Accident Analysis and Prevention 149 (2021) 105869

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Prediction of vehicle occupants injury at signalized intersections using real-time traffic and signal data

Emmanuel Kidando [a], Angela E. Kitali [b,*], Boniphace Kutela [c], Mahyar Ghorbanzadeh [d], Alican Karaer [d], Mohammadreza Koloushani [d], Ren Moses [d], Eren E. Ozguven [d], Thobias Sando [e]

[a] Mercer University, United States
[b] Department of Civil and Environmental Engineering, Florida International University, 10555 West Flagler Street, EC 3720, Miami, FL, 33174, United States
[c] Texas A&M Transportation Institute, United States
[d] Florida State University, United States
[e] University of North Florida, United States

ARTICLE INFO

Keywords:
Vehicle occupant injury
Real-time data
High-resolution
Event-based data
Random Forest
XGBoost classifiers

ABSTRACT

Intersections are among the most dangerous roadway facilities due to the existence of complex movements of traffic. Most of the previous intersection safety studies are conducted based on static and highly aggregated data such as average daily traffic and crash frequency. The aggregated data may result in unreliable findings because they are based on averages and might not necessarily represent the actual conditions at the time of the crash. This study uses real-time event-based detection records, and crash data to develop predictive models for the vehicle occupants' injury severity. The three-year (2017–2019) data were acquired from the arterial highways in the City of Tallahassee, Florida. Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) classifiers were used to identify the important factors on the vehicle occupants' injury severity prediction. The performance comparison of the two classifiers revealed that the XGBoost has a higher balanced accuracy score than RF. Using the XGBoost classifier, five topmost influential factors on injury prediction were identified. The factors are the manner of the collision, through and right-turn traffic volume, arrival on red for through and right-turn traffic, split failure for through traffic, and delays for through and right-turn traffic. Moreover, the partial dependency plots of the influential variables are presented to reveal their impact on vehicle occupant injury prediction. The knowledge gained from this study will be useful in developing effective proactive countermeasures to mitigate intersection-related crash injuries in real-time.

1. Background

Compared to other roadway facilities, intersections are widely recognized as the most dangerous locations due to the presence of complex conflicting movements, frequent stop-and-go traffic, and road users who disregard traffic controls (FHWA, 2009; Tay, 2015; Yuan and Abdel-Aty, 2018). Statistics in the United States highlight this problem. More than 40 % of all reported crashes and 50 % of fatal and injury crashes in 2019 occurred at intersections (USDOT, 2019; IIHS-HLDI, 2019). As such, transportation officials have devised several strategies to reduce the conflicting movements at intersections by introducing some forms of control, ranging from the stop, yield control to traffic signals.

Despite improved intersection control, the yearly cost of human life lost due to motor vehicle crashes at intersections has not substantially changed, and fatal motor vehicle crashes in urban intersections have increased by 15 % between 2004 and 2018 (FARS, 2020; FHWA, 2004). Moreover, signals at intersections are accompanied by several new safety challenges, including red-light running (RLR) and yellow light confusion (the dilemma zone). The RLR is a situation where a driver enters an intersection after the traffic signal has turned red. Meanwhile, the yellow light confusion describes a situation where some drivers decide to speed up attempting to beat the yellow light, while others decelerate and prepare to stop. With this confusion, a decision to pass through the intersection might result in a right-angle crash, whereas a

* Corresponding author.
E-mail addresses: kidando_ey@mercer.edu (E. Kidando), akita002@fiu.edu (A.E. Kitali), b-kutela@tti.tamu.edu (B. Kutela), mg17x@my.fsu.edu (M. Ghorbanzadeh), ak18k@my.fsu.edu (A. Karaer), mk18h@my.fsu.edu (M. Koloushani), moses@eng.famu.fsu.edu (R. Moses), eozguven@eng.famu.fsu.edu (E.E. Ozguven), t.sando@unf.edu (T. Sando).

https://doi.org/10.1016/j.aap.2020.105869
Received 8 July 2020; Received in revised form 22 October 2020; Accepted 2 November 2020
Available online 16 November 2020

decision to stop might result in a rear-end collision (Zhang et al., 2014). Considering the speeding factor, both manner of collisions tend to be significantly severe. Above all, intelligent transportation systems (ITS) provide a great opportunity to manage traffic at intersections. Proactive measures such as real-time safety applications could be the key to improving safety at signalized intersections, particularly in an urban environment (Kidando et al., 2018).

The advancements in traffic data collections and traffic signal controller capability have paved the way for researchers to analyze the safety of intersections using real-time data for traffic flow and signal timing. One of the recent advances in traffic signal controller capability is the introduction of the high-resolution data-logging capability. These event-based data are collected at high-resolution (10 Hz) with parameters such as detector status, pedestrian status, and phase state of change with their corresponding timestamp in real-time. Because of this capability, several measures of effectiveness for signalized intersection operations have been proposed, in real-time (Day et al., 2014). Examples of such metrics include approach volume, arrivals on red and green, delay, green time allocation, and split failure, among others. Many empirical studies have demonstrated the utility of these metrics (Day et al., 2010; Day and Bullock, 2012; Wu and Liu, 2014). As a result, AASHTO Innovation and USDOT have included the above-mentioned metrics in their daily counts (Kittelson and Associates Inc. and Purdue University, 2017). Besides, the event-based data effortlessly facilitates detector error diagnosis, vehicle classification, queue estimation, and environmental studies.

By providing the variation of traffic flow and signal timing within shorter periods, event-based data also have the potential to be used in the intersection safety analysis. Nonetheless, only a few studies have explored the use of this data in crash risk analysis at intersections (Chen et al., 2017; Yuan et al., 2019). In addition to being used for crash risk analysis, event-based data also have the potential for being used in injury severity analysis. That is, the prevailing traffic conditions and signal parameters before the occurrence of a crash may have a significant impact on the severity of occupants involved in a crash at the intersection.

Thus, the objective of this research is to develop a model for the injury prediction of the vehicle occupant with signal timing parameters and crash attributes. Moreover, a black-box visualization tool based on the partial dependency plots is used to examine the causality effect of the most influential signal timing parameters and crash attributes on crash probabilities of vehicle occupant injury. Data for the years 2017, 2018, and 2019 were acquired from the City of Tallahassee, the Capital of Florida. The novelty of this study is the fact that, to the authors' best knowledge, none of the existing studies have predicted occupants' injury severities using the traffic signal high-resolution event-based data. Thus, there is a need to study the real-time traffic signal event-based factors along with crash attributes on how to influence the injury prediction of vehicle occupants. The identified factors and the prediction model are useful to transportation engineers for developing proactive crash injury countermeasures.

2. Real-time safety analysis of signalized intersections

Following the advancements in data collection technologies, the investigation of road crash likelihood and severity by utilizing real-time data has recently received significant attention from researchers. Real-time traffic data replace the traditional aggregated data, including the annual average daily traffic (AADT). The aggregated data limit the reliability of the findings because they cannot reflect the actual traffic conditions at the time of the crash (Kitali et al., 2018b; Yuan and Abdel-Aty, 2018). Notably, aggregated data does not represent the variation of traffic flow within shorter periods (Essa and Sayed, 2019).

A large number of previous studies has focused on freeways in terms of real-time safety analysis (Kitali et al., 2018b, 2018a; Pande and Abdel-Aty, 2006; Yu and Abdel-Aty, 2014) while few studies have considered arterial highways (Essa and Sayed, 2019, 2018; Mussone et al., 2017; Theofilatos, 2017). The real-time traffic data from these studies were collected using different approaches, including video recording, Bluetooth detectors, and inductive loop detectors. As indicated in Table 1, two of the studies (Essa and Sayed, 2019, 2018) utilized real-time traffic data to develop conflict-based and crash-based safety performance functions, respectively.

The third used real-time traffic data to identify factors associated with crash risk (Yuan and Abdel-Aty, 2018). Real-time traffic data in all of the three studies were found to have a significant role in predicting crash risk.

Unlike freeways, arterial highways also require an additional data source in real-time, i.e., traffic signal data. As such, few other studies explored the use of both real-time traffic and signal data (Chen et al., 2017; Yuan et al., 2020, 2019; Yuan and Abdel-Aty, 2018). The real-time signal data were obtained from signal controllers. The three studies (Yuan et al., 2020, 2019; Yuan and Abdel-Aty, 2018) used real-time data to evaluate factors influencing crash likelihood, while the fourth study focused on the analysis of factors affecting RLR crash frequency. As indicated in Table 1, some real-time traffic and signal variables were found to have a significant influence on the likelihood of intersection crashes and frequency of RLR crashes. In summary, all of the seven studies presented in Table 1, show the need for incorporating real-time traffic and signal data in the intersection safety analysis. Nonetheless, none of these studies explored the use of this data in occupant injury analysis for intersection crashes.

3. Data description

This study used data from two sources: the first source is the Signal Four Analytics database, while the second is the Automated Traffic Signal Performance Measure (ATSPM) system. The Signal Four Analytics, which is a statewide interactive web-based geospatial crash database, developed and hosted by the University of Florida, was used to acquire crash data. On the other hand, the ATSPM system, which provides high-resolution traffic flow and signal timing data, was used to compute signal performance measures. The analysis used 22 intersections (Fig. 1) located along the US 90 located in Tallahassee, which comprises a corridor with approximately 7.7 miles long.

3.1. Crash data

The research team acquired all crashes that occurred along the study corridor between July 2017 (start date of the ATSPM system operations) and May 2019. A 300 feet buffer was used to filter the crashes that occurred within the intersection using ArcGIS software, as described in Fig. 2(a). A similar buffer threshold was also used in previous studies (Abdel-Aty and Wang, 2006; Megat-Johari et al., 2018). Extracting crashes that occurred while a vehicle was approaching an intersection – target crashes in Fig. 2(a) – was done by observing the direction attribute from the crash data.

