TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
1–15

DOI: 10.1177/03611981221096113
journals.sagepub.com/home/trr
SAGE

Short Duration Crash Prediction for Rural Two-Lane Roadways: Applying Explainable Artificial Intelligence

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

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. In this study, the combined effect of roadway geometry, speed distribution, and weather conditions on crash occurrence and severity was investigated on short duration daily level crash data. This study collected data from four different sources on rural two-lane roadways in Texas. A machine learning method, XGBoost (eXtreme Gradient Boosting), was applied to train the data. To mitigate imbalanced data problems, a synthetic minority oversampling technique (SMOTE) was applied. The XGBoost model was trained separately on all crash occurrences and severe crash occurrences. Finally, an explainable artificial intelligence (AI) technique, SHAP (SHapley Additive exPlanation), was applied to investigate the contribution of all variables to the model's output. The results show that annual average daily traffic has a significant impact on all crash occurrences and severe crash (fatal and incapacitating injury) occurrences on rural two-lane roadways. Moreover, weather condition factors including daily precipitation, average visibility, and the standard deviation of visibility show association with high crash occurrences. The short duration crash prediction models of this study can provide more insights into the relationships between crash, geometric variables, traffic exposure, weather, and operating speed.

Keywords

data and data science, artificial intelligence, safety, crash analysis, crash data, crash severity

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 change significantly during this time period are typically not considered because granular level 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 very important for crash analysis as well. Many studies have found that weather conditions, especially precipitation and visibility (*2–4*), and speed distribution (*5–7*) are closely related to crash occurrence. Aggregation of these variables over a long time period can introduce biased results because the variation within this time period cannot be fully explored. Short duration models may overcome these biases. In particular, data can be aggregated by day, hour, or even minute.

Note that the conventional explanatory variables such as road geometry data (i.e., curve, lane width, shoulder width, etc.) are relatively static in nature. 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 each other always share similar geometry 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*).

There are three main aggregation intervals previously applied by researchers. The first is the yearly aggregation interval, under which the effect of roadway geometry can

[1]Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX
[2]Texas A&M Transportation Institute, San Antonio, TX

Corresponding Author:
Zihang Wei, wzh96@tamu.edu

be analyzed. However, the effects of speed distribution and weather conditions cannot be analyzed under this. The second is the daily aggregation interval, under which the effects of roadway geometric and weather conditions can be analyzed, but it is not ideal for studying speed distribution. The third is hourly or minute by minute, under which the effect of speed distribution can be analyzed. However, it is not ideal for studying the effects of roadway geometry and weather conditions. This study chose to use daily aggregation intervals. As 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, data needs to be aggregated into daily intervals, but they should also include lots of roadway segments with various geometry features. Thus, this study includes a wide range of rural two-lane roadway segments to address this issue.

The size of the data included in this study is in the periphery of big data. Thus, artificial intelligence techniques are suitable for this data analysis. In this study, the XGBoost algorithm is trained on the data set. Moreover, 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. Thus, in this study, apart from applying machine learning algorithms to analyze the data set, the SHAP (SHapley Additive exPlanation) method is applied to unearth the relationship between explanatory variables and the model's outcomes.

With regard to crash occurrence with different severity levels, the contributing factors of severe (fatal and incapacitating injury) crash occurrences tend to be different from those of all crash occurrences. As a result, this study trains another machine learning model on severe crash occurrence to compare the different factors behind all crash occurrence and severe crash occurrence. The results of this study can reveal the collaborative effects of roadway geometry, weather conditions, and speed distribution on daily crash occurrence on Texas rural two-lane roadways and their effects on crash severity patterns.

Literature Review

Many previous studies explored 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 a monthly interval. 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 the one using annual data without speed information.

There are many studies that have analyzed the relationship between roadway geometric features and crash occurrence or severity. Miaou and Lum (*11*) used 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 the variation in crash severity levels.

Speed distribution is another important contributor to crash occurrence. Garber and Gadiraju (*14*) investigated the impact of mean speed and speed variation on crash rates. 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 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 (HPLN) 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); each time period is 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 also significantly affect crash occurrence. 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 dependency 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 using weather radar images. This novel approach offered improvements to the analysis of weather-related crashes 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 snowy weather is less likely to cause severe crashes and that lower temperatures increase the likelihood of severe crashes.

Although previous studies have investigated these three factors separately, fewer studies have considered these factors together and analyzed their collaborative effects on crash occurrence. One major problem when studying the collaborative effects on these three factors is which data aggregation level should be used. Previous studies have analyzed the relationship between crash and roadway geometric data based on the yearly aggregation interval (*11*, *12*), the relationship between crash and weather condition variables based on the daily aggregation interval or the monthly aggregation interval (*17–19, 22, 23*), and the relationship between crashes and speed distribution based on real-time data with the hourly interval 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 using daily aggregation interval because 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.

Rural two-lane roadway is the facility type analyzed in this study. There are some other previous studies that have analyzed crash occurrence and severity on rural two-lane roadways. Wu et al. (*24*) developed mixed logit models that examined driver injury severities in single-vehicle (SV) and multi-vehicle (MV) crashes on rural two-lane highways. They identified several factors that contribute to injury severity in these two types of crash including driver behavior, weather conditions, environmental features, roadway geometric features, and traffic composition. Ma et al. (*25*) developed a partial proportional odds model to examine the factors related to injury severity on rural two-lane highways in China. The results indicated that the at-fault driver's age, the at-fault driver having a license or not, alcohol use, speeding, the pedestrian involved, type of area, weather condition, pavement type, and collision type apparently affected injury severity. Persaud et al. (*26*) examined the effect of centerline and shoulder rumble strips on crash occurrence reduction. The results show that rumble strips can reduce all types of collision (head-on collisions the most). Das et al. (*27*) examined rural two-lane run-off-road (ROR) crash data from 2010 to 2016 from Louisiana to better understand ROR crashes and improve safety predictions for rural two-lane highways. Variables including annual average daily traffic (AADT), shoulder width, and pavement width are considered to predict yearly ROR crash number. Cubist models present a better performance on ROR crash prediction and the Cubist approach was applied to present rules-based safety performance functions (SPFs) for total, fatal, and injury crashes.

