Electronic Research Archive

http://www.aimspress.com/journal/ERA

ERA, 31(6): 3417–3434.
DOI: 10.3934/era.2023173
Received: 08 March 2023
Revised: 05 April 2023
Accepted: 11 April 2023
Published: 18 April 2023

Research article

A data-driven on-site injury severity assessment model for car-to-electric-bicycle collisions based on positional relationship and random forest

Ye Yu[1] and Zhiyuan Liu[2,*]

[1] Department of Public Security Management, Jiangsu Police Institute, Nanjing, Jiangsu, China
[2] Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, Nanjing, China

* **Correspondence:** Email: zhiyuanl@seu.edu.cn.

Abstract: Vulnerable road users (VRUs) are usually more susceptible to fatal injuries. Accurate and rapid assessment of VRU injury severity at the accident scene can provide timely support for decision-making in emergency response. However, evaluating VRU injury severity at the accident scene usually requires medical knowledge and medical devices. Few studies have explored the possibility of using on-site positional relationship to assess injury severity, which could provide a new perspective for on-site transportation professionals to assess accident severity. This study proposes a data-driven on-site injury severity assessment model for car-to-electric-bicycle accidents based on the relationship between the final resting positions of the car, electric bicycle and cyclist at the accident scene. Random forest is employed to learn the accident features from the at-scene positional relationship among accident participants, by which injury severity of the cyclist is assessed. Conditional permutation importance, which can account for correlation among predictor variables, is adopted to reflect the importance of predictor variables more accurately. The proposed model is demonstrated using simulated car-to-electric-bicycle collision data. The results show that the proposed model has good performance in terms of overall accuracy and is balanced in recognizing both fatal and non-fatal accidents. Model performance under partial information confirms that the position information of the electric bicycle is more important than the position information of the cyclist in assessing injury severity.

Keywords: traffic safety; data-driven; assessment; random forest; vulnerable road user

1. Introduction

Road traffic accidents cause a large number of causalities worldwide every year. According to the World Health Organization (WHO), road traffic accidents were ranked the eighth cause of death (2.5%) among people of all ages worldwide [1]. Deaths resulting from road traffic accidents reached 1.35 million in 2016 [1]. More than half of road traffic deaths are among vulnerable road users (VRUs), such as pedestrians, bicyclists and motorcyclists [1,2]. The safety issue of VRUs has been recognized as one of the critical traffic safety problems [3,4]. In general, VRUs are less protected and more vulnerable than car occupants in road traffic accidents [5–7]. As they are at greater risk for being killed in a crash than other road users, improving their safety is a top priority [8]. Therefore, there is an urgent need to study VRU-involved accidents and improve traffic safety for VRUs.

VRU-involved accidents have been the focus of much attention in the field of traffic safety analysis. Some studies have focused on pedestrian-involved accidents [9–11], while others analyzed cyclist-involved accidents [12–14]. In terms of modeling techniques, numerous statistical and machine learning methods have been developed and applied to discover injury patterns of VRU-involved accidents. Statistical methods are relatively easy to interpret, but they often involve pre-assumed relationships among variables [15]. In contrast, machine learning techniques require no a priori assumptions but usually work like a black box [2,16–20].

Among previous studies, many aim to identify the factors influencing VRU injury severity. For example, Sun et al. [21] examined influencing factors contributing to injury severity of VRU-involved crashes and investigated the differences across seasons. Islam et al. [22] investigated factors affecting injury severity in pedestrian crashes at signalized intersections. Fountas et al. [23] investigated the determinants of injury severity in single-bicycle and bicycle-motor vehicle crashes. Behnood et al. [24] examined factors influencing the injury severity in bicycle collisions and evaluated the temporal instability of factors. These studies contributed greatly to the identification of the relationships between VRU injury severity and various factors.

However, while numerous studies have investigated factors of VRU-involved traffic accidents, very few studies have been conducted in terms of on-site assessment of VRU injury severity, especially in the field of transportation engineering. Accurate and rapid assessment of VRU injury severity can provide timely support for decision-making in emergency response. For transportation professionals, emergency response and incident handling can differ between fatal accidents and other types of accidents in many ways.

Evaluating VRU injury severity at the accident scene usually requires medical knowledge and medical devices. It is a challenging task for transportation professionals at the accident scene. In practice, they often rely on medical first responders to make an initial assessment of VRU injury severity whenever possible. However, sometimes medical first responders may leave the accident scene before transportation professionals arrive. In this case, transportation professionals have to make an initial assessment fully based on personal experience and then make necessary corrections according to the final results of hospital examination and diagnosis. Although hospital examination is more accurate and reliable, it often takes a period of time and cannot fully meet the timeliness requirements of emergency response at the accident scene.

This study is therefore set out to investigate the possibility of using on-site positional relationship to assess injury severity. The motivation of this study is as follows. On the one hand, the positional relationship among the accident participants at the accident scene contains a large amount of information of accident characteristics. While this information is commonly used in forensic accident reconstruction, the potential to assess injury severity based on positional relationship at the accident scene has seldom been explored. On the other hand, the positional relationship among the accident participants is easy to measure and is available for transportation professionals at the accident scene. In practice, transportation professionals are often responsible for accident scene survey, including measuring and diagraming the accident scene.

The primary objective of this study is to propose a data-driven model for transportation professionals to conduct on-site assessment of VRU injury severity from the perspective of the positional relationship at the accident scene. Specifically, the study was focused on the assessment of the severity of the cyclist's head injuries in car-to-electric-bicycle accidents. Positional relationship among the car, electric bicycle and cyclist at the accident scene was used as model input. The severity of the cyclist's head injuries is then predicted as fatal or non-fatal based on a random forest (RF) model. In addition, given that the accident scene may change before the on-site measurement due to various reasons (such as secondary accidents, first aid, etc.), the performance of the proposed model under scenarios of partial information was also investigated.

The contributions of this study are twofold. First, a data-driven framework was proposed for on-site assessment of cyclist injury severity in car-to-electric-bicycle accidents based on the relationship among the final resting positions of the car, electric bicycle and cyclist at the accident scene. Using the proposed framework, transportation professionals at the accident scene can make a rapid assessment of VRU injury severity and adjust emergency response in time. This study might be the first effort to demonstrate the viability of assessing VRU injury severity based on positional relationship at the accident scene.

Second, the positional relationships among accident participants at the car-to-electric-bicycle accident scene were examined in terms of their associations with VRU injury severity and ranked according to their conditional permutation importances within the proposed model. Compared to permutation importance, conditional permutation importance can account for correlation among predictor variables and is more robust in reflecting the true importance. The results of variable importance analysis can provide guidance for the collection of distance information at the accident scene.

