TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–18

DOI: 10.1177/03611981221134629
journals.sagepub.com/home/trr
SAGE

Applying Explainable Machine Learning Techniques in Daily Crash Occurrence and Severity Modeling for Rural Interstates

Zihang Wei[1], Yunlong Zhang[1], and Subasish Das[2]

Abstract

Conventional traffic crash analysis methods often use highly aggregated data, making it difficult to understand the effects of time-varying factors on crash occurrence. Although studies have used data with small aggregation intervals, they typically analyze the effect of a single factor on crash occurrence. In this study, we investigate the collaborative effect of roadway geometry, speed distribution, and weather conditions on crash occurrence and severity using explainable machine learning methods on daily level crash data. The data were collected on rural Interstate highways in Texas. Four machine learning methods: random forest, AdaBoost, XGBoost, and deep neural network, were tested on the dataset. The results showed that XGBoost performs the best on the imbalanced dataset. The study used the synthetic minority oversampling technique (SMOTE) method to mitigate the data imbalance issue. The XGBoost model was trained separately on all crash occurrences and severe crash occurrences. Finally, the SHAP (SHapley Additive exPlanation) method was applied to investigate the contribution of all variables to the model's output. The results showed that weather condition factors have a significant contribution to all crash occurrences. Speed distribution factors have a stronger impact on severe crash occurrences.

Keywords

artificial intelligence and advanced computing applications, crash analysis, crash data, data and data science, safety performance and analysis

Conventional traffic crash analysis methods typically use highly aggregated data. Crash-related variables are often aggregated over some time period. Thus, some crash-related explanatory variables that may significantly change during this time period are typically not considered because detailed data are usually not available (*1*). However, for many time-varying explanatory variables such as speed and weather data, the variations within these intervals are also very important for crash analysis. Many studies have found that weather conditions, especially precipitation and visibility (*2–4*), and speed distribution (*5–7*) are closely related to crash occurrence. It is unreasonable to aggregate these variables over a long time period because the variations within this time period cannot be fully explored. Thus, ignoring the possible variation of these explanatory variables within the time period may result in the loss of valuable information. The best way to avoid this problem is to aggregate crash data into smaller time intervals. In particular, data can be aggregated by day, hour, or even minute.

On the other hand, other explanatory variables such as road geometry data (i.e., curve, lane width, shoulder width, etc.) are relatively static. For the same roadway segment, road geometry data rarely change over any time period. By aggregating roadway geometry data into smaller intervals, more identical observations will be generated. Moreover, roadway segments that are close to

[1]Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX
[2]Ingram School of Engineering, Texas State University, San Marcos, TX

Corresponding Author:
Zihang Wei, wzh96@tamu.edu

each other always share similar geometric features, which results in correlation over space. The temporal and spatial correlation will negatively affect the analysis of these relatively static explanatory variables (*8*, *9*).

Many studies have investigated the effect of roadway geometrics on crash occurrence, and some have studied the effect of speed distribution and weather conditions. However, few of them are conclusive enough, and even fewer researchers have investigated the collaborative effect of these three factors on crash occurrence. The main challenge in analyzing these three factors together is to choose an ideal aggregation time interval.

There are three main aggregation intervals that have previously been applied by researchers. The first one is the yearly aggregation interval, under which the effect of roadway geometry can be analyzed. However, the effects of speed distribution and weather conditions cannot be analyzed under this. The second one is the daily aggregation interval, under which the effects of roadway geometrics and weather conditions can be analyzed, but it is not ideal for studying speed distribution. The third one is hourly or minute by minute, under which the effect of speed distribution can be analyzed, but it is not ideal for studying the effects of roadway geometry or weather conditions. In this study, we chose to use daily aggregation intervals. Because speed distribution tends to vary differently throughout the day, this study introduces speed measurement variables for different time periods during the day, such as average daytime speed and nighttime speed standard deviation.

Moreover, to address the problem of the temporal and spatial correlation in crash frequency analysis, we need to aggregate the data into daily intervals but should also include several roadway segments with various geometry features. Thus, this study includes a wide range of rural Interstate highway segments to address this issue.

The size of the dataset included in this study is huge. Machine learning techniques can be applied to analyze this big data problem efficiently. In this study, four machine learning algorithms were trained on the dataset: random forest, AdaBoost, XGBoost, and deep neural network (DNN). The performances of these four models were compared with each other, and one optimal model was selected. To investigate the detailed relationships between the explanatory variables and crash occurrence and severity, training the machine learning model alone is not sufficient because many machine learning methods are the black-box type. Therefore, in this study, apart from applying machine learning algorithms to analyze the dataset, the SHAP (SHapley Additive exPlanation) method was applied to unearth the relationship between explanatory variables and the model's outcomes.

In crash occurrences with different severity levels, the contributing factors of severe crash occurrences tend to be different from those of all crash occurrences. As a result, this study trained another machine learning model on severe crash occurrences to compare the different factors behind all crash occurrences and severe crash occurrences. The results of this study can reveal the collaborative effects of roadway geometry, weather conditions, and speed distribution on daily crash occurrences on Texas rural Interstates and their effects on crash severity patterns.

Literature Review

Many previous researchers have studied the effects of roadway geometry, weather conditions, and speed distribution factors on crash occurrence. Some researchers have also studied the collaborative effects of two of the three factors. For example, Shankar et al. (*2*) studied highway crash frequency by analyzing the effect of geometric elements and weather conditions. The study was conducted by applying a negative binomial model, and the data were aggregated into monthly intervals. Dutta and Fontaine (*10*) introduced crash prediction modeling on a freeway segment using disaggregated speed data and roadway geometric data. The results indicated that by including hourly averaged speed data and selected roadway geometric data, the crash prediction performance improved compared with that using annual data without speed information.

Many studies have analyzed the relationship between roadway geometric features and crash occurrence or severity. Miaou and Lum (*11*) utilized two linear regression models and two Poisson regression models to study the relationship between roadway geometric factors and crash frequency. Anderson et al. (*12*) applied Poisson, negative binomial, and log normal regression analysis to study the relationship between rural two-lane highway crash frequency and roadway geometric design consistency. Haghighi et al. (*13*) investigated the effect of roadway geometric factors on crash severity with data collected from rural two-lane highways. They developed a multilevel ordered logit model to deal with the hierarchical structure of the crash data. They found that the introduction of crash type as a variable can better explain variation in crash severity level.

Speed distribution is another important contributor to crash occurrence. Garber and Gadiraju (*14*) investigated the impact of mean speed and speed variation on crash rate. They concluded that a higher mean speed does not necessarily increase the crash rate, but a higher speed variance can lead to a higher crash rate. Lee et al. (*15*) used real-time traffic flow data from loop detectors to predict crash occurrence. They applied an aggregated log-linear model to model crash occurrence and found that speed variation and traffic density are strong

indicators of crash frequency. Pei et al. (*16*) analyzed the effect of mean speed on crash occurrence using disaggregated speed and crash data with a 4 h interval from different time periods of a day. They found that the mean speed and crash occurrence are positively related when distance exposure is considered, but are negatively related when time exposure is considered. Wang et al. (*7*) studied the relationship between mean speed, speed variation, and crash frequency on arterials in urban areas. A hierarchical Poisson log normal model was applied to model the crash frequency. Since speed distribution tends to vary significantly during different time periods, this study aggregated crash data into three study periods (morning, midday, and evening), where each time period was 3 h long. The results revealed that a higher average speed and higher speed variation will lead to higher crash frequencies on urban arterial roads.

Weather condition factors can significantly affect crash occurrence as well. Scott (*17*) included temperature and rainfall as explanatory variables to model the time-series crash data. A regression model was applied to model single-vehicle crashes, and a Box-Jenkins model was applied to model two-vehicle crashes. Eisenberg (*18*) analyzed the effects of precipitation on traffic crashes by applying the negative binomial regression method. Two data aggregation intervals (monthly and daily) were studied. The results revealed a significant negative relationship between monthly fatal crash frequency and precipitation. However, the results indicated a significant positive relationship between daily fatal crashes and precipitation. Brijs et al. (*19*) applied an integer autoregressive model on daily crash data to model the time-dependent nature of the crash occurrence. Their results showed that the intensity of the rainfall is significantly related to the daily crash count. Jaroszweski and McNamara (*20*) analyzed the influence of precipitation on crashes by utilizing weather radar images. This novel approach offered improvements to the analysis of weather-related accidents by giving a more representative rainfall measure in urban areas. Yu and Abdel-Aty (*21*) analyzed the relationship between weather conditions and crash severity on mountainous freeways. The results indicated that (i) snowy weather is less likely to cause severe crashes and (ii) lower temperature increases the likelihood of severe crashes.

Although previous studies have investigated these three factors separately, fewer studies have considered these three factors together and analyzed their collaborative effects on crash occurrence. One major problem in studying the collaborative effects of these three factors is which data aggregation level should be used. Previous studies tended to analyze the relationship between crash and roadway geometric data based on yearly aggregation intervals (*11*, *12*), the relationship between crash and weather condition variables based on daily aggregation intervals or monthly aggregation intervals (*17–19*, *22*), and the relationship between crashes and speed distribution based on real-time data with hourly intervals or minute-by-minute aggregation (*7*, *15*, *16*).

By summarizing previous research, it is found that the daily aggregation interval seems ideal for collectively analyzing the effects of roadway geometry and weather conditions on crash occurrence. Although it is not ideal for analyzing the effect of speed distribution on crash occurrence as speed distribution tends to differ significantly throughout a day, variables such as average daytime speed and average nighttime speed can be included in the daily model to address this problem.

