Accident Analysis and Prevention 151 (2021) 105851

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Ensemble-based model selection for imbalanced data to investigate the contributing factors to multiple fatality road crashes in Ghana

Mahama Yahaya [a], Runhua Guo [b], Xinguo Jiang [a,*], Kamal Bashir [c], Caroline Matara [d], Shiwei Xu [e]

[a] School of Transportation and Logistics, Southwest Jiaotong University, West Park, High-Tech District, Chengdu, China 611756; National Engineering Laboratory of Integrated Transportation Big Data Application Technology, West Park, High-Tech District, Chengdu, 611756, China
[b] Department of Civil Engineering, Suite 217, Heshangheng Bldg, Tsinghua University, 100084, Beijing, China
[c] Department of Information Technology, Karare University, Omdurman, 12304, Sudan
[d] Department of Civil and Construction Engineering, University of Nairobi, 30197, Nairobi, Kenya
[e] Guangzhou Transportation Planning Institute, 510030, Guangzhou, China

ARTICLE INFO

Keywords:
Multiple fatal injury crash
Classification
Model selection
Class imbalance
Oversampling
Ensemble classifiers

ABSTRACT

The study aims to identify relevant variables to improve the prediction performance of the crash injury severity (CIS) classification model. Unfortunately, the CIS database is invariably characterized by the class imbalance. For instance, the samples of multiple fatal injury (MFI) severity class are typically rare as opposed to other classes. The imbalance phenomenon may introduce a prediction bias in favour of the majority class and affect the quality of the learning algorithm. The paper proposes an ensemble-based variable ranking scheme that incorporates the data resampling. At the data pre-processing level, majority weighted minority oversampling (MWMOTE) is employed to treat the imbalanced training data. Ensemble of classifiers induced from the balanced data is used to evaluate and rank the individual variables according to their importance to the injury severity prediction. The relevant variables selected are then applied to the balanced data to form a training set for the CIS classification modelling. An empirical comparison is conducted through considering the variable ranking by: 1) the learning of single inductive algorithm with imbalanced data where the relevant variables are applied to the imbalanced data to form the training data; 2) the learning of single inductive algorithm with MWMOTE data and the relevant variables identified are applied to the balanced data to form the training data; and 3) the learning of ensembles with imbalanced data where the relevant variables identified are applied to the imbalanced data to form the training data. Bayesian Networks (BNs) classifiers are then developed for each ranking method, where nested subsets of the top ranked variables are adopted. The model predictions are captured in four performance indicators in the comparative study. Based on three-year (2014–2016) crash data in Ghana, the empirical results show that the proposed method is effective to identify the most prolific predictors of the CIS level. Finally, based on the inference results of BNs developed on the best subset, the study offers the most probable explanations to the occurrence of MFI crashes in Ghana.

1. Introduction

The crash injury severity (CIS) data have long been utilized as the foundation for guiding roadway and automobile designs and directing regulatory policies for enhancing road safety (Mannering and Bhat, 2014). However, the CIS database contains many variables, among which some may be redundant or irrelevant to the CIS analysis (Delen et al., 2006). Analyzing data with a large number (high dimension) of variables may not only be computationally expensive but also result in an overly complex model that has the potential to overfit the training data (Kwon et al., 2015). Of note, a model with many parameters is harder to interpret than one with few parameters; in particular, parameters that have no real association with the target variable can be misleading when the final model is interpreted (Thammasiri et al., 2014). Recently, significant safety studies are dedicated to model selection, perhaps, due to the fact that the advances in safety data collection technologies such as camera and sensor applications come with some drawbacks. The use of these technologies may further

[*] Corresponding author.
E-mail address: ejiang@gmail.com (X. Jiang).

https://doi.org/10.1016/j.aap.2020.105851
Received 27 July 2020; Received in revised form 25 September 2020; Accepted 16 October 2020
Available online 28 December 2020

intensify the number of irrelevant variables in the CIS database (Guerrero-Ibáñez et al., 2018). To identify the relevant variables for the CIS analyses, several researches have adopted machine learning (ML) classification methods (Kwon et al., 2015; Jeong et al., 2018). Regarding the variable selection, two main approaches have been identified: (i) subset evaluation and (ii) individual evaluation (Kumar and Minz, 2014). Subset evaluation generates candidate variable subsets that are useful to develop a good predictor through a search criterion. In contrast, individual evaluation methods weight separate variables according to their relative levels of importance to the given one. It is widely preferred due to its simplicity, scalability, and good empirical success (Guyon and Elisseeff, 2003). In contrast to the individual evaluation, the subset evaluation constructs and identifies variable subsets that are useful to develop an accurate predictor. In traffic safety studies, accurate prediction is one side of modelling, while the other is the variables that are relevant to explain the data. Unfortunately, subset of useful variables may omit many redundant but relevant ones (Guyon and Elisseeff, 2003). The evidence in the literature also suggests that compared to the individual evaluation, the subset evaluation can be time-consuming and computationally burdensome (Tang et al., 2014). Thus, many studies have applied inductive learning mechanism to evaluate individual variables for developing efficient CIS models (Kwon et al., 2015; Shanthi and Ramani, 2012)

Unfortunately, learning with imbalanced data is a major challenge in many data mining applications. Evidently, the CIS data are severely imbalanced: the majority of samples belong to non-multiple fatal injury (nMFI) class and multiple fatal injury (MFI) ones are rare. The standard ML algorithms applied in the data analysis assume balanced training data and are predisposed towards identifying the samples in majority, whereas the ones in minority are often misclassified. Consequently, the relevant variables selected by these methods may fail to accurately predict the minority class samples (Khoshgoftaar et al., 2010). So, class imbalance has been a subject of growing concern to safety experts recently and quite a number of studies have been conducted accordingly. For instance, (Pei et al. (2016)) applied a bootstrap resampling technique together with Poisson regression to find a more accurate and reliable classifier for the risk factors analysis. Similarly, (Zheng et al. (2019)) employed Borderline SMOTE2 (BL-SMOTE2) balancing technique together with a deep learning based CNN network for enhancing the injury severity prediction. Furthermore, (Mujalli et al. (2016)) conducted a comparative study of three different resampling methods (i.e., oversampling, undersampling, and a mix technique that integrates both). The authors found that the Bayes network developed with the synthetic minority oversampling technique (SMOTE) was more productive for the data analysis. Recently, (Schlögl et al. (2019)) conducted a comparative study of statistical learning methods to determine the factors of traffic crash occurrence based on imbalanced high resolution dataset. The authors applied SMOTE oversampling together with maximum dissimilarity undersampling technique for the data imbalance treatment. The experimental results showed a satisfactory performance in favour of the tree-based classifier. However, the above-mentioned studies developed models based on the entire set of variables. With a huge number of variables (sometimes as many as 100), the models were time-consuming, computationally intractable, and analytically complex (Delen et al., 2006).

The present study seeks to address the data imbalance problem by selecting the relevant predictors of the injury severity from high dimensional crash data for the analysis. To overcome the imbalance issue, different approaches have been proffered and can be broadly categorized into two groups (He and Ma, 2013): (i) the data-level and (ii) the algorithm-level approaches. The data level techniques use data pre-processing to guarantee a uniform class distribution. Oversampling and undersampling techniques are typical methods in this category (Chawla et al., 2002; Han et al., 2005; Yen and Lee, 2006). With the undersampling, samples from the majority class are kicked out in the manner until the desired balance is achieved. Even though undersampling-based methods are easy to implement, some informative majority samples could be lost in the process, which makes the learning task difficult (Leevy et al., 2018). For the oversampling, additional samples are generated for the minority class. Oversampling-based techniques have the advantage of improving the classification performance without causing the information loss (Tantithamthavorn et al., 2018). In this category, the synthetic minority over-sampling (SMOTE) (Chawla et al., 2002) is widely used for its ability to deal with the issue of overfitting (Fernández et al., 2017). **Thus, recent road safety studies** (Lamba et al., 2020; Yahaya et al., 2019) **applied SMOTE to address the data imbalance issue and reported improved crash injury severity prediction results. However, SMOTE poses some disadvantages linked to its blind oversampling, in which the creation of additional synthetic minority (positive) examples only consider the proximity among these examples and the number of samples of each class, whereas other features of the data are overlooked – such as the distribution of majority class samples. Thus, over the years, some more intelligent oversampling methods have been proposed to extend SMOTE and overcome the drawbacks** (Sáez et al., 2015). Among them, majority weighted minority oversampling technique (MWMOTE) has demonstrated promising results in different applications (Barua et al., 2014; Bashir et al., 2020).

