Transportation Research Interdisciplinary Perspectives 10 (2021) 100378

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

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

A temporal investigation of crash severity factors in worker-involved work zone crashes: Random parameters and machine learning approaches

Seyedmirsajad Mokhtarimousavi [a,*], Jason C. Anderson [b], Mohammed Hadi [c], Atorod Azizinamini [d]

[a] Department of Civil and Environmental Engineering, Florida International University, 10555West Flagler Street ARC 1238, Miami, FL 33174, USA
[b] Department of Civil and Environmental Engineering, Portland State University, 1930 SW 4th Avenue Suite 300, Portland, OR 97201, USA
[c] Department of Civil and Environmental Engineering, Florida International University, 10555 W. Flagler Street EC 3605, Miami, FL 33174, USA
[d] Accelerate Bridge Construction University Transportation Center, Moss School of Construction, Infrastructure and Sustainability, Florida International University, 10555 W. Flagler Street EC 3600, Miami, FL 33174, USA

ARTICLE INFO

Keywords:
Work zone safety
Crash severity
Mixed Logit Model
Support vector machine
Cuckoo search optimization algorithm

ABSTRACT

In the context of work zone safety, worker presence and its impact on crash severity has been less explored. Moreover, there is a lack of research on contributing factors by time-of-day. To accomplish this, first a mixed logit model was used to determine statistically significant crash severity contributing factors and their effects. Significant factors in both models included work-zone-specific characteristics and crash-specific characteristics, where environmental characteristics were only significant in the daytime model. In addition, results from parameter transferability test provided evidence that daytime and nighttime crashes need to be modeled separately. Further, to explore the nonlinear relationship between crash severity levels and time-of-day, as well as compare the effects of variables to that of the logit model and assess prediction performance, a Support Vector Machines (SVM) model trained by Cuckoo Search (CS) algorithm was utilized. Opening the SVM black-box, a variable impact analysis was also performed. In addition to the characteristics identified in the logit models, the SVM models also included the impacts of vehicle-level characteristics. The variable impact analysis illustrated that the termination area of the work zone is most critical for both daytime and nighttime crashes, as this location has the highest increase in severe injury likelihood. In summary, results of this study demonstrate that work zone crashes need to be modeled separately by time-of-day with a high level of confidence. Furthermore, results show that the CS-SVM models provide better prediction performance compared to the SVM and logit models.

1. Introduction

The increasing demand for the renovation and reconstruction of aging transportation infrastructures in the United States have resulted in thousands of roadway construction projects throughout the United States. In the past few years, the number of work zones has increased due to the growth of highway renovations in the State of Florida. Florida is ranked as the top three states with the highest number of work zone crashes, with a total of 67, 73, and 71 fatal crashes resulting in 73, 80, and 76 fatalities from 2015 to 2017 (ARTBA 2018). Another important aspect of work zone crashes is worker safety. Worker safety is a key concern for transportation agencies. In 2017 alone, there were 132 worker fatalities in work zone sites in the U.S. (ARTBA 2018). Among the total number of crashes occurred at work zone locations in 2017 in the State of Florida (i.e., 11,286), around 43.4% were associated with worker presence, in which 16 workers were killed (Bejleri, 2018). The worker fatalities in 2017 are 33.3% and 45.45% higher compared to 2016 and 2015, respectively (ARTBA 2018). The significant loss of workers' lives and injuries resulting from work zone crashes shed light on the emergent need for a better understanding of work zone crash characteristics. This has been highlighted with emphasis in a review paper on work zone safety analysis and modeling (Yang et al. 2015).

The attributes of work zones have significant impacts on the risk of crash occurrence or increasing the severity of crashes in work zones (Garber and Zhao, 2002; Adomah et al., 2021a). Hence, in order to improve work zone safety, it is necessary to investigate the contributing factors involved in work zone crashes. From a logistics perspective, work zone activities can occur during nighttime hours to reduce adverse impacts on traffic operations and complaints by the traveling

* Corresponding author.
E-mail addresses: smokh005@fiu.edu (S. Mokhtarimousavi), jason.c.anderson@pdx.edu (J.C. Anderson), hadim@fiu.edu (M. Hadi), aazizina@fiu.edu (A. Azizinamini).

https://doi.org/10.1016/j.trip.2021.100378
Received 25 December 2020; Revised 9 March 2021; Accepted 29 April 2021

public (Srinivasan et al., 2011; Rahmani, 2018). However, this requires further attention to worker safety due to the more hazardous work conditions at night. Although daytime work zone crashes that involved workers resulted in injuries in Florida in 2016 was higher than the ones that occurred at nighttime (76.32% vs. 23.68% respectively), both day and nighttime crashes shared the same number of fatalities, in which 34 people were killed in total (Bejleri, 2018). Lower traffic volumes during nighttime hours increase driver maneuverability and yield higher operating speeds, which in turn increase safety risks for the construction crew. The visibility of drivers and workers at night is another issue that can greatly affect the relative daytime and nighttime work zone crash risk and severity (Arditi et al., 2007; Li and Bai, 2009; Srinivasan et al., 2011).

Although each has its own limitations and characteristics, different statistical modeling approaches and data mining techniques have been used in recent years to analyze the severity of traffic accidents. A growing number of crash severity studies, including work zone crashes, disaggregate crash records to further understand the impact of various contributing factors. This disaggregation, however, generally results in imbalanced datasets in which the most commonly used statistical models fail to correctly predict rare events (e.g., a severity outcome with a low number of observations) (Longadge and Dongre, 2013). Thus, in addition to identifying significant contributing factors and quantifying their effects, the model's predictive ability should also be evaluated. As it pertains to the prediction aspect, studies which incorporate machine learning models in crash severity analysis tend to overlook the potential improvement of prediction performance through hyperparameter tuning. On account of this, while providing a deeper examination of models' outcomes, this work also examines the application of a metaheuristic algorithm in hyperparameter tuning of a machine learning model for work zone crash severity prediction. Additionally, studies that implement a machine learning approach often do not focus on variable effects, but rather on prediction only. This work uniquely examines both of these aspects.

With that in mind, and considering that most of the existing safety research has focused on the traveling public and not on worker safety, the present study seeks to examine the impact of contributing factors that affect the severity of work zone crashes associated with worker presence by time-of-day. To the best of the authors' knowledge, this is the first attempt at analyzing the severity of work zone crashes associated with workers' presence by time-of-day through discrete choice and supervised machine learning models. As such, the primary contributions and objectives of the current study include

- Detailed investigation of contributing factors that affect work zone crashes involving workers using a mixed logit model and a machine learning approach.
- Investigation of whether daytime and nighttime work zone crashes should be modeled separately for safety analysis by conducting a parameter transferability test.
- Investigation of the capability of metaheuristic optimization in tuning SVM hyper parameters to enhance the model prediction performance.

Although the methodology of the current study was developed in the context of work zone safety, it can be applied to other transportation safety analysis applications.

2. Background

Statistical models are the primary methods for crash severity analysis. Among them, regression models are the most common techniques used to identify the relationship between crash severity and its contributing factors. Different modeling approaches such as binary logit and binary probit models (Haleem and Abdel-Aty, 2010; Mokhtarimousavi et al., 2019, 2020b) or ordered response models (Ye and Lord, 2014; Osman et al., 2016; Ghasemzadeh et al., 2018; Haghighi et al., 2018), have been applied for injury severity analysis. By allowing parameters to vary across observations, random parameter models can address an inherent shortcoming of non-mixed models by taking unobserved heterogeneity into account, which is commonly present in crash data (Chen and Tarko, 2014; Mannering et al., 2016; Anderson and Hernandez, 2017; Bakhshi and Ahmed, 2021). This will ultimately yield more reliable model results and inference when estimating the crash severity contributing factors (Kim et al. 2013). For this reason, a mixed logit modeling framework was used in this research. Moreover, conventional traffic safety analyses should generally not focus on model fit or be concerned with model prediction performance, but should be more focused on the estimation of contributing factors. This sometimes leads to weak prediction results, which are not very reliable (Mujalli and de Oña 2013). Pseudo R-squared measures have often been presented as logistic regression predictive power in many previous crash severity studies such as (Haleem et al. 2015) and (Kim et al. 2010). However, the interpretation of R-squared itself is not a straightforward task, like in linear regression, and the interpretations do not draw a sophisticated inference about model prediction performance, especially when dealing with an imbalanced dataset (Mittlböck and Schemper 1996). To do so, other prediction metrics, which will be discussed later in more details this paper, need to be taken into account.

An important consideration in this study is whether daytime and nighttime work zone crashes need to be modeled separately. Recent safety research has sought to provide a comparison of parameter stability of injury severity analysis among various time periods (Behnood and Mannering, 2015, 2016, 2019; Pahukula et al., 2015; Anderson and Dong, 2017; Mokhtarimousavi et al., 2020a). In previous work zone injury severity modeling and analysis works, time-of-day was typically considered as an indicator variable (Osman et al., 2016, 2018; Mokhtarimousavi et al., 2019). However, the issue of temporal stability, which indicates whether the model estimates are stable across the time periods in work zone crash severity models, has not been addressed. As such, disaggregating work zone crashes by time-of-day (daytime vs. nighttime) can provide additional insight into a better understanding of work zone scheduling. This would assist traffic safety engineers and transportation agencies in establishing a safer work zone and mitigate these types of crashes and their associated costs.

