Accident Analysis and Prevention 150 (2021) 105936

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Bayesian networks for imbalance data to investigate the contributing factors to fatal injury crashes on the Ghanaian highways

Mahama Yahaya [a], Runhua Guo [b], Wenbo Fan [a], Kamal Bashir [c], Yingfei Fan [d], Shiwei Xu [e], Xinguo Jiang [a,*]

[a] School of Transportation and Logisitics, Southwest Jiaotong University, National Engineering Laboratory of Integrated Transportation Big Data Application Technology, School of Transportation and Logistics, No. 999, Xi'an Road, Chengdu, Sichuan, PR China
[b] Department of Civil Engineering, Suit 217, Heshanheng Bldg, Tsinghua University, Beijing 10084, PR China
[c] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, PR China
[d] School of Transportation and Logistics Engineering, Taiyuan University of Science and Technology, 66 Waliu Road, Wanbailin District, Taiyuan, China 030024
[e] Guangzhou Transport Planning Research Institute, No. 10 Guangwei Road, Guangzhou, PR China

ARTICLE INFO

Keywords:
Crash injury severity
Imbalance data
Oversampling techniques
Bayesian networks
Classification

ABSTRACT

The crash data are often predominantly imbalanced, among which the fatal injury (or minority) crashes are significantly underrepresented relative to the non-fatal injury (or majority) ones. This unbalanced phenomenon poses a huge challenge to most of the statistical learning methods and needs to be addressed in the data preprocessing. To this end, we comparatively apply three data balance methods, i.e., the Synthetic Minority Oversampling Technique (SMOTE), the Borderline SMOTE (BL-SMOTE), and the Majority Weighted Minority Oversampling (MWMOTE). Then, we examine different Bayesian networks (BNs) to explore the contributing factors of fatal injury crashes. The 2016 highway crash data of Ghana are retrieved for the case study. The results show that the accuracy of the injury severity classification is improved by using the preprocessed data. Highest improvement is observed on the data preprocessed by the MWMOTE technique. Statistical verification is done by the Wilcoxon signed-rank test. The inference results of the best BNs show the significant factors of fatal crashes which include off-peak time, non-intersection area, pedestrian involved collisions, rural road environment, good tarred road, roads without shoulders, and multiple vehicles involved crash.

1. Introduction

Considerable studies have been conducted using different statistical and machine learning (ML) techniques to develop crash severity models based on historical traffic accident records for safety analysis. Considerable studies have been conducted using different statistical and machine learning (ML) techniques to develop crash severity models based on historical traffic accident records for safety analysis. Among them, statistical methods such as logit (Milton et al., 2008; Shankar and Mannering, 1996) and probit (Fountas et al., 2018; Zhang et al., 2018b) are the most used. However, statistical modeling techniques have intrinsic assumptions and pre-defined underlying relationship between the predictor and response variables, which are often hard to achieve with the complex disaggregate level CIS data (Chang and Wang, 2006; Savolainen et al., 2011). Notwithstanding this, road safety researchers who are in favor of statistical methods have long emphasized the estimation of causal parameters over the model fit. One consequence of following this prescription is that most statistical models of highway injury severity have failed to predict the target concept when tested with samples outside the training set (Yahaya et al., 2020; Zong et al., 2013). This fact should be a cause of grave concern for, at least, two apparent reasons. First, road crashes are incredibly destructive, and accurately predicting their outcome is a critical issue for policymakers who must try to anticipate severe injury crashes and find ways to prevent or minimize their frequency. Secondly, this fact may be a good indication that the existing estimates of causal parameters are not very reliable. Besides, the data issues associated with the CIS database, such as missing values, over- and under-dispersion, heterogeneity etc., coupled with theoretical underdevelopment may undermine the credibility of causal modeling assumptions and eventually lead to inaccurate parameter

* Corresponding author.
E-mail addresses: yahayamahama448@yahoo.com (M. Yahaya), guorh@tsinhua.edu.cn (R. Guo), wbfan@swjtu.edu.cn (W. Fan), kamalbashir1@yahoo.com (K. Bashir), fanyf@my.swjtu.edu.cn (Y. Fan), 390353624@qq.com (S. Xu), xjiang@swjtu.edu.cn (X. Jiang).

https://doi.org/10.1016/j.aap.2020.105936
Received 19 December 2019; Received in revised form 23 May 2020; Accepted 30 November 2020
Available online 17 December 2020

estimates and erroneous study conclusions (Savolainen et al., 2011).