The algorithm-level techniques function on the algorithms rather than the datasets, which can be further categorized into two types: (i) cost-sensitive and (ii) ensemble learning (Leevy et al., 2018). Cost sensitive approaches assign different misclassification costs to the classes involved in CIS classification task (Mujalli et al., 2016). However, it still remains a huge research challenge to obtain the optimal misclassification cost values for the individual classes (Zhang et al., 2018). In the case of the ensemble methods, several classifiers are developed on the training data and the injury classification outcome is combined to form a single composite classifier. Bagging and Boosting are commonly applied in this category. Bagging reduces the prediction error through creating numerous training sets from the given dataset. A classifier is then generated with each training set and the resultant independent CIS models are linked for the ultimate classification. Boosting also works by creating a number of training sets from the given dataset, where equal weights are initially assigned to all samples in the training data and subsequently altered according to their classification errors. The results of the individual classifiers are then combined into a unit composite classifier by a weighted approach. Compared to bagging, Freund et al. (Freund and Schapire, 1996) showed that boosting was a more effective way to obtain a small subset of variables that could predict the target concept almost as well as the complete set. To perform boosting, one commonly used algorithm is Adaptive Boosting (AdaBoost) (Freund and Schapire, 1996), in which a weak learning algorithm is compelled to modify its hypotheses according to the errors made by the previously generated hypotheses. Even though boosting has proven useful, the effect of the CIS data imbalance remains an important modelling limitation. In fact, there is a strong likelihood that ensembled classifiers generated through boosting will incline towards the non-fatal injury category that are in the majority and select variables that tend to favour same. Such learning bias can be more evident in applications where the target concept is scarce, such as the resulting outcome of multiple fatalities in a traffic crash. To account for the class imbalance problem associated with CIS data, a considerable portion of research has utilized oversampling techniques at the data level (Mujalli et al., 2016; Schlögl et al., 2019) and sampling methods at the algorithm level (Mafi et al., 2018) in imbalance data classification framework to model injury severity outcomes. **In a related study in the context of real-time crash prediction on expressways, Cai** et al. (Cai et al., 2020) **utilized the deep convolutional generative adversarial network (DCGAN) model to generate synthetic samples to address the crash data imbalance problem. The authors reported a better prediction accuracy in favour of the convolutional neural network model based on the DCGAN balanced data compared to SMOTE.**

The idea of ensemble learning has been extended to imbalanced data classification (Li, 2007) and variable selection (Saeys et al., 2008). **Even though MWMOTE and AdaBoost methods have been deployed in separate studies, the combined use of these methods has not been investigated in the previous studies as far as we know. In light of the above, the paper proposes a framework that integrates data resampling with boosting to identify the relevant determinants of MFI crashes. The proposed approach offers a unified framework that integrates ensemble feature selection and multiple sampling in a mutually beneficial way. Both methods operate on a uniform principle of first identifying the hard-to-learn informative minority samples and assigning them weights in some way to ease learning. Thus, the research seeks to explore the synergistic value of the proposed framework for feature selection in the context of CIS analysis. Specifically, MWMOTE and AdaBoost.M1 algorithms are integrated to address the data imbalance and create ensemble of classifiers for evaluating and ranking the individual predictors. The approach is implemented on several years of crash data in Ghana to identify the risk factors of multiple fatal injury (MFI) crashes. Although MFI crashes are socio-economically burdensome, few studies are available to unravel the most probable factors of their occurrence, perhaps due to insufficient data and the associated methodological limitations. We hypothesize that the bias imposed by the data skew coupled with the weakness of single inductive learning may obscure the identification of the useful predictors of CIS.**

Therefore, the main objective of the study is threefold: 1) to develop a model selection framework for the imbalanced CIS data, 2) to evaluate the learning impact of data imbalance for variable selection in the context of CIS analysis, and 3) to identify the contributing factors to multiple fatality crashes in Ghana.

The rest of the paper is organized as follows: next section presents the data and the statistical methodology on which the model selection is based. Section 3 describes the experimental framework that encompasses the detailed description of the experimental setup and settings and the evaluation metrics. Section 4 shows the empirical results. Finally, discussions and conclusions are summarized in section 5.

2. Methodology

2.1. System overview

A systematic machine learning framework known as ensemble based variable selection (EbVR) is proposed for variable selection as shown in Fig. 1. To begin with, the original imbalanced crash data is first obtained and rebalanced with MWMOTE algorithm. Ensembles of classifiers generated by boosting technique (AdaBoost) are induced from the balanced data to evaluate the importance of individual variables with respect to the CIS level. The selected variables are applied to the balanced data to form the training dataset for the CIS modelling. The application potential of two inductive algorithms is tested for the base learning in the ensemble algorithm. These include decision stump (DSt) and the multinomial logistic regression with a ridge estimator (MLRE).

For the comparative study, three existing variable evaluation techniques demonstrated in Fig. 1 as S1, S2 and S3 are considered: S1) a single learning algorithm used based on original imbalanced data and selected variables applied to the original imbalanced data to form the training set (Kwon et al., 2015; Shanthi and Ramani, 2012); S2) ensemble of learning algorithms induced from the imbalanced data and selected variables applied to the original imbalanced data to form the training set (Saeys et al., 2008); and S3) a single learning algorithm induced from balanced data and selected variables applied to imbalanced data to form the training set (Mujalli et al., 2016).

The selection subsets corresponding to the ranking methods are used independently to train the Bayesian networks (BNs) CIS classification models and the predictions captured in various metrics for the comparative analysis. We determine the optimal number of variables for each selection subset by progressively adding the top ranked variables of the selection schemes to the BN classifier until the addition of the next variable ceases to improve the model performance. One may argue that developing the prediction model based on the oversampled data is enough to improve the performance of the prediction model regardless

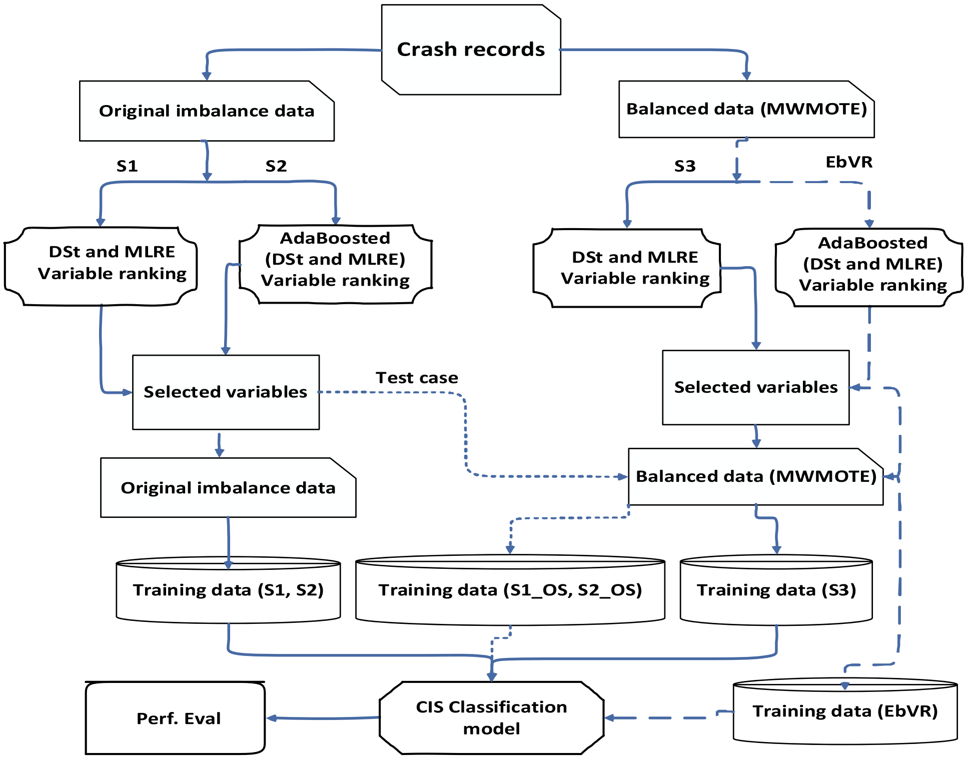

Fig. 1. The study framework.

of the quality of variables selected. **To further test the claim and validate our study hypothesis, variables selected by single and ensemble classifiers based on imbalanced data are applied to balanced data to form separate training sets, i.e., S1_OS and S2_OS, respectively, for the classification modelling (shown in dotted lines).** Finally, the injury severity prediction model is developed using cross validation strategy and the performance captured in several measures for evaluation. These steps are diagrammatically represented in Fig. 1.

2.2. *Data*

The traffic crashes between January 2014 and December 2016 in Ghana are obtained from the Building and Road Research Institute (BRRI). Each record consists of factors linked to the situations at the time of the crash, such as:

- Roadway: roadway features such as horizontal and vertical alignments, pavement surface condition, pavement type, shoulder type, shoulder condition, and location type
- Environmental characteristics: lighting, weather, and roadway conditions
- Crash characteristics and outcome severity: contributing factors such as the nature of the collision, collision pattern, and the number of vehicles involved. The crash severity is determined according to the level of injury sustained by the worst affected person in line with the practice (Han et al., 2005; Yen and Lee, 2006).

