Electronic Research Archive

ERA, 32(1): 584–607.
DOI: 10.3934/era.2024029
Received: 16 September 2023
Revised: 05 December 2023
Accepted: 27 December 2023
Published: 05 January 2024

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

Research article

Assessing crash severity of urban roads with data mining techniques using big data from in-vehicle dashcam

Nuri Park[1], Junhan Cho[2] and Juneyoung Park[3,*]

[1] Department of Smart City Engineering, Hanyang University, Ansan, Gyeonggi-do 15588, Korea
[2] Samsung Traffic Safety Research Institute, Samsung Fire & Marine Insurance, Seoul 06626, Korea
[3] Department of Transportation and Logistics Engineering/Smart City Engineering, Hanyang University, Ansan, Gyeonggi-do 15588, Korea

* **Correspondence:** Email: juneyoung@hanyang.ac.kr; Tel: +82314005151; Fax: +82314368147.

Abstract: The factors that affect the severity of crashes must be identified for pedestrian and traffic safety in urban roads. Specifically, in the case of urban road crashes, these crashes occur due to the complex interaction of various factors. Therefore, it is necessary to collect high-quality data that can derive these various factors. Accordingly, this study collected crash data, which included detailed crash factor data on the huge urban and mid-level roads. Using this, various crash factors including driver, vehicle, road, environment, and crash characteristics are constructed to develop a crash severity prediction model. Through this, this study identified more detailed factors affecting the severity of urban road crashes. The crash severity model was developed using both machine learning and statistical models because the insights that can be obtained from the latest technology and traditional methods are different. Therefore, the binary logit model, a support vector machine, and extreme gradient boosting were developed using key variables derived from the multiple correspondence analysis and Boruta-SHapley Additive exPlanations. The main result of this study shows that the crash severity decreased at four-street intersections and when traffic segregation facilities were installed. The findings of this study can be used to establish a traffic safety management strategy to reduce the severity of crashes on urban roads.

Keywords: crash severity model; in-vehicle dashcam video data; crash data; traffic safety; machine learning; urban road traffic management

1. Introduction

According to statistical reports from the Organization for Economic Cooperation and Development (OECD), the severity of crashes urban areas showed a gradual decrease until 2020. However, the number of deaths has seen a rapid increase after 2020. Consequently, a traffic safety strategy to reduce both the frequency and severity of crashes is required to achieve Vision Zero, which is the goal to reduce deaths and injuries resulting from traffic crashes to zero. On urban roads, there are many complex factors, including diverse road facilities, transportations, and signalized intersections, thereby affecting the occurrence of road crashes. Therefore, special attention must be paid to traffic safety on urban roads. One strategy for improving traffic safety in urban areas involves identifying contributing factors to crash severity and either implementing improvements or eliminating those factors. Previous studies predicted crash severity for traffic safety management on urban roads; based on severity prediction, they have sought to determine the main factors that contribute to crash severity [1–6]. Among these studies, there have been studies employed statistical models [1–3], and recently, studies have been conducted using machine learning models exhibiting remarkable predictive power [4,5]. Because both methodologies have different strengths and weaknesses in terms of data interpretability and model performance, efforts should be made to interpret the results of severity models using methodologies from both fields. A review study of various crash severity model development investigations based on methodologies has been conducted in one of the existing studies [6].

The existing national urban traffic crash database (DB) provided by the Korean National Police Agency can be subjective, as its data is collected by field personnel. Additionally, unlike highway crash data that provides detailed crash information, urban road crash data lacks detailed information except for a few categories. Thus, there are limitations in deriving the main factors that affect crash occurrence and severity in detail. Therefore, recently, studies have been conducted to capture the warning signs of crashes, develop crash severity models, and derive the main factors using the video data. Reportedly, these studies could afford new insights that could not be gained through approaches using existing crash data [7–11]. In this study, we aimed to derive the main factors that affect crash severity on urban roads and construct a model with high explanatory power using the video data of crashes from the dashcams installed in individual vehicles.

The crash DB was constructed by investigating dashcam video data collected for four months from January to April 2021; data from 381 crashes was collected, excluding those crashes wherein the personal information could not be collected. In this process, an information collection checklist was prepared to collect crash information manually and objectively. The crash information contains crash types, the personal information of offenders and victims, weather and road conditions at the time the crash occurred, and various geometries. Furthermore, information regarding the presence of road traffic facilities, detailed vehicle types, and whether the vehicles exceeded the speed limit at the time of the crash were collected. In this study, numerous variables were devised to develop crash severity prediction models.

However, an appropriate number of variables must be selected because too many variables may cause estimation errors in the regression model or machine learning (ML) development during the development of the crash severity model, thereby yielding incorrect results due to high correlations among the variables. Several recent studies have entailed feature selection based on various ML techniques. This study used two methodologies for deriving key variables: multiple correspondence analysis (MCA) and Boruta-SHapley Additive exPlanations (Boruta-SHAP). The key variables that

determined the characteristics of the dimensions within similar crash severities were derived through dimensional reduction and Boruta-SHAP; the additional key variables used for modeling were selected based on the variable importance that could be derived via ensemble techniques during modeling.

In previous studies, classification models for crash severity were constructed using statistical methods to identify crash severity-influencing factors. However, with current improvements in machine learning technology, the machine learning models can be interpreted through explainable artificial intelligence (XAI) techniques. Therefore, this study developed three crash severity models based on both statistics-based and ML-based models, including the binary logit model, a support vector machine (SVM) and extreme gradient boosting (XGBoost). By integrating the advantages of both statistical and ML-based models, this paper's approach aims to derive factors that effectively influence urban road crash severity.

The main contributions of this study are outlined as follows. First, in contrast to previous studies utilizing historical data that was collected after the crash occurred, this study applied in-vehicle dashcam data to capture detailed factors present at the crash scene. Through this approach, it was possible to examine the vehicle and road conditions at the precise time of the incident. Second, the study derived new insights and identified contributing factors to crash severity by analyzing the elements influencing crash severity through both statistical models and ML-based models.

The remaining chapters of this study are as follows. In Section 2, a literature review on the in-vehicle dashcam data used in traffic safety analysis and on the developing crash severity prediction model is provided. In Section 3, data and variables used in this study are described. The methodology for variable selection and modeling are introduced in Section 4. The results and discussion are summarized in Section 5. Finally, conclusions are summarized in Section 6.

2. Literature review

2.1. Research on traffic safety using in-vehicle dashcam data

Studies utilizing in-vehicle dashcam videos mainly aim to either obtain information or identify the precursors of crashes before they occur by reconstructing traffic crashes.

Giovannini et al. [12] analyzed dashcam videos in order to reconstruct traffic injury information. Taccari et al. [13] developed a model that classified crashes and crash risk events by using computer vision and convolutional neural networks. Paradana [14] developed an algorithm that predicted crash risks from a first-person perspective using in-vehicle dashcam video data. Moreover, the developed algorithm predicted the risk of traffic crashes by recognizing objects (e.g., vehicles, pedestrians, and obstacles) from videos and calculating the safety distance between them. Hairi and Fradi [15] identified crash risks based on vision transformer and artificial intelligence (AI) techniques. They converted video data into images for analysis and used the dashcam accident dataset (DAD) as video data.

2.2. Research on crash severity model development using video data

In-vehicle dashcam video data are also referred to as vehicle black-box data, dashcam videos, and in-vehicle videos. These data can aid in the development of crash severity models because they contain useful data such as images and each vehicle's velocity profiles.

Song et al. [7] used the video images and speed profile information extracted from in-vehicle

dashcams for their study and derived the key factors that affected pedestrian crash severity in the event of a taxi–pedestrian accident on urban arterial roads. Chung [8] used vehicle black-box data and crash DB provided by the Korean National Police Agency to analyze the severity of taxi-pedestrian crashes. An ordered probit model was constructed using crash severity as a dependent variable; the model construction results were examined based on the average marginal probability effects. In a similar study, Chung [9] developed a crash severity model for taxis and two-wheelers by using vehicle black-box data and crash data from the Korean National Police Agency. Using the model, the key factors that contributed to an increase in the severity of crashes between taxis and two-wheelers were derived. Their results indicated that an increase in the collision speed contributed to more serious crashes. Cho et al. [10] collected personal information and information on traffic conditions and crash situations from the dashcam video data at the time of traffic crashes collected from highways. Using the collected crash data, they developed crash severity prediction models based on a cluster analysis and a binary logit model. Loo et al. [11] conducted research to predict the bus crash frequency at various crash severity levels using dashcam video data and GPS data collected from buses. They extracted distance and behavioral factors from bus dashcam videos across the city by using a deep learning-based computer vision methodology for the analysis.

Table 1. Previous studies in crash severity analysis using black box video data.