Data Preparation

Data Acquisition and Processing

Data were collected from roadways in the State of Texas. First, a comprehensive dataset was developed by using the data conflation method. The dataset analyzed in this study contains four parts: (i) roadway geometry features and traffic information, (ii) weather condition data, (iii) speed measurement data, and (iv) crash data. These data were collected from four different sources respectively: (i) Texas Department of Transportation Road-Highway Inventory Network Offload 2018, (ii) Automated Surface Observing System, (iii) National Performance Management Research Dataset, and (iv) Crash Record Information System.

Base Roadway Network and Geometric Data. The base roadway network was collected from the Texas Department of Transportation (TxDOT) Road-Highway Inventory Network Offload (RHiNO) 2018, which contains roadway Geographic Information System (GIS) linework and roadway inventory attributes, including geometric features and traffic information. TxDOT submitted this dataset to the Federal Highway Administration (FHWA) as part of the Highway Performance Monitoring System program (*23*). This study selected rural Interstates from the base network.

Speed Distribution Data. Speed data were collected from FHWA's National Performance Management Research Dataset (NPMRDS). The NPMRDS contains travel time and speed data collected from a fleet of probe vehicles (cars and trucks). The NPMRDS can generate speed and travel time data by using probe vehicle location information. The data are aggregated in three time intervals: 5 min, 15 min, and 1 h. This study used the data with 5 min intervals, as more detailed information can be kept when calculating speed variation. Moreover, in one

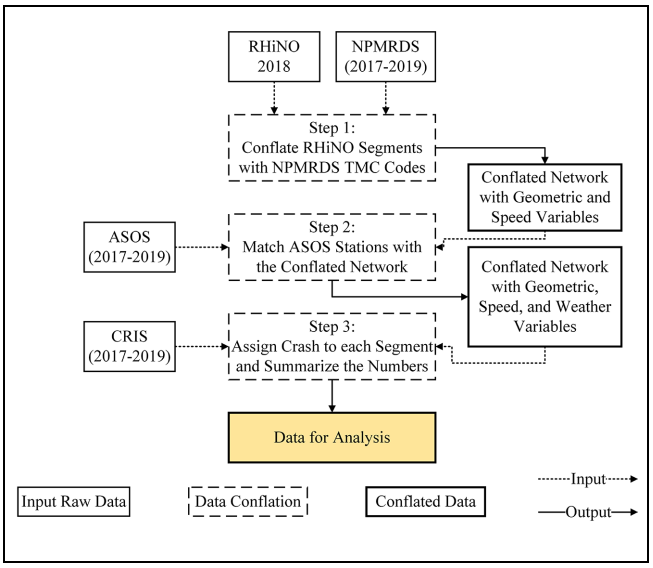

Figure 1. Flowchart of the data preparation process.

study, the results showed that the impact of different short-duration time intervals does not affect the modeling results (*24*). Speed data are available across the National Highway System, and the spatial resolution is set by different Traffic Message Channel (TMC) location codes (*25*). Daily speed distribution variables (i.e., average speed, speed standard deviation, 85th percentile speed, etc.) were calculated based on 5 min speed data at each TMC. Each TMC is conflated with its corresponding roadway segment by using GIS software.

Weather Condition Data. Weather condition data were collected from the Automated Surface Observing System (ASOS) of the National Centers for Environmental Information. Each roadway segment is matched with an ASOS station closest to it. Note that the Road Weather Information System (RWIS) or land-based weather station provides more accurate and granular information on weather data. The current study has not used RWIS data as we found that these data are sporadic and will not cover all networks used in this study. Using RWIS will also need extrapolation. Thus, we used ASOS as the data is extensive and covers almost all networks.

Crash Data. Crash data were collected within the State of Texas from 2017 to 2019 through the Crash Record Information System (CRIS). Each crash record includes location and date information. Through GIS software, all crashes were assigned to the roadway segments on which they occurred. The daily crash count of each roadway segment was then summarized. The dataset used in this study contains 2,601,106 non-crash observations and 26,210 crash observations, including 1,428 severe crash observations. The definition of a severe crash in this study is a crash that resulted in severe injury or fatality. In this process, crash severity is classified into five different levels according to the Highway Safety Manual (*26*): (i) K: Killed; (ii) A: Incapacitating Injury; (iii) B: Non-Incapacitating Injury; (iv) C: Possible Injury; (v) O: Not Injured or Unknown.

Table 1. Summary of Segments and Number of Crashes at Different Severity Levels

	Number of crashes (2017–2019)				
Number of segments	K	A	B	C	O
2,543	398	1,043	2,863	3,212	20,616

Note: K = Killed; A = Incapacitating Injury; B = Non-Incapacitating Injury; C = Possible Injury; O = Not Injured or Unknown.

The data from the four parts above were conflated by using ArcGIS software. All data were aggregated into a daily interval. The final dataset is made up of 26 variables calculated from the abovementioned four parts. The total number of segments, total segment length, and the number of crashes at all different severity levels (KABCO) are summarized in Table 1. In this study, "all crash" included the severity levels KABCO and "severe crash" only included the severity levels KA. The detailed definitions of all variables and their descriptive statistics are listed in Table 2. The data preparation process is shown in Figure 1.

In this study, all crash observations were defined as follows: at a particular roadway segment on a particular day, a crash of any type (KABCO) occurred. The opposite of an all crash observation is a non-crash observation, which means that no crashes occurred at a particular roadway segment on a particular day. Severe crash observations were defined as follows: crashes that resulted in incapacitating injuries and fatalities (K and A) happened at a particular segment on a particular day. The opposite of severe crash observation is a non-severe-crash observation, which means that there were no crashes that led to incapacitating injuries and fatalities at a particular segment on a particular day.

Feature Selection

In the prepared dataset, some explanatory variables may be highly correlated with others. When machine learning models are trained on the dataset, these variables do not have extra benefits in distinguishing the target variables. Thus, to improve modeling efficiency and accuracy, some highly correlated explanatory variables need to be removed.

Table 2. Variable Names, Definitions, and Descriptive Statistics

Variable names	Definition	Descriptive statistic			
		Mean	SD	Min.	Max.
Weather condition					
DailyPrecip	Daily precipitation	0.07	0.46	0.00	49.20
VsbyAve	Average visibility	9.37	1.32	0.28	177.22
VsbyStd	Visibility standard deviation	0.91	2.80	0.00	709.46
Speed distribution					
SpdAve	Average of daily speed	67.37	2.33	25.96	75.45
SpdStd	Standard deviation of daily speed	2.82	1.29	0.73	24.80
SpdCV	Coefficient of variation of daily speed	0.06	0.03	0.02	0.70
Spd85	85th percentile of daily speed	70.98	2.51	37.90	85.70
RefSpd	Reference speed	73.01	2.46	61.02	79.55
SpdAveDay	Average of daily speed using data during daytime (6:00 a.m.–6:00 p.m.)	67.67	2.64	25.33	77.34
SpdStdDay	Standard deviation of daily speed using data during daytime (6:00 a.m.–6:00 p.m.)	2.62	1.43	0.64	29.54
SpdCVDay	Coefficient of variation of daily speed using data during daytime (6:00 a.m.–6:00 p.m.)	0.06	0.04	0.01	1.08
SpdAveNight	Average of daily speed using data during nighttime (6:00 p.m.–6:00 a.m.)	67.02	2.15	18.98	77.22
SpdCVNight	Coefficient of variation (CV) of daily speed using data during nighttime (6:00 p.m.–6:00 a.m.)	0.06	0.03	0.02	0.92
SpdStdNight	Standard deviation of daily speed using data during nighttime (6:00 p.m.–6:00 a.m.)	2.86	1.19	0.74	24.86
SpdFFAve	Average of daily speed larger than reference speed	75.83	2.96	52.00	99.00
SpdFF85	85th percentile of daily speed larger than reference speed	77.41	3.49	52.00	99.00
Roadway geometry and traffic					
SpdMax	Maximum speed limit	74.47	3.77	45.00	85.00
MedWid	Median width	50.65	28.54	1.00	267.00
NumLanes	Number of through lanes	4.12	0.47	4.00	6.00
LaneWidth	Lane width	11.92	0.51	8.00	20.00
SWid_I	Inside shoulder width	10.73	3.81	0.00	32.00
SWid_O	Outside shoulder width	19.52	2.82	6.00	48.00
SrfType	Surface type (categorical)				
AADT	Average annual daily traffic	31,627.76	17,329.54	5,008.00	102,074.00
TrkAADTP	Truck AADT percentage	32.92	11.46	5.90	64.90
Length	Roadway segment length	0.66	0.61	0.00	2.00

Note: SD = standard deviation; Min. = minimum; Max. = maximum; AADT = average annual daily traffic.

First, the Pearson correlation coefficient (PCC) is applied to evaluate feature correlation. The PCC of a variable pair is calculated as in Equation 1. Figure 2 shows the PCC heatmap of the Rural Interstates dataset. Brighter cells represent a higher correlation between two explanatory variables. Two explanatory variables are considered highly correlated if their PCC is higher than 0.7. In general, two variables are considered correlated if their PCC is higher than 0.5 (*27*). However, since the decision tree algorithms (including the XGBoost and AdaBoost algorithms) are very robust to correlated variables, the threshold is increased to 0.7 in this study (*28*).

