Journal of Safety Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Safety Research

journal homepage: www.elsevier.com/locate/jsr

Analysis of the severity of vehicle-bicycle crashes with data mining techniques

Siying Zhu

School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 22 March 2020
Received in revised form 19 July 2020
Accepted 23 November 2020
Available online xxxx

Keywords:
Safety
Severity
Vehicle-bicycle crashes
Integrated data mining framework
Gradient boosting algorithm

ABSTRACT

Although cycling is increasingly being promoted for transportation, the safety concern of bicyclists is one of the major impediments to their adoption. A thorough investigation on the contributing factors to fatalities and injuries involving bicyclist is in need. This paper designs an integrated data mining framework to determine the significant factors which contribute to the severity of vehicle-bicycle crashes based on the crash dataset of Victorian, Australia (2013–2018). The framework integrates imbalanced data resampling, learning-based feature extraction with gradient boosting algorithm and marginal effect analysis. The top ten significant predictors of the severity of vehicle-bicycle crashes are extracted which gives an area under ROC curve (AUC) value of 0.8236 and computing time as 37.8 s. The findings provide insights for understanding and developing countermeasures or policy initiatives to reduce severe vehicle-bicycle crashes.

1. Introduction

Cycling is becoming increasingly popular in recent years as the mode of transportation is healthy, low-cost and environmentally friendly. Many transportation decision makers aim to make cycling a lifestyle in order to support the car-lite vision and enhance the overall livability of the city or country. However, the safety concern on cycling is one of the major impediments to its adoption, as bicyclists are more vulnerable in comparison with auto-mobile occupants. To achieve the long-term goal of alleviating congestion and pollution by engaging more bicyclists, it is necessary to solve commuters' fear of being involved in a crash (Kaplan & Giacomo Prato, 2015).

Enhancing the safety level of bicyclists is a different challenge compared with motorized traffic which has been well studied in the literature. The crashes involving bicyclists are rare and often severe; bicyclists exposure is different from vehicle exposure which is difficult to quantify; and the crash trends of bicyclists are quite distinctive which depend on land use, existing bicycle infrastructure, socio-economic factors, etc. Raihan, Alluri, Wu, and Gan (2019). A thorough investigation on different characteristics contributing to fatalities and injuries involving bicyclists is necessary.

For road safety analysis, the application of non-parametric and data mining technique becomes increasingly popular in recent years, which refers to the analytic process designed to explore big data, searching for structures, commonalities, and hidden patterns or rules (Prati, Pietrantoni, & Fraboni, 2017; Han, Pei, & Kamber, 2011). The data mining techniques can handle large and complicated datasets with relatively short data preparation time and provides satisfiable accuracy (Ding, Chen, & Jiao, 2018).

This paper designs an integrated data mining framework to determine the significant factors which contribute to the severity of vehicle-bicycle crashes based on the crash dataset of Victorian, Australia (2013–2018). The framework integrates imbalanced data resampling, learning-based feature extraction with gradient boosting algorithm and marginal effect analysis to determine the significant contributing factors to vehicle-bicycle crashes.

Specifically, in terms of traffic safety, the main contributions of this paper are elaborated as follows: This paper is dedicated to vehicle-bicycle crash severity modeling to address the safety concerns on bicyclists who are vulnerable road users. In vehicle-bicycle crash severity analysis, the class imbalance issue of the crash dataset exists as the proportion of fatal or severe crashes is relatively small (Prati et al., 2017). The problem can be handled by the imbalanced data resampling process in the integrated data mining framework. The complexity of crash dataset can also be addressed with the learning-based feature extraction process in an iterative manner, in order to determine the most significant contributing factors to the severity of vehicle-bicycle crashes and cater for the trade-off between computation time and model performance. Moreover, the vehicle-bicycle crash dataset in this paper

E-mail address: siying001@e.ntu.edu.sg

https://doi.org/10.1016/j.jsr.2020.11.011
0022-4375/© 2020 Published by Elsevier Ltd.

contains a large number of discrete variables. The large number of categories can be handled by gradient boosting algorithm, relying on no strict statistical assumption (Saha, Alluri, & Gan, 2015; Ding, Cao, & Næss, 2018; Zheng, Lu, & Lantz, 2018). The impact of the most significant contributing factors on the severity of vehicle-bicycle crashes are explained with the marginal effect analysis, and the result from the integrated data mining framework can provide some implications for policies and counter-measures for fatal and serious vehicle-bicycle crashes.

2. Related work

Research on crash severity modeling has been conducted to identify the most significant predictors of the severity of vehicle-bicycle crashes. Klop and Khattak (1999) identified the contributing factors to bicycle crash severity on two-lane, undivided roadways with ordered probit model. Kim, Kim, Ulfarsson, and Porrello (2007) analyzed the determinants of bicyclists injury severity in vehicle-bicycle crashes with multinomial logit model. Yan, Ma, Huang, Abdel-Aty, and Wu (2011), Bahrololoom, Moridpour, and Tay (2016) analyzed the interrelationship of irregular maneuver, crash patterns, etc., and cyclist injury severity with binary logit model. Kaplan, Vavatsoulas, and Prato (2014), Bahrololoom, Moridpour, Tay, and Sobhani (2017) analyzed the determinants of cyclist injury severity level with generalized ordered probit model and generalized ordered logit model. Helak et al. (2017) utilized univariate and multiple regression analyses to study the influence of bike lanes, alcohol, lighting, speed, and helmet on the injury severity of bicyclists. Robartes and Chen (2017) identified the factors that affect cyclist injury severity in the case of single bicycle-single vehicle crashes, in consideration of the cyclist, auto-mobile driver, vehicle, environmental and roadway characteristics with an ordered probit model. Behnood and Mannering (2017) analyzed the factors that significantly affect bicycle injury severities in vehicle-bicycle crashes, utilizing a random parameters multinomial logit model with heterogeneity in means and variances. Bahrololoom, Young, and Logan (2018) investigated the effect of factors related to the three pillars of the Safe System approach including 'safe roads and roadsides', 'safe speeds' and 'safe road users' on bicycle crash severity with random parameter binary logit model. Yasmin and Eluru (2018) analyzed the total crash count and crash proportion by various crash severity levels based on a joint negative binomial-ordered logit fractional split econometric model framework. Sivasankaran and Balasubramanian (2020) applied latent class clustering algorithm to analyse the crash severity of vehicle-bicycle crashes for each cluster. Liu, Khattak, Li, Nie, and Ling (2020) investigated bicyclists injury severity with geographically weighted ordinal logistic regression model to address the spatial heterogeneity.

