To Balance or Not to Balance? Applying a Machine Learning Technique to Oversample Severe Injury Crashes in Work Zones

Muhammad Adeel
Graduate Research Assistant
Department of Civil & Environmental Engineering,
The University of Tennessee Knoxville, TN 37996, USA
Tel: 1-865-371-5740; Email: madeel1@vols.utk.edu
ORCID: https://orcid.org/0000-0002-4097-3953

Asad J. Khattak, Ph.D., Corresponding Author
Beaman Distinguished Professor & Transportation Program Coordinator
Department of Civil & Environmental Engineering,
The University of Tennessee Knoxville, TN 37996, USA
Tel: 1-865-974-7792; Email: akhattak@utk.edu
ORCID: https://orcid.org/0000-0002-0790-7794

Sabyasachee Mishra, Ph.D., PE
Department of Civil Engineering
University of Memphis, 3815 Central Avenue, Memphis, TN 38152
Tel: 1-901-678-5043, Email: smishra3@memphis.edu
ORCID: https://orcid.org/0000-0002-7198-3548

Diwas Thapa, Ph.D.
Department of Civil Engineering
University of Memphis, 3815 Central Avenue, Memphis, TN 38152
Tel: 1-901-691-9243, Email: dthapa@memphis.edu
ORCID: https://orcid.org/0000-0003-4747-6797

Word Count: 6,995 words text + 2 tables x 250 words (500) = 7,495 words

Submission Date: August 1, 2024

ABSTRACT

Road work zones (WZs) are increasingly common due to aging infrastructure and the need for capacity enhancement, presenting significant safety risks characterized by narrow lanes, uneven traffic flow, lower speeds, and reduced visibility. This study focuses on understanding the role of human behavioral factors in WZ crash injury severity and addressing the data imbalance caused by the lower incidence of high-cost fatal and serious injuries. A unique dataset comprising 7,855 WZ crashes in Tennessee from 2018 to 2022 was examined. The study applies the Synthetic Minority Over-sampling Technique (SMOTE) combined with a Random Forest (RF) model (a machine learning technique) to balance the dataset and improve prediction accuracy. Results indicate that aggressive driving, overspeeding, and drunk driving significantly escalate injury severity. Additionally, balancing the minority categories of crash injury severity levels (fatal and serious injuries) shifts the importance of contributing factors, emphasizing those more closely associated with higher injury categories. The application of SMOTE proved effective, significantly enhancing the prediction performance across various categories. The accuracy of the RF model improved from 71.88% to 74.36%, while the balanced accuracy increased substantially from 51.58% to 80.97%. These findings offer valuable insights for traffic safety engineers, transportation agencies, and policymakers to enhance WZ design and management. The study provides a framework for analyzing imbalanced crash data, highlighting critical behavioral factors, and recommending additional signage, speed limit reductions, and increased enforcement against unsafe driving behaviors. This approach aims to mitigate injury severity and improve road user safety in work zones.

Keywords: Work zones (WZs), Crash injury severity, Synthetic minority oversampling technique (SMOTE), Data imbalance, Random forest (RF) model, Predictive modeling, Human behavioral factors, Aggressive driving, Overspeeding, Drunk driving

INTRODUCTION

Road networks are important in providing mobility and boosting a country's economy. The aging characteristics of the highway infrastructure, especially in the US, 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 (*1; 2*). 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 (*3*). Additionally, approximately 42,151 injuries and 61,893 instances of property damage were reported, totaling 105,000 crashes in WZs in 2021 (*4*). Despite significant technological advancements that have enhanced the operations of WZs, such as smart work zones (*5*), societal concerns regarding safety and mobility near WZs persist, and Tennessee highways are no exception. In Tennessee, 3,855 WZ crashes were reported in 2022 (*6*), 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 (*7-9*) and machine learning techniques (*10; 11*) 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) (*12*). Therefore, effectively addressing the imbalance in crash data is crucial for the accuracy and overall efficacy of the model.

Considering the limitations mentioned above, 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 WZ types of WZs using a unique dataset of Tennessee WZ crashes. The study uses a machine learning (ML) technique (Random Forest algorithm) to make predictions. Importantly, proficiently implementing the SMOTE is integrated into the RF 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.

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 literature synthesis 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 (*9; 13; 14*). WZ interventions aim to curb speeding and ensure consistent traffic flow. Additionally, driver behaviors, including the age and gender of the driver (*13; 15*), alcohol impairment (*16*), and distraction or recklessness (*13; 17*), significantly influence the severity of injuries from WZ crashes. Vehicle characteristics, such as the number of vehicles involved (*10*) and the type of vehicles (light/ heavy) (*8; 9; 15; 18*), also substantially affect injury severity. Environmental factors like weather and lighting often influence WZ crash severity (*10; 13; 15; 19*). 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 (*9; 16-20*). Moreover, crash attributes such as the type of crash (*13; 18*), the nature of vehicle collisions (*21-23*), the design and duration of the WZ (*2; 7; 21*), and the presence of vulnerable road users like pedestrians (*24*) 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 the partial proportional odds model (*7-9*). More recently, machine learning (ML) techniques have emerged for the analysis of WZ crash injury severity (*10; 25-28*). Among these, the random forest (RF) 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 (*26; 27; 29-31*). The SMOTE is acknowledged as an effective method for correcting imbalanced class distributions, generating additional data for the minority class to balance the dataset (*32-35*). 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 the SMOTE.

METHODOLOGY

Data Source and 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), United States. 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 to merge the data. The initial 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 dataset of 7,855 observations representing crashes within the WZs in Tennessee.

Study Design and Methodology

Figure 1 provides an overview of the study's design. The distribution of the dataset 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 injury,' 'minor and possible injury,' 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. The Random Forest algorithm is employed in the study to predict key variables based on their significance and importance. Additionally, the application of SMOTE reinforces and enhances the RF model's predictions for the minority category (Fatal/Serious Injury) of the outcome variable (crash injury severity). The study concludes by presenting findings along with valuable recommendations (Figure 1).