For random forest and DNN models, this feature selection process helped to remove highly correlated variables and boost the model's performance, whereas for XGBoost and AdaBoost models this process only means removing the redundant variables in the dataset to facilitate the model training speed. Subsequently, it is not a necessary step for training an XGBoost or an AdaBoost model. There are 23 pairs of highly correlated explanatory variables identified from the dataset (see Table 3).

$$c_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

After the PCCs were calculated for all explanatory variable pairs, a random forest model was trained using

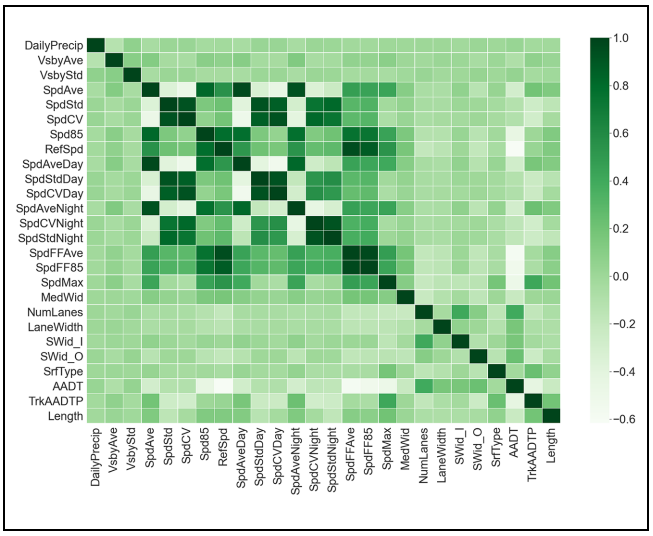

Figure 2. Pearson correlation coefficient heatmap.

Table 3. Correlated Variable Pairs

Identified correlated variable pairs		Absolute Pearson correlation coefficient
SpdFF85	Spd85	0.743
SpdCV	SpdStdNight	0.744
SpdCVNight	SpdStd	0.750
SpdFFAve	Spd85	0.760
Spd85	SpdAveNight	0.778
SpdAveDay	Spd85	0.784
Spd85	RefSpd	0.788
SpdStd	SpdStdNight	0.813
SpdCV	SpdCVNight	0.813
Spd85	SpdAve	0.822
SpdAveDay	SpdAveNight	0.822
SpdStdDay	SpdCV	0.862
SpdCVDay	SpdStd	0.864
RefSpd	SpdFF85	0.880
SpdStdDay	SpdStd	0.924
SpdCVNight	SpdStdNight	0.924
SpdCVDay	SpdCV	0.929
SpdCVDay	SpdStdDay	0.930
SpdCV	SpdStd	0.930
SpdAveNight	SpdAve	0.934
SpdFFAve	RefSpd	0.950
SpdAveDay	SpdAve	0.969
SpdFFAve	SpdFF85	0.973

Note: Variables defined in Table 2.

all the explanatory variables on the dataset and the feature importance values of all available explanatory variables. All variables were ranked based on their feature importance values. The feature selection criterion is that for each highly correlated explanatory variable pair, the one with a lower feature importance value is removed (see Table 2). Finally, nine explanatory variables were removed from the dataset: SpdCV; SpdStdDay; SpdStd; SpdCVNight; SpdAveNight; SpdAve; SpdFF85; Spd85; RefSpd. See Table 2 for definitions of the variables.

Resampling Imbalanced Dataset

Because crash occurrence is aggregated into daily intervals, a significant number of observations will not have crash occurrences. This is because of the rare nature of crash events. Thus, in the prepared dataset, the number of non-crash observations is significantly larger than that of the crash observations. This results in a highly imbalanced dataset. The machine learning model cannot be properly directly trained by using the imbalanced dataset.

This study applied the synthetic minority oversampling technique (SMOTE) to rebalance the original dataset. Many previous studies have applied resampling methods. Abdel-Aty (*29*) applied a matched case-control method that manually matched crash samples with non-crash samples. Chawla et al. (*30*) first proposed SMOTE to address the problem of imbalanced datasets. SMOTE is an oversampling method that only oversamples the minority class and is only applied to the training dataset. Data points from the minority group are oversampled by creating "synthetic" samples along the line segments joining the k minority class nearest neighbors. The number of neighbors is randomly chosen based on the required amount of oversampling numbers. Since SMOTE is not applied to the testing dataset, the testing result can still be considered to reflect reality. Many previous studies have applied SMOTE to address imbalanced datasets (*27*, *31*, *32*). Before resampling, there was a huge difference between the number of crash observations and the number of non-crash observations (see Figure 3*a*). After resampling, the number of crash observations in the training set is now equal to the number of non-crash observations (see Figure 3*b*). Note that in Figure 3, only two dimensions ("DailyPrecip" and "SpdCVDay") are selected to visualize the SMOTE process. However, the true dataset is multi-dimensional, which is impossible to plot.

Methodology

XGBoost

XGBoost (eXtreme gradient boosting) is a scalable end-to-end tree boosting system. It implements machine learning algorithms under the gradient boosting framework (*33*). XGBoost is an additive boosting tree package that is built by k essential tree functions implemented with regularization, missing value imputation, shrinkage and column subsampling, sparsity-aware split finding, and column block for parallel learning. Compared with

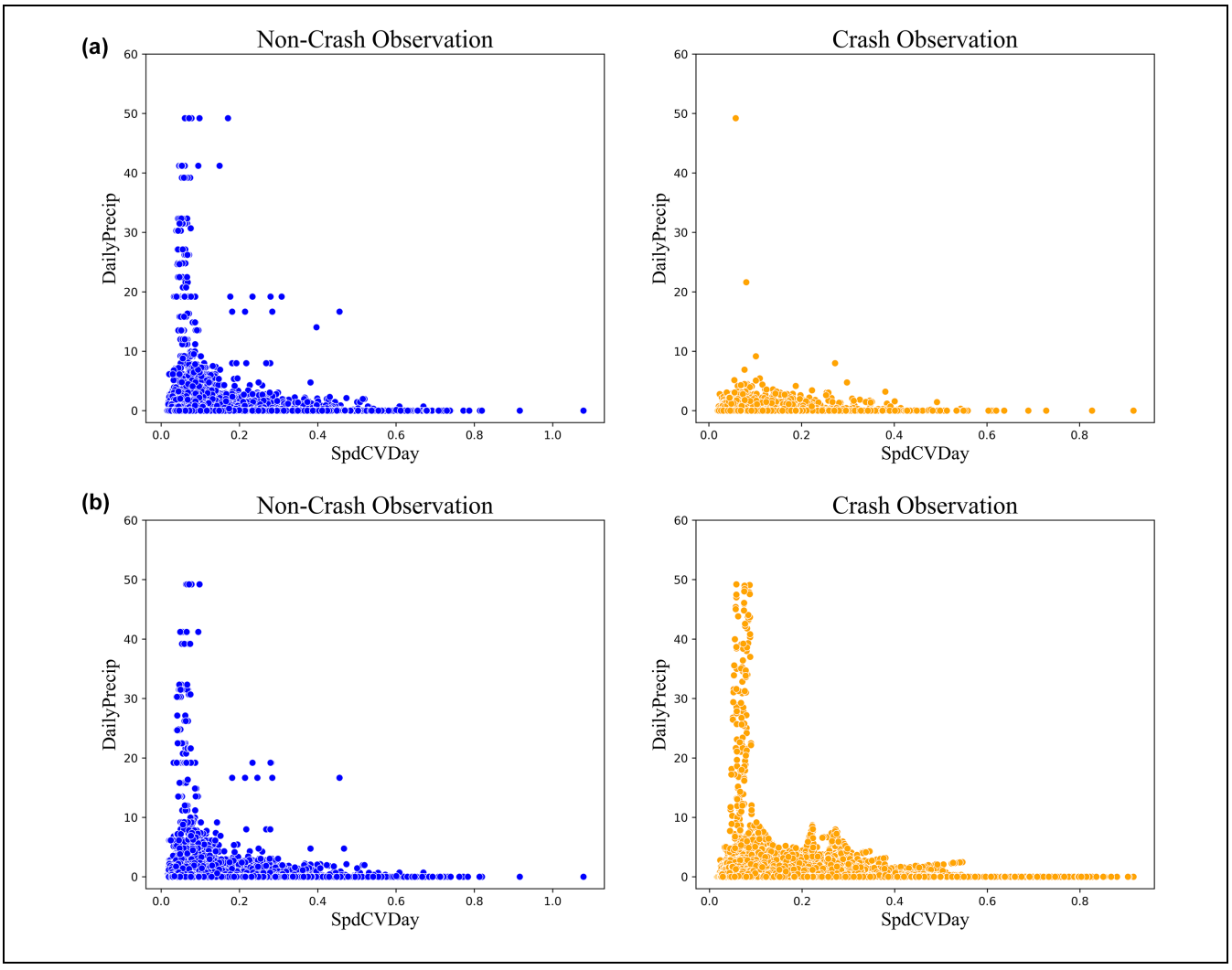

Figure 3. Synthetic minority oversampling technique (SMOTE) method: (*a*) before SMOTE oversampling (1,950,829 non-crash observations and 19,658 crash observations) and (*b*) after SMOTE oversampling (1,950,829 non-crash observations and 1,950,829 crash observations).

gradient boosting, XGBoost can deliver more accurate approximations by using the strengths of the second-order derivative of the loss function, L1 and L2 regularization, and parallel computing. It can run more than 10 times faster than existing popular machine learning solutions, making it suitable for big data problems. XGBoost can solve real-world-scale problems by using relatively few resources. It is currently one of the fastest and best open-source boosting tree tools for modeling and prediction analyses. The detailed mathematical process of XGBoost is as follows.