In particular, research has been conducted for crash severity modeling of vehicle-bicycle crashes on specific types of infrastructure, such as bike lanes and intersections. Klassen, El-Basyouny, and Islam (2014) investigated the severity level of vehicle-bicycle intersection-related and mid-block-related crashes with spatial mixed logit model. Wall et al. (2016) evaluated the influence of sharrow, painted bicycle lane and physically protected path on bicyclist injury severity with negative binomial model. Moore, Schneider, Savolainen, and Farzaneh (2011) examined bicyclists injury severity in vehicle-bicycle crashes at intersection and non-intersection respectively with mixed logit model. Wang, Lu, and Lu (2015) analyzed the contributing factors bicyclists' injury severity level in vehicle-bicycle crashes at unsignalized intersections with partial proportional odds model. Stipancic, Zangenehpour, Miranda-Moreno, Saunier, and Granié (2016) investigated the contributing factors to the severity of vehicle-bicycle conflicts at urban intersections with ordered logit model based on video data and post-encroachment time, taking the gender differences into account. Asgarzadeh, Verma, Mekary, Courtney, and Christiani (2017) analyzed the effects of intersection and street design on vehicle-bicycle crash severity with multivariate log-binomial regression model, and the results indicated that non-orthogonal intersections and non-intersection segments are associated with higher crash severity. Bahrololoom, Young, and Logan (2018), Bahrololoom, Young, and Logan (2018) investigated the impact of kinetic energy on crash severity of bicyclists at intersections. Rash-ha Wahi, Haworth, Debnath, and King (2018) studied the effects of various traffic control types at intersection with the application of separate mixed logit models for bicyclist injury severity.

Ordinal regression models has been commonly applied to formulate the crash severity model since the injury outcomes are ordinal from no injury to fatal. More recently, to address the limitation of the assumption that all parameters estimated in the model are constant across observations and the heterogeneity of the crash outcomes, some multi-nomial logit models and mixed logit models have been applied for crash severity modeling (Li, Ma, Zhu, Zeng, & Wang, 2018). Regression models relies on strict statistical assumptions, for example, linearity in modeling the relation, which can hardly be satisfied in most crash circumstances. Moreover, the performance of the regression model is poor when handling mass complicated crash data with many discrete variables or variables with a large number of categories satisfactorily (Prati et al., 2017; Li et al., 2018; Ding et al., 2018). To overcome the shortcomings of statistical models, data mining techniques which examines the pre-existing large database have been applied for crash severity modeling. Prati et al. (2017) applied the CHi-squared Automatic Interaction Detection (CHAID) decision tree and Bayesian network analysis to predict the severity of bicycle crashes corresponding to the factors related to crash characteristics. The Bayesian network analysis was further applied to identify the most significant predictors. However, when the complexity of the crash dataset gets larger, feature extraction process is necessary to be applied to address trade-off between computing time and model performance.

Based on the literature review, the research gaps in terms of traffic safety are summarized as follows: (1) In the field of traffic safety, limited research was dedicated to crash severity modeling for vehicle-bicycle crashes in comparison with vehicle-vehicle crashes. (2) The crash severity levels in the vehicle-bicycle crash dataset is highly imbalanced, affecting the performance of crash severity classification model. (3) The data mining techniques which can overcome some shortcomings of statistical models such as the reliance on strict statistical assumptions have rarely been used for the analysis of crash severity of vehicle-bicycle crashes. (4) The learning-based feature extraction process has not been applied in the literature to address the complexity of the crash dataset.

This paper aims to address the research gap by determining the significant factors which contribute to the severity of vehicle-bicycle crashes with a data mining framework, integrating imbalanced data resampling, learning-based feature extraction and marginal effect analysis. The framework introduced in this paper uses gradient boosting as the key algorithm for feature extraction, which can handle different types of predictor attributes, require little data preprocessing effort, and can fit complex nonlinear relationship (Elith, Leathwick, & Hastie, 2008; Zhang & Haghani, 2015; Zheng et al., 2018).

Table 1
Descriptive statistics.

