Accident Analysis and Prevention 208 (2024) 107794

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Enhancing work zone crash severity analysis: The role of synthetic minority oversampling technique in balancing minority categories

Muhammad Adeel[a], Asad J. Khattak[a,*], Sabyasachee Mishra[b], Diwas Thapa[b]

[a] Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, United States
[b] Department of Civil Engineering, University of Memphis, TN, United States

ARTICLE INFO

Keywords:
Work zones (WZ) crashes
Injury severity
Synthetic Minority Oversampling Technique (SMOTE)
Partial Proportional Odds (PPO) model
Random Forest (RF) model
Aggressive driving
Drunk driving
Posted speed limit
Weather conditions
Lighting conditions

ABSTRACT

Road work zones are becoming increasingly common due to the aging infrastructure and the need for capacity enhancement. They present significant safety risks due to narrow lanes, uneven traffic flow, lower speed, and reduced visibility. It is particularly important to understand the role of human behavioral factors in WZ crash injury severity due to difficulty navigating such areas. Furthermore, the crash injury data available is mostly imbalanced, primarily due to the lower incidence of high-cost fatal and severe injuries, and can benefit from the use of emerging analysis techniques. This research study examines a unique dataset comprising 7,855 WZ crashes in Tennessee from 2018 to 2022 as a case study to provide useful insight into the behavioral factors associated with injury severity and how they change after adjusting for the underrepresented fatal and serious injuries within the dataset. The study applies frequentist methods and a machine learning technique enhanced with the Synthetic Minority Oversampling Technique (SMOTE), addressing the data imbalance (relatively fewer fatal and serious injuries) for useful inferences and predictions. The study results indicate that aggressive driving, overspeeding, and drunk driving significantly elevate injury severity. Additionally, after balancing the minority categories of crash injury severity levels, the importance of contributing factors changes. The study offers engineers and data analysts a framework for analyzing imbalanced data, a prevalent issue in crash injury severity analysis. By exploring key behavioral factors responsible for injury severity in WZ crashes, the study provides useful insight and valuable information to traffic safety engineers, transportation agencies, and policymakers to implement enhanced safety measures in WZ design and management, ultimately aiming to mitigate injury severity and to improve overall safety for road users.

1. Introduction

Road networks are important in providing mobility and boosting a country's economy. The aging characteristics of the highway infrastructure, especially in the United States, and the construction or maintenance works have led to increasing road work zones (WZs). However, these WZs pose significant risks to the safety of drivers, passengers, pedestrians, and workers due to various factors such as variations in traffic flow, lane closures, posted speed limits, and reduced visibility. To construct, maintain, and rehabilitate a road network, activities are sometimes carried out during active traffic hours, necessitating careful planning and execution of WZs for the safety of road users (Muhammad et al., 2018; Yang et al., 2015). WZ crashes are a significant concern for transportation agencies, construction agencies, and road users because they are high-risk, resulting in fatalities, injuries, property damage, delays, and traffic congestion.

According to the National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS), the Federal Highway Administration (FHWA) reported a total of 956 traffic fatalities in the WZs in the United States in 2021. Of these, 778 fatalities were drivers and passengers, 173 were bicyclists and pedestrians, and 5 were categorized as others (FHWA, 2021). Additionally, approximately 42,151 injuries and 61,893 instances of property damage were reported, totaling 105,000 crashes in WZs in 2021 (Crashes, 2021). Despite significant technological advancements that have enhanced the operations of WZs, such as smart work zones (Venthuruthiyil et al., 2023; Tang and Hu, 2024), societal concerns regarding safety and mobility near WZs persist, and Tennessee highways are no exception. In Tennessee, 3,855

* Corresponding author.
E-mail addresses: madeel1@vols.utk.edu, adeelm@ornl.gov (M. Adeel), akhattak@utk.edu (A.J. Khattak), smishra3@memphis.edu (S. Mishra), dthapa@memphis.edu (D. Thapa).

https://doi.org/10.1016/j.aap.2024.107794
Received 5 March 2024; Received in revised form 18 September 2024; Accepted 19 September 2024
Available online 27 September 2024

work zone crashes were reported in 2022 (TDOT, 2022), the majority of which involved distracted driving and other human errors. This significant number emphasizes the need to investigate behavioral factors contributing to the severity of injuries sustained in such incidents. Given the new developments in analytical methods, understanding the role of these factors on the entire spectrum of injury severity is critical due to the challenges of navigating WZs and the evolving nature of human behavior.

Significant efforts and studies have been carried out in the past by researchers to identify, analyze, and quantify the key contributing factors responsible for the severity of WZ crash injuries. Various frequentist methods (Khattak et al., 2002; Osman et al., 2019; Sze & Song, 2019) and machine learning techniques (Hasan et al., 2022) have been used to study and correlate the different factors associated with it. Most of the past studies contain plausible, intuitive, and logical results, yet many aspects of the WZ crash injury severity remain unexplored. Moreover, in the realm of machine learning and data analysis, it is common to encounter imbalanced crash data, which often includes instances with a lower occurrence of fatal or serious injury crashes. Despite their lower frequency, these types of crashes carry a significantly higher social cost than crashes involving property damage only (PDO). According to the United States Department of Transportation, Federal Highway Administration (FHWA) safety program, the social cost of a fatal injury resulting from a crash is estimated at approximately 11 million (2016 dollars). In contrast, a PDO crash costs around 12,000 (2016 dollars) (Harmon et al., 2018). Therefore, effectively addressing the imbalance in crash data is crucial for the accuracy and overall efficacy of the model.

Considering the aforementioned limitations, this research study intends to identify and explore the key behavioral factors responsible for the severity of WZ crash injuries while controlling for crash attributes, geometric features of the roadway, environment, and types of WZs using a unique dataset of Tennessee WZ crashes. The study utilizes frequentist methods (Ordered Logistic Regression and Proportional Odds models) and a machine learning (ML) technique (Random Forest algorithm) for estimating relationships and making predictions, followed by a comparison to gain valuable insights. Importantly, proficiently implementing the Synthetic Minority Oversampling Technique (SMOTE) is integrated into the Random Forest model to correct the dataset's imbalance. This approach strategically augments the representation of the minority class, resulting in a more robust and reliable analysis that yields meaningful and intuitive results. The study findings provide valuable inferences for traffic safety engineers, construction engineers, transportation agencies, policymakers, planners, and researchers facilitating further enhancements in WZ safety.

2. Literature review

The literature review indicates that WZs have remained the topic of interest for many researchers due to their importance in road safety. Many studies have considered WZ crash frequency as a safety performance measure to examine the impact of WZ, whereas others have focused on its injury severity. While WZs are associated with higher crash frequencies, they may have lower injury severity compared with non-WZs. A synthesis of the literature indicates numerous factors contribute to WZ crashes, such as over-speeding and significant speed variation within WZs, which are major factors. Higher speeds raise the probability of crashes and their potential severity (Osman et al., 2018; Thapa et al., 2024; Zhang & Hassan, 2019a). WZ interventions aim to curb speeding and ensure consistent traffic flow. Additionally, driver behaviors, including the age and gender of the driver (Li & Bai, 2009; Zhang & Hassan, 2019a), alcohol impairment (Weng et al., 2016), and distraction or recklessness (Liu et al., 2016; Zhang & Hassan, 2019a), significantly influence the severity of injuries from WZ crashes. Vehicle characteristics, such as the number of vehicles involved (Hasan et al., 2022) and the type of vehicles (light/heavy) (Khattak & Targa, 2004; Li & Bai, 2009; Osman et al., 2018; Vieira et al., 2023) also substantially affect injury severity. Environmental factors like weather and lighting often influence WZ crash severity (Ghasemzadeh & Ahmed, 2019; Hasan et al., 2022; Li & Bai, 2009; Zhang & Hassan, 2019a). Roadway features impacting WZ crash severity include the number of lanes, the road's functional classification, traffic control devices, the presence of a curve on the road, and the posted speed limits within the WZs (Ghasemzadeh & Ahmed, 2019; Liu et al., 2016; Vieira et al., 2023). Moreover, crash attributes such as the type of crash (Vieira et al., 2023; Zhang & Hassan, 2019a), the nature of the vehicle collisions (Yu et al., 2020), the design and duration of the WZ (Garber & Zhao, 2002; Khattak et al., 2002; Muhammad et al., 2018), and the presence of vulnerable road users like pedestrians (Sze & Song, 2019) play a pivotal role in determining crash severity in WZs.

Researchers in the past have employed a variety of statistical tools for analyzing WZ crash severity, including but not limited to Ordered Logistic Regression, Multinomial Logistic Regression, Mixed Generalized Ordered Response Probit models, Hierarchical models, Bayesian models, and Partial Proportional Odds model (Khattak et al., 2002; Osman et al., 2019; Sze & Song, 2019; Yu et al., 2022). More recently, machine learning (ML) techniques have emerged for the analysis of WZ crash injury severity (Chen et al., 2016; Hasan et al., 2022, 2023; Santos et al., 2022). Among these, the Random Forest algorithm has proven to be highly efficient and recommended for crash injury severity analysis due to its robustness against overfitting, its ability to handle various types of data, and its capacity to rank the importance of variables (Ahmad et al., 2023; Ashqar et al., 2021; Breiman, 2001; Hasan et al., 2023; Santos et al., 2022; Usman et al., 2024).

To address the class imbalance in datasets, various techniques are employed, such as Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE, Adaptive Synthetic Sampling (ADASYN), Random Over-Sampling Examples (ROSE), Random Under-sampling of the Majority Class (RUMC), Cluster-Based Under-Sampling (CBUS), and mixed resampling, etc. SMOTE generates new minority class instances by interpolating between existing samples and their nearest neighbors, and it has been widely used since 2002 due to its simplicity and effectiveness (Chawla et al., 2002). Borderline-SMOTE focuses on borderline minority samples, reducing overlap with majority class samples (Han et al., 2005), while ADASYN generates more synthetic samples around minority samples that are harder to learn, based on their density distribution (He et al., 2008). ROSE, a bootstrap-based technique, generates synthetic examples from a conditional density estimate (Lunardon et al., 2013). RUMC involves randomly under-sampling the majority class to balance the class distribution, which can be effective but may lead to loss of information. CBUS clusters the majority class into subsets and then under-samples within each cluster, aiming to retain the original data structure. Mixed resampling combines over-sampling the minority class and under-sampling the majority class to achieve a balanced dataset, utilizing the strengths of both approaches (Chen et al., 2022; Ding et al., 2022). Each method has strengths and weaknesses, and their effectiveness varies depending on the dataset and classification task (Brandt & Lanzén, 2021). Recent studies indicate that using SMOTE with Random Forest provides higher accuracy compared to other techniques used for addressing the class imbalance in the datasets (Brandt & Lanzén, 2021; Demir & Şahin, 2022; Dey & Pratap, 2023; Kuo et al., 2024). Using emerging analytical techniques, the current study addresses the significant gap in comprehending how the evolving nature of driver behavior in WZs contributes to the spectrum of injury severity. Specifically, this study handles imbalanced injury severity data by applying SMOTE.