Additionally, the police report for each of the target crashes was reviewed to identify vehicle occupant information, including the occupant injury severity, age, and gender, along with the vehicle-manufacturing year additional to the crash direction. The final dataset included 344 crashes that involved 918 vehicle occupants. For analysis, each observation in the dataset is a record of the level of injury sustained by each vehicle occupant involved in the crash. Thus, the term "vehicle occupant" herein refers to either drivers or passengers.

3.2. Signal timing and detection events

As indicated earlier, the traffic signal timing, along with the detector events, were retrieved from the city's ATSPM database. Fig. 2(b) shows the approach used to retrieve the data.

As shown in Fig. 2(b), crash timestamp, intersection ID, and direction

Table 1
Summary of studies that used real-time traffic and signal event data to explore the safety of the signalized intersection.

Study	Objective	Study Location	Real-time Data (Source)	Methodology	Key Findings
Yuan et al. (2020)	Crash likelihood	42 intersection approaches in Seminole, Florida	• Signal timing and loop detector data (ATSPM database)	Conditional logistic and binary logistic models	Significant factors associated with cycle-level crash risk at signalized intersections: • Traffic volume • Signal timing • Headway and occupancy • Traffic variation • Shockwave characteristics
(Essa and Sayed, 2019)	Conflict based SPF	6 signalized intersections in Alberta and British Columbia, Canada	• Traffic data (traffic video-data)	Full Bayes Poisson log-normal models	• Highest conflict frequency was noticed at the beginning of the green time • Highest conflict severity was noticed at the beginning of the red time • Traffic conflicts were found to be higher for signal cycles with bigger shock waves, and lower platoon ratios
(Yuan et al., 2019)*	Crash likelihood	44 signalized intersections in Central Florida	• Travel speed data (BLUETOAD detectors) • Signal timing and loop detector data (ATSPM database)	Long short-term memory recurrent neural network algorithm	Significant factors associated with real-time crash risk: • Arrival on yellow (+) • Arrival on green (+)
(Essa and Sayed, 2018)	Signal cycle level SPF	6 signalized intersections in Alberta and British Columbia, Canada	• Traffic data (traffic video-data)	Generalized linear models	Important factors that affect the number of rear-end conflicts at the signal cycle: • shockwave area, maximum queue length, shockwave speed, platoon ratio • Most rear-end conflicts occur during: • the beginning of the red light which represents the dilemma zone • the start of the green time where the stopped-flow starts to discharge gradually at low speed while other vehicles are arriving at higher speeds to the end of the queue
(Yuan and Abdel-Aty, 2018) *	Crash likelihood	23 signalized intersections in Central Florida	• Travel speed data (IterisVelocity Bluetooth detectors) • Signal phasing and 15 min interval traffic volume (adaptive signal controllers)	Bayesian conditional logistic models	Factors that significantly influenced crash risk within the intersection: • Through volume of the traveling approach of the at-fault vehicle (+) • Left-turn volume of the near side crossing approach (+) • Overall average flow ratio of the far-side crossing approach (+) • Increased adaptability for the left-turn signal of the near side crossing approach (-) • More priority for the traveling approach of at-fault vehicle (-) Factors that significantly influenced crash risk at the entrance of the intersection: • Average queue length for through vehicles • Average speed (-) • Average green time and average waiting time for the left-turn phase (-) • Green ratio for the through phase (-) • Higher adaptability for the through phase (-)
(Chen et al., 2017)*	RLR crash frequency	44 signalized intersections in Central Florida	• High-resolution traffic and signal event data (loop detectors)	Non-linear regression model	RLR crashes were more likely to occur: • Large traffic demand and longer signal cycles • 1.5 s after the onset of the red phase
(Mussone et al., 2017)	Crash severity	5 intersections in Minneapolis, Minnesota	• Traffic data (induction-loop traffic sensors)	Back propagation neural network model and a generalized linear mixed model	Real-time traffic flow characteristics were found to have a relevant role in predicting crash severity

Note: ATSPM = Automated traffic signal performance measure; SPF = Safety performance function; RLR = Red light running; (+) = variable increase crash risk; (-) = variable reduce crash risk; * = used both real-time traffic and signal data.

at which a crash occurred on the highway are the attributes that were used in the algorithm to filter the data from the ATSPM database. All recorded intersection-related activities covering 15 min before a crash occurrence were downloaded. Note that ATSPM data at shorter time intervals will contain a large amount of noise (Guo et al., 2018), and previous literature has recommended using a minimum of 15-min measurement intervals to obtain stable traffic flow related data (Smith and Ulmer, 2003).

The ATSPM event dictionary, detector dictionary, and intersection dictionary were used to decode the downloaded events data, and then the computations of different performance measures followed. The next paragraphs discuss how the signal timing and detection events related

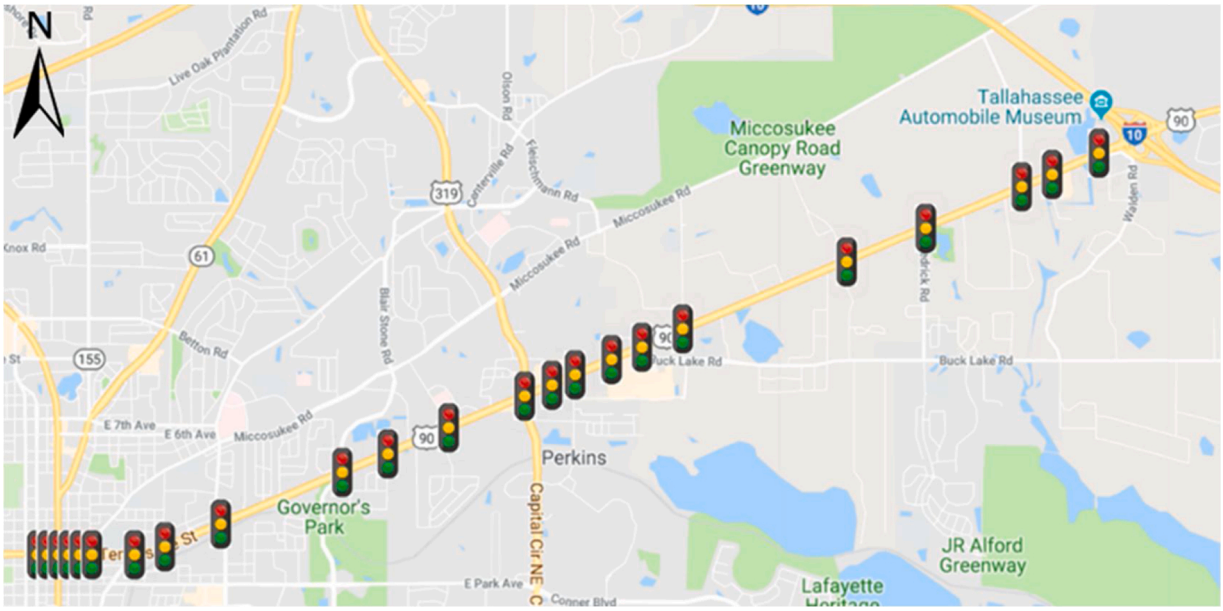


Fig. 1. Study corridor.

variables were computed and incorporated into the safety model.

3.2.1. Approach volume

The approach volume of traffic that arrived 15 min before the crash was estimated using the detector events and the time of crash attributes. An algorithm was written to automate the estimation process for all crashes. More specifically, the estimation of approach volume was implemented by counting the total number of detector actuations during the 15 min before the crash.

3.2.2. Arrivals on green (AOG)

This metric was estimated using the beginning of the green phase, the end of the green phase, the detector event, and the time of a crash. The number of vehicles that arrived during the green signal was estimated by counting the detector actuations during the green interval phase.

3.2.3. Arrivals on red (AOR)

The total number of vehicles arrived during the yellow and red clearance and all red interval was identified as the AOR in this study. Similar to AOG estimation, the detector actuations during the yellow and red time interval were counted, representing the number of arrivals when the traffic signal was yellow or red.

3.2.4. Split failure

It provides a measure on a signalized intersection when the amount of demand for services is too large to be served with a certain green time. In this study, this attribute was computed using the "Max Out" or "Force Off" events. Similar to Mahajan et al. (2019), if the Max Out or Force Off events are identified to occur in any green phase during the 15 min interval, a split failure is alleged to occur in this period of analysis. If a split failure was identified, a value of 1 was recorded. Naturally, a value of 0 was recorded if otherwise. To avoid recording false split failure due to less demand on the approach, the detector event was used as a criterion to filter the true split failure. More specifically, a true split failure to occur, a detector should be actuated preceding the split failure event recorded.

3.2.5. Average approach delay

This measure can also be referred to as a simple approach delay because it does not incorporate either start-up loss time, deceleration, or queue length (Dakic et al., 2018). The approach delay was estimated using the stop bar detector and phase status. It was estimated as a difference between the time when a stop bar detector is first actuated during the red phase and the departure time when the phase turns green.

3.2.6. Average Occupancy ratio (OR)

This metric quantifies phase utilization at an intersection (Day et al., 2014). It is estimated as the percentage of the time a stop bar detector is actuated during the green intervals. This measure was computed using the following formula:

$$OR = \frac{T_{On}}{T_{On} + T_{Off}} \; x \; 100\ \% \tag{1}$$

where, T_{On} and T_{Off} are total durations of the detector were "on" and "off" during the green interval.