Category	Variable	Count	%
Severity	Fatal accident	33	0.53
	Serious injury accident	1467	23.52
	Other injury accident	4737	75.95
	Accident type		
Collision type	Right through	994	16.01
	Cross traffic	793	12.78
	Vehicle strikes door of parked/stationary vehicle	714	11.5
	Left turn side wipe	458	7.38
	Vehicle off footpath strikes another vehicle while emerging from driveway	413	6.65
	Vehicle strikes another vehicle while emerging from driveway	408	6.57
	Left near	359	5.78
	Right near	308	4.96
	Rear end	279	4.49
	Lane side swipe	243	3.91
	Out of control on carriageway	163	2.63
	Entering parking	108	1.74
	Right far	84	1.35
	Y turn	71	1.14
	Right turn side swipe	65	1.05
	Head on	63	1.01
	...	...	...
Alcohol related	No	6217	99.68
	Yes	20	0.32
	Time factor		
Year	2013	584	9.36
	2014	1298	20.81
	2015	1221	19.58
	2016	1104	17.70
	2017	1054	16.90
	2018	976	15.65
Month	January	442	7.09
	February	547	8.77
	March	683	10.95
	April	499	8.00
	May	517	8.29
	June	397	6.37
	July	477	7.65
	August	481	7.71
	September	437	7.01
	October	626	10.04
	November	563	9.03
	December	568	9.11
Day of week	Monday	883	14.39
	Tuesday	1044	17.02
	Wednesday	1081	17.62
	Thursday	1081	17.62
	Friday	881	14.36
	Saturday	582	9.49
	Sunday	583	9.50
	Vehicle characteristics		
No. of vehicles involved	2	5987	95.99
	3	226	3.62
	4	11	0.18
	5	5	0.08
	6	4	0.06
	7	2	0.03
	8	1	0.02
	14	1	0.02
No. of heavy vehicles involved	0	6143	98.49
	1	94	1.51
No. of passenger vehicle involved	0	224	3.59
	1	5878	94.24
	2	126	2.02
	3	3	0.05
	4	1	0.02
	5	3	0.05
	6	1	0.02
	13	1	0.02
No. of public vehicles	0	6180	99.09
	1	57	0.91
Involve vehicle run off-road	No	14210	95.87
	Yes	206	1.43
	Environment condition characteristics		
Light condition	Dark no street lights	40	0.64

(*continued on next page*)

Table 1 (*continued*)

Category	Variable	Count	%
	Dark street lights off	3	0.05
	Dark street lights on	647	10.37
	Dark street lights unknown	83	1.33
	Day	4392	70.42
	Dusk/Dawn	865	13.87
Road Geometry	Cross intersection	1914	25.87
	Multiple intersection	137	2.20
	Not at intersection	2237	35.87
	T intersection	1922	30.82
	Y intersection	16	0.26
	Dead end	2	0.03
Speed zone	110 km/h	3	0.05
Category	Variable	Count	%
	100 km/h	93	1.49
	90 km/h	7	0.11
	80 km/h	249	3.99
	75 km/h	1	0.02
	70 km/h	195	3.13
	60 km/h	2391	38.34
	50 km/h	1579	25.32
	40 km/h	1149	18.42
	30 km/h	17	0.27
	Campus ground or off-road	59	10.95
	Other speed limit	17	0.27
Node type	Intersection	3823	61.31
	Non-intersection	2387	38.28
	Off-road	26	0.42
Urbanized area	Melbourne Urban	4983	79.89
	Small cities	315	5.05
	Melbourne CBD	303	4.86
	Large provincial cities	276	4.43
	Rural Victoria	197	3.16
	Towns	148	2.37
	Small towns	15	0.24
Road classification	Arterial highway	657	10.75
	Arterial other	2378	39.91
	Freeway	48	0.79
	Local road	3028	49.55
Road condition by the Statewide Route Numbering Scheme	A	80	12.82
	B	125	20.03
	C	369	59.13
	M	50	8.01
Crash occur on a divided portion of road	Divided	4324	70.76
	Undivided	1787	29.24
Where crash occur	Metro region	5426	87
	Country region	811	15.86
	Demographics		
Hit-and-run	No	5911	94.77
	Yes	326	5.23
No. of males involved	0	655	10.50
	1	2772	44.44
	2	2580	41.37
	3	193	3.09
	4	22	0.35
	5	14	0.22
	7	1	0.02
No. of females involved	0	2705	43.37
	1	2750	44.09
	2	724	11.61
	3	41	0.66
	4	10	0.16
	5	5	0.08
	6	1	0.02
	7	1	0.02
No. bicyclists involved	1	6146	98.54
	2	79	1.27
	3	6	0.10
	4	4	0.06
	5	1	0.02
	6	1	0.02
No. of pedestrians involved	0	6227	99.84
	1	9	0.14
	4	1	0.02
No. of drivers involved	1	6127	98.24

Table 1 (*continued*)

Category	Variable	Count	%
	2	103	1.65
	3	3	0.05
	4	1	0.02
	5	2	0.03
	7	1	0.02
No. of vehicle passengers involved	0	5601	89.80
	1	509	8.16
	2	78	1.25
	3	31	0.50
	4	15	0.24
	5	3	0.05
No. of 5–12 year old cyclists involved	0	6043	96.89
	1	191	1
	2	3	0.05
No. of 13–18 year old cyclists involved	0	6236	99.98
	1	1	0.02
No. of 65 years and older pedestrians involved in the crash	0	5929	95.06
	1	308	
No. of 65 years and older drivers involved	0	5929	95.06
	1	308	
No. of 18–25 year old young drivers involved	0	5448	87.35
	1	787	12.62
	2	2	0.03
Unlicensed driver	0 (no)	6155	98.69
	1 (yes)	82	1.31

3. Methodology

3.1. Integrated data mining framework

An integrated data mining framework is designed to extract the key crash-related features and predict vehicle-bicycle crash severity level, which integrates imbalanced data resampling, learning-based feature extraction, and marginal effect analysis. The framework is illustrated in Algorithm 1.