Data Preparation

Data Acquisition and Processing

Data are collected from roadways in the state of Texas. First, a comprehensive data set is developed by using the data conflation method. The data set analyzed in this study contains four parts: (1) roadway geometry features and traffic information; (2) weather condition data; (3) speed measurement data; and (4) crash data. These data are collected from four different sources respectively: (1) Texas Department of Transportation (Texas DOT) Road-Highway Inventory Network Offload (RHiNO) 2018; (2) the Automated Surface Observing System (ASOS); 3) the National Performance Management Research Dataset (NPMRDS); and (4) the Crash Record Information System (CRIS).

Base Roadway Network and Geometric Data. The base roadway network is collected from the Texas DOT 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. Texas DOT submitted this data set to the Federal Highway Administration (FHWA) as part of the Highway Performance Monitoring System (HPMS) program (*28*). This study only selects rural two-lane roadways from the base network.

Speed Distribution Data. Speed data are 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 of 5 min, 15 min, and 1 h. This study uses data with 5-min intervals because more detailed information can be kept when calculating speed variation. Speed data are available across the National Highway System (NHS), and the spatial resolution is set by different Traffic Message Channel (TMC) location codes (*29*). Daily speed distribution variables (i.e., average speed, speed standard deviation, 85th percentile speed, etc.) are 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 are collected from the Automated Surface Observing System (ASOS) of the National Centers for Environmental Information (NOAA). Each roadway segment is matched with an ASOS station closest to it.

Crash Data. Crash data are 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 are assigned to the roadway segments on which they occurred. The daily crash count of each roadway segment can then be summarized. The data set used in this study contains 7,443,107 non-crash observations and 9,554 crash observations, including 466 severe crash observations. The definition of a severe crash in this study is a crash that resulted in severe injuries or fatalities. In this process, crash severity is classified into five different levels: (1) K: Killed or Fatal; (2) A: Incapacitating Injury; (3) B: Non-Incapacitating Injury; (4) C: Possible Injury; and (5) O: Not Injured or Unknown. Yannis et al. (*30*) also applied a crash and non-crash approach in their study to investigate road accident severity and likelihood in urban areas by using real-time traffic data.

The data from the four parts above are conflated by using ArcGIS software. The data preparation process is shown in Figure 1. All data are aggregated into a daily

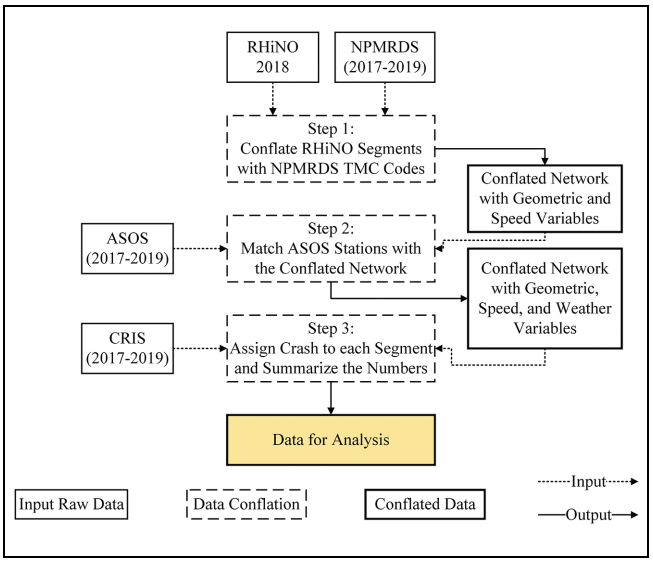

Figure 1. Flowchart of the data preparation process.
Note: RHiNO = Road-Highway Inventory Network Offload; NPMRDS = National Performance Management Research Dataset; TMC = Traffic Message Channel; ASOS = Automated Surface Observing System; CRIS = Crash Record Information System.

interval. The final data set is made up of 26 variables calculated from the above-mentioned four parts. The total number of segments, total segment length, and the number of crashes at different severity levels (KABCO) are summarized in Figure 2. In this study, all crashes include KABCO severity levels and severe crashes only include severity levels KA. The detailed definitions of all variables and their descriptive statistics are listed in Table 1.

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

Feature Selection

In the prepared data set, some explanatory variables may be highly correlated with others. When machine learning models are trained on the data set, 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.

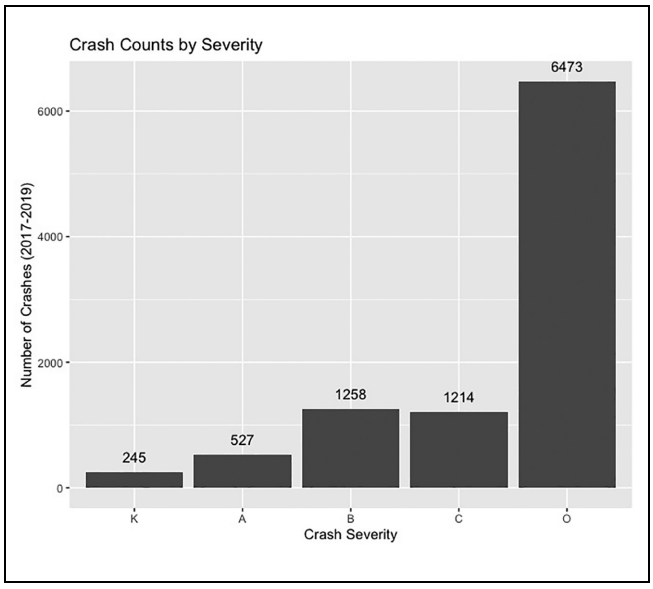

Figure 2. Crash counts for different severity levels.

At 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 3 shows the PCC heatmap of the Rural Instates data set. Brighter cells represent a higher correlation between two explanatory variables. Two explanatory variables are considered highly correlated if their PCC is higher than 0.8. In general, two variables are considered correlated if the PCC is higher than 0.5 (*31*). However, since the decision tree algorithm (including the XGBoost and AdaBoost algorithm) are very robust to correlated variables, the threshold is increased to 0.8 in this study (*32*).