Given a dataset with n observations, each observation has multiple features, x_i, and a corresponding response variable, y_i. $\hat{y}_i^{(t)}$ is the predicted response value after tth iterations by adding one tree function $f(x_i)$ to the predicted value of the (t–1)th iteration corresponding to the ith observation. The boosting process is shown in Equation 2.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

The objective of this process is to minimize Equation 3. $l(y_i, \hat{y}_i)$ is a loss function and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ represents the penalty for the complexity of the model where T is the number of leaves and w_j^2 is the L2 norm of jth leaf scores. This term is used to avoid overfitting.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) \tag{3}$$

By solving Equations 2 to 3, the optimal value of w_j is:

$$w_j^* = -\frac{\sum_i \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\sum_i \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) + \lambda} \quad (4)$$

And the corresponding minimum object value is:

$$Obj^{min} = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_i \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)\right)^2}{\sum_i \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) + \lambda} + \gamma T \quad (5)$$

Random Forest

The random forest model is an ensemble method that can be used for both classification and regression (*34*). The random forests are a combination of individual tree predictors. The final decision made by a random forest model is based on the decisions made by all individual decision trees in the model. The random forest model has been applied to many crash risk assessment problems in previous studies (*35–38*).

AdaBoost

The AdaBoost model was developed by Freund and Schapire (*39*). It is a machine learning algorithm for classification and regression problems. The core idea of the AdaBoost model is to train several weak learners on the training set, and then the weaker learners are grouped together to build a stronger learner. There are several previous studies that have applied AdaBoost to address the crash risk assessment problem (*40*).

Deep Neural Networks (DNNs)

The structure of artificial neural networks resembles the structure of biological neurons in the human brain. It has been widely applied to speech recognition, image classification, and data classification. A DNN is an artificial neural networks that contains more than one hidden layer. In this study, a DNN model with three hidden layers was built to classify crash observations and non-crash observations. The structure of the DNN model was designed as shown in Figure 4. Note that the number of neurons and input nodes displayed in this figure do not match the actual neutron numbers since the numbers are too large to display. The basic information of the DNN model structure is listed as follows:

- Input Layer: 17 variables
- Hidden Layer 1: 60 neutrons
- Hidden Layer 2: 600 neutrons
- Hidden Layer 3: 60 neutrons

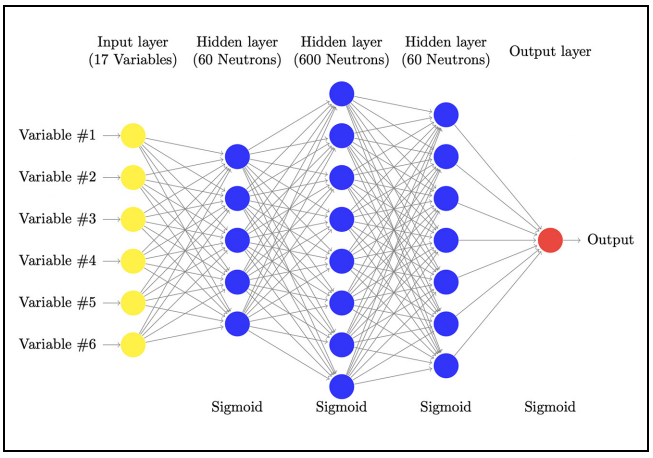

Figure 4. Deep neural network structure.

Shapley Additive exPlanations (SHAP)

The results from the XGBoost model are explained by the SHAP method. Machine learning methods used to be criticized as black boxes since it is hard to interpret the contribution of each individual variable to the model's output. Lundberg and Lee (*41*) proposed the SHAP method, which can explain tree-based machine learning models by estimating the individual contribution of each feature on the model prediction based on a game theory approach. For any particular prediction, the Shapley value of a feature i can be calculated as Equation 6:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}}^{T} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] \quad (6)$$

where

ϕ_i = Shapley value of feature i
S = a possible feature subset
$|S|$ = number of features in subset S
F = the set of all features
$|F|$ = the number of features in set F
$f_{S\cup\{i\}}(x_{S\cup\{i\}})$ = model prediction based on the features in subset S and feature i
$f_S(x_S)$ = model prediction based on the features in subset S.

Here, $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ is the difference between the model prediction made with features in subset S and feature i and the model prediction made only with features in subset S. This difference can reflect how the presence of feature i can change the model's prediction output. For any given prediction, the Shapley value of a particular feature is calculated as the weighted average of all differences across all possible subsets S.

Results and Discussion

Model Tuning

For the tuning process, the grid search method is used to find the set of hyperparameters that perform the best. The performance measure for the XGBoost model is the Receiver Operating Characteristic-Area under the ROC curve (ROC-AUC) score from the fine-tuning process. After fine-tuning, the hyperparameters for the XGBoost model are set as follows:

- Learning rate: 0.2
- Maximum depth: 6
- Number of estimators: 1,000

The performance measure used for fine-tuning the AdaBoost model is balanced accuracy. After fine-tuning, the hyperparameters for the AdaBoost model are set as follows:

- Learning rate: 1
- Number of estimators: 300

The performance measure used for fine-tuning the random forest model is balanced accuracy. After fine-tuning, the hyperparameters for the AdaBoost model are set as follows:

- Maximum depth: 50
- Number of estimators: 250

Model Comparison

To select the best model for classifying the highly imbalanced crash dataset in this study, the performance measures of four models (XGBoost, random forest, AdaBoost, and DNN) are compared in this section. The following performance measures are selected for model comparison:

1. $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
2. $Sensitivity = \frac{TP}{TP + FN}$
3. $Specificity = \frac{TN}{TN + FP}$
4. $Weighted\ Accuracy = \frac{Sensitivity + Specificity}{2}$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Accuracy is the number of correctly classified observations ($TP + TN$) over the number of all observations ($TP + TN + FP + FN$) in the testing dataset. This score ideally reflects the model performance when the data from two classes are nearly balanced. Sensitivity is the number of correctly classified positive observations (TP) over the number of all positive observations ($TP + FN$) in the testing dataset. The sensitivity score reflects a model's ability to effectively identify positive observations. The sensitivity score is very important for crash datasets because positive observations are in the minority class. Specificity is the number of correctly classified negative observations (TP) over the number of all negative observations ($TN + FP$) in the testing dataset. Specificity scores show a model's capability of correctly identifying negative observations. The weighted accuracy is the average of the sensitivity score and the specificity score. For an unbalanced dataset, the accuracy score cannot reflect the true performance of a model if it tends to classify most of the observations as the majority class. In these cases, the weighted accuracy score can be used to show the true performance of a model since it has to perform well on both the majority and minority to achieve a high score.

Moreover, this study also introduces Copen's kappa statistics to measure the agreement between the true values of the testing set and the predicted values by the models. The Copen's kappa statistics can better reflect the models' performance on imbalanced data. The Copen's kappa statistic can be calculated as follows:

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{7}$$

$$p_e = p_{e1,target} * p_{e1,pred} + p_{e2,target} * p_{e2,pred} \tag{8}$$

where

K = Copen's kappa statistic

p_0 = overall accuracy

p_e = a measure of the agreement between the model predictions and the actual class values

$p_{e1,target}$ = the actual proportion of the first class

$p_{e1,pred}$ = the predicted proportion of the first class

$p_{e2,target}$ = the actual proportion of the second class

$p_{e2,pred}$ = the predicted proportion of the second class

For binary classification, the Copen's kappa can be calculated as:

$$K = \frac{2 * (TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)} \tag{9}$$

The four models were trained on the dataset with all crash (KABCO) occurrences as the target variable. Table 4 summarizes the performance measures of these four models.

The random forest model has the highest accuracy score and specificity score. However, its sensitivity scores and weighted accuracy scores are low. As the SMOTE oversampling method is only applied to the training dataset, the testing dataset is still highly imbalanced. The random forest model tends to classify most of the observations in the testing set as the majority class (non-

Table 4. Performance Measures Comparison Across Four Models

Performance measures	XGBoost	AdaBoost	Random forest	Deep neural network
Accuracy	78.7%	80.4%	97.4%	76.5%
Sensitivity	64.2%	58.6%	14.2%	64.2%
Specificity	78.8%	80.6%	98.2%	76.6%
Weighted accuracy	71.5%	69.6%	56.2%	70.4%
Copen's kappa	0.038	0.085	0.038	0.033

Table 5. Confusion Matrices (XGBoost, Random Forest, AdaBoost, DNN)

	XGBoost		Random forest		AdaBoost		Deep neural network	
Predicted	0	1	0	1	0	1	0	1
True								
0	512,731	137,546	638,668	11,609	524,090	126,140	498,027	152,250
1	2,343	4,209	5,622	930	2,735	3,865	2,344	4,208

Table 6. Confusion Matrix of the XGBoost Models

	All crash model		Severe crash model	
Predicted labels	Non-crash	Crash	Non-crash	Crash
True labels				
Non-crash	512,731	137,546	552,184	104,288
Crash	2,343	4,209	114	243

crash), which gives the model higher accuracy and specificity values (see Table 5). The sensitivity score of the random forest model is low, at 14.2%, which means it cannot effectively identify crash observations. The AdaBoost model has the highest specificity score (80.6%) after the random forest model. However, its sensitivity score is 58.6%, which is lower than the XGBoost and DNN model scores. The sensitivity score is more important for imbalanced crash datasets as it reflects a model's ability to identify crash observations; therefore, the performance of the AdaBoost model is considered to be slightly inferior to the XGBoost model and the DNN model. The XGBoost model and the DNN model both have the same sensitivity score (64.2%). For the specificity score, the XGBoost model performs better than the DNN model. Based on the comparisons of these four models, it is concluded that the XGBoost model has the best performance for the daily level imbalanced crash dataset of this study. In the following sections, the results of the XGBoost model will be presented and analyzed further. Table 5 summarizes the confusion matrices of these four models. Finally, with regard to the Copen's kappa, since the training dataset is still highly imbalanced, the Copen's kappa statistics of all four models are not ideal and are not a good indicator of model performance in this case.