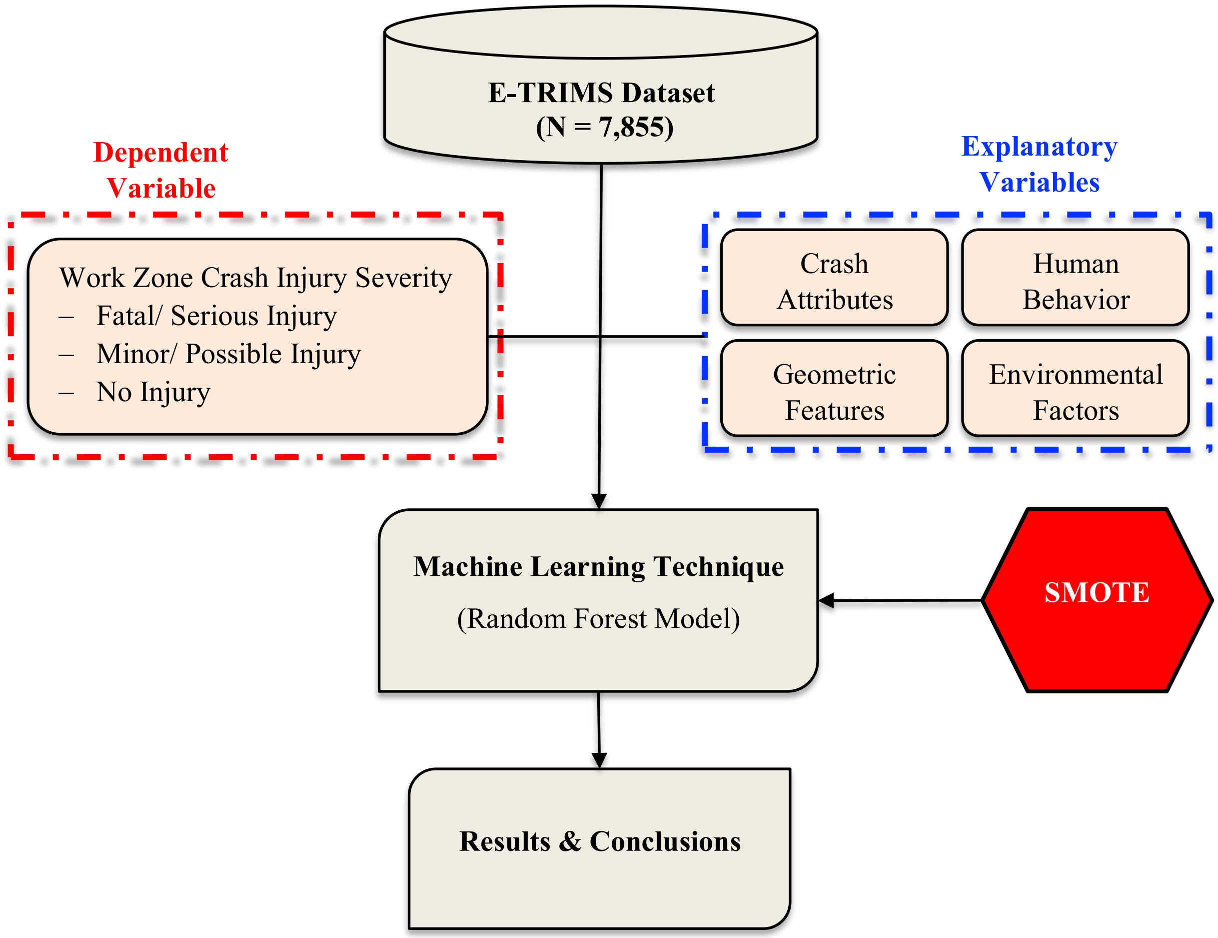

Figure 1 Design of the Study

Machine Learning Technique – Random Forest Model

Introduced by Breiman in 2001, the Random Forest 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 (*26; 27; 29-31; 36-38*). The predictive performance of the RF model is evaluated using the following metrics, with the test dataset serving as a holdout sample:

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:

$$\text{Accuracy of the Model} = \left(\frac{\text{Number of correctly classified observations}}{\text{Total number of observations}}\right) \text{ x } 100 \tag{1}$$

Balanced Accuracy of the Model

Balanced Accuracy is a metric used to evaluate the performance of a classification model, especially when dealing with imbalanced datasets. A higher balanced accuracy value indicates that the model is balanced with respect to both the minority and majority classes. It is calculated as:

$$\text{Balanced Accuracy of the Model} = \left(\frac{\text{Sensitivity} + \text{Specificity}}{2}\right) \text{ x } 100 \tag{2}$$

Cohen's Kappa Value

Cohen's Kappa is a statistical measure used to evaluate the inter-rater agreement for categorical items. It considers the agreement occurring by chance. A higher Kappa value indicates that the agreement between the raters is substantially greater than what would be expected by chance alone, reflecting more reliable and consistent ratings (*39*). It is calculated as:

$$\text{Kappa Value } (k) = \left(\frac{P_O - P_e}{1 - P_e}\right) \tag{3}$$

Where:

P_o is the observed agreement among raters (the actual agreement)

P_e is the expected agreement (agreement by chance)

Sensitivity/ Recall of the Class

Sensitivity, or recall, for a particular class is the ratio of correctly classified observations in that class to all observations of that class, calculated as:

$$\text{Sensitivity} = \frac{\text{Number of correctly classified observations in a class (True Positives)}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

Precision of the Class

Precision for a class is the proportion of correctly classified observations in that class to all correctly predicted observations, calculated as:

$$\text{Precision} = \frac{\text{Number of correctly classified observations in a class (True Positives)}}{\text{True positives for all classes}} \tag{5}$$

F-1 Score of the class

The F-1 score, which is the harmonic mean of precision and recall, is crucial in contexts with class imbalance issues. A higher F-1 score indicates better predictive performance for a specific class. This score measures how effectively SMOTE improves the model's ability to classify minority class instances. A higher F-1 score post-SMOTE indicates enhanced model performance in identifying minority class instances while maintaining precision. It is computed as follows:

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

Synthetic Minority Over-sampling Technique

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. Widely favored for rectifying imbalanced classification issues, SMOTE employs over-sampling of the minority class, capitalizing on its unique feature space. The algorithm is finely tuned to meet specific sampling needs by carefully selecting the number of neighbors. By drawing connections between minority class data points and these neighbors, SMOTE manages dataset imbalances, ensuring equitable representation of classes. The 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. 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:

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

Where u is a random number between 0 and 1, drawn from a uniform distribution $U\ (0,1)$. Recent studies indicate that SMOTE, when used with the Random Forest model, provides high accuracy compared to other methods such as ADASYN, Borderline-SMOTE, and ROSE (*40; 41*). Extensive discussions and analyses of SMOTE are present in prior research, contributing to its established reputation in the field (*32; 34; 40-45*).

