be the leading causes of fatal accidents based on the proposed framework. GIS is then used to conduct the spatial analysis on the relationships between these eight factors and the fatality rate.

The remaining paper is organized as follows: Section II describes the proposed methodology framework, Section III presents a case study in Los Angeles County, Section IV and Section V show the results and discussion, and Section VI concludes the work.

II. METHODOLOGY FRAMEWORK

Figure 1 shows the proposed methodology framework. It consists of three parts. The first part is data preprocessing. Data collection, cleaning, formatting, and balancing are conducted. The second part is model training and optimization. XGBoost is applied to build the classification model and calculate the feature importance. The last part is post-mining. This is accomplished by conducting the grid-based spatial analysis in GIS.

A. XGBOOST

The classification modeling using Extreme Gradient Boosting Decision Tree (XGBoost) is an essential component in the proposed framework. It plays the role of modeling between the urban features and the outputs. This algorithm has been reported to achieve good performance in different research domains. For example, Zhang and Zhan [30] adopted XGBoost to classify the rock faces. Torlay *et al.* [31] utilized XGBoost to identify atypical language patterns and differentiate patients with epilepsy. Ma and Cheng [32] improved the identification accuracy using XGBoost when modeling features that affect the green building markets. Zheng *et al.* [33] applied XGBoost to evaluate the feature importance in short-term electric load forecasting.

XGBoost is an ensemble technique developed based on the Gradient Boosting proposed by Friedman [34]. It learns a set of regression trees (CARTs) in parallel and obtains the result by summing up the score of each CART. Compared with the original GBDT algorithm [34], Chen and Guestrin [35] added some improvements in 2016 and named it as XGBoost. One of special improvements is the regularized objective to the loss function. Calculation of the regularized objective L_k for the k^{th} iteration is shown in Equation (1).

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

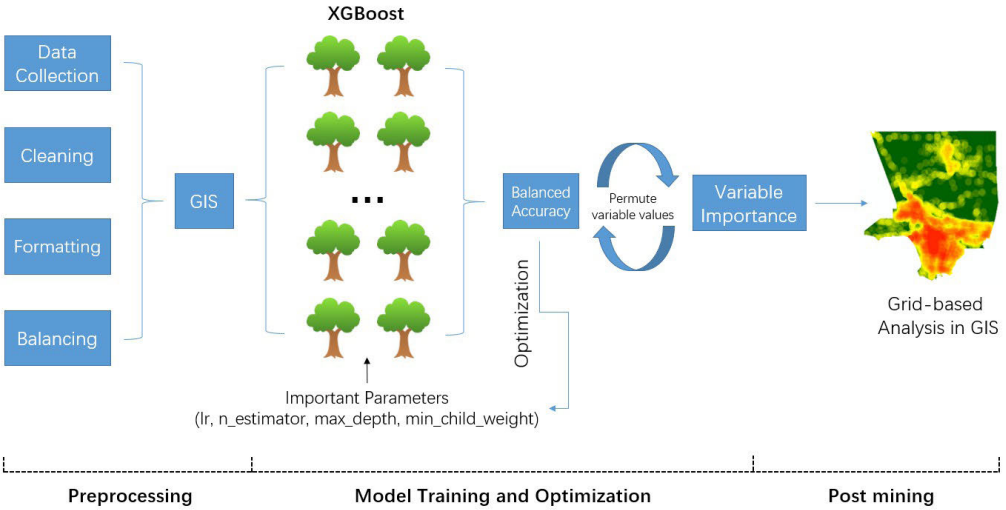


FIGURE 1. Methodology framework.

where n represents the number of samples, $\hat{y}_k^{(i)}$represents the prediction of the sample i at iteration k, $l(\cdot)$represents the original loss function. Ω is the regularization term, and is calculated by Equation (2).

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

where T represents the number of leaf nodes. γ and λ are two constants used to control the regularization degree.

Another improvement of XGBoost is the utilization of additive learning strategy [36]. Instead of applying stochastic gradient descent method to complement the corresponding optimization procedure, XGBoost adds the best tree model $f_k(x^{(i)})$ into the current classification model to give prediction result for the mth iteration. In this case, Equation (1) can be further formulated as follows:

$$L_k = \sum_{i=1}^{n} l(y^{(i)}, \hat{y}_{k-1}^{(i)} + f_k(x^{(i)})) + \Omega(f_k) + \sum_{j=1}^{k-1} \Omega(f_j) \tag{3}$$

Furthermore, XGBoost adopts the Taylor Expansion second order to the objective function, and Equation (3) can be further expanded to Equation 4.

$$L_k = \sum_{i=1}^{n} [l(y^{(i)}, \hat{y}_{k-1}^{(i)}) + g_i \cdot f_k(x^{(i)}) + \frac{1}{2} h_i \cdot f_k(x^{(i)})] + \Omega(f_k) + C \tag{4}$$

where $g_i = \partial_{\hat{y}_{k-1}} l(y^{(i)}, \hat{y}_{k-1})$ and $h_i = \partial^2_{\hat{y}_{k-1}} l(y^{(i)}, \hat{y}_{k-1})$ are the first and second order derivatives on the loss function respectively, C is a constant.

These advantages make XGBoost potentially to obtain better results than traditional GBDT. According to the results of the experiments in our case study, it also outperforms other commonly-used machine learning algorithms. More details will be provided later.

B. VARIABLE IMPORTANCE

Except the ability in modeling non-linear classification and regression problems, XGBoost is also capable of ranking the variable importance by using the weight. It is given by the frequency of a feature used in splitting the data across all CARTs. Importance of a variable v (W_v) based on the weights can be calculated by Equation 5 and Equation 6.

$$W_v = \sum_{k=1}^{K} \sum_{l=1}^{L-1} I(V_k^l, v) \tag{5}$$

$$I(V_k^l, v) = \begin{cases} 1, & \text{if } V_k^l = v \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where K represents the number of trees or iterations, L represents the number of leaf nodes of the k^{th} tree, V_k^lrepresents the feature related to the node l, I(·) represents the indicator function.

C. GRID-BASED ANALYSIS IN GIS

A geographical information system (GIS) is a framework for gathering, integrating, managing, and analyzing data. It can analyze the spatial locations and organize different layers of information into visualizations using maps and 3D scenes. With this capability, GIS reveals deeper and broader spatial insights in data.

In this article, GIS is applied to further detect the spatial relationships between the fatality rate and the influential factors. However, since we cannot directly calculate the fatality rate using the traffic data points, the gridding technique in GIS is then applied to tackle the problem. It divides the studied region into many fishnets (grids), and then maps the traffic data points into these grids. The fatality rate R_f^i of the ith grid can be calculated using Equation 7.

$$R_f^i = \frac{N_f^i}{N_a^i} \tag{7}$$

where N_f^i means the total number of fatal accidents, and N_a^i means the total number of accidents.

III. CASE STUDY

A. DATA COLLECTION

To validate the proposed framework, we conducted a case study in Los Angeles County, USA. The dataset was collected from California Statewide Integrated Traffic Records System (SWITRS). It records every traffic accident reported in Los Angeles County during 2010-2012, totaling 307,971 accidents. Each record contains the information of the accident severity, the number of parties involved, the surrounding environment, and the location. Table 1 presents an overall summary of the data. The data contains 28 collision features, 18 time and location features, and 7 environment features, so the data dimension is 53. Besides, since the study focuses on analyzing the leading causes of traffic fatalities, we set whether the case involves fatality as the classification target. Therefore, 3,146 fatal accidents are marked as negative cases while 304,825 non-fatal accidents are marked as positive cases.

TABLE 1. Data summary.

Item	Description
Negative cases (Fatal accidents)	3,146 cases
Positive cases (Non-fatal accidents)	304,825 cases
Total number of accidents	307,971 cases
Number of collision description variables	28 features
Number of time and location variables	18 features
Number of environment variables	7 features
Total variable dimension	53 dimensions
Time span	2010-2012
Geographic area	Los Angeles County

Table 2 shows the detailed description of the data features. The second column in Table 2 shows the feature abbreviation and description. The third column presents the data type. Here C represents categorical features, N represents numeric features, B represents binary features, and S represents string features. For example, VIOLCAT in Collision Description means the Violation Category, and this categorical feature recorded what kind of traffic violation has the vehicle committed, including speeding, impeding traffic, traffic signals etc. POP in the Environment represents the surrounding population density of the accident spot. For more details, please refer to the official SWITRS website.