All Severity Level Model

This model considers crash occurrences with all severity levels (KABCO). A value of zero indicates all crash observations and one indicates non-crash observations. The prepared dataset is split into a training set (70% of all observations) and a testing set (30% of all observations). The SMOTE oversampling method is applied to the training dataset to balance the minority group and the majority group. The training dataset contains 1,950,829 non-crash observations and 19,658 crash observations. After the oversampling process, the total numbers of both non-crash and crash observations are 1,950,829. The model evaluation is made with the testing dataset. Table 6 is the confusion matrix of the XGBoost model. The model performance is evaluated by four measurements. Figure 5*a* is the ROC plot of the all crash occurrence model.

SHAP is applied to interpret the feature importance in the XGBoost model. Figure 6 is the SHAP summary plot

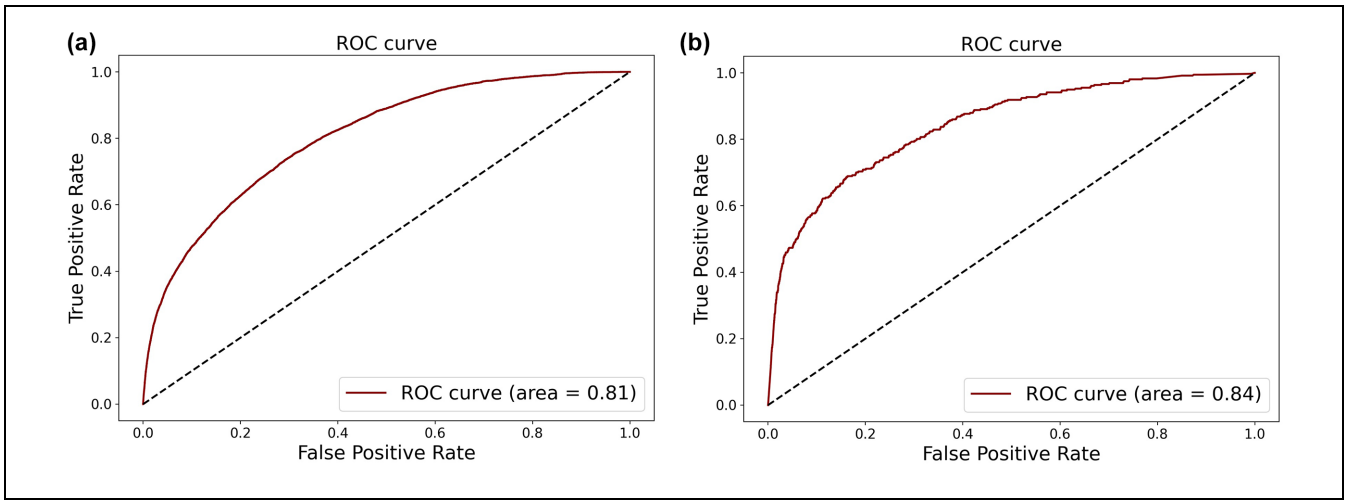

Figure 5. ROC plots for all and severe crash occurrence models: (*a*) ROC curve of all crash occurrence model and (*b*) ROC curve of severe crash occurrence model.
Note: ROC: Receiver Operating Characteristic.

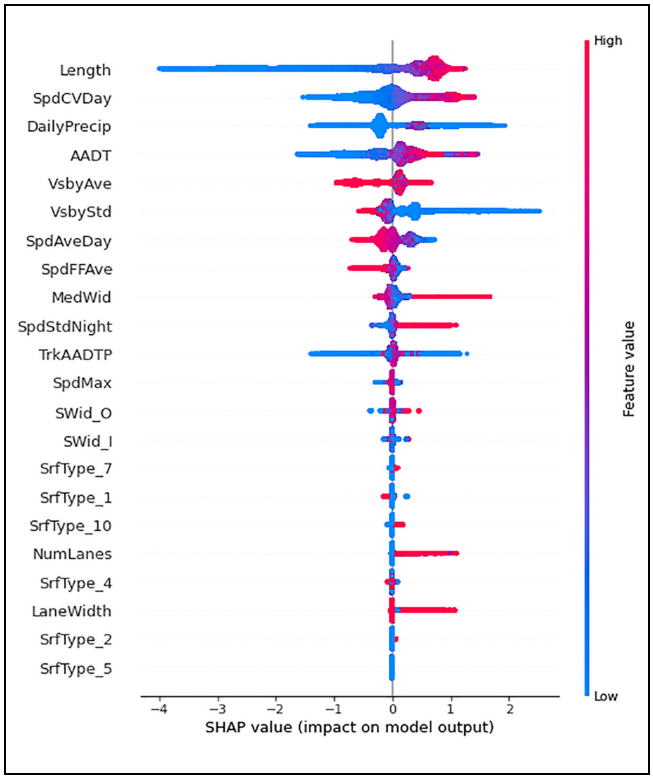

Figure 6. SHAP (SHapley Additive exPlanation) summary plot (all crash occurrence model).

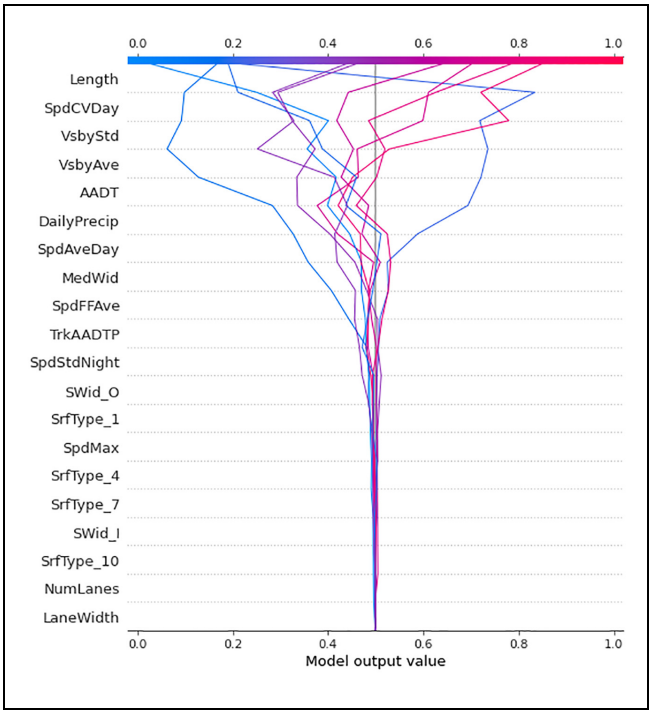

Figure 7. SHAP (SHapley Additive exPlanation) decision plot (all crash occurrence model).

of the all crash occurrence model. It ranks all explanatory variables based on their impact on the model output. Higher feature importance indicates that the variable has a greater weight in determining the classification of observations as crash or non-crash. Figure 7 is the SHAP decision plot of the all crash occurrence model. The decision plot shows how the model reaches its decision about an observation based on its explanatory variables' values. The SHAP decision plot's straight vertical line marks the model's base value. Each prediction is marked as colored lines. In Figure 7, several predictions of the model were randomly selected. Starting at the bottom of the plot, the

prediction line shows how the SHAP values accumulate from the base value to arrive at the model's final score at the top of the plot.

The most important feature identified by SHAP is segment length. This is no surprise because longer roadway segments have a greater likelihood of daily crash occurrence. This study uses the RHiNO database as the base network. In RHiNO, roadways are segregated into different lengths; this is a limitation of this study. Generally, an equal length segment is used in the roadway segmentation process (*2*). Another possible way to address the problem of unequal segment length is to normalize crash frequency by measuring crashes per mile. However, since this study applies a machine learning method to solve a binary classification problem (i.e., crash and non-crash), normalizing crash frequency does not make any difference. The second most important feature is the coefficient of variation (CV) of daily speed during the daytime. The third most important feature is daily precipitation. A higher precipitation level has a positive impact on the model's outcome. Daily precipitation level is a strong indicator of daily crash occurrence. Other top important features are average annual daily traffic (AADT), average visibility, the standard deviation of visibility, and average daily speed during the daytime. The results show that weather condition variables, especially daily precipitation, have significant impacts on daily crash occurrence. In previous studies, the importance of weather conditions seems to be overlooked in comparison with geometric and speed factors (*42*). The SHAP dependence plots in Figure 8 are presented to show the collaborative effect of roadway geometry, speed distribution, and weather conditions on crash occurrence. Figure 8*a* presents the SHAP dependency plot between daily precipitation and average daytime speed. When daily precipitation levels are near zero, they make little contribution to distinguishing crash observations and non-crash observations because there are crashes that happen on non-raining days as well. However, as the value of daily precipitation increases, the impact of this explanatory variable becomes positive. This indicates that precipitation tends to cause an increase in daily crash occurrences. Interestingly, when daily precipitation is greater than zero, a higher daily precipitation level does not increase the chance of crash occurrence. This indicates that as long as the daily precipitation is greater than zero, crashes are equally likely to occur at whatever the precipitation level is. As for average daytime speed, as shown in Figure 8*a*, when the precipitation level is greater than zero, higher average daytime speeds are more likely to cause daily crash occurrences. This is different from the general contribution of this explanatory variable to the model output. It is clearly shown in Figure 6 that higher average daytime speeds tend to have a negative effect on the model's output. Figure 8*b* is the SHAP dependency plot between average visibility and the standard deviation of visibility. Similar to daily precipitation, when the value of visibility is near 10 (the maximum value), it makes little contribution to distinguishing between crash and non-crash observations because there are crashes that occur on clear days as well. When average visibility starts to decrease, it tends to have a positive impact on the model's output. It is noteworthy that on the left side of the SHAP dependence plot, a lower visibility standard deviation tends to make a positive contribution to the model's output when average visibility is low. This is because, if average visibility is low and the standard deviation is also low, the adverse visibility condition barely changes throughout the day, which is a hazardous condition for drivers.