RESULTS

Descriptive Statistics

Table 1 presents the descriptive statistics of the key variables in the dataset, comparing the original data of the WZ crashes with the SMOTE-enhanced data. In the original dataset, there are a total of 7,855 WZ crashes, out of which 5,656 (72.0%) are 'no injury,' 1,916 (24.4%) are 'minor or possible injury,' and 283 (3.6%) are 'fatal or serious injury' crashes. After applying SMOTE, the total number of crashes increased to 16,968, with an equal distribution across the injury severity levels: 5,656 (33.33%) for 'no injury,' 5,656 (33.33%) for 'minor or possible injury,' and 5,656 (33.33%) for 'fatal or serious injury.' Interestingly, the distribution of the explanatory variables remains similar after applying SMOTE, indicating that the technique has preserved the distribution of the original variables while addressing class imbalance. Minor changes have been observed in the distribution of various factors. For instance, in the original data, 10.0% of crashes involved more than two vehicles, which slightly increased to 10.34% in the SMOTE-enhanced data. Aggressive driving/overspeeding was reported in 12.41% of crashes originally, which increased to 14.62% after SMOTE application. Similarly, the descriptive statistics of other key explanatory variables are presented in Table 1. Figure 1 shows a graphical visualization of the sample distribution for the outcome variable 'WZ crash injury severity' across original and SMOTE-enhanced data.

TABLE 1 Descriptive Statistics for Key Variables: Original vs. SMOTE-Enhanced Data

Key Variables *(Including Dummy Variables)*	**Original Data**		**SMOTE-Enhanced Data**	
	Frequency	**Percentage (%)**	**Frequency**	**Percentage (%)**
WZ Crash Injury Severity *(Dependent Variable)*				
Total WZ **crashes**	7,855	100.00	16,968	100.00
Fatal/ serious Injury	283	3.60	5,656	33.33
Minor/ possible injury	1,916	24.39	5,656	33.33
No injury	5,656	72.01	5,656	33.33
Crash Factors				
Total vehicles involved in the crash				
More than two vehicles involved	786	10.00	1,755	10.34
Two vehicles involved	1,972	25.11	5,209	30.70
One vehicle involved	5,097	64.89	9,924	58.49
Area of the vehicle damaged				
Front end damaged	3,689	46.96	8,283	48.82
Rear end damaged	1,227	15.62	3,220	18.98
Left/ right sides damaged	1,321	16.82	2,268	13.37
Vehicle not damaged	1,618	20.60	3,175	18.71
Human Behaviour Factors				
Driver actions during the crash				
Aggressive driving/overspeeding	975	12.41	2,480	14.62
Improper maneuver/ braking	2,428	30.91	6,265	36.92
Driver other actions	1,934	24.62	3,037	17.90
No contributing action	2,518	32.06	5,063	29.84
Driver Physical Condition				
Drunk driving/ positive blood alcoholic concentration (BAC)	361	4.60	981	5.78
Driver other physical	1,403	17.86	2,200	12.97

conditions				
Appeared normal	6,091	77.54	13,911	81.98
Roadway Geometric Factors				
Roadway surface condition				
Wet road surface/ water standing	1,771	22.55	3,246	19.13
Snowy road surface	73	0.93	112	0.66
Dry road surface	6,011	76.52	13,621	80.27
Posted speed limit (mph)				
Speed limit greater than 60	625	7.96	1,974	11.63
Speed limit 31 – 60 mph	6,168	78.52	12,851	75.74
Speed limit 0 – 30 mph	1,062	13.52	2,210	13.02
Environmental Factors				
Weather condition				
Cloudy/ foggy/ windy weather	1,426	18.15	2,067	12.18
Rainy/ snowy weather	837	10.66	2,876	16.95
Clear weather	5,592	71.19	12,019	70.83
Light condition				
Dark not-lighted condition	1,089	13.86	2,541	14.98
Dark lighted condition	1,569	19.97	2,791	16.45
Daylight	5,197	66.16	11,586	68.28
Work Zone Factors				
Type of work zone				
Construction work zone (long duration)	1,532	19.50	2,570	15.15
Maintenance work zone (short duration)	4,131	52.59	9,034	53.24
Utility work zone (short duration)	2,192	27.91	5,349	31.52

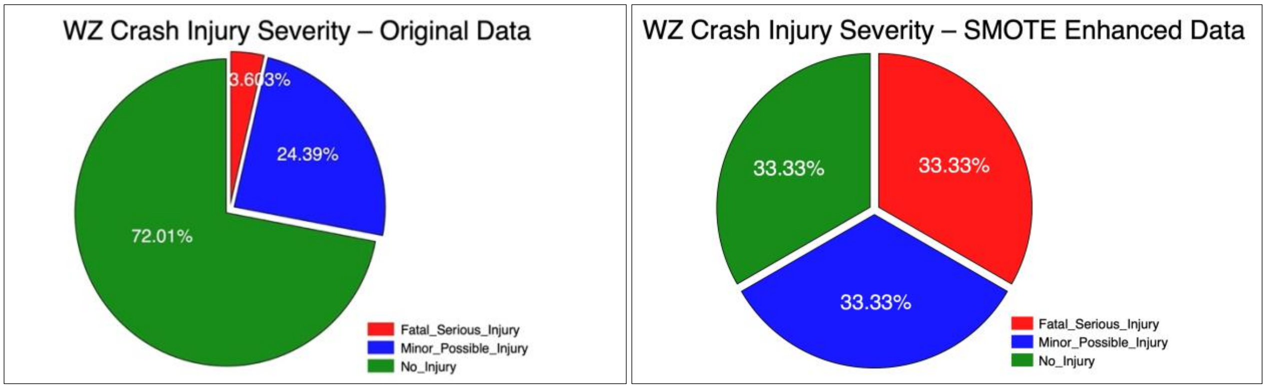

Figure 2 Outcome Variable (Crash Injury Severity): Original vs. SMOTE-Enhanced Data