Algorithm 1: Integrated data mining framework

Input: Vehicle-bicycle crash severity dataset D
Output: Vehicle-bicycle crashes severity level prediction
Step 1: Resample imbalanced dataset
(a) Synthetic Minority Over-sampling Technique (SMOTE)
(b) Resampled crash dataset $D\prime$
Step 2: Learning-based feature selection
(a) Train gradient boosting model based on $D\prime$
(b) Determine relative feature importance
(c) Recursive feature elimination
(d) Dataset with key vehicle-bicycle crash features
Step 3: Marginal effect analysis

As the vehicle-bicycle crash dataset is with highly imbalanced classes, data resampling process is applied. For learning-based feature selection, gradient boosting algorithm which is an ensemble learning technique is applied as the key algorithm. The key features for bicycle-vehicle severity prediction is extracted recursively, in consideration of the trade-off between AUC (area under the receiver operating characteristic curve) value and computing time. The receiver operating characteristics (ROC) is a probability curve which demonstrates a comparison of two operating characteristics, namely, specificity and sensitivity, as the threshold changes (Beshah & Hill, 2010). Then, the AUC value serves as a measure of separability, quantifying the overall capability of the model in distinguishing between classes (Narkhede, 2018; Bradley, 1997). The AUC value of 0.5 represents an entirely random test, while the AUC value of 1 represents a perfect classification test. The features extracted are further included in marginal effect analysis to investigate their impact on the vehicle-bicycle crash occurrences.

3.2. Resample imbalanced data

The vehicle-bicycle crash dataset is imbalanced as the classes of crash severity levels are not approximately equally represented. Although predictive accuracy is commonly applied to evaluate the performance of machine learning techniques, it is not suitable when the dataset is imbalanced and the cost of different errors varies greatly. In such situations, the Receiver Operating Characteristic (ROC) curve typically serves as the performance measurement to optimize, which is a probability curve calculated by the true positive rate on the y-axis against the false positive rate on the x-axis. Every point on the ROC curve corresponds to a pair of sensitivity and specificity values based on specific decision threshold. The area under the ROC curve (AUC) serves as a measure of separability between the two classes of crash severity in this paper, such that the value closer to 1 indicates a better separability of model. This paper applies Synthetic Minority Over-sampling Technique (SMOTE) to resample the original crash dataset, which synthesises new minority instances between existing minority instances instead of over-sampling to replacement (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Moreover, the majority instances are also under-sampled, whose output is a more balanced dataset.

3.3. Gradient boosting algorithm & learning-based feature selection

The gradient boosting algorithm is a type of ensemble learning technique, which sequentially fits a simple parameterized function or base learner into current 'pseudo'-residuals by least squares in each iteration in order to construct additive regression models. At each step, the weightage of observations is adjusted as the sub-

sequent predictors learn from the mistakes of previous predictors. Regarding the model values at each training data point evaluated in the current step, the pseudo-residuals are the gradient of the loss functions being minimized (Friedman, 2002).

Let N be the total number instances in the dataset, and M be the total number of trees to be generated. Let x be the feature vector with a set of predictors and $F(x)$ be an approximation function of the target variable (i.e. severity level of vehicle-bicycle crashes). The gradient boosting algorithm estimates the function $F(x)$ as an additive expansion based on the base learner function $b(x; a_m)$ (Ding et al., 2018; De'Ath, 2007; Saha et al., 2015; Chung, 2013; Zhang & Haghani, 2015):

$$F(x) = \sum_{m=1}^{M} F_m(x) = \sum_{m=1}^{M} \lambda_m b(x; a_m) \quad (1)$$

$$b(x; a_m) = \sum_{d=1}^{D} \gamma_{dm} I(x \in R_{dm}) \quad (2)$$

where each decision tree m divides the input space into D disjoint regions $R_{1m}, \ldots, R_{Dm}$ and predicts a constant value γ_{dm} for each region R_{dm};

$$I(x \in R_{dm}) = \begin{cases} 1, & \text{if } x \in R_{dm} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

a_m represents the mean of split locations and the leaf node for each splitting variable in tree m; λ_m represents weights given to the nodes of decision tree and determines how predictions from the individual decision trees are combined. λ_m is calculated by minimizing a specified loss function, which is a squared error function:

$$L(y, F(x)) = (y - F(x))^2 \quad (4)$$

The gradient boosting model is built in a stage-wise fashion, which is updated by minimizing the expected value of the loss function. To avoid over-fitting and improve accuracy, the learning rate or shrinkage, is used to scale the contribution of each base tree learner by introducing a factor of $\varepsilon (0 < \varepsilon \leqslant 1)$ as below (Ding et al., 2018; Friedman, 2002):

$$F_m(x) = F_{m-1}(x) + \varepsilon \cdot \lambda_m b(x; a_m) \quad (5)$$

A hybrid two-step learning-based feature selection model is applied to select the key features for vehicle-bicycle crash severity modelling, which is summarized in Algorithm 2. The procedure firstly train and tune the model to rank the feature importance, then the process is permuted to determine an optimal subset of features with Recursive Feature Elimination (RFE), in consideration of the trade-off between the area under ROC curve and computing time.

Algorithm 2: Learning-based feature selection

Input: Crash severity dataset with the set of all predictors as S
Output: Model corresponding to subset S^* of predictors
\\Feature importance ranking:
Train the model with all predictors
Calculate model performance
Rank predictors according to feature importance
\\RFE:
for Subset $S_i \in S$ $(i = 1, ..., |S|)$ **do**
  Keep i most important predictors
  Train the model based on the subset of predictors S_i
  Calculate model performance
end
Generate the performance profile over all subsets of predictors S_i
Select the appropriate number of predictors S^*

3.4. Marginal effect analysis

The sensitivity analysis of traditional linear regression models can only evaluate one predictor at one time such that it ignores the correlation among other predictors. In this paper, the non-linear effects of a set of predictors on the severity of vehicle-bicycle crashes can be illustrated with partial dependence plots generated based on the integrated data mining framework. The partial dependence plot of each predictor demonstrates its marginal effect on the target variable in consideration of the average influences of all other predictors (Saha et al., 2015; Ding et al., 2018).