The use of Machine Learning (ML) techniques in safety studies has recently increased (Moghaddam et al., 2011; Li et al., 2012; Weng et al., 2013; Chen et al., 2016; Alkheder et al., 2017; Kitali et al., 2021; Mohammadnazar et al., 2021; Vasebi et al., 2021). SVM is one of the more popular ML approaches, which has shown better prediction outcomes for injury severity analysis compared to conventional statistical models (Li et al., 2012; Iranitalab and Khattak, 2017; Mokhtarimousavi, 2019; Mokhtarimousavi et al., 2019). In this paper, the SVM modeling framework is employed to predict injury severity outcomes of work zone crashes for further investigation of the contributing factors.

A recent demonstration showed that the prediction performance of SVM models can be substantially improved by a proper model parameter tuning (Mokhtarimousavi et al. 2019). Traditionally, non-heuristic algorithms such as grid-search and gradient descent were applied to set SVM parameters (Chapelle et al., 2002; Keerthi, 2002; Wang et al., 2005). These methods, however, are vulnerable to local optimum and cannot guarantee convergence to a global optimum. On the other hand, biologically-inspired metaheuristics, such as the Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Fruit Fly Optimization Algorithm (FOA), etc., are more likely to result in finding the global optimum solution compared to the traditional aforementioned methods (Shen et al., 2016; Taghiyeh and Xu, 2016; Mokhtarimousavi et al., 2018, 2021). Hence, in this study, a new

member of the Swarm Intelligence (SI) algorithms, the Cuckoo Search (CS), is employed to tune SVM models' hyperparameters. The CS was proved to be an efficient algorithm in parameter estimation of complex nonlinear problems and illustrated better and competitive performance as compared to GA and PSO algorithm respectively (Aly 2013). It was also shown that it can handle a tradeoff between problem complexity and the quality of solutions (Aly, 2013; Yang and Deb, 2013). Additionally, the effectiveness and efficiency of the method that combines SVM with CS optimization, which is referred to as CS-SVM in this paper, is rigorously evaluated in terms of classification accuracy, sensitivity, specificity, and AUC (the area under the receiver operating characteristic (ROC) curve) criterion.

The SVM models often attain high quality predictions, but generally work like a black-box in that the obtained models are difficult to interpret. On account of such, after the CS-SVM was trained, a two-stage sensitivity analysis was employed to measure the impacts of the contributing factors on crash severity outcomes from a statistical perspective.

As observed, there are inherent gaps in the work zone safety literature. Therefore, the current study seeks to uniquely fill these gaps through specific data disaggregation (daytime and nighttime crashes) and provide specific methodological ideas embraced by the comparison of traditional econometric techniques versus enhanced SVM models to contribute to the body of traffic safety knowledge.

3. Empirical setting and data

In this study, a three-year period of statewide crash data, including a total of 2,112,783 crash records, was collected from the Signal Four crash database from January 1, 2015 to December 31, 2017 (Bejleri 2018). Then, work zone crashes which consisted of 1.55% of the total crash records were extracted. Finally, 37.48% of the total number of work zone crashes were associated with the worker presence, which was considered for the analysis.

After data cleaning, a total of 12,042 crash records were included in the severity models, which consisted of 64 fatal crashes, 3476 injury crashes, and 8,502 no-injury crashes. The combining of the low frequency of fatal crashes with injury crashes resulted in a severity level called "Fatality/Injury." Property damage only (PDO) was the other considered severity level in this study and marked as "No Injury." According to the average times for sunset and sunrise conditions for the State of Florida (Timeanddate, 2019), two time periods, from 6:00 to 19:59 and 20:00 to 05:59, were considered the daytime and nighttime conditions. The frequency of potential crash contributing factors by severity levels is illustrated in Table 1.

4. Methodology

This section will detail the methodologies used throughout the study, including the mixed logit model, parameter transferability, SVM, and the CS algorithm used to tune the SVM parameters. Fig. 1 illustrates the methodology used in the current work.

The premise behind the application of the logit model and the corresponding parameter transferability test is two-fold. First, the logit model is used to determine significant factors that contribute to work zone crash severity. Second, the parameter transferability test statistically confirms whether the contributing factors to work zone crash severity are different by daytime and nighttime conditions. The identified variables in the logit model are then investigated in detail using the enhanced SVM approach.

4.1. Binary mixed logit

Mixed logit models (MXL) were developed in this study to estimate the statistically significant contributing factors involved in work zone crash severity. The MXL is perhaps one of the most popular and widely used econometric models for crash severity analysis, which can take unobserved heterogeneity into account when estimating the relationship between explanatory variables and crash severity outcomes (Haleem et al., 2015; Behnood and Mannering, 2017; Seraneeprakarn et al., 2017; Mokhtarimousavi et al., 2019). It extends the standard fixed parameter logistic regression model by allowing parameters to be randomly distributed across observations, which in fact enhances the model's estimate reliability (Washington et al., 2010; Cerwick et al., 2014). As an example, although it may be possible to estimate various crash characteristics and environmental characteristics based on crash data, there are several data items that influence crash occurrence and severity that are difficult to collect and are normally not available. The application of a random parameters model attempts to account for this unobservable heterogeneity which, if disregarded, can result in biased parameter estimates (i.e., estimations are not true representations of the population parameters).

The binary mixed logit modeling framework (BMXL) utilized in this study is derived from the traditional binary logit model, as shown in in Eq. (1):

$$P_n(i) = \frac{e^{(\beta_i X_{in})}}{1 + e^{(\beta_i X_{in})}} \tag{1}$$

where $P_n(i)$ is the probability of crash n resulting in crash severity i, $\boldsymbol{X_{in}}$ is a vector of observable crash-related factors (i.e., variables shown in Table 1), and $\boldsymbol{\beta_i}$ is a vector of estimable parameters corresponding to crash severity i.

To estimate random parameters, a mixing distribution is introduced to Equation (1), such that (McFadden and Train, 2000; Train, 2009; Washington et al., 2010):

$$P_n(i|\Omega) = \int \frac{e^{(\beta_i X_{in})}}{1 + e^{(\beta_i X_{in})}} f(\beta|\Omega) d\beta \tag{2}$$

where $P_n(i|\Omega)$ is the mixed logit probability (weighted average of the MNL probabilities). The weights used to determine the probability are determined by density function $f(\beta|\Omega)$ (the density function of β). The density function takes on a distribution as defined for distributional parameter Ω, where both a mean and variance are estimated. This distribution is defined by the analyst and, for the current work, is defined to normally distributed. If the estimated variance is statistically significant, the parameter is random and accounts for crash-specific variation due to unobservables (Mannering and Bhat, 2014; Mannering et al., 2016). For this work, Ω was specified to be normally distributed. Under this formulation, parameter estimates can account for crash-specific variations on crash severity.

Parameters from the logit model are not readily interpretable; therefore, measures must be taken to interpret the effects of independent variables on crash severity. For this work, marginal effects are used to determine these effects based on a change in the probability of an observed crash severity outcome (this change may be an increase or decrease in probability). This interpretation is based on a one-unit change in an independent variable, while all other variables remain equal to their means. With all variables in this work being indicators, marginal effects are determined as (Greene 2018):

$$ME_{X_{ink}}^{P_n(i)} = \text{Prob}\left[P_n(i) = 1 | X_{(X_{ink})}, X_{ink} = 1\right] - \text{Prob}\left[P_n(i) = 1 | X_{(X_{ink})}, X_{ink} = 0\right] \tag{3}$$

4.2. Parameter transferability

The next step is to determine if daytime and nighttime crashes need to be analyzed independently. To determine if such steps need to be taken, this work utilizes a parameter transferability test. The premise behind the parameter transferability test, in this context, is to determine if the estimated parameters in work zone crash severity models

Table 1
Data description.