3. Data description

The study is based on a unique dataset, manually extracted from the Enhanced Tennessee Roadway Information Management System (E-TRIMS), compiled and maintained by the Tennessee Department of Transportation (TDOT). E-TRIMS is a single integrated, reliable system that includes an inventory of the state and local roadways, structures, traffic, and crash data. The data used in the study are the road crashes within the WZs in Tennessee from 2018 to 2022. The initial dataset, extracted from the E-TRIMS, comprised Microsoft Excel files and geographical information system (GIS) shapefiles (having longitudes/ latitudes) of the WZ crashes. Since E-TRIMS data are not available in a consolidated form, various data files containing key factors/variables associated with WZ crashes were integrated. The information/column of case ID (crash identification number as per police record) was used in all data files for merging the data. The original dataset included 8,088 WZ crash observations covering details about crash events, geometric features of the roadway, human behavior, and vehicle characteristics. After carefully screening and cleaning the data, 233 observations were removed due to missing, unknown, or incorrect information. This resulted in a final dataset of 7,855 observations representing crashes within the WZs in Tennessee.

The distribution of 7,855 WZ crashes across different injury severity levels (assessed using the KABCO scale) is as follows: Fatal Injuries (K): 59 (0.75 %); Serious Injuries (A): 224 (2.85 %); Minor Injuries (B): 1,325 (16.87 %); Possible Injuries (C): 591 (7.52 %); and No Injuries (O): 5,656 (72.01 %). Due to the relatively low number of fatal injury crashes, the dataset is reorganized into three broader categories: KA, BC, and O, grouping 'fatal and serious injuries,' 'minor and possible injuries,' while retaining 'no injury' as the lowest category. The WZ crash injury severity is selected as the 'dependent/outcome variable,' representing a categorical and ordinal scale (in ordered form). The dataset includes various potential factors that describe attributes of the crash, human behavior, roadway geometry, environment, WZ type, etc. These factors are classified into 15 distinct categories of independent/explanatory variables for the recorded crashes. To facilitate analysis, dummy variables (0–1) are generated for each categorical explanatory variable, resulting in a final dataset containing 48 explanatory/predictor variables.

Table 1 represents the descriptive statistics of the final dataset used in the study, comprising detailed information on the distribution of the injury severity levels and the key predictor/explanatory variables for all the WZ crashes. The base category is identified for each variable category. The percentages of injury severity levels across various characteristics/categories of key explanatory variables are shown in detail in the table. There are a total of 7,855 WZ crashes in the final dataset, out of which 5,656 (72 %) are 'no injury'; 1,916 (24.4 %) are 'minor or possible injury'; and 283 (3.6 %) are 'fatal or serious injury' crashes. Considering predictor 'total vehicles involved in the crash'; 786 (10.0 %) crashes involve 'more than two vehicles'; 1,972 (25.1 %) crashes involve 'two vehicles'; and 5,097 (64.9 %) crashes involve 'one vehicle in the crash'. Similarly, the results of other key predictors are presented clearly in detail in Table 1. Graphical visualization of the dependent variable and key explanatory variables is presented in Fig. 1.

Correlation among explanatory variables in statistical modeling and machine learning is a concern, as it can lead to 'collinearity' during model estimation. Cramer's V is used to assess the correlations among categorical variables with multiple unique values per category. This method quantifies the association strength on a scale from '0′ (no association) to '1′ (perfect association). In this study, Cramer's V values for the exploratory variables were low to moderate, all below + 0.75. Notable exceptions include the relationships between 'weather condition' and 'roadway surface condition,' 'dark lighted condition' and 'crash injury severity,' and 'dark lighted condition' and 'short-duration maintenance WZs,' which ranged between 0.50 and 0.75, indicating a moderate association. To further investigate, variance inflation factors (VIF) were calculated for these variables, all of which were below 2.5, confirming moderate correlations that do not necessitate any corrective measures.

4. Methodology

Fig. 2 presents an overview of the study's design, integrating two

Table 1
Descriptive statistics of WZ crash injury severity outcomes and key explanatory variables.

Variables *(Including Dummy Variables)*	Crash Injury Severity *(Dependent Variable)*			
	Total	No Injury*	Minor/ Possible Injury	Fatal/ Serious Injury
Work zone crashes	7,855 (100 %)	5,656 (72 %)	1,916 (24.4 %)	283 (3.6 %)
Crash Factors				
Total vehicles involved in the crash				
More than two vehicles involved	786 (10.0 %)	435 (55.3 %)	308 (39.2 %)	43 (5.5 %)
Two vehicles involved	1,972 (25.1 %)	1,403 (71.1 %)	475 (24.0 %)	94 (4.8 %)
One vehicle involved *	5,097 (64.9 %)	3,818 (74.9 %)	1,133(22.2 %)	146 (2.9 %)
Area of the vehicle damaged				
Front end damaged	3,689 (46.96 %)	2,537 (68.8 %)	974 (26.4 %)	178 (4.83 %)
Rear end damaged	1,227 (15.62 %)	880 (71.7 %)	318 (25.9)	29 (2.4 %)
Left/right sides damaged	1,321 (16.82 %)	995 (75.3 %)	283 (21.4 %)	43 (3.3 %)
Vehicle not damaged*	1,618 (20.6 %)	1,244 (76.9 %)	341 (21.1 %)	33 (2.0 %)
Human Behaviour Factors				
Driver actions during the crash				
Aggressive driving/ overspeeding	975 (12.4 %)	663 (68.0 %)	263 (27.0 %)	49 (5.0 %)
Improper maneuver/ braking	2,428 (30.9 %)	1,638 (67.5 %)	680 (28.0 %)	110 (4.5 %)
Driver other actions	1,934 (24.6 %)	1,538 (79.5 %)	353 (18.3 %)	43 (2.2 %)
No contributing action*	2,518 (32.1 %)	1,817 (72.2 %)	620 (24.6 %)	81 (3.2 %)
Driver physical condition				
Drunk driving/positive blood alcoholic concentration (BAC)	361 (4.6 %)	186 (51.5 %)	137 (38.0 %)	38 (10.5 %)
Driver other physical conditions	1,403 (17.86 %)	1,176 (83.8 %)	181 (12.9 %)	46 (3.3 %)
Appeared normal*	6,091 (77.5 %)	4,294 (70.5 %)	1,598(26.2 %)	199 (3.2 %)
Roadway Geometric Factors				
Roadway surface condition				
Wet road surface/water standing	1,771 (22.5 %)	1,416 (80.0 %)	315 (17.8 %)	40 (2.2 %)
Snowy road surface	73 (0.9 %)	56 (76.7 %)	17 (23.3 %)	0 (0.0 %)
Dry road surface*	6,011 (76.5 %)	4,184 (69.6 %)	1,584(26.4 %)	243 (4.0 %)
Posted speed limit (mph)				
Speed limit 31 – 60 mph	6,168 (78.5 %)	4,421 (71.6 %)	1,527(24.7 %)	220 (3.6 %)
Speed limit greater than 60 mph	625 (7.9 %)	396 (63.3 %)	185 (29.6 %)	44 (7.0 %)
Speed limit 0 – 30 mph*	1,062 (13.5 %)	839 (79.0 %)	204 (19.2 %)	19 (1.8 %)

(*continued on next page*)

Table 1 (*continued*)

Variables *(Including Dummy Variables)*	Crash Injury Severity *(Dependent Variable)*			
	Total	No Injury*	Minor/ Possible Injury	Fatal/ Serious Injury
Environmental Factors				
Weather condition				
Cloudy/foggy/windy weather	1,426 (18.2 %)	1,168 (81.9 %)	236 (16.5 %)	22 (1.54 %)
Rainy/snowy weather	837 (10.6 %)	546 (65.2 %)	250 (30.0 %)	41 (4.8 %)
Clear weather*	5,592 (71.2 %)	3,942 (70.5 %)	1,430(25.6 %)	220 (3.9 %)
Light condition				
Dark lighted condition	1,569 (19.9 %)	1,169 (74.5 %)	339 (21.6 %)	61 (3.9 %)
Dark not-lighted condition	1,089 (13.9 %)	752 (69.1 %)	291 (26.7 %)	46 (4.2 %)
Daylight*	5,197 (66.2 %)	3,735 (71.9 %)	1,286(24.7 %)	176 (3.4 %)
Work Zone Factors				
Type of work zone				
Maintenance work zone (short duration)	4,131 (52.6 %)	2,796 (67.7 %)	1,163(28.2 %)	172 (4.2 %)
Utility work zone (short duration)	2,192 (27.9 %)	1,716 (78.3 %)	403 (18.4 %)	73 (3.3 %)
Construction work zone (long duration) *	1,532 (19.5 %)	1,144 (74.7 %)	350 (22.8 %)	38 (2.5 %)

* Selected as a base among each variable category.

distinct analytical methodologies: a frequentist statistical approach and a machine learning technique enhanced by SMOTE. The statistical analysis employs regression models, specifically Ordered Logit and Partial Proportional Odds models, chosen for their suitability in handling the ordinal nature of the dependent variable, crash injury severity. These models allow a nuanced understanding of the relationships between predictors and crash outcomes. In parallel, the machine learning analysis utilizes the Random Forest algorithm, reinforced by SMOTE, to address the class imbalance and enhance prediction accuracy. The Random Forest model identifies key variables based on their importance, and these are compared with the standardized significances derived from the regression models, alongside the accuracies of both models. This dual approach provides a comprehensive analysis, blending the strengths of inferential statistics and predictive modeling. The study concludes by presenting the findings from both methodologies, offering invaluable insights and recommendations for future research (Fig. 2).