3.5. Data description

The crash dataset for model estimation comes from Victoria police crash reports across the entire state included 6237 crashes involving bicyclists and motorists between 2013 and 2018 in the entire Victoria State in south-eastern Australia (VicRoads, 2019). Victoria is the smallest mainland state and the second densely populated state in Australia. All crashes included in the analysis are finished cases, while reopened cases are excluded. The attributes of dataset which describe the characteristics of the vehicle-bicycle crashes are summarized in Table 1. Vehicle-bicycle crash injury severity is the dependent variable, which is classified into three categories, namely, fatal accident, serious injury accident and other injury accident. As the percentage of fatal accident is

extremely low in comparison with the other types of injuries, the injury severity is finally categorized as Fatal and serious injury accident (1), Other injury accident(0). 30 independent variables are selected for analysis which are classified into five categories, including the type of accident, time factor, characteristics of vehicles, characteristics of environment condition, and human factors. The definitions of some variables are further clarified as below: The road condition classification is based on the State-wide Route Numbering Scheme (SRNS) (data.vic, 2019): 'M' represents the roads which provide a consistent high standard of driving conditions, with divided carriageways, four traffic lanes, sealed shoulders and line marking easily visible in all weather conditions; 'A' represents the roads with similar high standard of driving conditions on a single carriageway; 'B' represents the sealed roads, which are wide enough for two traffic lines, with good centre line and edge line marking, shoulders, and a high standard of guidepost delineation; 'C' represents the roads that are generally two lane sealed with shoulders; others are not classified. Different types of collisions are defined based on Victoria State's local definitions for classifying accidents (DCA code) (VicRoads, 2013).

4. Results analysis

4.1. Model optimization

The vehicle-bicycle crash dataset is randomly separated into two sets, namely training set (80%) and test set (20%). To optimize the model, this paper applied a ten-fold cross-validation procedure which repeated three times on the training set to determine the optimal combination of parameters. The training set is randomly partitioned into ten sub-samples, and each of them is used as the test set while the remaining sub-samples serve as the training set.

The parameters which have been tuned from grid search based on AUC values are explained as below: The shrinkage value or learning rate is introduced to reduce the influence of each individual tree structure and leave space for future trees for improving the model (Friedman, 2002), which is set as 0.1. The lower learning rate can lead to longer computation time. The number of trees indicate the number of gradient boosting iterations, and if the value is too high, it may lead to overfitting, and the value is set as 3. The interaction depth indicate the maximum number of splits of each tree, which is set as 150. The minimum number of observations in the terminal nodes of the trees is also tuned, which is set as 10.

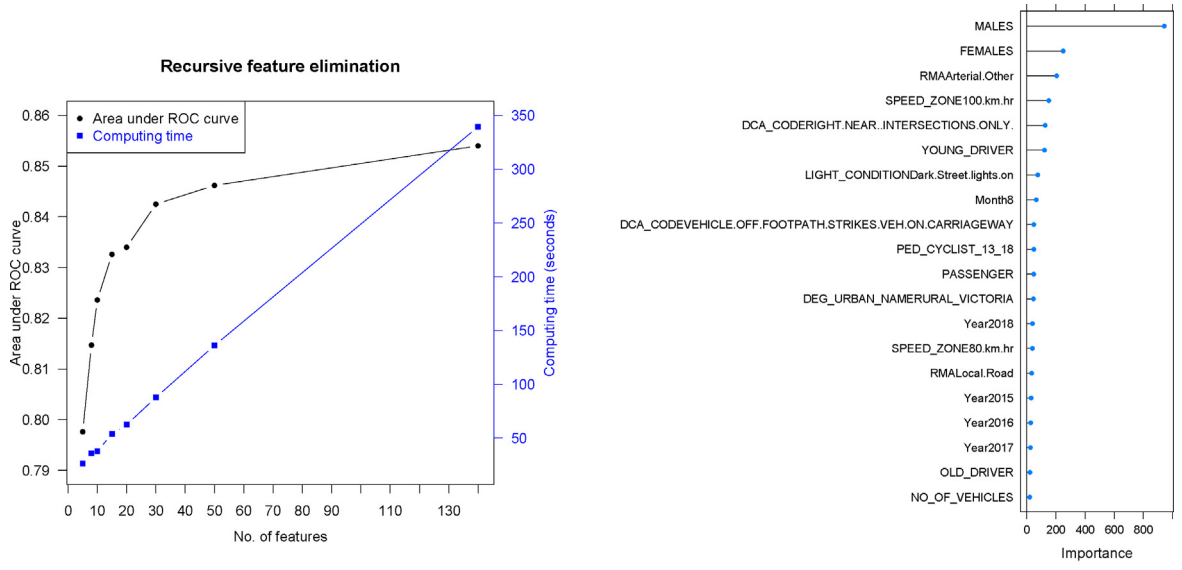
4.2. Recursive feature selection and importance ranking

After model training, recursive feature selection process has been carried out in consideration of the trade-off between AUC value and computing time, as shown in Fig. 1a. The relative contribution rankings or relative importance of the twenty most significant explanatory variables in predicting the severity of vehicle-bicycle crashes are summarized with Fig. 1b. The calculated importance ranking scores demonstrate the association between the crash-related predictors and vehicle-bicycle crash severity level. The higher the score, the more significant the predictor. 140 vehicle-bicycle crash features are included in the analysis after sparsifying the crash dataset. As the number of features increases from 5 to 140, the AUC value increases with decreasing gradient from 0.7976 to 0.8540, while the computing time also increases from 26.3 s to 339.4 s.