Variable Description*	Variable* Name	Crash Severity Levels						Total
		Fatality		Injury		No Injury		
		Percent	Freq.	Percent	Freq.	Percent	Freq.	
Crash Severity	SEV	0.53%	64	28.87%	3476	70.60%	8502	12,042
Crash-Level Variables								
Crash Time	TOD							
Daytime	DAYT	56.25%	36	74.97%	2606	76.89%	6537	9179
Nighttime	NIGHTT	43.75%	28	25.03%	870	23.11%	1965	2863
Crash Type	CRSHTYP							
Backed Into	CRSHTBI	1.56%	1	0.66%	23	2.40%	204	228
Left Entering	CRSHTLE	4.69%	3	5.58%	194	2.62%	223	420
Left-Rear	CRSHTLR	1.56%	1	1.78%	62	0.91%	77	140
Off Road	CRSHTOR	10.94%	7	7.77%	270	8.07%	686	963
Parked Vehicle	CRSHTPV	7.81%	5	2.93%	102	4.54%	386	493
Pedestrian	CRSHTPDS	20.31%	13	2.42%	84	0.15%	13	110
Rear-End	CRSHTRE	32.81%	21	61.25%	2129	49.59%	4216	6366
Right Angle	CRSHTRA	4.69%	3	3.57%	124	2.38%	202	329
Rollover	CRSHTROLO	1.56%	1	2.19%	76	0.62%	53	130
Same Direction Sideswipe	CRSHTSDS	3.13%	2	6.39%	222	20.17%	1715	1939
Single Vehicle	CRSHTSV	10.94%	7	5.47%	190	8.55%	727	924
Road Surface Condition	RDSURF							
Dry	RDSURDR	93.75%	60	90.22%	3136	90.41%	7687	10,883
Wet	RDSURWT	6.25%	4	9.78%	340	9.59%	815	1159
Weather Condition	WETHR							
Clear	WTHRCLR	79.69%	51	74.31%	2583	74.82%	6361	8995
Cloudy	WTHRCLD	15.63%	10	20.40%	709	19.44%	1653	2372
Rain	WTHRRIN	4.69%	3	5.29%	184	5.74%	488	675
Road Sys Identifier	RDWTYP							
County	RDWTCNT	7.81%	5	10.33%	359	10.41%	885	1249
Interstate	RDWTINTS	40.63%	26	36.39%	1265	37.36%	3176	4467
Local	RDWTLOC	12.50%	8	10.70%	372	13.57%	1154	1534
State	RDWTST	28.13%	18	29.80%	1036	27.44%	2333	3387
Turnpike/Toll	RDWTTRNT	4.69%	3	3.42%	119	3.93%	334	456
U.S.	RDWTUS	6.25%	4	9.35%	325	7.29%	620	949
Number of Vehicle Involved	NOVINV							
Single Vehicle	NOVINVS	60.94%	39	83.86%	2915	83.75%	7120	10,074
Multi Vehicle	NOVINVM	39.06%	25	16.14%	561	16.25%	1382	1968
Vehicle-Level Variables								
Number of Passengers	NUMPAS							
Driver Only	NUMPSDO	60.94%	39	55.93%	1944	68.27%	5804	7787
Single Occupant	NUMPSSO	7.81%	5	19.79%	688	12.22%	1039	1732
Multi Occupant	NUMPSMO	31.25%	20	24.28%	844	19.51%	1659	2523
Alcohol Related	ALCH							
Yes		89.06%	57	96.14%	3342	98.32%	8359	11,758
No		10.94%	7	3.86%	134	1.68%	143	284
Distraction Related	DISTRL							
Yes		89.06%	57	79.49%	2763	82.73%	7034	9854
No		10.94%	7	20.51%	713	17.27%	1468	2188
Work Zone Variables								
Type of Work Zone	WZTYP							
Intermittent or Moving Work	WZTIMW	7.81%	5	5.41%	188	4.70%	400	593
Lane Closure	WZTLCL	29.69%	19	35.53%	1235	37.07%	3152	4406
Lane Shift/Crossover	WZTLSHC	1.56%	1	8.72%	303	10.23%	870	1174
Work on Shoulder or Median	WZTSHLM	60.94%	39	50.35%	1750	47.99%	4080	5869
Crash Location in Work Zone	WZLOC							
Activity Area	WZLACA	90.63%	58	70.11%	2437	68.71%	5842	8337
Advance Warning Area	WZLADWA	4.69%	3	9.67%	336	8.75%	744	1083
Before the First Work Zone Warning Sign	WZLBFWS	1.56%	1	4.09%	142	4.76%	405	548
Termination Area	WZLCTRA	0%	0	1.99%	69	1.75%	149	218
Transition Area	WZLTRA	3.13%	2	14.15%	492	16.02%	1362	1856
Law Enforcement in Work Zone	LAWINF							
Yes		84.38%	54	81.33%	2827	79.22%	6735	9616
No		15.63%	10	18.67%	649	20.78%	1767	2426

are significantly different between daytime and nighttime conditions; hence, their effects. Parameter transferability consists of a log-likelihood test, as shown in Eq. (4) (Washington et al. 2010):

$$x^2 = -2\left[LL\left(\beta_{MX1_{MX2}}\right) - LL\left(\beta_{MX1}\right)\right] \quad (4)$$

where $LL\left(\beta_{MX1_{MX2}}\right)$ is the log-likelihood at convergence of model MX_1 based on using time-period data for model MX_2, and $LL\left(\beta_{MX1}\right)$ is the log-likelihood at convergence of model MX_1. To provide an example, the model for daytime crashes is fit using the data for nighttime crashes, and the nighttime model is fit using the data for daytime crashes. The resulting log-likelihood values are $LL\left(\beta_{MX1_{MX2}}\right)$. Considering the degrees of freedom (the number of estimated parameters in the model using the other model's data), significance is determined through the use of Eq. (4). To be specific, this log-likelihood ratio test tests the null hypothesis that daytime and nighttime crashes should be

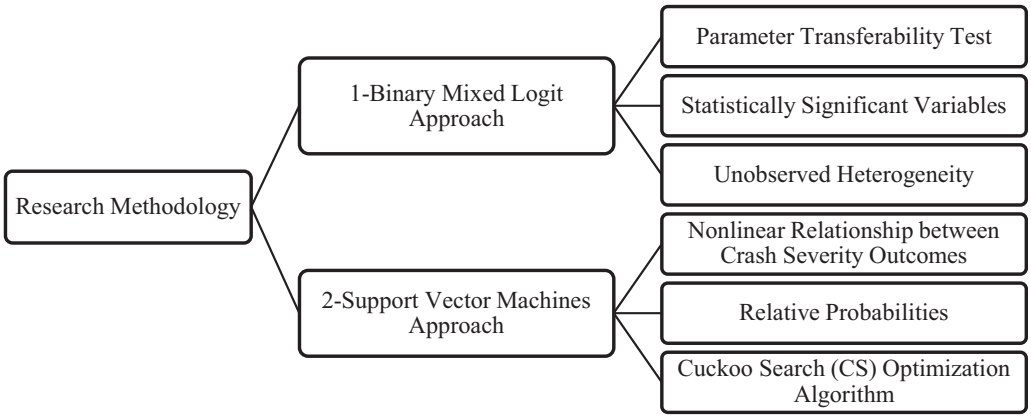

Fig. 1. Research Methodological Flowchart.

modeled together and that their contributing factors, or parameter estimates, are not statistically different. On this premise, this work aims to determine if this hypothesis is rejected.

4.3. Support vector machine (SVM)

Upon determination of significant contributing factors through the logit model and examination of the results from the parameter transferability test, the SVM model is applied to capture crash severity patterns among all explanatory variables. SVM is a non-parametric supervised learning classification model introduced and developed in the 1990s (Boser et al., 1992; Vapnik, 1998). The SVM algorithm is originally designed for binary classification and aims to find (n-1) dimensional separating hyperplanes (one hyperplane in binary classification problems) while simultaneously maximizing the distances of the nearest data points to the decision boundary (i.e., the margin). The hyperplane defines a decision boundary as a set of points, $\boldsymbol{x}$, as illustrated in Equation (5).

$$y(\boldsymbol{x}) = \boldsymbol{w}.\boldsymbol{x} + b = 0 \tag{5}$$

where $\boldsymbol{w}$ (a normal vector of weights) and b (bias) are determined and optimized through the learning process on a training set (x_1, y_1) to (x_n, y_n) with the criterion to maximize the distances of the nearest data points to the decision boundary (i.e., margin). To reach the optimal separating hyperplane, given a training set of explanatory variables and severity outcomes pairs (x_i, y_i), the SVM algorithm solves the quadratic optimization problem shown in Eq. (6) (Bottou and Lin 2007):

$$\begin{aligned} &\min Q(w, b, \xi) = \tfrac{1}{2}\|w\|^2 C\textstyle\sum_{i=1}^{n}\xi_i + \\ &\text{Subjectto}, \forall_i y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \end{aligned} \tag{6}$$

where ϕ is a feature vector, C controls margin violations as a penalty variable, and the misclassification errors is measured by parameter slack ξ. Ultimately, the SVM contains a subset of points of the two classes (crash severity outcomes) called support vectors. Along with the support vectors are a corresponding set of weights $\boldsymbol{w}$ (one for each feature), also called alpha, on an optimal hyperplane in which the distance to the origin is determined by parameter bias. Furthermore, transformation into a higher-dimensional space for data which are not linearly separable in the original space is implemented by introducing the following kernel function: $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$. Among the proposed kernel functions, the Gaussian Radial Basis Function (RBF) showed its capability to capture the non-linear relationship between the response and explanatory variables and is one of the more commonly used kernel functions. It has demonstrated better results in related works of crash severity analysis (Yu and Abdel-Aty, 2014; Mokhtarimousavi et al., 2019) and thus was used in this study. The RBF kernel is defined as:

$$K_{Gaussian}(\boldsymbol{x}_i \times \boldsymbol{x}_j) = exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{7}$$

In Eq. (7), the parameter σ which controls the width of the Gaussian is set to be 0.4. Since the prediction performance of SVM in safety analysis can be significantly enhanced by tuning the model parameters (Mokhtarimousavi et al. 2019), the CS, a powerful metaheuristic algorithm for global optimization, was employed to tune the SVM parameters. A critical SVM parameter is b, the bias term. This term allows the SVM to pass the origin in order to determine a separating hyperplane with the maximum margin. Without bias, the SVM will always go through the origin of the feature space. Another critical parameter is alpha, which forms the hyperplane, and the last parameter that will be tuned using the CS-SVM is the number of support vectors.

4.4. Cuckoo Search (CS) optimization algorithm

The final step, as stated previously, is the application of the CS algorithm. The CS algorithm is a swarm-intelligence based optimization algorithm developed by Yang and Deb (2009), and in the case of the current work, was used to tune the SVM parameters. This nature-inspired metaheuristic mimics the breeding behavior of a specific bird family called "cuckoo." In order to understand how the algorithm is inspired by cuckoo's unique breeding behavior to find a global optimal solution in optimization problems, two concepts need to be explained. These two concepts will be discussed in the following subsections.

4.4.1. The Cuckoo's reproduction strategy

The cuckoo follows a unique reproduction system called "brood parasitism." This strategy makes them dependent on other birds to hatch their eggs. The female cuckoo tries to find the nest of another species that recently laid eggs so that it will lay and hide its own eggs. If the eggs are identified by the host bird, they may either be thrown away or the host bird will abandon the nest and make new ones (Li and Yin 2013).

4.4.2. Lévy flights

Lévy flights is basically a random forward-step technique for finding a new nest or food. A cuckoo's random walk to explore surrounding areas near the current location is derived from the Lévy distribution, which is a transition probability with an infinite variance and mean. Such behavior was applied to different optimization algorithms, and the results demonstrated its superiority and capability over other distributions, specifically in CS (Yang 2010).