4.1. Statistical analysis – Partial Proportional Odds (PPO) model

The statistical analysis within the frequentist framework involves estimating Ordered Logistic (Ologit) and Partial Proportional Odds (PPO) regression models. The Ologit model is based on the assumption of proportional odds—or parallel lines (pl)—meaning the relationship between the dependent and explanatory variables is consistent across the ordered categories of the dependent variable, which in this study is crash injury severity. However, there is an alternative, the 'Generalized Ordered Logit' model, which relaxes the parallel lines assumption of the Ordered Logit model, allowing model coefficients to vary across different injury severity levels (Osman et al., 2016; Sasidharan & Menéndez, 2019; Williams, 2006, 2016). The PPO model is a middle-ground approach: it is less restrictive than the Ordered Logit model, relaxing the parallel lines assumption only where necessary, allowing for more parsimonious models than those generated by non-ordinal models (Peterson & Harrell Jr, 1990; Sasidharan & Menéndez, 2014, 2019; Yu et al., 2022). In the PPO model, the probability of injury severity (j) for a given WZ crash (i) can be mathematically written as (Sasidharan & Menéndez, 2014; Williams, 2006):

$$P(Y_i > j) = P_{ij} = \frac{e^{(\propto_j + X_i\beta_j)}}{1 + e^{(\propto_j + X_i\beta_j)}}\, j = 1, 2, \cdots, J-1 \tag{1}$$

Where,

P_{ij} = Probability of injury severity (j) for a given WZ crash (i)

j = WZ Crash injury severity levels (1 = PDO (No Injury), 2 = Minor/ Possible Injury, 3 = Fatal/Serious Injury, and J is the number of severity levels (in this study J=3).

X=Matrix of independent variables.

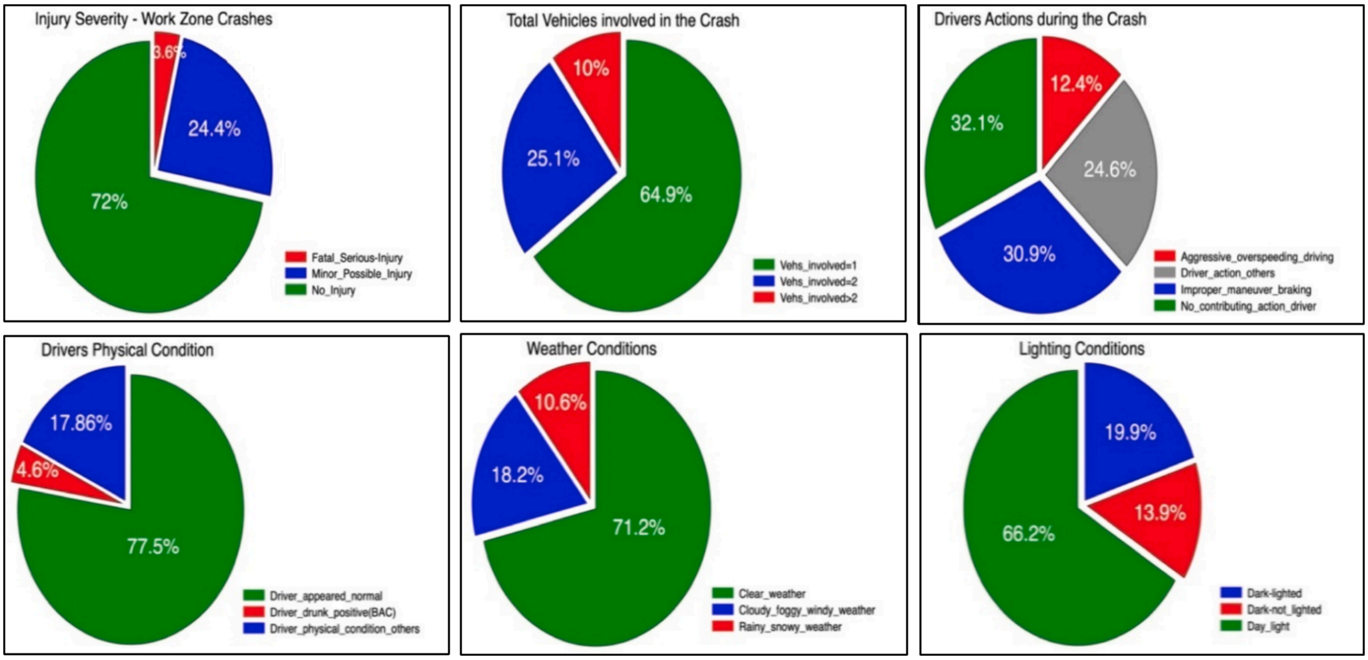

Fig. 1. Graphical visualization of dependent and key explanatory variables.

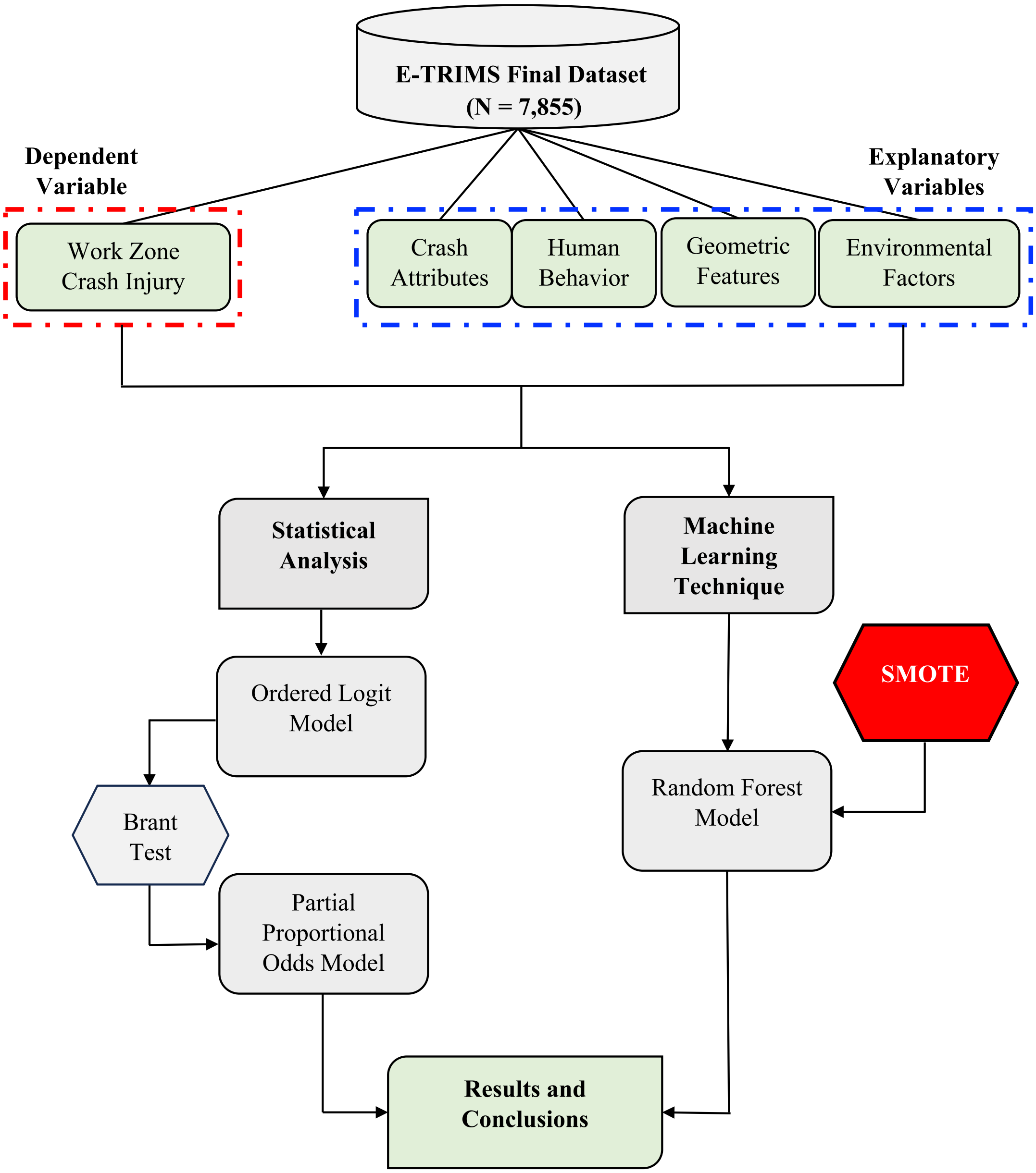

Fig. 2. Design of the study.

$\propto$ = Intercept of jth logit.

β = Regression Coefficients for X (difference in the log odds of having severity level j vs. other j − 1 severity levels).

The PPO model considers different variables based on compliance with the proportional odds assumption. For Example, in the PPO model shown in Equation (1), the variables X_1 and X_2, which adhere to this assumption, their coefficients (β_1 and β_2) remain constant across all dependent variable levels. However, for variable X_3, which does not adhere to the assumption, its coefficients (β_{3j}) vary at each level of the dependent variable, as given below:

$$P_{ij} = \frac{e^{(\propto_j + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_{3j})}}{1 + e^{(\propto_j + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_{3j})}} j = 1, 2, \cdots, J-1 \tag{2}$$

The PPO model requires careful interpretation, particularly for intermediate categories. This is because the sign for these categories does not always clearly indicate the direction of their effect (Washington et al., 2020; Wooldridge, 2010). In this study, the model's marginal effects are used to interpret the results, which effectively measure the impact of changes in independent variables on the probability of each dependent variable category. Detailed explanations of Ordered Logit and Generalized Ordered Logit models are available in numerous studies (Li & Fan, 2019; Peterson & Harrell Jr, 1990; Sasidharan & Menéndez, 2014, 2019; Williams, 2006, 2016), which can be consulted for understanding. For brevity, this paper does not include the explanations of these models and the estimates from the Ologit model. Instead, the study's findings are based on the results derived from the PPO model, primarily due to its ability to address unobserved heterogeneity (to some extent) within the data. This decision is further supported by the results of the 'Brant test,' which indicated a violation of the proportional odds/parallel line assumption (i.e., chi-square (18) = 68.6 with $p < 0.0001$). Hence, employing the PPO model is justified, as the Ordered Logit model's proportional odds assumption does not hold. Furthermore, we emphasize the interpretation of the results through marginal effects, which are crucial for understanding the impact of variables on crash severity outcomes, providing a more nuanced and accurate representation of the effects, especially for intermediate injury severity categories. By focusing on marginal effects, we provide a more accurate and practical insight into how changes in independent variables influence the probability of different crash severity levels, thereby enhancing the clarity and applicability of our findings.