2.2.1. *Data preprocessing*

The original data are made up of 27,602 records with 42 variables. We remove the records with the critical missing values, so 27,387 are left for the analysis. To detect the main influential variables of crash severity prediction, eighteen independent variables are analyzed (Table 1). The variable selection is based on the completeness of records for the given variable in the original data and recommended in the literature (Leevy et al., 2018; Tantithamthavorn et al., 2018). **Out of the eighteen variables, thirteen are used as they appear in the original data. Others such as time, date, number of vehicles and day of the week are discretized and reduced from multiple to binary class levels to facilitate the implementation of BNs. For example, date and time variables are transformed into two levels and relabelled as seasons (i.e., rainy & dry) and time (i.e., peak & off-peak), respectively. Similarly, the number of vehicles is binarized to include one and more than one.** As the study aims to find the factors contributing to multiple fatality crashes, we binarize the variable depicting the number of casualties into "multiple fatality crashes where the number of fatalities for the given crash exceeds one (MFI)" = 1 and "non-multiple fatality crashes in which the number of fatalities is one or less (nMFI)" = 0. Binarizing the original data eventually results in a sample distribution of 98:2 (Table 1) for the nMFC and MFC, respectively. This huge proportion difference could cause problem in the statistical modelling.

2.3. *Variable ranking*

The principle of variable ranking is to select variables based on their individual prediction capacity. Traditionally, a classifier is developed on a given variable to estimate its prediction capacity with respect to the class variable. However, this may not be appropriate under highly imbalanced data due to the classification bias and other limitations of the inductive learning algorithm.

2.3.1. *Our proposed approach*

Consider an input dataset $D\{(x_1, y_1), ..., (x_m, y_m)\}$, $x_i \in X$, with

Table 1
Variable description and classification according to injury severity.

CODE: Variable	Categories	Description	Count	Crash severity	
				nMFI	MFI
N_VEH: Number of Vehicles involved	0	less than 2 vehicles	12,028	98.30%	1.70 %
	1	2 or more vehicles	15,359	97.90%	2.10 %
CSON: Season	0	Dry season	11,332	98.10%	1.90 %
	1	Rainy season	16,055	98.10%	1.90 %
D_WK: Day	0	Weekend	8297	97.77%	2.23 %
	1	Week day	19,090	98.20%	1.80 %
TME: Time	0	Non peak	9682	97.69%	2.31 %
	1	Peak	27,387	98.07%	1.93 %
WTHR: Weather	1	Clear	22,747	98.33%	1.67 %
	2	Fog/Mist	96	92.71%	7.29 %
	3	Rain	188	97.87%	2.13 %
	4	Dust/Smoke	61	100.00 %	0.00 %
	5	Dazzle	94	94.68%	5.32 %
	6	Others	4201	96.83%	3.17 %
LT_C: Light condition	0	No lighting	4657	97.12%	2.88 %
	1	Lighting	22,202	98.27%	1.73 %
RDES: Road description	1	Straight and flat	24,135	98.35%	1.65 %
	2	Curve only	1480	96.35%	3.65 %
	3	Incline only	541	96.12%	3.88 %
	4	Curve & Incline	1194	95.56%	4.44 %
	5	Bridge (name river)	37	94.59%	5.41 %
RSURF_T: Road surface type	1	Tar Good	22,228	98.22%	1.78 %
	2	Tar with Potholes	4244	97.60%	2.40 %
	3	Gravel	115	96.52%	3.48 %
	4	Earth with potholes	800	96.75%	3.25 %
SLD_C: Shoulder condition	1	Tared	14,333	97.54%	2.46 %
	2	Untarred	1637	96.46%	3.54 %
	3	No Shoulder	11,417	98.97%	1.03 %
RSEP: Road separation	1	Median	8005	99.21%	0.79 %
	2	No median	19,382	97.60%	2.40 %
SURF_C: Surface condition	1	Dry	27,119	98.09%	1.91 %
	2	Wet	224	95.98 %	4.02 %
	3	Muddy	44	95.45%	4.55 %
LOC_T: Location type	1	Not at junction	21,241	97.74%	2.26 %
	2	Plus-intersection	1587	99.68%	0.32 %
	3	T-intersection	3497	98.91%	1.09 %
	4		277	99.64%	

(*continued on next page*)

Table 1 (*continued*)

CODE: Variable	Categories	Description	Count	Crash severity	
				nMFI	MFI
		Staggered intersection			0.36 %
	5	Y-intersection	76	100.00 %	0.00 %
	6	Roundabout	412	99.51%	0.49 %
	7	Railway crossing	6	100.00 %	0.00 %
	8	Other	291	98.97%	1.03 %
TRF_C: Traffic control	1	None	18,736	97.77%	2.23 %
	2	Pedestrian crossing	157	98.73%	1.27 %
	3	Signals	3323	99.64%	0.36 %
	4	Stop sign	60	98.33%	1.67 %
	5	Give Way	352	99.15%	0.85 %
	6	Other	4759	98.07%	1.93 %
COL_T: Collision type	1	Head on	2762	93.41%	6.59 %
	2	Rear end	6250	99.02%	0.98 %
	3	Right Angle	1743	99.43%	0.57 %
	4	Side Swipe	3478	99.11%	0.89 %
	5	Ran off road /Overturn	3823	96.68%	3.32 %
	6	Hit Object on road	218	98.62%	1.38 %
	7	Hit Object off road	789	98.23%	1.77 %
	8	Hit parked vehicle	935	96.90%	3.10 %
	9	Hit Pedestrian	7059	99.01%	0.99 %
	10	Hit Animal	81	98.77%	1.23 %
	11	Other	249	100.00 %	0.00 %
HT_RN: Hit and Run	1	Not Hit & Run	26,278	98.01%	1.99 %
	2	Hit & Run	1109	99.55%	0.45 %
RD_WKS: Road works	1	Not at Roadworks	27,145	98.10%	1.90 %
	2	At Roadworks	242	97.10%	2.90 %
RD_EVMT: Road environment	1	Urban	17,306	99.16%	0.84 %
	2	village	1206	96.77%	3.23 %
	3	Rural/no settlement	8875	96.12%	3.88 %
CIS: Crash severity	0	nMFI	26,859	98.100 %	
	1	MFI	528		1.90 %

class labels $y_i \in Y = \{1, \ldots., K\}$ where $K_m < K$, relates to the minority class. To deal with the classification challenge associated with the imbalanced data, we incorporate data pre-processing in two stages. At the data level, we balance the training data using MWMOTE oversampling algorithm. The oversampling process is carried out in three main stages as follows: firstly, the most relevant and hard-to-learn minority class samples are chosen from the original minority to build a set of important minority samples; secondly, different weights are given to the selected minority based on their relevance to the learning task; finally, additional synthetic samples are created from the weighted informative minority class samples by a clustering technique to augment the minority class. The process outputs a balanced dataset $D^{'}$.

It is noted that learning algorithms that output only a single hypothesis are faced with statistical, computational, and representational challenges (Dietterich, 2002).The resulting hypothesis h_t induced from $D^{'}$ may present a high degree of training error ε_t that may impact the classification decision. To account for the effects, we induce the ensemble learning (Freund and Schapire, 1996) from the balanced data $D^{'}$, where a weak algorithm is repeatedly invoked to learn from various weighted distributions of the training data. The individual predictions of the resulting classifiers are combined through a voting scheme to form the final hypothesis for the variable evaluation. For instance, consider the training set $D^{'} = [(x_i, y_i), \ldots(x_m, y_m)]$, where a crash record x_i represented as a vector of variable values, and y_i εY is the injury severity level (in our case, Y = 2) associated with x_i. AdaBoost generates different weighted distributions of the datasets and train any weak learner h_t repeatedly in several rounds $t - T$. In each round t, h_t focuses on the hard-to-learn samples to find the one that reduces the training error ε_t. The predictions of the individual classifiers developed thereof are combined through a voting scheme to form a more robust hypothesis h_{fin} expressed as follows:

$$h_{fin}(x) = \underset{y\in Y}{argmax} \sum_{t=1}^{T} \left(log \frac{1}{\beta_t} \right) h_t(x, y) \tag{1}$$

$\beta_t \in \{0, 1\}$ is calculated as:

$$\beta_t = \varepsilon_t / (1 - \varepsilon_t) \tag{2}$$

The training error ε_t of given h_t is expressed as:

$$\varepsilon_t = \frac{1}{2} \sum_{(i,y)\in B} D_t^{'}(i, y)(1 - h_t(x_i, y_i) + h_t(x_i, y_i)) \tag{3}$$

Here, B is the set of all mislabels and $D_t^{'}$ represents the distribution of misclassified samples. The proposed approach is expected to ultimately reduce the inherent classification error associated with h_t. Hence, the relevant variables identified in the process may favour all classes in the CIS classification task. The proposed ensemble-based variable ranking is referred to as EbVR. To verify its suitability under different learning assumptions, the decision stump (DSt) and multinomial logistic regression with ridge estimator (MLRE) algorithms are independently used for induction in the EbVR framework. Even though decision tree algorithms are unstable under the least data perturbation, they are still preferred as the base classifiers in ensemble learning as they are fast to train (Dietterich, 2000). This is one important advantage of DSt to the proposed method as we need to evaluate individual variables using multiple classifiers in a more efficient manner. Also, it is shown that decision tree and logistic classifiers are susceptible to the data imbalance (Liu et al., 2010; Muchlinski et al., 2016). MLRE algorithm applies a ridge estimator to reduce overfitting by penalizing the large coefficient based on the paper (Le Cessie and Van Houwelingen, 1992). Thus, using these classifiers in our experimental setting has both theoretical and practical significances.