TABLE 2. Feature description.

	Feature abbreviation and description	Data type[3]
Collision Description	CHPTYPE:CHP[1] Beat Type	C
	VIOLCODE:PCF[2] Violation Code	C
	VIOLCAT: PCF Violation Category	C
	CRASHTYP:Type of Collision	C
	INVOLVE:Motor Vehicle Involved With	C
	PED:Pedestrian Action	C
	CHPFAULT:CHP Vehicle Type is at fault	C
	SHIFT:CHP effect of new 12-hour shifts	C
	VIOLSUB:PCF Violation Subsection	C
	BEATTYPE:Beat Type	C
	PARTIES:Party Count	N
	PEDCOL:whether involved a pedestrian	B
	BICCOL:whether involved a bicycle	B
	MCCOL:whether involved a motorcycle	B
	TRUCKCOL:whether involved a big truck	B
	ETOH:whether involved drinking party	B
	STFAULT:indicates who is at fault	C
	HITRUN:Hit And Run	C
	PCF:Primary Collision Factor	C
	RIGHTWAY:Control Device	C
	NOTPRIV:whether on private property	B
	DIRECT:Direction of the offset distance	C
	INTERSECT_:whether at an intersection	B
	KILLED:counts of victims with 1-degree injury	N
	INJURED: counts of victims with 2,3,4 of injury	N
	CRASHSEV: the severity level of the collision	N
	BEATNUMB: Beat Number	S
	VIOL: PCF Violation	S
Time and Location	YEAR_:Collision Year:	C
	MONTH_:The month of the year	C
	DAYWEEK:The Day of Week	C
	TIMECAT:3-hour categories time	C
	DATE_: the date when the collision occurred	C
	TIME_: the time (24 hour time)	N
	LAPDDIV:City Division LAPD	S
	STATEHW:whether on a state highway	B
	POINT_X: The longitude of the geocoded location	N
	POINT_Y: The latitude of the geocoded location	N
	JURDIST: Reporting District	S
	DISTANCE: Offset distance from the secondary road	N
	LOCATION: the location code PRIMARYRD	S
	JURIS: Jurisdiction	S
	POSTMILE: markers that indicate the mileage in California	N
	SECONDRD: A secondary reference road	S
	SPECIAL:Special Condition	C
	RAMP:Ramp Intersection	C
Environment	WEATHER:the weather condition	C
	WEATHER2:the additional weather condition	C
	LIGHTING:lighting condition	C
	POP:Population level	C
	ROADSURF:Road Surface	C
	CHPRDTYP:CHP Road Type	C
	RDCOND1:Road Condition 1	C

1: California Highway Patrol
2: Primary Collision Factor
3: C-categorical, B-binary, N-numeric, S-string

B. DATA PREPROCESSING

Since the raw data always has some flaws, preprocessing is usually necessary before applying the data into the mathematical model. The preprocessing process in this study includes three parts: data formatting, data cleaning, and data balancing.

1) DATA FORMATTING

The first part is data formatting. The features collected in this study are mostly categorical but not ordinal. However, some machine learning algorithms like logistic regression cannot operate on categorical values directly. They require the input variables and the output variables to be numeric. Therefore, these categorical data were converted into dummy variables in this study. Note that a dummy variable means

to use numeric value 0 or 1 to represent a binary variable. For example, in the original dataset, categorical data DIRECT has four independent labels, including north, east, south, and west. After one hot encoding, four dummy variables, DIRECT_N, DIRECT_E, DIRECT_S, and DIRECT_W, are given to indicate the driving direction of the vehicle. In this way, 29 categorical features were formatted into 359 dummy variables, and the data dimension is expanded to 383.

2) DATA CLEANING

Data cleaning usually refers to the process of excluding useless data. Two types of data, including noisy data and high correlational data [37], [38], need to be deleted from the dataset. Noisy data means the data is irrelevant to our problem or has limited contribution to the model but increases the computation complexity. For instance, POINT_X, POINT_Y, LAPDDIV, and JURIS recorded the location and jurisdiction information. In this study, they are not the potential causes of accident fatality. Therefore, they are excluded from the data mining process.

High correlational data includes two kinds of data. The first kind is the data that are high correlational to the target. For example, CRASHSEV has the similar meaning with our target ''fatality''. It will influence the accuracy of our model. Therefore, it should be deleted. The second kind of data is those closely related to each other. For example, TIME means the record of accident time. TIMECAT has the similar meaning but divides the time into nine intervals. As TIMECAT is more convenient for further treatment and analysis, TIME is deleted. Pearson correlation coefficient is deployed to remove high correlational data. It is a method based on co-variance and can give information about the magnitude of the association or correlation. In each pair that has a correlation coefficient higher than 0.9, we would delete one feature in order to mitigate the multicollinearity problem.

In summary, 15 noisy features and 18 high correlational features are excluded from this study. The remained number of features reduces to 350.

3) DATA BALANCING

As shown in Table 1, there are 3,146 negative cases and 304,825 positive cases. This means the dataset is very imbalanced and could make our analysis biased. Therefore, the dataset needs to be balanced. Two sampling schemes namely oversampling and undersampling are commonly used to address this problem [39]. However, both of them have disadvantages. Oversampling will induce more massive computation and may cause overfitting, while undersampling may discard potentially useful information. In this study, we combine both undersampling and parts of oversampling concept to balance their disadvantages. This is achieved by conducting the under-sampling ten times and calculate the average performance. Each time, all the 3,146 negative cases were selected, and the same amount of positive cases was sampled and selected without replacement. Combine them to form a new dataset of 6,292 cases, and then cross-validate the mathematical model and get the result. Averaging the ten results gives the final result.

IV. RESULTS

A. MODELING AND OPTIMIZATION

The experiment was run on a computer with 16 GB ram, Intel (R) Xeon CPUL5640, Windows 7 operating system. The coding environment is Python 3.6. Parameter optimization of XGBoost is firstly conducted. As stated in Section II, parameter optimization refers to the adjustment of the algorithm parameters so that the algorithm can better model the problem. After referring some literature [30], [40]–[42], the following parameters of XGBoost are optimized in this section:

- *lr:* Learning rate. This parameter adjusts the size of the learning steps. Too small will lead to local optimum and slow down the calculation, while too large may miss the optimal value and not converge.
- *n_estimators:* Number of boosting rounds(or trees). This parameter represents the number of training iterations on the data. Too small will lead to under fitting, while too large will cause overfitting.
- *max_depth:* Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
- *min_child_weight:* Minimum sum of instance weight (hessian) needed in a child.

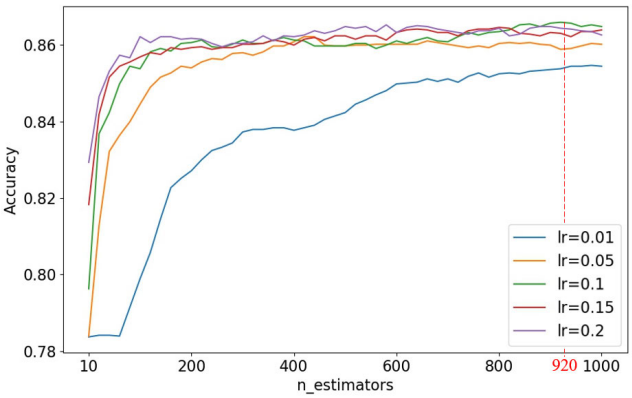


FIGURE 2. Optimization of lr and n_estimators.

Note that all the optimization process was conducted using 5-fold cross-validation to obtain stable results. To begin with, we set lr to {0.01, 0.05, 0.1, 0.15, 0.2} and n_estimators to {10, 20, 30, . . . , 1000}. The optimization result is shown in Figure 2. It can be seen that the pair {lr, n_estimators} = {1.0, 920} offered an optimal value. After this, max_depth and min_child_weight were then processed to be tuned. As shown in the Figure 3, when these two parameters were set as max_depth = 10 and min_child_weight = 1, the model was able to obtain an optimal result with 0.8673 in Accuracy. Note that the optimization criteria used here is Accuracy. Its calculation is shown in Figure 4.