4.2. Machine learning technique – Random Forest (RF) model

Introduced by Breiman in 2001, the Random Forest (RF) technique amalgamates multiple independent base classifiers known as 'decision trees.' Each tree contributes a vote towards the classification of a test sample, with the final category label determined by majority voting. The model's performance is optimized by fine-tuning hyperparameters, including the number of trees. RF enhances its efficacy through random processes at various stages, such as selecting training set subsets via the 'bagging' technique. Additionally, RF offers valuable feature ranking, identifying the most influential variables on the outcome. This integration of randomness markedly improves the RF algorithm's classification accuracy. The RF model has been extensively discussed in numerous prior studies, which provide a comprehensive understanding that can be consulted for further insight (Ahmad et al., 2023; Ashqar et al., 2021; Breiman, 2001; Hasan et al., 2023; Zarei Yazd et al., 2024). The predictive performance of the RF model is evaluated using the following metrics, with the test dataset serving as a holdout sample:

4.2.1. Accuracy of the model

Model accuracy is the proportion of correctly classified observations (True Positives + True Negatives) to the total observations in the dataset, calculated as:

$$AccuracyoftheModel = \left(\frac{Numberofcorrectlyclassifiedobservations}{Totalnumberofobservations}\right) \times 100 \tag{3}$$

4.2.2. Sensitivity/recall of the class

Sensitivity, or recall, for a particular class is the ratio of correctly classified observations in that class (True Positives) to all observations of that class (True Positives + False Negatives), calculated as:

$$Sensitivity = \frac{Numberofcorrectlyclassifiedobservationsinaclass(TruePositives)}{TruePositives + FalseNegatives} \tag{4}$$

4.2.3. Precision of the class

Precision for a class is the proportion of correctly classified observations in that class (True Positives) to all correctly predicted observations (True Positives across all classes), calculated as:

$$Precision = \frac{Numberofcorrectlyclassifiedobservationsinaclass(TruePositives)}{Truepositivesforallclasses} \tag{5}$$

4.2.4. F-1 score of the class

The F-1 score, the harmonic mean of precision and recall, is crucial, especially where class imbalances are present because accuracy alone can be misleading due to the dominance of the majority class. By incorporating precision and recall, the F-1 score provides a more comprehensive evaluation of the model's performance. The higher the F-1 score of a class, the better the predictive performance of the model specific to that class. It can be calculated as:

$$F-1Score = \frac{2^{*}Precision^{*}Recall}{(Presicion + Recall)} \tag{6}$$

In this study, the F-1 score is used to assess the performance of the RF model. It is particularly relevant when applying SMOTE, as it allows us to evaluate how well the oversampling technique has enhanced the model's ability to correctly classify minority class instances. By comparing the F-1 scores of the RF model before and after applying SMOTE, we can quantify the improvement in classification performance. A higher F-1 score post-SMOTE application indicates that the model has become better at identifying minority class instances without compromising too much on precision.

4.3. Addressing class imbalance – Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE can effectively capture the entire spectrum of injury severity. Given that fatal and severe injury crashes have relatively high costs but are relatively rare in the data, the SMOTE algorithm can serve to balance this minority class. Renowned for its effectiveness in machine learning and data mining, SMOTE augments the presence of the minority class through interpolation, fostering stronger generalization in classification models. It transcends mere duplication of minority class instances by generating synthetic examples, interpolating among various minority class instances within a neighborhood. This method prioritizes the 'feature space' over the 'data space,' formed by the instance and its K-nearest neighbors, focusing on feature values and their interrelations. Besides SMOTE, various other techniques are employed to address the class imbalance in datasets. The pros and cons of a few methods, along with a comparison to SMOTE, are presented in Table 2.

To balance the class distribution in the dataset, the SMOTE algorithm generates synthetic samples for the minority class as follows: for a given minority class sample x, identify its 5 nearest neighbors within the same class based on the smallest Euclidean distance or $n_{min} - 1$ neighbors if the minority class count n_{min} is less than or equal to 5. Randomly select one of these nearest neighbors, denoted as x_R. Construct a new synthetic sample, S, using the formula:

$$S = x + u^{*}(x_R - x)\frac{2^{*}Precision^{*}Recall}{(Presicion + Recall)} \tag{7}$$

Where u is a random number between 0 and 1, drawn from a uniform distribution $U(0,1)$. The value of u is constant for all attributes of a particular synthetic sample but varies across different SMOTE samples. This procedure ensures that each synthetic sample is positioned on the line segment connecting the original sample and its selected neighbor. The application of SMOTE expands the minority class, resulting in a balanced training set for subsequent analyses. Fig. 3 provides a visual representation of the SMOTE process. The left panel shows the original dataset, with black circles representing the majority class and blue circles representing the minority class, highlighting the imbalance between them. The right panel illustrates the application of SMOTE to the data. Synthetic samples, shown as green circles, are generated at random distances along the straight lines between the nearest neighbors of the minority class. This results in a more balanced distribution between the majority and minority classes in the dataset. Extensive discussions and analyses of SMOTE are present in prior research, contributing to its established reputation in the field (Ali et al., 2024; Chawla et al., 2002; Dey & Pratap, 2023; Joloudari et al., 2023; Kuo et al., 2024; Luo, 2023; Soundrapandiyan et al., 2023; Waqar et al., 2021).

5. Results

5.1. Statistical analysis – Partial Proportional Odds (PPO) model

The PPO model's estimates for different injury severity levels are presented in Table 3. Three levels of injury severity for WZ crashes are considered: 'fatal or serious injury,' 'minor or possible injury,' and the lowest category 'no injury.' The PPO model produces two sets of results corresponding to the three severity levels. The first set (Panel I) aligns with the Ordered Logistic Regression model, categorizing the dependent variable (injury severity) as 'no injury' versus 'minor/possible + fatal/serious injury.' In the second set (Panel II), the dependent variable is grouped as 'no injury + minor/possible injury' versus 'fatal/serious injury.' Explanatory or predictor variables that meet the proportionality assumption have identical coefficients and t-statistics in both panels and are indicated in Table 3. The model's likelihood ratio (LR) chi-square

Table 2
Comparison of SMOTE with other class imbalance methods.

Method	Pros	Cons	Comparison to SMOTE
Borderline-SMOTE (Han et al., 2005)	Focuses on borderline cases, potentially improving the classification of hard-to-classify instances.	May not be effective for datasets with significant class overlap or noise.	While Borderline-SMOTE refines SMOTE by focusing on borderline cases, it may be less effective in scenarios where noise is prevalent, making standard SMOTE more reliable in such cases.
ADASYN (He et al., 2008)	Enhances learning by focusing on harder-to-learn minority samples.	Computationally intensive and can produce more complex synthetic samples that may complicate model learning.	ADASYN offers targeted sampling but at the cost of increased complexity, which can make SMOTE a more straightforward and efficient choice.
ROSE (Lunardon et al., 2013)	Maintains class balance through a smoothed bootstrap-based approach, making it particularly suitable for binary classification problems	Potential for overfitting, especially in smaller datasets.	ROSE is effective for binary classification problems but may lead to overfitting, particularly in smaller datasets. In our study, SMOTE's approach may be more advantageous given the ordinal nature of the outcome variable.
RUMC (Chen et al., 2022; Ding et al., 2022)	Simple to implement; directly reduces imbalance by under-sampling the majority class.	Can lead to the loss of valuable information from the majority class.	RUMC's simplicity is a strength, but the loss of majority class data can be a significant drawback, which SMOTE avoids by preserving the entire dataset.
CBUS (Chen et al., 2022; Ding et al., 2022)	Retains original data structure by clustering before under-sampling.	Computationally expensive; may struggle with complex datasets.	While CBUS retains data structure, its complexity and resource demands may make SMOTE a more accessible and efficient option.
Mixed Resampling (Chen et al., 2022; Ding et al., 2022)	Combines over-sampling and under-sampling to balance the dataset, utilizing the strengths of both methods.	Complexity in implementation; potential to introduce both noise and data loss.	Mixed Resampling offers flexibility but at the cost of increased complexity, whereas SMOTE's single-method approach can be more streamlined and easier to implement.

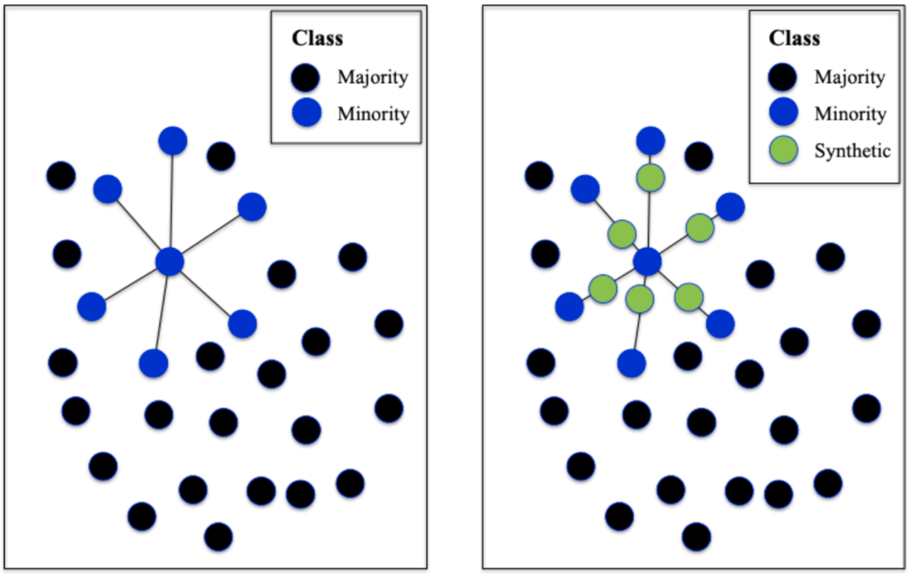

Fig. 3. Visual representation of the SMOTE process.

(χ^2) statistic is 538.69 with 25 degrees of freedom, and the probability > Chi-square = 0.0000, which indicates that the model is statistically significant at a 95 % confidence level (0.05 significance level). The model's prediction error measures, as assessed by Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are 10518.24 and 10706.40, respectively. Furthermore, the model achieves a predictive accuracy of 71.15 %, highlighting its effectiveness in predicting crash severity outcomes. Positive coefficients of explanatory or predictor variables suggest that higher values increase the probability of more severe crashes, whereas negative coefficients imply a decreased likelihood of severe crashes.