Author	Data	Variable	Methodology
Song et al. [7]	black box video data	pedestrian, crash, and driver characteristics (4 severity levels)	multiple indicator and multiple cause (MIMIC) model
Chung [8]	vehicle black-box (VBB) data and crash DB	time, time to collision (TTC), speed, location, and crash characteristics collected from the VBB, vehicle, pedestrian, environmental, road, and crash characteristics collected from the crash DB (4 severity levels)	ordered probit model
Chung [9]	in-vehicle video recording (IVVR)	taxi, two-wheeled vehicle, environmental, road, two-wheeled vehicle rider, and crash characteristics including taxi speed, helmet wearing, etc. (3 severity levels)	ordered probit model
Cho et al. [10]	black box video data	road, crash, and driver characteristics	latent class analysis and binary logit model
Loo et al. [11]	bus dashcam video and GPS data	demographic data, risk factors, pedestrian exposure coefficient, pedestrian jaywalking index, bus stop congestion, sidewalk railings, and etc.	negative binomial model, random forest, and XGBoost

The use of dashcam video data is suitable for analyzing factors that contribute to crash severity in remarkable detail, as it facilitates the collection and utilization of more accurate information at the time of crash as well as the collection of various data, such as speed profiles, in contrast to the traffic crash DB provided by public agencies. However, in previous studies, additional information, such as detailed offender and victim vehicle type, was not collected and used for detailed crash severity analyses. In regard to the methodology, statistical methods were previously used; recent studies have

employed various latest ML techniques to improve the explanatory power and accuracy of crash severity analyses. Additionally, no study thus far has focused on urban roads. Table 1 summarizes studies on crash severity that used dashcam video data.

2.3. Research on deriving influencing factors for urban crash severity based on ML

In this study, a traditional statistical technique, such as the binary logit model, and ML techniques, such as SVM and XGBoost, were used to develop crash severity models with a high accuracy and explanatory power and derive factors that affect the severity. Such ML techniques have been used in previous studies on factors that affected the crash severity.

Table 2. Previous studies on the developing crash severity model using ML techniques.

Author	Data	Variable	Methodology
Mussone et al. [16]	detector, traffic crash, and weather data collected every 5 minutes (total of 1838 crashes)	vehicle, driver, and environmental characteristics	back-propagation neural network, generalized linear mixed model
Iranitalab et al. [17]	vehicle-to-vehicle crash data (4 years, total of 68,448 crashes)	road, driver, vehicle, land use, crash, and environment characteristics	Multinomial Logit (MNL), Nearest Neighbor Classification (NNC), SVM and RF
Mafi et al. [18]	crash data (5 years, total of 32,730 crashes)	driver, road, traffic, environmental, vehicle, and crash characteristics	C4.5 algorithm, RF, nearest-neighbor instance-based
Liu [19]	crash data (6 years)	various variables including traffic control, weather, lighting, road characteristics etc.	XGBoost, AdaBoost, RF, Gradient Boost Decision Tree, SVM, KNN
Islam et al. [20]	crash data (4 years, total of 4093 crashes)	faulty tires, shock moving veh, not giving way, etc.	RF, XGBoost, GIS spatial autocorrelation analysis
Yan et al. [21]	crash data (4 years, total of 30,426 crashes)	traffic. temporal, weather characteristics, and points of interest (POI, crosswalk, etc.)	hybrid model integrating RF and Bayesian Optimization
Afshar et al. [22]	rural road crash data (5 years)	traffic, vehicle, land use, temporal, environmental characteristics	Extremely Randomized Tree
Alrumaidhi et al. [23]	elderly driver accident data (8 years)	crash, traffic signal, weather, road, construction, etc.	logistic regression, linear discriminant analysis, RF
Astarita et al. [24]	crash data (2 years, total of 202 crashes)	traffic, road, driver, environmental, crash characteristics	artificial neural network (ANN), hybrid grey wolf optimization-based ANN

Mussone et al. [16] analyzed factors that affected crash severity in urban intersections by using the vehicle detection system data, traffic crash data, and weather data. They developed crash severity models by utilizing neural networks (NN) and generalized linear mixed models. Iranitalab et al. [17] developed a crash severity model by using vehicle-to-vehicle crash data and compared four statistical

and ML methodologies for crash severity prediction. Particularly, they investigated the effects of unsupervised learning-based clustering methodologies on the performance of a crash severity prediction model. Mafi et al. [18] analyzed crash severity according to age and gender in terms of accident cost by using various ML methodologies, including random forest (RF) and NN techniques. Liu [19] utilized ensemble techniques, such as XGBoost and Adaptive Boosting tree (AdaBoost), and ML models, such as SVM and K-Nearest Neighbors (KNN), to predict the severity of large truck crashes. They found that the gradient boost decision tree methodology yielded the highest performance in predicting crash severity. Islam et al. [20] derived factors that affected crash severity on urban roads by using RF and XGBoost. Additionally, they conducted research on identifying crash hotspots through a spatial autocorrelation analysis. Yan et al. [21] developed an urban road crash severity model using a hybrid model that combined RF and Bayesian optimization and identified the important factors that affected crash severity. Afsher et al. [22] analyzed the crash severity in rural areas by using extremely randomized trees and analyzed models based on a feature importance analysis, partial dependence plots, and individual conditional expectations. Alrumaidhi et al. [23] utilized ML models to predict the severity of elderly driver crashes. Astarita et al. [24] developed an urban road traffic crash severity model using AI techniques and conducted a sensitivity analysis to determine the most important variables that affected the road crash severity level.

Crash severity models have been developed using ML techniques for various road and crash types, and studies have been conducted to derive the main factors that affect crash severity. Most research results showed that ML-based methodologies improved the predictive performance of models. Table 2 summarizes previous studies on the development of crash severity models using ML techniques.

2.4. Summary of the literature review

This study reviewed research on traffic safety using in-vehicle dashcam data, research on the crash severity model development using video data, and research deriving the influencing factors of urban crash severity based on ML.

Previous studies mainly utilized dashcam data to reconstruct crash situations; since possibilities for data collection have increased, recent efforts have emerged to derive crash severity factors from dashcam data. However, these studies have limitations since they focus on either highways or specific vehicles such as taxis and buses; thus, they do not generally consider crashes that occur in urban areas. With the advancement of technology, ML techniques have recently been widely applied to crash severity analyses, and in most cases, only limited variables were utilized from the historical crash data. Additionally, previous studies do not devote much effort in identifying key variables that influence crash severity on urban roads among the large number of variables.

In this study, dashcam video data were collected for crashes that occurred on urban huge-level and mid-level roads, and numerous variables including detailed vehicle type information were devised to encompass information that may affect the severity, such as car body types. Furthermore, various ML techniques were applied to variable selection, model development, and model interpretation to develop crash severity prediction models with a high accuracy and explanatory power compared to traditional statistical models and to derive key factors that affect crash severity. The results were comparatively analyzed to derive the main factors that affect crash severity on urban roads and identify the models that are suitable for crash severity model development and exhibit high predictive performance.

3. Data description

In this study, video data before the crash occurrence and at the time of the crash were collected through the in-vehicle dashcam. Then, detailed contents on crash situations were collected by preparing a checklist for the driving environment, road conditions, and driving situation (Figure 1). For instance, researchers collected data by reviewing a single dashcam video and checked crash-related road, traffic, environmental, and crash information from the video against the items on a checklist. However, personal information such as "age", "gender" and "vehicle type" was collected using data provided by the insurance companies that manage dashcam video data.

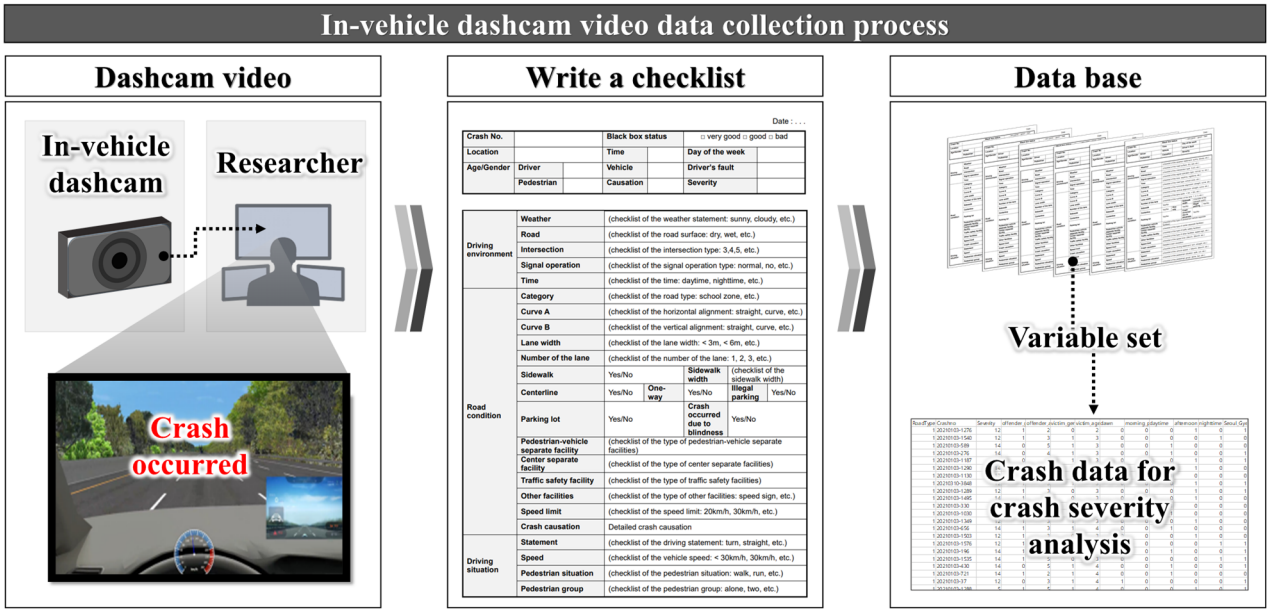

Figure 1. Dashcam video data collection process.