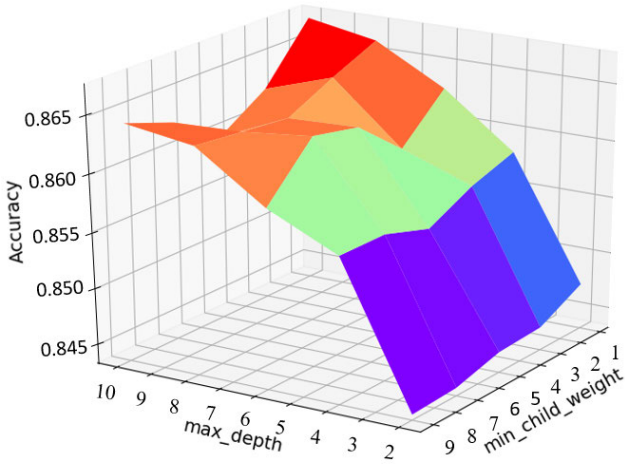


FIGURE 3. Optimization of max_depth and min_child_weight.

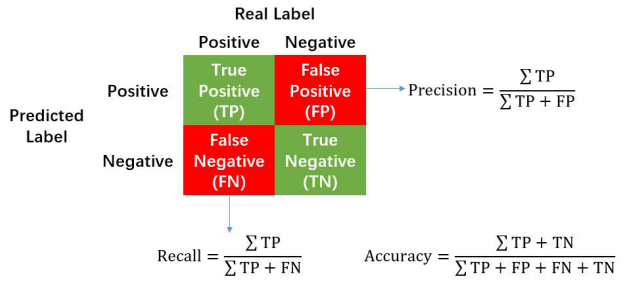


FIGURE 4. Calculation of Precision, Recall and Accuracy in the confusion matrix.

In addition, to further test the performance of XGBoost, we compared its performance with five other commonly seen models: Multiple Linear Regression (MLR), Logistic Regression (LR), Multi-Layer Perceptron (MLP, one typical Artificial Neural Network model), Support Vector Machine (SVM) and Random Forest (RF). 5-fold cross validation is applied to stabilize the results. Also, two more commonly used criteria for binary classifications are tested [42]. Their calculations are shown in Figure 4. Precision reflects the percentage of corrected predictions among the positively predicted cases, while recall represents the percentage of corrected predictions among the real positive cases. All Accuracy, Precision, and Recall are within [0, 1], and the higher values mean better predictions.

The performances of the models are shown in Table 3. As it can be seen, the indicators of the non-linear algorithms, MLP, SVM, RF, and XGBoost are significantly higher than the linear-based algorithms, MLR and LR. XGBoost outperforms MLP, SVM and RF with the highest numbers in all the three indicators. This proves that selecting XGBoost to model our problem and further analyze the variable importance is a reasonable choice.

B. VARIABLE IMPORTANCE

As introduced in Section II, XGBoost can help calculate the variable importance. Figure 5 presents the 15 features that obtained the highest importance. The most influential factor is ETHO, followed by PARTIES, CRASHTYPE_C, INVOLVE_C, LIGHTING_A, etc. Explanations of these features can be seen in Table 2.

TABLE 3. Comparisons of models.

Models	MLR	LR	MLP	SVM	RF	XGBoost
Accuracy	0.7739	0.7854	0.8278	0.8280	0.8463	0.8673
Precision	0.7249	0.7901	0.8335	0.8080	0.8674	0.8758
Recall	0.8490	0.7790	0.8205	0.8510	0.8178	0.8562

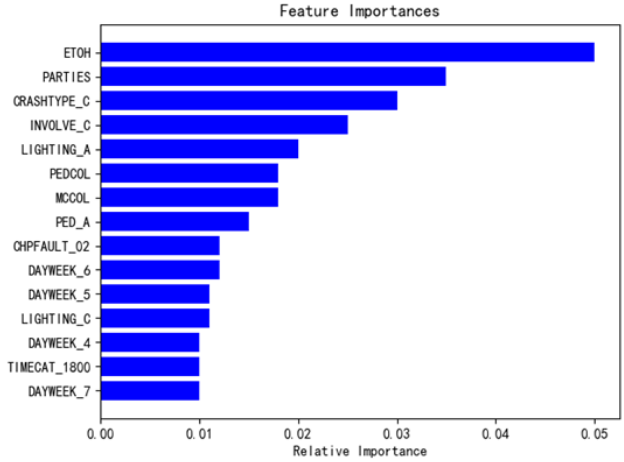


FIGURE 5. Top 15 influential factors.

It can be observed from Figure 5 that some features have similar meanings or belong to the same category of data. For example, MCCOL and CHPFAULT_02 both mean whether the accidents involve motorcycles. DAYWEEK_6, DAYWEEK_5, DAYWEEK_4, and DAYWEEK_7 all refer to the date category. LIGHTING_A and LIGHTING_C both refer to the lighting condition. PED_A and PEDCOL both relate to pedestrian accidents. INVOLVE_C refers to motor-motor collision, and it could be discussed when analyzing motor-pedestrian collisions. Therefore, we grouped the 15 features from Figure 5 into eight representative ones for a more explicit analysis and discussion. Table 4 lists these eight features and presents the fatality rates with and without the relevant feature. More details will be discussed in the following section.

V. DISCUSSION

Before diving into the eight features, identifying the overall distribution of the traffic accidents in Los Angeles County can help us better understand the spatial relationship between the fatality rate and the eight features. Figure 6 plots the distribution of the accident density, the fatal accident density, and the fatality rate. Figure 6 (1) and (2) were plotted using the kernel density in GIS. It can be seen that most accidents and fatal accidents happened in urban areas. Figure 6 (4) was calculated using the grid-based analysis introduced

TABLE 4. Fatality rates of the selected eight features.

Feature	Description	The fatality rate of the accidents that not involve the feature (%)	The fatality rate of the accidents that involve the feature (%)
ETOH	Drink-driving	0.714	4.645
PARTIES	Number of parties in the collision	0.986 (when the number <=4)	2.708 (when the number >4)
CRASHTYP_C	Rear-end collision	1.408	0.385
LIGHTING_A	Daylight	2.038	0.604
PEDCOL	Pedestrian accidents	0.673	4.547
MCCOL	Motorcycle accidents	0.912	3.24
DAYWEEK	Day of the week	0.904 (weekday)	1.414 (weekend)
TIMECAT	Time of the day	0.817 (6:00-24:00)	3.724 (00:00-6:00)

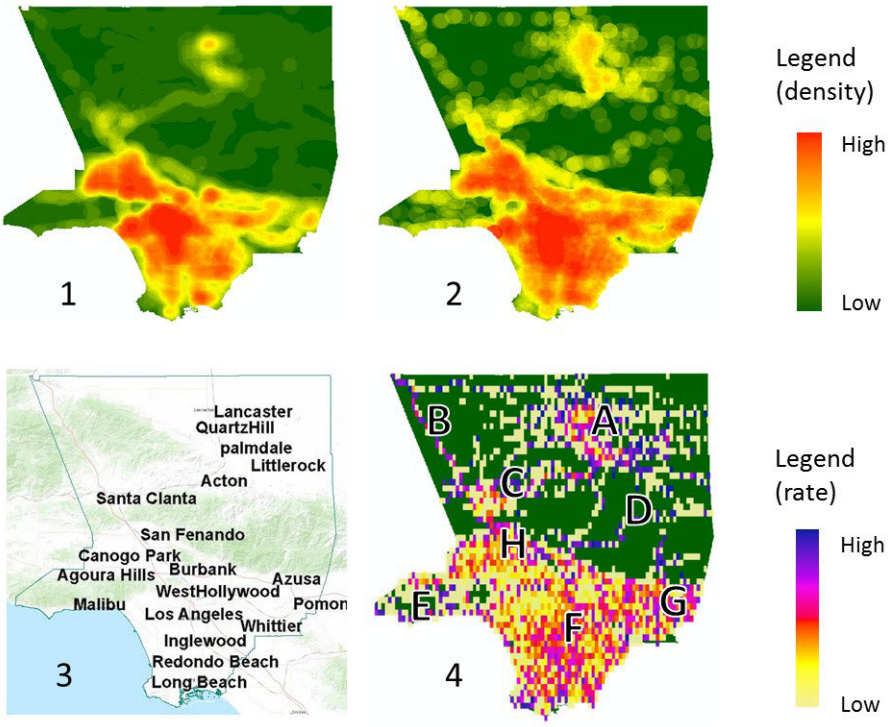


FIGURE 6. The distributions of (1) accident density, (2) fatal accident density, (3) main cities in Los Angeles County, and (4) fatality rates (grid-based).