Measure of Collinearity

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. A notable exception is the relationship between 'weather condition' and 'roadway surface condition,' with a higher Cramer's V of 0.606, indicating a moderate association. To explore this further, variance inflation factors (VIF) were calculated for these variables. The obtained VIF of 1.32 for both 'weather condition' and 'roadway surface condition' confirms the moderate correlation without necessitating any corrective action.

Machine Learning Technique – Random Forest Model

To estimate the Random Forest (RF) model, the WZ crashes original dataset was randomly divided into two distinct subsets: the training set, comprising 70% (N_train = 5,540 crashes), and the test set, comprising the remaining 30% (N_test = 2,315 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 10-fold validation optimized the model's hyperparameters. 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 cross-validation error of 0.536. With these optimal parameters, the model reached a training accuracy of 74.95% and an out-of-sample prediction accuracy of 71.88% for the test data (Table 2), with other performance metrics such as balanced accuracy, Cohen's kappa value, precision, sensitivity/recall, and F-1 score shown in Table 2. Feature importance analysis revealed each predictor's individual contribution to model accuracy. Figure 3 displays the list of 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 achieved an overall accuracy of 71.88% with the original data, which is acceptable. However, it significantly underperformed in predicting the minority class of 'Fatal/Serious Injury.' The precision, recall, and F1 score for this class were extremely low at 0.000, 0.009, and 0.00, respectively. These values indicate that the model failed to correctly identify instances of 'Fatal/Serious Injury.' Furthermore, the balanced accuracy and Cohen's kappa values for the RF model were 51.58% and 0.0611, respectively. These metrics further highlight the model's inadequacy in handling the minority class, emphasizing the need for improved techniques to address class imbalance. Given the significant underrepresentation of 'Fatal/Serious Injury' crashes (only 283 out of 7,855 total WZ crashes), the RF model's performance was enhanced by applying SMOTE. After applying SMOTE, the total number of crashes increased to 16,968 with an equal distribution across the injury severity levels (Table 1). The WZ crashes original SMOTE-enhanced dataset was then randomly divided into two distinct subsets: a training set comprising 70% (N_train = 11,781 crashes) and a test set comprising the remaining 30% (N_test = 5,187 crashes). The RF model applied to the SMOTE-enhanced data achieved an overall accuracy of 74.36% and a balanced accuracy of 80.97%, indicating improved performance over the original data. Precision, recall, and F-1 scores for the 'Fatal/Serious Injury' class significantly improved to 0.894, 0.950, and 0.921, respectively, demonstrating enhanced model capability in classifying this previously underrepresented class. The Minor/ Possible Injury class also saw improvements, with precision (0.772), recall (0.617), and F-1 score (0.686) being higher compared to the original data. The No Injury class saw a decrease in precision (0.578) and F-1 score (0.636), but recall (0.708) remained strong. Cohen's Kappa value improved to 0.6158, indicating a substantial agreement beyond chance as shown in Table 2.

Interestingly, after applying SMOTE, the variables associated with the 'Fatal/Serious Injury' category became more prominent among the top ten influential predictors compared to other variables, as demonstrated in Figure 4. 'Aggressive driving/overspeeding' emerged as the most impactful predictor, with a maximum relative importance score of 1.00, closely followed by 'more than two vehicles involved,' which scored 0.94. Most of the top ten predictors, based on their relative importance in the SMOTE-enhanced Random Forest model, are significant contributors to fatal or serious injuries. Additionally, human behavior-related predictors such as 'aggressive driving/overspeeding' and 'drunk driving/positive BAC' maintained their prominence, with relative importance scores of 1.00 and 0.75, respectively, similar to those in the original data RF model.

TABLE 2 Performance Metrics of Random Forest Models with Original and SMOTE-Enhanced Data

Observed Outcomes (Original Data)	**Predicted Outcomes**			**Total**
	Fatal/ Serious Injury	**Minor/ Possible Injury**	**No Injury**	
Fatal/ Serious Injury	0	1	0	1
Minor/ Possible Injury	7	46	57	110
No Injury	67	519	1618	2204
Total	74	566	1675	2315
Performance Metrics				
Precision	0.000	0.081	0.966	-
Recall/ Sensitivity	0.009	0.175	0.983	-
F-1 Score	0.000	0.111	0.975	-
Overall Accuracy	71.88 %			
Balanced Accuracy	51.58 %			
Cohen's Kappa Value	0.0611			
Observed Outcomes (SMOTE-Enhanced Data)	**Predicted Outcomes**			**Total**
	Fatal/ Serious Injury	**Minor/ Possible Injury**	**No Injury**	
Fatal/ Serious Injury	1493	23	56	1572
Minor/ Possible Injury	110	1317	709	2136
No Injury	67	365	1047	1479
Total	1670	1705	1812	5187
Performance Metrics				
Precision	0.894	0.772	0.578	-
Recall/ Sensitivity	0.950	0.617	0.708	-
F-1 Score	0.921	0.686	0.636	-
Overall Accuracy	74.36 %			
Balanced Accuracy	80.97 %			
Cohen's Kappa Value	0.6158			