For random forest and DNN models, this feature selection process helped to remove highly correlated variables and boost the model's performance. In the case of the XGBoost and AdaBoost models, this process only means removing the redundant variables in the data set to facilitate the model training speed. This is not a necessary step for training an XGBoost or an AdaBoost

Table 1. Variable Names, Definitions, and Descriptive Statistics

Variable names	Definition	Descriptive statistic			
		Mean	SD	Min.	Max.
Weather condition					
DailyPrecip	Daily precipitation	0.07	0.49	0.00	133.01
VsbyAve	Average visibility	9.35	1.21	0.14	12.50
VsbyStd	Visibility standard deviation	0.92	1.15	0.00	14.67
Speed distribution					
SpdAve	Average of daily speed	60.24	7.50	9.77	76.93
SpdStd	Standard deviation of daily speed	5.82	2.12	1.00	30.24
SpdCV	Coefficient of variation of daily speed	0.14	0.07	0.02	1.45
Spd85	85th Percentile of daily speed	67.28	6.38	10.90	87.25
RefSpd	Reference speed	70.80	5.09	36.00	78.53
SpdAveDay	Average of daily speed using data during daytime	60.28	7.79	5.58	81.40
SpdStdDay	Standard deviation of daily speed using data during daytime	5.72	2.26	0.00	39.33
SpdCVDay	Coefficient of variation of daily speed using data during daytime	0.14	0.07	0.00	1.57
SpdAveNight	Average of daily speed using data during nighttime	60.30	7.17	6.12	81.29
SpdCVNight	Coefficient of variation of daily speed using data during nighttime	0.14	0.07	0.00	1.34
SpdStdNight	Standard deviation of daily speed using data during nighttime	5.60	2.25	0.00	42.37
SpdFFAve	Average of daily speed larger than reference speed	74.09	5.04	37.00	99.00
SpdFF85	85th percentile of daily speed larger than reference speed	75.46	5.30	37.00	99.00
Roadway geometry and traffic					
SpdMax	Maximum speed limit	66.90	10.26	30.00	75.00
MedWid	Median width	0.00	0.00	0.00	0.00
NumLanes	Number of through lanes	2.00	0.00	2.00	2.00
LaneWidth	Lane width	12.53	1.67	6.00	32.00
SWid_I	Inside shoulder width	8.24	2.51	0.00	25.00
SWid_O	Outside shoulder width	8.48	2.49	0.00	25.00
SrfType	Surface type (categorical)				
AADT	AADT	4,786.79	3,037.81	134.00	26,674.00
TrkAADTP	Truck AADT percentage	20.54	10.77	1.20	66.60
Length	Roadway segment length	0.38	0.49	0.00	2.00

Note: Min. = minimum; Max. = maximum; SD = standard deviation; AADT = annual average daily traffic. The unit of speed is miles per hour (mph) and the unit of shoulder and lane width is feet (ft).

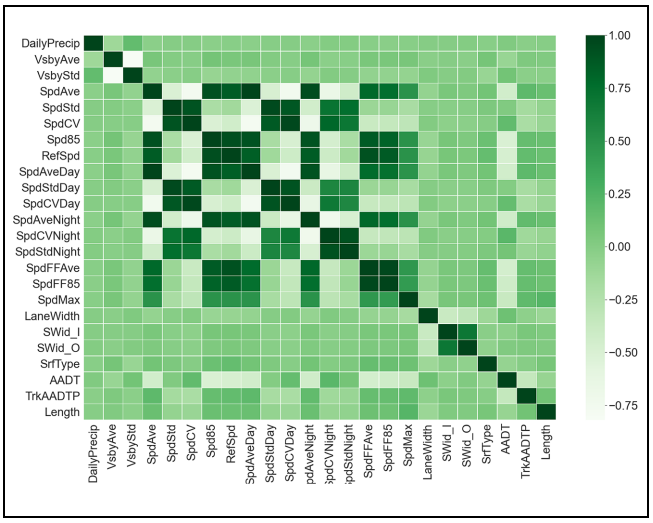

Figure 3. Pearson correlation coefficient heatmap.

Table 2. Correlated Variable Pairs

Identified correlated variable pairs		Absolute PCC
Spd85	SpdFF85	0.828
SpdAveDay	RefSpd	0.845
SpdAveNight	RefSpd	0.862
SpdAve	RefSpd	0.862
SpdFFAve	Spd85	0.867
SpdFF85	RefSpd	0.869
SpdCV	SpdStdDay	0.873
SpdStd	SpdCVDay	0.893
SpdAveNight	Spd85	0.909
SpdStdDay	SpdCVDay	0.912
SpdStd	SpdCV	0.912
SpdAveDay	SpdAveNight	0.921
SpdAveDay	Spd85	0.921
SpdCVNight	SpdStdNight	0.924
RefSpd	SpdFFAve	0.926
Spd85	SpdAve	0.932
RefSpd	Spd85	0.939
SpdAve	SpdAveNight	0.955
SpdStdDay	SpdStd	0.961
SpdCV	SpdCVDay	0.975
SpdFFAve	SpdFF85	0.976
SpdAveDay	SpdAve	0.994

Note: PCC = Pearson correlation coefficient.

model. There are 23 pairs of highly correlated explanatory variables identified from the data set (see Table 2).

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

After the Pearson correlation coefficients were calculated for all explanatory variable pairs, a random forest (RF) model was trained using all explanatory variables on the data set 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 (Spd85, RefSpd, SpdCV, SpdStd, SpdStdDay, SpdAveNight, SpdStdNight, SpdFFAve, SpdAveDay) were removed from the data set.

Resampling Imbalanced Data Set

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

This study applies the synthetic minority oversampling technique (SMOTE) to rebalance the original data set. Many previous studies have applied resampling methods. Abdel-Aty et al. (*33*) applied a matched case-control method that manually matched crash samples with non-crash samples. Chawla et al. (*34*) first proposed SMOTE to address the problem of imbalanced data sets. SMOTE is an oversampling method that only oversamples the minority class and is only applied to the training data set. 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 data set, the testing result can still be considered to reflect reality. Many previous studies have applied SMOTE to address imbalanced data sets (*31*, *35*, *36*). Before resampling, the number of crash observations and the number of non-crash observations had a huge difference (see Figure 4*a*). After resampling, the number of crash observations in the training set is now equal to the number of non-crash observations (see Figure 4*b*). Note that in Figure 4, only two dimensions ("DailyPrecip" and "SpdCVDay") are selected to visualize the SMOTE process.

Methodology

Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable end-to-end tree boosting system.