in Section II-C. It is drawn by firstly cutting the Los Angeles County map in Figure 6 (3) into 60*60 grids. Then the R_f of each grid was summarized and calculated. The grids with $R_f = 0$ were presented using the green color, while others are plotted based on their R_f values from low to high. It can be observed that the fatality rates in eight areas are relatively higher. They are:

- *Area A:* the city of Lancaster, the city of Palmdale and their surrounding areas.
- *Area B:* Interstates 5 Highway between Santa Clarita Valley and the Pyramid Lake.
- *Area C:* the southern portion of the California State Route 14 (SR 14).
- *Area D:* the middle portion of the California State Route 2 (SR 2).
- *Area E:* the city of Malibu, the Malibu Creek State Park and their. surrounding areas.
- *Area F:* the east of Los Angeles city and its surrounding areas.
- *Area G:* city of Pomona, the city of West Covina and their surrounding areas.
- *Area H:* interchange of Interstates 405, 5 and 210 Highways and its surrounding areas.

It can be observed from Figure 6 that the distribution of the fatality rate is not highly correlated with the accident density. This is because the density of accidents mostly relies on traffic volumes. Higher traffic volumes in urban areas will naturally lead to more traffic accidents [9]. However, compared with other places, the higher fatality rates in area A-H mean that these places are more dangerous, and the

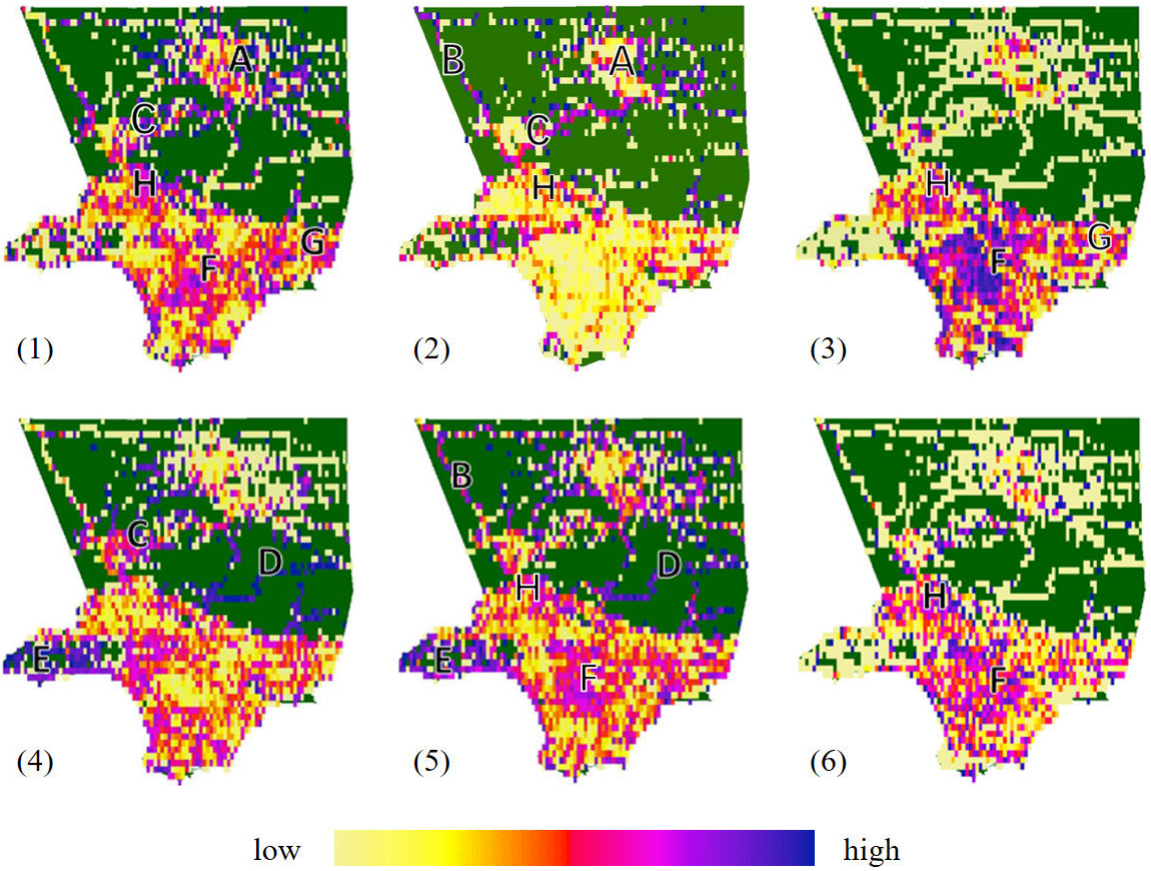


FIGURE 7. The grid-based distribution of the percentage of (1) the drink-driving accidents, (2) the accidents under poor lighting condition (no street lights or the lights are not functioning), (3) the accidents involving pedestrians, (4) the accidents involving motorcycles, (5) the accidents happened on weekends and (6) the accidents happened during 0am-6am.

traffic control and safety supervision are not doing well there. Therefore, one object of this study is to look into the reasons behind these dangerous areas and discuss possible measurements to improve the safety performance. According to the results shown in Table 4, the eight features may reveal some clues.

A. DRINK-DRIVING

Drink-driving is the most important feature on the fatality of the accidents according to our results. It can be seen from Table 4 that when under the influence of alcohol, the fatality rate is 4.645%, while the number drops to 0.714% for those without alcohol. Previous studies have demonstrated that driving under the influence of alcohol is considered an elevated risk of traffic accidents [18], [43]. The reason behind is that the response time and the awareness of the drivers will be impaired under the influence of alcohol. When the accidents are imminent, intoxicated drivers cannot judge the dangers or take evasive actions in time. This highly increases the potential fatality of traffic accidents.

Figure 7 (1) shows the distribution of the percentage of traffic accidents associated with drink-driving. It indicates that the numbers are higher in area A, C, G, H and F. This might be caused by the relatively looser drink-driving control there. Compared with Figure 6 (4), it can be inferred that drink-driving is one of the crucial reasons why areas A, C, G, H, and F have higher fatality rates. Hence, in these areas, drink-driving enforcement should be strengthened.

B. PARTIES INVOLVEMENT

The second feature is the number of parties involved in the accidents. It can be seen from Table 4 that when this number is lower than or equal to 4, the fatality rate is 0.986%. While the number is higher than 4, the fatality rate increases to 2.708%. More details about the number and the fatality rate are shown in Figure 8. It can be seen that when the number is larger than one, the fatality rate goes up as the number of parties increases. This is reasonable because multi-vehicle accidents are more likely to be serious traffic accidents, which have a higher probability to cause fatality. Previous studies also agreed that the more parties involved in an accident, the higher the number of people to be involved, which in turn increases the potential fatality rate [44]. Also, in a multi-vehicle accident, those who are not wounded at first and decide to escape at once from their vehicles are still at risk of being hit by other upcoming vehicles. However, the cause of