When generating new solutions, or in other words, choosing a new random nest, a Lévy flight is performed, as follows:

$$X_i^{t+1} = X_i^t + \alpha \oplus Lévy\ (\lambda) \tag{8}$$

$$Lévy(\lambda)\ u = t^{-\lambda} \tag{9}$$

where X_i^{t+1} is a new solution, α is the step size, the product ⊕ denotes entry-wise multiplications, and lastly, the Lévy (λ) distribution is shown in Equation (9) (Yang and Deb 2009).

The procedure of CS algorithms to find global optimum solutions is based on three main rules, as described below (Yang and Deb 2009):

- Each cuckoo dumps eggs on a nest that is randomly selected.
- Nests with the highest quality eggs (i.e., solutions) will be kept in the model for the next generation.
- The laid eggs in a fixed number of available host nests can be discovered with probability $p \in [0, 1]$.

5. Model estimation results

5.1. Binary mixed logit model

Separate models were generated for worker-involved work zone crashes. The estimation results of mixed logit models for the daytime and nighttime periods are presented in Table 2 and Table 3, respectively. The results include the estimation of corresponding marginal effects, while the random parameters were selected based on the statistical significance of the estimated standard deviation under a normal distribution. In addition, the marginal effects were used to illustrate the change in the probability of crash severity due to a one-unit increase in the explanatory variables.

5.2. Parameter transferability test results

In regard to examining parameter transferability across time periods (daytime vs. nighttime), applying Equation (4) results in x^2 values of 6867.2 and 7350.8, with 20 and 15 degrees of freedom for MX_1 and MX_2, respectively. The obtained chi-square statistics indicate that the null hypothesis that daytime and nighttime work zone crashes involved workers needs to be modeled together for safety analysis, and can be rejected with well over 99% confidence. This indicates that a single model, including daytime and nighttime crashes for the given data, is not appropriate, and the parameter estimates are statistically different. It has been well-documented in a number of recent traffic safety studies demonstrating that instability exists in models estimated for different time periods when analyzing injury-severity (Behnood and Mannering, 2015; Anderson and Dong, 2017; Mokhtarimousavi et al., 2020a).

5.3. SVM results

In this study, the RBF kernel function was utilized to develop SVM models that were coded in the MATLAB R2018b programming environment. In order to assess the prediction performance of the classification models, the whole daytime and nighttime datasets were randomly separated into three training and testing sets with ratios of 6:4, 7:3, and 8:2, respectively. As shown in the confusion matrices in Fig. 2, the preliminary performance test results reveal that SVM models with the split of 70% for training and 30% for testing result in better prediction performance in both daytime and nighttime models. This split ratio was therefore selected for further model prediction

Table 2
Daytime Mixed Logit Model Specifications and Marginal Effects.

Variable	Coefficient	Std. Error	*t*-statistic	Marginal Effects
Constant	−0.410	0.093	−4.41	
(Std. Dev. Of the Random Parameter)	*(0.442)*	*(0.030)*	*(14.50)*	
Work Zone Characteristics				
Law Enforcement Indicator	−0.219	0.082	−2.67	−0.041
Work Zone Type: Intermittent/Moving Work	0.295	0.099	2.98	0.056
Work Zone Type: Work on Shoulder/Median	0.112	0.046	2.43	0.021
Crash Characteristics				
Crash Type: Rear-End	0.032	0.096	0.34	0.006
(Std. Dev. Of the Random Parameter)	*(1.842)*	*(0.057)*	*(31.92)*	
Crash Type: Pedestrian Related	2.572	0.253	10.16	0.490
Crash Type: Same Direction Sideswipe	−0.853	0.110	−7.78	−0.162
Crash Type: Left Entering	0.879	0.127	6.94	0.168
Crash Type: Rollover	1.311	0.174	7.52	0.250
Crash Type: Single Vehicle	−0.178	0.104	−1.72	−0.034
Crash Type: Backed-Into	−1.00	0.222	−4.51	−0.190
Crash Type: Left-Rear	0.763	0.181	4.20	0.145
Crash Type: Right Angle	0.510	0.136	3.75	0.097
Alcohol Related Indicator	0.890	0.217	4.10	0.170
Distraction Related Indicator	0.154	0.053	2.87	0.029
Number of Vehicle Involved: Multiple Vehicles	−0.330	0.114	−2.90	−0.062
(Std. Dev. Of the Random Parameter)	*(1.039)*	*(0.038)*	*(27.33)*	
Number of Passengers: Multi Occupant	0.219	0.067	3.26	0.042
Number of Passengers: Driver Only	−0.567	0.052	−10.86	−0.108
(Std. Dev. Of the Random Parameter)	*(0.932)*	*(0.042)*	*(21.96)*	
Road Sys Identifier: Local	−0.363	0.065	−5.61	−0.069
Environmental Characteristics				
Weather Condition: Rainy	−0.353	0.100	−3.54	−0.067
(Std. Dev. Of the Random Parameter)	*(1.521)*	*(0.150)*	*(10.15)*	
Model Summary*				
Number of Observations	9,179			
Log-Likelihood at Zero	−5509.21			
Log-Likelihood at Convergence	−5108.09			
Overall Prediction Accuracy	62.45%			
Sensitivity	34.78%			
Specificity	73.64%			
AUC	0.7014			

*Analysis of Binary Choice Model Predictions Based on Threshold = 0.5000

Table 3
Nighttime Mixed Logit Model Specifications and Marginal Effects.

Variable	Coefficient	Std. Error	*t*-statistic	Marginal Effect
Constant	−0.726	0.120	−6.04	
(Std. Dev. Of the Random Parameter)	*(0.377)*	*(0.046)*	*(8.24)*	
Work Zone Characteristics				
Law Enforcement Indicator	−0.303	0.067	−4.51	−0.067
Work Zone Location: Advance Warning Area	0.179	0.094	1.90	0.040
Work Zone Type: Lane Shift or Crossover	−0.355	0.124	−2.86	−0.079
Work Zone Type: Work on Shoulder or Median	0.173	0.088	1.94	0.038
Crash Characteristics				
Number of Vehicle Involved: Multi Vehicle	−0.672	0.152	−4.41	−0.150
(Std. Dev. Of the Random Parameter)	*(0.113)*	*(0.049)*	*(2.31)*	
Crash Type: Left Entering	1.250	0.288	4.34	0.278
(Std. Dev. Of the Random Parameter)	*(2.586)*	*(0.655)*	*(3.95)*	
Crash Type: Left-Rear	1.156	0.115	10.01	0.258
Crash Type: Pedestrian Related	3.138	0.764	4.10	0.698
Crash Type: Right Angle	1.340	0.247	5.41	0.298
Crash Type: Parked Vehicle	0.626	0.181	3.45	0.140
Crash Type: Rear-End	1.156	0.115	10.01	0.258
Crash Type: Single Vehicle	−0.745	0.177	−4.19	−0.166
(Std. Dev. Of the Random Parameter)	*(1.261)*	*(0.197)*	*(6.41)*	
Alcohol Related Indicator	0.729	0.131	5.54	0.162
(Std. Dev. Of the Random Parameter)	*(1.726)*	*(0.241)*	*(7.17)*	
Number of Passengers: Single Occupant	0.170	0.076	2.22	0.038
Model Summary*				
Number of Observations	2,863			
Log-Likelihood at Zero	−1780.77			
Log-Likelihood at Convergence	−1623.55			
Overall Prediction Accuracy	61.37%			
Sensitivity	38.42%			
Specificity	71.86%			
AUC	0.671			

*Analysis of Binary Choice Model Predictions Based on Threshold = 0.5000

performance improvement through the application of CS metaheuristic optimization in parameter tuning.

Considering the fact that, like any other metaheuristic algorithms, the performance of CS is considerably impacted by the proper set of its parameters, a Taguchi method was used to obtain the optimal combination of parameters that have the most principal influence on the algorithm performance (for detailed information regarding the Taguchi method, readers are referred to (Peace 1993)). In performing the Taguchi test, a number of 1000 iterations, population size of 100, step size equal to 0.1 and discovery rate equal to 0.6 were utilized when using CS algorithms.

The following criteria were considered to assess the effectiveness of the proposed CS-SVM models:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{10}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \tag{12}$$

In the above equations, TP, FN, TN, and FP refer to counts for true positive, which represents the crash records with a 'Fatality/Injury' level of crash severity and are correctly classified as Fatality/Injury crashes; false negative represents crash records with a 'No Injury' severity level, which is mistakenly classified as Fatality/Injury; true negative represents the 'No Injury' crashes, which are correctly classified as No Injury crashes, and false positive represents crash records with a 'Fatality/Injury' severity level, which are mistakenly classified as No Injury. In addition, the area under the receiver operating characteristic curve (AUC) criterion, one of the popular criteria for evaluating binary classifiers, is calculated as proposed in (Bradley 1997). AUC reflects the model performance in avoiding false classification. The detailed classification results of the final CS-SVM model are included in Table 4.

5.4. Variable impact analysis

In order to measure the impacts of contributing factors from the probability distribution of work zone crash severities, a two-stage sensitivity analysis was conducted. This method was recently adopted in SVM safety studies to quantify and analyze the association between severity outcomes and crash explanatory variables (Li et al., 2012; Yu and Abdel-Aty, 2013, 2014; Chen et al., 2016). In this method, a contribution of each explanatory variable is extracted by replacing the original value with a user-defined value (the same value is used for all input variables). Then, the deviation of the corresponding probabilities of the severity outcomes (No Injury and Fatality/Injury in this study) before and after these changes are calculated and recorded for CS-SVM models. The results are shown in Tables 5 and 6 for daytime and nighttime models, respectively.