3.2.7. Average platoon ratio

This metric quantifies the level of progression between intersections. A higher value of this metric indicates a higher degree of platooning. The current study estimated the platoon ratio using the following Equation:

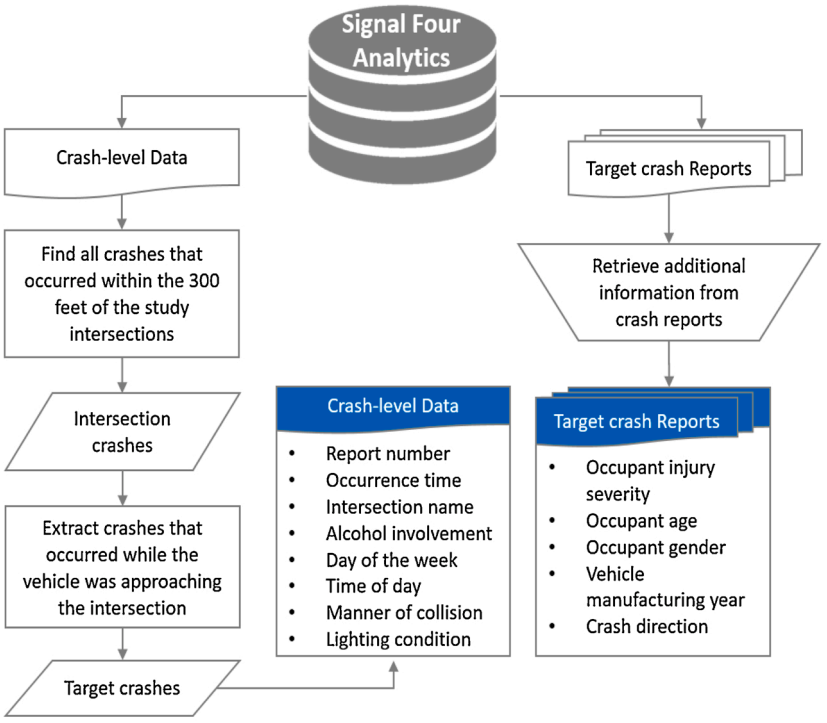
$$Platoon\ ratio = \frac{P}{g/C} \tag{2}$$

where P is the ratio of the number of vehicles arriving when the signal is green to the total volume in a cycle, g is the average green allocation in the cycle in 15 min, and C is the average cycle length in 15 min.

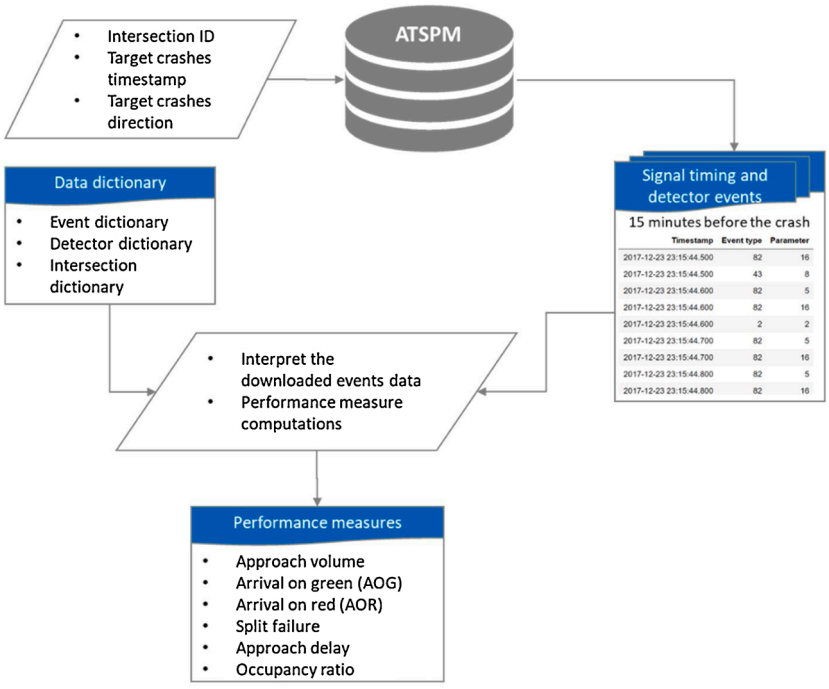
3.3. Descriptive statistics

Crash severity in Florida is categorized into five levels, i.e., fatal injury, incapacitating injury, non-incapacitating injury, possible injury, and PDO. None of the occupants involved in the crashes considered in the study died, 0.44 % of the occupants were incapacitated, 3.59 % sustained non-incapacitating injuries, 9.48 % sustained possible injuries, and the remaining 86.49 % sustained no injury. Considering this distribution, in this study, the data were further grouped into two injury severity levels, i.e., no injury and injury which encompasses all injury cases (i.e., incapacitating, non-incapacitating, or possible injury).

Table 2 describes the categorical variables that were considered in the analysis. As stated earlier, the response variable is the occupant



(a) Signal Four Analytics



(b) ATSPM

Fig. 2. Flow chart indicating the approach used to collect study data.

injury severity, which is divided into two groups: no-injury and injury or fatality. Approximately 14 % of occupants involved in crashes on the study corridor sustained either injury or fatality. Female vehicle occupants are overrepresented in the injury or fatality category accounting for 72 % of all injury/fatality events. Three types of manners of the collision, rear-end, sideswipe, and angle, were included in the analysis. The highest proportion of injured or dead occupants was involved in the angle crashes (34 %) compared to the proportion of uninjured occupants (15 %). Other crash-related explanatory variables included in the study include occupant age, day of the week, time of day, lighting condition, and vehicle manufacturing year.

The variable time of the day was categorized into peak hours and off-peak hours based on the observations of typical traffic conditions in the study corridor. Specifically, the time of the day attribute is categorized

Table 2
Categorical variables definition and data description.

Attribute	Attribute Category	No injury		Injury		Total
		Count	Percent	Count	Percent	
Occupant gender	Male	386	48.61	35	28.23	421
	Female	408	51.39	89	71.77	497
Occupant age	Adult (20−64)	548	69.02	97	78.23	645
	Young (19 or younger)	136	17.13	16	12.90	152
	Senior (65 or older)	110	13.85	11	8.87	121
Day of the week	Weekday	674	84.89	111	89.52	785
	Weekends	120	15.11	13	10.48	133
Time of day	Off-peak	498	62.72	77	62.10	575
	Peak hour	296	37.28	47	37.90	343
Manner of collision	Rear-end	539	67.88	78	62.90	617
	Sideswipe	135	17.00	4	3.23	139
	Angle	120	15.11	42	33.87	162
Lighting condition	Daytime	643	80.98	88	70.97	731
	Nighttime	151	19.02	36	29.03	187
Vehicle manufacturing year	Before 2010	501	63.10	72	58.06	573
	2010 and after	293	36.90	52	41.94	345
Split Failure L	Yes	86	10.83	10	11.11	86
	No	708	89.17	80	88.89	788
Split failure T & R	Yes	263	33.12	30	33.33	293
	No	531	66.88	60	66.67	591
Response variable	Occupant injury severity	794	86.49	124	13.51	918

into peak-hours for the crashes that occurred between 7:00 AM to 9:30 AM, and between 4:00 PM to 6:30 PM. Other crashes are represented in the off-peak hours class.

Vehicles involved in the crashes used in the study were manufactured between the years 1990 and 2019. Most of these vehicles (58.06 %) were manufactured after 2010. Meanwhile, about 40 % of the vehicles were manufactured between 2001 and 2010, and less than 10 % of these vehicles were manufactured between 1990 and 2000. Based on the data distribution, the year 2010 was considered as a threshold for categorizing the vehicle manufacturing year variable.

Regarding the ATSPM performance measures, the right turn and through traffic performance measures were combined due to low traffic on the right turns in most of the intersections within the study corridor. On the other hand, the left turns approach volume was separated. Table 3 shows the descriptive statistics of the ATSPM performance measures estimated for safety analysis. From Table 3, it can be concluded that left-turning volumes are lower than through and right turns combined movements. A similar observation was observed in other estimated metrics, including approach delay, AoG, split failure, occupancy ratio, average green allocation, and yellow and red actuations attributes.

Table 3
ATSPM performance measure definition and data description.

Variable	Mean	SD	Minimum	Maximum
Approach Delay L (Seconds)	6.71	25.29	0	150
Approach Delay T And R (Seconds)	20.00	37.32	0	183
Approach Volume L (Vehicles/15 min)	9.71	5.71	0	40
Approach Volume T And R (Vehicles/15 min)	220.68	133.20	0	574
AoR L (Vehicles/15 min)	4.66	10.70	0	143
Aor T & R (Vehicles/15 min)	18.01	30.34	0	374
OR L (Vehicles/15 min)	0.34	1.69	0	21
OR T & R (Vehicles/15 min)	6.31	13.86	0	70
AoG L (Vehicles/15 min)	13.98	31.66	0	358
AoG T & R (Vehicles/15 min)	88.15	88.37	0	449
Average Platoon Ratio L	0.12	0.26	0	3.12
Average Platoon Ratio T & R	0.28	0.36	0	2.76

Note: L represents left turn and T represents through movement while R represents right turns; SD is the standard deviation.

4. Methodology

In this study, a random forest (RF) and eXtreme Gradient Boosting Model (XGBoost) classifiers were used to study the contributing factors of the vehicle occupant injury prediction at intersections. In this section, the correlation analysis of predictor variables and the principle for classification of the RF and XGBoost algorithms are illustrated. Moreover, hyper-parameter tuning, feature importance, and sensitivity analysis through partial dependence analysis are presented.