It implements machine learning algorithms under the Gradient Boosting framework (*37*). XGBoost is an additive-boosting tree package that is built by *k* essential

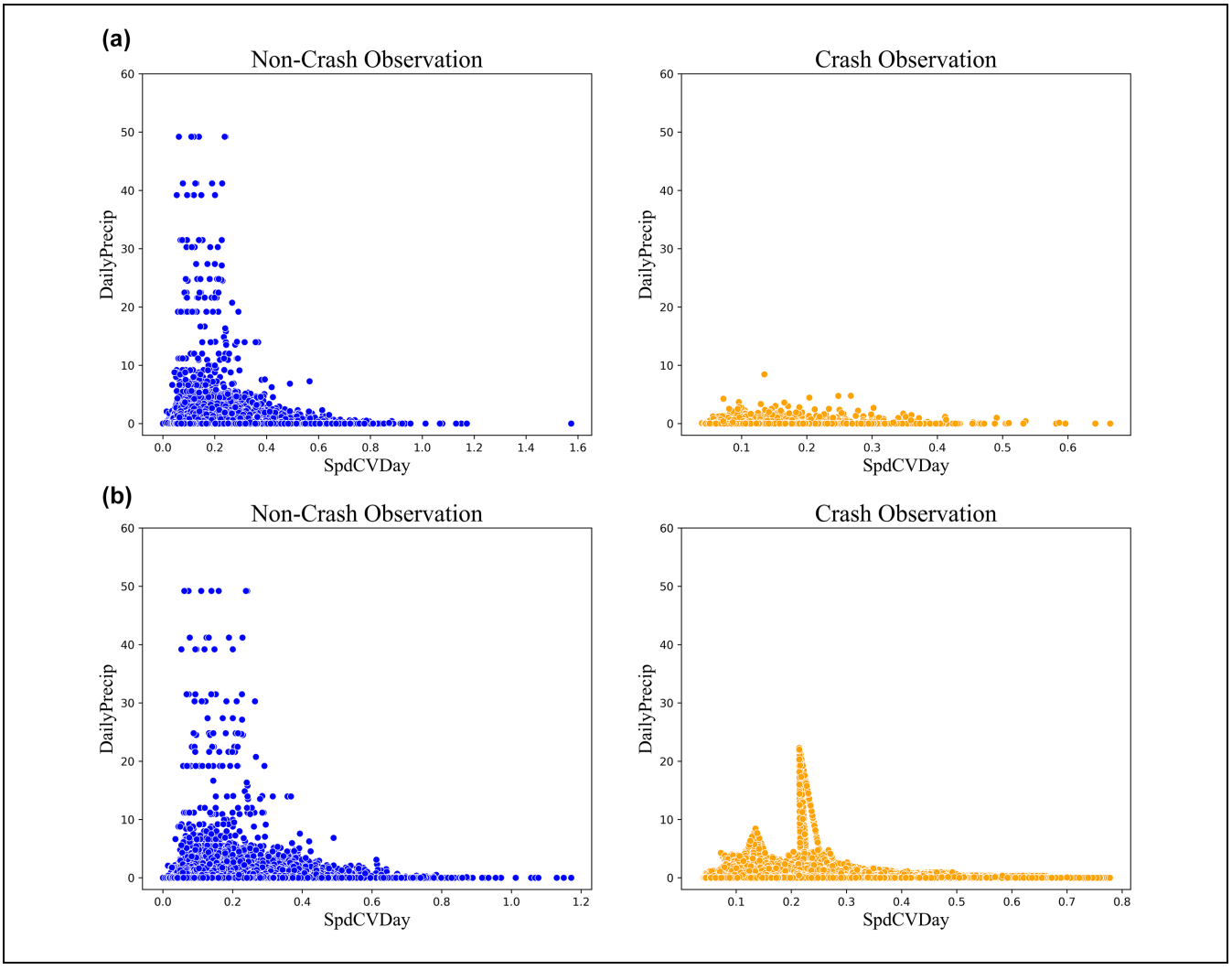

Figure 4. Synthetic minority oversampling technique (SMOTE) oversampling method: (*a*) before SMOTE oversampling (5,582,330 non-crash observations and 7,165 crash observations); and (*b*) after SMOTE oversampling (5,582,330 non-crash observations and 5,582,330 crash observations).
Note: SpdCVDay = Coefficient of variation of daily speed using data during daytime.

tree functions implemented with regularization, missing value imputation, shrinkage, column subsampling, sparsity-aware split finding, and column block for parallel learning. Compared with 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 a small number of resources. It is currently one of the fastest and best open-source boosting tree tools for modeling and prediction analysis. The detailed mathematical process of XGBoost is as follows. Interested readers can also refer to Chen and Guestrin (*37*) for more detailed information.

Given a data set 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 t^{th} iterations by adding one tree function $f(x_i)$ to the predicted value of $(t-1)^{th}$ iteration corresponding to the i^{th} 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 j^{th} leaf scores. This term is used to avoid overfitting.

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

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

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

And the corresponding minimum object value is:

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

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 on the model's output. Lundberg and Lee (*38*) 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 game-theoretic approach. For any particular prediction, the Shapely value of a feature i can be calculated as Equation 6:

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

where
ϕ_i: shapely value of feature i,
S: a possible feature subset,
$|S|$: number of features in subset S,
F: the set of all features,
$|F|$: 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, and
$f_S(x_S)$: 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 in 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 shapely value of a particular feature is calculated as the weighted average of all differences across all possible subset S.

Ma et al. (*39*) applied a similar analytical framework in their study. They applied XGBoost and SHAP to predict the occurrence of distraction-affected crashes using phone-use data.

Results and Discussion

Model Tuning

For the tuning process, the grid search method is used to find the set of hyperparameters that performs best. The performance measure for the XGBoost model is the receiver operating characteristic-Area Under Curve (ROC-AUC) score from the fine-tuning process. After fine-tuning, the hyperparameters for the XGBoost model are set as follows:

All crash model:

- Learning Rate: 0.2
- Max Depth: 4
- Number of Estimators: 500

Severe Crash model:

- Learning Rate: 0.2
- Max Depth: 3
- Number of Estimators: 600

Learning rate is the parameter that can be set to control the weight of new trees being added to the model. Max depth can be set to control the maximum depth of each tree in the model. Generally, higher max depth can result in overfitting which makes the model perform better on the training data set while the performance on the testing data set is not ideal. On the other side, lower max depth can enhance the model's performance on testing data set. However, if the max depth is too low, it can cause underfitting. Number of estimators is the parameter to control the total number of trees in the model. The performance measures of the two models (all crash model and severe crash model) on the testing data set are summarized in Table 3.