2. Methods

2.1. Injury severity assessment model for car-to-electric-bicycle accidents

Random forest is an ensemble learning technique proposed by Breiman [25]. A RF model consists of many decision trees as its base learners. Randomness is introduced in the training process of base learners using both random sample selection and random attribute selection. On the one hand, each tree is constructed from a bootstrap sample from the training dataset. In other words, different decision trees are trained using different data sets with the same sample size. On the other hand, when splitting each node, a subset is randomly selected from the whole attribute set, and then an optimal attribute is chosen for partitioning from that subset. In this way, RF can diversify base learners and thereby improve model performance. RF has been widely used in many real-world tasks and has shown strong performance [26–29].

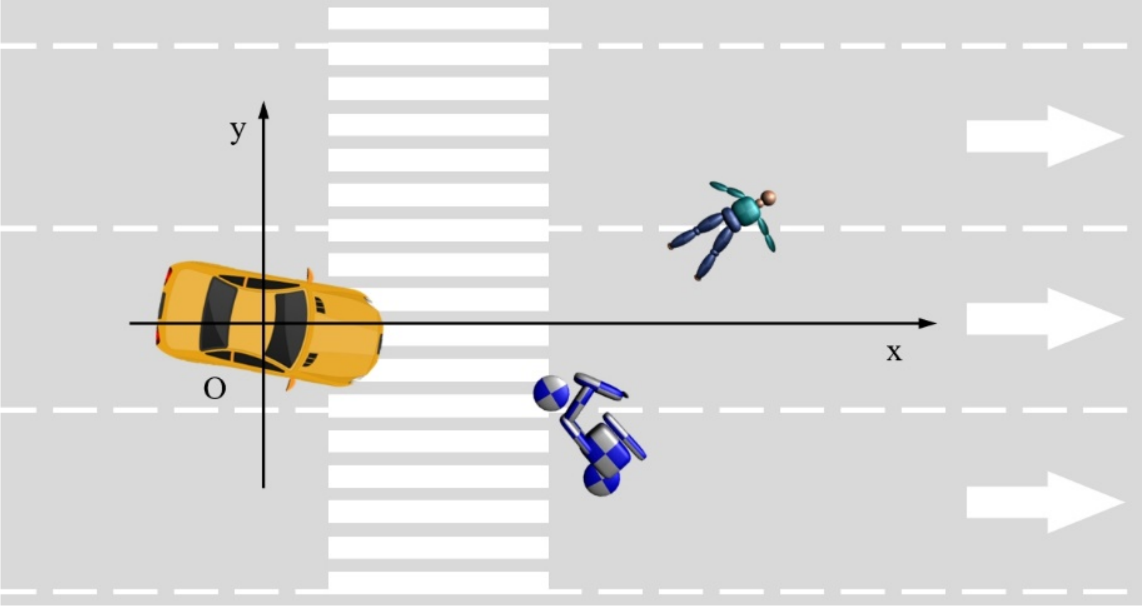

Figure 1. Schematic diagram of a car-to-electric-bicycle accident scene.

In this study, a data-driven injury severity assessment model is established based on random forest. Figure 1 presents a scene diagram of a car-to-electric-bicycle accident. As shown in Figure 1, a Cartesian coordinate system is established with the center of gravity of the car as the origin and the forward direction of the motorway as the positive direction of the X-axis. This coordinate system is used in the following analysis in this study.

To capture the final resting positions and orientations of the cyclist and the electric bicycle at the accident scene, both the cyclist and the electric bicycle are seen as made up of multiple components. X and Y coordinates of a component can represent the longitudinal and lateral distance between the component and the center of gravity of the car. For each component, its X and Y coordinates are obtained and used as inputs for the injury severity model. In this way, positional relationships among accident participants at the accident scene are incorporated into the proposed injury severity assessment model.

The output of the proposed model is cyclist injury severity. When performing an on-site assessment of injury severity, the primary concern is whether the cyclist has been fatally injured. For real-world traffic accidents, injury severity can be obtained from crash records. For simulated traffic accidents, injury severity can be evaluated using the head injury criterion (HIC). The higher the HIC score is, the greater the likelihood of a head injury. The reason for using HIC to represent overall injury severity for the cyclist is that head injuries have been confirmed as the primary and most fatal type of injury for cyclists [30]. HIC value can be derived from the acceleration measurements of the test dummy's head during the collision. Consistent with the Federal Motor Vehicle Safety Standards (FMVSS), the tolerable limit for HIC value is set to 700 in this study [31,32]. Cyclists with HIC values more than 700 are considered fatally injured in the accident simulation.

Based on the training data set, the proposed injury severity assessment model attempts to learn the accident features from the at-scene positional relationship among accident participants, by which injury severity of the cyclist is evaluated.

2.2. Variable importance measure

Variable importance is a measure to rank variables in the predictor set based on their importance in producing accurate predictions [33,34]. RFs have been widely used in many applications for identifying and ranking relevant predictor variables.

2.2.1. Permutation importance

Traditionally, a RF model evaluates the importance of each predictor variable by permuting its values randomly [25,34]. Prediction accuracy before and after permutation is then obtained and compared. The difference in prediction accuracy before and after permuting the predictor variable, averaged over all trees, is used as an importance measure. It can be formalized as follows.

Let Ω_k be the out-of-bag sample for tree k in a random forest model. Then, the variable importance of variable X_j in tree k is

$$PI_k(X_j) = \frac{\sum_{i\in\Omega_k} I\left(\gamma_i = \widehat{\gamma}_i^{(k)}\right)}{|\Omega_k|} - \frac{\sum_{i\in\Omega_k} I\left(\gamma_i = \widehat{\gamma}_{i,\pi_j}^{(k)}\right)}{|\Omega_k|} \tag{1}$$

where $PI_k(X_j)$ is permutation importance for predictor variable X_j in tree k, $\widehat{\gamma}_i^{(k)}$ is the predicted class for observation i before permuting X_j, $\widehat{\gamma}_{i,\pi_j}{}^{(k)}$ is the predicted class for observation i after permuting X_j, I represents the indicator function, and $|\Omega_k|$ denotes the number of elements in Ω_k.

Let K be the number of trees in the RF model, and variable importance for each predictor variable is then computed as the mean importance for all trees.

$$PI(X_j) = \frac{\sum_k PI_k(X_j)}{K} \tag{2}$$

where $PI(X_j)$ is permutation importance for predictor variable X_j averaged over all trees.

2.2.2. Conditional permutation importance for correlated predictor variables

While permutation importance has been extensively used to identify and rank relevant predictor variables based on RF models, it tends to overestimate the importance of correlated predictor variables [35–37]. A predictor variable that is weakly or not associated with the response variable but highly correlated with another predictor variable can still be misjudged as an important variable [35]. In this study, the issue of correlated predictor variables is prominent and unignorable. For example, positioning data of body parts of the same vehicle or the same victim are highly correlated. Under such circumstances, permutation importance is not suitable for ranking variable importance.