2.3.2. *Score function*

In the variable ranking, inductive learning algorithm is commonly used to generate the variable importance indices. Consider a set of m samples $[x_k, y_k]$ $(k = 1, \ldots m)$ comprising of n input variables $x_{k,i}$ $(i = 1, \ldots n)$ and one output variable y_k. Variable ranking utilizes a scoring function $\varphi(i)$ calculated from the values $x_{k,i}$ and y_k, $k = 1, \ldots m$. When a single inductive algorithm is applied, the predictions of the classifier for each sample are sorted directly and the variable importance measure estimated according to a chosen performance metric. Since we are dealing with the imbalanced classification, the area under ROC curve (AUC) is a more reliable performance measure (Mujalli et al., 2016).

Regarding the ensemble classifier, the predictions for each sample in various weighted distributions are brought together and normalized among all base classifiers and the AUC value. Accordingly, we define the worth of a variable as follows:

$$Score(x_i) = AUC\left\{ \underset{y_i \in Y}{\operatorname{argmax}} \sum_{t=1}^{T} \left(log \frac{1}{\beta_t} \right) h_t(x_i, y_i) \right\} \tag{10}$$

where the function $AUC()$ estimates the AUC value.

2.4. Bayesian network classification

BN is a directed acyclic graph which allows inference to be made based on the causality or direct dependence among variables (Mujalli et al., 2016). Given a training dataset, BN classifier establishes the conditional probability of each variable and the corresponding injury severity level. Classification is then conducted by applying Bayes rule to estimate the likelihood of a severity level for specific instances of a given variable and then selecting the injury level with the highest posterior probability.

3. Experimental Setup

3.1. Algorithms and parameter settings

Several software tools are employed to generate and examine the empirical results, such as WEKA 3.6.13 (Hall et al., 2009), KEEL (Alcalá-Fdez et al., 2009), and R statistics program. For instance, we utilized the “imbalance” package (Barua et al., 2014) for the MWMOTE oversampling where the number of synthetic minority samples required to balance the data is estimated and generated. **In resampling, we followed the approach proposed in studies** (Mujalli et al., 2016)**, in which the additional synthetic samples are generated such that the minority class is augmented to the size of the majority. Thus, in the study the ratio used in the MWMOTE oversampling is 50:50 for the majority and minority classes.**

In the study, the ClassifierAttributeEval (WEKA function) is employed to implement the proposed method and ensembles of the base classifiers (i.e., DSt and MLR) developed on the balanced data are used to rank the CIS variables independently. The inductive algorithm is invoked in a series of 10 rounds for the ensembles, creating a classifier in each round for the final prediction. The ultimate ranking score is then estimated based on the average AUC of the k-folds cross validated (CV) results. For this experiment, we choose $k = 10$. Extensive experimental studies on different datasets, with diverse learning approaches, have demonstrated that 10 is nearly the best number of partitions to obtain a reliable error estimate (Witten et al., 2016).

To further assess the variable selection schemes, the BN injury severity classification models are developed on the corresponding selection subsets by the schemes. For more reliable BN prediction results, we conduct the Simulated Annealing search in 10,000 iterations with a start temperature and δ values of 10 and 0.999, respectively. Finally, the BN model result is calculated by 10-fold CV method. Here, the essence of CV is to assess how well the CIS model can generalize when applied to independent dataset. To score the network structure, Akaike Information Criterion (AIC) is chosen. Statistically potent subsets are also determined based on the output AIC values. Note that the lower values point to a better statistical fit. It is important to point out that the above settings for the BN have been proven effective in the literature (Mujalli et al., 2016).

3.2. Performance metrics

Since the synthetic examples generated belong to one class, it is practically infeasible to measure the quality in isolation. However, the widely used approach of examining data quality in machine learning domain, which is adopted in the paper to assess the quality of the various datasets, including the original and oversampled datasets, has been demonstrated in Fig. 1 **as "Perf. Eval." For the performance evaluation, four widely-used metrics are considered in this paper, including accuracy, recall, precision, and the receiver operating characteristic curve area (AUC)** (Fawcett, 2006). These measures are calculated as follows:

$$Accuracy = \left(\frac{tMFI + tnMFI}{tMFI + tnMFI + fMFI + fnMFI} \right) 100\% \tag{11}$$

$$True\ positive\ (TP) rate = recall = \frac{tMFI}{tMFI + fnMFI} \tag{12}$$

$$False\ positive\ (FP) rate = \frac{fnMFI}{tMFI + tnMFI} \tag{13}$$

$$Precision = \frac{tMFI}{tMFI + fMFI} \tag{14}$$

Where tMFI is true multiple fatal injury instances, tnMFI indicates true non-multiple injury cases, fMFI is false multiple fatal injury cases, and fnMFI defines false non-multiple fatal injury cases.

The accuracy measure has been criticized extensively in many studies (Chawla et al., 2002; Hossin and Sulaiman, 2015), especially for the classification tasks with the imbalanced data. It is self-evident that the CIS datasets are severely unbalanced with extremely few samples for the MFI (1.9 % of total samples) compared to the samples in the other class (98.1 %). Consequently, it is possible to develop a CIS model that yields an overall accuracy of over 98 % for nMFI severity crashes but a negligible accuracy of 2% for MFI ones. Such a model offers inadequate insight into how MFI severity crashes occur.

Therefore, the receiver operating characteristic curve (ROC) is typically used to provide unbiased assessment of ML algorithms in the imbalanced learning (Leevy et al., 2018). For every possible classification threshold, the ROC curve shows the variation in the number of correctly classified positive (MFI) cases and the number of incorrectly classified negative cases (nMFI). To compare classifiers, a ROC curve is often reduced to a single scaler quantity such as the area under ROC (AUC), which has a maximum value of 1 describing a perfect prediction and a value of 0.50 corresponding to a prediction that is not better than a mere guess. For AUC, one important statistical property of interest in this study is that of its equivalence to the probability of the classifier to rank a randomly selected positive sample (MFI) higher than a randomly selected negative sample (nMFI) (Fawcett, 2006).

Further, we apply the Aligned Friedman test (García et al., 2010) to compare the variable selection techniques in a pair-wise manner. Accordingly, the set of ranks that represent the effectiveness associated with each technique and the p-value related to the significance of the differences will be estimated. Here, it is noted that lower rank values indicate better results. The confidence of 95 % is used under the null hypothesis that the performance difference for any paired variable selection methods is not statistically significant.

4. Results

4.1. Variable rankings

Fig. 2 ranks the relative importance of the injury severity risk factors according to the scores estimated by the proposed EbVR and the compared methods. It should be noted that, compared to DSt, MLRE induction to the schemes, especially, the proposed one, yields slightly higher AUC values for the variables evaluated. In general, the top three most important variables including collision type, road environment, and shoulder type, remain the same across all the ranking schemes. However, some changes in the variable importance scores for the various ranking methods have been noted. With EbVR, the variables

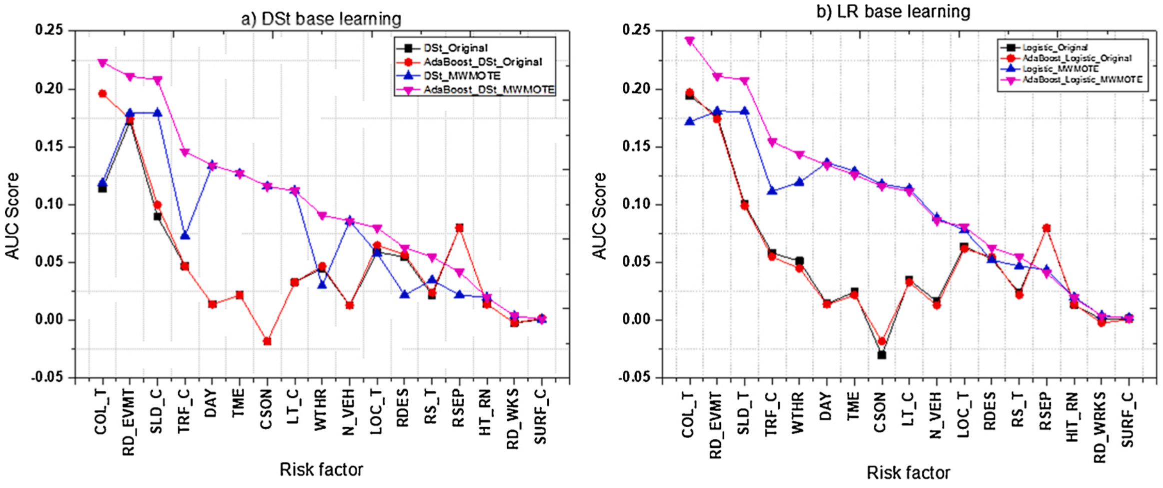

Fig. 2. Ranking Scores of the risk factors.

demonstrate higher AUC values relative to those of the existing ranking approaches. Besides, some inconsistencies in the variable's relative importance have been noted across the different ranking approaches. For instance, according to EbVR, collision type is the most important variable, followed by road environment. However, when variable ranking by a single inductive learning algorithm is based on the imbalanced data (e.g., S1), the importance of collision type is underestimated and ranked erroneously. Similarly, when a single classifier induced from the balanced data (e.g., S3) is applied for the variable ranking, we find the AUC values lower and the relative importance of the variables inconsistent compared to the proposed EbVR method. It is also worth noting that when ensemble classifiers induced from imbalance data (i.e., S2) is applied to evaluate the crash factors, the AUC scores are under estimated. The analytical results clearly demonstrate that the risk factor ranking is influenced by the data distribution and the potency of the learning algorithm.