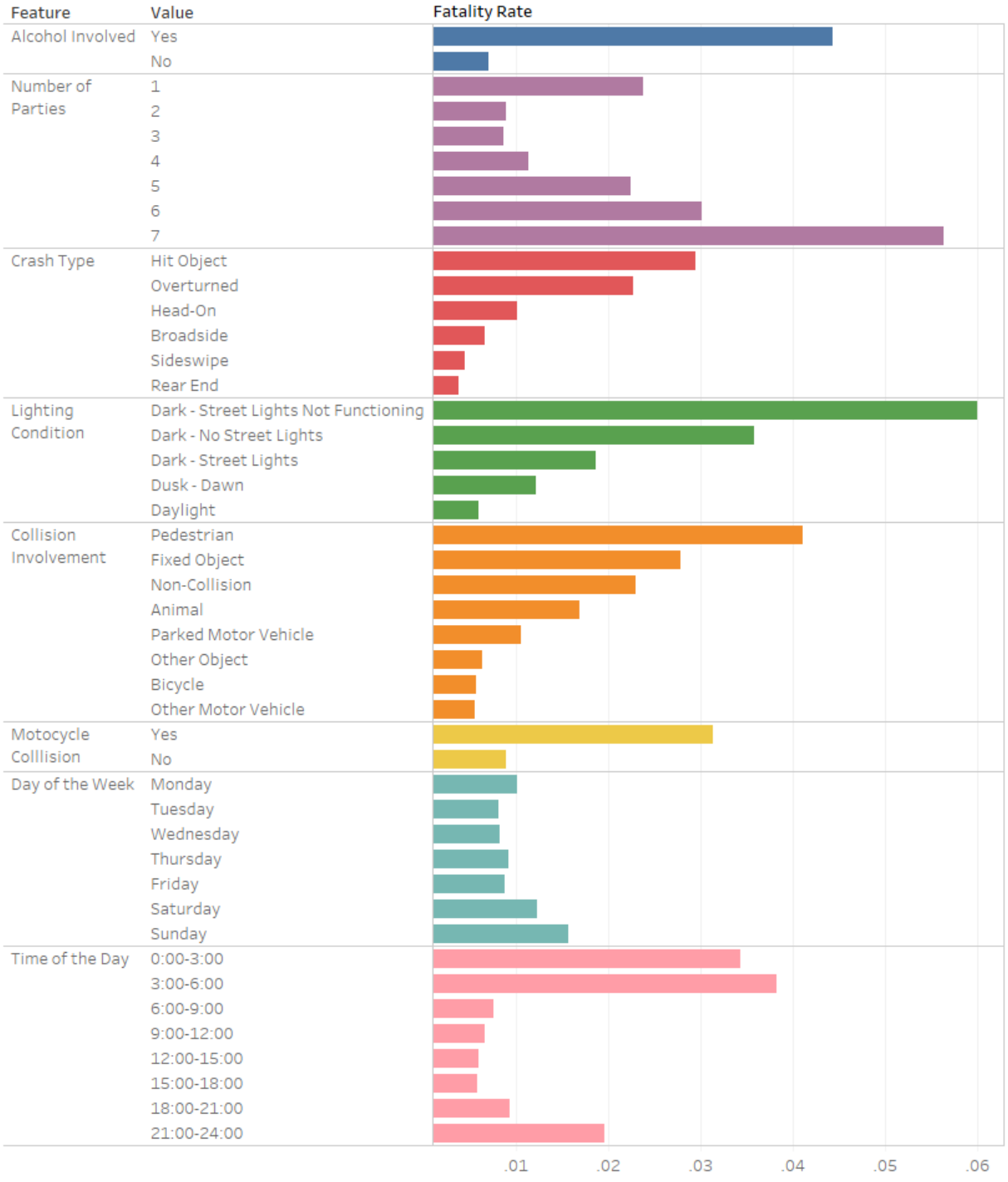


FIGURE 8. Fatality rate under different conditions.

multi-vehicle collisions is often hard to determine. Increasing the road speed limit can be one choice to control the rate of multi-vehicle accidents.

Another thing can be seen from Figure 8 is that when the number of parties equals to one, the fatality rate is also high. Common factors contributing to single-vehicle collisions include excessive speed, driver fatigue, driving under influence and age of vehicle [45], [46]. Environmental and roadway factors can also contribute to single-vehicle accidents. These include inclement weather, falling rocks, poor lighting condition, narrow lanes and shoulders, insufficient curve banking and sharp curves [47], [48]. These can be the reasons why single-vehicle collisions have a high fatality rate.

C. REAR-END COLLISION

Whether the accident is a rear-end collision is the third factor that differentiates the fatal and the non-fatal accidents. Rear-end collisions are considered the most frequently occurring type of traffic accidents [11]. In our data, there are more than 108 thousand rear-end crashes, accounting for around 35.3% of the total accidents. It can be observed from Table 4 that if the type of accidents is rear-ended, the fatality rate is 0.385%. Otherwise, the fatality rate is 1.408%. Fatality rates of different crash types are shown in Figure 8. It can also be observed that the rear-end crashes have the lowest fatality rate while overturned and hit object crashes are the highest. Overturned means the vehicles flip over its side or roof and usually happens when the vehicle makes a high-speed sharp turn. This apparently will lead to a higher fatality rate. Hit object means leaving the roadway and hitting a fixed object alongside the road. Such consequences always result from behaviors like excessive speeds, drowsy driving, drink driving, and reveal a higher fatality rate [11], [49].

D. LIGHTING CONDITION

Lighting condition has been concluded to be one of the most critical factors affecting traffic accidents in previous literature [50], [51]. Poor lighting condition will significantly increase the likelihood of fatal accidents. In can be seen in Table 4 that the fatality rate during the daytime is only 0.604%, but it increases to 2.038% under other lighting conditions. Detailed fatality rates under different lighting conditions are presented in Figure 8. It can be seen that when there is no street light, or the street lights are not functioning, the fatality rates are higher. This is reasonable since the drivers may not see the road condition clearly and make wrong judgments. Besides, when the surrounding is dark, drivers are more likely to be influenced or dazzled by the strong headlights from other vehicles. This is even more dangerous as the drivers may not see the road at all.

Figure 7 (2) shows the distribution of the percentage of accidents happened under poor lighting conditions (no street lights or the lights are not functioning). It can be observed that the numbers in area B, area C, area H and the main roads around area A are higher than other places. Compared with Figure 6 (4), it can be inferred that poor lighting conditions may be one of the reasons of the higher fatality rate in area B, area C, area H and the surrounding areas of A. Therefore, it is suggested that the maintenance of the street lights should be strengthened in these areas.

E. PEDESTRIAN INVOLVEMENT

The fifth factor is the pedestrian involvement. Pedestrians are referred to as one typical kind of vulnerable road users (VRUs), which means they are at higher risk in the traffic compared with other road users [52]. This is because pedestrians do not have sufficient protections when traffic accidents happen. Also, drinking and traffic-rule violation is a severe risk not only to motor drivers, but also to pedestrians, especially when they crossing the roads [55].

It can be observed from Table 4 that for accidents involving pedestrians, the fatality rate reaches 4.547%, while for accidents without the involvement of pedestrians, the rate is only 0.673%. More kinds of fatality rates of accidents between motor vehicles and other parties are shown in Figure 8. It can be seen that the fatality rate of pedestrian accidents is almost eight times higher than motor-motor accidents (Other Motor Vehicle). This could be the reason why INVOLVE_C (motor-motor collisions) stands out in Figure 5 since its relatively low fatality rate differentiates the fatal and non-fatal accidents.

Distribution of the percentage of pedestrian accidents is presented in Figure 7 (3). It indicates that area F, H and G have a higher percentage of pedestrian accidents. This is because that in these areas, the population density are higher. Compared with Figure 6 (4), it can be inferred that pedestrian accidents might be one of the reasons why area F, H, and G have higher fatality rates. Therefore, it is suggested that, in these three areas, traffic operation on pedestrians should be strengthened, for example, reducing the vehicle speed limit, installing more raised crosswalks, adding pedestrian lights, etc. [53]–[55].

F. MOTORCYCLE INVOLVEMENT

Motorcycle accidents are the sixth factor. It can be observed from Table 4 that for motorcycle accidents, the fatality rate is 3.240%, but for accidents without the involvement of motorcycles, the fatality rate decreases to 0.912%. Reasons behind the high fatality rate of motorcycle accidents can be divided into two aspects. On one hand, compared with automobile drivers, motorcycle drivers are not protected by seatbelts or airbags. Therefore, when the accidents happen, they are more likely to be ejected from the seats, and their heads might be directly bumped without protection form any cushions. These will increase the likelihood of fatality. On the other hand, the stability of motorcycles is poor. They are easily influenced by elements like debris, uneven road, wet road surface and other small objects [56].