In this study, the conditional permutation importance proposed by Strobl et al. [35] is therefore employed to account for the correlation among predictor variables, thus reflecting true impact of the predictor variables. The key difference between permutation importance and conditional permutation importance is the last term in Eq (1). A permutation grid is defined within which the values of X_j are permuted for each tree based on the partition of the feature space induced by that tree. For all variables

Z to be conditioned on, the cut points that split X_j in the current tree are extracted and used to create the permutation grid. The out-of-bag prediction accuracy after conditional permutation is then computed based on conditionally permuting the values of X_j within groups of Z. Conditional permutation importance can be formalized as follows.

$$CPI_k(X_j) = \frac{\sum_{i\in\Omega_k} I\left(\gamma_i=\hat{\gamma}_i^{(k)}\right)}{|\Omega_k|} - \frac{\sum_{i\in\Omega_k} I\left(\gamma_i=\hat{\gamma}_{i,\pi_j|Z}^{(k)}\right)}{|\Omega_k|} \tag{3}$$

$$CPI(X_j) = \frac{\sum_k CPI_k(X_j)}{K} \tag{4}$$

where $CPI_k(X_j)$ is conditional permutation importance for predictor variable X_j in tree k, $CPI(X_j)$ is conditional permutation importance for predictor variable X_j averaged over all trees, and $\hat{\gamma}_{i,\pi_j|Z}^{(k)}$ is the predicted classes for observation *i* after permuting X_j within the grid defined by the variables *Z*. For details on computing conditional permutation importance, see Strobl et al. [35].

2.3. Model performance evaluation

In this study, traffic accidents were classified into fatal and non-fatal accidents according to injury severity. It is a two-class classification task. Fatal accidents were designated as the positive class, with the other as the negative class. Obviously, the positive class is the class of focus. To evaluate the performance of the proposed data-driven injury severity assessment model, the following three indexes were used: namely, accuracy, sensitivity and specificity. The latter two indexes are more focused on each class. These indexes were also utilized to compare the performances of multiple models.

The classification performance of a classifier f with respect to test data D can be summarized using a confusion matrix [38]. As shown in Table 1, the four cells of the confusion matrix are designated as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). In this context, the performance measures used in this study can be defined as follows.

Table 1. Confusion matrix for a two-class classification task.

Reference class	Predicted class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Accuracy is defined as the ratio of the number of correct predictions to the total number of predictions. It is an intuitive performance measure for a classification problem and can reflect the overall classification performance.

$$accuracy(f;D) = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

Sensitivity is the probability of a positive test result, conditioned on the individual truly being positive. It can measure the fraction of traffic accidents with fatal injuries (i.e., positive examples) which have a positive test result. In this study, it can reflect how well a classifier can identify fatal accidents.

$$sensitivity(f;D) = \frac{TP}{TP+FN} \tag{6}$$

Specificity is the probability of a negative test result, conditioned on the individual truly being negative. It can reflect a classifier's ability to identify non-fatal accidents.

$$specificity(f;D) = \frac{TN}{FP+TN} \tag{7}$$

3. Data description

The proposed method is demonstrated using car-to-electric-bicycle collision data simulated by well-known accident reconstruction software, namely, PC-Crash. It can simulate vehicle collisions and generate highly accurate reconstructions of various accident scenarios. Specifically, traffic collisions between cars and electric bicycles crossing the road in front of them are simulated. For simplicity, electric bicycles always cross the road from the right side of the cars in the simulation experiments. The situations of crossing the road from the left side are exactly the mirror-image scenarios and not simulated in this study. As a result, cars always hit the left side of the electric bicycles first in the simulation experiments in this study.

To increase the coverage of car-to-electric-bicycle collision scenarios, the speeds of cars, the speeds of electric bicycles, the collision angles and the contact positions when collisions happen are adjusted in a given range. The speeds of the car and electric bicycle before collision vary in the ranges of 30–70 km/h and 15–35 km/h, respectively. The collision angle varies within 45 to 135 degrees, and the collision position varies from the left front corner to the right front corner of the car. In this way, a total of 315 collision scenarios were simulated.

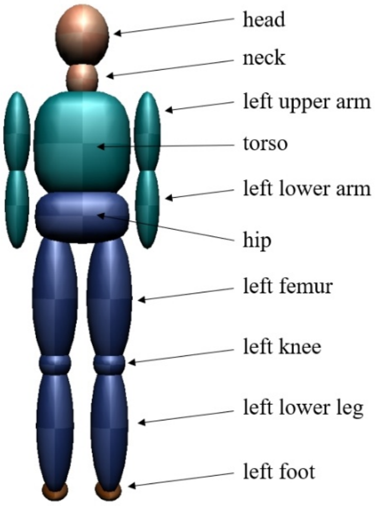

Figure 2. The components of the cyclist model.

Table 2. Description of position variables for the cyclist.

Component	Position	Mean	Standard deviation	Median	Min	Max
Head	longitudinal	1.14	9.91	3.27	-29.54	20.00
	lateral	6.98	4.07	6.42	-3.85	18.63
Hip	longitudinal	1.19	9.79	3.11	-29.29	20.20
	lateral	7.02	4.03	6.40	-3.58	18.02
Left femur	longitudinal	1.20	9.78	3.20	-28.95	20.24
	lateral	7.00	4.04	6.36	-3.50	17.85
Left foot	longitudinal	1.27	9.74	3.42	-28.49	20.48
	lateral	7.00	4.06	6.25	-3.55	17.80
Left knee	longitudinal	1.22	9.78	3.26	-28.74	20.20
	lateral	6.99	4.04	6.29	-3.51	17.72
Left lower arm	longitudinal	1.14	9.87	3.24	-29.00	19.99
	lateral	6.91	4.09	6.39	-3.84	18.57
Left lower leg	longitudinal	1.24	9.76	3.36	-28.64	20.33
	lateral	7.00	4.05	6.32	-3.52	17.65
Left upper arm	longitudinal	1.14	9.87	3.14	-29.16	20.16
	lateral	6.94	4.08	6.44	-3.70	18.46
Neck	longitudinal	1.15	9.88	3.15	-29.48	20.04
	lateral	6.99	4.06	6.49	-3.78	18.48
Right femur	longitudinal	1.22	9.75	3.23	-29.34	20.08
	lateral	7.04	4.02	6.49	-3.67	17.79
Right foot	longitudinal	1.30	9.71	3.42	-29.08	20.22
	lateral	7.07	4.06	6.41	-3.70	18.08
Right knee	longitudinal	1.25	9.73	3.37	-29.31	19.99
	lateral	7.05	4.02	6.42	-3.71	17.64
Right lower arm	longitudinal	1.22	9.96	3.45	-29.40	19.92
	lateral	7.01	4.07	6.45	-4.30	18.71
Right lower leg	longitudinal	1.27	9.72	3.36	-29.16	20.11
	lateral	7.06	4.04	6.44	-3.66	17.84
Right upper arm	longitudinal	1.19	9.91	3.38	-29.59	19.81
	lateral	7.02	4.06	6.47	-4.07	18.52
Torso	longitudinal	1.17	9.84	3.26	-29.39	20.12
	lateral	7.00	4.04	6.39	-3.68	18.26

Note that both the cyclist and the electric bicycle are modeled as multibody systems consisting of multiple components in the simulation experiments. Figures 2 and 3 depict the components of the cyclist model and the e-bike model, respectively. In Figure 2, for the components that exist on both the left and right sides of the human body, only one side is marked for simplicity. The final resting positions of cars, electric bicycles and cyclists were extracted and saved as a TXT file. The positions of the electric bicycle and the cyclist were obtained at the component level. Tables 2 and 3 describe the position variables for the cyclist and the electric bicycle, respectively.