The temporal range of the collection was from January to April 2021, and the spatial range was set to urban huge-level and mid-level roads across the country. The spatial range of this study is shown in Figure 2. Overall, data from 417 crashes were collected. Among them, data without basic information were excluded (e.g., time and information). Finally, although the offender's and victim's vehicle speed information were utilized as independent variables, the traffic flow was not considered in this study. Thus, it will be possible to estimate the traffic on the offender. Consequently, data from 381 crashes were used for modeling. The duration of each dashcam video data ranged from 30 seconds to 1 minute, thus encompassing the period from immediately before the crash occurrence until the actual crash. The crash severity ranged from 1 to 16, as supplied by an insurance company. Injuries were classified into grades 1 to 14, with 15 denoting a death at the crash location and 16 indicating a death during treatment. For the injury level (1 to 14), a lower number indicates a high severity. In the crash data collected for this study, since the severity levels range only includes 1 to 14, the levels were categorized into two groups: 1 to 13 were classified as serious crashes, while level 14 was categorized as relatively less serious crashes.

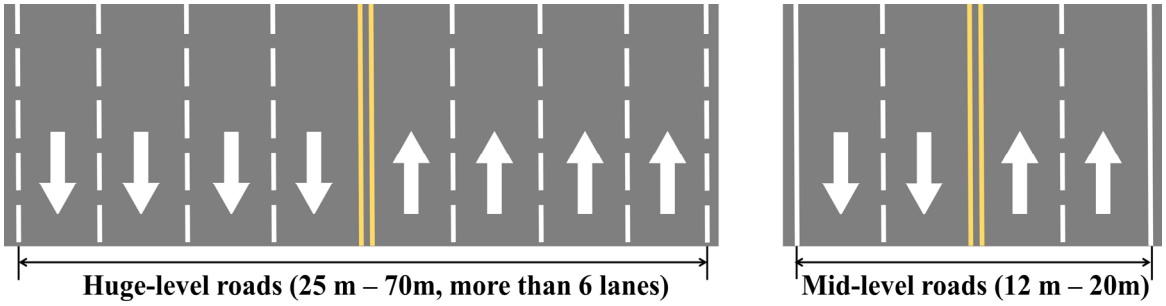

Figure 2. Huge-level roads and mid-level roads.

The data used in this study include unreported traffic crashes that are not included in the existing police traffic crash data. It was possible to collect relatively accurate and objective crash scene information compared to the crash DB of the Korean National Police Agency by examining the video data. Collected crash scene information included the following: the driving behavior of the offenders and victims immediately before crashes, alongside whether the speed limit was exceeded and whether there was a turn; environmental conditions that included the weather at the time of the crash; road alignment information (e.g., horizontal alignment and vertical alignment); road characteristics, such as types of intersections, special roads (e.g., school zones, crosswalks, tunnels and underground roads), road surface condition, and one-way traffic; and road facility information, such as sidewalks, centerlines, on-road parking areas, facilities for pedestrian-vehicle segregation, safety facilities and other facilities.

Table 3. Variable description.

Category	Variable	Descriptions
Driver characteristic		
Offender driver		
Offender gender	Male, Female	1 if the offender driver is a male
Offender age	Youth (< 19), Younger (19–29), Middle-aged (30–49), Older (50–64), Others (> 64)	Categories 1 to 5 from Youth to Others
Offender vehicle type	Sedan, SUV, Hatchback, Two wheel, Truck, Wagon, Coupe, Bus, Others	1 if offender vehicle type is each vehicles (Sedan, SUV, …)
Victim driver		
Victim gender	Male, Female	1 if the victim driver is a male
Victim age	Youth (< 19), Younger (19–29), Middle-aged (30–49), Older (50–64), Others (> 64)	Categories 1 to 5 from Youth to Others
Victim vehicle type	Sedan, SUV, Hatchback, Two wheel, Truck, Coupe, Bus, Bicycle, Others	1 if victim vehicle type is each vehicles (Sedan, SUV, …)
Crash information		
Time	Dawn (0:00–6:59), Morning peak (7:00–8:59), Daytime (9:00–16:59), Afternoon (17:00–19:59), Night time (20:00–23:59)	
Week	Weekday, Weekend	1 if the crash occurred at weekday

Continued on next page

Category	Variable	Descriptions
Region	Seoul_Gyeonggi, Gangwon, Chungcheong, Jeolla, Gyeongsang, Jeju	1 if crash occurred at each region (Seoul_Gyeonggi, Gangwon, …)
Road type	Wide road, Middle road	1 if the crash occurred at wide road
Crash type	Side collision, Head-on collision, Others	1 if the crash type is each type
Multiple crash	Multiple crash, Non-multiple crash	1 if the multiple crash
Offender vehicle behavior		
Vehicle driving behaviour	straight, back, lane change, left turn, right turn, parking, u-turn, others	1 if the vehicle driving behaviour is each type
Speed	Unclassified, Low, Middle, High speed	Categories 0 to 4
Turn signal	No, Yes, Unclassified	Offender turn signal (yes = 1, no = 0)
Victim vehicle behavior		
Vehicle driving behaviour	straight, back, lane change, left turn, right turn, parking, u-turn, crossing, others	1 if the vehicle driving behaviour is each type
Speed	Unclassified, Low, Middle, High speed	Categories 0 to 4
Turn signal	No, Yes, Unclassified	Victim turn signal (yes = 1, no = 0)
Road characteristic		
Road surface	Dry, Wet, Frost/Freezing, Snow, Unclassified	Road surface state (yes = 1, no = 0)
Intersection	Three-legged, Four-legged, Five-or-more-legged, Roundabout, Other intersection, Segment	Intersection type (yes = 1, no = 0)
Special road type	children's zone, school zone, crosswalk, tunnel, underpass, silver zone, Others	Special road type (yes = 1, no = 0)
Horizontal curve	Straight section, Right curve section, Left curve section, Others	Horizontal curve type (yes = 1, no = 0)
Vertical curve	Flat road, Uphill road (low), Uphill road (high), Downhill road (low), Downhill road (high), Others	Vertical curve type (yes = 1, no = 0)
Number of lane (1–12)		Number of lane
Existence of the outside lane	No, Yes, Unclassified	-
Existence of the one-way lane	No, Yes, Unclassified	-
Existence of the illegal parking	No, Yes, Unclassified	-
Existence of the on-street parking	No, Yes, Unclassified	-
Road facility		
Traffic lights	Normal traffic light, Flashing yellow, Flashing red, Flashing yellow and red, Others	Traffic lights state
Existence of the sidewalk	No, Yes, Unclassified	-
Width of the sidewalk	Unclassified, Less than 2m, More than 2m	Sidewalk width

Continued on next page

Category	Variable	Descriptions
Existence of the center line	No, Yes, Unclassified	-
Pedestrian-vehicle segregation	Curb only, Fence only, Mark only, Tree only, Curb and fence, Curb and tree, Curb and marking, Curb & fence & tree, Curb & fence & marking, Curb & fence & mark & tree, No	Pedestrian-vehicle segregation type in crash occurred location
Center segregation	Flowerbed, Median strip, Road sign, No, Unclassified	Center segregation type in crash occurred location
Safety facility	CCTV only, Marking only, Speed sign only, CCTV and Marking, Marking and Speed bump, CCTV and Speed sign, Marking and Speed sign, No	Safety facility type in crash occurred location
Traffic characteristic		
Speed limit	Others, 30, 40, 50, 60, 70, > 70	Categories 0 to 6
Environment characteristic		
Weather	Sunny, Cloudy , Rainy, Snowy, Unclassified	1 if the crash occurred in each weather

Additionally, variables were also set for traffic conditions (e.g., speed limit) and personal information (e.g., ages and genders of victims and offenders). Moreover, information on the vehicle type of the victims and offenders was collected because the crash severity may vary depending on the vehicle size and shape. The vehicle type was classified into seven categories (sedan, SUV, hatchback, two-wheel, truck, wagon and coupe) to be reflected in the analysis. The variables used in this study were constructed as either categorical or continuous variables according to the characteristics of the data. Table 3 shows the variables and their description.