For further analysis, the top ten predictors have been selected, including the number of males and females involved in the crash, the road type is arterial (other than highway), the speed zone of 100 km/h, the number of young drivers involved in the crash, the light condition (dark with street lights on), the month (August), the collision type (right turn near intersections,bicycle off footpath strikes the vehicle on the carriageway), the number of bicyclists from 13 to 18 years old, with AUC value as 0.8236 and computing time as 37.8 s.

4.3. Marginal effects of key predictors

The partial dependence plots of the significant predictors on the severity of vehicle-bicycle crashes has been plotted for marginal effect analysis, where the relative logit contribution of the predictor on the class probability of 'Fatal & serious injury accident' is plotted as the y-axis. More detailed information on the concept of partial dependence plot can be found in Friedman (2002). The general positive or negative effects of the top ten predictors on the vehicle-bicycle crash severity level reflected by the partial dependence plots are described and summarized in Table 2.



(a) Trade-off of RFE performance (b) Feature importance ranking

Fig. 1. Learning-based feature selection.

Table 2
Top ten predictors.

No.	Predictor	Effect
1	No. of males involved in the crash	Positive
2	No. of females involved in the crash	Positive
3	Road classification (arterial other)	Positive
4	Speed zone (100 km/h)	Positive
5	Collision type (right turn near intersections)	Positive
6	No. of young drivers involved in the crash	Negative
7	Light condition (dark with street lights on)	Positive
8	Month (August)	Positive
9	Collision type (bicycle off footpath strikes the vehicle on the carriageway)	Positive
10	No. of bicyclists from 13 to 18 years old	Negative

4.4. Discussion

4.4.1. Demographics

As for the impact of demographics, the result shows that the crash is more likely to be a fatal or severe injury accident as the number of males involved in the accident increases. The result is consistent with Kim et al. (2007), Eluru, Bhat, and Hensher (2008), Behnood and Mannering (2017), which suggested that male bicyclists are more likely to be involved in severer crashes. In this paper, we also found that as the number of females involved in the crash increases, the crash is more likely to be severe or fatal. However, the impact of the number of females involved in the crash is less significant than the number of males involved in the crash on the prediction of vehicle-bicycle crash severity.

On the other hand, the results also show that the vehicle-bicycle crash is less likely to be fatal or severe when young drivers or bicyclists are involved in the crash. The result is consistent with the literature (Prati et al., 2017; Yan et al., 2011; Bíl, Bílová, & Müller, 2010), which suggested that bicyclists' injury severity increases with age. The result can be explained by the physical fragility of elder bicyclists, longer perception and reaction time during collision and more inattentive during cycling. It can also be explained that the elder drivers are more susceptible to injury and the higher crash involvement rate due to unsafe driving (Li, Braver, & Chen, 2003). Therefore, more attention should be directed to elder bicyclists' group and driver group to alleviate the vehicle-bicycle crash severity level.

4.4.2. Environment condition characteristics

The results illustrate that vehicle-bicycle crashes on arterials (not highway) are more like to be fatal or serious injury accidents, hence more attention should be paid to developing effective countermeasures on arterials (not highway) to improve the safety level of bicyclists. It is also found that the vehicle speed zone of 100 km/h are more likely to result in fatalities or serious injuries in comparison with other speed zones. The finding is consistent with Robartes and Chen (2017), Kim et al. (2007) which suggested that high speed can significantly affect the severity level of bicyclist-vehicle crashes. It is also found that when the light condition is dark with street lights on, the likelihood of severe and fatal vehicle-bicycle crash increases. This is consistent with the literature, as poor light condition were likely to be associated with more severe consequences of bicycle crashes due to the limited range of visibility (Eluru et al., 2008; Prati et al., 2017).

Although (Robartes & Chen, 2017; Liu et al., 2020) suggested that crashes are less likely to be severe at intersections, the intersection type is not found to be important to crash severity level in this paper. This is understandable, as the intersections still represent major conflict points for bicyclists, despite increasing alertness of bicyclists and drivers at intersections. This may also be explained by the increased numbers of cycle lanes constructed, which can reduce the sensitivity of intersection type to crash severity. In addition, the change in time and space may also cause the difference in the relationship between crash severity and the influential factors (Liu, Hainen, Li, Nie, & Nambisan, 2019).

4.4.3. Time factor

The vehicle-bicycle crash is more likely to be serious or fatal in the August, which is the winter season in Australia. According to Liu, Shen, and Huang (1995), Kaplan and Prato (2013), bicycle crashes is influenced by season and weather conditions, and the result in this paper can be explained by the unpredictability of the weather condition in the specific month.

Unlike (Behnood & Mannering, 2016), year is not identified as a significant contributing factor to crash severity. This may be explained by the fact that (Behnood & Mannering, 2016) utilized the eight-year (2005–2012) crash dataset, covering pre-recession, recession and post-recession period, such that the significant temporal instability was found. On the other hand, this paper only utilized crash dataset for post-recession period such that the long-term effect of time factor is not apparent.

4.4.4. Accident type

Previous research has also addressed the impact of collision types on crash severity level (Kim et al., 2007; Bíl et al., 2010; Yan et al., 2011; Behnood & Mannering, 2017; Prati et al., 2017). In this paper, it is found that when the vehicle turns right near intersections and when bicycle off footpath strikes the vehicle on the carriageway, the vehicle-bicycle crash severity is more likely to increase. The result can be explained by the unexpected event and higher level of kinetic energy during the particular types of collision in comparison with others.