Table 3 indicates that the predictors (dummy variables) namely 'more than two vehicles involved in the crash,' 'rear-end of the vehicle damaged,' 'aggressive driving/overspeeding,' 'drunk driving/positive BAC,' 'posted speed limit of WZs between 31 and 60 mph,' 'rainy/snowy weather,' 'cloudy/foggy/windy weather,' 'dark not-lighted condition,' 'dark lighted condition,' 'maintenance work zone,' and 'utility work zone' follow the proportional odds/parallel lines assumption. Consequently, the coefficients for these predictors remain consistent across both Panel I and Panel II, indicating equal intervals/differences between injury severity levels for these variables. The interpretation of coefficients for these indicator variables is similar to the Ordered Logit model. For instance, the ordered log odds of more severe injuries in WZ crashes increase by 0.697 when more than two vehicles are involved, keeping other predictors constant. WZ crashes involving 'rear-end damage of the vehicle' result in a 21.5 % increase in the likelihood of severe injuries, with other predictors kept constant. The ordered log odds of higher injury severity are associated with an increase of 0.330 if the drivers involved in the crashes are aggressive or overspeeding. Likewise, the odds of severe injuries in WZ crashes increase by 2.14 times ($e^{0.760}$) if the drivers are drunk or have a positive BAC. Detailed results for other predictors are presented in Table 3.

The predictors (dummy variables) in the PPO model, which do not follow the parallel line assumption, include 'two vehicles involved in the

Table 3
Partial Proportional Odds model estimates for WZ crash injury severity in Tennessee.

Variables	Panel I – (No Injury) versus (Minor/Possible + Fatal/ Serious Injury)		Panel II – (No Injury + Minor/Possible Injury) versus (Fatal/Serious Injury)	
	Coefficient	**t-stats**[#]	**Coefficient**	**t-stats**[#]
Total vehicles involved in the crash				
More than two vehicles involved*	0.697	8.70**	0.697	8.70**
Two vehicles involved	0.120	1.80	0.516	3.73**
One vehicle involved (base)				
Area of the vehicle damaged				
Front end damaged	0.496	6.78**	1.028	6.22**
Rear end damaged*	0.215	2.24**	0.215	2.24**
Left/right sides damaged	0.115	1.30	0.569	2.75**
Vehicle not damaged (base)				
Driver actions during the crash				
Aggressive driving/ overspeeding*	0.330	3.73**	0.330	3.73**
Improper maneuver/ braking	0.025	0.33	− 0.351	− 2.29**
Driver other actions	− 0.199	− 2.34**	− 0.756	− 3.84**
No contributing action (base)				
Driver physical condition				
Drunk driving/positive blood alcoholic concentration (BAC)*	0.760	6.67**	0.760	6.67**
Driver other physical conditions	− 0.562	− 6.39**	0.231	1.33
Appeared normal (base)				
Posted speed limit (mph)				
Speed limit greater than 60	0.531	4.54**	0.909	4.84**
Speed limit 31 – 60 mph*	0.240	2.87**	0.240	2.87**
Speed limit 0 – 30 mph (base)				
Weather condition				
Rainy/snowy weather*	0.160	2.00**	0.160	2.00**
Cloudy/foggy/windy weather*	− 0.450	− 5.70**	− 0.450	−5.70**
Clear weather (base)				
Light condition				
Dark not-lighted condition*	0.205	2.71**	0.205	2.71**
Dark lighted condition*	0.018	0.25	0.018	0.25
Daylight (base)				
Type of work zone				
Maintenance work zone (short duration)*	0.248	3.58**	0.248	3.58**
Utility work zone (short duration)*	0.0360	0.44	0.0360	0.44
Construction work zone (long duration) (base)				
Constant	− 1.671	− 14.23	− 4.504	− 25.81
Summary Statistics				
Number of observations	7,855			
Likelihood Ratio (LR) chi-sq (25)	538.69			
Prob > chi-square	0.0000			
Log-Likelihood (LL) at convergence	− 5232.119			
LL at Null (Zero)	− 5501.462			
Akaike Information Criterion (AIC)	10518.24			
Bayesian Information Criterion (BIC)	10706.40			
Predictive Accuracy of the Model	71.15 %			

* Indicates that the predictor satisfies the parallel line assumption.
** Indicates parameter is significant at 0.05.
[#] *t-stats are calculated at a 95% confidence level.*

crash,' 'front end of the vehicle damaged,' 'left/right sides of the vehicle damaged,' 'improper maneuver/braking,' 'driver other actions,' 'driver other physical condition,' and 'posted speed limit in the WZ greater than 60 mph.' For these predictors, the sign of the coefficients does not uniformly indicate the direction of effect for the intermediate injury severity categories. Consequently, where the proportionality assumption does not hold, marginal effects are employed to explain the impacts on these intermediate categories. Table 4 presents the marginal effects of the PPO model as estimated for the predictors across various levels of injury severity in WZ crashes.

Table 4 reveals that the likelihood of fatalities and other injuries resulting from WZ crashes is higher for incidents involving 'two vehicles,' 'front-end of the vehicle damaged,' and 'WZ posted speed limits greater than 60 mph.' For instance, the marginal effects of the predictor 'posted speed limit greater than 60 mph' are positive for both fatal/ serious and minor/possible injury crashes (0.031 and 0.070, respectively). This indicates that the probabilities of fatal/serious and minor/ possible injuries occurring in crashes are 0.031 and 0.070 times higher, respectively, when the posted speed limit in the WZ exceeds 60 mph compared to WZs with posted speed limits under 30 mph, keeping other variables constant or at their mean. Regarding the driver's physical conditions other than 'drunk driving/positive BAC,' the likelihood of fatal/serious injuries is increased by 0.008 times, while the probability of minor/possible injuries is reduced by 0.114 times, relative to the driver's normal condition, with other predictors held constant.

The variable importance analysis highlights the individual contributions of each predictor to the model's accuracy. Fig. 4 displays the most influential predictors in the model along with their relative importance values, which are standardized on a scale from 0 to 1 and sorted in descending order. These standardized coefficients measure the impact on the response variable due to a change of one standard deviation in the predictor's value, providing a benchmark to gauge each predictor's significance in the model. As shown in Fig. 4, the predictor 'front end of the vehicle damaged' emerges as the most consequential, with a relative importance score of 1.00, followed by 'more than two vehicles involved in the crash,' which holds a significance of 0.80. Notably, two predictors of driver behavior—'drunk driving/positive BAC' and 'aggressive driving/overspeeding'—rank among the top ten in relative importance, scoring 0.62 and 0.42, respectively.

5.2. *Machine learning technique – Random Forest (RF) model*

To estimate the RF model, the WZ dataset was randomly divided into

Table 4
Marginal effects of explanatory variables for different WZ crash injury severity levels.

Variables	Marginal Effects (dy/dx) for Crash Injury Severity		
	No Injury	**Minor/ Possible Injury**	**Fatal/ Serious Injury**
Total vehicles involved in the crash			
More than two vehicles involved	– 0.132	0.108	0.024
Two vehicles involved	– 0.022	0.005	0.018
Area of the vehicle damaged			
Front end damaged*	– 0.094	0.059	0.035
Rear end damaged	– 0.041	0.033	0.007
Left/right sides damaged	– 0.022	0.002	0.019
Driver actions during the crash			
Aggressive driving/overspeeding	– 0.062	0.051	0.011
Improper maneuver/braking	– 0.005	0.017	– 0.012
Driver other actions*	0.038	– 0.012	– 0.026
Driver physical condition			
Drunk driving/positive blood alcoholic concentration (BAC)	– 0.144	0.118	0.026
Driver other physical conditions*	0.106	– 0.114	0.008
Posted speed limit (miles per hour)			
Speed limit greater than 60*	– 0.101	0.070	0.031
Speed limit 31 – 60 mph	– 0.045	0.037	0.008
Weather condition			
Rainy/snowy weather	– 0.0301	0.025	0.005
Cloudy/foggy/windy weather	0.0854	– 0.070	– 0.015
Light condition			
Dark not-lighted condition	– 0.039	0.031	0.007
Dark lighted condition	– 0.003	0.003	0.001
Type of work zone			
Maintenance work zone (short duration)	– 0.047	0.039	0.008
Utility work zone (short duration)	– 0.007	0.006	0.001

* Indicates the significant explanatory variables in the PPO model at a 0.05 significance level, which do not satisfy the proportionality assumption.

two distinct subsets: the training set, comprising 70 % (N_train = 5540 crashes), and the test set, comprising the remaining 30 % (N_test = 2315 crashes), using the 'ranger' package in RStudio statistical software. The training set was utilized to train the RF model, while the test set served as a holdout sample for predictions. A grid search with a 10-fold cross-validation (CV) process optimized the model's hyperparameters. This process involved dividing the data into ten equal parts, using nine parts for training and one part for validation to gauge predictive accuracy, with each part serving as the validation set in turn. The best model parameters included several trees (1500), variables for splitting the node mtry (10), minimum node size (15), and sample fraction (0.7). These resulted in a minimum CV error of 0.536. With these optimal parameters, the RF model achieved a training accuracy of 76.91 % and an out-of-sample prediction accuracy of 74.80 % for the test data (Table 5). This out-of-sample accuracy surpasses the 71.15 % predictive accuracy of the PPO model (Table 3), highlighting the RF model's superior performance in predicting crash severity outcomes. Additional performance metrics of the RF model, such as precision, sensitivity/recall, and F-1 score, are detailed in Table 5. Feature importance analysis revealed each predictor's contribution to model accuracy. Fig. 5 displays the top ten most influential predictors with their relative importance values on a standardized scale from 0 to 1, sorted in descending order. 'More than two vehicles involved in a crash' is the most significant predictor, with a relative importance of 1.00, followed by the predictor 'cloudy/foggy/windy weather,' with a value of 0.85. Notably, two human behavior-related predictors, 'drunk driving/positive BAC' and 'aggressive driving/overspeeding' are among the top 10 most important features, with a relative importance of 0.74 and 0.66, respectively.

The RF model yielded an overall acceptable accuracy (74.80 %); however, it underperformed in predicting the minority class of 'Fatal/Serious Injury,' with precision, recall, and F-1 score values at 0.227, 0.230, and 0.228, respectively. Given the imbalance in crash injury severity categories, with 'Fatal/Serious Injury' crashes being significantly underrepresented (only 283 out of 7,855 total WZ crashes), the RF model's performance was enhanced by applying SMOTE. The creation of synthetic samples for the minority category through SMOTE improved the RF model's prediction of 'Fatal/Serious Injury' crashes, increasing the precision, recall, and F-1 score to 0.480, 0.635, and 0.547, respectively, with an overall accuracy of 72.10 %, as shown in Table 6. Although the overall accuracy of the SMOTE-enhanced model slightly decreased (by 2.7 %) compared to the RF model without SMOTE, it significantly improved the representation of the minority category, and the performance metrics specific to the 'Fatal/Serious Injury' category notably improved: precision increased from 0.227 to 0.480, recall from 0.230 to 0.635, and the F-1 score from 0.228 to 0.547, respectively. This trade-off is often acceptable because it ensures a model that performs well across all classes rather than disproportionately favoring the majority class. Our study further revealed that variables associated with 'Fatal/Serious Injury' became more influential among the top ten predictors when SMOTE was applied, underscoring the method's effectiveness in highlighting critical predictors for severe crash outcomes, as demonstrated in Fig. 6. 'Aggressive driving/overspeeding' emerged as the most impactful predictor, with the maximum relative importance score of 1.00, closely followed by 'more than two vehicles involved,' which scored 0.94. The majority of the top ten predictors, based on their relative importance in the SMOTE-enhanced RF model, are anticipated to be significant contributors to fatal or serious injuries. Additionally, human behavior-related predictors such as 'aggressive driving/overspeeding' and 'drunk driving/positive BAC' maintain prominence, registering relative importance scores similar to those in the original RF model, specifically 1.00 and 0.75, respectively.