Meanwhile, some scholars posit that a model has the explanatory power to the degree that it makes accurate predictions about observations in out-of-sample data (Holland, 1986). Accordingly, a number of studies have shown that ML techniques often demonstrate a better prediction performance than the statistical methods (Abdel-Aty and Abdelwahab, 2004; Zong et al., 2013). Thus, recently, more scholars have advocated the use of machine learning (ML) methods to develop CIS models for the data analysis (Cigdem and OZDEN, 2018; Mafi et al., 2018; Schlögl et al., 2019). In this regard, decision trees (DT) (Chong et al., 2004; Cigdem and OZDEN, 2018; Moral-García et al., 2019; Oña et al., 2013), Neural Network (NN) (Chong et al., 2005; Jeong et al., 2018), Bayesian Networks (BNs) (de Oña et al., 2011; Mujalli et al., 2016) have been employed to analyse accident patterns. Nevertheless, there are a number of quality problems associated with the crash database, among which class imbalance is prominent (Montella et al., 2012; Mujalli et al., 2016; Vilaça et al., 2019). In binary class data, class imbalance occurs when the samples in one class (majority) far outnumbers the samples in the other class (minority). It is self-evident that the vast majority of traffic crashes that occur in most parts of the world record relatively fewer fatal injuries than non-fatal ones, making the appearance of these samples rare in the crash database (Montella et al., 2012). Thus, the crash database is predominantly and extremely imbalanced.

The data imbalance phenomenon poses a considerable challenge to the statistical learning methods that are used for the data analysis (Vilaça et al., 2019). These methods inherently assume that the training data are balanced (Leevy et al., 2018), and therefore, are more inclined to accurately predict the majority class samples whereas the minority class ones often tend to be misclassified (Leevy et al., 2018). To overcome this problem, two broad approaches are adopted, namely: the algorithm level and the data level (He and Ma, 2013). The algorithm level techniques are based on the cost-sensitive learning, in which different misclassification costs are assigned to the classes involved in a classification task (He and Garcia, 2009). However, it is still an enormous research challenge to obtain the optimal misclassification cost values for the individual classes (Zhang et al., 2018a).

With the data level approaches, data preprocessing is conducted to ensure that the classes are approximately equally represented. Typical methods in this category are based on the oversampling, undersampling, and case-control methods (Chawla et al., 2002; Han et al., 2005; Yen and Lee, 2006). Undersampling approaches randomly eliminate samples from the majority class until the required level of balance between classes is achieved. While the random undersampling (RUS) is easy to implement, it may cause the removal of informative majority samples to the detriment of classification performance of the model (Leevy et al., 2018). In case-control sampling, examples are selected homogeneously from each class while regulating the mixture of the classes to augment the minority and save the computational cost. Statistical analysis of case-control sampling has been extensively conducted in the literature. However, in highly imbalanced crash data with very few minority cases, case control method may lead to the significant loss of informative majority samples and deteriorate the classification performance. This assertion is supported by the work of Mujalli et al. (2016), in which the Bayes classifiers developed for imbalanced accident data failed to improve the CIS prediction in case-control scenario. Comparatively, oversampling techniques generate additional samples for the minority class to balance the training data. Methods that are based on oversampling have the advantage of enhancing the classification performance without causing the information loss (Tantithamthavorn et al., 2018). Some well-known techniques in this category include synthetic minority over-sampling (SMOTE) (Chawla et al., 2002) and Borderline SMOTE (Han et al., 2005). SMOTE oversamples the minority class by randomly taking any minority class sample and introducing synthetic ones along the line segment linking any of the minority class to the nearest neighbor. With BL-SMOTE, only the borderline minority samples identified are used for the up-sampling.

To address the data imbalance problem, Leevy et al. (2018) proposed that different methods ought to be investigated for the given domain datasets as there was no single best technique for all scenarios. Accordingly, Mujalli et al. (2016) conducted a comparative study of three different resampling techniques: under sampling, synthetic minority oversampling technique (SMOTE), and a hybrid of the two, together with three different Bayes classifiers to analyze traffic injury severity. The authors showed that the balanced data obtained by SMOTE technique was more powerful to improve the injury severity classification performance of the Bayes classifiers. Similarly, Mussone et al. (2017) developed two classification models based on SMOTE balanced data to determine the influential factors of crash injury severity at the urban intersections. Also, Vilaça et al. (2019) employed three different resampling methods including random undersampling (RUS), random oversampling (ROS), and SMOTE, together with the decision tree and logistic classifiers, to identify the risk factors of crash injury severity for vulnerable roadway users. The authors concluded that the classifiers developed by using SMOTE data were capable of determining the most probable factors of the injury severity. Recently, Zhang et al. (2018a) employed the BL-SMOTE technique to tackle the data imbalance issue before developing a deep-learning based convolutional neural network for traffic crash severity prediction.