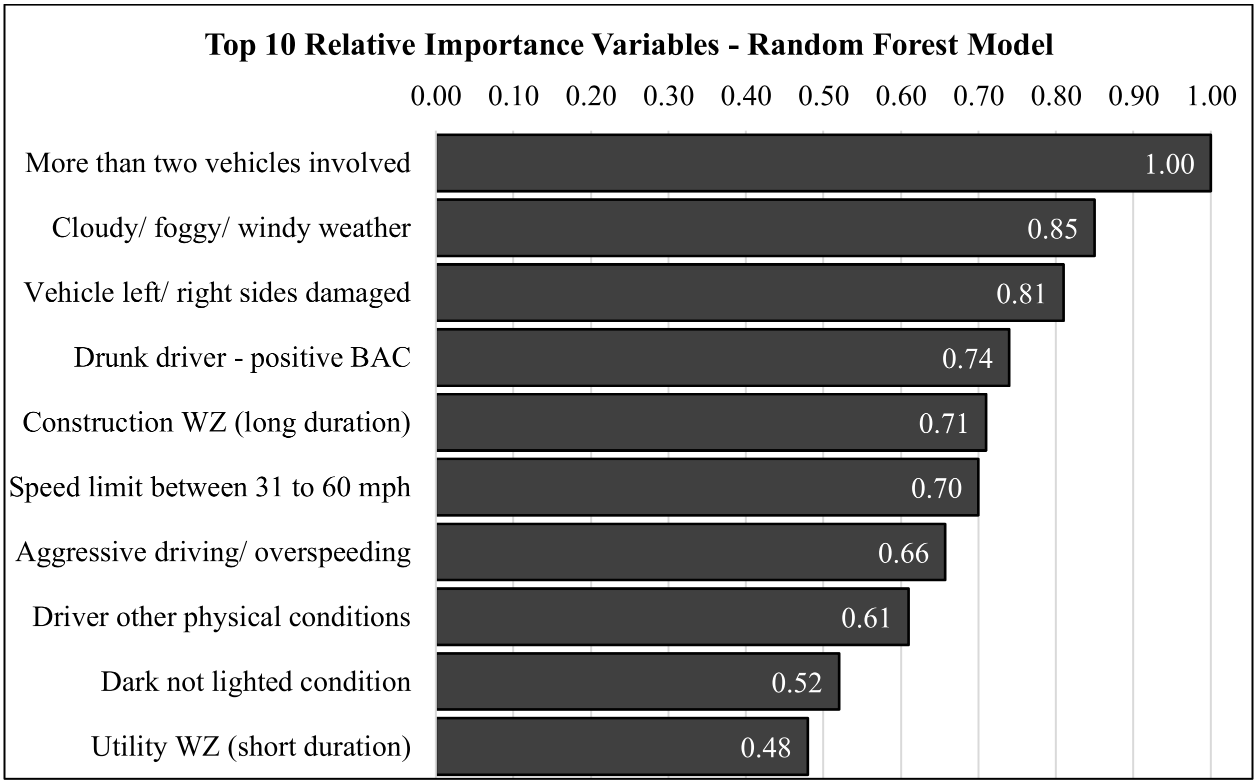

Figure 3 Top 10 Relative Importance Variables – Original Data Random Forest Model

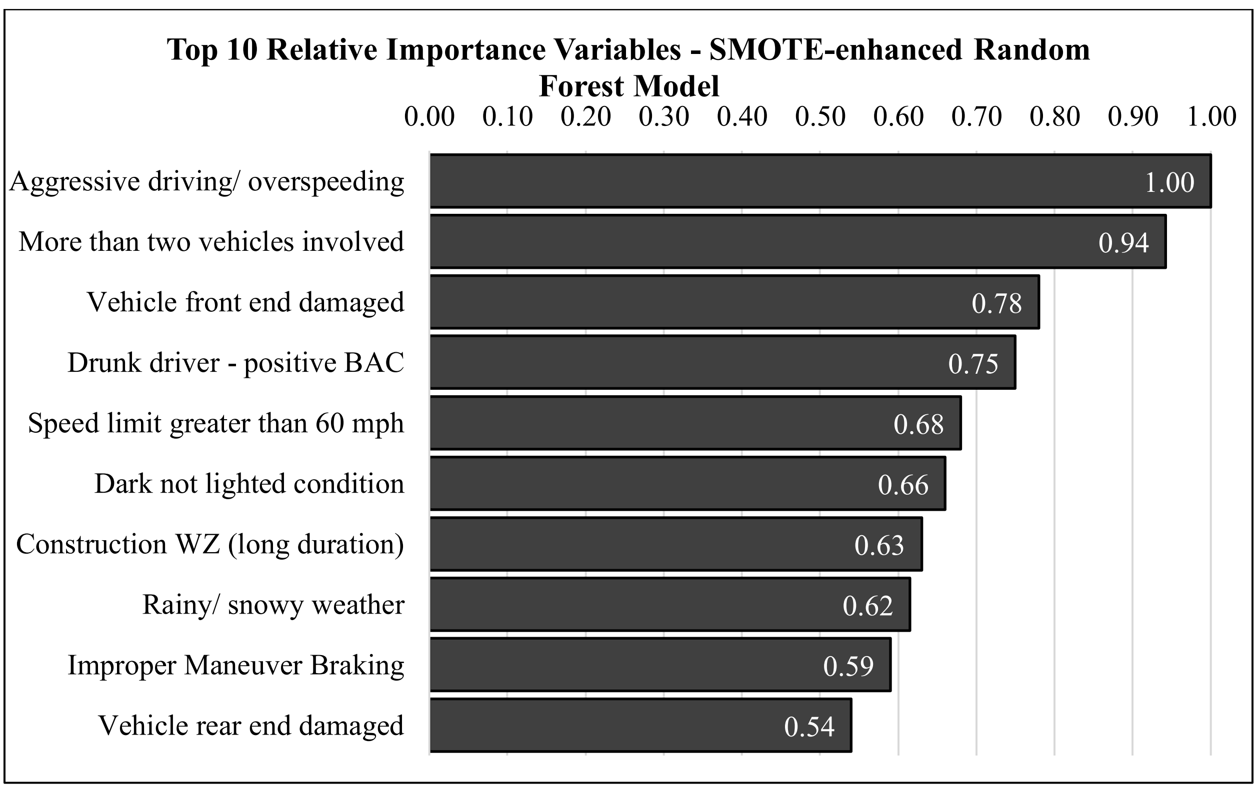

Figure 4 Top 10 Relative Importance Variables – SMOTE-enhanced Random Forest Model