RF models were estimated for both sets of predictors—those following and not following the parallel line assumption—using crash injury severity as the dependent variable to identify any differences in the results due to the different nature of these predictors. In both models, SMOTE was applied to address the underrepresentation of the minority class (namely Fatal/Serious Injury). The results of these models reveal that the model inclusive of all predictors achieves the highest accuracy of 72.10 % (Table 6), followed by the model with predictors not following the parallel line assumption at 71.36 %. In contrast, the RF model estimated with predictors following the parallel line assumption resulted in the lowest accuracy of 69.72 %. For brevity, the detailed results of the RF models following and not following the parallel line assumptions are not presented in the paper.

6. Discussion

6.1. Statistical analysis – Partial Proportional Odds model

The PPO model's results and findings are discussed for each key factor associated with WZ crash severity as follows:

6.1.1. Human behavior factors

The significant and positive coefficient of aggressive driving/overspeeding suggests that such behaviors by drivers are likely to result in severe injury crashes (Table 2). This is attributed to the reduced reaction

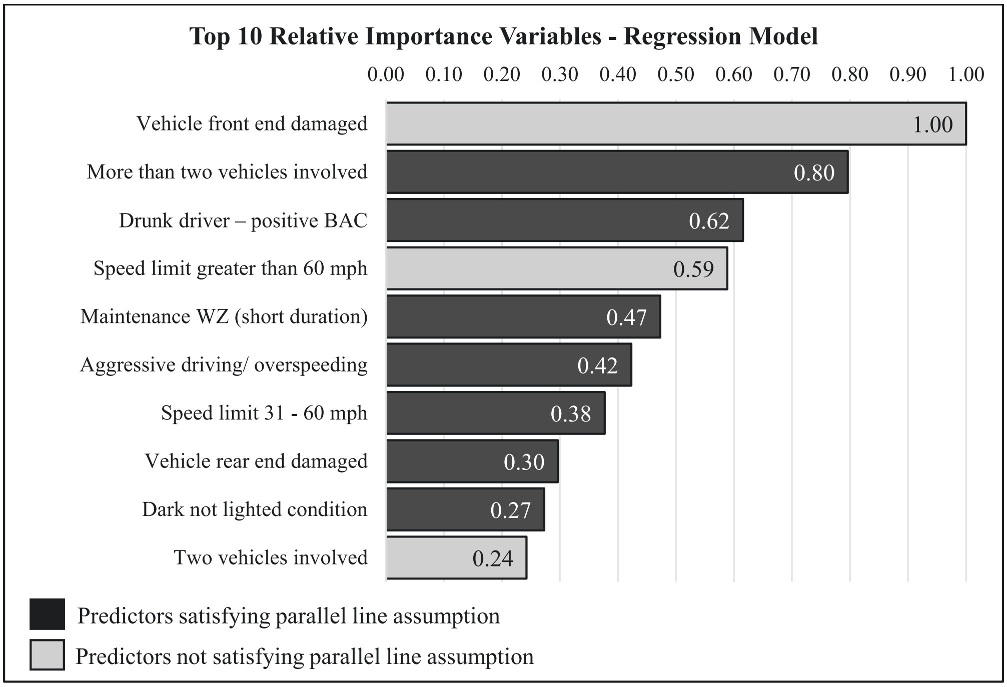

Fig. 4. Top 10 relative importance features/variables – Regression model.

Table 5
Confusion matrix for test data – Random Forest model.

Observed Outcomes	Predicted Outcomes			Total
	Fatal/Serious Injury	Minor/Possible Injury	No Injury	
Fatal/Serious Injury	17	13	44	74
Minor/Possible Injury	21	289	256	566
No Injury	37	213	1425	1675
Total	75	515	1725	2315
Performance Metrics				
Overall Accuracy	74.80 %			
Precision	0.227	0.561	0.826	–
Recall/Sensitivity	0.230	0.511	0.851	–
F-1 Score	0.228	0.535	0.838	–

time for drivers to respond to unforeseen circumstances (Zhang & Hassan, 2019b). Additionally, if the driver is under the influence of alcohol, indicated by a positive BAC, the severity of injuries is expected to increase, as alcohol impairs vision, coordination, distance judgment, and reaction times, which are further compromised by the complexities of a WZ. These findings align with those documented by other researchers (Liu et al., 2016). Fig. 4 emphasizes the significance of these two human behavior factors—'aggressive driving/overspeeding' and 'drunk driving/positive BAC'—with importance scores of 0.42 and 0.62, respectively.

6.1.2. *Other factors*

The modeling results indicate that higher posted speed limits within WZs are associated with an increased likelihood of crash injury severity. Positive coefficients in the model results imply that any posted speed limit exceeding 30 mph will likely escalate injury severity, with speeds over 60 mph posing even higher risks (Table 3). The marginal effects shown in Table 4 suggest that the probability of fatal/serious injury crashes in WZs is nearly triple when the posted speed limit surpasses 60 mph, compared to limits between 31 and 60 mph. These results are consistent with prior studies on WZ crash severity (Ghasemzadeh & Ahmed, 2019). The total number of vehicles involved in the WZ crash has also been identified as a significant determinant of injury severity. More than two vehicles involved in a crash increase the likelihood of severe injuries, confirming findings from earlier research (Khattak & Targa, 2004). Additionally, the nature of vehicle damage impacts injury severity, with front or rear-end damage likely resulting in more severe injuries (Table 3). This aligns with the understanding that head-on and rear-end collisions are typically more severe, as supported by previous studies (Khattak & Targa, 2004; Liu et al., 2016).

Weather and lighting conditions significantly influence the severity of injuries in WZ crashes. Rainy or snowy weather conditions associated with slippery roads increase the severity of injuries as compared to clear weather (Table 3), as found in past studies (Ghasemzadeh & Ahmed, 2019). While cloudy, foggy, or windy conditions may lead to more cautious driving due to reduced visibility, potentially reducing severe crashes (Zhang & Hassan, 2019b). This interpretation must be viewed cautiously, as some studies have reported increased rates of fatal crashes in foggy conditions (Ahmadi et al., 2020). Moreover, dark, not-lighted conditions are positively associated with WZ crash injury severity, indicating that severe injury crashes occur during dark conditions compared to daylight or dark-lighted conditions. The finding is intuitive as the darkness reduces visibility, leading to a shorter time for drivers to make evasive maneuvers to avoid collisions with vehicles or objects, thus increasing the likelihood of severe injury crashes, as reported in the past (Ghasemzadeh & Ahmed, 2019).

The model suggests that short-duration maintenance WZs are associated with higher injury severity compared to long-duration construction WZs, potentially due to several factors. Maintenance WZs, with their temporary nature, often catch road users off guard giving them a surprise, and generally have fewer safety measures, including less illumination at night. In contrast, construction WZs typically have more robust safety features. This is supported by the descriptive statistics in Table 1, which show that 60.78 % (172 out of 283) of fatal crashes occurred in maintenance WZs. This finding contrasts with some previous studies that report a higher incidence of fatal or serious crashes in construction WZs (Daniel et al., 2000; Weng & Meng, 2011). However, (Weng & Meng, 2011) noted that drivers are more likely to be injured in

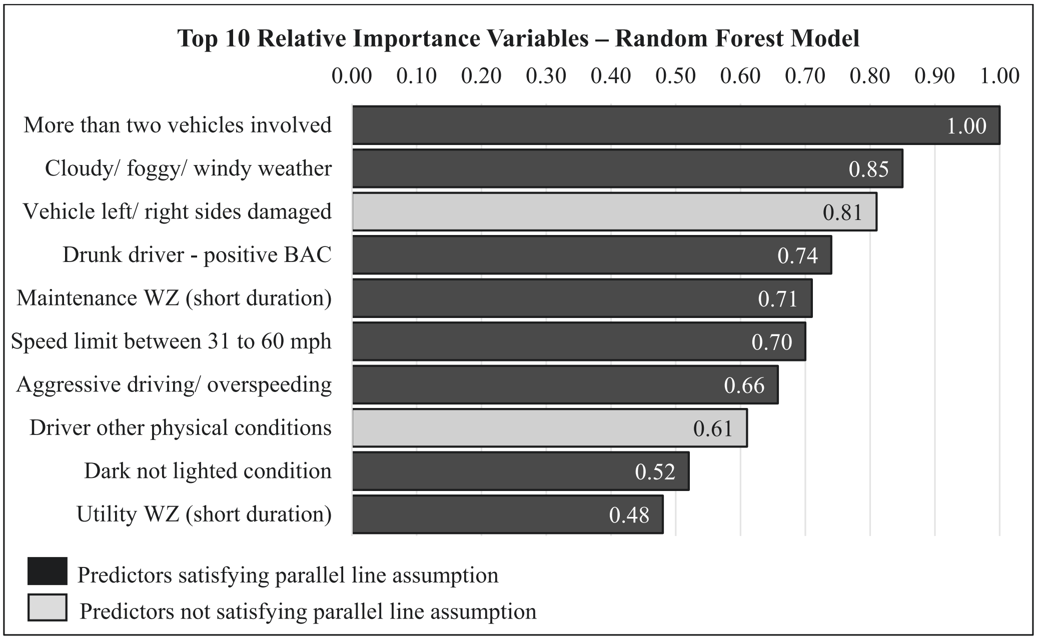

Fig. 5. Top 10 relative importance variables – Random Forest model.

Table 6
Confusion matrix for SMOTE-enhanced test data – Random Forest model.