All Severity Level Model

This model considers crash occurrences with all severity levels (KABCO). A value of 0 indicates all-crash observations, and 1 indicates non-crash observations. The prepared data set is split into a training set (70% of all observations) and a testing set (30% of all observations).

Table 3. Models' Performance Measures

Performance measures (on testing data set)	All crash model (%)	Severe crash model (%)
Accuracy	75.4	81.1
Sensitivity	68.0	64.2
Specificity	75.4	81.1
Weighted accuracy	71.7	72.7

Table 4. Confusion Matrix of the XGBoost Models

		All crash model		Severe crash model	
Predicted labels		Non-crash	Crash	Non-crash	Crash
True labels	Non-crash	1,402,733	458,044	1,510,139	352,834
	Crash	764	1,625	69	124

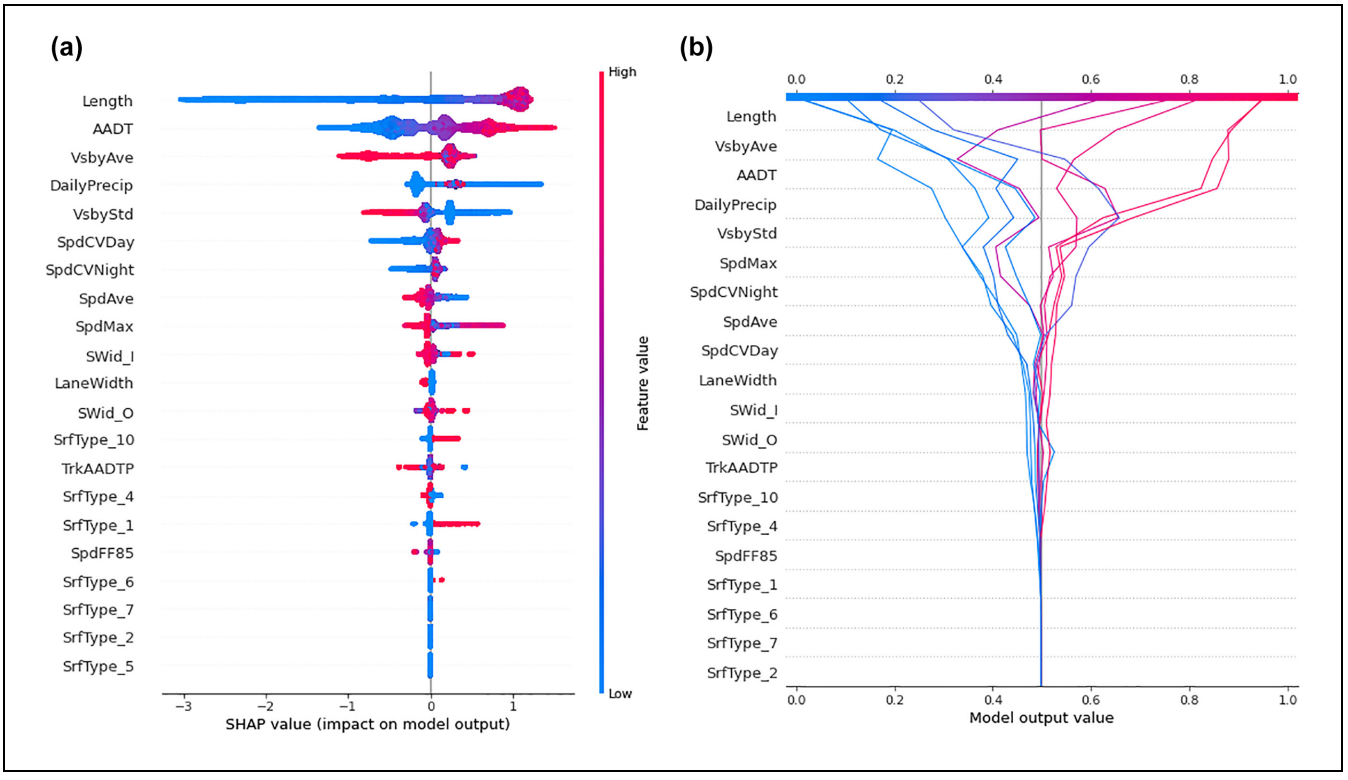

Figure 5. SHapley Additive exPlanation (SHAP) summary plot and decision plot (all crash occurrence model): (*a*) SHAP summary plot; and (*b*) SHAP decision plot.

Since the training set and the testing set are split randomly, the distributions of all variables in these two sets are supposed to be similar. The SMOTE oversampling method is applied to the training data set to balance the minority and majority groups. The training data set contains 5,582,330 non-crash observations and 7,165 crash observations. After the oversampling process, the total numbers of both non-crash and crash observations are 5,582,330. The model evaluation is made with the testing data set. Table 4 is the confusion matrix of the XGBoost model. The model performance is evaluated by four measurements (see Table 3). Figure 7*a* is the ROC plot of all crash occurrence model.

SHAP (SHapley Additive exPlanation) is applied to interpret the feature importance in the XGBoost model. Figure 5*a* is the SHAP summary plot of the all-crash occurrence model. It ranks all explanatory variables based on their impact on the model output. A higher feature importance indicates that the variable has a greater weight in determining the classification of observations as crash or non-crash. Figure 5*b* 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 the values of its explanatory variables.

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 AADT. Higher AADT tends to be more likely to cause crash occurrence. The third most important feature is average visibility. The SHAP summary plot shows that lower daily average visibility on rural two-lane roadways tends to increase the probability of all crash occurrences. Daily precipitation is the fourth most important variable. Higher precipitation tends to be more likely to cause crash occurrence on rural two-lane roadways. Other top important features are visibility standard deviation, daytime speed variation, and nighttime speed variation.