4. Methodology

In this study, various road, traffic, environmental, and human factor variables for the crash severity prediction were constructed using dashcam video data to develop crash severity models on each urban huge-level and mid-level roads to analyze factors that affected the severity. Additionally, key variables for the development of a severity prediction model were selected through ML and statistics-based feature selection.

In previous studies, various methods were employed for feature selection. Among these, researchers sought to identify key variables that effectively explained the data using dimensionality reduction methods [25]. Some studies utilized ensemble techniques among ML techniques to assess the importance of variables in the model learning process. Results of variable importance were calculated and applied for variable selection [26,27].

In this study, we incorporated both methodologies for key variable selection to bring the advantages of each and mitigate bias in the variable selection process. To develop models with a high accuracy and explanatory power, prediction models based on SVM and XGBoost were developed along with binary logistic models, which are commonly used for binary classification [17–21,23]. The

ML-based prediction model development results were analyzed through an XAI technique. Figure 3 shows the overall research flow of this study.

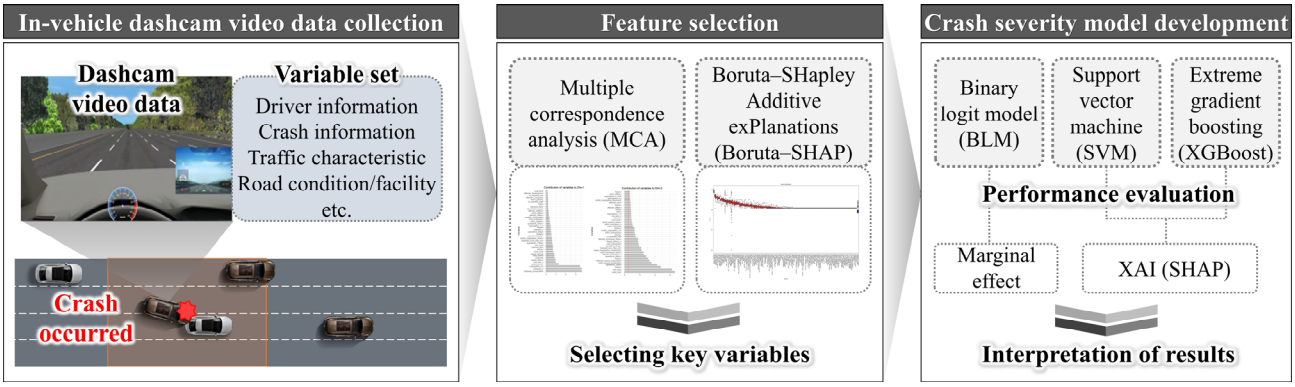

Figure 3. Research framework.

4.1. Feature selection

This paper employed both the MCA and the Boruta-SHAP methodologies to extract key independent variables for developing the crash severity model. Through MCA, data dimensions can be reduced, and variables that effectively explain each dimension can be identified. In other words, this technique allows for the identification of variables that efficiently explain the data. The Boruta-SHAP method is based on ensemble techniques and can select variables that significantly influence the prediction performance of the model. Therefore, in this study, both methodologies were adopted and utilized to consider both variables that can efficiently explain the data and those that have a major influence on predicting the crash severity. Through MCA, the top 20 variables with high R–square values for each dimension were extracted. Furthermore, the methodology was configured to additionally reflect variables deemed significant based on the Boruta-SHAP variable importance assessment.

4.1.1. Multiple correspondence analysis (MCA)

A principal component analysis (PCA) is a methodology used to reduce the dimensions of continuous data, whereas an MCA is a methodology used to reduce the dimensions of categorical data (nominal data). As with PCA and correspondence analyses, MCA can be considered a multivariate methodology that can analyze systematic variation patterns by using categorical data.

An MCA can determine the number of dimensions through the scree plot, which is a graph that shows the change in the dispersion of the eigenvalue and principal components, to select the number of principal components. Typically, the number of dimensions is chosen up to the point where the graph levels off. In this study, the number of principal components was set to two for the analysis, as it decreases the change rate of the eigenvalue, thereby indicating that there are two significant dimensions, as shown in Figure 4.

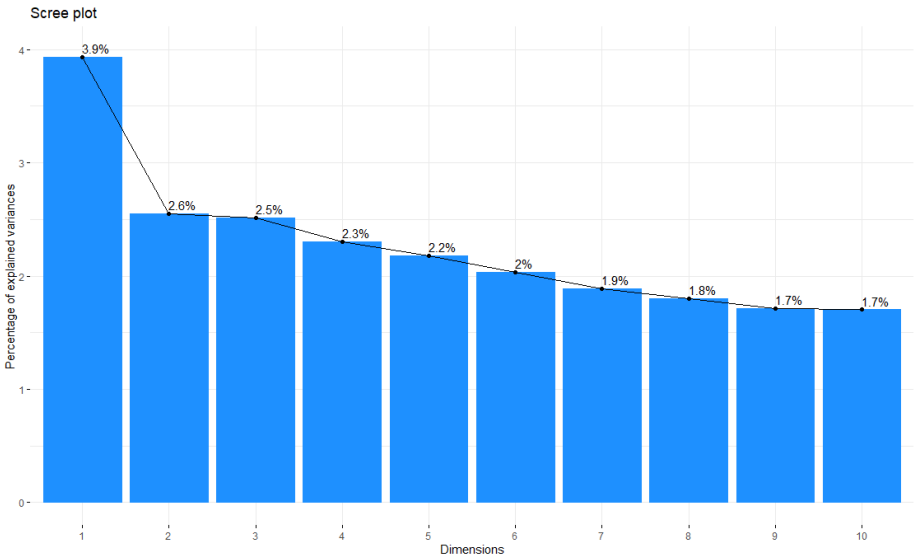

Figure 4. Scree plot results.

4.1.2. Boruta-SHapley additive exPlanations (Boruta-SHAP)

The Boruta, which is a feature selection methodology for choosing key variables during model development, is an improved technique to derive the importance of variables that are calculated from RF [26,27]. In other words, Boruta is a variable selection methodology based on ensemble techniques (e.g., RF and XGBoost) that removes irrelevant variables through multiple iterations. Unlike RF-based feature selection methods, wherein the importance rankings may vary depending on the iteration, Boruta exhibits relatively small variations in importance rankings because the importance is calculated after multiple iterations. In this study, the key variables and their influence were derived using the Boruta-SHAP methodology that combined the Boruta technique and SHAP, which is one of the XAI techniques.

4.2. Crash severity prediction

4.2.1. Binary logit model (BLM)

The logit model is a probability distribution model that uses a logistic distribution. In this study, a binary logit model was used, and fatal crash and non-fatal crash events were classified. In the logit model, the suitability of the model can be tested through the log-likelihood ratio test and pseudo-R2. It is possible to identify significant variables and their effects on model prediction by testing the significance probability of variables. If the logit model is expressed using an equation, the probability of predicting the dependent variable as 1 (i.e., a fatal crash) is given by Eqs (1) and (2), where Y refers to the dependent variable (crash severity) and x refers to the independent variable (traffic, road, environment, and crash information variables). Each β_x are model parameters.

$$P(Y = 1|x) = \frac{\exp(f(x))}{1+\exp(f(x))} \tag{1}$$

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_n x_n \quad (2)$$

4.2.2. Support vector machine (SVM)

SVM, an ML technique, is a supervised learning model that can either be used for classification or a regression analysis [28]. SVM classifies data by generating hyperplanes using the margin, and data in different categories are classified by adopting a hyperplane that maximizes the margin. Here, the margin is the distance between the hyperplane and support vectors. The hyperplane is a plane that classifies the data, and support vectors are points that are closest to the hyperplane.

In this study, the tuned SVM hyperparameters are as follows: kernel, which is a function for mapping low-dimension data to high dimensions; C, which indicates the error tolerance; and the gamma value, which can adjust the precision of the decision boundary. In this process, considering that overfitting may occur, the gamma value was fixed at 0.01 and the C was fixed at 100.

4.2.3. Extreme gradient boosting (XGBoost)

XGBoost, an ensemble technique, is a boosting technique-based methodology that creates strong classifiers from weak classifiers by combining multiple decision trees, unlike RF, which is a bagging technique [29,30]. XGBoost, a type of gradient boost, exhibits a high speed and high prediction reliability because the training and estimation are performed through parallel processing. However, it is sensitive to the parameter setting and overfitting may occur if the number of samples is small. Therefore, attention must be paid to the setting of XGBoost hyperparameters. In this study, the maximum depth (i.e., the number of trees) was set among multiple XGBoost hyperparameters. Additionally, a binary logistic classification was set as a learning parameter because a binary classification model had to be constructed. Table 4 presents the XGBoost hyperparameters.