4.1. Variable correlation

Before developing the predictive models, a variable correlation analysis was conducted. A Pearson correlation matrix was built to identify and exclude highly correlated variables. A correlation threshold of 0.5 was used to identify highly correlated variables (Kobelo et al., 2008). The results of the correlation analysis are presented in Fig. 3. From the correlation analysis results in this figure, the following variables were omitted in the models because they were found to be highly correlated with other predictors. The ORs were omitted, which were found correlated with split failure; AoGs, which was correlated with the AoRs; average platoon ratios, which were found correlated with the AoRs.

4.2. RF classification algorithm

The RF model is one the powerful, fast in computation, and robust ensemble machine-learning algorithm (Yao et al., 2018), which can be used in either classification or regression analysis. This model has excellent performance even with a small dataset, unlike other machine learning algorithms such as artificial neural networks and support vector machines (Mafi et al., 2018). The RF model combines many weak learners, such as simple decision trees, to improve the outcome of the learning model by aggregating the results from the learned tree predictors (Breiman, 2001; Mafi et al., 2018). In doing so, the RF model controls the overfitting problem, an issue common to most machine-learning algorithms (Mafi et al., 2018). Overfitting generally rises when the developed model has a lower prediction performance to the new dataset that has not been used in the model development compared to training performance.

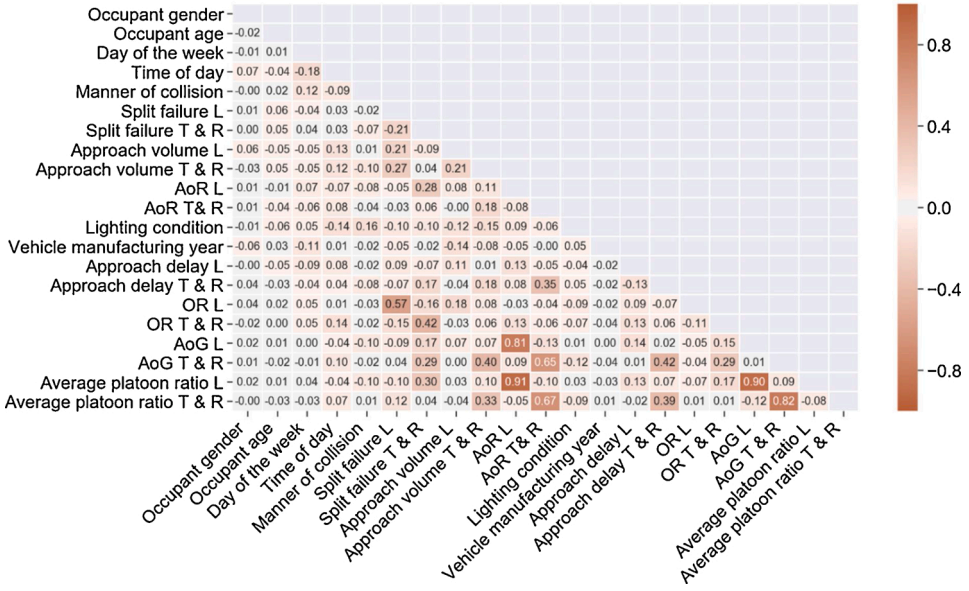


Fig. 3. Variable correlation matrix plot.

4.3. XGBoost classification algorithm

Like RF, the XGBoost classifier also combines multiple weak learners to improve prediction accuracy. On the other hand, the XGBoost is based on gradient boosted decision trees, where each integrated decision tree learns from the previous tree to build a stronger learner in a sequential training process (Zhu and Zhu, 2019). The new classifier tree is then added to the fitted model to update the residuals. The mathematical expression for the XGBoost model with the final tree model, $\widehat{y}_i^{(t)}$ which depends on the previously generated tree model, $\widehat{y}_i^{(t-1)}$ can be presented as follows:

$$\widehat{y}_i^{(t)} = \sum_{k=1}^{t} f_k\,(x_i) = \widehat{y}_i^{(t-1)} + f_t\,(x_i)\, f_k \in t \tag{3}$$

where,

$f_t\,(x_i)$ is the newly generated tree model,
t is the total number of base tree modes, and
x_i are inputs

Table 4
Set of parameters optimized in the XGBoost and RF classifiers.

Model	Parameters	Range of grid	Final Parameters
RF	Number of estimators	10 - 500	240
	Maximum depth of a tree	1 - 32	28
	Maximum features	1 - 20	16
	Criterion	Entropy, Gini	Entropy
	Min samples split	2 - 20	2
	Minimum samples leaf	1 - 10	1
XGBoost	Number of estimators	10 - 500	18
	Maximum depth of a tree	1 - 32	31
	Learning rate	0.01−0.05	0.5
	Minimum sum of instance weight needed in a child	1 - 10	1
	Subsample ratio of the training instances	0.1 - 1	0.782
	Gamma	0.1 - 1	0.198
	Subsample ratio of columns when constructing each tree	0.1 - 1	0.994

4.4. Hyperparameter tuning

The hyperparameters tuning involves searching for the best algorithm by adjusting parameters to optimize prediction accuracy. The hyperparameter tuning for the RF and XGBoost models is an experimental basis (Yu et al. 2017; Dong et al. 2018; Fan et al. 2018). The best parameters are identified through trials of several different combinations. The parameters that yield the best performing model are selected for developing a model that is used in the prediction. The current study used a Bayesian optimization algorithm implemented in the *Optuna* package to search for the best parameters for both RF and XGBoost classifiers (Akiba et al., 2019). The Bayesian optimization algorithm achieves a better performance than random search as it uses past evaluation results to choose the next hyperparameters in the analysis. The objective function in the algorithm minimizes or maximizes a selected metric of the validation dataset.

In hyperparameter tuning, cross-validation with 10-fold was employed, whereby nine sets were used for training while the resulting model was validated with the remaining set of the data. The optimized hyperparameters of the RF and XGBoost classifier, range of searches, and final or best parameter values are presented in Table 4.

Developing predictive models in highway safety, particularly for classification problems, is challenged by the imbalanced nature of outcome classes, whether it is crash injury severity analysis or crash occurrences analysis. The developed classifier with class imbalance outcomes usually is biased in prediction power toward the more frequent class. Many approaches have been proposed to address this challenge in machine learning algorithms. One approach, for example, is the use of re-sampling technique: oversampling of the small class or under-sampling of the large class to re-establish balance on the dataset (Brodersen et al., 2010; Shi et al., 2019). This technique, however, can eliminate useful data in the analysis leading to a biased classifier being created. An alternative to this approach is to use the appropriate misclassification cost, i.e., performance measure. The current study used a balanced accuracy score to reduce biases in the fitted classifier toward the class, with the majority observed outcomes (Brodersen et al., 2010). Mathematically, this metric is the average accuracy of the true negative and true positive predictions. The balanced accuracy score can be estimated from a confusion matrix of predicted versus actual labels of the target variable (see Table 5 and Eq. 4).

Table 5
Confusion matrix for estimating balanced accuracy.

Actual label	Predicted label	
	None vehicle occupant injury	Vehicle occupant injury
None vehicle occupant injury	True Negatives (TN)	False Positives (FP)
Vehicle occupant injury	False Negatives (FN)	True Positives (TP)

$$\text{Balanced accuracy score} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = \frac{1}{2}(sensitivity + \text{Specificity}) \quad (4)$$

To summarize, the following procedures were performed in the analysis:

- Tune the hyperparameters using the 10-cross-validation on entire data using balanced accuracy as an objective score. In this optimization process, the Optuna package was used.
- Randomly split the dataset into training and testing data with a proportion of 0.7/0.3.
- Use retrieved parameters in step one to fit the training set.
- Test the model on the testing set.

4.5. Sensitivity analysis

To understand the association of the most influential predictor variables on the injury prediction of vehicle occupants, the partial dependence plots were prepared. These plots give a graphical depiction of the impact that each predictor has on the target variable. Principally, the partial dependence function estimates the average trend of a given predictor against the probability of vehicle occupant injury occurrence for every value of the predictor variable after accounting for the average impacts of all other input variables in the model (Ding et al., 2018; Saha et al., 2015). In this study, the partial dependencies were calculated using the testing dataset.

The partial dependence plots of categorical influential predictors were centered at a particular point of a variable to provide a clear comparison between two or more classes of the same variable. The centering was done by computing the difference of the average prediction of each group from a base group (Eq. 5).

$$\text{Marginal effect}, ME = \left(\frac{P[x = 1] - P[x = 0]}{P[x = 0]}\right) \times 100\ \% \quad (5)$$

Where, *ME* is the average marginal effect or impact of the categorical predictor variable x on the vehicle occupant probability, and $x = 0$ is set as the base condition. For a predictor variable with more than two groups, e.g., manner of collision and age groups, the marginal effect was computed for each of its groups, where one of its class probability values was set as the base condition.

5. Results

The average balanced accuracy scores of the two implemented predictive models are presented in Fig. 4. This figure portrays the distributions of the balanced accuracy metric derived from 5000 iterations of the parameter tuning process through a cross-validation approach. The majority of the iterations for the XGBoost classifier have a median balanced accuracy 0.63, while RF is 0.54. Based on this observation, it can be concluded that the XGBoost consistently outperforms the RF classifier in most of the iterations.