Although head-on vehicle-bicycle collision or facing the traffic was found to be important in Kim et al. (2007), Liu et al. (2020), it is not identified as a significant predictor to crash severity in this paper. This can be explained by the fact that head-on interaction causes higher relative speed as well as more rapid response of drivers and bicyclists since they are able to see their conflict party before collision. Higher relative speed may increase the crash severity while more proactive reaction may reduce it. Moreover, the type of opponent vehicles involved in the vehicle-bicycle crash is not identified as a significant contributing factor to crash severity level in this study, even though it was identified as a significant predictor in Robartes and Chen (2017), Yan et al. (2011). The result can be explained by the difference in the dataset used for analysis and the way of feature extraction.

5. Conclusions

In this paper, an integrated data mining framework which includes imbalanced data resampling, learning-based feature extraction and marginal effect analysis is designed to determine the significant factors contributing to the severity level of vehicle-bicycle crashes based on the crash dataset of Victoria, Australia from year 2013 to 2018.

This paper has been dedicated to crash severity modeling for vehicle-bicycle crashes which has been less commonly addressed in the literature in comparison with vehicle-vehicle crashes. The learning-based feature selection technique based on gradient boosting algorithm has been applied for key feature extraction on the mass complicated crash dataset which contains a large amount of categorical variables, empty entries, etc. The most significant predictors that affect the severity of vehicle-bicycle crashes are extracted which include the number of males and females involved in the crash, the road type is arterial (other than highway), the speed zone of 100 km/h, the number of young drivers involved in the crash, the light condition (dark with street

lights on), the month (August), the collision type (right turn near intersections, bicycle off footpath strikes the vehicle on the carriageway), the number of bicyclists from 13 to 18 years old.

Although cycling is being increasingly promoted as transportation mode, the safety concerns is one of the main obstacles to their adoption. It is a challenge to enhance the safety level of bicyclists to solve the major fear of them, as their safety depends on various factors, including land use, socio-economic factors, etc. The findings of this paper highlight the necessity to identify the contributing factors to fatal and serious vehicle-bicycle crashes to enhance the safety level and guide safety improvements. Several recommendations have been made to improve the safety level of bicyclists. More attention should be paid to the road types and speed zone that are determined to be more prone to severe vehicle-bicycle crashes. For the road user groups, collision types, and time period which are identified to be significant contributing factors to the fatal and serious injury crashes, targeted education campaign should be carried out to enhance the road safety level. The method of this paper can also be extended and applied to other datasets with higher dimensionality for feature selection in order to extract the most significant features and reduce the computation time. Further study can also be carried out to study the crash severity and frequency together and identify the contributing factors to various types of crashes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Asgarzadeh, M., Verma, S., Mekary, R. A., Courtney, T. K., & Christiani, D. C. (2017). The role of intersection and street design on severity of bicycle-motor vehicle crashes. *Injury Prev., 23*, 179–185.

Bahrololoom, S., Moridpour, S., Tay, R., 2016. Factors affecting bicycle fatal and serious injury crashes in victoria, Australia, in: The 38th Australasian Transport Research Forum, Australian Transport Research Forum, pp. 1–12. .

Bahrololoom, S., Moridpour, S., Tay, R., Sobhani, A., 2017. Exploring the Factors Affecting Bicycle Crash Severity in Victoria, Australia. Technical Report.

Bahrololoom, S., Young, W., Logan, D., 2018a. Exploring the effect of kinetic energy on bicycle crash severity at intersections, in: ARRB International Conference, 28th, 2018, Brisbane, Queensland, Australia. .

Bahrololoom, S., Young, W., Logan, D., 2018b. The role of kinetic energy in bicyclist's injury severity at intersections, in: Australasian Road Safety Conference, 2018, Sydney, New South Wales, Australia. .

Bahrololoom, S., Young, W., Logan, D., 2018c. A Safe System Based Investigation of Factors Influencing Bicycle Crash Severity in Victoria, Australia. Technical Report.

Behnood, A., & Mannering, F. (2017). Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances. *Anal. Methods Accid. Res., 16*, 35–47.

Behnood, A., & Mannering, F. L. (2016). An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. *Anal. Methods Accid. Res., 12*, 1–17.

Beshah, T., Hill, S., 2010. Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia, in: 2010 AAAI Spring Symposium Series. .

Bíl, M., Bílová, M., & Müller, I. (2010). Critical factors in fatal collisions of adult cyclists with automobiles. *Accid. Anal. Prev., 42*, 1632–1636.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn., 30*, 1145–1159.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res., 16*, 321–357.

Chung, Y. S. (2013). Factor complexity of crash occurrence: an empirical demonstration using boosted regression trees. *Accid. Anal. Prev., 61*, 107–118.

data.vic, 2019. Crashes Last Five Years. http://data.vicroads.vic.gov.au/metadata/Crashes_Last_Five_Years%20-%20Open%20Data.html. Online; accessed 22 Apr 2019. .

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology, 88*, 243–251.

Ding, C., Cao, X. J., & Næss, P. (2018). Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in oslo. *Transp. Res. A Policy Pract., 110*, 107–117.

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach. *Accid. Anal. Prev., 112*, 116–126.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *J. Anim. Ecol., 77*, 802–813.

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accid. Anal. Prev., 40*, 1033–1054.

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal., 38*, 367–378.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Helak, K., Jehle, D., McNabb, D., Battisti, A., Sanford, S., & Lark, M. C. (2017). Factors influencing injury severity of bicyclists involved in crashes with motor vehicles: Bike lanes, alcohol, lighting, speed, and helmet use. *Southern Med. J., 110*, 441–444.