However, Sáez et al. (2015) pointed out that SMOTE presented many drawbacks related to its blind up-sampling, whereby the creation of new minority samples failed to consider the distribution of samples from the majority. This weakness can cause too many synthetic samples to be generated in the neighborhood of needless minority samples, which may lead to the increased boundary overlapping and aggravate the classification task (Sáez et al., 2015). Thus, BL-SMOTE was intended to address the weakness of SMOTE related to the blind oversampling and boundary line disruption issue. Yet, in a related development, Barua et al. (2014) asserted that the k-Nearest Neighbor (kNN)-based criteria employed by SMOTE and its modified versions might fail to identify the useful minority class instances for oversampling, particularly so for a complex data scenario where the minority samples were organized in smaller disjuncts. Therefore, the authors proposed the majority weighted minority oversampling scheme (MWMOTE). With MWMOTE, a weighted scheme that incorporates a clustering technique is applied to identify the important minority samples based on which the oversampling is conducted. These imbalance treatment methods cited above have been extensively studied and applied in several domains (Kamal et al., 2017; Mathew et al., 2015). However, some scholars argue that oversampling may lead to overfitting issues (He and Garcia, 2009). Since the focus of this study is to identify the factors that are significantly related to fatal injury severity crashes, and given that overfitting does not alter the relationship between dependent and independent variables (Mussone et al., 2017), we chose to oversample the data to run the **Bayesian Networks (BNs)** classifications and analyse the crash injury severity (CIS) thereof. To analyze crash injury severity, several past studies have highlighted the advantages of Bayesian classifiers where the relationship between crash factors and the injury severity can be learned easily (de Oña et al., 2011; Gregoriades, 2007; Mujalli, et al., 2016). BNs have the advantages of bi-directional induction, incorporation of missing variables and probabilistic inference, among others. By using BNs, it is comparatively easy to ascertain the underlying patterns of data, to explore the relationships between variables and to make predictions using these relationships (de Oña et al., 2011).

Therefore, the objective of the study is in three folds; (1) to evaluate the suitability of different oversampling methods (i.e., SMOTE, BL-SMOTE, and MWMOTE) for the crash injury severity data balancing, (2) to examine the Bayesian networks (BNs) for the CIS, and (3) to determine the factors contributing to fatal injury crashes on the Ghanaian highways.

The rest of the paper is organized as follows: Section 2 presents the methodology, the data and data preprocessing techniques used, brief

description of BNs classifiers applied, and a description of the assessment metrics used to evaluate the models. The experimental results are presented in Section 3. Finally, the results discussions, conclusions and prospects for the future work are given in Section 4

2. Methodology

In this study, an imbalanced crash dataset is obtained and preprocessed by applying SMOTE, BL-SMOTE, and MWMOTE oversampling techniques. This process results in four different datasets consisting of three balanced datasets and the original imbalanced one. Moreover, four Bayesian classifiers are separately developed using the original imbalanced data as well as the balanced versions. For the BN classification, different scores and search algorithms are considered. The performance of the models estimated by the 10-fold cross-validation is captured in several metrics, including receiver operating characteristic curve (ROC) area, Mathew correlation coefficient (MCC), accuracy, and sensitivity. Overall, 36 models are developed for the comparative study. Finally, the best performing classifier is employed to identify the key factors of fatal injury crashes.

The study applies KEEL (García et al., 2015) data mining tool to implement SMOTE and BL-SMOTE algorithms, whereas MWMOTE is applied by the source code published in R statistical open source program (Harris et al., 2014).

To assess the experimental results for the Bayesian classifiers, we use WEKA software version 3.8 (Witten et al., 2016) with the default settings for all the classifiers except K2 and HillClimbing. For these search methods, the maximum number of parents is set to 2. The k-fold cross validation approach is adopted to calculate the classification results. Here, the value of k is set to 10 in accordance with the literature (Mujalli et al., 2016). The study framework is shown in Fig. 1.

2.1. Data

The traffic crashes that occurred on the Ghanaian highways between January and December 2016 are obtained from the Building and Road Research Institute (BRRI) of Ghana. Each record consists of features related to the situations at the time of the crash such as:

- Roadway features: describing the characteristics of roadway, such as horizontal and vertical alignments, pavement surface condition, pavement type, shoulder type, shoulder condition, location type, etc.
- Environmental features: lighting and weather conditions
- Crash characteristics and outcome severity: contributing circumstances 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 (Jeong et al., 2018; Oña et al., 2013).