DISCUSSION

The application of SMOTE has effectively addressed the imbalance in WZ crash data by enhancing the representation of the minority class (Fatal/Serious Injury) within the dataset. Interestingly, the distribution of the variables remains similar after applying SMOTE, indicating that the technique has preserved the distribution of the original variables while addressing class imbalance (Table 1). The improved performance metrics of the RF model, particularly the F-1 score of the minority class after applying SMOTE, demonstrate its efficacy. The overall accuracy improved from 71.88% to 74.36%, and the balanced accuracy increased significantly from 51.58% to 80.97%. The Cohen's Kappa value, which measures the agreement between observed and predicted classifications, also improved from 0.0611 to 0.6158, indicating better model performance after SMOTE application (Tables 2). Previous studies have observed similar outcomes in other contexts (*32; 34; 41-45*). Another significant finding of the study is that SMOTE alters the importance of contributing variables, making them more prominent in the list of variables of relative importance, particularly those closely associated with higher injury severity classes, such as fatal or serious injury crashes (Figure 4). This adjustment in variable importance significantly contributes to the existing literature on SMOTE, showcasing its effectiveness in refining predictive models for traffic safety analysis.

Human Behavior Factors

The SMOTE-enhanced RF model demonstrates superior performance metrics and predictive accuracy, making it more reliable for identifying critical factors contributing to WZ crash injury severity. Human behavior factors, namely 'aggressive driving/overspeeding' and 'drunk driving' indicated by a positive BAC, have been found significant in both the original data and SMOTE-enhanced RF models' feature importance analyses (Figures 3 & 4). This highlights the robustness of SMOTE in retaining key variable characteristics while improving the model's focus on severe injury predictors. With the highest standardized coefficient of 1.00 in the SMOTE-enhanced RF model (Figure 4), aggressive driving and overspeeding emerge as the most critical factors in crash severity. This suggests that such behaviors by drivers are likely to result in severe injury crashes due to reduced reaction time for drivers to respond to unforeseen circumstances. This finding underscores the importance of enforcing speed limits and implementing measures to curb aggressive driving behaviors in WZs (*46-49*). Additionally, if the driver is under the influence of alcohol, indicated by a positive BAC, the severity of injuries is expected to increase. 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 (*17; 50-52*).

Other Key Factors

Other significant factors observed in the SMOTE-enhanced RF model's feature importance analysis include 'more than two vehicles involved in a crash,' 'vehicle front end damaged,' 'speed limits greater than 60 mph,' 'dark not-lighted condition,' 'construction WZ (long duration),' 'rainy/snowy weather,' and 'vehicle rear end damaged' (Figure 4). The predictor 'more than two vehicles involved in a crash' indicates that if more than two vehicles are involved in a crash, the likelihood of severe injuries increases, as observed in past studies (*8; 52*). Similarly, the modeling results suggest that higher posted speed limits within WZs are associated with an increased likelihood of crash injury severity, with speeds over 60 mph posing higher risks. This result is consistent with prior studies on WZ crash severity (*8; 19; 53; 54*). Additionally, the nature of vehicle damage impacts injury severity, with front or rear-end damage likely resulting in more severe injuries (Figure 4). This aligns with the understanding that head-on and rear-end collisions are typically more severe, as supported by previous studies (*8; 17*). 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, as observed in past studies (*19; 55-57*). 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 (*19; 24; 55-57*). The model indicates that long-duration construction WZs are likely to cause more injury severity as compared to short-duration maintenance and utility WZs due to their prolonged exposure and complexity, which lead to increased driver confusion and higher chances of accidents, Additionally, extended traffic disruptions, changing environmental conditions, and driver overconfidence with repeated exposure exacerbate these risks, resulting in more severe crashes as noted by previous researchers (*57; 58*).

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 balancing the dataset and improving the prediction performance of the RF model, the findings are based on data localized to Tennessee. This regional focus may limit the generalizability of the results. The applicability of the findings to other states or regions may be constrained by differences in socio-economic, geographical, and traffic conditions. Therefore, caution should be exercised when extrapolating these results to broader contexts.

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 SMOTE to address the issue of data imbalance in crash injury data, primarily arising from the lower incidence of costly fatal and serious injuries. To provide useful insights 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. The study applies a Random Forest model enhanced by SMOTE to comprehensively analyze WZ crash data. Interestingly, SMOTE has not only proved effective in addressing the imbalance in WZ crash data by enhancing the representation of the minority class (Fatal/Serious Injury) within the dataset but it also improved the model's results by emphasizing factors associated with fatalities and severe injuries. The accuracy of the RF model improved from 71.88% to 74.36%, while the balanced accuracy increased substantially from 51.58% to 80.97%, enhancing the model's prediction performance across various categories. The Cohen's Kappa value also improved from 0.0611 to 0.6158. This refinement ensures that the infrequent but severe fatal and serious injuries are not overlooked, providing a clearer and more comprehensive picture of safety issues.

This study reveals important findings. It determines that driver behaviors such as aggressive driving, overspeeding, and drunk driving significantly escalate the severity of injuries in WZ crashes. Geometric features of WZs, like the posted speed limit, have been recognized as significant contributors, with injury severity escalating when the speed limit exceeds 60 miles per hour. Additionally, 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. Environmental factors such as weather and lighting conditions are directly related to injury severity in WZ crashes. Wet conditions, such as during rain or snow, aggravate injury severity due to slippery roads. Moreover, WZ crashes occurring in unlit (dark, not-lighted) conditions are more likely to result in severe injuries than daylight or dark-lighted conditions due to the reduced reaction time available to drivers. Long-duration construction zones are more prone to severe injuries compared to maintenance and utility zones.