Kaplan, S., & Giacomo Prato, C. (2015). A spatial analysis of land use and network effects on frequency and severity of cyclist–motorist crashes in the copenhagen region. *Traffic Injury Prev., 16*, 724–731.

Kaplan, S., & Prato, C. G. (2013). Cyclist–motorist crash patterns in denmark: a latent class clustering approach. *Traffic Injury Prev., 14*, 725–733.

Kaplan, S., Vavatsoulas, K., & Prato, C. G. (2014). Aggravating and mitigating factors associated with cyclist injury severity in denmark. *J. Saf. Res., 50*, 75–82.

Kim, J. K., Kim, S., Ulfarsson, G. F., & Porrello, L. A. (2007). Bicyclist injury severities in bicycle–motor vehicle accidents. *Accid. Anal. Prev., 39*, 238–251.

Klassen, J., El-Basyouny, K., & Islam, M. T. (2014). Analyzing the severity of bicycle-motor vehicle collision using spatial mixed logit models: a city of Edmonton case study. *Saf. Sci., 62*, 295–304.

Klop, J. R., & Khattak, A. J. (1999). Factors influencing bicycle crash severity on two-lane, undivided roadways in north carolina. *Transp. Res. Record, 1674*, 78–85.

Li, G., Braver, E. R., & Chen, L. H. (2003). Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among older drivers. *Accid. Anal. Prev., 35*, 227–235.

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accid. Anal. Prev., 111*, 354–363.

Liu, J., Hainen, A., Li, X., Nie, Q., & Nambisan, S. (2019). Pedestrian injury severity in motor vehicle crashes: an integrated spatio-temporal modeling approach. *Accid. Anal. Prev., 132* 105272.

Liu, J., Khattak, A. J., Li, X., Nie, Q., & Ling, Z. (2020). Bicyclist injury severity in traffic crashes: a spatial approach for geo-referenced crash data to uncover non-stationary correlates. *J. Saf. Res.*.

Liu, X., Shen, L., & Huang, J. (1995). Analysis of bicycle accidents and recommended countermeasures in Beijing, China. *Transp. Res. Record, 1487*, 75–83.

Moore, D. N., Schneider, W. H., IV, Savolainen, P. T., & Farzaneh, M. (2011). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. *Accid. Anal. Prev., 43*, 621–630.

Narkhede, S. (2018). Understanding auc-roc curve. *Towards Data Sci., 26*.

Prati, G., Pietrantoni, L., & Fraboni, F. (2017). Using data mining techniques to predict the severity of bicycle crashes. *Accid. Anal. Prev., 101*, 44–54.

Raihan, M. A., Alluri, P., Wu, W., & Gan, A. (2019). Estimation of bicycle crash modification factors (cmfs) on urban facilities using zero inflated negative binomial models. *Accid. Anal. Prev., 123*, 303–313.

Robartes, E., & Chen, T. D. (2017). The effect of crash characteristics on cyclist injuries: an analysis of virginia automobile-bicycle crash data. *Accid. Anal. Prev., 104*, 165–173.

Saha, D., Alluri, P., & Gan, A. (2015). Prioritizing highway safety manual's crash prediction variables using boosted regression trees. *Accid. Anal. Prev., 79*, 133–144.

Sivasankaran, S. K., & Balasubramanian, V. (2020). Exploring the severity of bicycle–vehicle crashes using latent class clustering approach in India. *J. Saf. Res., 72*, 127–138.

Stipancic, J., Zangenehpour, S., Miranda-Moreno, L., Saunier, N., & Granié, M. A. (2016). Investigating the gender differences on bicycle-vehicle conflicts at urban intersections using an ordered logit methodology. *Accid. Anal. Prev., 97*, 19–27.

VicRoads, 2013. Crashstats User Guide. http://data.vicroads.vic.gov.au/metadata/crashstats_user_guide_and_appendices.pdf. Online; accessed 04 Jun 2019. .

VicRoads, 2019. Crashes Last Five Years. https://vicroadsopendata-vicroadsmaps.opendata.arcgis.com/datasets/crashes-last-five-years. Online; accessed 22 Apr 2019. .

Rash-ha Wahi, R., Haworth, N., Debnath, A. K., & King, M. (2018). Influence of type of traffic control on injury severity in bicycle–motor vehicle crashes at intersections. *Transp. Res. Record, 2672*, 199–209.

Wall, S., Lee, D., Frangos, S., Sethi, M., Heyer, J., Ayoung-Chee, P., & DiMaggio, C. (2016). The effect of sharrows, painted bicycle lanes and physically protected paths on the severity of bicycle injuries caused by motor vehicles. *Safety, 2*, 26.

Wang, C., Lu, L., & Lu, J. (2015). Statistical analysis of bicyclists' injury severity at unsignalized intersections. *Traffic Injury Prev., 16*, 507–512.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M., & Wu, C. (2011). Motor vehicle–bicycle crashes in beijing: irregular maneuvers, crash patterns, and injury severity. *Accid. Anal. Prev., 43*, 1751–1758.

Yasmin, S., & Eluru, N. (2018). A joint econometric framework for modeling crash counts by severity. *Transportmetr. A Transp. Sci., 14*, 230–255.
Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transp. Res. C Emerg. Technol., 58*, 308–324.
Zheng, Z., Lu, P., & Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. *J. Saf. Res., 65*, 115–124.

Siying Zhu is a Ph.D. student in the School of Civil and Environmental Engineering, Nanyang Technological University, Singapore. She received the bachelor?s degree from Nanyang Technological University, Singapore, in 2017. Her research involves network modeling, urban infrastructure, data analysis, public transport, and transportation safety.