Table 4. XGBoost hyperparameters.

Road type	Maximum depth	The number of decision trees	Objective
mid-level roads	3	100	Binary logistic
huge-level roads	10	100	Binary logistic

4.2.4. Explainable artificial intelligence (XAI)

As ML techniques are black-box models, it is not possible to determine influential variables within the algorithm. Ensemble techniques can derive the importance of variables because they can calculate the impurity during node generation; however, they cannot determine whether the variables have positive or negative effects. Therefore, in this study, XAI was used for interpreting ML models. XAI has various methodologies, such as local interpretable model-agnostic explanation (LIME) and SHAP. In this study, the SHAP methodology was used. The SHAP algorithm measures changes in the Shapley value according to the presence/absence of variables. It can identify the contribution and importance of variables along with their degree of influence and whether they have positive or negative effects based on the Shapley value [31].

5. Results and discussion

5.1. Feature selection results

An MCA was employed to identify variables that explain the data itself, while Boruta-SHAP was utilized to derive variables that have a major impact on the model when modeling crash severity. In this study, two methodologies were used to select variables that reflect the advantages of both methodologies.

In the MCA results, the top 20 variables that contributed to each dimension were derived based on R-squared values, as shown in Table 5. The R-squared indicates how highly correlated a variable is with a given dimension. The 20 derived variables can effectively explain the dimension and crash data.

Table 5. MCA results with R-squared.

Dimension 1		Dimension 2	
Variable	R^2	Variable	R^2
Width of the sidewalk	0.631	Offender behavior (Turn signal-No)	0.343
Sidewalk (No)	0.622	Offender behavior (Turn signal-Unclassified)	0.317
Sidewalk (Yes)	0.607	Victim behavior (Turn signal-Unclassified)	0.217
Pedestrian-vehicle segregation (No)	0.599	Victim behavior (Turn signal-No)	0.162
Intersection (segment)	0.345	Road surface (Dry)	0.154
Traffic lights (Others)	0.335	Offender behavior (Straight)	0.150
Traffic lights (Normal traffic light)	0.294	Number of lane	0.137
Speed limit	0.273	Speed limit	0.110
Existence of the center line (No)	0.247	Existence of the center line (Yes)	0.109
Pedestrian-vehicle segregation (Curb only)	0.218	Weather (Snowy)	0.105
Intersection (Four-legged intersection)	0.213	Victim behavior (Speed)	0.102
Existence of the one-way lane (Yes)	0.200	Week	0.098
Existence of the center line (Yes)	0.193	Offender behavior (left turn)	0.094
Existence of the one-way lane (No)	0.179	Road surface (Frost/Freezing)	0.093
Number of lane	0.139	Existence of the one-way lane (Yes)	0.093
Special road type (Road tunnel)	0.100	Existence of the center line (No)	0.086
Center segregation (Unclassified)	0.082	Pedestrian-vehicle segregation (Curb & mark)	0.086
Victim behavior (Victim-left turn)	0.073	Crash type (Head-on collision)	0.084
Victim behavior (Victim-right turn)	0.070	Offender behavior (Speed)	0.081
Width of the sidewalk	0.631	Offender behavior (Turn signal-No)	0.343

Among the behaviors of the offender's and the victim's vehicles, variables related to turning were found to be able to explain the dimensionality of the data, and speed-related variables such as the vehicle's driving speed and speed limit were also in the top 20 R-square values. In terms of road facilities and road conditions, variables such as sidewalk, pedestrian-vehicle segregation, and traffic lights were found to explain each dimension, and environmental variables such as road surface, weather, and week were also found to explain the data.

In addition to the MCA, Boruta-SHAP was used in this study to find key variables that can be additionally reflected, and 500 iterations were performed. The Boruta-SHAP results are shown in

Figure 5. The variable of pedestrian-vehicle segregation using curb-only and the speed limit were accepted in Boruta-SHAP.

In Boruta-SHAP, except for the two variables in Figure 5, the remaining variables were not deemed significant. However, when considering the MCA results collectively, it was found that 32 variables exhibited a high correlation with the data and its dimensions. These variables were utilized to develop a crash severity prediction model.

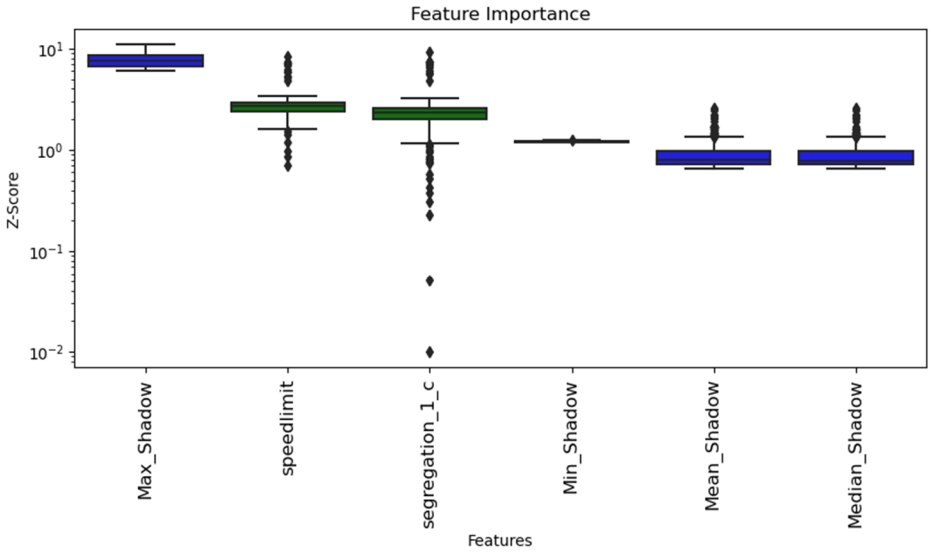

Figure 5. Boruta-SHAP results.

5.2. Crash severity prediction results

For crash severity models, crash severity was predicted for urban huge-level and mid-level roads because road and traffic characteristics were expected to be different depending on the road type. Based on the confusion matrix for the crash severity prediction results, predictive performance indicators, such as accuracy, recall, precision, and F1 score, were calculated. The accuracy of the SVM was found to be 0.696 for huge-level roads, which was higher compared to the BLM and XGBoost. In addition, the accuracy of the SVM was found to be 0.695 and 0.696 for mid-level roads and huge-level roads, respectively, which were higher compared to other models. Table 6 compares the predictive performances.

Table 6. Crash severity predictions performance results.

Model	Road type	Accuracy	Recall	Precision	F1 score
BLM	mid-level roads	0.509	0.667	0.294	0.408
	huge-level roads	0.661	0.400	0.375	0.387
SVM	mid-level roads	0.695	0.467	0.412	0.438
	huge-level roads	0.696	0.533	0.444	0.485
XGBoost	mid-level roads	0.644	0.313	0.333	0.323
	huge-level roads	0.679	0.400	0.400	0.400

Therefore, the BLM results and the SVM model results were interpreted to derive factors that

contributed to the severity of crashes that occurred in urban huge-level and mid-level roads by comparing the influencing factors derived from both statistical and ML models. As the SVM model is a black-box model, the contribution and effects of the variables used in the model were interpreted through SHAP, which is an XAI technique.

Table 7. Crash severity prediction results with BLM.