The study provides valuable insights into the human behavioral factors and other elements affecting WZ safety, emphasizing the importance of using SMOTE for more reliable analyses of imbalanced crash data. Transportation agencies should consider developing tools to balance data before analysis, which can significantly improve traffic safety research quality. Additionally, insights from this study can inform the implementation of protective measures in urban WZs, such as more signage, additional barriers, and speed limit reductions. Law enforcement should increase efforts against drunk driving, aggressive driving, and overspeeding, while public awareness campaigns can educate the public on the dangers of unsafe behavior in work zones.

ACKNOWLEDGMENT

The authors thank the Tennessee Department of Transportation (TDOT) for providing the data for the study. During the preparation of this manuscript, the authors used ChatGPT in a few instances to improve grammar, writing, and readability. However, all authors approved the final version of the manuscript and take full responsibility for the manuscript's content.

AUTHORS CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: *Study Conception and Design:* Muhammad Adeel, Asad J. Khattak; *Data Collection and Preliminary Analysis:* Muhammad Adeel, Diwas Thapa; *Analysis and Interpretation of Results:* Muhammad Adeel, Asad J. Khattak; *Draft Manuscript Preparation:* Muhammad Adeel, Asad J. Khattak, Sabya Mishra, Diwas Thapa. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] Yang, H., K. Ozbay, O. Ozturk, and K. Xie. Work zone safety analysis and modeling: a state-of-the-art review. *Traffic injury prevention,* Vol. 16, No. 4, 2015, pp. 387-396.

[2] Muhammad, A., K. M. Bilal, and S. M. Kamran. Work zone traffic management in rehabilitation of M-2. *Journal of Sustainable Development of Transport and Logistics,* Vol. 3, No. 3 (6), 2018, pp. 99-108.

[3] US Department of Transportation, F. H. A. F. Work Zone Facts and Statistics. 2021.

[4] Crashes, W. Z. https://www.workzonebarriers.com/work-zone-crash-facts.html. Accessed October 28, 2023.

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

[6] (TDOT), T. D. o. T. TDOT Reminds Motorists to Work with Us – Move Over, Slow Down in Work Zones. 2022.

[7] Khattak, A. J., A. J. Khattak, and F. M. Council. Effects of work zone presence on injury and non-injury crashes. *Accident Analysis & Prevention,* Vol. 34, No. 1, 2002, pp. 19-29.

[8] Khattak, A. J., and F. Targa. Injury severity and total harm in truck-involved work zone crashes. *Transportation research record,* Vol. 1877, No. 1, 2004, pp. 106-116.

[9] Osman, M., R. Paleti, and S. Mishra. Analysis of passenger-car crash injury severity in different work zone configurations. *Accident Analysis & Prevention,* Vol. 111, 2018, pp. 161-172.

[10] Hasan, A. S., M. A. B. Kabir, M. Jalayer, and S. Das. Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety & Security*, 2022, pp. 1-32.

[11] Mokhtarimousavi, S., J. C. Anderson, A. Azizinamini, and M. Hadi. Improved support vector machine models for work zone crash injury severity prediction and analysis. *Transportation research record,* Vol. 2673, No. 11, 2019, pp. 680-692.

[12] Harmon, T., G. B. Bahar, and F. B. Gross. Crash costs for highway safety analysis. In, United States. Federal Highway Administration. Office of Safety, 2018.

[13] Zhang, K., and M. Hassan. Crash severity analysis of nighttime and daytime highway work zone crashes. *PLoS one,* Vol. 14, No. 8, 2019, p. e0221128.

[14] Thapa, D., S. Mishra, A. Khattak, and M. Adeel. Assessing driver behavior in work zones: a discretized duration approach to predict speeding. *Accident Analysis & Prevention,* Vol. 196, 2024, p. 107427.

[15] Li, Y., and Y. Bai. Highway work zone risk factors and their impact on crash severity. *Journal of Transportation engineering,* Vol. 135, No. 10, 2009, pp. 694-701.

[16] Weng, J., J.-Z. Zhu, X. Yan, and Z. Liu. Investigation of work zone crash casualty patterns using association rules. *Accident Analysis & Prevention,* Vol. 92, 2016, pp. 43-52.
[17] Liu, J., A. Khattak, and M. Zhang. What role do precrash driver actions play in work zone crashes?: Application of hierarchical models to crash data. *Transportation research record,* Vol. 2555, No. 1, 2016, pp. 1-11.
[18] Vieira, A., B. Santos, and L. Picado-Santos. Modelling Road Work Zone Crashes' Nature and Type of Person Involved Using Multinomial Logistic Regression. *Sustainability,* Vol. 15, No. 3, 2023, p. 2674.
[19] Ghasemzadeh, A., and M. M. Ahmed. Exploring factors contributing to injury severity at work zones considering adverse weather conditions. *IATSS research,* Vol. 43, No. 3, 2019, pp. 131-138.
[20] Meng, Q., J. Weng, and X. Qu. A probabilistic quantitative risk assessment model for the long-term work zone crashes. *Accident Analysis & Prevention,* Vol. 42, No. 6, 2010, pp. 1866-1877.
[21] Garber, N. J., and M. Zhao. Distribution and characteristics of crashes at different work zone locations in Virginia. *Transportation research record,* Vol. 1794, No. 1, 2002, pp. 19-25.
[22] Weng, J., S. Xue, Y. Yang, X. Yan, and X. Qu. In-depth analysis of drivers' merging behavior and rear-end crash risks in work zone merging areas. *Accident Analysis & Prevention,* Vol. 77, 2015, pp. 51-61.
[23] Yu, M., C. Zheng, and C. Ma. Analysis of injury severity of rear-end crashes in work zones: A random parameters approach with heterogeneity in means and variances. *Analytic methods in accident research,* Vol. 27, 2020, p. 100126.
[24] Sze, N., and Z. Song. Factors contributing to injury severity in work zone related crashes in New Zealand. *International journal of sustainable transportation,* Vol. 13, No. 2, 2019, pp. 148-154.
[25] Chen, C., G. Zhang, J. Yang, and J. C. Milton. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident Analysis & Prevention,* Vol. 90, 2016, pp. 95-107.
[26] Santos, K., J. P. Dias, and C. Amado. A literature review of machine learning algorithms for crash injury severity prediction. *Journal of safety research,* Vol. 80, 2022, pp. 254-269.
[27] Hasan, A. S., M. A. B. Kabir, M. Jalayer, and S. Das. Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety & Security,* Vol. 15, No. 6, 2023, pp. 604-635.
[28] Li, Z., P. Liu, W. Wang, and C. Xu. Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention,* Vol. 45, 2012, pp. 478-486.
[29] Ashqar, H. I., Q. H. Shaheen, S. A. Ashur, and H. A. Rakha. Impact of risk factors on work zone crashes using logistic models and Random Forest.In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, 2021. pp. 1815-1820.
[30] Breiman, L. Random forests. *Machine learning,* Vol. 45, 2001, pp. 5-32.
[31] Ahmad, N., B. Wali, and A. J. Khattak. Heterogeneous ensemble learning for enhanced crash forecasts–A frequentist and machine learning based stacking framework. *Journal of safety research,* Vol. 84, 2023, pp. 418-434.