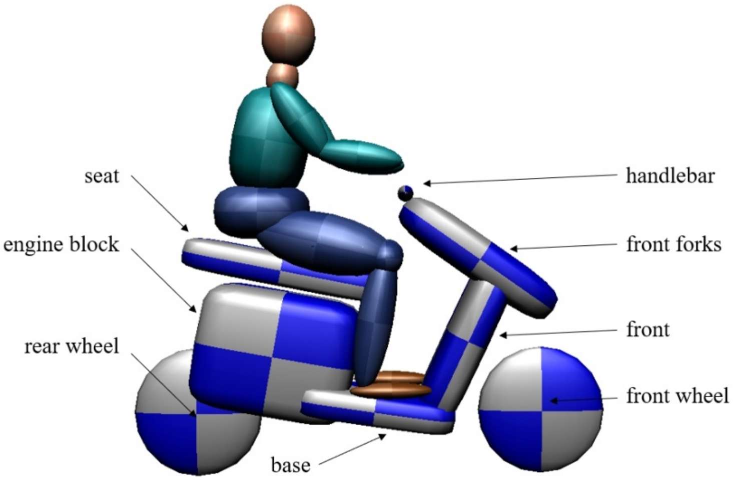

Figure 3. The components of the electric bicycle model.

In reality, injury severity of the cyclist can only be evaluated after a detailed medical examination. Sometimes, the injured may die after a period of hospitalization. However, in the simulation experiments, the injury severity of the cyclists can be determined directly based on dynamic responses of their heads during the collision in terms of HIC. As mentioned above, cyclists with HIC value more than 700 are judged as fatally injured. In the dataset generated by simulation, there are 67 fatal examples and 248 non-fatal examples. Therefore, it is an unbalanced dataset.

Table 3. Description of position variables for the electric bicycle.

Component	Position	Mean	Standard deviation	Median	Min	Max
Base	longitudinal	12.18	18.38	11.22	-29.15	70.47
	lateral	2.30	7.83	3.70	-17.85	17.89
Engine block	longitudinal	12.13	18.34	11.22	-28.95	70.16
	lateral	2.29	7.77	3.68	-17.49	17.51
Front	longitudinal	12.24	18.41	11.47	-28.93	70.82
	lateral	2.31	7.91	3.68	-17.92	18.00
Front forks	longitudinal	12.25	18.41	11.61	-28.67	70.89
	lateral	2.32	7.93	3.77	-18.14	17.87
Front wheel	longitudinal	12.25	18.40	11.72	-29.10	71.02
	lateral	2.29	7.91	3.72	-18.22	18.35
Handlebar	longitudinal	12.25	18.42	11.50	-28.53	70.69
	lateral	2.34	7.94	3.57	-18.13	17.55
Rear wheel	longitudinal	12.08	18.32	11.32	-29.14	69.87
	lateral	2.28	7.72	3.87	-17.48	17.57
Seat	longitudinal	12.13	18.34	11.48	-28.67	70.23
	lateral	2.30	7.79	3.58	-17.65	17.54

4. Results

4.1. Model performance

The whole dataset generated by collision simulations was partitioned into a training set and a testing set randomly. Stratified sampling was utilized, ensuring that the ratio of fatal and non-fatal accidents in both the training and testing datasets remains proportional to the overall dataset. Among them, 75% of the original data entered the training set. Based on the training set, 10-fold cross-validation was employed to avoid overfitting. During each iteration, 10% of the training set was left out for model performance validation. The other 90% of the training set was used to train the proposed model. When training the proposed model, downsampling was employed to get a balanced dataset for both majority and minority classes (namely, non-fatal accidents and fatal accidents). Unbalanced datasets can bias classification models toward the majority class. Down sampling ensures that the model is trained on a balanced dataset, which can improve its ability to correctly predict the minority class. The model achieving the highest accuracy with the optimal parameters was obtained through validation. The optimal model was then employed and further tested against the testing set. For comparison, three models were established in the same way using decision tree (DT), neural network (NN) and support vector machine (SVM).

Table 4 presents the performances of these four models based on out-of-bag evaluation. In other words, the proposed model was evaluated using the data which it has never seen during the whole training process. As shown in Table 4, the proposed RF model achieves an overall accuracy of 83.33% with a sensitivity of 87.50% and a specificity of 82.26%. It outperforms all the other models in terms of overall accuracy. Comparing to decision tree, using random forest improves model performance in all of accuracy, sensitivity and specificity, just as expected. While the neural network achieves a higher sensitivity (93.75%) than the random forest model, its specificity (75.81%) is much lower than that of the random forest model (82.26%).

Accuracy measures how well a test can correctly classify all test results, both positive and negative. Sensitivity measures a test's ability to correctly identify true positives, while specificity measures its ability to correctly identify true negatives. Sensitivity and specificity are inversely related. As sensitivity increases, specificity typically decreases, and vice versa. However, in many cases, a balance between sensitivity and specificity is necessary. The model’s performance in this study will be primarily evaluated based on its accuracy while also taking into account the balance between sensitivity and specificity. In general, the proposed RF model outperforms all other models and is more balanced in recognizing both fatal and non-fatal accidents, indicating that it is a promising method for on-site injury severity assessment.

Table 4. Model performance based on out-of-bag evaluation.

Model	Accuracy	Sensitivity	Specificity
Random forest (RF)	83.33%	87.50%	82.26%
Decision tree (DT)	74.36%	68.75%	75.81%
Neural network (NN)	79.49%	93.75%	75.81%
Support vector machine (SVM)	70.51%	87.50%	66.13%

4.2. Variable importance analysis

Another advantage of using the RF model is its ability to evaluate the importance of multiple predictor variables. As mentioned above, conditional permutation importance is employed to account for correlation among predictor variables in this study. To eliminate the effect of the random seed on the results, a group of random forest models was established using 10 different random seeds. The 48 predictor variables were ranked according to their conditional permutation importances. For each variable, the average rank across the 10 models was used as the final importance ranking. Table 5 provides the top 10 most important predictor variables.

Table 5. The top 10 most important predictor variables.