Figure 7 (4) shows the distribution of the percentage of motorcycle accidents. It can be seen that area C, D and E have higher percentages of motorcycle accidents. These three areas are located in the mountainous areas of Los Angeles County. Steep grades and curves of highways built on mountainous areas pose serious safety threats on passing vehicles, especially motorcycles [57]. Compared with Figure 6 (4), it can be inferred that motorcycle accidents can be one of the causes of the high fatality rates in area C, D and E. Thus, in these areas, traffic control on motorcycles should be strengthened, speed limits should be lowered, and more warning signs should be installed to remind the motorcyclists.

G. DAY OF THE WEEK

Day of the week appears to be an essential factor differentiating the fatal and non-fatal accidents according to our

calculation results. It can be observed from Table 4 that the fatality rate is 0.904% and 1.414% for accidents in weekdays and weekends respectively. Fatality rates from Monday to Sunday within the whole county are presented in Figure 8. It also indicates that the fatality rates in weekends are higher than those on weekdays. This might be because that accidents involving alcohol use, speeding, not using seatbelts, etc., are more likely to happen when people are having fun on weekends [10].

The distribution of the percentage of accidents happened on weekends is presented in Figure 7 (5). It can be seen that the percentage of weekend accidents are higher in area B, D, E, F and H. Compared with Figure 6 (4), it can be inferred that the problems in weekends may be one of the reasons why area B, D, E, F, and H have higher rates of fatality. Therefore, in these areas, traffic control during the weekend should be strengthened.

H. TIME OF THE DAY

Different times within a day have different fatality rates. It can be seen from Table 4 that compared with other time periods, the fatality rate is significantly higher from 0 am to 6 am, reaching 3.724%. Detailed fatality rates of different time periods within a day are shown in Figure 8. It shows that the fatality rate reaches the highest during 3 am and 6 am. Although the traffic volume during this period is not as large as other periods, factors like fatigue driving, speeding, poor lighting conditions, and loose traffic control all could be possible reasons for exacerbating the severity of the accidents after mid-night [58].

Figure 7 (6) presets the distribution of the percentage of accidents happened from 0 am to 6 am. It can be seen that the rate is higher in area F and H. Compared with Figure 6 (4), it can be inferred that one of the reasons of the high fatality rate in area F and H might be the problems during 0 am to 6 am. Therefore, in these two areas, the traffic control should be strengthened during this period.

I. SUMMARY

Based on Table 4, Figure 6, Figure 7 and Figure 8, possible reasons behind the high fatality rate in area A-H are analyzed and discussed. According to the results revealed by our methodology, relevant countermeasures to those areas can be summarized as follows:

- *Area A:* In this area, drink-driving and poor lighting condition are two crucial reasons for the high fatality rate. Therefore, in area A, drink-driving control should be emphasized, more street lights should be installed, and regular maintenance should be strengthened.
- *Area B:* Poor lighting condition is one of the reasons why area B has a high fatality rate. Thus, the local government should increase the budget and improve the lighting condition there. High percentage of weekend accidents is another possible reason for the high fatality rate in area B. Therefore, traffic control such as speed limit should be strengthened in weekends there.
- *Area C:* Similar to area A, drink-driving and poor lighting conditions are two principal causes of the high fatality rate in area C. Furthermore, the high frequency of motorcycle accidents also worsens the road safety there. Hence, in area C, besides the drink driving control and the improvement of lighting infrastructures, proper managements on motorcycles should also be emphasized.
- *Area D and E:* The high fatality rates in area D and E may be caused by motorcycle accidents and weekend accidents. Therefore, in these two areas, the control of motorcycles should be stricter, and the traffic management in weekends need to be enhanced.
- *Area F and H:* Reasons behind the high fatality rate in area F and H include drink-driving, pedestrian accidents, weekend accidents as well as the high accident rate from 0 am to 6 am. In this case, it is suggested that in these two areas, the management for drink-driving control should be stricter. Road infrastructures for pedestrians should be further improved. Traffic control during weekends and the time from 0 am to 6 am should be given more attention. Also, area H also reveals poor lighting conditions, and relevant enhancements should be delivered.
- *Area G:* In area G, drink-driving and pedestrian accidents are the main driving factors of the high fatality rate there. Therefore, in this area, drink-driving control should be improved, and more road infrastructures for pedestrians should be installed.

Note that this study is analyzing the influential factors on fatality rates rather than accident rates. Fatality rates represent the rates of the fatal accidents over all the accidents, while accident rates mean the proportion of accidents over traffic volume. These are two different problems. The former problem assumes that if an accident happens, what are the factors that will likely make it fatal. While the latter one refers to the study of the factors that may lead to an accident from normal driving. Both problems are important and typical in accident analysis [21]. Due to data availability, this study focuses on the grid-based fatality rate. Future work will try to integrate traffic volume data and analyze the influential factors on accident rates.

In addition, driving speed is a potentially important factor that had been mentioned by some literature [11], [55]. However, the speed related factor ''whether the vehicle is driving in an unsafe speed (speeding)'' derived from the ''VIOLCAT'' feature in Table 2 did not rank into top 15 in our experiment in Los Angeles. This means it is not as important as the top 15 in leading an accident into fatal according to our methodology. This is understandable because, on one hand, due to data availability, this speed related factor may not perfectly represent the influence of speed on accident fatality. On the other hand, in many fatal accidents, it is not because that the drivers violated the traffic rules but the drunk pedestrians or other vulnerable road users did. A deeper investigation on the influence of driving speed on fatality rates should be conducted in the future.

VI. CONCLUSION

To conclude, this paper proposed a GIS-based data mining methodology framework to analyze the influential factors of fatal traffic accidents. The idea of this study is firstly using XGBoost to build a binary classification model between fatal and non-fatal accidents. Based on the identified important factors using XGBoost models, the grid-based analysis (or the area analysis) is then used to conduct the spatial analysis on fatality rates. A case study in Los Angeles County was conducted to validate the proposed method. Five commonly seen machine learning algorithms, including MLR, LR, MLP, SVM, and RF, were applied to model the influential features and the traffic fatality. Results showed that XGBoost obtained the highest modeling accuracy. It was then applied to investigate the variable importance of the studied features. Eight factors were found to be the most influential. They were drink-driving, the number of parties involved, rear-end crash, lighting condition, pedestrian involvement, motorcycle involvement, the day of the week and time of the day. Through the grid-based analysis in GIS, their spatial relationships between the fatality rate were analyzed.

A. CONTRIBUTIONS

The strengths and contributions of the study lie in two aspects. First is that we proposed an effective methodology framework in analyzing the influential factors on the fatality of traffic accidents. The methodology not only provided accurate modeling performance compared with traditional linear methods, but also calculated the variable importance more objectively, and then specifically located the relationships between the factors and the target. Such a process can systematically analyze the most influential features on traffic fatalities in a numerical way. Of course, the current experiments only supported the effectiveness of the method in analyzing the accidents and fatalities in Los Angeles, whether the method fits other places and scopes need further studies to verify.

The second contribution is that the case study conducted in Los Angeles County uncovered eight areas with higher traffic fatality rate. According to our numerical analysis, the higher fatality rate in these areas may result from the top influential factors discovered in our model. Based on the spatial analysis, specific practical suggestions on how to improve road safety in these eight areas are provided. These can be useful references for the governments during policy-making.

B. LIMITATIONS AND FUTURE WORK

Still, there are limitations in this paper. Due to the availability of the data, some possible influential factors, such as traffic volume, education, and road width, are not considered for different grids. Also, the authors did not obtain the accident data in more recent years like 2017-2018 with sufficient features and factors, this restricted us from analyzing and comparing the difference and improvements on the traffic fatalities in Los Angeles. Further studies can be expanded to consider those factors and data for a more comprehensive analysis.