Besides, the area under the receiver operating characteristic curve (AUC) and confusion matrix (i.e., TP, FP, TN, and FN) were computed of the XGBoost and RF classifier. As presented in Fig. 5, The TP values of the XGBoost and RF classifiers are 18 and 9, respectively. Also, the AUC for XGBoost classifier (AUC = 0.79) is higher than that of RF classifier (AUC = 0.77). A similar conclusion can be made using the AUC and confusion matrix that XGBoost performs well in prediction than the RF classifier with the data at hand.

This result indicates that the XGBoost classifier can discriminate the occupant injury from non-injury with higher accuracy than the RF classifier. Thus, the XGBoost classifier was used for further analysis.

The XGBoost can be interpreted using the relative importance value of each variable in the model to identify the critical factors contributing to injury prediction of vehicle occupants. Two feature importance algorithms are widely used: mean-decrease-impurity based on Gini importance and mean-decrease-accuracy based on permutation importance (Kang and Ryu, 2019). The later algorithm was used in this study because the mean-decrease-impurity is biased, preferring continuous variables and variables with more categories (Hastie et al., 2008).

Also, the impurity-based importance is computed on training set statistics, and therefore, it does not reflect the ability of a fitted model to generalize with a new dataset (Machado et al., 2015). On the other hand, the permutation importance can be used on the testing dataset to estimate the relative rank of the attributes used in the model. Similar to model tuning, the current study used the balanced accuracy scoring in the permutation algorithm. Fig. 6 shows the relative feature importance value of each factor derived from the XGBoost classifier.

A higher value of relative importance indicates stronger influences of a predictor variable on the likelihood of the vehicle occupant sustaining an injury (Ding et al., 2016). The top five ranked predictors that critically affect the vehicle occupant injury probability estimation are the amount of through traffic volume (0.073), the manner of collision (0.063), arrival on red volume for though and right turn traffic (0.046), volume for left-turn traffic (0.036), and vehicle year (0.031). This is followed by the delays for through and right turn traffic (0.022), age attribute (0.016), split failure for left-turn traffic (0.004), arrival on red volume for left-turn traffic (0.011), lighting condition (0.003), and day

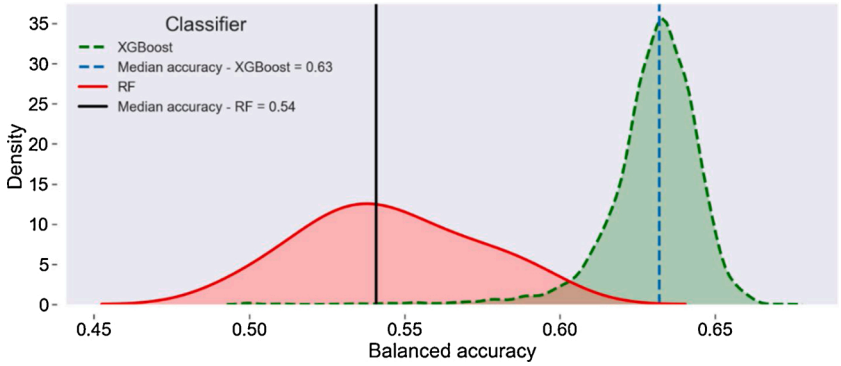
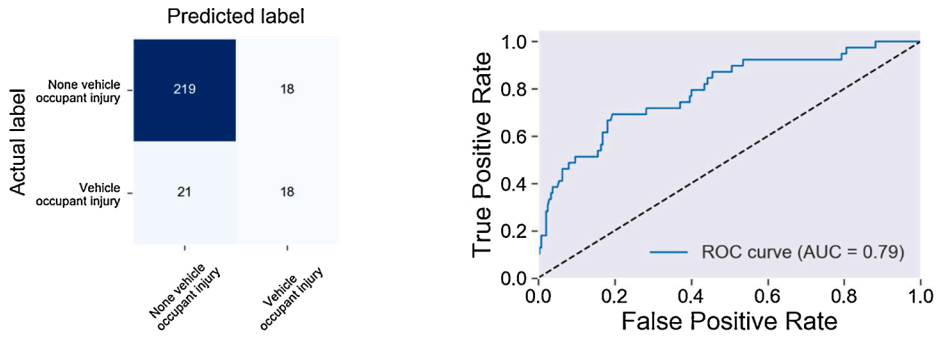
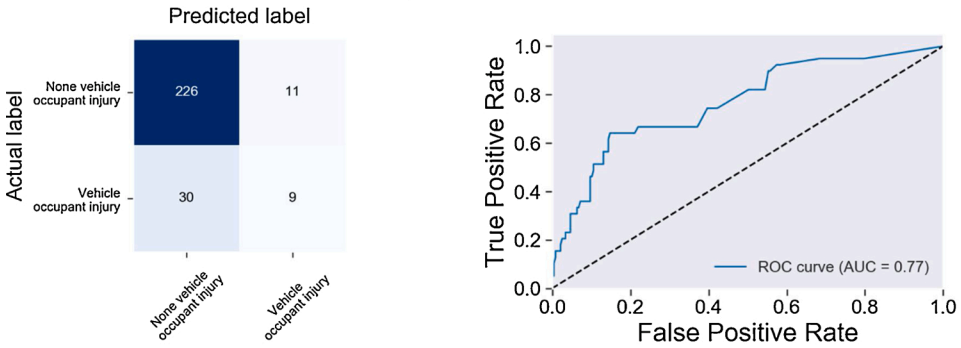


Fig. 4. Distribution of Balanced Accuracy Obtained from Bayesian Optimization Algorithm.



(a) XGBoost classifier



(b) RF classifier

Fig. 5. Confusion Matrix and Receiver Operating Characteristic Curve of Fitted Models.

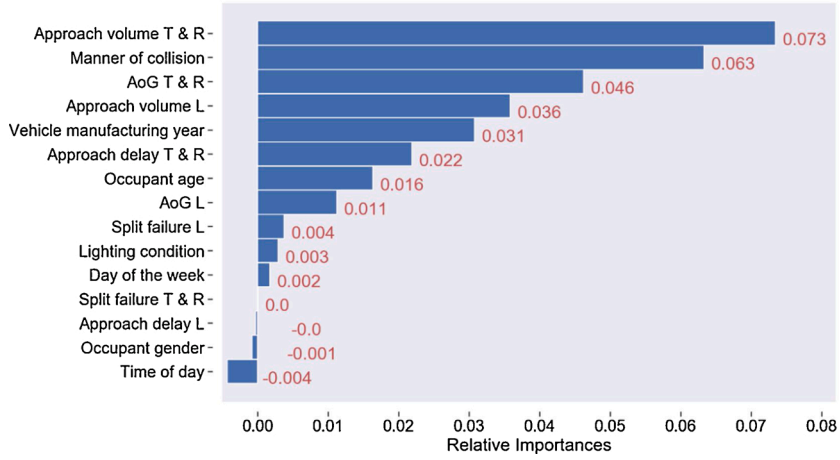


Fig. 6. Relative Importance of Extracted using Permutation Method from XGBoost Classifier.

of the week (0.002), respectively.

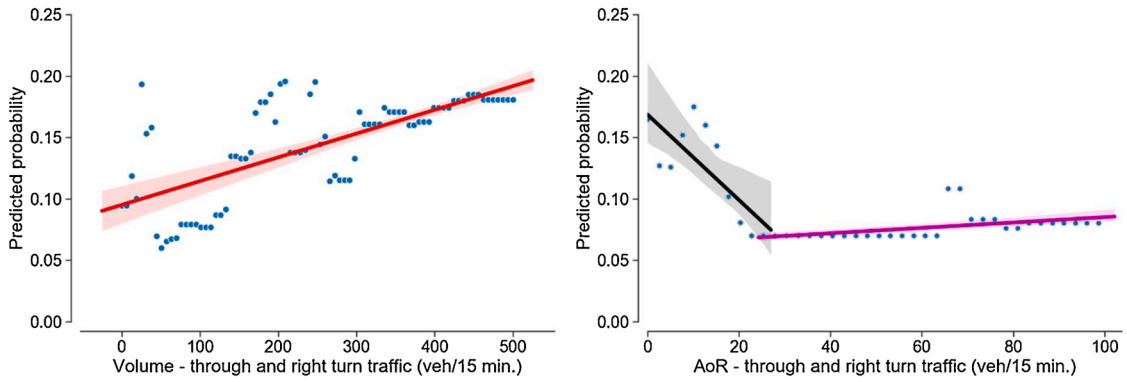
On the other hand, the results from the feature importance analysis indicate that an occurrence of a split failure for right-turn and through traffic (0.000) at an intersection has no role in predictions of vehicle occupant injury. Also, delays for left-turn traffic (-0.001), gender occupant being a male or female (-0.001) and time of day, peak hour versus off-peak hour (-0.004) were identified to reduce the accuracy of the predictions when used in the model. In other words, these three factors are recognized as noise, suggesting that removing from the model improves the accuracy of the fitted model.

The following sections discuss the eleven explanatory variables that were identified, using the feature importance analysis, to influence the prediction of occupant injury severity. The discussion of results is further supported by the results of the marginal effect presented in Figs. 7 and 8. Signal timing and detection events related variables are discussed first, followed by crash-related variables.