Observed Outcomes	**Predicted Outcomes**			**Total**
	Fatal/Serious Injury	**Minor/Possible Injury**	**No Injury**	
Fatal/Serious Injury	47	9	18	74
Minor/Possible Injury	19	234	313	566
No Injury	32	256	1387	1675
Total	98	499	1718	2315
Performance Metrics				
Overall Accuracy	72.10 %			
Precision	0.480	0.469	0.807	–
Recall/Sensitivity	0.635	0.413	0.828	–
F-1 Score	0.547	0.439	0.818	–

maintenance WZs under dark conditions with illumination, whereas the risk is lower in construction WZs under similar conditions, possibly due to the use of retroreflective signs at night in construction WZs that enhance visibility and driver awareness (MUTCD, 2009). Although the current study did not find a significant association between 'Dark lighted condition' and crash injury severity in the model, this factor might still influence the observed results. This interpretation must be viewed cautiously, as variations in WZ characteristics, traffic patterns, driver behavior, and safety measures may all contribute to these differing outcomes.

6.2. *Machine learning technique – Random Forest model enhanced with the SMOTE*

SMOTE has effectively addressed the imbalance in WZ crash data. The improved performance metrics of the RF model, particularly the F-1 score for the minority class, demonstrate its efficacy (Table 6). Past studies have observed similar outcomes in other contexts (Dey & Pratap, 2023; Kuo et al., 2024; Sarkar et al., 2020). Another significant finding of the study is that SMOTE changes the importance of contributing variables, thereby making those predictors more prominent in the list of variables of relative importance, which are associated with fatal/serious injury crashes (Fig. 6). This adjustment in the importance of variables represents a substantial contribution of the study, highlighting the effectiveness of SMOTE in refining predictive models for traffic safety analysis. A comparative analysis of the three RF models post-SMOTE application reveals that the model inclusive of all variables achieves the highest accuracy at 72.10 %, followed by the model with variables that do not adhere to the parallel line assumption at 71.36 %. In contrast, the RF model estimated with variables that conform to the parallel line assumption shows the lowest accuracy at 69.72 %. This reduction in accuracy may indicate the limitations of assuming a uniform effect across all injury severity levels, potentially overlooking the varied impact of certain predictors under different conditions. Due to the complex nature of traffic and safety data, these findings underscore the importance of model selection and variable treatment in predicting crash injury severity.

The outcomes of the RF model, a machine learning technique, complement the findings of the PPO model, which employs a conventional frequentist approach. Human behavior factors, namely aggressive driving/overspeeding and drunk driving indicated by a positive BAC, have been found significant in both RF and SMOTE-enhanced RF model's feature importance analyses (Figs. 5 & 6). Additionally, most features deemed important in the RF models correspond with those identified in the PPO model, including 'more than two vehicles involved in a crash,' 'front end of the vehicle damaged,' 'dark not-lighted condition,' 'speed limits greater than 60 mph,' 'maintenance WZ (short duration),' 'rainy/snowy weather,' and 'utility WZ (short duration),' among others.

6.3. *Limitations*

The results and findings of this study are subject to several limitations inherent in the available data, which may affect the outcomes and their interpretation. Although the study employs a unique and reliable dataset for model estimation, it lacks essential variables relevant to WZ crashes, such as the traffic volume (daily or hourly) at the time of the crash and precise crash locations within the WZs. Future research should aim to collect and integrate more detailed data into the existing database, which could enhance model calibration and provide a more in-depth analysis of risk factors across various crash scenarios. Additionally, while the study demonstrates the effectiveness of SMOTE in

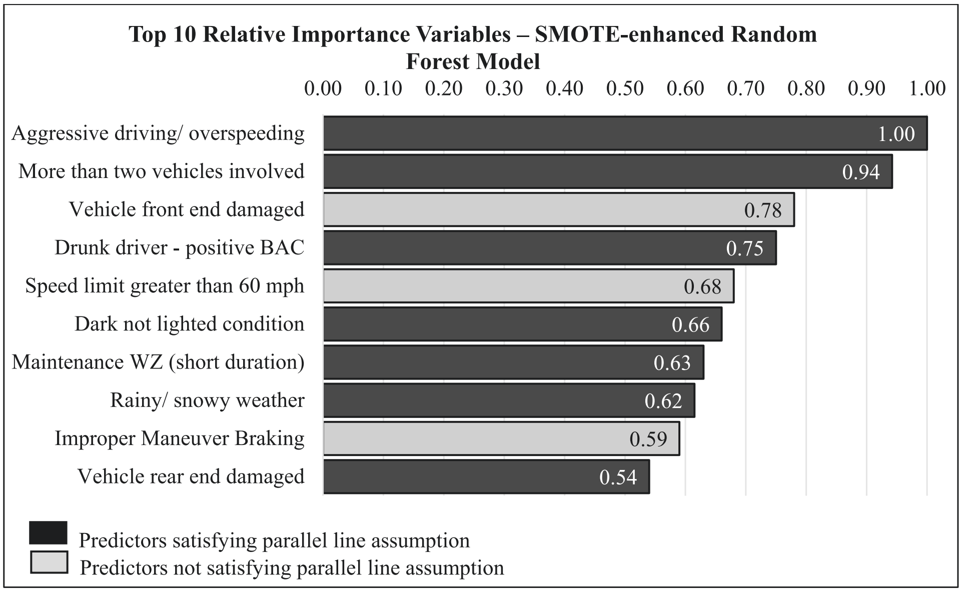

Fig. 6. Top 10 relative importance variables – SMOTE-enhanced Random Forest model.

balancing the dataset and improving the prediction performance of the RF model, the findings are based on data localized to Tennessee. Although we acknowledge that a comparative analysis with datasets from other regions could provide additional insights, such an analysis is beyond the scope of the current study. However, we believe that our models have captured generalizable patterns effectively, as they are based on five years of data collected statewide. Tennessee is a diverse state with substantial variations in transportation infrastructure, road maintenance, weather, geography, and socioeconomics. Using a comprehensive dataset from Tennessee allows for a robust analysis within the context of this state.

7. Conclusions

With the rise of road WZs necessitated by aging infrastructure and the need for maintenance and increased capacity, there is an inherent elevation in safety risks characterized by narrow lanes, irregular traffic flow, reduced speeds, and limited visibility. These factors emphasize the need to thoroughly understand the association of human behavioral factors on the entire spectrum of injury severity of WZ crashes, especially given the challenges in navigation and the emergence of new methods for data analytics. This research study focuses primarily on two aspects: firstly, to explore the key human behavioral factors responsible for WZ crash injury severity, and secondly, to effectively apply the Synthetic Minority Over-sampling Technique (SMOTE) to address the issue of data imbalance in crash injury data, primarily arising from the lower incidence of costly fatal and serious injuries. The SMOTE improves the imbalance in WZ crash data by enhancing the representation of the minority class (Fatal/Serious Injury) within the dataset. Note that accounting for the imbalance can change the relative importance of contributing variables, emphasizing the variables that are correlated with fatal and serious injuries. It is a significant finding of the study, as it ensures that the fatal and serious injuries that happen less frequently but cause more impact are not overlooked. To explore and provide useful insight into the behavioral factors associated with WZ crash injury severity, the study examines a unique dataset comprising 7,855 WZ crashes that occurred in Tennessee from 2018 to 2022 after adjusting for the underrepresented fatal and serious injuries within the dataset using SMOTE. By applying frequentist methods (statistical analysis using Ordered Logit and Partial Proportional Odds (PPO) models), and machine learning technique (Random Forest (RF) model) enhanced by the SMOTE, the study provided a multidimensional analysis of WZs crash data.

In the frequentist approach to statistical analysis, this study's findings are based on estimates from the PPO model, which mitigates the parallel line assumption and yields more insightful results than the Ordered Logit model. The outcomes from the RF model, a machine-learning approach, supplement the results of the frequentist methods. Notably, the SMOTE has demonstrated efficacy in balancing the dataset for the RF model by enhancing the representation of the minority class (fatal and serious injuries) within the dependent variable, thereby refining the model's results by emphasizing correlates of fatalities and severe injuries.

This study reveals important findings; it determines that driver behaviors such as aggressive driving, overspeeding, and drunk driving can escalate the severity of injuries in WZ crashes. Regarding the geometric features of WZs, the posted speed limit within the WZs has been recognized as a significant contributor. Injury severity escalates when the posted speed limit exceeds 30 miles per hour and further intensifies if it exceeds 60 miles per hour. Injury severity from WZ crashes intensifies when more than two vehicles are involved. Regarding vehicular damage during a crash, front and rear-end damage are more likely to result in severe injuries. The findings further reveal that environmental factors such as weather and lighting conditions are directly related to the severity of injuries in WZ crashes. Wet conditions, such as during rain or snow, are anticipated to aggravate WZ injury severity due to slippery roads. Conversely, cloudy, foggy, or windy weather correlates with lower WZ injury severity. Moreover, WZ crashes occurring in unlit (dark, not-lighted) conditions are more likely to result in severe injuries than in daylight or dark-lighted conditions due to the reduced reaction time available to drivers due to reduced visibility. Concerning the types of work zones, maintenance WZs are more prone to severe injuries compared to construction WZs.

8. Practical applications

The study, being part of the Tennessee Department of Transportation (TDOT) project, holds significant potential to enhance WZ safety at both state and national levels. It substantially contributes to understanding human behavioral factors and other elements that affect WZ safety, thereby setting a foundation for future investigations, such as assessing various WZ strategies. Applying SMOTE to address data imbalance is a key advancement. Crash databases often exhibit significant imbalance, particularly in the most costly 'minority class' data. Using SMOTE ensures more reliable analyses and outcomes. This advancement is critical for traffic safety research as it enhances our predictive capabilities and helps develop solutions based on solid data, especially for severe crashes that are infrequent but have serious consequences and high societal costs. Thus, transportation departments and agencies should consider developing tools or software with capabilities similar to SMOTE to balance data before analysis. As evidenced by this study, such tools can significantly improve the quality of traffic safety research.

The study offers valuable insights for traffic safety engineers, construction engineers, transportation agencies, and policymakers by investigating the critical human behavioral factors contributing to the severity of injuries in WZ crashes. These insights can guide the implementation of additional protective measures in urban WZs, particularly in areas with higher posted speed limits or maintenance activities. Recommended safety enhancements include deploying more signage, installing additional barriers or cones, and considering speed limit reductions where feasible. Such interventions could mitigate the severity of injuries and enhance the overall safety of all roadway users. Furthermore, law enforcement can increase efforts against drunk driving, aggressive driving, and overspeeding to prevent serious injuries, especially during adverse weather conditions or in dark, unlit WZs. Additionally, public awareness campaigns can educate the public about the dangers of unsafe behavior in WZs and the potential severe consequences of such actions.

Funding

This study was part of a project partially supported and funded by the Tennessee Department of Transportation (TDOT).