REFERENCES

[1] *WHO|10 Facts on Global Road Safety*. Accessed: Oct. 10, 2018. [Online]. Available: http://www.who.int/features/factfiles/roadsafety/en/
[2] (Oct. 2017). *USDOT Releases 2016 Fatal Traffic Crash Data*. [Online]. Available: https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data
[3] Y. Shen, E. Hermans, Q. Bao, T. Brijs, G. Wets, and W. Wang, "International benchmarking of road safety: State of the art," *Transp. Res. C, Emerg. Technol.*, vol. 50, pp. 37–50, Jan. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X14002083. doi: 10.1016/j.trc.2014.07.006.
[4] P. St-Aubin, N. Saunier, and L. Miranda-Moreno, "Large-scale automated proactive road safety analysis using video data," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 363–379, Sep. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X15001485. doi: 10.1016/j.trc.2015.04.007.
[5] S. E. Shladover and C. Nowakowski, "Regulatory challenges for road vehicle automation: Lessons from the California experience," *Transp. Res. A, Policy Pract.*, vol. 122, pp. 125–133, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0965856417300952. doi: 10.1016/j.tra.2017.10.006.
[6] G. Ginzburg, S. Evtiukov, I. Brylev, and S. Volkov, "Reconstruction of road accidents based on braking parameters of category L3 vehicles," *Transp. Res. Procedia*, vol. 20, pp. 212–218, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146517300546. doi: 10.1016/j.trpro.2017.01.054.
[7] C. P. Castañeda and J. G. Villegas, "Analyzing the response to traffic accidents in Medellín, Colombia, with facility location models," *IATSS Res.*, vol. 41, no. 1, pp. 47–56, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0386111216300437. doi: 10.1016/j.iatssr.2016.09.002.
[8] Y. Liu, Z. Li, J. Liu, and H. Patel, "A double standard model for allocating limited emergency medical service vehicle resources ensuring service reliability," *Transp. Res. C, Emerg. Technol.*, vol. 69, pp. 120–133, Aug. 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0968090X16300602. doi: 10.1016/j.trc.2016.05.023.
[9] M. A. Abdel-Aty and A. E. Radwan, "Modeling traffic accident occurrence and involvement," *Accident Anal. Prevention*, vol. 32, no. 5, pp. 633–642, Sep. 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457599000949. doi: 10.1016/S0001-4575(99)00094-9.
[10] E. K. Adanu, A. Hainen, and S. Jones, "Latent class analysis of factors that influence weekday and weekend single-vehicle crash severities," *Accident Anal. Prevention*, vol. 113, pp. 187–192, Apr. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457518300411. doi: 10.1016/j.aap.2018.01.035.
[11] S. A. Mohamed, K. Mohamed, and H. A. Al-Harthi, "Investigating factors affecting the occurrence and severity of rear-end crashes," *Transp. Res. Procedia*, vol. 25, pp. 2098–2107, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S235214651730710X. doi: 10.1016/j.trpro.2017.05.403.
[12] F. Netjasov, D. Crnogorac, and G. Pavlović, "Potential safety occurrences as indicators of air traffic management safety performance: A network based simulation model," *Transp. Res. C, Emerg. Technol.*, vol. 102, pp. 490–508, May 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X1831060X. doi: 10.1016/j.trc.2019.03.026.
[13] J. Lee, J. Chae, T. Yoon, and H. Yang, "Traffic accident severity analysis with rain-related factors using structural equation modeling—A case study of Seoul city," *Accident Anal. Prevention*, vol. 112, pp. 1–10, Mar. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457517304517. doi: 10.1016/j.aap.2017.12.013.
[14] W. H. Schneider and P. T. Savolainen, "Comparison of severity of motorcyclist injury by crash types," *Transp. Res. Rec.*, vol. 2265, no. 1, pp. 70–80, Jan. 2011. [Online]. Available: http://journals.sagepub.com/doi/10.3141/2265-08. doi: 10.3141/2265-08.
[15] S. Hashimoto, S. Yoshiki, R. Saeki, Y. Mimura, R. Ando, and S. Nanba, "Development and application of traffic accident density estimation models using kernel density estimation," *J. Traffic Transp. Eng.*, vol. 3, no. 3, pp. 262–270, Jun. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2095756415305808. doi: 10.1016/j.jtte.2016.01.005.

[16] H. A. Aziz, S. V. Ukkusuri, and S. Hasan, "Exploring the determinants of pedestrian–vehicle crash severity in New York city," *Accident Anal. Prevention*, vol. 50, pp. 1298–1309, Jan. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0001457512003533. doi: 10.1016/j.aap.2012.09.034.

[17] B. J. Russo and P. T. Savolainen, "A comparison of freeway median crash frequency, severity, and barrier strike outcomes by median barrier type," *Accident Anal. Prevention*, vol. 117, pp. 216–224, Aug. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457518301763. doi: 10.1016/j.aap.2018.04.023.

[18] F. Dapilah, B. Y. Guba, and E. Owusu-Sekyere, "Motorcyclist characteristics and traffic behaviour in urban northern ghana: Implications for road traffic accidents," *J. Transp. Health*, vol. 4, pp. 237–245, Mar. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214140516000232. doi: 10.1016/j.jth.2016.03.001

[19] A. Karimi and E. Kashi, "Investigating the effect of geometric parameters influencing safety promotion and accident reduction (Case study: Bojnurd-Golestan national park road)," *Cogent Eng.*, vol. 5, no. 1, 2018, Art. no. 1525812. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/23311916.2018.1525812

[20] D. Onozuka, K. Nishimura, and A. Hagihara, "Full moon and traffic accident-related emergency ambulance transport: A nationwide case-crossover study," *Sci. Total Environ.*, vol. 644, pp. 801–805, Dec. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048969718325361. doi: 10.1016/j.scitotenv.2018.07.053.

[21] L. Mussone, M. Bassani, and P. Masci, "Analysis of factors affecting the severity of crashes in urban road intersections," *Accident Anal. Prevention*, vol. 103, pp. 112–122, Jun. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457517301355. doi: 10.1016/j.aap.2017.04.007.

[22] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Anal. Prevention*, vol. 45, pp. 478–486, Mar. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457511002363. doi: 10.1016/j.aap.2011.08.016.

[23] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," Apr. 2018, *arXiv:1804.06620*. [Online]. Available: https://arxiv.org/abs/1804.06620

[24] (Aug. 2018). *Geographic Information System*. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Geographic_information_system&oldid=853176113

[25] J. Ma and J. C. Cheng, "Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology," *Appl. Energy*, vol. 183, pp. 182–192, Dec. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306261916311679

[26] S. Hong, J. Heo, and A. P. Vonderohe, "Simulation-based approach for uncertainty assessment: Integrating GPS and GIS," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 125–137, Nov. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0968090X13001721. doi: 10.1016/j.trc.2013.08.008.

[27] J. H. Kazmi and S. Zubair, "Estimation of vehicle damage cost involved in road traffic accidents in karachi, Pakistan: A geospatial perspective," *Procedia Eng.*, vol. 77, pp. 70–78, Jan. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877705814009850. doi: 10.1016/j.proeng.2014.07.008.

[28] M. A. M. Din, S. Paramasivam, N. M. Tarmizi, and A. M. Samad, "The use of geographical information system in the assessment of level of service of transit systems in Kuala Lumpur," *Procedia Social Behav. Sci.*, vol. 222, pp. 816–826, Jun. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877042816302567. doi: 10.1016/j.sbspro.2016.05.181.

[29] A. Zangeneh, F. Najafi, S. Karimi, S. Saeidi, and N. Izadi, "Spatial-temporal cluster analysis of mortality from road traffic injuries using geographic information systems in West of Iran during 2009–2014," *J. Forensic Legal Med.*, vol. 55, pp. 15–22, Apr. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1752928X18300258. doi: 10.1016/j.jflm.2018.02.009.

[30] L. Zhang and C. Zhan, "Machine learning in rock facies classification: An application of XGBoost," in *Proc. Int. Geophys. Conf.*, Qingdao, China, May 2017, pp. 1371–1374. [Online]. Available: http://library.seg.org/doi/10.1190/IGC2017-351. doi: 10.1190/IGC2017-351.

[31] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain Inf.*, vol. 4, no. 3, pp. 159–169, Sep. 2017. [Online]. Available: http://link.springer.com/10.1007/s40708-017-0065-7. doi: 10.1007/s40708-017-0065-7.