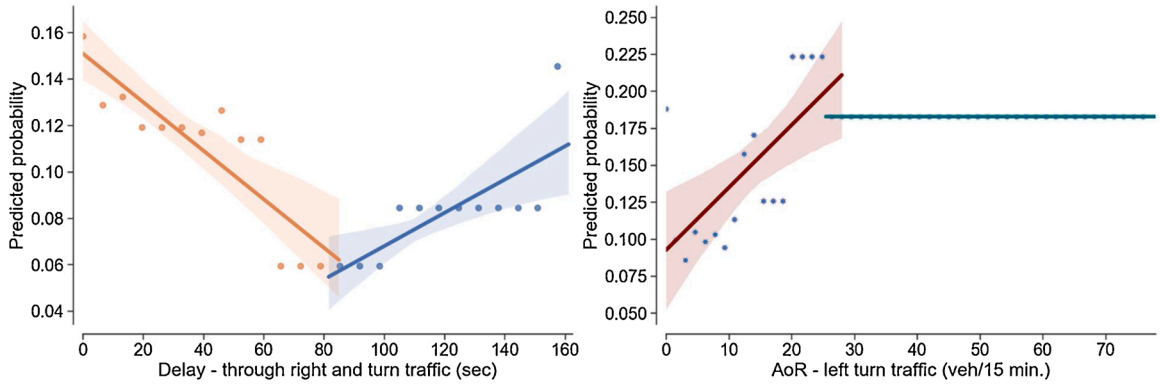
5.1. Signal timing and detection events related variables

Six signal timing and detection events related variables were found to influence the risk prediction of the vehicle occupant sustaining an injury: (1) through/right-turn traffic volume, (2) left-turn traffic volume, (3) arrival on red volume for through and right-turn traffic, (4) arrival on red volume for left-turn traffic, (5) delay for through, and right-turn traffic, and (6) split failure for left-turn traffic.

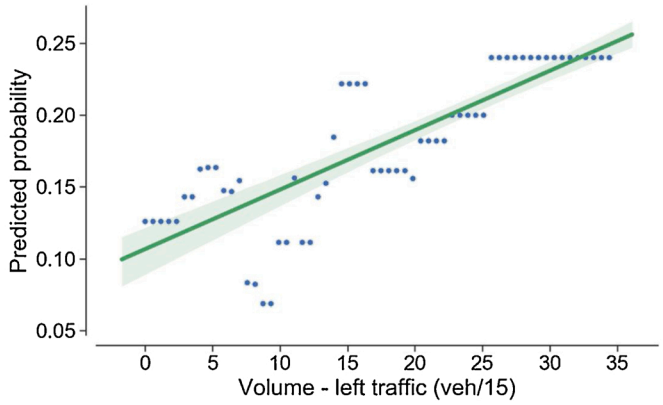
In this study, the number of approaching vehicles that continued through or turned right 15 min before the occurrence of a crash was considered as the through/right-turn traffic volume. As indicated in



(a) Total Volume – through and right turn traffic (b) Total AoR – through and right turn traffic



(c) Delay – through and right turn traffic (d) AoR – left turn traffic



(e) Total Volume – left turn traffic

Fig. 7. The Partial Dependency Plots of continuous predictor variables.

Fig. 7, generally, it can be inferred that the increase in through/right-turn traffic volume is associated with a higher probability of vehicle occupant injury. A similar pattern was found on the left-turn traffic volume where higher left-turning traffic volumes were associated with an increased risk of occupants injury severity (Fig. 7(e)). At the intersection, higher approaching traffic volume may be considered as a precursor for congestion where due to longer waiting times, drivers may be impatient, end up being more aggressive. Longer waiting times may also result in red-light running a tendency which presents a higher risk of occurrence of severe crashes. Chen et al. (2017) also associated large traffic demands and longer signal cycles with red-light running.

The increased number of left-turn vehicles that arrived on red is associated with an increased risk of occupants' injury severity.

The predicted probability of injury severity of the occupants reached 22 % when left-turn volume that arrives on red reaches 25 vehicles per 15 min (Fig. 7(d)). Beyond that volume, the risk of injury severity remains constant. The two scenarios can be explained by the situations that are commonly observed at the signalized intersections. First, the intersections with a low volume of left-turning vehicles that arrive on red are more likely to also have low through traffic volume as well as low

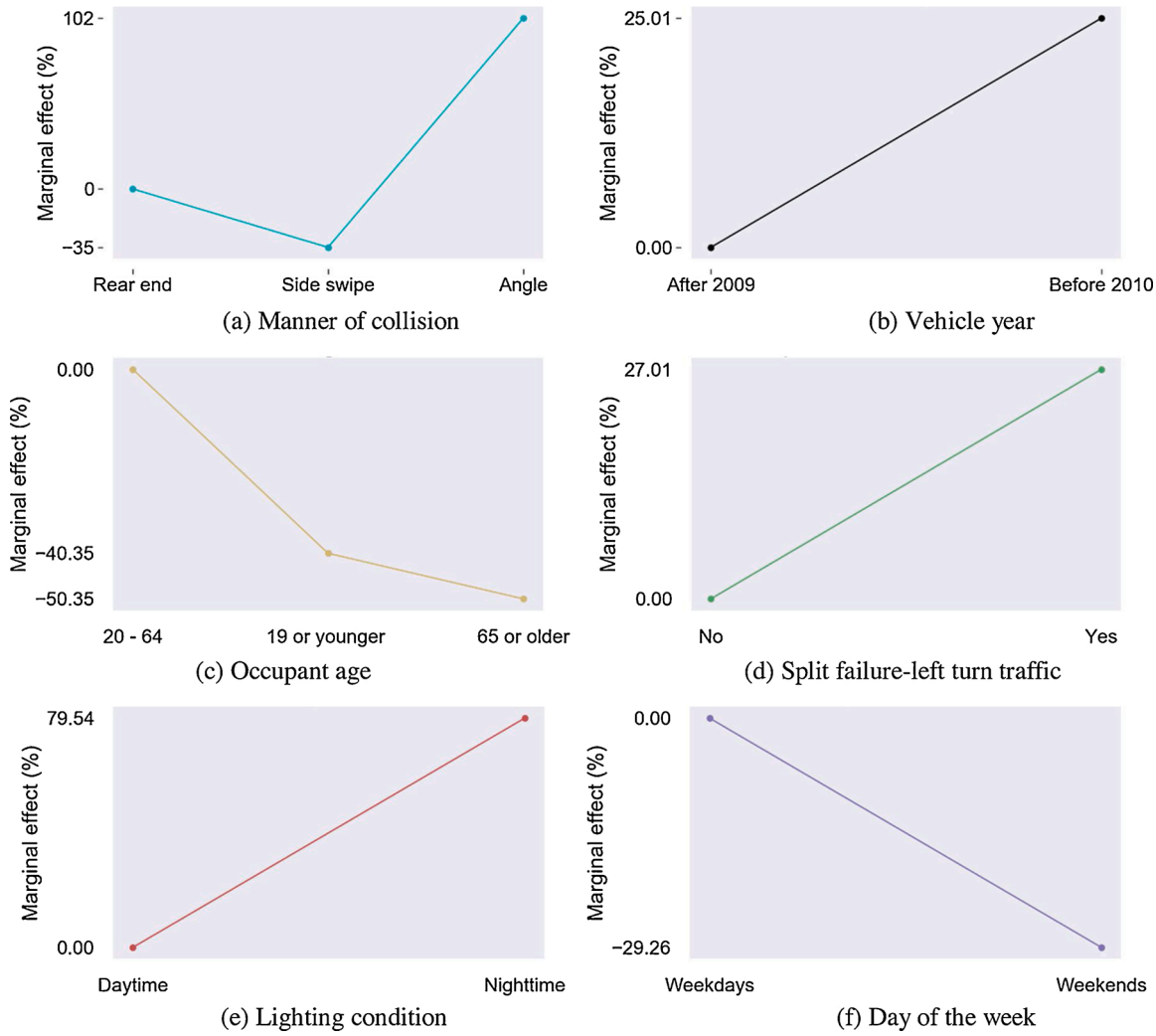


Fig. 8. The Partial Dependency Plots of Categorical Predictor Variables.

speeds, which simply imply more gaps. The presence of gaps motivates the left-turning vehicles that arrive on red to run the red light. Since the vehicle will be turning left, if a crash occurs, it is more likely to be an angle crash, which can have more impact on occupants than on a driver. However, the risk of injury is less likely to be high at low speeds. Second, as the traffic volume increases, the travel speed also increases up to a certain limit. At this point, the attempt to run red light decreases, but the probability of occupant injury increases. Such a scenario explains the constant part of Fig. 7(b).

As revealed in the correlation analysis, the AoR and the average platoon ratio are correlated. These two variables indicate the level of signal progression of traffic between intersections. The higher AoR at intersection reveals a poor signal progression of traffic on the street.

The predicted probability of occupants' injury severity decreases as AoR of the through and right turn vehicles increase up to 25 veh/15 min (Fig. 7(b)). Afterward, the increase of the AoR for the through and right turn vehicles increases the risk of vehicle occupants being injured. The predicted probability drops from 17 % at low volume to 7% at 25 veh/15 min volume. The AoR above 25 veh/15 min, the predicted probability increases gently with an increase in AoR volume. Notably, the increasing pattern in the predicted injury probability of occupants can be explained by the fact that through vehicles are more likely to be involved in angle crashes with the left or crossing vehicle on the side. These types of crashes are more likely to be severe but high volume and low speed reduce the severity of such crashes.

The predicted probability of injury severity is high at the intersections with short delays. It decreases as the delay increases but starts increasing when the delay reaches 80 s. The scenario can be linked to the drivers' behaviors. Longer delays are likely to be a tempting factor for a driver to run a red light, which might result in severe crashes. On the other hand, shorter delay encourages drivers not to run red on light because the traffic signal will change to green light in a few seconds, particularly for commuter drivers who are familiar with the intersections.

Referring to Fig. 8(d), the estimate for the marginal effect of the split failure for the left turn movement predictor reveals that intersections experiencing split failure for the left-turn traffic are more likely (27.01 %) to be associated with injury than those without split failure. Note that, the split failure quantifies the shortage of intersection capacity. The finding of this measure suggests that intersections operating with overcapacity are associated with a higher probability of crash injury than those operating below capacity.