Rank	Multibody system	Component	Position
1	cyclist	left foot	longitudinal
2	electric bicycle	handlebar	longitudinal
3	electric bicycle	handlebar	lateral
4	cyclist	right lower arm	longitudinal
5	electric bicycle	rear wheel	longitudinal
6	electric bicycle	seat	lateral
7	cyclist	left lower leg	longitudinal
8	cyclist	left knee	longitudinal
9	electric bicycle	seat	longitudinal
10	electric bicycle	rear wheel	lateral

Among the top 10 most important predictor variables, four of them are variables describing the final resting position of the cyclist, namely, the longitudinal positions of the cyclist’s left foot, right lower arm, left lower leg and left knee. It should be noted that, in the simulation experiments in this study, the electric bicycle always crosses the road from the right side of the car. Therefore, the cyclist was always hit first on the left side. Left foot, left lower leg and left knee of the cyclist are the lower extremities on the impact side. It appears that the longitudinal distance between the car's final resting position and the lower extremity on the impact side of the cyclist plays a crucial role in assessing the severity of head injuries.

Interestingly, the longitudinal position of the right lower arm is also included in the top 10 most important predictor variables in Table 5. The reason for this phenomenon may be that, as the human body stretches after being hit, the posture of cyclist can be better positioned by combining the position of the left lower extremity and right upper extremity. Figure 4 depicts the relationship between the position of right lower arm of the cyclist and injury severity. As shown in Figure 4, the marginal distribution of the longitudinal position of the right lower arm differs substantially in the two accident categories. Therefore, the longitudinal position of the right lower arm can provide information for assessing accident injury severity. However, the marginal distribution of the lateral position of the right lower arm is very similar in the two accident categories. As a result, the lateral position of the right lower arm is not included in the top 10 most important predictor variables.

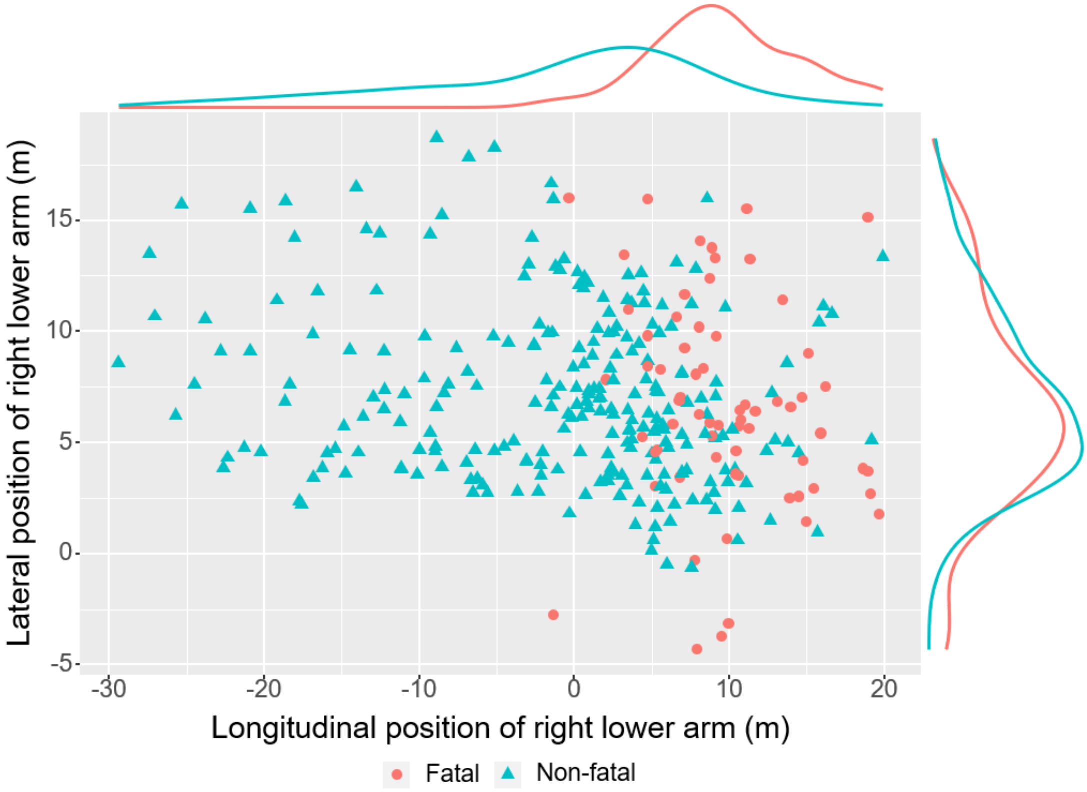

Figure 4. The position of right lower arm and the injury severity.

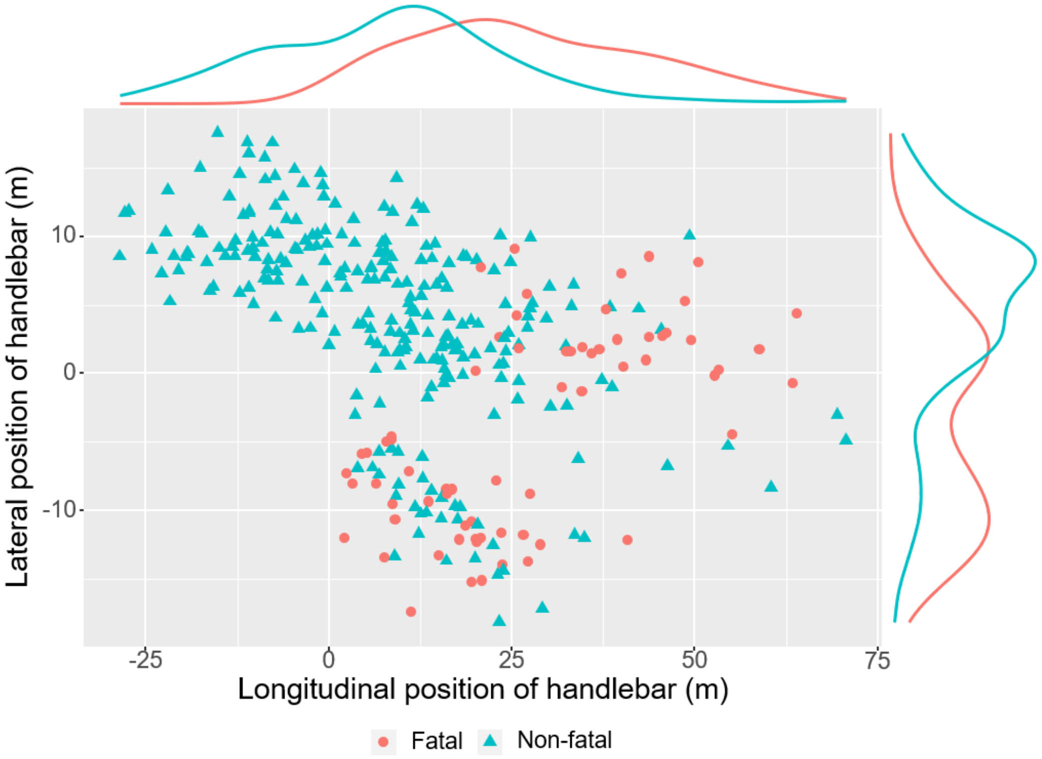

Figure 5. The position of handlebar and the injury severity.