[32] J. Ma and J. C. P. Cheng, "Identification of the numerical patterns behind the leading counties in the U.S. local green building markets using data mining," *J. Cleaner Prod.*, vol. 151, pp. 406–418, May 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959652617305176

[33] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, p. 1168, Aug. 2017. [Online]. Available: http://www.mdpi.com/1996-1073/10/8/1168. doi: 10.3390/en10081168.

[34] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167947301000652. doi: 10.1016/S0167-9473(01)00065-2.

[35] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. San Francisco, CA, USA, 2016, pp. 785–794. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2939672.2939785. doi: 10.1145/2939672.2939785.

[36] 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," *Appl. Soft Comput.*, vol. 80, pp. 57–79, Jul. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494619301358. doi: 10.1016/j.asoc.2019.03.017.

[37] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, Jan. 2006.

[38] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing* (Springer Topics in Signal Processing), I. Cohen, Y. Huang, J. Chen, and J. Benesty, Eds. Berlin, Heidelberg: Springer, 2009, pp. 1–4. doi: 10.1007/978-3-642-00296-0_5.

[39] A. Y.-C. Liu, "The effect of oversampling and undersampling on classifying imbalanced text datasets," M.S. thesis, Univ. Texas Austin, Austin, TX, UAS, 2004.

[40] J. C. P. Cheng and L. J. Ma, "A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects," *Building Environ.*, vol. 93, pp. 349–361, Nov. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132315300676

[41] J. Ma and J. C. P. Cheng, "Data-driven study on the achievement of LEED credits using percentage of average score and association rule analysis," *Building Environ.*, vol. 98, pp. 121–132, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132316300051

[42] M. A. Jun and J. C. P. Cheng, "Selection of target LEED credits based on project information and climatic factors using data mining techniques," *Adv. Eng. Inform.*, vol. 32, pp. 224–236, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474034616301124

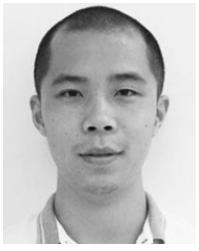
[43] B. Assemi and M. Hickman, "Relationship between heavy vehicle periodic inspections, crash contributing factors and crash severity," *Transp. Res. A, Policy Pract.*, vol. 113, pp. 441–459, Jul. 2018.[Online]. Available: http://www.sciencedirect.com/science/article/pii/S096585641630502X. doi: 10.1016/j.tra.2018.04.018.

[44] K. K. Yau, H. Lo, and S. H. Fung, "Multiple-vehicle traffic accidents in Hong Kong," *Accident Anal. Prevention*, vol. 38, no. 6, pp. 1157–1161, Nov. 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0001457506000807. doi: 10.1016/j.aap.2006.05.002.

[45] W. Haddon and V. A. Bradess, "Alcohol in the single vehicle fatal accident: Experience of westchester county, New York," *JAMA*, vol. 169, no. 14, pp. 1587–1593, Apr. 1959. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/325699. doi: 10.1001/jama.1959.03000310039009.

[46] J.-K. Kim, G. F. Ulfarsson, S. Kim, and V. N. Shankar, "Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender," *Accident Anal. Prevention*, vol. 50, pp. 1073–1081, Jan. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457512002990. doi: 10.1016/j.aap.2012.08.011.

[47] V. Shankar and F. Mannering, "An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity," *J. Saf. Res.*, vol. 27, no. 3, pp. 183–194, Sep. 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022437596000102. doi: 10.1016/0022-4375(96)00010-2.
[48] K. K. Yau, "Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong," *Accident Anal. Prevention*, vol. 36, no. 3, pp. 333–340, May 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457503000125. doi: 10.1016/S0001-4575(03)00012-5.
[49] M. Hiramatsu, H. Obara, and K. Umezaki, "Broadside collision scenarios for different types of pre-crash driving patterns," *JSAE Review*, vol. 24, no. 3, pp. 327–334, Jul. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S038943040300033X. doi: 10.1016/S0389-4304(03)00033-X.
[50] M. Jalayer, R. Shabanpour, M. Pour-Rouholamin, N. Golshani, and H. Zhou, "Wrong-way driving crashes: A random-parameters ordered probit analysis of injury severity," *Accident Anal. Prevention*, vol. 117, pp. 128–135, Aug. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0001457518301635. doi: 10.1016/j.aap.2018.04.019.
[51] R. O. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accident Anal. Prevention*, vol. 88, pp. 37–51, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457515301548. doi: 10.1016/j.aap.2015.12.003.
[52] M. Vilaça, N. Silva, and M. C. Coelho, "Statistical analysis of the occurrence and severity of crashes involving vulnerable road users," *Transp. Res. Procedia*, vol. 27, pp. 1113–1120, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146517310104. doi: 10.1016/j.trpro.2017.12.113.
[53] 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," *Saf. Sci.* , vol. 54, pp. 27–37, Apr. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925753512002664. doi: 10.1016/j.ssci.2012.11.001.
[54] S. Jung, X. Qin, and C. Oh, "Improving strategic policies for pedestrian safety enhancement using classification tree modeling," *Transp. Res. A, Policy Pract.*, vol. 85, pp. 53–64, Mar. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0965856416000021. doi: 10.1016/j.tra.2016.01.002.
[55] M. Soilán, B. Riveiro, A. Sánchez-Rodríguez, and P. Arias, "Safety assessment on pedestrian crossing environments using MLS data," *Accident Anal. Prevention*, vol. 111, pp. 328–337, Feb. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0001457517304475. doi: 10.1016/j.aap.2017.12.009.
[56] Z. Hui, Y. Guang-yu, L. Sheng-xiong, Y. Zhi-yong, W. Zheng-guo, H. Wei, Y. Yong-min, C. Rong, and L. Gui-e, "Analysis of 86 fatal motorcycle frontal crashes in Chongqing, China," *Chin. J. Traumatol.*, vol. 15, no. 3, pp. 170–174, Jun. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1008127515302984. doi: 10.3760/cma.j.issn.1008-1275.2012.03.009.
[57] F. Chen and S. Chen, "Differences in injury severity of accidents on mountainous highways and non-mountainous highways," *Procedia Social Behav. Sci.*, vol. 96, pp. 1868–1879, Nov. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877042813023380. doi: 10.1016/j.sbspro.2013.08.212.
[58] G. Zhang, K. K. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Anal. Prevention*, vol. 87, pp. 34–42, Feb. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0001457515301159. doi: 10.1016/j.aap.2015.10.033.



JUN MA received the Ph.D. degree from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2016. He is currently the Chief Research Officer with the Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include smart city, urban computing, data mining, and artificial intelligence.

YUEXIONG DING received the B.S. degree from the School of Pharmaceutical Information Engineering, Guangdong Pharmaceutical University, Guangdong, China, in 2017. He is currently pursuing the M.S. degree with the School of Technology, Shantou University, Guangdong. He is also a Researcher with the Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong. His research interests include data mining, artificial intelligence, and smart city.

JACK C. P. CHENG received the Ph.D. degree from Standford University, Stanford, CA, USA. He is currently the Associate Director of the GREAT Smart Cities Institute, Director of BIM Lab, and an Associate Professor with The Hong Kong University of Science and Technology. His research interests include Construction IT and knowledge management, building information modeling (BIM), the Internet of Things (IoT), computer vision and AI, reality capture, green buildings and sustainable construction, and smart cities.

YI TAN received the M.Sc. degree from the University of Southern California, in 2014, and the Ph.D. degree from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology (HKUST), in 2018. He was an Assistant Professor of construction management with Shenzhen University. His research interests include building information modeling (BIM), digital twin, smart construction and smart operation, and maintenance (O&M) of both buildings and infrastructures.

VINCENT J. L. GAN received the M.Sc. and Ph.D. degrees from the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology (HKUST), in 2016, where he was a Research Assistant Professor of civil engineering. His research interests include tall building, optimum design, performance optimization, energy efficiency, indoor human comfort, and construction informatics.

JINGCHENG ZHANG received the M.S. degree from the School of Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2014. He is currently a Researcher with Shenzhen Mariocode Science and Technology Company Ltd., Shenzhen, China. His research interests include the Internet of Things, smart city, and urban computing.