6. Discussion

Out of the total 39 indicator variables, the impact of 33 variables were found statistically significant throughout the daytime and nighttime mixed logit models. Among significant variables, eight are common in both models, in which one has heterogeneous effects in both: the indicator for a multi-vehicle crash. In the daytime model, the indicators for rear-end crashes, rainy weather, and vehicles that had no passengers were found to be normally distributed random parameters. In the nighttime model, the indicators for left-entering crashes, single vehicle crashes, and alcohol consumption were found to have heterogeneous effects.

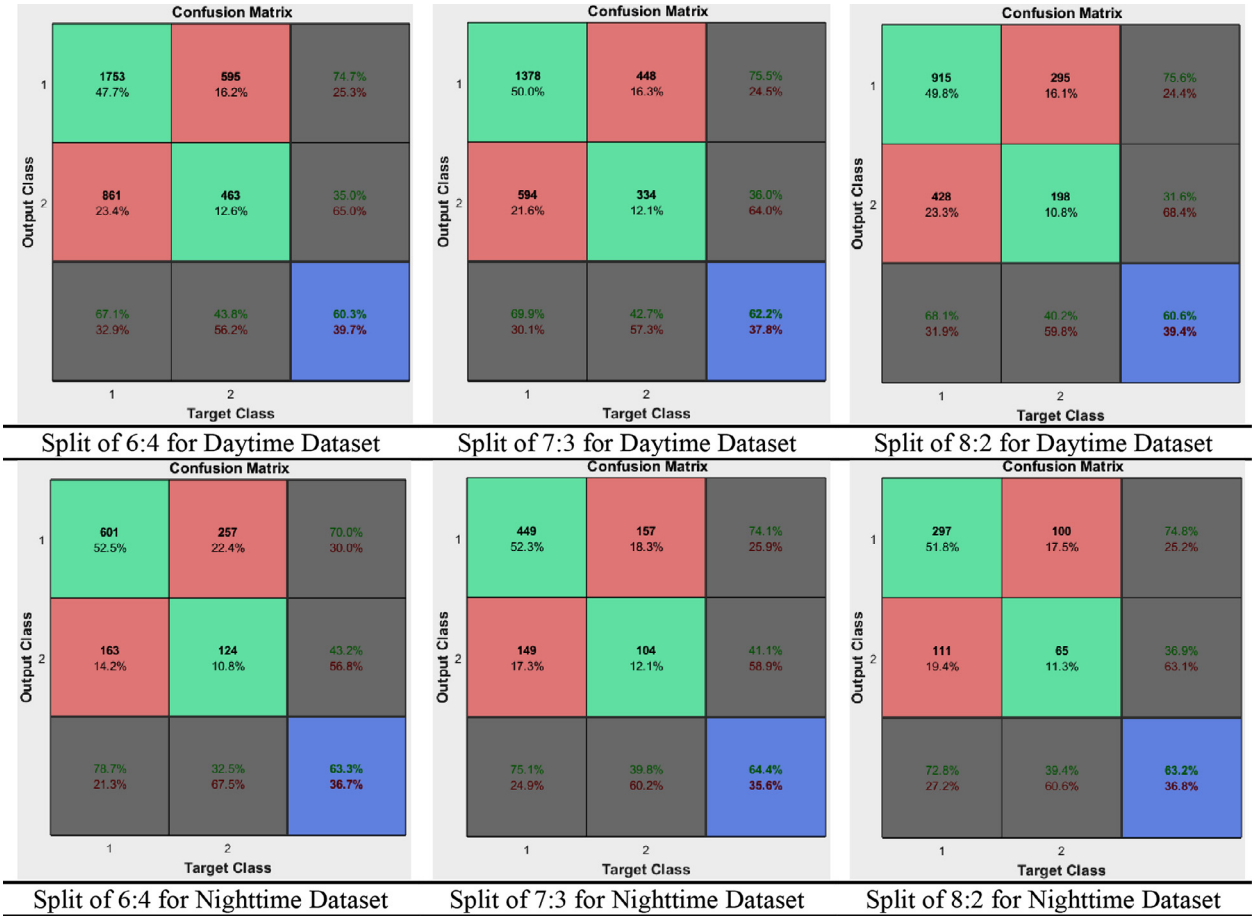

Fig. 2. SVM Confusion Matrices of Different Data Split.

Table 4
Results of CS-SVM Models.

CS-SVM	Confusion Matrices			Accuracy	Sensitivity	Specificity	AUC
Daytime Model		*No Injury*	*Injury*	84.00%	98.71%	48.71%	0.7371
	No Injury	1917 (69.6%)	416 (15.1%)				
	Injury	25 (0.9%)	395 (14.3%)				
Nighttime Model		*No Injury*	*Injury*	89.40%	97.32%	71.37%	0.8434
	No Injury	580 (67.6%)	75 (8.7%)				
	Injury	16 (1.9%)	187 (21.8%)				

The following section discusses each of the impacts of contributing factors according to the daytime and nighttime models, while comparison of the results from the two models are provided.

6.1. Daytime crash severity models

6.1.1. Crash characteristics

As for random parameters, the indicator for rear-end crashes is a normally distributed estimated random parameter. Model estimates indicate that with a mean of 0.032 and standard deviation of 1.842, 50.69% of rear-end crashes (greater than zero) are more likely to result in a fatal/injury crash, and 49.31% (less than zero) are less likely. As stated by (Wang et al. 1996), rear-end crashes increased significantly in work zone locations. In addition, rear-end crashes were found to be the prominent crash type in work zones (Srinivasan et al. 2007). Sudden stops, following too closely while drivers are distracted due to cell phone use, and distraction with worker presence or work zone equipment, are all factors more likely to be the reported cause for rear-end crashes (Osman et al. 2018). Therefore, the significance of this variable in a work zone context is anticipated. As for the heterogeneous effects on crash severity, a potential reason may stem from the differences in speed limits, driver compliance, and location where the crash occurred. Specifically, work zones often have a lower speed limit (i.e., speed drop). Subsequently, if a rear-end crash occurs at lower speeds, a rear-end crash in which no injury is sustained can be expected. However, if drivers are distracted at the start of the work zone or do not comply with the lower speed limit, rear-end crashes will occur at higher speeds, resulting in an increased likelihood of a more severe crash involving injuries.

Results from other crash-type-related variables found to have statistically significant impacts on crash severity in the daytime model are pedestrian-related crashes, rollover crashes, and left-entering crashes. Each of these, according to marginal effects, have the greatest impact on observing a fatal/injury crash. Specifically, pedestrian-related crashes have a 0.49 higher probability of resulting in an injury, rollover crashes have a 0.25 higher probability, and left-entering crashes have a 0.17 higher probability.

The abovementioned results are consistent with the CS-SVM (referred to as SVM for the remainder of this section) daytime output. Based on the variable impact analysis of the SVM model, it was found

Table 5
CS-SVM Daytime Variable Impact Analysis.

Variable	Severity		Variable	Severity	
	No Injury	Fatality/Injury		No Injury	Fatality/Injury
Crash-Level Variables			Dry	0.862	0.138
Crash Type			Wet	0.806	0.194
Backed Into	0.829	0.171	**Vehicle-Level Variables**		
Left Entering	0.777	0.223	*Number of Passengers*		
Left-Rear	0.785	0.215	Driver Only	0.883	0.117
Off-Road	0.792	0.208	Single Occupant	0.821	0.179
Parked Vehicle	0.827	0.173	Multi Occupant	0.776	0.224
Pedestrian	0.767	0.233	*Alcohol Related*		
Rear-End	0.849	0.151	Yes	0.785	0.215
Right Angle	0.798	0.202	No	0.867	0.133
Rollover	0.771	0.229	*Distraction Related*		
Same Direction Sideswipe	0.863	0.137	Yes	0.818	0.182
Single Vehicle	0.819	0.181	No	0.867	0.133
Weather Condition			**Work Zone Variables**		
Clear	0.857	0.143	*Type of Work Zone*		
Cloudy	0.825	0.175	Intermittent or Moving Work	0.787	0.213
Rain	0.808	0.192	Lane Closure	0.838	0.162
Road Sys Identifier			Lane Shift/Crossover	0.811	0.189
County	0.828	0.172	Work on Shoulder or Median	0.854	0.146
Interstate	0.843	0.157	*Crash Location in Work Zone*		
Local	0.838	0.162	Activity Area	0.861	0.139
State	0.829	0.171	Advance Warning Area	0.814	0.186
Turnpike/Toll	0.813	0.187	Before the First Work Zone Warning Sign	0.806	0.194
U.S.	0.793	0.207	Termination Area	0.795	0.205
Number of Vehicle Involved in Crash			Transition Area	0.827	0.173
Single Vehicle	0.801	0.199	*Law Enforcement in Work Zone*		
Multi Vehicle	0.857	0.143	Yes	0.813	0.187
Road Surface Condition			No	0.864	0.136

Table 6
CS-SVM Nighttime Variable Impact Analysis.