Figure 6 summarizes the dependency plots between the topmost important variables. Figure 6*a* is the SHAP dependency plot between AADT and average visibility. Higher AADT significantly increases the probability of crash occurrence on rural two-lane roadways. When AADT is low, lower visibility tends to increase the probability of crash occurrence, while when AADT is high, lower visibility tends to decrease the probability of crash occurrence. Figure 6*b* presents the SHAP dependency plot between nighttime speed CV and average visibility. According to the plot, a higher nighttime speed CV significantly increases the probability of crash occurrence. The level of average visibility does not make a difference on crash occurrence probability at any nighttime speed CV levels.

Figure 6*c* is the dependency plot between daytime speed CV and average visibility. Higher daytime speed CV increases the probability of crash occurrence on rural two-lane roadways. Lower daytime speed CV and lower average visibility decrease the probability of crash occurrence. When daytime speed CV increases to higher levels, lower average visibility tends to increase the probability of crash occurrence. Figure 6*d* is the dependency plot between daytime speed CV and daily precipitation. Higher precipitation tends to decrease the probability of crash occurrence when daytime speed CV is low. Similarly, when daytime speed CV is higher, higher precipitation still decreases the crash occurrence probability. A higher precipitation level increases crash occurrence probability only when daytime speed CV is in between. Figure 6*e* presents the SHAP dependency plot between daily precipitation and average speed. When the daily precipitation level is near 0, it makes little contribution to distinguish crash observation and non-crash observation because there are crashes that happened 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 occurrence probability. Interestingly, when daily precipitation is greater than 0, 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 0, crashes are equally likely to occur whatever the precipitation level is. As for average speed, as shown in Figure 6*e*, when the precipitation level is greater than 0, a larger average speed is more likely to cause daily crash occurrence. This is different from the general contribution of this explanatory variable to the model output. It is clearly shown in Figure 6*f* that larger average daytime speeds tend to negatively affect the model's output.

Severe Crash Occurrence Model

This model considers crash occurrence with severe levels (KA). A value of 0 indicates a non-severe crash observation, and 1 indicates a severe crash observation. The confusion matrix for the performance of this model is also in Table 4. Figure 7*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 8*a*, most of the top important features remain the same. However, several features' rankings have changed significantly. The first thing to notice is that the importance rank of daily precipitation dropped. This indicates that precipitation level does not contribute to severe crash occurrences as much as it contributes to all crash occurrences. The importance rank of average visibility, however, stays unchanged. Another noticeable point is that the importance of lane width increases in the severe crash model compared with all crash models. Lower lane width tends to increase the probability of severe crash occurrence. As for the maximum speed limit, its importance also grows in the severe crash model. In the SHAP summary plot, it shows that lower speed limit roadways are associated with more severe injury crashes. Moreover, the importance of outside shoulder width is lower than that of inside shoulder width in the all-crash model, while in the severe crash model, the importance of outside shoulder width increases and is higher than that of inside shoulder width. Figure 8*b* is the SHAP decision plot of the severe crash model.

Conclusions

This study investigated the collaborative effect of roadway geometry, speed distribution, and weather conditions on daily crash occurrence with different severity levels on rural two-lane highways. The results show that AADT, average visibility, daily precipitation, and speed variation are the main factors that affect daily crash occurrence. Moreover, these factors tend to have different impacts on severe crash occurrences compared with all crash occurrences. The key findings of this study are:

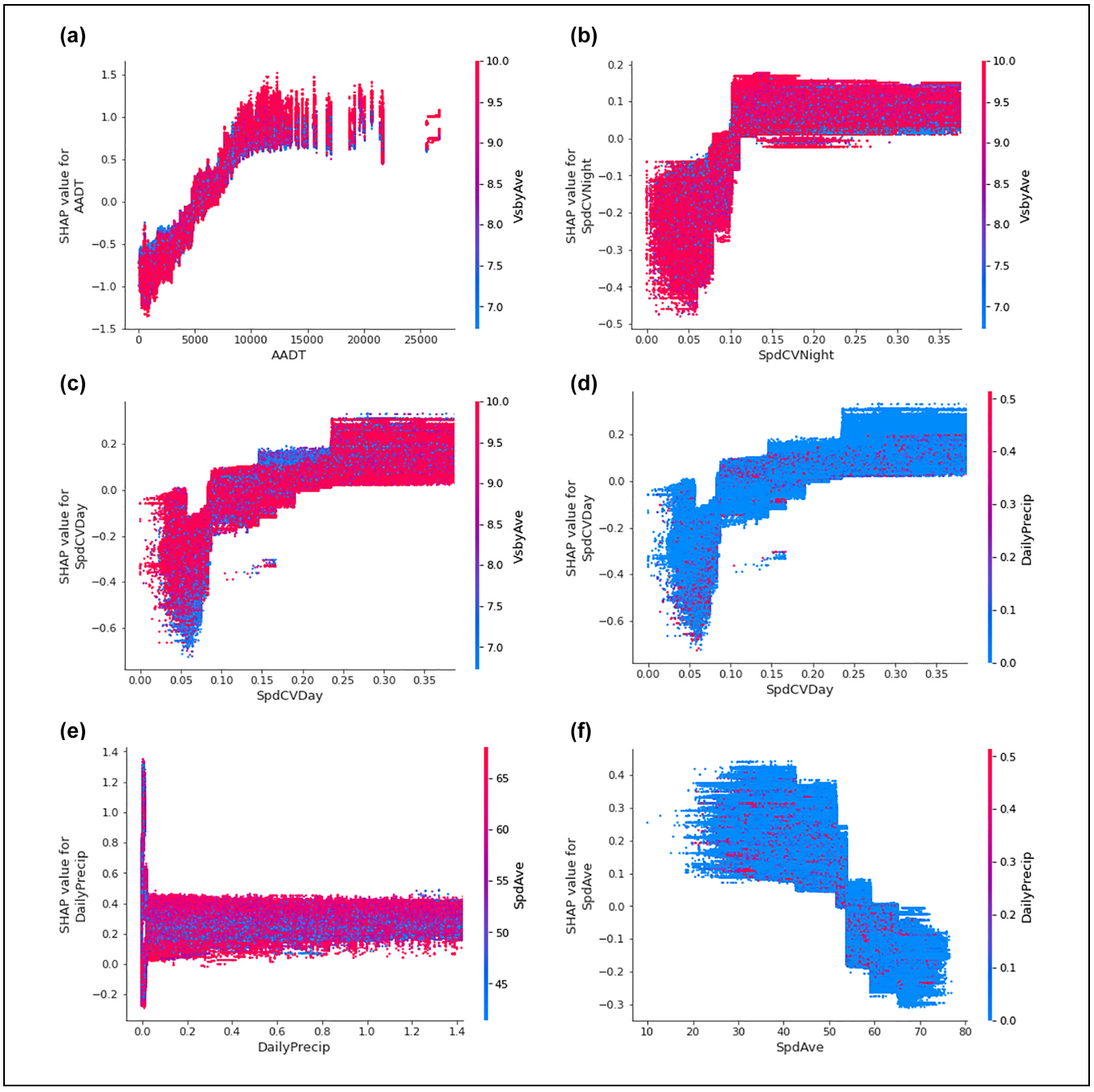

Figure 6. SHapley Additive exPlanation (SHAP) dependence plots (all crash occurrence model): (*a*) AADT and average visibility; (*b*) nighttime speed CV and average visibility; (*c*) daytime speed CV and average visibility; (*d*) daytime speed CV and daily precipitation; (*e*) daily precipitation and average speed; and (*f*) average speed and daily precipitation.
Note: CV = coefficient of variation; SpdCVDay = coefficient of variation of daily speed using data during daytime.

- AADT, average visibility, and daily precipitation are the main influential factors of roadway daily crash occurrence.
- Daily precipitation is ranked highly 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 roadway crashes, but it does not necessarily lead to severe crash occurrences.
- The importance ranking of average visibility stayed unchanged in both the all-crash and severe crash models.
- Lower average visibility and lower visibility standard deviation (i.e., the visibility remains at a low

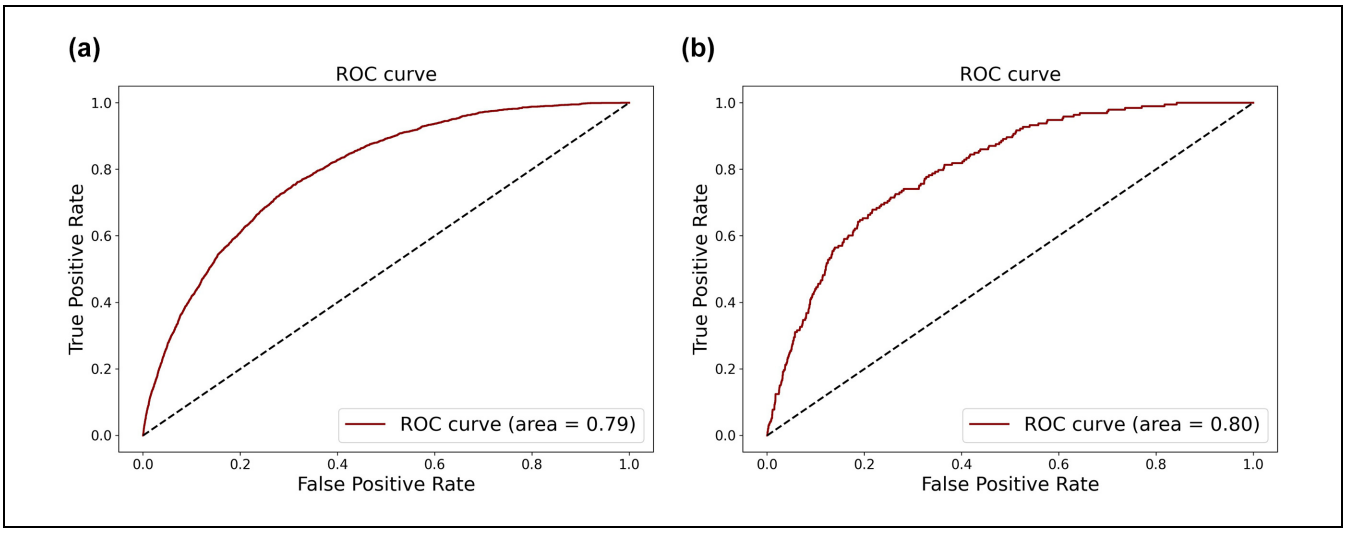

Figure 7. Receiver operating characteristic (ROC) plots for all and severe crash occurrence models: (*a*) ROC curve of all crash occurrence model; and (*b*) ROC curve of severe crash occurrence model.