To further highlight the worth of the various evaluation schemes, aligned Friedman test is conducted on the AUC scores assigned. Table 2 presents the results for each scheme and performance indicator in the form of ranks/p-value, based on Aligned Friedman test. Those cases where the null hypothesis is not rejected are demonstrated with star (*). It is observed that the proposed EbVR with either of the base learners (i.e., MLRE or DSt) is significantly better than any of the compared methods except S3. In the case of S3, it is found that EbVR with the DSt base learning is not statistically significantly different. Also, the difference among S1, S2, and S3 is not statistically notable, even though S3 is better in the ranking.

Table 2
Aligned Friedman test results for the comparison of different variable selection schemes.

EbVR-MLRE vs.	R/ PHochberg	EbVR-DSt vs.	R/ PHochberg	S3 vs.	R/ PHochberg
S1-DSt	**107.50/ 0.007**	S1-DSt	**91.56/ 0.008**	S1	35.68/ 0.025*
S2-DSt	**99.88/ 0.008**	S2-DSt	**82.53/ 0.013**	S2	32.68/0.05*
S3-DSt	**56.68/ 0.017**	S3-DSt	39.21/ 0.025*	S3	9.65
EbVR-DSt	28.06/ 0.050*	S1-MLRE	**81.00/ 0.017**		
S1-MLRE	**98.30/ 0.013**	S2-MLRE	**82.91/ 0.010**		
S2-MLRE	**99.85/ 0.010**	S3-MLRE	24.18/ 0.050*		
S3-MLRE	**35.94/ 0.025**	EbVR-DSt	18.62		
EbVR-MLRE	21.79				
PAlignedFriedman: 0.036		PAlignedFriedman: 0.023		PAlignedFriedman: 0.003	

Note: R = rank, P = p-value, PAlignedFriedman = p- value computed by Aligned Friedman test, significant performance differences are in bold.

4.2. BN classification results

Figs. 3 and 4 show the ROC curve improvement for the subsets of top ranked variables by each ranking method with DSt and MLRE base learning, respectively. It is interesting to note that predicting the injury severity the BN model based on EbVR method is far superior to the compared approaches according to AUC measure. The result of S3 is also noteworthy, whereas S2 and S1 obtain the least but comparable prediction performance. This observation is also valid for precision and recall measures (Table 3). However, when the accuracy metric is used for the performance assessment, S1 and S2 appear to be the optimal way for the CIS modelling (Fig. 5).

Besides, the data imbalance seems to obscure variables that are important to the CIS level prediction. For instance, it can be seen that where the data imbalance prevails in the training set, such as S1 and S2 (Figs. 3a, b, and 4 a, b), the optimal subsets obtained are limited to two variables (i.e., collision type and road environment). However, when the data skew is treated in S3 and EbVR (Figs. 3c, d and 4 c, d), we find that the optimal subsets extend to 6 and 10 variables for DSt and MLRE, respectively. Clearly, more relevant variables which could have been left out in the CIS modelling are captured when variables ranking takes the data imbalance into account. Also, relative to DSt, the induction of MLRE in the proposed EbVR is helpful to identify many useful predictors. Comparing the methods based on the AIC goodness-of-fit measure (Table 4), we find that the proposed EbVR is statistically superior to the other methods, particularly when MLRE is chosen for the base learning (Table 5).

For the test case, Fig. 2 shows the average AUC values of the BN classifier for DSt and MLRE induction from the different variable selection schemes studied. Evidently, the proposed EbVR demonstrates higher AUC values relative to the other schemes. The finding shows that variables selected by EbVR approach are the most authentic ones for the CIS prediction.

4.3. Injury severity analysis

To obtain a more reliable subset of the risk factors, we integrate the top 6 variables obtained by EbVR based on DSt and MLRE inductions. The integration results in the discovery of 7 most probable factors of the

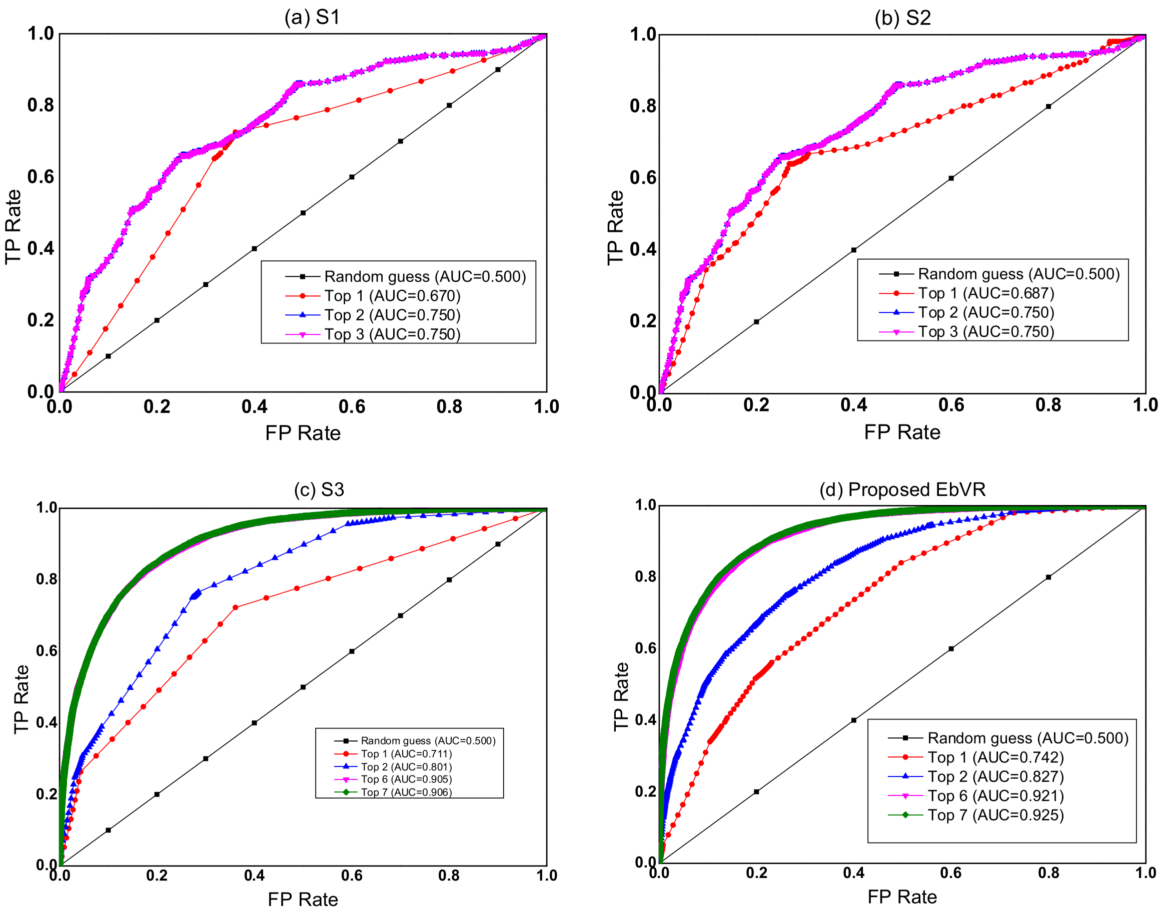

Fig. 3. BN classifiers ROC curve improvements for the top-ranked risk factors identified based on DSt induction in the various ranking methods.

CIS prediction including: collision type (COL_T), road environment (RD_EVMT), shoulder type (SLD_T), traffic control (TRFC_C), weather (WTHR), day of week (DAY), and time (TME). These variables are employed to determine the most probable factors of MFI crashes based on the BN inference results (Table 4). Table 4 depicts the values of variables in which the probability of a MFI crash is higher (in bold) than an nMFI one. For instance, the table demonstrates that, assigning a probability of 1 to the following variable values: collision type (head on), shoulder condition (untarred), weather (dazzle), and road environment (rural), the likelihood that the crash will be a multiple fatality one is 0.531. Accordingly, we find that five most probable explanations can be offered to a multiple fatality crash in Ghana, including: 1) under the dazzle weather, head on crashes on rural highways with untarred shoulders 2) collision in a village area of the highway with tarred shoulders, 3) collision on a rural highway with untarred shoulders and without traffic control, 4) crashes in peak time at urban locations where the road has no shoulders, and 5) head-on collisions during the off peak.