Figure 8*c* presents the SHAP dependency plot between median width and daily precipitation. Larger median width tends to decrease slightly the probability of all crash occurrence probability. Moreover, higher precipitation seems to decrease all crash occurrence probability when the median width is narrow and increase all crash occurrence probability when the median width is wide. Figure 8*d* is the SHAP dependency plot between median width and average visibility. A similar pattern can also be observed here. When the median width is wide, lower average visibility tends to increase the probability of all crash occurrences.

In Figure 8*e* the impact of speed CV and daily precipitation on the model's output is presented. As shown on the SHAP dependency plot, larger daytime speed CV values tend to push the model's output toward positive. Interestingly, the red dots in the dependency plot indicate that when daily precipitation levels are high, larger daytime speed CV values are more likely to cause daily crash occurrence. This shows that the effect of speed CV on daily crash occurrence becomes more significant under rainy weather conditions. Figure 8*f* is the dependence plot between AADT and daily precipitation. A higher AADT increases all crash occurrence probability. However, higher daily precipitation seems to decrease the probability of all crash occurrences when the AADT is at higher levels.

Severe Crash Occurrence Model

This model considers crash occurrence with severe levels (KA). A value of zero indicates a severe crash observation and one indicates a non-severe crash observation. The confusion matrix for this model is also in Table 6. Figure 5*b* is the ROC plot of the severe crash occurrence model. This study defines severe crashes as those that led to death or severe injuries. In Figure 9, most of the top

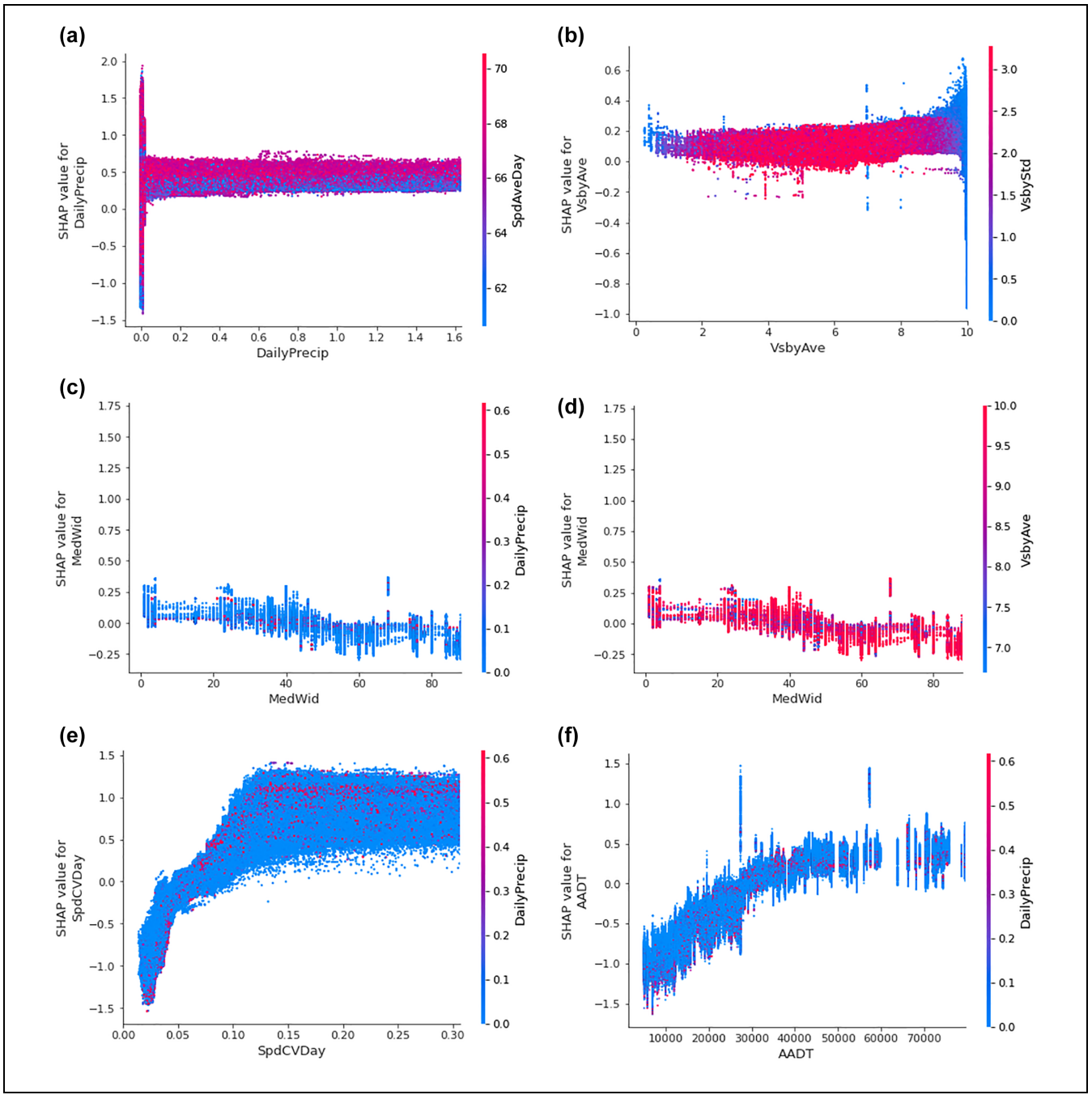

Figure 8. SHAP (SHapley Additive exPlanation) dependency plots (all crash occurrence model): (*a*) SHAP dependency plot between daily precipitation and average daytime speed, (*b*) SHAP dependency plot between average visibility and the standard deviation of visibility, (*c*) SHAP dependency plot between median width and daily precipitation, (*d*) SHAP dependency plot between median width and average visibility, (e) SHAP dependency plot between daytime speed coefficient of variation (CV) and daily preciptation, and (*f*) SHAP dependency plot between average annual daily traffic (AADT) and daily precipitation.

important features remain the same. However, several features' rankings have changed significantly, including daily precipitation, nighttime speed standard deviation, and average daytime speed. In the severe crash occurrence model, daily precipitation ranks in 14th place. This means that daily precipitation makes little contribution to the model's output. Figure 10 is the SHAP decision plot of the severe crash occurrence model. It also shows that the daily precipitation variable does not affect the model's decision too much, and the nighttime speed

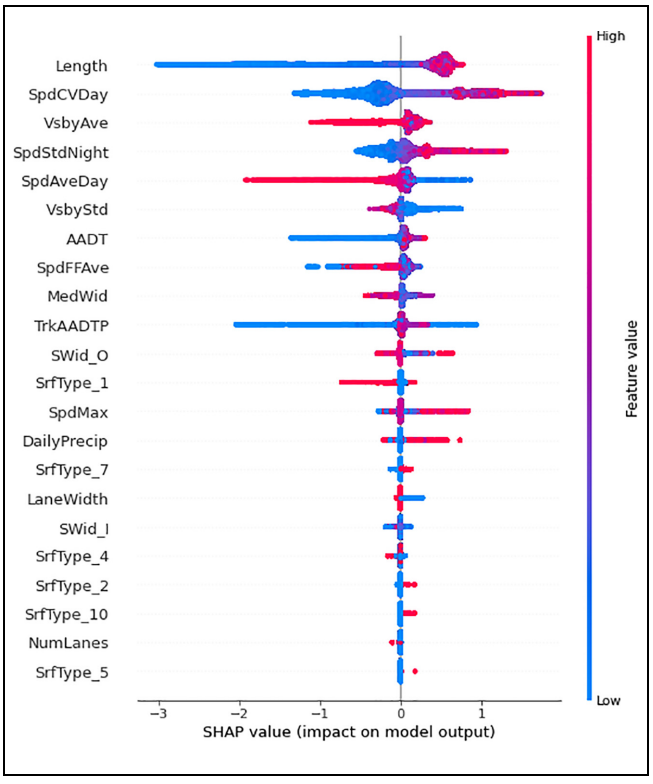

Figure 9. SHAP (Shapley Additive exPlanation) summary plot (severe crash occurrence model).

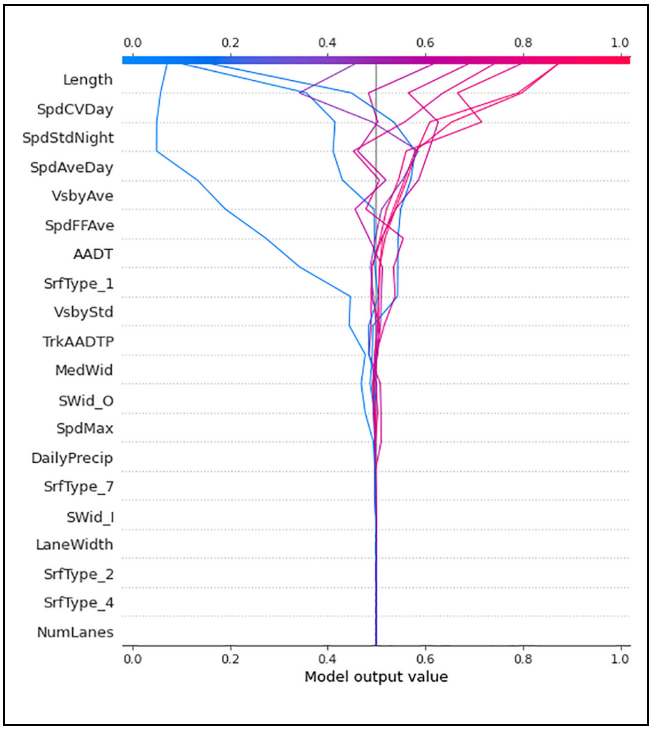

Figure 10. SHAP (SHapley Additive exPlanation) decision plot (severe crash occurrence model).

standard deviation variable has a much stronger influence on the model decision in comparison to the all crash occurrence model.