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used ChatGPT in a few instances to improve its grammar, writing, and readability. However, all authors approved the final version of the manuscript and take full responsibility for the manuscript's content.

CRediT authorship contribution statement

Muhammad Adeel: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Asad J. Khattak:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – review & editing. **Sabyasachee Mishra:** Data curation, Project administration, Resources, Writing – review & editing. **Diwas Thapa:** Data curation, Resources.

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.

Data availability

Data will be made available on request.

Acknowledgment

The authors thank the Tennessee Department of Transportation (TDOT) for providing the data for the study. The views expressed here are solely those of the authors and do not necessarily represent TDOT. Support was also provided by the Collaborative Sciences Center for Road Safety (CSCRS), a USDOT-sponsored university transportation center through Grant No. 69A3551747113.

References

Ahmad, N., Wali, B., Khattak, A.J., 2023. Heterogeneous ensemble learning for enhanced crash forecasts–A frequentist and machine learning based stacking framework. J. Saf. Res. 84, 418–434.

Ahmadi, A., Jahangiri, A., Berardi, V., Machiani, S.G., 2020. Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. J. Transp. Saf. Security 12 (4), 522–546.

Ali, Y., Hussain, F., Haque, M.M., 2024. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. Accid. Anal. Prev. 194, 107378.

Ashqar, H. I., Shaheen, Q. H., Ashur, S. A., & Rakha, H. A. (2021). Impact of risk factors on work zone crashes using logistic models and Random Forest. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC),.

Brandt, J., & Lanzén, E. (2021). A comparative review of SMOTE and ADASYN in imbalanced data classification.

Breiman, L., 2001. Random Forests. Machine Learn. 45, 5–32.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, T., Lu, Y., Fu, X., Sze, N., Ding, H., 2022. A resampling approach to disaggregate analysis of bus-involved crashes using panel data with excessive zeros. Accid. Anal. Prev. 164, 106496.

Chen, C., Zhang, G., Yang, J., Milton, J.C., 2016. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. Accid. Anal. Prev. 90, 95–107.

Crashes, W. Z. (2021). Retrieved October 28, 2023 from https://www.workzonebarriers.com/work-zone-crash-facts.html.

Daniel, J., Dixon, K., Jared, D., 2000. Analysis of fatal crashes in Georgia work zones. Transp. Res. Rec. 1715 (1), 18–23.

Demir, S., Şahin, E.K., 2022. Evaluation of oversampling methods (OVER, SMOTE, and ROSE) in classifying soil liquefaction dataset based on SVM, RF, and Naïve Bayes. Avrupa Bilim Ve Teknoloji Dergisi(34), 142–147.

Dey, I., & Pratap, V. (2023). A comparative study of SMOTE, borderline-SMOTE, and ADASYN oversampling techniques using different classifiers. 2023 3rd international conference on smart data intelligence (ICSMDI),.

Ding, H., Lu, Y., Sze, N., Chen, T., Guo, Y., Lin, Q., 2022. A deep generative approach for crash frequency model with heterogeneous imbalanced data. Analytic Methods in Accident Research 34, 100212.

FHWA. (2021). *U.S. Department of Transportation, Federal Highway Administration (FHWA), Work Zone Facts and Statistics.* Retrieved September 17, 2024 from https://ops.fhwa.dot.gov/wz/resources/facts_stats.htm.

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., Ahmed, M.M., 2019. Exploring factors contributing to injury severity at work zones considering adverse weather conditions. IATSS Research 43 (3), 131–138.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing,.

Harmon, T., Bahar, G. B., & Gross, F. B. (2018). *Crash costs for highway safety analysis.*

Hasan, A.S., Kabir, M.A.B., Jalayer, M., Das, S., 2022. Severity modeling of work zone crashes in New Jersey using machine learning models. Journal of Transportation Safety & Security 1–32.

Hasan, A.S., Kabir, M.A.B., Jalayer, M., Das, S., 2023. Severity modeling of work zone crashes in New Jersey using machine learning models. Journal of Transportation Safety & Security 15 (6), 604–635.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence),.

Joloudari, J.H., Marefat, A., Nematollahi, M.A., Oyelere, S.S., Hussain, S., 2023. Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. Appl. Sci. 13 (6), 4006.

Khattak, A.J., Khattak, A.J., Council, F.M., 2002. Effects of work zone presence on injury and non-injury crashes. Accid. Anal. Prev. 34 (1), 19–29.

Khattak, A.J., Targa, F., 2004. Injury severity and total harm in truck-involved work zone crashes. Transp. Res. Rec. 1877 (1), 106–116.

Kuo, P.-F., Hsu, W.-T., Lord, D., Putra, I.G.B., 2024. Classification of autonomous vehicle crash severity: Solving the problems of imbalanced datasets and small sample size. Accid. Anal. Prev. 205, 107666.

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

Li, Y., Fan, W.D., 2019. Modelling severity of pedestrian-injury in pedestrian-vehicle crashes with latent class clustering and partial proportional odds model: A case study of North Carolina. Accid. Anal. Prev. 131, 284–296.

Liu, J., Khattak, A., Zhang, M., 2016. What role do precrash driver actions play in work zone crashes?: Application of hierarchical models to crash data. Transp. Res. Rec. 2555 (1), 1–11.

Lunardon, N., Menardi, G., & Torelli, N. (2013). R Package`ROSE': Random Over-Sampling Examples.

Luo, S., 2023. Synthetic Minority Oversampling Technique Based on Adaptive Noise Optimization and Fast Search for Local Sets for Random Forest. Int. J. Pattern Recognit Artif Intell. 37 (01), 2259038.

Muhammad, A., Bilal, K. M., & Kamran, S. M. (2018). Work zone traffic management in rehabilitation of M-2. *Journal of Sustainable Development of Transport and Logistics*, *3* (3 (6)), 99-108.

MUTCD. (2009). Manual on Uniform Traffic Control Devices (MUTCD). *US Department of Transportation, Federal Highway Administration (FHWA)*..

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.

Osman, M., Mishra, S., Paleti, R., Golias, M., 2019. Impacts of work zone component areas on driver injury severity. Journal of Transportation Engineering, Part a: Systems 145 (8), 04019032.

Peterson, B., Harrell Jr, F.E., 1990. Partial proportional odds models for ordinal response variables. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 39 (2), 205–217.

Santos, K., Dias, J.P., Amado, C., 2022. A literature review of machine learning algorithms for crash injury severity prediction. J. Saf. Res. 80, 254–269.

Sarkar, S., Pramanik, A., Maiti, J., Reniers, G., 2020. Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. Saf. Sci. 125, 104616.

Sasidharan, L., Menéndez, M., 2014. Partial proportional odds model—An alternate choice for analyzing pedestrian crash injury severities. Accid. Anal. Prev. 72, 330–340.

Sasidharan, L., Menéndez, M., 2019. Application of partial proportional odds model for analyzing pedestrian crash injury severities in Switzerland. Journal of Transportation Safety & Security 11 (1), 58–78.

Soundrapandiyan, R., Manickam, A., Akhloufi, M., Murthy, Y.V.S., Sundaram, R.D.M., Thirugnanasambandam, S., 2023. An Efficient COVID-19 Mortality Risk Prediction Model Using Deep Synthetic Minority Oversampling Technique and Convolution Neural Networks. BioMedInformatics 3 (2), 339–368.

Sze, N., Song, Z., 2019. Factors contributing to injury severity in work zone related crashes in New Zealand. Int. J. Sustain. Transp. 13 (2), 148–154.

TDOT. (2022). *TDOT Reminds Motorists to Work with Us – Move Over, Slow Down in Work Zones.* https://www.tn.gov/tdot/news/2023/4/17/tdot-reminds-motorists-to-work-with-us—move-over–slow-down-in-work-zones.html#:~:text=In%202022%2C%20there%20were%203%2C855,see%20vehicles%20with%20flashing%20lights.

Tang, Q., Hu, X., 2024. A multi-state merging based analytical model for an operation design domain of autonomous vehicles in work zones on two-lane highways. Journal of Intelligent Transportation Systems 28 (3), 372–385.

Thapa, D., Mishra, S., Khattak, A., Adeel, M., 2024. Assessing driver behavior in work zones: a discretized duration approach to predict speeding. Accid. Anal. Prev. 196, 107427.

Usman, S.M., Khattak, A.J., Chakraborty, S., Mahdinia, I., Tavassoli, R., 2024. Detection of distracted driving through the analysis of real-time driver, vehicle, and roadway volatilities. Journal of Transportation Safety & Security 1–22.

Venthuruthiyil, S. P., Thapa, D., & Mishra, S. (2023). Towards smart work zones: Creating safe and efficient work zones in the technology era. *Journal of safety research*.

Vieira, A., Santos, B., Picado-Santos, L., 2023. Modelling Road Work Zone Crashes' Nature and Type of Person Involved Using Multinomial Logistic Regression. Sustainability 15 (3), 2674.

Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., Alharbey, R., 2021. An efficient SMOTE-based deep learning model for heart attack prediction. Sci. Program. 2021, 1–12.

Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and econometric methods for transportation data analysis. CRC Press.

Weng, J., Meng, Q., 2011. Analysis of driver casualty risk for different work zone types. Accid. Anal. Prev. 43 (5), 1811–1817.

Weng, J., Zhu, J.-Z., Yan, X., Liu, Z., 2016. Investigation of work zone crash casualty patterns using association rules. Accid. Anal. Prev. 92, 43–52.

Williams, R., 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. Stata J. 6 (1), 58–82.

Williams, R., 2016. Understanding and interpreting generalized ordered logit models. J. Math. Sociol. 40 (1), 7–20.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

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.

Yu, M., Zheng, C., Ma, C., 2020. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. Anal. Meth. Acc. Res. 27, 100126.

Yu, M., Ma, C., Zheng, C., Chen, Z., Yang, T., 2022. Injury severity of truck-involved crashes in work zones on rural and urban highways: Accounting for unobserved heterogeneity. J. Transp. Safety Secur. 14 (1), 83–110.

Zhang, K., & Hassan, M. (2019b). Identifying the factors contributing to injury severity in work zone rear-end crashes. *Journal of advanced transportation*, *2019*.

Zarei Yazd, M., Taheri Sarteshnizi, I., Samimi, A., Sarvi, M., 2024. A robust machine learning structure for driving events recognition using smartphone motion sensors. Journal of Intelligent Transportation Systems 28 (1), 54–68.

Zhang, K., Hassan, M., 2019. Crash severity analysis of nighttime and daytime highway work zone crashes. PLoS One 14 (8), e0221128.