5. Discussions and conclusions

Multiple fatality crashes are extremely rare events compared to other injury categories in the crash database. A body of research suggests that the data imbalance may pose challenges to the conventional machine learning classifiers in CIS analysis. The current study proposes an ensemble-based variable ranking (EbVR) approach for the imbalance data learning. The technique integrates the MWMOTE oversampling with the adaptive boosting to develop the ensemble classifiers to evaluate and select useful variables in a high dimensional and severely imbalanced crash data. Important variables identified are applied to the balanced to form the training data for the CIS modelling. The main objective is to find the parsimoniously fitted model to analyze the risk factors of multiple fatality crashes.

The empirical results in the case study of Ghana suggest that the ranking scores by the EbVR, especially with MLRE induction, significantly outperform the variables obtained by applying single classifier induced from the imbalanced data (S1), ensemble classifiers induced from balanced data (S2) and single classifiers induced from the balanced data (S3). This is due to the fact that the proposed EbVR incorporates data balance and ensemble learning in a single framework to deal with the learning bias and reduce the pseudo-loss of the final hypothesis. This leads to the selection of relevant predictors of the CIS level based on the BN classification. The fact that S1 and S2 perform the least suggest that the failure to account for data imbalance associated with the crash data can eventually bring about classification biases and suboptimal results (Mujalli et al., 2016; Jiang et al., 2019). Even though S3 records better AUC values compared to S1 and S2, the result shows no significant difference. It illustrates that oversampling alone is insufficient for the accurate evaluation of the variables. The disaggregate-level crash data are complex with injury severity patterns that are hard to learn (Savolainen et al., 2011). Even in an oversampled data, a single learning algorithm might be insensitive to the changes in the training samples upon the least perturbation (Freund and Schapire, 1996). This may cause the formation of wrong prediction rules, resulting in incorrect variable importance measure. Also, the fact that S2 is not significantly different from S1 gives reason to believe that ensemble learning in highly imbalanced data is worthless for the variable ranking. In imbalanced training data, the inherent bias of standard ML algorithm towards the majority samples may still remain irrespective of the number of learning algorithms induced, and eventually affect the classification accuracy (Han et al., 2005). Although the proposed ranking technique with DSt

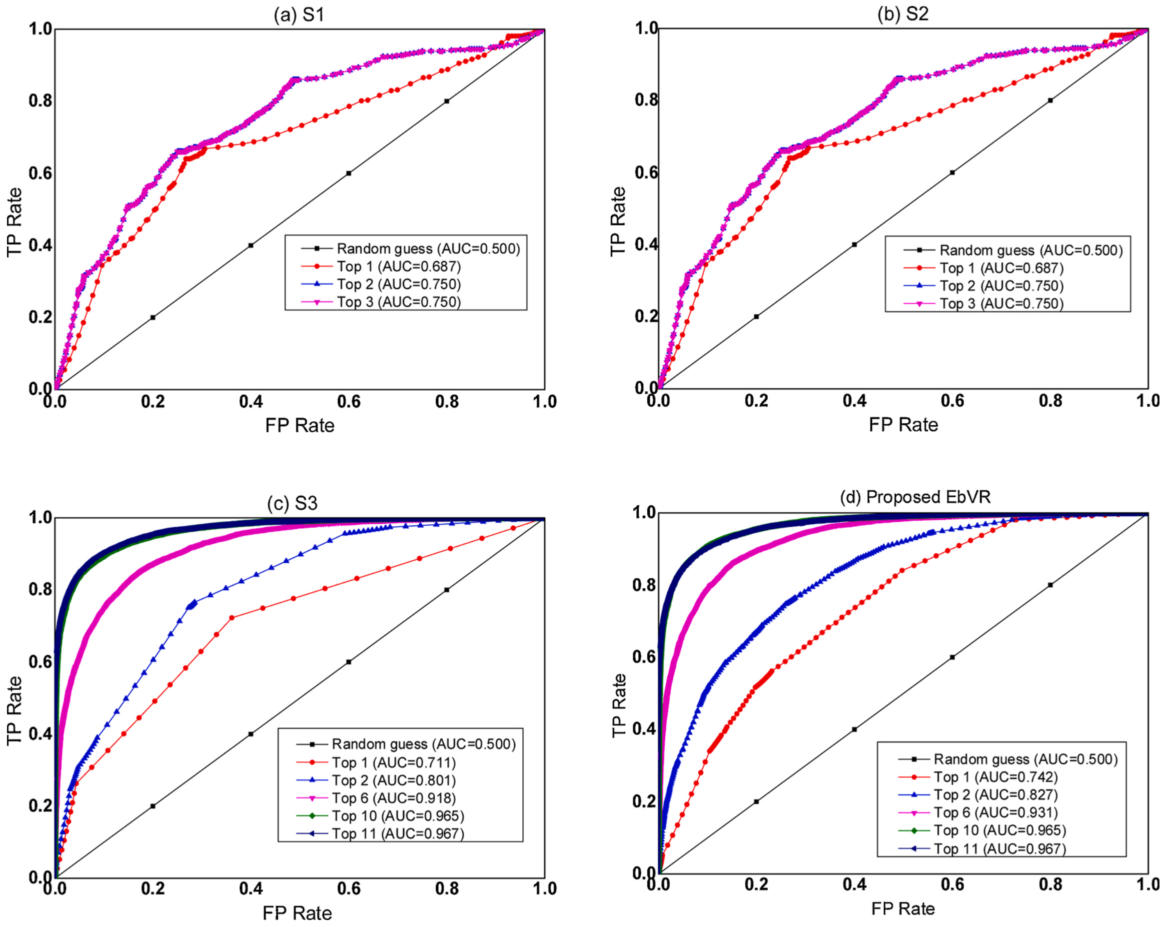

Fig. 4. BN classifiers ROC curve improvements for the top-ranked risk factors identified based on MLRE induction in the various ranking methods.

induction is remarkably better than S1 and S2, relative to S3, it is insignificantly better. However, in the case of EbVR with MLRE base learning, the results show that there is a statistically significant difference relative to any of the ranking methods. It suggests that the MLRE algorithm is more suitable to define the decision boundary with lower error rate than DSt. Unlike the decision trees, MLRE fits a single linear boundary for the classification, which makes it relatively less prone to over-fitting even when the data classes are not well separated. Thus, for the variable ranking, it appears MLRE is the better choice for induction with the proposed scheme. MWMOTE technique have been proven effective to deal with the data imbalance issue for the enhancing classification, particularly with the minority samples (Barua et al., 2014). Also, unlike single inductive learning, ensemble learning can respond to the training data perturbation by generating various hypotheses on different distributions that focus on the hard-to-learn samples and compel a weak learner to improve the performance. Therefore, the synergy between MWMOTE and AdaBoost techniques is desirable and has proven useful for the variable selection task in the current study.

The superior prediction power of the BN classifier developed on the variable subset obtained by proposed EbVR method further demonstrate its advantage over the compared methods. However, the model comparison with the overall accuracy measure is inconsistent with the other performance indicators such as AUC, precision, and recall. According to the overall accuracy measure, S1 and S2 are the best routes for the variable selection. Meanwhile, the variable subsets found therein perform abysmally in the minority class, i.e., recall and precision, suggesting that the accuracy measure is biased towards the samples with the most observations. However, with AUC, performance of ML methods is measured across the entire probability threshold in an impartial manner, making it a more appropriate performance measure in imbalance learning (Vilaça et al., 2019). Thus, using accuracy as the performance metric in the imbalanced learning may conceal the important risk factors of a road crash and generate misleading study conclusions (Jiang et al., 2019).

The analytical results suggest that under the dazzle weather condition, head-on crashes on rural highways with untarred shoulders are of the most influencing factors of multiple fatality crashes. The climate of Ghana is tropical and yearlong hot with an average temperature of 30 °C. It is plausible to assume that the dazzling sunlight impairs drivers' visibility and the heating effect on the road surface resulting from such weather could cause sudden burst to car tires. The situation could be more dangerous when the vehicle involved is speeding. Moreover, untarred road shoulders are often found with overgrown shrubs and trees, particularly in rural highways sections without residents, leaving the effective road width narrow for safe overtaking manoeuvres and recovery of errant vehicles. The combined effect of these conditions may increase the likelihood of a head-on collision, which is one major cause of multiple deaths on the Ghanaian highways. For the safety improvement, a detailed study is recommended to develop estimated temperature thresholds for safe vehicle and driving conditions in Ghana. Based on the results of such studies, variable message system can be adopted to pre-inform motor users of the weather conditions and corresponding precautionary measures. It is also important to intensify routine maintenance activities along the entire stretch of the highways to keep the complete road width clear and intact for safer driving operations. It is shown that head-on collision rate for divided highways is significantly lower than the rate on undivided highways, particularly for rural two-lane highway (Fitzpatrick et al., 2005). Obviously, with non-traversable median, out-of-control vehicles could manoeuvre to a safe stop without getting in the way of opposing traffic to cause a deadly

Table 3
Predictions of BN classifiers developed on nested subsets identified according to the different selection approaches.