Mid-level roads	Mid-level roads			Huge-level roads		
Variable	coef	std error	P > |z|	coef	std error	P > |z|
Crash type (Head-on collision)	**0.43**	**0.26**	**0.10***	-0.282	0.278	0.310
Exist of the center line (No)	-0.13	0.57	0.83	-2.917	-	1.000
Exist of the center line (Yes)	-0.67	0.46	0.14	-2.407	-	1.000
Center seg (Unclassified)	0.75	0.49	0.12	-3.549	-	0.999
Intersection (Four-legged)	-0.27	0.29	0.35	0.366	0.303	0.227
Number of lane	-0.18	0.29	0.52	0.335	0.242	0.166
Offender behavior (left turn)	-0.13	0.30	0.67	0.101	0.291	0.728
Offender (Speed)	-0.01	0.29	0.97	0.372	0.264	0.159
Offender (Straight)	-0.38	0.27	0.16	-0.259	0.307	0.400
Offender (Turn sig-No)	0.07	0.65	0.91	0.222	0.747	0.766
Offender (Turn sig-Unclass)	-0.33	0.62	0.59	0.099	0.719	0.890
Exist of the one-way lane (No)	0.07	0.54	0.90	-11.716	-	1.000
Exist of the one-way lane (Yes)	-0.54	0.62	0.39	-	-	-
Road surface (Dry)	-0.99	0.49	0.04	2.812	-	1
Road surface (Frost/Freezing)	-0.35	0.40	0.38	5.492	-	1
Special road type (tunnel)	**-0.89**	**0.38**	**0.02***	-	-	-
Pede-veh seg (Curb only)	**0.60**	**0.31**	**0.05***	0.257	0.318	0.419
Pede-veh seg (Curb & mark)	-0.16	0.28	0.57	0.113	0.319	0.722
Pede-veh (No)	0.14	0.65	0.83	-0.599	0.477	0.209
Sidewalk (No)	0.40	1.13	0.73	0.396	-	1
Width of the sidewalk	-0.73	0.46	0.11	0.809	0.684	0.237
Sidewalk (Yes)	0.65	0.89	0.47	-0.397	-	1
Traffic lights (Normal)	**-1.30**	**0.58**	**0.03***	-0.389	0.617	0.529
Traffic lights (Others)	**-1.15**	**0.66**	**0.08***	-0.330	0.634	0.602
Weather (Snowy)	**-0.86**	**0.35**	**0.01***	-	-	-
Speed limit	0.25	0.25	0.33	0.383	0.347	0.269
Victim (left turn)	-0.46	0.29	0.12	**0.634**	**0.348**	**0.069***
Victim (right turn)	-0.10	0.24	0.67	0.013	0.324	0.968
Victim (Speed)	-0.18	0.28	0.53	0.068	0.261	0.794
Victim (Turn sig-No)	0.48	0.50	0.33	0.489	0.513	0.341
Victim (Turn sig-Unclassified)	0.56	0.54	0.30	-0.025	0.458	0.957
Week	-0.14	0.27	0.61	-0.534	0.402	0.184

Note: *: Significant variable in a significance probability of 90%

**: Center seg = Center segregation; Pede-veh seg = Pedestrian-vehicle segregation

In the BLM results, the pseudo-R-squared was found to be 0.184 and 0.08 for huge-level roads and mid-level roads, respectively, thus indicating that the model has an explanatory power of approximately 10 to 20%. When significant variables were derived based on a significance probability of 90%, the variables 'Crash type (Head-on collision)', 'Special road type (tunnel)', 'Pedestrian-vehicle segregation (Curb only)', 'Traffic lights (Normal)', 'Traffic lights (Others)' and 'Weather (Snowy)' were significant for mid-level roads. In the huge-level roads model, only 1 variable, namely 'Victim (left turn)', was a significant variable. Table 7 shows the overall BLM coefficient estimation results.

According to the detailed analysis results for the mid-level road model, the probability of high-crash severity decreased in tunnel sections and snowy weather. Since the mid-level road tunnel is a relatively short segment with no vulnerable road users, such as pedestrians or bicycles, it can be inferred that the probability of a high-crash severity in this tunnel section is low. Furthermore, while the probability of a high-crash severity typically increases during adverse weather conditions, an accurate interpretation becomes challenging in this dataset. This is due to the limited occurrence of crashes in snowy weather, with only 10 out of 212 mid-level road crashes happening under snowy weather conditions. Additionally, the probability of high-crash severity was found to decrease at the normal traffic lights signalized intersection. Moreover, high-crash severity probability increased when there was only one pedestrian-vehicle segregation facility, namely the curb. This suggests that road markings and trees, as well as curbs, should be installed as a pedestrian-vehicle segregation facility to effectively reduce the probability of high-severity crashes occurring. In the detailed analysis results for huge-level roads, a probability of a high-crash severity was found to increase when the victim's vehicle attempted a left turn.

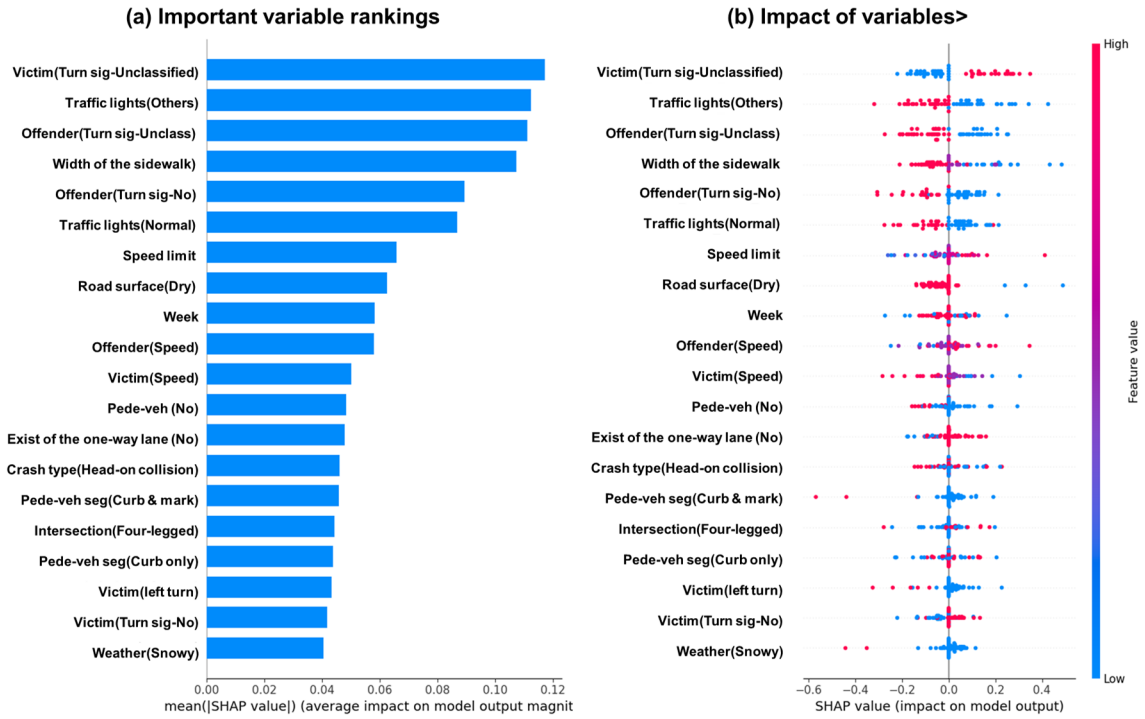

Figure 6. SHAP results in huge-level roads model.

The results of interpreting the SVM results using SHAP and the visualization results are shown

in Figures 6 and 7.

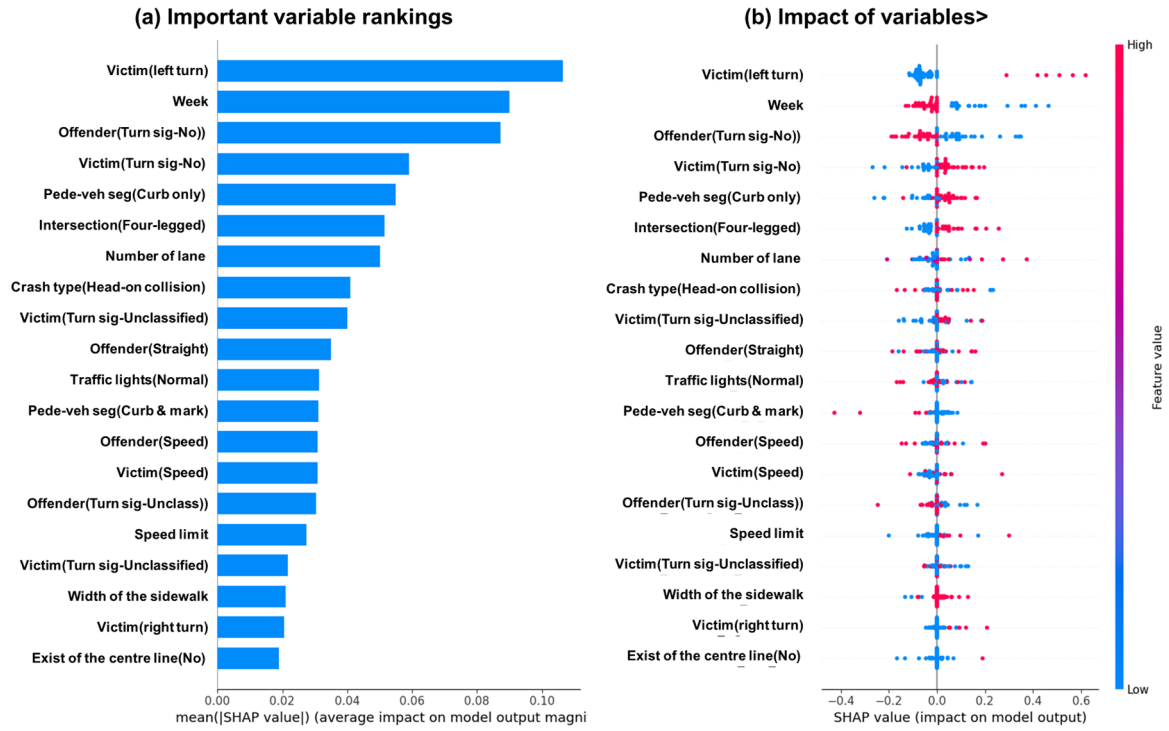

Figure 7. SHAP results in mid-level roads model.