Variable	Severity		Variable	Severity	
	No Injury	Fatality/Injury		No Injury	Fatality/Injury
Crash-Level Variables			Dry	0.804	0.196
Crash Type			Wet	0.782	0.218
Backed Into	0.787	0.213	**Vehicle-Level Variables**		
Left Entering	0.774	0.226	*Number of Passengers*		
Left-Rear	0.778	0.222	Driver Only	0.830	0.170
Off-Road	0.792	0.208	Single Occupant	0.777	0.223
Parked Vehicle	0.790	0.210	Multi Occupant	0.750	0.250
Pedestrian	0.764	0.236	*Alcohol Related*		
Rear-End	0.768	0.232	Yes	0.739	0.261
Right Angle	0.776	0.224	No	0.814	0.186
Rollover	0.780	0.220	*Distraction Related*		
Same Direction Sideswipe	0.819	0.181	Yes	0.786	0.214
Single Vehicle	0.795	0.205	No	0.802	0.198
Weather Condition			**Work Zone Features**		
Clear	0.805	0.195	*Type of Work Zone*		
Cloudy	0.784	0.216	Intermittent or Moving Work	0.765	0.235
Rain	0.777	0.223	Lane Closure	0.813	0.187
Road Sys Identifier			Lane Shift/Crossover	0.797	0.203
County	0.795	0.205	Work on Shoulder or Median	0.760	0.240
Interstate	0.810	0.190	*Crash Location in Work Zone*		
Local	0.756	0.244	Activity Area	0.806	0.194
State	0.778	0.222	Advance Warning Area	0.772	0.228
Turnpike/Toll	0.810	0.190	Before the First Work Zone Warning Sign	0.788	0.212
U.S.	0.758	0.242	Termination Area	0.756	0.244
Number of Vehicle Involved in Crash			Transition Area	0.797	0.203
Single Vehicle	0.795	0.205	*Law Enforcement in Work Zone*		
Multi Vehicle	0.793	0.207	Yes	0.823	0.177
Road Surface Condition			No	0.763	0.237

that crashes are more likely to result in a fatality/injury in pedestrian-related, rollover, and left-entering work zone crashes, but less likely in backed-into, same direction sideswipe, parked vehicle, single vehicle, and rear-end crashes. Sideswipe crashes in the same direction have the lowest probability of fatality/injury crashes with 0.137. On the other hand, with 4.48% and 2.69% lower probabilities when compared to pedestrian-related and rollover crashes, left-entering crashes are among the top three types of crash resulting in fatal/injury. Fig. 3

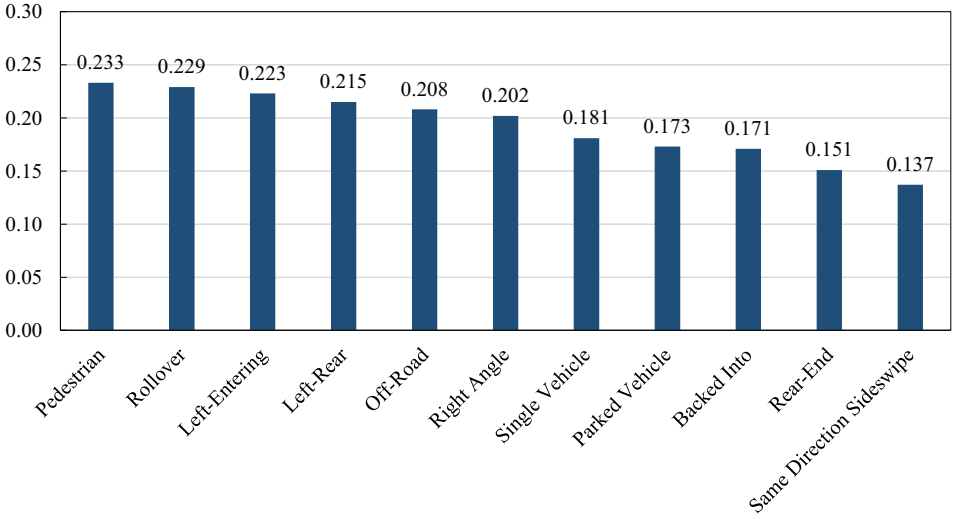

Fig. 3. Effect of Crash Type on Crash Severity, Results of CS-SVM Daytime Model.

illustrates the effects of different crash types sorted from those with the highest to lowest impacts on fatality/injury crashes obtained from the daytime SVM model.

As for other crash-related variables with estimated random parameters, a mean of −0.330 and a standard deviation of 1.039 implies that 62.46% of the multi-vehicle crashes are less likely to result in a fatal/injury crash.

6.1.2. Environmental characteristics

The rainy weather indicator was also found to be statistically significant, with a normally distributed random parameter with mean −0.353 and standard deviation 1.521. This corresponds to 59.18% of the crash occurrences under rainy weather being less likely to result in a fatal/injury crash. In addition, 72.85% of crashes where the driver was the only occupant were less likely to be a fatal/injury crash. Results from CS-SVM model also illustrate that the probability of driver-only vehicles being fatal/injury crashes is 52.99% and 91.45% lower than that with single and multi-occupants, respectively.

6.2. Nighttime crash severity models

6.2.1. Crash characteristics

A total of 14 indicator variables were found to be significant in the nighttime model, where four were found to have heterogeneous impacts on crash severity outcomes (all four were crash characteristics). As previously stated, multi-vehicle involvement, left-entering crash type, alcohol consumption, and single vehicle involvement were found to have normally distributed random parameters. With a mean of 1.250 and standard deviation of 2.586, 68.56% (greater than zero) of left-entering crashes in worker-involved work zone crashes are more likely to result in a fatal/injury crash. Simultaneously, 31.44% (less than zero) of left-entering crashes are associated with crashes in which no injury was sustained. The heterogeneous nature may have linked with unobservables related to nighttime conditions, such as the level of lighting present in the work zone or the ability to see reflective vests worn by workers.

The next estimated variable with a normally distributed random parameter is the indicator for drivers under the influence of alcohol. With a mean of 0.729 and corresponding standard deviation of 1.726, 33.64% of crashes involving a driver under the influence of alcohol are less likely to result in a fatal/injury crash, and 66.36% are more likely. With a higher likelihood of alcohol consumption during nighttime hours, the significance of this variable is expected (Yasmin et al. 2014). The heterogeneous nature is consistent with findings from previous works. For example, (Xie et al. 2012) found driving under the influence to increase the likelihood of a no-injury crash. The majority of work, however, found alcohol to increase the likelihood of a severe injury crash (Kockelman and Kweon, 2002; Qi et al., 2005; Bai and Li, 2007; Harb et al., 2008; Morgan and Mannering, 2011; Xiong et al., 2014; Chen et al., 2015). This is in agreement with the results from the present work that found that the majority have an increase in the severe injury likelihood. In addition, these results are in-line with the results of the developed SVM model, where the SVM model shows a 40.32% higher probability of a severe crash if the driver was driving under the influence of alcohol.

The single vehicle indicator variable is the last indicator to have a normally distributed estimated random parameter, with a mean of −0.745 and a standard deviation of 1.261. This indicates that 27.73% (greater than zero) of single vehicle crashes are more likely to result in a fatal/injury crash, whereas 72.27% (less than zero) are less likely. To be more specific, the majority of single vehicle crashes were less probable to result in a fatal/injury crash in work zone crashes with workers. The same result obtained from the SVM model shows that with a probability of 0.205, single vehicle crashes were less likely to be in a fatal/injury crash at nighttime, compared to multi-vehicle crashes with a probability of 0.207.

Of the remaining significant variables, left-rear crashes, pedestrian-related crashes, right angle crashes, parked vehicle crashes, and rear-end crashes have moderate impacts on the probability of a crash resulting in a fatal/injury crash. The analysis of marginal effects shows that pedestrian-related and right-angle crashes have 0.698 and 0.298 higher probabilities, respectively, of resulting in a fatal/injury crash. This result is not only consistent with the SVM output that shows a lower probability of single vehicle involvement compared to multi-vehicle in fatal/injury crashes, but is also consistent with the findings of previous studies on work zone injury severity (Katta, 2013; Dias, 2015).

6.2.2. Work zone characteristics

Previous work zone safety studies in the area of crash severity lack an important contributing factor called law enforcement. The results from this study shed light on the importance of law enforcement and were found to have a statistically significant effect on crash severity. According to the marginal effects, the presence of law enforcement decreases in the probability of fatal/injury crashes by 0.067, holding all other independent variables constant. The same conclusion can

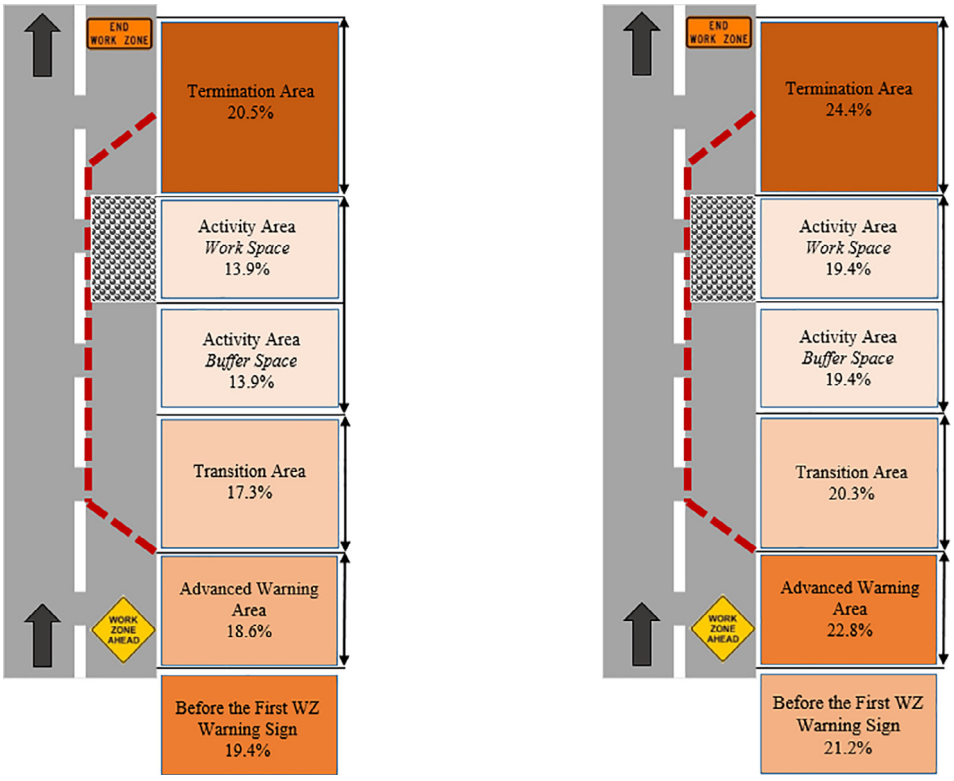

Fig. 4. Critical Locations in Work Zone([1]).

be inferred from the SVM results. It was shown that of the crash occurrences in nighttime work zones while workers are present, the absence of law enforcement is associated with approximately a 34.0% higher probability of a fatal/injury crash. This result emphasizes the need for the appropriate managing of work zone operations and the role of traffic control devices such as stationary or circulating enforcements to warn drivers when approaching a work zone, especially where workers are present.