Base learner	Risk factor sets	S1	S2	S3	EbVR
	Perf. metrics	Accuracy			
	Top 1	**98.072**	**98.072**	68.082	67.166
	Top 2	**98.072**	**98.072**	74.048	74.411
	Top 6	–	–	82.27	83.987
	top 10	–	–	89.052	89.753
	Perf. metrics	Precision			
	Top 1	0.000	0.000	**0.667**	0.628
DSt	Top 2	0.000	0.000	0.729	**0.743**
	Top 6	–	–	0.811	**0.827**
	top 10	–	–	0.895	**0.911**
	Perf. metrics	Recall			
	Top 1	0	0	0.723	**0.84**
	Top 2	0	0	**0.766**	0.747
	Top 6	–	–	0.842	**0.86**
	top 10	–	–	**0.885**	0.881
	Perf. metrics	Accuracy			
	Top 1	**98.072**	**98.072**	68.082	67.166
	Top 2	**98.072**	**98.072**	74.048	74.411
	Top 6	–	–	83.784	85.461
	top 10	–	–	89.754	89.754
	Perf. metrics	Precision			
	Top 1	0	0	**0.667**	0.628
MLRE	Top 2	0	0	0.729	**0.743**
	Top 6	–	–	0.83	**0.849**
	top 10	–	–	**0.911**	**0.911**
	Perf. metrics	Recall			
	Top 1	0	0	0.723	**0.84**
	Top 2	0	0	**0.766**	0.747
	Top 6	–	–	0.849	**0.863**
	top 10	–	–	**0.881**	**0.881**

Note: The bold numbers indicate that the values obtained are the highest for the respective variable when compared to others within the category.

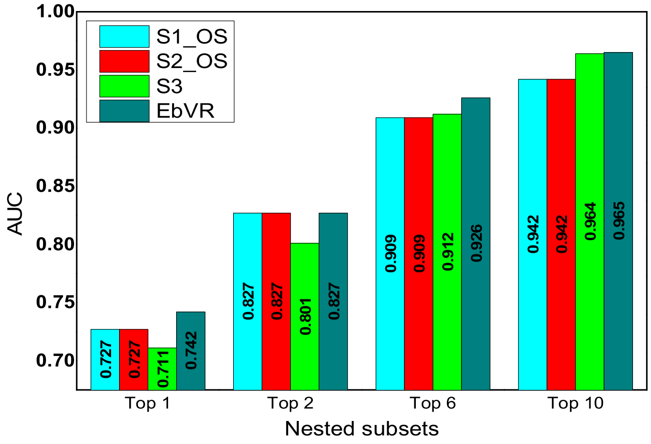

Fig. 5. The BN classification performance on selected subsets according to the ranking methods.

head-on crash. Despite the potential safety impact, road medians on the majority of rural two-lane highways in Ghana are defined by faded line markings that are generally hard to see. Thus, as an engineering solution to the road carnage, non-traversable medians and visible road line markings should be considered for the major highways linking to the neighbouring countries.

Head-on collisions during the off-peak hours are significantly linked to MFI. The prevalence of head-on collisions on rural and village segments can be explained by several factors. First, the majority of long-distance passenger transport shifts the journey time towards the evening mainly to evade the deadlock traffic and reduce travel time. However, driving continuously throughout the night often causes fatigue which is one major contributing factor to the gory accidents in recent times in Ghana. Evidently, low traffic volume during the night time motivates dangerous driving behaviours such as speeding and improper overtaking manoeuvres. A large body of literature shows that speeding is associated with the rising frequency and severity of crashes. Speed limit increases are linked to frequent and severity of road crashes in Michigan (Wagenaar et al., 2020), Australia (Taylor et al., 2002), and across Europe (Helfenstein, 2020). On undivided segments of rural highways with posted speed of 80 km/h, 66 % of the vehicles observed in Ghana are found speeding in excess (Damsere-Derry et al., 2020). Thus, the finding is intuitively linked to fatigue-driving and the propensity towards excessive speeding that may lead to MFI severity crashes (Mujalli et al., 2016; Moral-García et al., 2019). The observation indicates that reducing fatigue-related crashes may require measures to ensure more than one driver for each of such passenger transport vehicles. Alternatively, rest areas should be provided at suitable locations along the highways for drivers to take intermediate breaks. Also, the implementation of speed limit regulations must be further strengthened by equipping the relevant agencies with required logistics and human resources to clamp down on the speeding violations.

Furthermore, we find that crashes occurring in a village settlement area with tarred road shoulders have a higher probability of MFI. Ordinarily, in the study background highway sections within the settlement areas have good tarred shoulders with reasonably wider pavement. Unfortunately, in these areas the road shoulders are misused and tend to be fertile grounds for the commercial activities, thus, narrowing

Table 5
Inference results for multiple evidences in BN CIS model.

No. Var	Condition satisfied	Prob. MFI
1	P(CIS = MFI|COL_T = 1, SLD_C = 2, WTHR = 5, RD_EVMT = 3)	**0.531**
2	P(CIS = MFI|SLD_C = 1, RD_EVMT = 2, WTHR = 6)	**0.895**
3	P(CIS = MFI|TRF_C = 1, RD_EVMT = 3, WTHR = 6, SLD_C = 2)	**0.978**
4	P(CIS = MFI|DAY = 0, TRF_C = 5, RD_EVMT = 1, TME = 1, SLD_C = 3)	**0.965**
5	P(CIS = MFI|TME = 0, WRTH = 6, COL_T = 1)	**0.998**
6	P(CIS = MFI|RD_EVMT = 3)	0.460
7	P(CIS = MFI|WTHR = 1, RD_EVMT = 1)	0.729

Table 4
Goodness-of-fit measures (AIC) for the CIS models found by different model selection methods.

Learner	Data/model	Top 1	Top 2	Top 6	Top 10
	S1	−24187.241	−76175.729	−146259.182	−201474.670
	S2	−55782.004	−76175.729	−146259.182	−201474.670
DSt	S3	−86960.642	−128293.269	−337040.178	−443619.951
	EbVR	**−142425.176**	**−190538.750**	**−348842.060**	**−447731.228**
	S1	−55782.004	−76175.729	−153676.756	−200568.018
MLRE	S2	−55782.004	−76175.729	−153676.756	−200568.018
	S3	−86960.642	−128293.269	−338124.999	−447731.228
	EbVR	**−142425.176**	**−190538.750**	**−358267.160**	**−447731.228**

The bold numbers indicate that the values obtained are the lowest for the respective variable when compared to others within the category.

the effective road width and inducing unsafe movement of persons. It is common to find that errant vehicles that try to use the road shoulders for the recovery purposes end up running into traders and causing more deaths. To this effect, planners and city authorities must earmark suitable locations away from the road corridors for the commercial activities and traders should be compelled to use these designated locations rather than the road shoulders.

Urban crashes at peak time on roads without shoulders at a location where a give-way sign is installed are also associated with MFI. Road shoulder is a portion of the roadway contiguous with the carriage way that accommodates the stopped vehicles and serves to provide space for errant vehicles to recover safely. Without shoulders the lane width might be insufficient to accommodate overtaking operations especially on two-lane bidirectional highways. Unfortunately, a considerable number of urban networks are without shoulders in Ghana, leaving vehicles with no choice but to stop on the carriage way. At locations of intersecting roads, these vehicles tend to obstruct driving visibility for entering and exiting traffic, which together with the speeding that characterizes rush hours may lead to MFI. Similarly, it is worthwhile to note that illegal structures such as kiosks and indoor advertising billboards at the Ghanaian urban intersections also create a significant obstruction to driver visibility. Moreover, traffic sign violations in the urban centres are more obvious during the peak time (Fountas and Anastasopoulos, 2018). In Ghana, traffic violations are common among drivers of "trotro" passenger vehicles which are the dominant mode for the urban commute in Ghana. In trying to increase the number of trips and meet the minimum daily income target, operators of these vehicles engage in risky driving behaviours such as wrongful overtaking and failure to give way. These actions often result in severe crashes, particularly when the traffic volume rises (Frost and Morrall, 1998). Sadly, the rate of traffic violations in the cities is on the rise due to the perceived personal time gains associated with the violations (Awialie Akaateba and Amoh-Gyimah, 2013). Consistent efforts by law enforcement agencies to implement the laws on traffic violations are still below the belt, perhaps, due to logistics and human resource constraints. It is also plausible to argue that the statutory penalty for traffic violations is not deterring enough considering the perceived benefits accrued to the individual perpetrators as well as the socio-economic loss of a fatal crash resulting from these violations. In view of the study findings, we suggest the competent roadway authorities to set up a traffic violation charging unit to handle traffic violation charges where different violations would attract penalties commensurate the likelihood of severe injuries. In this way, the frequency of those violations with the potentials to cause MFI may reduce as prospective offenders might be deterred by the associated austere penalties. Besides, law enforcement agencies must be adequately resourced to enforce the traffic rules for the user safety.