Figure 11*a* presents the SHAP dependence plot between daily precipitation and average daytime speed. There is no obvious pattern observed in this SHAP dependency plot. At different daily precipitation levels, the impacts on the model's output are evenly distributed along the y-axis. This indicates that the level of daily precipitation is not related to severe crash occurrences. Figure 11*b* also shows this finding. Similar to the all crash occurrence model, a higher speed CV increases the probability of severe crash occurrence. However, when the speed CV is high, the daily precipitation level seems to be unrelated to the model's output. Figure 11*c* presents the dependence plot between nighttime speed standard deviation and median width. Higher nighttime speed variation significantly increases the probability of severe crash occurrence. Figure 11*d* shows that when the average speed is lower than a given threshold (around 68 mph), the probability of a severe crash occurrence increases and remains around the same level.

Although higher daily precipitation levels have a significant impact on the occurrence of crashes of all severity levels, the level of daily precipitation has little impact on the daily occurrence of severe crashes. This is likely to be because when people drive in rainy weather, they tend to drive more carefully. Even though crashes are more likely to happen during rainy weather, it does not necessarily cause severe crashes. On the other hand, visibility factors (both average and standard deviation) still seem to play an important role in distinguishing the occurrence of severe crashes.

The other two features whose rankings changed significantly in the severe crash occurrence model are nighttime speed standard deviation and average daytime speed. In the all crash occurrence model, the standard deviation of nighttime speed is ranked in 10th place. In the severe crash occurrence model, the rank of nighttime speed standard deviation rises to fourth place, and the rank of the average daytime speed changes from seventh to fifth. These changes indicate that the nighttime speed standard deviation and average daytime speed become more important in contributing to severe crash occurrences.

Conclusion

This study investigates the collaborative effect of roadway geometry, speed distribution, and weather conditions on daily crash occurrences with different severity levels on rural Interstate highways. The results from machine learning models show that speed distribution

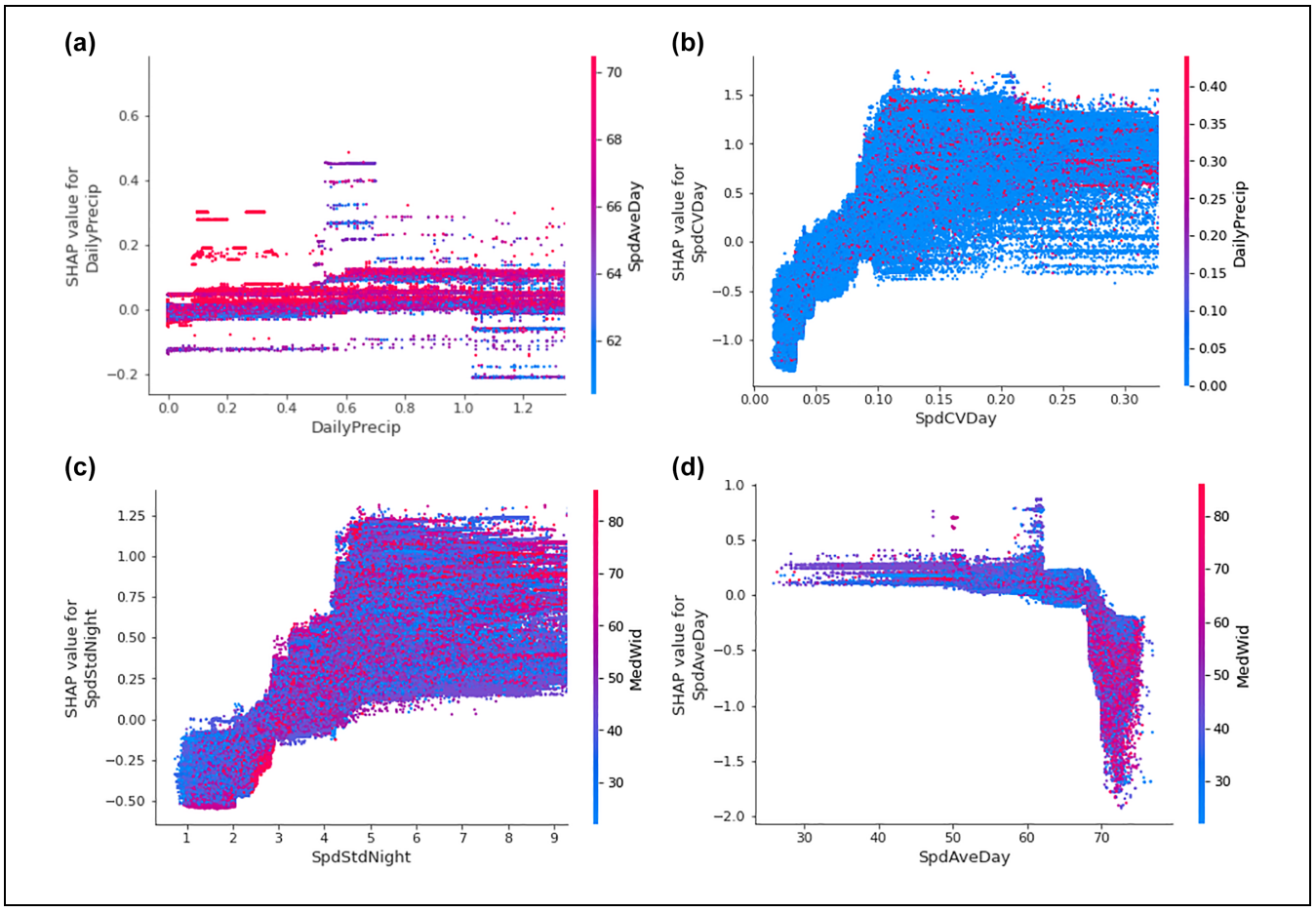

Figure 11. SHAP (SHapley Additive exPlanation) dependency plots (severe crash occurrence model): (*a*) SHAP dependency plot between precipitation and daytime average speed, (*b*) SHAP dependency plot between speed coefficient of variation (CV) and daily precipitation, (*c*) SHAP dependency plot between nighttime speed standard deviation and median width, and (*d*) SHAP dependency plot between daytime average speed and median width.

and weather conditions are the main factors that result in crash occurrence, and moreover, these factors tend to have various impacts on severe crash occurrences compared with all crash occurrences. The key findings of this study include:

- Weather factors (precipitation and visibility) are the main influential factors of rural Interstate highway daily crash occurrence.
- Daily precipitation is highly ranked in the all crash occurrence model. However, its rank falls significantly in the severe crash occurrence model. This indicates that precipitation is more likely to cause crashes on rural Interstate highways, but it does not necessarily lead to severe crash occurrences.
- Generally, a high average daytime speed has a negative effect on crash occurrence. However, along with a higher daily precipitation level, a high average daytime speed is actually more likely to cause crashes on rural Interstate highways.
- Low daily visibility and low standard deviation (i.e., the visibility remains at a low level throughout the day) are more likely to lead to higher crash occurrences on rural Interstate highways.
- The nighttime speed standard deviation is a strong contributor to severe crash occurrences. A higher nighttime speed standard deviation is more likely to cause severe crashes. However, the nighttime speed standard deviation does not show the same importance in the all crash occurrence model.
- In general, speed distribution factors are more significant contributors in the severe crash model than in the all crash model. This indicates that severe crash occurrence is more closely related to speed distribution factors in comparison to all crash occurrences on rural Interstate highways.

For precipitation, many previous studies concluded that higher daily precipitation levels have a positive relationship with crash occurrence (*19*). For fatal crashes, Eisenberg (*18*) concluded that precipitation has a negative relationship with the monthly fatal crash occurrence and a positive relationship with daily fatal crash occurrence. In this study, the results agree that precipitation is more likely to cause all crash occurrences. However, a higher level of precipitation does not necessarily increase the crash occurrence likelihood. This means that once precipitation is greater than zero for one day, the crash occurrence likelihood is almost the same, whatever the precipitation level. As for fatal crash occurrence, the findings of this study show neither a positive relationship nor a negative relationship between daily precipitation and fatal crash occurrence. The importance of daily precipitation is low in the severe crash model. Many previous studies concluded that higher speed variation is more likely to result in crash occurrence (*14*, *15*). However, previous researchers have different conclusions on average speed. The finding of this paper is that, on rural Interstate highways, a lower average speed is more likely to cause crash occurrence. This finding echoes the results of the study conducted by Pei et al. (*16*), in which the authors found that average speed has a negative relationship with crash occurrence when time exposure is considered. As for roadway geometry and traffic factors, the overall importance of the features from this category is less compared with the other two. AADT is the most important geometric and traffic feature, and it has a positive relationship with crash occurrence. Other important geometric and traffic features are median width and percentage of truck AADT.