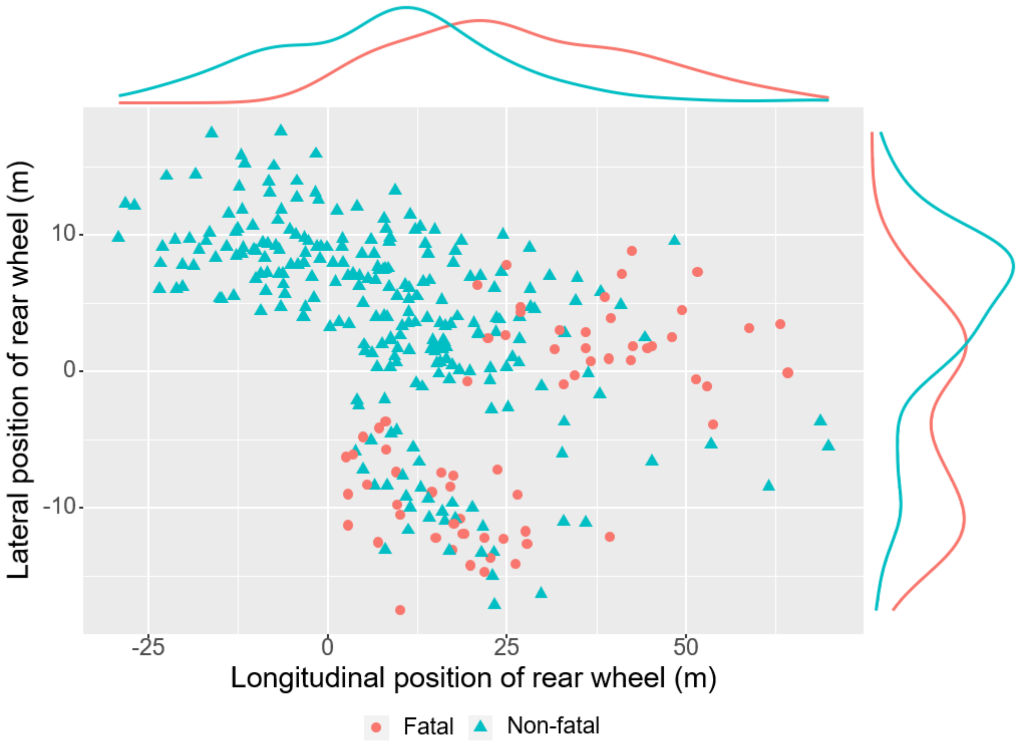

Figure 6. The position of rear wheel and the injury severity.

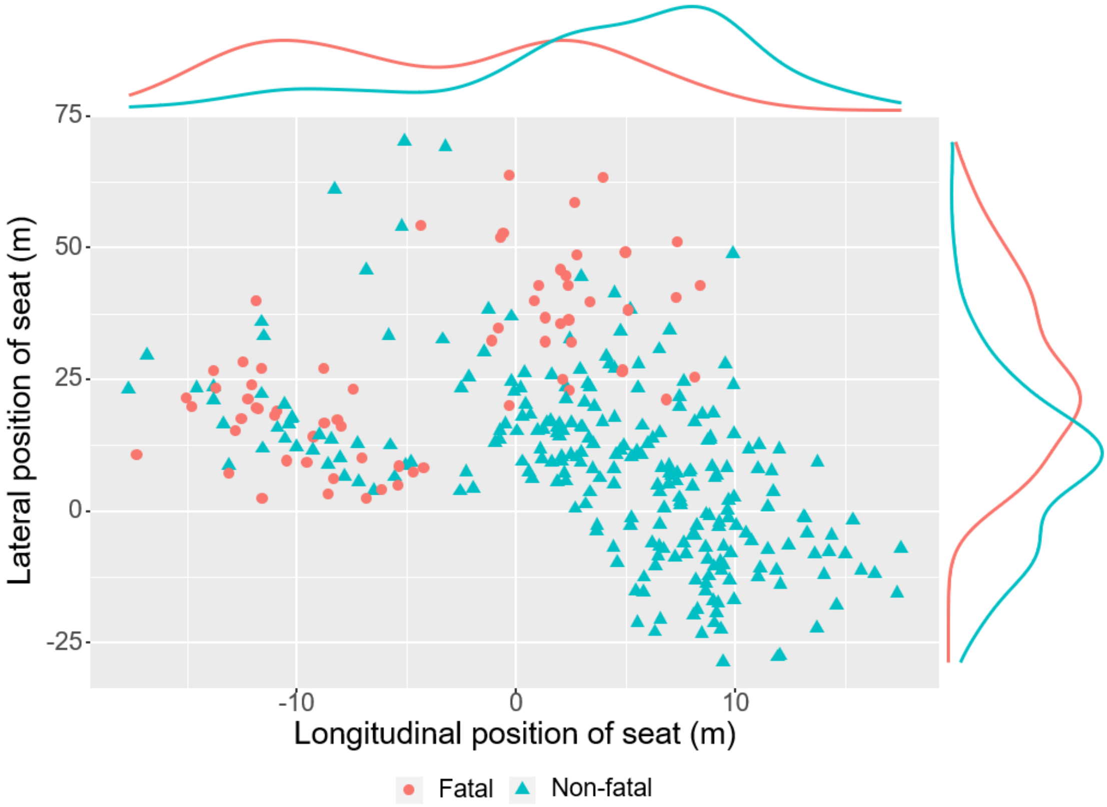

Figure 7. The position of seat and the injury severity.

The other 6 variables in Table 5 are variables that characterize the final resting position of the electric bicycle. Interestingly, these 6 variables can be grouped into 3 pairs. Each pair can describe the longitudinal and lateral position of a component of the electric bicycle. The three components of the electric bicycle are the handlebar, seat and rear wheel. Both longitudinal and lateral positions were considered important for the 3 components of the electric bicycle, while only longitudinal position was identified to be important for the 4 components of the cyclist in Table 5.

Figures 5–7 depict the longitudinal and lateral positions of the handlebar, seat and rear wheel, respectively. The corresponding marginal distributions are also provided. The distributions of each of the 6 predictor variables differ noticeably between fatal and non-fatal accidents, which is why these predictors have been ranked among the top 10 most important variables. Additionally, unlike the results for rider variables, the differences in lateral variables are more noticeable for electric bicycle variables. Nevertheless, assessing injury severity based on only one or two of the top 10 most important variables remains very challenging.