Finally, with the use of the results from the SVM variable impact analysis, the impacts of work zone crash location variables to the probability distribution of experiencing Fatal/Injury crashes was investigated. To this end, a heat map was created on a typical work zone layout to visualize the critical locations of work zone configuration in the format of relative probabilities for worker safety. The heat map is illustrated in Fig. 4.

As observed in Fig. 4, the termination area is the most critical location because it increases the likelihood of severe crashes in both daytime and nighttime work zones. This area, in terms of impact on severity, is followed by the area before the first work zone sign in daytime work zones, and the advance warning area in nighttime work zones. This finding may be attributed to a driver's intention to speed as they are exiting the work zone area, which is in line with the previous work zone crash analysis results (Osman et al., 2018; Adomah et al., 2021b). The effects of speed variation on crash frequency were also recently investigated (Kamrani et al., 2018; Arvin et al., 2019; Parsa et al., 2019), and results demonstrated that higher speed volatility is associated with a higher likelihood of crash occurrence.

7. Summary and conclusions

Lack of awareness of worker safety in construction work zones represents a significant concern in roadway safety because it could lead to worker casualties. While research has been conducted to investigate the crash characteristics of nighttime and daytime construction activities, the statistical reasons were unknown. In addition, worker presence and its impact on severity in work zone crashes has remained largely unexplored. To address this gap in research, this study was undertaken to empirically examine the crash severity contributing factors by time-of-day (daytime vs. nighttime) for worker-involved work zone crashes.

Likelihood ratio tests were performed to determine parameter transferability of model estimates among daytime and nighttime crashes. Parameter transferability test results demonstrated with a high level of confidence that parameter estimates among daytime and nighttime crashes are not transferable (i.e., crash severity factors are contingent on these time periods), and such crashes should be modeled separately.

Due to the limitations of parametric models, such as the pre-assumption of data distribution and linear form of utility functions, which may not necessarily be applicable for crash data, non-parametric SVM models were also utilized to predict the entire set of explanatory variables in both models. When comparing the model performance, CS-SVM produced a higher percentage of correctly predicted crash severity levels by 35.04% for daytime and 38.81% for nighttime compared to the SVM models, which were also higher than that produced by the BMXL model by 62.45% and 61.36%, respectively. This implies the ability of applying SI optimization techniques in SVM parameter selection to achieve higher prediction performance. In addition to prediction accuracy, other prediction metrics were also considered regarding goodness-of-fit. For instance, the values of the AUC metric for the CS-SVM daytime and nighttime models are 0.7371 and 0.8434, respectively. Both of these are higher compared to the BMXL model. These improvements may also be associated with consideration of the non-linearity between the explanatory variables and crash severity outcomes, which is in-line with the findings of previous studies (Yu and Abdel-Aty, 2014; Chen et al., 2016). The results obtained from a two-stage sensitivity analyses demonstrate that driver

[1] The darker the color, the higher probability of severe crashes in that location.

alcohol involvement, rainy weather condition, wet road surface, multi-occupant for vehicle occupancy, and distraction are the most significant causes of fatalities/injuries in work zone crash-involved workers in both daytime and nighttime models. By taking into consideration the number of vehicles involved and law enforcement indicators, different effects were found between daytime and nighttime conditions.

Non-parametric models like SVM do not have the ability to recognize significant variables affecting the response variable (outcome). On the other hand, the results from statistical methods do not provide any idea of where the variable effect stands among all of the variables within each category. Taking this into consideration, the integration of the traditional statistical model and machine learning technique results enhance the understandings of work zone crash characteristics to interpret the effects of work zone presence on crash severity outcome. This may eventually lead to valuable comparative information about these types of crash characteristics and provide safety experts and decision makers with the ability to prioritize work zone operations based on different time periods, and environmental and geospatial conditions which will lead to a better outcome in roadway user and worker safety.

Lastly, previous work has shown limitations as it pertains to kernel function selection or the appropriate split of training and testing datasets in crash data analysis (Li et al., 2012; Yu and Abdel-Aty, 2013; Chen et al., 2016). With that in mind, this work sheds additional light on parameter tuning; more specifically, its importance on the prediction performance of ML models and the ability to enhance with the application of swarm intelligence algorithms. Although this line of research is very promising, the amount of studies that address this issue is still relatively scarce. Recommendations for future investigation should focus on applying different feature and parameter selection techniques on different machine learning methods. From a statistical modeling perspective, the ability of a model to accurately predict outcomes is just as important as its ability to explain causal factors. Current traffic safety literature lacks such a discussion. Thus, deeper examination of model outcomes is necessary in traffic safety analysis to avoid any misunderstanding of the impact of contributing factors. Moreover, investigation of the similarities and differences of risk factors in work zone crash severities with or without worker presence by time of day may be of interest for future studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adomah, E., Bakhshi, A.K., Ahmed, M.M., 2021b. Safety impact of connected vehicles on driver behavior in rural work zones under foggy weather conditions. in: Proceedings of the 100th Annual Meeting Transportation Research Board, pp. No. TRBAM-21-03428.

Adomah, E., Bakhshi, A.K., Ahmed, M.M., 2021a. Safety impact of connected vehicles on driver behavior in rural work zones under foggy weather conditions. In: Proceedings of the Transportation Research Board 100th Annual Meeting, Washington, DC. (No. TRBAM-21-03428).

Alkheder, S., Taamneh, M., Taamneh, S., 2017. Severity prediction of traffic accident using an artificial neural network. J. Forecasting 36 (1), 100–108.

Aly, W.M., 2013. Evaluation of cuckoo search usage for model parameters estimation. Int. J. Comput. Appl. 78 (11).

Anderson, J.C., Dong, S., 2017. Heavy-vehicle driver injury severity analysis by time of week: a mixed logit approach using hsis crash data. Inst. Transp. Eng. ITE J. 87 (9), 41–49.

Anderson, J., Hernandez, S., 2017. Roadway classifications and the accident injury severities of heavy-vehicle drivers. Anal. Methods Accident Res. 15, 17–28.

Arditi, D., Lee, D.-E., Polat, G., 2007. Fatal accidents in nighttime vs. daytime highway construction work zones. J. Saf. Res. 38 (4), 399–405.

Artba,Work zone fatal crashes and fatalities. American Road & Transportation Builders Association, National Work Zone Safety Information Clearinghouse, https://www.workzonesafety.org/, Accessed December, 10, 2018.

Arvin, R., Kamrani, M., Khattak, A.J., 2019. How instantaneous driving behavior contributes to crashes at intersections: extracting useful information from connected vehicle message data. Accid. Anal. Prev. 127, 118–133.

Bai, Y., Li, Y., 2007, Determining major causes of highway work zone accidents in kansas, phase ii. Kansas Department of Transportation, Final Report: K-TRAN: KU-06-1.

Bakhshi, A.K., Ahmed, M.M., 2021. Practical advantage of crossed random intercepts under bayesian hierarchical modeling to tackle unobserved heterogeneity in clustering critical versus non-critical crashes. Accid. Anal. Prev. 149, 105855.

Behnood, A., Mannering, F.L., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: some empirical evidence. Anal. Methods Accident Res. 8, 7–32.

Behnood, A., Mannering, F.L., 2016. An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. Anal. Methods Accident Res. 12, 1–17.

Behnood, A., Mannering, F., 2017. Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances. Anal. Methods Accident Res. 16, 35–47.

Behnood, A., Mannering, F., 2019. Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes. Anal. Methods Accident Res. 23, 100102.

Bejleri, I., Signal four analytics. The GeoPlan Center, Department of Urban & Regional Planning, University of Florida, https://s4.geoplan.ufl.edu/analytics/, Accessed December, 10, 2018.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the the fifth annual workshop on Computational learning theory, pp. 144–152.

Bottou, L., Lin, C.-J., 2007. Support vector machine solvers. Large Scale Kernel Mach. 3 (1), 301–320.

Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Anal. Methods Accident Res. 3, 11–27.

Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. Mach. Learning 46 (1–3), 131–159.

Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random parameters and random effects models. Anal. Methods Accident Res. 1, 86–95.

Chen, C., Zhang, G., Tarefder, R., Ma, J., Wei, H., Guan, H., 2015. A multinomial logit model-bayesian network hybrid approach for driver injury severity analyses in rear-end crashes. Accid. Anal. Prev. 80, 76–88.

Chen, C., Zhang, G., Qian, Z., Tarefder, R.A., Tian, Z., 2016. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid. Anal. Prev. 90, 128–139.

Dias, I.M., 2015. Work Zone Crash Analysis and Modeling to Identify Factors Associated with Crash Severity and Frequency Ph.D. Dissertation. Kansas State University.

Garber, N.J., Zhao, M., 2002. Distribution and characteristics of crashes at different work zone locations in Virginia. Transp. Res. Rec. 1794 (1), 19–25.