In the SHAP result graph, the blue bar graphs (Figures 6(a) and 7(a)) are presented in the order of variables that had a significant influence on predicting the crash severity. The magnitude of the SHAP value indicates the level of importance: the larger the value, the more influential the variable. The scatter plot graph (Figures 6(b) and 7(b)) illustrates the impact of each variable in predicting the crash severity, with red indicating higher values. Values distributed to the right side of the plot indicate a greater influence on the model prediction. For instance, in the case of the 'Victim (Turn Signal-Unclassified)' variable in Figure 6(b), a larger value (i.e., indicating greater difficulty in checking the turn signal of the victim's vehicle) can be interpreted as having a positive effect on predicting a higher crash severity.

Table 8. Marginal effect results.

Mid-level roads			Huge-level roads		
Variable	Marginal effect	P > |z|	Variable	Marginal effect	P > |z|
Crash type(Head-on collision)	0.076	0.082	Victim (left turn)	0.097	0.057
Special road type(tunnel)	-0.158	0.012	-	-	-
Pede-veh seg(Curb only)	0.105	0.043	-	-	-
Traffic lights(Normal)*	-0.230	0.018	-	-	-
Traffic lights(Others)*	-0.203	0.069	-	-	-
Weather(Snowy)*	-0.152	0.007	-	-	-

When examining the detailed results of the mid-level model predictions, it was observed that

'Traffic lights (Others)', 'Offender (Turn Signal-Unclassified)', 'Width of the sidewalk', and 'Offender (Turn sig-No)' had negative effects on the crash severity prediction. Conversely, 'Victim (Turn Signal-Unclassified)' had a positive effect on the crash severity prediction.

For the huge-level road model, the top five variables that are important for crash severity classification were found to be 'Victim (left turn)', 'Week', 'Offender (Turn sig-No)', 'Victim (Turn sig-No)', 'Pedestrian-vehicle segregation (Curb only)'. Using the SHAP graph, the effects of each variable on the crash severity classification were examined in detail. It was found that the 'Victim (left turn)' affected the predicted crash severity values with a positive correlation. Moreover, other variables were found to have a negative effect.

Table 9 presents the integration of results from the BLM and SVM models, which displays the key variables obtained from each model and their respective impact on the crash severity, categorized into mid-level and huge-level roads.

Table 9. Integrated results of crash severity models.

Mid-level roads			Huge-level roads		
Model	Variable	Impact on crash severity	Model	Variable	Impact on crash severity
BLM	Crash type (Head-on collision)	(+)	BLM	Victim vehicle (left turn)	(+)
	Special road type (tunnel)	(-)			
	Pede-veh seg (Curb only)	(+)			
	Traffic lights (Normal)	(-)			
	Traffic lights (Others)	(-)			
	Weather (Snowy)	(-)			
SVM	Victim (Turn Signal-Unclassified)	(+)	SVM	Victim (left turn)	(+)
	Offender (Turn Signal-Unclassified)	(-)		Week	(-)
	Traffic lights (Others)	(-)		Offender (Turn sig-No)	(-)
	Width of the sidewalk	(-)		Victim (Turn sig-No)	(-)
	Offender (Turn sig-No)	(-)		Pedestrian-vehicle segregation (Curb only)	(-)

In summary, the findings indicate that on mid-level roads, factors such as head-on collisions, pedestrian-vehicle segregation with curbs-only, and an inability to confirm the victim vehicle's turn signal are associated with a high crash severity. Therefore, enhancing safety measures, including more effective vehicle separation and improved pedestrian-vehicle segregation facilities, is crucial on mid-level roads. Additionally, there is a need to review signal configurations in locations with turning traffic. Furthermore, the association of wider sidewalks with a decreased severity suggests the importance of reviewing and potentially widening narrow sidewalks in certain road segments.

The findings for huge-level roads showed that a high crash severity is associated with the victim's vehicle making a left turn, as observed in both the BLM and SVM models. Consequently, it was concluded that safety management at intersections with left-turn traffic is imperative. Moreover, the result indicating a decreased severity when both the offender and the victim's vehicle do not use turn

signals suggests that driving straight is comparatively safer than engaging in turning traffic. Therefore, similar to the results obtained for the mid-level road model, caution is advised in the safety management of turning traffic. However, unlike mid-level road modeling results, the variable of pedestrian-vehicle segregation with curbs-only is associated with a decreased severity on huge-level roads. This result is interpreted to be due to the large number of lanes and relatively low number of pedestrians jaywalking on huge-level roads.

6. Conclusions

In this study, the data before, after, and at the time of the traffic crashes were collected from dashcam videos on urban huge-level and mid-level roads, and variables were devised to identify the factors that contributed to crash severity. In the analysis process, critical variables for crash severity classification were derived from 152 variables based on an MCA-based dimensional reduction and Boruta-SHAP. Through this process, variables that can reflect the characteristics of the data were derived using the MCA-based dimensional reduction method; simultaneously, variables with a significant influence on the model construction in the ensemble technique were considered through Boruta-SHAP. With this approach, key variables, that are both meaningful for modeling and explaining the data, were employed to develop a crash severity model. The model construction results showed that the SVM exhibited the highest performance in terms of accuracy and F1 score. The BLM results and SVM model development results were examined to comparatively analyze the results of constructing statistics-based and ML-based severity prediction models.

The primary findings of this paper are as follows:

1) On urban roads, the study demonstrates that the severity of crashes rose in the presence of left-turn traffic. Therefore, it is essential to reassess the design of the signal phase and the overall configuration of intersections. Furthermore, particular attention should be given to addressing this issue (left-turning traffic management) and improving road safety.

2) On mid-level roads with a lane width of 12 to 20 m, the study indicates that crash severity is higher in sections where only the curb separates pedestrians and vehicles compared to tunnel sections where pedestrians and vehicles are separated. To mitigate the severity of crashes, there is a need for effective pedestrian and vehicle separation facilities.

3) Finally, in contrast to mid-level roads, the separation of pedestrians and vehicles on huge-level roads was not identified as having a substantial impact on the crash severity. However, similar to mid-level roads, it is still crucial to formulate a safety management strategy specifically addressing turning traffic.

These findings provide valuable insights for road design, safety measures, and traffic management on different types of roads.

This study is significant in that it entailed the development of detailed crash severity models by collecting road conditions, traffic conditions, vehicle and personal information, and vehicle behavior at the time of the crash from dashcam video data. Notably, critical variables were selected using ML techniques, and models with a high accuracy and explanatory power were constructed. The findings of this study can be used to devise measures that reduce crash severity on urban huge-level and mid-level roads in the future. Furthermore, highly accurate crash severity models may potentially be constructed using the proposed methodologies. In addition, the findings of this study provide transportation planners and policymakers with insights to identify crucial factors that influence crash

severity on urban areas, especially on mid-level and large-level roads. This information enables efforts to either eliminate or mitigate factors that contribute to a high crash severity. For instance, one of the key results of the study indicates that crashes on mid-level roads that only using a curb to separate pedestrians and vehicles can result in high-severity outcomes. Transportation and road operators can utilize these findings to review their infrastructure, update relevant manuals or instructions, and take actions such as adding road markings with curbs to separate pedestrians and vehicles and enhance safety.

However, this study has some limitations. The analysis in this study was conducted using 381 crash data points, owing to the limitation in collecting dashcam video data; however, the dataset is too small to be used when constructing ML and statistical models. Therefore, in future research, it is necessary to create datasets suitable for ML model development through data augmentation or by collecting additional crash data. Furthermore, to mitigate human errors in the data collection process, it is necessary to explore the possibility of automating the process through the utilization of AI technology. Additionally, in this study, crash severity levels of 1 to 14 were classified into two categories, and a binary classification model for classifying accidents with high severity and low severity was constructed.

In the future, it is necessary to appropriately classify crash severity levels using methodologies, such as unsupervised learning-based clustering, and identify the factors that affect crash severity for each crash severity level by constructing detailed crash severity prediction models. Moreover, the capabilities of the proposed framework can be improved by incorporating various advanced optimization algorithms. Similar to existing studies utilizing advanced methodologies and integrated techniques [32–37], a novel and integrated methodology can be formulated to develop crash severity prediction models and derive key factors.

Furthermore, this study can be reframed as categorizing huge-level and mid-level roads rather than building a model that classifies high and low crash severity. Through this approach, the severity of crashes on each type of road, including huge-level and mid-level roads, can be determined, thereby enabling a distinction between the degree and the characteristics of each road by interpreting each variable.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

This research was supported by a grant (2021-MOIS38-001) of Proactive Technology Development on Safety Accident for Vulnerable Group and Facility funded by the Ministry of the Interior and Safety (MOIS, South Korea).