The current study has several limitations. First, the collected data have missing variable information such as the presence of a ramp, the presence of an interchange, and other data about the surroundings of an Interstate segment. Second, since this study only focuses on the data of rural Interstate roadways, future studies can explore daily level modeling using other roadway functional classes such as rural two-lane roadways, rural multilane roadways, and urban roadways. A more comprehensive study can be conducted by developing daily level models for all rural and urban facility types to provide more generalized results. Third, since daily aggregation intervals still average out a lot of information, in future studies data with smaller aggregation intervals, such as hourly intervals, can be applied. Future studies could explore using day-based sunset and sunrise times to determine a more accurate distinction between day and night. Fourth, we applied only a few machine learning models based on our engineering judgment and preliminary analysis. Other advanced machine learning and deep learning models can be applied to this dataset too. Future studies can explore the usage of other advanced algorithms. Moreover, the relationships discovered in this paper between crash occurrence and roadway geometry, speed distribution, and weather condition factors do not indicate direct causality, they only indicate that crash events are more likely to occur under certain conditions.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Z. Wei; data collection: Z. Wei, S. Das; analysis and interpretation of results: Z. Wei; draft manuscript preparation: Z. Wei, Y. Zhang, S. Das. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Zihang Wei https://orcid.org/0000-0002-1790-022X
Subasish Das https://orcid.org/0000-0002-1671-2753

References

1. Washington, S., M. G. Karlaftis, F. Mannering, and P. Anastasopoulos. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC, New York, NY, 2020.
2. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis & Prevention*, Vol. 27, No. 3, 1995, pp. 371–389. https://doi.org/10.1016/0001-4575(94)00078-Z.
3. Das, S., A. Dutta, and X. Sun. Patterns of Rainy Weather Crashes: Applying Rules Mining. *Journal of Transportation Safety & Security*, Vol. 12, No. 9, 2020, pp. 1083–1105. https://doi.org/10.1080/19439962.2019.1572681.
4. Theofilatos, A., and G. Yannis. A Review of the Effect of Traffic and Weather Characteristics on Road Safety. *Accident Analysis & Prevention*, Vol. 72, 2014, pp. 244–256.
5. Choudhary, P., M. Imprialou, N. R. Velaga, and A. Choudhary. Impacts of Speed Variations on Freeway Crashes by Severity and Vehicle Type. *Accident Analysis & Prevention*, Vol. 121, 2018, pp. 213–222.
6. Quddus, M. Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, Vol. 5, No. 1, 2013, pp. 27–45. https://doi.org/10.1080/19439962.2012.705232.

7. Wang, X., Q. Zhou, M. Quddus, T. Fan, and S. Fang. Speed, Speed Variation and Crash Relationships for Urban Arterials. *Accident Analysis & Prevention*, Vol. 113, 2018, pp. 236–243. https://doi.org/10.1016/j.aap.2018.01.032.
8. Lord, D., and B. N. Persaud. Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. *Transportation Research Record: Journal of the Transportation Research Board*, 2000. 1717: 102–108.
9. Mountain, L., M. Maher, and B. Fawaz. The Influence of Trend on Estimates of Accidents at Junctions. *Accident Analysis & Prevention*, Vol. 30, No. 5, 1998, pp. 641–649. https://doi.org/10.1016/S0001-4575(98)00009-8.
10. Dutta, N., and M. D. Fontaine. Improving Freeway Segment Crash Prediction Models by Including Disaggregate Speed Data From Different Sources. *Accident Analysis & Prevention*, Vol. 132, 2019, p. 105253. https://doi.org/10.1016/j.aap.2019.07.029.
11. Miaou, S.-P., and H. Lum. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis & Prevention*, Vol. 25, No. 6, 1993, pp. 689–709. https://doi.org/10.1016/0001-4575(93)90034-T.
12. Anderson, I. B., K. M. Bauer, D. W. Harwood, and K. Fitzpatrick. Relationship to Safety of Geometric Design Consistency Measures for Rural Two-Lane Highways. *Transportation Research Record: Journal of the Transportation Research Board*, 1999. 1658: 43–51.
13. Haghighi, N., X. C. Liu, G. Zhang, and R. J. Porter. Impact of Roadway Geometric Features on Crash Severity on Rural Two-Lane Highways. *Accident Analysis & Prevention*, Vol. 111, 2018, pp. 34–42. https://doi.org/10.1016/j.aap.2017.11.014.
14. Garber, N. J., and R. Gadiraju. Factors Affecting Speed Variance and its Influence on Accidents. *Transportation Research Record: Journal of the Transportation Research Board*, 1989. 1213: 64–71.
15. Lee, C. K., F. Saccomanno, and B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 2002. 1784: 1–8.
16. Pei, X., S. C. Wong, and N. N. Sze. The Roles of Exposure and Speed in Road Safety Analysis. *Accident Analysis & Prevention*, Vol. 48, 2012, pp. 464–471.
17. Scott, P. P. Modelling Time-Series of British Road Accident Data. *Accident Analysis & Prevention*, Vol. 18, No. 2, 1986, pp. 109–117. https://doi.org/10.1016/0001-4575(86)90055-2.
18. Eisenberg, D. The Mixed Effects of Precipitation on Traffic Crashes. *Accident Analysis & Prevention*, Vol. 36, No. 4, 2004, pp. 637–647. https://doi.org/10.1016/S0001-4575(03)00085-X.
19. Brijs, T., D. Karlis, and G. Wets. Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time-Series Model. *Accident Analysis & Prevention*, Vol. 40, No. 3, 2008, pp. 1180–1190. https://doi.org/10.1016/j.aap.2008.01.001.
20. Jaroszweski, D., and T. McNamara. The Influence of Rainfall on Road Accidents in Urban Areas: A Weather Radar Approach. *Travel Behaviour and Society*, Vol. 1, No. 1, 2014, pp. 15–21.
21. Yu, R., and M. Abdel-Aty. Analyzing Crash Injury Severity for a Mountainous Freeway Incorporating Real-Time Traffic and Weather Data. *Safety Science*, Vol. 63, 2014, pp. 50–56.
22. Wei, Z., S. Das, and Y. Zhang. Short Duration Crash Prediction for Rural Two-Lane Roadways: Applying Explainable Artificial Intelligence. *Transportation Research Record: Journal of the Transportation Research Board*, 2022.
23. Texas Department of Transportation. Roadway Inventory. https://www.txdot.gov/inside-txdot/division/transportation-planning/roadway-inventory.html. Accessed February 23, 2021.
24. Das, S., S. Geedipally, R. Avelar, L. Wu, K. Fitzpatrick, M. Banihashemi, and D. Lord. *Rural Speed Safety Project for USDOT Safety Data Initiative: [Supporting Dataset]*. Texas A&M Transportation Institute, College Station, March 1, 2020.
25. Federal Highway Administration. The National Performance Management Research Data Set (NPMRDS) and Application for Work Zone Performance Measurement. https://ops.fhwa.dot.gov/publications/fhwahop20028/index.htm. Accessed February 23, 2021.
26. American Association of State Highway and Transportation Officials (AASHTO). *Highway Safety Manual*. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
27. Li, P., M. Abdel-Aty, and J. Yuan. Real-Time Crash Risk Prediction on Arterials Based on LSTM-CNN. *Accident Analysis & Prevention*, Vol. 135, 2020, p. 105371. https://doi.org/10.1016/j.aap.2019.105371.
28. Chen, T., T. He, M. Benesty, and Y. Tang. *Understand Your Dataset With Xgboost*. 2018. https://cran.r-project.org/web/packages/xgboost/vignettes/discoverYourData.html
29. Abdel-Aty, M., N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia. Predicting Freeway Crashes From Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. 1897: 88–95.
30. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357. https://doi.org/10.1613/jair.953.
31. Yuan, J., M. Abdel-Aty, Y. Gong, and Q. Cai. Real-Time Crash Risk Prediction Using Long Short-Term Memory Recurrent Neural Network. *Transportation Research Record: Journal of the Transportation Research Board*, 2019. 2673: 314–326.
32. Parsa, A. B., H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian. Real-Time Accident Detection: Coping With Imbalanced Data. *Accident Analysis & Prevention*, Vol. 129, 2019, pp. 202–210. https://doi.org/10.1016/j.aap.2019.05.014.
33. Chen, T., and C. Guestrin. XGBoost: A Scalable Tree Boosting System. *Proc., 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, 2016.

34. Breiman, L. Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32. https://doi.org/10.1023/A:1010933404324.
35. Pham, M.-H., A. Bhaskar, E. Chung, and A.-G. Dumont. Random Forest Models for Identifying Motorway Rear-End Crash Risks Using Disaggregate Data. *Proc., 13th International IEEE Conference on Intelligent Transportation Systems*, Funchal, Portugal, 2010.
36. You, J., J. Wang, and J. Guo. Real-Time Crash Prediction on Freeways Using Data Mining and Emerging Techniques. *Journal of Modern Transportation*, Vol. 25, No. 2, 2017, pp. 116–123.
37. Zhou, X., P. Lu, Z. Zheng, D. Tolliver, and A. Keramati. Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared With Decision Tree. *Reliability Engineering & System Safety*, Vol. 200, 2020, p. 106931. https://doi.org/10.1016/j.ress.2020.106931.
38. Mondal, A. R., M. A. E. Bhuiyan, and F. Yang. Advancement of Weather-Related Crash Prediction Model Using Nonparametric Machine Learning Algorithms. *SN Applied Sciences*, Vol. 2, No. 8, 2020, p. 1372. https://doi.org/10.1007/s42452-020-03196-x.
39. Freund, Y., and R. E. Schapire. *Experiments With a New Boosting Algorithm*. In Proc.,13th International Conference on International Conference on Machine Learning (ICML'96). Morgan Kaufmann Publishers Inc., San Francisco, CA, 1996, pp. 148–156.
40. Tang, J., J. Liang, C. Han, Z. Li, and H. Huang. Crash Injury Severity Analysis Using a Two-Layer Stacking Framework. *Accident Analysis & Prevention*, Vol. 122, 2019, pp. 226–238. https://doi.org/10.1016/j.aap.2018.10.016.
41. Lundberg, S., and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*, 2017.
42. Parsa, A. B., A. Movahedi, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accident Analysis & Prevention*, Vol. 136, 2020, p. 105405.