Ghasemzadeh, A., Hammit, B.E., Ahmed, M.M., Young, R.K., 2018. Parametric ordinal logistic regression and non-parametric decision tree approaches for assessing the impact of weather conditions on driver speed selection using naturalistic driving data. Transp. Res. Rec. 2672 (12), 137–147.

Greene, W.H., 2018. Econometric Analysis. Pearson, New York, NY.

Haghighi, N., Liu, X.C., Zhang, G., Porter, R.J., 2018. Impact of roadway geometric features on crash severity on rural two-lane highways. Accid. Anal. Prev. 111, 34–42.

Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. J. Saf. Res. 41 (4), 347–357.

Haleem, K., Alluri, P., Gan, A., 2015. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accid. Anal. Prev. 81, 14–23.

Harb, R., Radwan, E., Yan, X., Pande, A., Abdel-Aty, M., 2008. Freeway work-zone crash analysis and risk identification using multiple and conditional logistic regression. J. Transp. Eng. 134 (5), 203–214.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108, 27–36.

Kamrani, M., Arvin, R., Khattak, A.J., 2018. Extracting useful information from basic safety message data: an empirical study of driving volatility measures and crash frequency at intersections. Transp. Res. Rec. 2672 (38), 290–301.

Katta, V., 2013. Development of Crash Severity Model for Predicting Risk Factors in Work Zones for Ohio Master's Thesis. University of Toledo.

Keerthi, S.S., 2002. Efficient tuning of svm hyperparameters using radius/margin bound and iterative algorithms. IEEE Trans. Neural Networks 13 (5), 1225–1229.

Kim, J.-K., Ulfarsson, G.F., Shankar, V.N., Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accid. Anal. Prev. 42 (6), 1751–1758.

Kim, J.-K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. Accid. Anal. Prev. 50, 1073–1081.

Kitali, A.E., Mokhtarimousavi, S., Kadeha, C., Alluri, P., 2021. Severity analysis of crashes on express lane facilities using support vector machine model trained by firefly algorithm. Traffic Inj. Prev. 22 (1), 79–84.

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34 (3), 313–321.

Li, Y., Bai, Y., 2009. Highway work zone risk factors and their impact on crash severity. J. Transp. Eng. 135 (10), 694–701.

Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486.
Li, X., Yin, M., 2013. A hybrid cuckoo search via lévy flights for the permutation flow shop scheduling problem. Int. J. Prod. Res. 51 (16), 4732–4754.
Longadge, R., Dongre, S., 2013. Class imbalance problem in data mining review. Int. J. Comput. Sci. Network (IJCSN) 2 (1), 83–87.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Accident Res. 1, 1–22.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accident Res. 11, 1–16.
Mcfadden, D., Train, K., 2000. Mixed mnl models for discrete response. J. Appl. Econ. 15 (5), 447–470.
Mittlböck, M., Schemper, M., 1996. Explained variation for logistic regression. Stat. Med. 15 (19), 1987–1997.
Moghaddam, F.R., Afandizadeh, S., Ziyadi, M., 2011. Prediction of accident severity using artificial neural networks. Int. J. Civ. Eng. 9 (1), 41.
Mohammadnazar, A., Arvin, R., Khattak, A.J., 2021. Classifying travelers' driving style using basic safety messages generated by connected vehicles: application of unsupervised machine learning. Transp. Res. Part C: Emerg. Technol. 122, 102917.
Mokhtarimousavi, S., 2019. A time of day analysis of pedestrian-involved crashes in California: investigation of injury severity, a logistic regression and machine learning approach using hsis data. Inst. Transp. Eng. ITE J. 89 (10), 25–33.
Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2019. Improved support vector machine models for work zone crash injury severity prediction and analysis. Transp. Res. Rec. 2673 (11), 680–692. https://doi.org/10.1177/0361198119845899.
Mokhtarimousavi, S., Anderson, J.C., Azizinamini, A., Hadi, M., 2020a. Factors affecting injury severity in vehicle-pedestrian crashes: a day-of-week analysis using random parameter ordered response models and artificial neural networks. Int. J. Transp. Sci. Technol. 9 (2), 100–115.
Mokhtarimousavi, S., Azizinamini, A., Hadi, M., 2020b. Severity of worker-involved work zone crashes: A study of contributing factors. In: Proceedings of the International Conference on Transportation and Development 2020, pp. 47–59.
Mokhtarimousavi, S., Hadi, M., Sadeghvaziri, E., Azizinamini, A., 2021. Evolutionary training approaches for machine learning models for analyzing classification imbalanced crash datasets. in: Proceedings of the Transportation Research Board 100th Annual Meeting, Washington, DC. , pp. No. TRBAM-21-03834.
Mokhtarimousavi, S., Talebi, D., Asgari, H., 2018. A non-dominated sorting genetic algorithm approach for optimization of multi-objective airport gate assignment problem. Transp. Res. Rec. 2672 (23), 59–70.
Morgan, A., Mannering, F.L., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accid. Anal. Prev. 43 (5), 1852–1863.
Mujalli, R.O., De Oña, J., 2013. Injury severity models for motor vehicle accidents: a review. Proc. Inst. Civ. Eng. http://hdl.handle.net/10481/24455.
Osman, M., Paleti, R., Mishra, S., Golias, M.M., 2016. Analysis of injury severity of large truck crashes in work zones. Accid. Anal. Prev. 97, 261–273.
Osman, M., Paleti, R., Mishra, S., 2018. Analysis of passenger-car crash injury severity in different work zone configurations. Accid. Anal. Prev. 111, 161–172.
Pahukula, J., Hernandez, S., Unnikrishnan, A., 2015. A time of day analysis of crashes involving large trucks in urban areas. Accid. Anal. Prev. 75, 155–163.
Parsa, A.B., Taghipour, H., Derrible, S., Mohammadian, A.K., 2019. Real-time accident detection: coping with imbalanced data. Accid. Anal. Prev. 129, 202–210.
Peace, G.S., 1993. Taguchi Methods: A Hands-on Approach. Addison Wesley Publishing Company, New York, NY.
Qi, Y., Srinivasan, R., Teng, H., Baker, R.F., 2005, Frequency of work zone accidents on construction projects. New York State Department of Transportation and U.S. Department of Transportation Final Report: 55657-03-15.
Rahmani, R., 2018. Statistical and Simulation Methods for Evaluating Stationary and Mobile Work Zone Impacts Ph.D. Dissertation. University of Missouri-Columbia.
Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: a random parameters approach with heterogeneity in means and variances. Anal. Methods Accident Res. 15, 41–55.
Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., Yang, B., Liu, D., 2016. Evolving support vector machines using fruit fly optimization for medical data classification. Knowl.-Based Syst. 96, 61–75.
Srinivasan, S., Carrick, G., Zhu, X., Heaslip, K., Washburn, S., 2007, Analysis of crashes in freeway work-zone queues: A case study. Southeastern Transportation Center, University of Tennessee, Knoxville, Tennessee, Final Report: 07-UF-R-S3.
Srinivasan, R., Ullman, G., Finley, M., Council, F., 2011. Use of empirical bayesian methods to estimate crash modification factors for daytime versus nighttime work zones. Transp. Res. Rec. 2241 (1), 29–38.
Taghiyeh, S., Xu, J., 2016. A new particle swarm optimization algorithm for noisy optimization problems. Swarm Intell. 10 (3), 161–192.
Timeanddate, Sunrise and sunset in florida, united states. Timeanddate.com, https://www.timeanddate.com/astronomy/usa/florida, Accessed Jannuary 10, 2019.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, England.
Vapnik, V., 1998. Statistical Learning Theory. John Wiley & Sons, New York, NY.
Vasebi, S., Hayeri, Y.M., Jin, P.J., 2021. Surrounding vehicles' contribution to car-following models: A deep-learning based analysis. in: Proceedings of the 100th Annual Meeting Transportation Research Board, pp. No. TRBAM-21-03521.
Wang, J., Hughes, W.E., Council, F.M., Paniati, J.F., 1996. Investigation of highway work zone crashes: what we know and what we don't know. Transp. Res. Rec. 1529 (1), 54–62.
Wang, J., Wu, X., Zhang, C., 2005. Support vector machines based on k-means clustering for real-time business intelligence systems. Int. J. Bus. Intelligence Data Mining 1 (1), 54–64.
Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC, Boca Raton, FL.
Weng, J., Meng, Q., Wang, D.Z., 2013. Tree-based logistic regression approach for work zone casualty risk assessment. Risk Anal. 33 (3), 493–504.
Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accid. Anal. Prev. 47, 36–44.
Xiong, Y., Tobias, J.L., Mannering, F.L., 2014. The analysis of vehicle crash injury-severity data: a Markov switching approach with road-segment heterogeneity. Transp. Res. Part B: Methodological 67, 109–128.
Yang, X.-S., 2010. Suash,“” engineering optimisation by cuckoo search“,”. Int. J. Math. Modelling Numer. Optim. 1 (4), 330–343.
Yang, X.-S., Deb, S., 2009. Cuckoo search via lévy flights. In: Proceedings of the 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), pp. 210–214.
Yang, X.-S., Deb, S., 2013. Multiobjective cuckoo search for design optimization. Comput. Oper. Res. 40 (6), 1616–1624.
Yang, H., Ozbay, K., Ozturk, O., Xie, K., 2015. Work zone safety analysis and modeling: a state-of-the-art review. Traffic Inj. Prev. 16 (4), 387–396.
Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal. Methods Accident Res. 1, 23–38.
Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Anal. Methods Accident Res. 1, 72–85.
Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accid. Anal. Prev. 51, 252–259.
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.