Developing a parsimoniously fitted model in high dimensional data is a scientific objective. However, the findings here suggest that data imbalance in high dimensional data may cause highly predictive factors to be left out of the crash severity model, leading to a suboptimal performance and biased study conclusions. Fortunately, the proposed method has demonstrated outstanding results and been capable of addressing the data imbalance and issues associated with single inductive learning for selecting effective variables. The technique has been applied in the case study of MFI severity in Ghana and the relevant contributory factors are identified and discussed. The study results are crucial to guide safety design and policy improvements that may reduce the frequency of horrific crashes on the Ghanaian highways.

Our evaluation has several limitations related to data quality and completeness. The original data contain a considerable amount of missing data which may render some intuitively relevant variables hard to include for the analysis. For instance, road width and crash coordinate variables which incorporate a large number of missing values become less informative and impossible to use. Besides, class noise which influences the performance of inductive learning algorithms is not accounted for in the current study. Also, the absence of driver and vehicle characteristics in the data database, which have possibly effects on crash severity, may not allow for the systematic study to reveal their personal influences on fatal injury crashes. Hence, there is an acute need for the national road safety commission (NRSC) to develop an efficient database that incorporates relevant variables for the comprehensive study. This is an important step towards reducing the road carnage in Ghana.

CRediT authorship contribution statement

Mahama Yahaya: Conceptualization, Methodology, Software. **Runhua Guo:** Data curation, Writing - original draft. **Xinguo Jiang:** Visualization, Investigation. **Kamal Bashir:** Supervision. **Caroline Matara:** Software, Validation. **Shiwei Xu:** Writing - review & editing.

Declaration of Competing Interest

None.

Acknowledgments

The funding support of the National Nature Science Foundation of China (NSFC, 71771191 and 51608455) and Sichuan Provincial Science and Technology Innovation Talents Fund (2019JDRC0023) is deeply acknowledged.

References

Alcalá-Fdez, J., et al., 2009. KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft comput. 13 (3), 307–318.

Awialie Akaateba, M., Amoh-Gyimah, R., 2013. Driver attitude towards traffic safety violations and risk taking behaviour in kumasi: the gender and age dimension. Int. J. Traffic Transp. Eng. 3 (4).

Barua, S., et al., 2014. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowl. Data Eng. 26 (2), 405–425.

Bashir, K., et al., 2020. SMOTEFRIS-INFFC: handling the challenge of borderline and noisy examples in imbalanced learning for software defect prediction. J. Intell. Fuzzy Syst. 38 (1), 917–933.

Cai, Q., et al., 2020. Real-time crash prediction on expressways using deep generative models. Transp. Res. Part C Emerg. Technol. 117, 102697.

Chawla, N.V., et al., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Damsere-Derry, J., et al., Assessment of vehicle speeds on different categories of roadways in Ghana. International Journal of Injury Control & Safety Promotion. 15 (2): p. 83-91.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434–444.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. 40 (2), 139–157.

Dietterich, T.G., 2002. ensemble learning. The Handbook of Brain Theory and Neural Networks, pp. 110–125, 2.

Lamba, D., et al., Coping with Class Imbalance in Classification of Traffic Crash Severity based on Sensor and Road Data: A Feature Selection and Data Augmentation Approach.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.

Fernández, A., et al., 2017. An insight into imbalanced big data classification: outcomes and challenges. Complex Intell. Syst. 3 (2), 105–120.

Fitzpatrick, K., Schneider, W.H., Park, E.S., 2005. Comparisons of Crashes on Rural Two-lane and Four-lane Highways in Texas. Texas Transportation Institute, Texas A & M University System.

Fountas, G., Anastasopoulos, P.C., 2018. Analysis of accident injury-severity outcomes: The zero-inflated hierarchical ordered probit model with correlated disturbances. Anal. Methods Accid. Res.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. Icml. Citeseer.

Frost, U., Morrall, J., 1998. A comparison and evaluation of the geometric design practices with passing lanes, wide-paved shoulders and extra-wide two-lane highways in Canada and Germany. Transp. Res. Part B Methodol. 34, 1–15.

García, S., et al., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf. Sci. (Ny) 180 (10), 2044–2064.

Guerrero-Ibáñez, J., Zeadally, S., Contreras-Castillo, J., 2018. Sensor technologies for intelligent transportation systems. Sensors 18 (4), 1212.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. March.

Hall, M., et al., 2009. The WEKA data mining software: an update. Acm Sigkdd Explor. Newsl. 11 (1), 10–18.

Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing. Springer.

He, H., Ma, Y., 2013. Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons.

Helfenstein, U., When did a reduced speed limit show an effect? Exploratory identification of an intervention time. Accident Analysis & Prevention. 22(1): p. 79-87.

Hossin, M., Sulaiman, M., 2015. A review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manag. Process. 5 (2), 1.

Jeong, H., et al., 2018. Classification of motor vehicle crash injury severity: a hybrid approach for imbalanced data. Accid. Anal. Prev. 120, 250–261.

Jiang, L., Xie, Y., Ren, T., 2019. Modelling highly unbalanced crash injury severity data by ensemble methods and global sensitivity analysis. In: Proceedings of the Transportation Research Board 98th Annual Meeting. Washington, DC, USA.

Khoshgoftaar, T.M., Gao, K., Seliya, N., 2010. Attribute selection and imbalanced data: problems in software defect prediction. International Conference on Tools With Artificial Intelligence.

Kumar, V., Minz, S., 2014. Feature selection. SmartCR 4 (3), 211–229.

Kwon, O.H., Rhee, W., Yoon, Y., 2015. Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. 75, 1–15.

Le Cessie, S., Van Houwelingen, J.C., 1992. Ridge estimators in logistic regression. J. R. Stat. Soc. Ser. C Appl. Stat. 41 (1), 191–201.

Leevy, J.L., et al., 2018. A survey on addressing high-class imbalance in big data. J. Big Data 5 (1), 42.

Li, C., 2007. Classifying imbalanced data using a bagging ensemble variation (BEV). In: Proceedings of the 45th Annual Southeast Regional Conference. ACM.

Liu, W., et al., 2010. A robust decision tree algorithm for imbalanced data sets. In: Proceedings of the 2010 SIAM International Conference on Data Mining. SIAM.

Mafi, S., Abdelrazig, Y., Doczy, R., 2018. Machine learning methods to analyze injury severity of drivers from different age and gender groups. Transp. Res. Rec. 2672 (38), 171–183.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.

Moral-García, S., et al., 2019. Decision tree ensemble method for analyzing traffic accidents of novice drivers in urban areas. Entropy 21 (4), 360.

Muchlinski, D., et al., 2016. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Anal. 24 (1), 87–103.

Mujalli, R.O., López, G., Garach, L., 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 88, 37–51.

Pei, X., et al., 2016. Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong. Accid. Anal. Prev. 95, 512–520.

Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer.

Sáez, J.A., et al., 2015. SMOTE–IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf. Sci. (Ny) 291, 184–203.

Savolainen, P.T., et al., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.

Schlögl, M., et al., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Accid. Anal. Prev. 127, 134–149.

Shanthi, S., Ramani, R.G., 2012. Feature relevance analysis and classification of road traffic accident data through data mining techniques. Proceedings of the World Congress on Engineering and Computer Science.

Tang, J., Alelyani, S., Liu, H., 2014. Feature selection for classification: a review. Data classification: Algorithms and applications 37.

Tantithamthavorn, C., Hassan, A.E., Matsumoto, K., 2018. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. Ieee Trans. Softw. Eng.

Taylor, M.A.P., Woolley, J.E., Zito, R., 2002. Immersion Reality: Combining Microsimulation Modelling and Probe Vehicles in Traffic Studies.

Thammasiri, D., et al., 2014. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Expert Syst. Appl. 41 (2), 321–330.

Vilaça, M., Macedo, E., Coelho, M.C., 2019. A rare event modelling approach to assess injury severity risk of vulnerable road users. Safety 5 (2), 29.

Wagenaar, A.C., F.M. Streff, and R.H. Schultz, Effects of the 65 mph speed limit on injury morbidity and mortality. Accident Analysis & Prevention. 22(6): p. 571-585.

Witten, I.H., et al., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Yahaya, M., et al., 2019. Enhancing crash injury severity prediction on imbalanced crash data by sampling technique with variable selection. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE.

Yen, S.-J., Lee, Y.-S., 2006. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. Intelligent Control and Automation. Springer, pp. 731–740.

Zhang, C., et al., 2018. A cost-sensitive deep belief network for imbalanced classification. IEEE Trans. Neural Netw. Learn. Syst. (99), 1–14.

Zheng, M., et al., 2019. Traffic accident's severity prediction: a deep-learning approach-based CNN network. IEEE Access 7, 39897–39910.