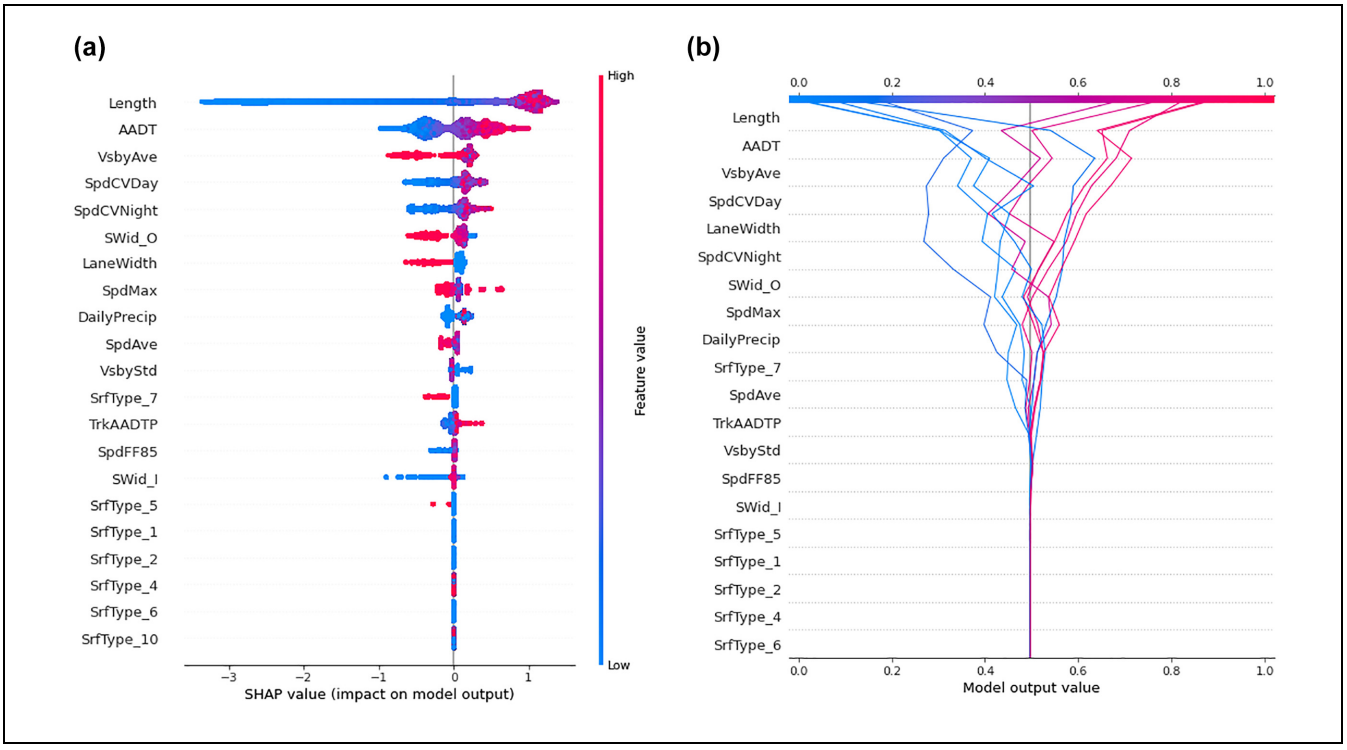

Figure 8. SHapley Additive exPlanation (SHAP) summary plot and decision plot (severe crash occurrence model): (*a*) SHAP summary plot; and (*b*) SHAP decision plot.

level throughout the day) are more likely to lead to higher crash occurrence probability.

- The results show that average operating speed is not positively associated with crash frequencies. However, average operating speed during rainy conditions is positively associated with crash counts.
- When AADT is at lower levels, lower visibility increases the probability of all crash occurrences. When AADT is at higher levels, lower visibility decreases the probability of all crash occurrences.
- The contribution of lane width is more obvious in the severe crash model compared with the

all-crash model, with lower lane width being more likely to cause severe crash occurrence.

For precipitation, many previous studies concluded that higher daily precipitation levels have a positive relationship with crash occurrence (*19*). Looking at 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 increase 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 0 for one day, the crash occurrence likelihood is almost the same, regardless of the precipitation level. As for fatal crash occurrence, the findings of this study show neither a positive 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, for average speeds, previous researchers have reached different conclusions. The finding of this paper is that on rural two-lane highways, an increase in average operating speed is not positively associated with the crash frequencies. This finding echoes that of 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. The 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 outside shoulder width and lane width. Das et al. (*27*) found that shoulder width, pavement width, and AADT are closely related to yearly ROR crash counts on rural two-lane highways. This study echoes their findings. Moreover, we found that outside shoulder width, and lane width make more contribution on the severe crash occurrences compared with all crash occurrences.

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 two-lane highways, future studies can explore daily level modeling using other roadway functional classes such as rural interstate roadways, rural multilane roadways, and urban roadways. A more comprehensive study can be conducted by developing daily level models for all rural and urban roadway types to provide more comprehensive results.

Author Contributions

The authors confirm the contribution to the paper as follows: study conception and design: Zihang Wei; data collection: Zihang Wei, Subasish Das; analysis and interpretation of results: Zihang Wei; draft manuscript preparation: Zihang Wei, Subasish Das, Yunlong Zhang. 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, P. Anastasopoulos, M. G. Karlaftis, F. Mannering, and P. Anastasopoulos. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 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., and B. Saccomanno Hellinga, and Transportation Research Board. 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. Das, S., S. Geedipally, R. Avelar, L. Wu, K. Fitzpatrick, M. Banihashemi, and D. Lord. *Rural Speed Safety Project for USDOT Safety Data Initiative*. FHWA, Washington, D.C., 2020.
23. Das, S., S. Geedipally, K. Fitzpatrick, E. Park, L. Wu, Z. Wei, I. Tsapakis, and S. Paal. *Develop a Real-Time Decision Support Tool for Rural Roadway Safety Improvements*. Texas Department of Transportation, Austin, 2022.
24. Wu, Q., F. Chen, G. Zhang, X. C. Liu, H. Wang, and S. M. Bogus. Mixed Logit Model-Based Driver Injury Severity Investigations in Single- and Multi-Vehicle Crashes on Rural Two-Lane Highways. *Accident Analysis & Prevention*, Vol. 72, 2014, pp. 105–115. https://doi.org/10.1016/j.aap.2014.06.014.
25. Ma, Z., W. Zhao, S. I. -J. Chien, and C. Dong. Exploring Factors Contributing to Crash Injury Severity on Rural Two-Lane Highways. *Journal of Safety Research*, Vol. 55, 2015, pp. 171–176.
26. Persaud, B., C. Lyon, K. Eccles, and J. Soika. Safety Effectiveness of Centerline Plus Shoulder Rumble Strips on Two-Lane Rural Roads. *Journal of Transportation Engineering*, Vol. 142, No. 5, 2016, p. 04016012. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000821.
27. Das, S., X. Sun, and M. Sun. Rule-Based Safety Prediction Models for Rural Two-Lane Run-Off-Road Crashes. *International Journal of Transportation Science and Technology*, Vol. 10, No. 3, 2021, pp. 235–244.
28. Texas Department of Transportation. Roadway Inventory. https://www.txdot.gov/inside-txdot/division/transportation-planning/roadway-inventory.html. Accessed February 23, 2021.
29. 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.
30. Yannis, G., A. A. Theofilatos, A. Ziakopoulos, and A. Chaziris. Investigation of Road Accident Severity and Likelihood in Urban Areas with Real-Time Traffic Data. *Traffic Engineering and Control*, Vol. 55, No. 1, 2014, pp. 31–35.
31. 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.
32. Chen, T., T. He, M. Benesty, and Y. Tang. *Understand Your Dataset with XGBoost*. R Document, 2018.
33. 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.
34. 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.
35. 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.
36. 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.
37. Chen, T., and C. Guestrin. *XGBoost: A Scalable Tree Boosting System*. New York, NY, 2016.
38. Lundberg, S., and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *arXiv Preprint arXiv:1705.07874 [cs, stat]*, 2017.
39. Ma, C., Y. Peng, L. Wu, X. Guo, X. Wang, and X. Kong. Application of Machine Learning Techniques to Predict the Occurrence of Distraction-Affected Crashes With Phone-Use Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2022. 2676: 692–705.