Conflict of interest

Juneyoung Park is a guest editor of the special issue for ERA and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

1. A. A. Jahangeer, S. S. Anjana, V. R. Das, A hierarchical modeling approach to predict pedestrian crash severity, in *Transportation Research: Proceedings of CTRG 2017*, **45** (2020), 355–366. https://doi.org/10.1007/978-981-32-9042-6
2. A. Sheykhfard, F. Haghighi, T. Nordfjærn, M. Soltaninejad, Structural equation modelling of potential risk factors for pedestrian accidents in rural and urban roads, *Int. J. Inj. Control Saf. Promot.*, **28** (2020), 46–57. https://doi.org/10.1080/17457300.2020.1835991
3. X. Yan, J. He, C. Zhang, Z. Liu, B. Qiao, H. Zhang, Single-vehicle crash severity outcome prediction and determinant extraction using tree-based and other non-parametric models, *Accid. Anal. Prev.*, **153** (2021), 106034. https://doi.org/10.1016/j.aap.2021.106034
4. I. Dash, M. Abkowitz, C. Philip, Factors impacting bike crash severity in urban areas, *J. Saf. Res.*, **83** (2022), 128–138. https://doi.org/10.1016/j.jsr.2022.08.010
5. Y. Yu, Z. Liu, A data-driven on-site injury severity assessment model for car-to-electric-bicycle collisions based on positional relationship and random forest, *Electron. Res. Arch.*, **31** (2023), 3417–3434. https://doi.org/10.3934/era.2023173
6. 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
7. T. J. Song, J. So, J. Lee, B. M. Williams, Exploring vehicle-pedestrian crash severity factors on the basis of in-car black box recording data, *Transp. Res. Rec.*, **2659** (2017), 148–154. https://doi.org/10.3141/2659-16
8. Y. Chung, Injury severity analysis in taxi-pedestrian crashes: An application of reconstructed crash data using a vehicle black box, *Accid. Anal. Prev.*, **111** (2018), 345–353. https://doi.org/10.1016/j.aap.2017.10.016
9. Y. Chung, An application of in-vehicle recording technologies to analyze injury severity in crashes between taxis and two-wheelers, *Accid. Anal. Prev.*, **166** (2022), 106541. https://doi.org/10.1016/j.aap.2021.106541
10. J. Cho, S. Lee, S. Park, J. Park, Classification and prediction of highway accident characteristics using vehicle black box data, *J. Korea Inst. Intell. Trans. Syst.*, **21** (2022), 132–145. https://doi.org/10.12815/kits.2022.21.6.132
11. B. P. Loo, Z. Fan, T. Lian, F. Zhang, Using computer vision and machine learning to identify bus safety risk factors, *Accid. Anal. Prev.*, **185** (2023), 107017. https://doi.org/10.1016/j.aap.2023.107017
12. E. Giovannini, A. Giorgetti, G. Pelletti, A. Giusti, M. Garagnani, J. P. Pascali, et al., Importance of dashboard camera (Dash Cam) analysis in fatal vehicle-pedestrian crash reconstruction, *Forensic Sci., Med. Pathol.*, **17** (2021), 379–387. https://doi.org/10.1007/s12024-021-00382-0
13. L. Taccari, F. Sambo, L. Bravi, S. Salti, L. Sarti, M. Simoncini, et al., Classification of crash and near-crash events from dashcam videos and telematics, in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, *IEEE*, (2018), 2460–2465. https://doi.org/10.1109/ITSC.2018.8569952
14. H. Pradana, An end-to-end online traffic-risk incident prediction in first-person dash camera videos, *Big Data Cognit. Comput.*, **7** (2023), 129. https://doi.org/10.3390/bdcc7030129

15. F. Hajri, H. Fradi, Vision transformers for road accident detection from dashboard cameras, in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE*, (2022), 1–8. https://doi.org/10.1109/AVSS56176.2022.9959545
16. L. Mussone, M. Bassani, P. Masci, Analysis of factors affecting the severity of crashes in urban road intersections, *Accid. Anal. Prev.*, **103** (2017), 112–122. https://doi.org/10.1016/j.aap.2017.04.007
17. A. Iranitalab, A. Khattak, Comparison of four statistical and machine learning methods for crash severity prediction, *Accid. Anal. Prev.*, **108** (2017), 27–36. https://doi.org/10.1016/j.aap.2017.08.008
18. S. Mafi, Y. AbdelRazig, R. Doczy, Machine learning methods to analyze injury severity of drivers from different age and gender groups, *Transp. Res. Rec.*, **2672** (2018), 171–183. https://doi.org/10.1177/0361198118794292
19. J. Liu, *Severity Analysis of Large Truck Crashes-Comparision Between the Regression Modeling Methods with Machine Learning Methods*, Ph.D thesis, Texas Southern University, 2021.
20. M. K. Islam, I. Reza, U. Gazder, R. Akter, M. Arifuzzaman, M. M. Rahman, Predicting road crash severity using classifier models and crash hotspots, *Appl. Sci.*, **12** (2022), 11354. https://doi.org/10.3390/app122211354
21. M. Yan, Y. Shen, Traffic accident severity prediction based on random forest, *Sustainability*, **14** (2022), 1729. https://doi.org/10.3390/su14031729
22. F. Afshar, S. Seyedabrishami, S. Moridpour, Application of Extremely Randomised Trees for exploring influential factors on variant crash severity data, *Sci. Rep.*, **12** (2022), 11476. https://doi.org/10.1038/s41598-022-15693-7
23. M. Alrumaidhi, M. M. Farag, H. A. Rakha, Comparative analysis of parametric and non-parametric data-driven models to predict road crash severity among elderly drivers using synthetic resampling techniques, *Sustainability*, **15** (2023), 9878. https://doi.org/10.3390/su15139878
24. V. Astarita, S. S. Haghshenas, G. Guido, A. Vitale, Developing new hybrid grey wolf optimization-based artificial neural network for predicting road crash severity, *Transp. Eng.*, **12** (2023), 100164. https://doi.org/10.1016/j.treng.2023.100164
25. S. Das, R. Avelar, K. Dixon, X. Sun, Investigation on the wrong way driving crash patterns using multiple correspondence analysis, *Accid. Anal. Prev.*, **111** (2018), 43–55. https://doi.org/10.1016/j.aap.2017.11.016
26. M. B. Kursa, W. R. Rudnicki, Feature selection with the Boruta package, *J. Stat. Software*, **36** (2010), 1–13. https://doi.org/10.18637/jss.v036.i11
27. M. B. Kursa, A. Jankowski, W. R. Rudnicki, Boruta—a system for feature selection, *Fundam. Inform.*, **101** (2010). https://doi.org/271-285.10.3233/FI-2010-288
28. B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, (1992), 144–152. https://doi.org/10.1145/130385.130401
29. T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, et al., Xgboost: extreme gradient boosting, *R Package Version 0.4-2*, **1** (2015), 1–4.
30. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785

31. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.*, **30** (2017).
32. M. Chen, Y. Tan, SF-FWA: A Self-Adaptive Fast Fireworks Algorithm for effective large-scale optimization, *Swarm Evol. Comput.*, **80** (2023), 101314. https://doi.org/10.1016/j.swevo.2023.101314
33. M. A. Dulebenets, An adaptive polyploid memetic algorithm for scheduling trucks at a cross-docking terminal, *Inf. Sci.*, **565** (2021), 390–421. https://doi.org/10.1016/j.ins.2021.02.039
34. J. Pasha, A. L. Nwodu, A. M. Fathollahi-Fard, G. Tian, Z. Li, H. Wang, et al., Exact and metaheuristic algorithms for the vehicle routing problem with a factory-in-a-box in multi-objective settings, *Adv. Eng. Inf.*, **52** (2022), 101623. https://doi.org/10.1016/j.aei.2022.101623
35. P. Singh, J. Pasha, R. Moses, J. Sobanjo, E. E. Ozguven, M. A. Dulebenets, Development of exact and heuristic optimization methods for safety improvement projects at level crossings under conflicting objectives, *Reliab. Eng. Syst. Saf.*, **220** (2022), 108296. https://doi.org/10.1016/j.ress.2021.108296
36. M. A. Dulebenets, A Diffused Memetic Optimizer for reactive berth allocation and scheduling at marine container terminals in response to disruptions, *Swarm Evol. Comput.*, **80** (2023), 101334. https://doi.org/10.1016/j.swevo.2023.101334
37. E. Singh, N. Pillay, A study of ant-based pheromone spaces for generation constructive hyper-heuristics, *Swarm Evol. Comput.*, **72** (2022), 101095. https://doi.org/10.1016/j.swevo.2022.101095

AIMS Press