4.3. Model performance under partial information

In practice, the final resting position of a cyclist or electric bicycle after a traffic accident may become unmeasurable due to the accident scene being changed prior to the site survey. For example, the cyclist may be moved by first-aiders. In addition, the position of the cyclist or the position of the electric bicycle may be impacted by a secondary accident. Under such circumstances, the position of the cyclist or the position of the electric bicycle may become unmeasurable at the accident scene.

Therefore, this study also investigated the performance of the proposed RF model under partial information. Model 1A was established using the proposed method without the final resting position of the cyclist after traffic accidents. It should be noted that the position information of the cyclist was neither used for model training nor used for testing for Model 1A. Similarly, Model 1B was developed without the final resting position of the electric bicycle. As shown in Table 6, the performances of Model 1A, Model 1B and Model 1 were compared using out-of-bag samples.

Table 6. Model performance under partial information.

Model	Accuracy	Sensitivity	Specificity
Random forest (Model 1)	83.33%	87.50%	82.26%
Random forest without the cyclist position (Model 1A)	82.05%	93.75%	79.03%
Random forest without the e-bike position (Model 1B)	76.92%	81.25%	75.81%

It was found that, in comparison to Model 1, the accuracy of Model 1A drops slightly from 83.33% to 82.05%. It appears that the lack of the information about the final resting position of the cyclist does not substantially affect the prediction performance. This is not surprising as the majority of the top 10 most important predictor variables in Model 1 are variables for the final resting position of the electric bike. Although the sensitivity of Model 1A increases to some extent, the specificity decreases from 82.26% to 79.03%. In terms of overall accuracy, Model 1A performs slightly worse than Model 1.

In terms of Model 1B, its accuracy, sensitivity and specificity all drop substantially compared to Model 1. This phenomenon further confirms that the position information of the electric bicycle is more important than the position information of the cyclist in assessing injury severity. Therefore, when evaluating injury severity of car-to-electric-bicycle accidents based on positional relationship at the accident scene, attention needs to be paid to the accuracy of the position information of the electric bicycle.

5. Conclusions

This study proposes a data-driven on-site injury severity assessment model for car-to-electric-bicycle accidents based on the relationship between the final resting positions of the car, electric bicycle and cyclist at the accident scene. Random forest was employed to learn the accident features from the at-scene positional relationship among the accident participants, by which injury severity of the cyclist is assessed. Conditional permutation importance, which can account for correlation among predictor variables, was adopted to reflect the importance of predictor variables more accurately. The proposed model can be used by on-site transportation professionals to assess injury severity in traffic accidents as no medical knowledge or medical device is required. For example, when first responders are not on the scene, traffic police officers can still assess the severity of injuries to victims by measuring the accident site. However, it should be noted that the model proposed in this study needs to be calibrated with local accident data when applied specifically. Whenever possible, it is advisable that qualified medical personnel conduct a thorough examination of the injured individual.

The proposed model was demonstrated using simulated car-to-electric-bicycle collision data. The results show that the proposed model can achieve good classification accuracy and is balanced in recognizing both fatal and non-fatal accidents. In terms of variable importance, it was found that the longitudinal distance between the car's final resting position and the lower extremity on the impact side of the cyclist plays a crucial role in assessing injury severity. However, six of the top 10 most important variables are the ones describing the final resting position of the electric bicycle. Model performance under partial information also confirms that, overall, the position information of the electric bicycle is more important than the position information of the cyclist in assessing injury severity. This may be because the final resting position of the electric bicycle can provide more information about accidents.

There are several potential enhancements that could be considered in future works. First, further research may validate the proposed method using real world traffic accident data and perform an in-depth comparison between the predicted injury severity and that judged by medical examination. Apart from categorizing accidents into fatal and non-fatal, it is also feasible to further classify them based on the severity of the injury (such as minor, serious or fatal). Second, researchers may incorporate other measurable factors at the accident scene (such as heading of the car and angle of steering wheel) into the data-driven model to achieve better performance. Third, although this study was mainly focused on car-to-electric-bicycle collisions, further research may extend to other types of traffic accidents, especially those related to vulnerable road users.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. *World Health Organization*, Global Status Report on Road Safety 2018, 2018. Available from: https://www.who.int/publications/i/item/9789241565684.
2. K. Santos, J. P. Dias, C. Amado, A literature review of machine learning algorithms for crash injury severity prediction, *J. Saf. Res.*, **80** (2022). 254–269. https://doi.org/10.1016/j.jsr.2021.12.007
3. J. Park, M. Abdel-Aty, Application of random effects nonlinear model for analyzing motorized and nonmotorized traffic safety performance, *J. Transp. Eng. Part A. Syst.*, **147** (2021), 04020147. https://doi.org/10.1061/jtepbs.0000485
4. J. A. Alagbe, H. H. Han, S. Jin, Effect of technological distractions on pedestrian safe-crossing performance during mixed pedestrian-bicycle flow overlapping with turning vehicles: A case study of Hangzhou, China, *J. Transp. Eng. Part A. Syst.*, **149** (2023), 05022007. https://doi.org/10.1061/jtepbs.Teeng-7597
5. W. Chen, F. Zhu, Discussions on pedestrian delay models and applications at signalized crosswalks, *Multimodal Transp.*, **1** (2022), 100039. https://doi.org/10.1016/j.multra.2022.100039
6. H. Ding, N. N. Sze, Effects of road network characteristics on bicycle safety: A multivariate Poisson-lognormal model, *Multimodal Transp.*, **1** (2022), 100020. https://doi.org/10.1016/j.multra.2022.100020
7. Q. Yuan, X. Xu, T. Wang, Y. Chen, Investigating safety and liability of autonomous vehicles: Bayesian random parameter ordered probit model analysis, *J. Intell. Connected Veh.*, **5** (2022), 199–205. https://doi.org/10.1108/JICV-04-2022-0012
8. Y. J. Hu, Y. Zhang, K. S. Shelton, Where are the dangerous intersections for pedestrians and cyclists: A colocation-based approach, *Transp. Res. Part C Emerging Technol.*, **95** (2018), 431–441. https://doi.org/10.1016/j.trc.2018.07.030
9. M. G. Mohamed, N. Saunier, L. F. Miranda-Moreno, 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.*, **54** (2013), 27–37. https://doi.org/10.1016/j.ssci.2012.11.001
10. A. Kumar, M. Paul, I. Ghosh, Analysis of pedestrian conflict with right-turning vehicles at signalized intersections in India, *J. Transp. Eng. Part A. Syst.*, **145** (2019), 04019018. https://doi.org/10.1061/jtepbs.0000239
11. R. O. Mujalli, L. Garach, G. Lopez, T. Al-Rousan, Evaluation of injury severity for pedestrian-vehicle crashes in Jordan using extracted rules, *J. Transp. Eng. Part A. Syst.*, **145** (2019), 04019028. https://doi.org/10.1061/jtepbs.0000244
12. I. Isaksson-Hellman, J. Toreki, The effect of speed limit reductions in urban areas on cyclists' injuries in collisions with cars, *Traffic Inj. Prev.*, **20** (2019). S39–S44. https://doi.org/10.1080/15389588.2019.1680836
13. C. Leo, C. Klug, M. Ohlin, N. M. Bos, R. J. Davidse, A. Linder, Analysis of Swedish and Dutch accident data on cyclist injuries in cyclist-car collisions, *Traffic Inj. Prev.*, **20** (2019). S160–S162. https://doi.org/10.1080/15389588.2019.1679551
14. A. K. Hoye, O. Johansson, I. S. Hesjevoll, Safety equipment use and crash involvement among cyclists - Behavioral adaptation, precaution or learning?, *Transp. Res. Part F Psychol. Behav.*, **72** (2020), 117–132. https://doi.org/10.1016/j.trf.2020.05.002