[32] Waqar, M., H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey. An efficient SMOTE-based deep learning model for heart attack prediction. *Scientific Programming,* Vol. 2021, 2021, pp. 1-12.

[33] Özdemir, A., K. Polat, and A. Alhudhaif. Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. *Expert Systems with Applications,* Vol. 178, 2021, p. 114986.

[34] Luo, S. Synthetic Minority Oversampling Technique Based on Adaptive Noise Optimization and Fast Search for Local Sets for Random Forest. *International Journal of Pattern Recognition and Artificial Intelligence,* Vol. 37, No. 01, 2023, p. 2259038.

[35] Elreedy, D., and A. F. Atiya. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences,* Vol. 505, 2019, pp. 32-64.

[36] Parmar, A., R. Katariya, and V. Patel. A review on random forest: An ensemble classifier.In *International conference on intelligent data communication technologies and internet of things (ICICI) 2018*, Springer, 2019. pp. 758-763.

[37] Iranitalab, A., and A. Khattak. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention,* Vol. 108, 2017, pp. 27-36.

[38] Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems,* Vol. 26, 2013.

[39] Blackman, N. J. M., and J. J. Koval. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in medicine,* Vol. 19, No. 5, 2000, pp. 723-741.

[40] Dey, I., and V. Pratap. A comparative study of SMOTE, borderline-SMOTE, and ADASYN oversampling techniques using different classifiers.In *2023 3rd international conference on smart data intelligence (ICSMDI)*, IEEE, 2023. pp. 294-302.

[41] Kuo, P.-F., W.-T. Hsu, D. Lord, and I. G. B. Putra. Classification of autonomous vehicle crash severity: Solving the problems of imbalanced datasets and small sample size. *Accident Analysis & Prevention,* Vol. 205, 2024, p. 107666.

[42] Soundrapandiyan, R., A. Manickam, M. Akhloufi, Y. V. S. Murthy, R. D. M. Sundaram, and S. Thirugnanasambandam. An Efficient COVID-19 Mortality Risk Prediction Model Using Deep Synthetic Minority Oversampling Technique and Convolution Neural Networks. *BioMedInformatics,* Vol. 3, No. 2, 2023, pp. 339-368.

[43] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* Vol. 16, 2002, pp. 321-357.

[44] Ali, Y., F. Hussain, and M. M. Haque. Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review. *Accident Analysis & Prevention,* Vol. 194, 2024, p. 107378.

[45] Joloudari, J. H., A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain. Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences,* Vol. 13, No. 6, 2023, p. 4006.

[46] Zhang, K., and M. Hassan. Identifying the factors contributing to injury severity in work zone

rear-end crashes. *Journal of advanced transportation,* Vol. 2019, 2019.
[47] Bharadwaj, N., P. Edara, and C. Sun. Risk factors in work zone safety events: a naturalistic driving study analysis. *Transportation research record,* Vol. 2673, No. 1, 2019, pp. 379-387.
[48] Conner, K. A., and G. A. Smith. The impact of aggressive driving-related injuries in Ohio, 2004–2009. *Journal of safety research,* Vol. 51, 2014, pp. 23-31.
[49] Silveira, R., S. Ferreira, L. Cunha, and M. Rebelo. Driver behavior on road work zones: a systematic review. *Transportation research procedia,* Vol. 72, 2023, pp. 1318-1325.
[50] Miatudila Sr, A. S. Modeling the crash injury severity in work zone areas on freeways.In, The University of North Carolina at Charlotte, 2016.
[51] Martin, T. L., P. A. Solbeck, D. J. Mayers, R. M. Langille, Y. Buczek, and M. R. Pelletier. A review of alcohol-impaired driving: The role of blood alcohol concentration and complexity of the driving task. *Journal of forensic sciences,* Vol. 58, No. 5, 2013, pp. 1238-1250.
[52] Osman, M., S. Mishra, R. Paleti, and M. Golias. Impacts of work zone component areas on driver injury severity. *Journal of Transportation Engineering, Part A: Systems,* Vol. 145, No. 8, 2019, p. 04019032.
[53] Hou, Y. Speed-based safety analysis for work zones.In, University of Missouri--Columbia, 2011.
[54] Rathnasiri, N., N. De Silva, and J. Wijesundara. State of the art in work zone safety: a systematic review. *International journal of transportation science and technology*, 2023.
[55] Wei, X., X. Shu, B. Huang, E. L. Taylor, and H. Chen. Analyzing traffic crash severity in work zones under different light conditions. *Journal of advanced transportation,* Vol. 2017, No. 1, 2017, p. 5783696.
[56] Zhang, K., and M. Hassan. Injury severity analysis of nighttime work zone crashes.In *2019 5th International Conference on Transportation Information and Safety (ICTIS)*, IEEE, 2019. pp. 1301-1308.
[57] Weng, J., and Q. Meng. Analysis of driver casualty risk for different work zone types. *Accident Analysis & Prevention,* Vol. 43, No. 5, 2011, pp. 1811-1817.
[58] Daniel, J., K. Dixon, and D. Jared. Analysis of fatal crashes in Georgia work zones. *Transportation research record,* Vol. 1715, No. 1, 2000, pp. 18-23.