5.2. *Crash-related variables*

Five crash-related variables were found to influence the risk prediction of the vehicle occupant sustaining an injury: (1) manner of collision, (2) vehicle manufacturing year, (3) vehicle occupant age, (4) lighting condition, and (5) day of the week. Three manners of the collision, angle, rear-end, and sideswipe, were included in the analysis. As shown in Fig. 8, the marginal effects of the manner of collision variable were estimated by setting the rear-end crashes as a reference group

in the analysis. The sideswipe crashes (-35 %) are found to be less likely associated with the vehicle occupant injury compared to the rear-end crashes. On the other hand, the angle crash type is found to be 102 % higher to be involved in a severe crash compared to the rear-end crash. This result is consistent with the study that was conducted in Florida, which found that 45 % of the angle crashes at intersections were involved with injury while the rest of the crash types had only 25 % (Abdel-Aty et al., 2009). Several other studies observed a similar finding where angle crashes were identified as the most dangerous crash type in terms of injury severity (Abdel-Aty et al., 2009). On the other hand, rear-end crashes were found to impose the least adverse impacts since rear-end crashes mostly cause vehicle structural damage and often occur during congested periods when vehicles are moving at relatively lower speeds (Huang et al., 2011; Uddin and Huynh, 2018; Zeng et al., 2019, 2017).

Vehicle manufacturing year was included as one of the potential variables that could influence injury severity prediction of vehicle occupants. Specifically, this study assessed the relationship between vehicle manufacturing year to its occupant's injury severity. The variable was divided into two categories, i.e., vehicles manufactured before 2010 and those manufactured in 2010 and later years. As can be inferred from Fig. 8(b), occupants involved in a crash with vehicles manufactured before 2010 were more likely to sustain injuries compared to those involved in a crash with vehicles manufactured in 2010 and later years by 25.01 %. This finding is consistent with a previous study where the severity of vehicle occupants involved in a crash increases with vehicle age (NHTSA, 2018).

Surprisingly, vehicle occupant age revealed contradicting results, especially in the comparison between adult (20–64 years old) and senior adult (65 + years old). The findings suggest that senior adults are 50.35 % less likely to be involved in the crash that they sustain an injury than the adult age group. Similar to senior adults, vehicle occupants who younger than 20-year-old are 40.35 % less likely to sustain an injury than the adult group. This could be attributed to the number of crashes recorded with the adult group occupants, which is four times higher than the younger group or, the older group, as indicated in Table 1. This finding suggests that more data is needed for this variable to have a meaningful model result.

Day of the week (weekday or weekend) on which the crash occurred was found to affect the severity outcome of a vehicle occupant. In particular, the model results of the weekdays versus weekend days reveal that weekdays are associated with a higher likelihood of a person to be involved with a crash injury than weekends. The average marginal effect is estimated to be 29.26 % lower for weekends than weekdays on the injury probability prediction. Similar findings were observed in previous studies (Adebisi et al., 2019; Uddin and Huynh, 2018). One possible reason could be the fact that intersections experience higher volumes of traffic during weekdays, and consequently, a higher chance of intersection crashes to occur.

The study corridor is in the city of Tallahassee, which is among the college-oriented cities with a great number of students enrolled at junior colleges, colleges, universities, and professional schools. As such, more youth drivers during weekdays usually make school/university trips and try to be on-time for their class, a condition that results in more risky maneuvers. Similar results were obtained in previous research (Harbeck and Glendon, 2018; Shinar and Compton, 2004). Shinar and Compton (2004) indicated that drivers were most likely to behave aggressively during the weekday rush hours and least likely to behave aggressively during the weekend.

The lighting conditions were also found to be influential in the injury outcome of occupants. Specifically, nighttime was found to have higher risks associated with the vehicle occupant injury. The corresponding marginal effect of nighttime condition reveals that there is an increasing trend on the probability of a person being injured by 79.54 %. A possible explanation for this finding may be related to the fact that under daylight, most drivers' vision is better, and, eventually, they have more time to perceive and comprehend the road environment and to react correspondingly (Christoforou et al., 2010).

6. Conclusions

Deployment of traffic signals at intersections is one of the common strategies used by transportation agencies to improve the efficiency of an intersection by reducing conflicting movements. One of the most recent developments of traffic signal applications is the Automated Traffic Signal Performance Measure (ATSPM) system. This system provides a detailed way of monitoring traffic flow at intersections in real-time with high-resolution event-based data. Many empirical studies have demonstrated the utility of the ATSPM in the mobility aspect; however, it has have not been well utilized in the traffic safety analysis.

This study attempted to develop a predictive model for vehicle occupant injury using the real-time data acquired from the high-resolution traffic sensors and signal events at intersections. About two years of real-time traffic data were collected to determine the traffic condition 15-min before the crash occurrence. Also, crash-related attributes such as age and gender of occupants were incorporated into the analysis. The random forest (RF) and eXtreme Gradient Boosting Model (XGBoost) classifiers were developed to analyze the data and identify the contributing factors on injury prediction of vehicle occupants. The results of the model comparison revealed that the XGBoost consistently outperformed the RF in the 50,000 iterations used for hyper-parameter tuning. The balanced accuracy score was used as a metric for comparison to account for the imbalanced samples in each class of the response variable, occupant injury vs non-injury. Accordingly, the XGBoost classifier was used to develop a predictive model and analyze the influential variables in predicting the vehicle occupant injury. By ranking based on XGBoost feature importance, 11 predictors out of 15 were identified as mostly influencing the injury severity prediction of vehicle occupants. The top five variables include the manner of the collision, through traffic, arrival on red for though and right turn traffic, split failure for through traffic, and delays for through and right turn traffic.

The partial dependency plots of the influential predictor variables were used to reveal the relationship between the predictor variables and occupant injury. The partial dependency analysis proves that a higher approach volume is associated with a higher risk of the injured occupant in a crash. Moreover, compared to rear-end crashes, angle crashes are 102 % more likely to be associated with occupant injury. On the other hand, sideswipe crashes are less likely (-35 %) to involved with result in severe occupant injury than rear-end crashes. It is important to note that the through/right-turn traffic volume and manner of the collision were identified as the top two leading factors influencing the prediction of injury. The findings from this study are expected to provide insights on the important real-time traffic signal event-based factors influencing the prediction of vehicle occupant injury at intersections. The findings can be used by transportation agencies to link occupant injuries with different traffic signal performance measures from the ATSPM and thus developing effective proactive crash countermeasures.

7. Limitations and future work

It is important to note that though this study substantiated the influence of real-time data on the severity of occupants involved in signalized intersection crashes, there are some limitations. It is well established that the approaching speed of vehicles involved in a crash plays a significant role in the outcome of the occupant injury. However, it was not possible to derive this parameter from the detector data. Thus, future research can investigate the role of approaching real-time speed together with other risk factors on the occupant injury. It will be an opportunity for future work to investigate the impact of time intervals such as 2, 5, and 15 min in the real-time safety analysis. Also, the current study investigated the influence of a downstream traffic signal only on

the occupant injury. The analysis did not consider the effects of the upstream intersection. Further study can incorporate the nearby upstream intersection as well in the analysis.

Several classification metrics are known to address the issue of imbalanced data. The current study only adopted the balanced accuracy metric. It will be an opportunity for future work to evaluate other metrics such as F-score, class-weighted evaluation metrics in the analysis of an imbalanced dataset.

It merits referencing that the underreporting of crashes, particularly for least severe crashes is one of the attributes that may bias the results of models developed using historical crash data. Future work may consider variables that relate to vehicle occupants including seating position and occupant status. Besides, pedestrian compliance particularly at signalized intersections is normally impacted by the prevailing traffic and signal characteristics. Thus, it will be an opportunity for future research to investigate factors affecting pedestrian compliance at signalized intersections based on high-resolution event-based data.

Credit author statement

Emmanuel Kidando, Angela Kitali, and Boniphace Kutela: Conceptualization; Emmanuel Kidando, Angela Kitali, Boniphace Kutela, Mahyar Ghorbanzadeh, Alican Karaer, and Mohammadreza Koloushani: Data Collection; Emmanuel Kidando, Angela Kitali, Boniphace Kutela: Analysis and Interpretation of Results; Emmanuel Kidando, Angela Kitali, Boniphace Kutela, Mahyar Ghorbanzadeh, Alican Karaer, Mohammadreza Koloushani, Ren Moses, Eren E. Ozguven, and Thobias Sando: Draft Manuscript Preparation. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declares no conflict of interest.

Acknowledgments

The authors thank the City of Tallahassee Traffic Management Center for their technical support and providing access to ATSPM data. The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented and do not necessarily, reflect the official views or policies of the sponsoring organizations.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aap.2020.105869.

References

Abdel-Aty, Mohamed, Wang, Xuesong, 2006. Crash Estimation at Signalized Intersections along Corridors: Analyzing Spatial Effect and Identifying Significant Factors. Transp. Res. Rec. 1953 (1), 98–111.

Abdel-Aty, M., Lee, C., Wang, X., Nawathe, P., Keller, J., Kowdla, S., Prasad, H., 2009. Identification of Intersections' Crash profiles/patterns. Tallahassee, Florida.

Adebisi, A., Ma, J., Masaki, J., Sobanjo, J., 2019. Age-related differences in motor-vehicle crash severity in California. Safety 5. https://doi.org/10.3390/safety5030048.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: a next-generation hyperparameter optimization framework. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2623–2631. https://doi.org/10.1145/3292500.3330701.