15. G. Li, Z. Yang, Y. Y. Pan, J. X. Ma, Analysing and modelling of discretionary lane change duration considering driver heterogeneity, *Transportmetrica B: Transport Dyn.*, **11** (2023), 343–360. https://doi.org/10.1080/21680566.2022.2067599.
16. K. Huang, C. Jiang, P. Li, A. Shan, J. Wan, W. H. Qin, A systematic framework for urban smart transportation towards traffic management and parking, *Electron. Res. Arch.*, **30** (2022), 4191–4208. https://doi.org/10.3934/era.2022212
17. Y. Liu, R. Jia, J. Ye, X. Qu, How machine learning informs ride-hailing services: A survey, *Commun. Transp. Res.*, **2** (2022), 100075. https://doi.org/10.1016/j.commtr.2022.100075
18. S. Li, Y. Liu, X. B. Qu, Model controlled prediction: A reciprocal alternative of model predictive control, *IEEE/CAA J. Autom. Sin.*, **9** (2022), 1107–1110. https://doi.org/10.1109/jas.2022.105611
19. Y. Liu, F. Y. Wu, C. Lyu, S. Li, J. P. Ye, X. B. Qu, Deep dispatching: A deep reinforcement learning approach for vehicle dispatching on online ride-hailing platform, *Transp. Res. E: Logist. Transp. Rev.*, **161** (2022). https://doi.org/10.1016/j.tre.2022.102694
20. Y. Zhang, Q. Cheng, Y. Liu, Z. Liu, Full-scale spatio-temporal traffic flow estimation for city-wide networks: a transfer learning based approach, *Transportmetrica B: Transport Dyn.*, **11** (2022), 869–895. https://doi.org/10.1080/21680566.2022.2143453
21. Z. Y. Sun, Y. X. Xing, J. Y. Wang, X. Gu, H. P. Lu, Y. Y. Chen, Exploring injury severity of vulnerable road user involved crashes across seasons: A hybrid method integrating random parameter logit model and Bayesian network, *Saf. Sci.*, **150** (2022), 105682. https://doi.org/10.1016/j.ssci.2022.105682
22. A. Islam, M. Mekker, P. A. Singleton, Examining pedestrian crash frequency, severity, and safety in numbers using pedestrian exposure from utah traffic signal data, *J. Transp. Eng. Part A. Syst.*, **148** (2022), 04022084. https://doi.org/10.1061/jtepbs.0000737
23. G. Fountas, A. Fonzone, A. Olowosegun, C. McTigue, Addressing unobserved heterogeneity in the analysis of bicycle crash injuries in Scotland: A correlated random parameters ordered probit approach with heterogeneity in means, *Anal. Methods Accid. Res.*, **32** (2021), 100181. https://doi.org/10.1016/j.amar.2021.100181
24. A. Behnood, S. H. Hosseini, S. R. Davoodi, Bicyclists injury severities: An empirical assessment of temporal stability, *Accid. Anal. Prev.*, **168** (2022), 106616. https://doi.org/10.1016/j.aap.2022.106616
25. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/a:1010933404324
26. C. Zhang, J. Bin, W. Wang, X. Peng, R. Wang, R. Halldearn, et al., AIS data driven general vessel destination prediction: A random forest based approach, *Transp. Res. Part C Emerging Technol.*, **118** (2020), 102729. https://doi.org/10.1016/j.trc.2020.102729
27. B. Dadashova, B. Arenas-Ramires, J. Mira-McWillaims, K. Dixon, D. Lord, Analysis of crash injury severity on two trans-European transport network corridors in Spain using discrete-choice models and random forests, *Traffic Inj. Prev.*, **21** (2020), 228–233. https://doi.org/10.1080/15389588.2020.1733539
28. H. Bai, L. Li, Y. Wu, C. Liu, Z. Gong, G. Feng, et al., Study on the influence of meteorological elements on growing season vegetation coverage in Xinjiang, China, *Electron. Res. Arch.*, **30** (2022), 3463–3480. https://doi.org/10.3934/era.2022177

29. M. X. Xu, H. Y. Lin, Y. Liu, A deep learning approach for vehicle velocity prediction considering the influence factors of multiple lanes, *Electron. Res. Arch.*, **31** (2022), 401–420. https://doi.org/10.3934/era.2023020
30. W. R. Gao, Z. H. Bai, F. Zhu, C. C. Chou, B. H. Jiang, A study on the cyclist head kinematic responses in electric-bicycle-to-car accidents using decision-tree model, *Accid. Anal. Prev.*, **160** (2021), 106305. https://doi.org/10.1016/j.aap.2021.106305
31. Y. Meng, C. Untaroiu, Numerical investigation of occupant injury risks in car-to-end terminal crashes using dummy-based injury criteria and vehicle-based crash severity metrics, *Accid. Anal. Prev.*, **145** (2020), 105700. https://doi.org/10.1016/j.aap.2020.105700
32. J. Xu, S. Shang, G. Z. Yu, H. S. Qi, Y. P. Wang, S. C. Xu, Are electric self-balancing scooters safe in vehicle crash accidents?, *Accid. Anal. Prev.*, **87** (2016), 102–116. https://doi.org/10.1016/j.aap.2015.10.022
33. N. R. Garge, G. Bobashev, B. Eggleston, Random forest methodology for model-based recursive partitioning: the mobForest package for R, *BMC Bioinf.*, **14** (2013), 125. https://doi.org/10.1186/1471-2105-14-125
34. R. Genuer, J. M. Poggi, C. Tuleau-Malot, Variable selection using random forests, *Pattern Recognit. Lett.*, **31** (2010), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014
35. C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, *BMC Bioinf.*, **9** (2008), 307. https://doi.org/10.1186/1471-2105-9-307
36. K. J. Archer, R. V. Kirnes, Empirical characterization of random forest variable importance measures, *Comput. Stat. Data Anal.*, **52** (2008), 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015
37. C. Strobl, A. L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinf.*, **8** (2007), 25. https://doi.org/10.1186/1471-2105-8-25
38. C. Sammut, G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*, 2nd edition, Springer, New York, 2017.