Breiman, L., 2001. ST4_Method_Random_Forest. Mach. Learn. 45, 5–32.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. Proc. - Int. Conf. Pattern Recognit. 3121–3124.

Chen, P., Yu, G., Wu, X., Ren, Y., Li, Y., 2017. Estimation of red-light running frequency using high-resolution traffic and signal data. Accid. Anal. Prev. 102, 235–247.

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2010. Vehicle occupant injury severity on highways: an empirical investigation. Accid. Anal. Prev. 42, 1606–1620.

Dakic, I., Mladenović, M.N., Stevanovic, A., Zlatkovic, M., 2018. Upgrade evaluation of traffic signal assets: high-resolution performance measurement framework. Promet - Traffic - Traffico. https://doi.org/10.7307/ptt.v30i3.2518.

Day, C.M., Bullock, D.M., 2012. Calibration of platoon dispersion model with high-resolution signal event data. Transp. Res. Rec. 2311, 16–28. https://doi.org/10.3141/2311-02.

Day, C.M., Haseman, R., Premachandra, H., Brennan, T.M., Wasson, J.S., Sturdevant, J. R., Bullock, D.M., 2010. Evaluation of arterial signal coordination: methodologies for visualizing high-resolution event data and measuring travel time. Transp. Res. Rec. 37–49.

Day, C.M., Bullock, D.M., Li, H., Remias, S.M., Hainen, A.M., Freije, R.S., Stevens, A.L., Sturdevant Jr., J.R., T.M.B, 2014. Performance measures for traffic signal systems: an outcome-oriented approach. JTRP Affiliated Reports.

Ding, C., Wu, X., Yu, G., Wang, Y., 2016. A gradient boosting logit model to investigate driver's stop-or-run behavior at signalized intersections using high-resolution traffic data. Transp. Res. Part C Emerg. Technol. 72, 225–238.

Ding, C., Cao, X., Jason, Næss, P., 2018. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. Transp. Res. Part A Policy Pract. 110, 107–117. https://doi.org/10.1016/j.tra.2018.02.009.

Essa, M., Sayed, T., 2018. Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. Transp. Res. Part C Emerg. Technol. 89, 289–302.

Essa, M., Sayed, T., 2019. Full Bayesian conflict-based models for real time safety evaluation of signalized intersections. Accid. Anal. Prev. 129, 367–381.

FARS (Fatality Analysis Reporting System), 2020. 2004-2017 Final File and 2018 Annual Report File (ARF) [WWW Document]. URL https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars (accessed 4.10.20).

FHWA (Federal Highway Administration), 2004. The National Intersection Safety Problem. U.S. Department of Transportation, Washington, D.C.

FHWA (Federal Highway Administration), 2009. The National Intersection Safety Problem (FHWA-SA-10-005). U.S. Department of Transportation, Washington, D.C.

Guo, J., Liu, Z., Huang, W., Wei, Y., Cao, J., 2018. Short-term traffic flow prediction using fuzzy information granulation approach under different time intervals. IET Intell. Transp. Syst. 12 (2), 143–150.

Harbeck, E.L., Glendon, A.I., 2018. Driver prototypes and behavioral willingness: young driver risk perception and reported engagement in risky driving. J. Safety Res. 66, 195–204.

Hastie, T., Tribshirani, R., Friedman, J., 2008. The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics 593–594.

Huang, H., Siddiqui, C., Abdel-Aty, M., 2011. Indexing crash worthiness and crash aggressivity by vehicle type. Accid. Anal. Prev. 43, 1364–1370.

IIHS-HLDI (Insurance Institute for Highway Safety: Highway Loss Data), 2019. Fatality Facts 2018 Urban/rural Comparison [WWW Document]. URL https://www.iihs.org/topics/fatality-statistics/detail/urban-rural-comparison (accessed 4.22.20).

Kang, K., Ryu, H., 2019. Predicting types of occupational accidents at construction sites in Korea using random forest model. Saf. Sci. 120, 226–236.

Kidando, E., Moses, R., Ghorbanzadeh, M., Ozguven, E.E., 2018. Traffic operation and safety analysis on an arterial highway: implications for connected vehicle applications. In: IEEE Conference on Intelligent Transportation Systems, Proceedings. ITSC.

Kitali, A.E., Alluri, P., Sando, T., Haule, H., Kidando, E., Lentz, R., 2018a. Likelihood estimation of secondary crashes using Bayesian complementary log-log model. Accid. Anal. Prev. 119, 58–67. https://doi.org/10.1016/J.AAP.2018.07.003.

Kitali, A.E., Kidando, E., Martz, P., Alluri, P., Sando, T., Moses, R., Lentz, R., 2018b. Evaluating factors influencing the severity of three-plus multiple-vehicle crashes using real-time traffic data. Transp. Res. Rec. 2672, 128–137.

Kittelson & Associates Inc. and Purdue University, 2017. Performance-Based Management of Traffic Signals (NCHRP Project 03-122).

Kobelo, D., Patrangenaru, V., Mussa, R., 2008. Safety analysis of Florida urban limited access highways with special focus on the influence of truck lane restriction policy. J. Transp. Eng. 134, 297–306.

Machado, G., Mendoza, M.R., Corbellini, L.G., 2015. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Vet. Res. 46, 1–15.

Mafi, S., Abdelrazig, Y., Doczy, R., 2018. Analysis of gap acceptance behavior for unprotected right and left turning maneuvers at signalized intersections using data mining methods: a driving simulation approach. Transp. Res. Rec.

Mahajan, D., Banerjee, T., Rangarajan, A., Agarwal, N., Dilmore, J., Posadas, E., Ranka, S., 2019. Analyzing Traffic Signal Performance Measures to Automatically Classify Signalized Intersections. https://doi.org/10.5220/0007714701380147.

Megat-Johari, Megat-Usamah, Bazargani, Bahareh, Kirsch, Trevor J., Barrette, Timothy P., Savolainen, Peter T., 2018. An examination of the safety of signalized intersections in consideration of nearby access points. Transp. Res. Rec. 2672 (17), 11–21. https://doi.org/10.1177/0361198118795997.

Mussone, L., Bassani, M., Masci, P., 2017. Analysis of factors affecting the severity of crashes in urban road intersections. Accid. Anal. Prev. 103, 112–122.

NHTSA (National Highway Traffic Safety Administration), 2018. Passenger Vehicle Occupant Injury Severity by Vehicle Age and Model Year in Fatal Crashes (DOT HS 812 528). U.S. Department of Transportation, Washington, D.C.

Pande, A., Abdel-Aty, M., 2006. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. Transp. Res. Rec. 1953, 31–40. https://doi.org/10.1177/0361198106195300104.

Saha, D., Alluri, P., Gan, A., 2015. Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. Accid. Anal. Prev. 79, 133–144.

Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C., Chai, C., 2019. A feature learning approach based on XGBoost for driving assessment and risk prediction. Accid. Anal. Prev. 129, 170–179. https://doi.org/10.1016/j.aap.2019.05.005.

Shinar, D., Compton, R., 2004. Aggressive driving: an observational study of driver, vehicle, and situational variables. Accid. Anal. Prev. 36 (3), 429–437.

Smith, L.B., Ulmer, M.J., 2003. Freeway traffic flow rate measurement: investigation into impact of measurement time interval. J. Transp. Eng. 129 (3), 223–229.
Tay, R., 2015. A random parameters probit model of urban and rural intersection crashes. Accid. Anal. Prev. 84, 38–40. https://doi.org/10.1016/j.aap.2015.07.013.
Theofilatos, A., 2017. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. J. Safety Res. 61, 9–21.
Uddin, M., Huynh, N., 2018. Factors influencing injury severity of crashes involving HAZMAT trucks. Int. J. Transp. Sci. Technol. 7, 1–9.
USDOT (United States Department of Transportation), 2019. Intersection Safety [WWW Document]. URL https://highways.dot.gov/research/research-programs/safety/intersection-safety (accessed 4.15.20).
Wu, X., Liu, H.X., 2014. Using high-resolution event-based data for traffic modeling and control: an overview. Transp. Res. Part C Emerg. Technol. 42, 28–43.
Yao, Y., Zhao, X., Du, H., Zhang, Y., Rong, J., 2018. Classification of distracted driving based on visual features and behavior data using a random forest method. Transp. Res. Rec. https://doi.org/10.1177/0361198118796963.
Yu, R., Abdel-Aty, M., 2014. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Saf. Sci. 63, 50–56.
Yuan, J., Abdel-Aty, M., 2018. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 119, 274–289.
Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. Transp. Res. Rec. 2673, 314–326.
Yuan, J., Abdel-Aty, M.A., Yue, L., Cai, Q., 2020. Modeling real-time cycle-level crash risk at signalized intersections based on high-resolution event-based data. IEEE trans. Intell. Transp. Syst. 1–16. https://doi.org/10.1109/TITS.2020.2994126.
Zeng, Q., Wen, H., Huang, H., Abdel-Aty, M., 2017. A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments. Accid. Anal. Prev. 100, 37–43. https://doi.org/10.1016/J.AAP.2016.12.023.
Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J., Hao, W., 2019. Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors. Accid. Anal. Prev. 127, 87–95.
Zhang, Y., Fu, C., Hu, L., 2014. Yellow light dilemma zone researches: a review. J. Traffic Transp. Eng. (English Ed. 1, 338–352.
Zhu, S., Zhu, F., 2019. Cycling comfort evaluation with instrumented probe bicycle. Transp. Res. Part A Policy Pract. 129, 217–231. https://doi.org/10.1016/j.tra.2019.08.009.