2.1.1. Data preprocessing

The original dataset contains 8651 records with 42 variables. Variables such as road width, X and Y coordinates, which have substantial number of missing values are deleted, leaving a total of 6837 records for the analysis. Eighteen independent variables are analyzed to identify the key influential factors of crash severity prediction (Table 1). The variable selection is based on the completeness of records for a given variable in the original data and recommended in the literature (Theofilatos et al., 2012; Wahab and Jiang, 2019). Out of the eighteen, thirteen variables are applied as they appear in the original data. Other variables such as time, date, number of vehicles, and day of the week are discretized to facilitate the BNs implementation. For instance, the date and time are transformed into two levels and relabeled as a season (i.e., rainy & dry) and time (i.e., peak & off-peak), respectively. Similarly, the variable depicting the number of vehicles involved in the crash has been binarized to include one and multiple vehicles involved. As the research aims to identify the relevant factors of fatal injury severity, we binarize the injury severity into "Fatal Injury (FI)" = 1 and "non-fatal Injury (nFI)" = 0. The nFI is the aggregation of "injured hospitalized," "injured not hospitalized," and "damage only" categories in the original data. Binarizing the original data leads to a sample distribution of 80:20 (Table 1) for the nFI and FI severities, respectively. From the viewpoint of effective problem solving, Triguero et al. (Triguero et al., 2015) posed the view that any class imbalance level that made modeling and prediction of the minority class a complex and challenging task could be considered a high-class imbalance by the domain experts. Accordingly, the academic researchers typically define imbalanced data as those in which the proportion of minority class samples constitutes less than 35 % of the dataset (Li and Sun, 2012). The assertion proposed by Li and Sun (2012) has been confirmed in the experimental study to improve CIS prediction on imbalanced crash data (Yahaya et al., 2019). From the foregoing discussion, it is palpable that the data used in the study is imbalance. Also, it must be stated that we concentrate our investigation on class imbalance in CIS data in the context of binary classification problems, since typically non-binary (i.e., multi-class) classification challenges can be exemplified using a series of multiple binary classification tasks.

2.1.2. Resampling techniques

Resampling is widely used as a preprocessing step to address the data imbalance problem, in which the class distribution of samples is altered to augment the minority class records in the training data (Thammasiri et al., 2014). In this work, we investigate the application potential of

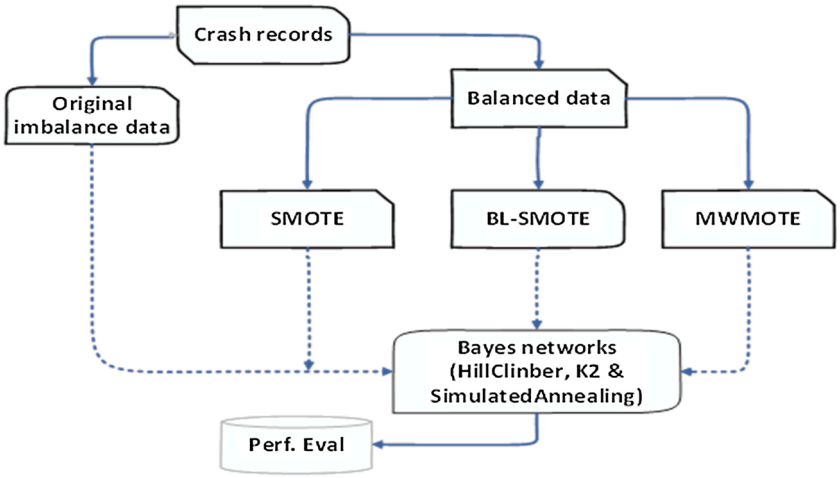

Fig. 1. The Experimental Framework.

Table 1
Variable Description and Classification According to Injury Severity.

Variable	Categories	Description	Count	Crash severity	
				nFI (%)	FI (%)
No. of Vehicles involved	0	single vehicle	2996	72.2	27.8
	1	2 or more vehicles	3841	86.4	13.6
Season	0	Dry season	2713	79	21
	1	Rainy season	4124	81	19
Day of week	0	Weekend	2075	78	22
	1	Week day	4762	81.2	18.8
Time	0	Off-peak	5120	79.4	20.6
	1	Peak hour	1717	82.5	17.5
Weather	1	Clear	5346	82.2	17.8
	2	Fog/Mist	7	71.4	28.6
	3	Rain	41	80.5	19.5
	4	Dust/Smoke	15	26.7	73.3
	5	Dazzle	6	83.3	16.7
	6	Others	1422	73.1	26.9
Light condition	1	Day	4435	82.9	17.1
	2	Night- No light	736	68.9	31.1
	3	Night- Lights OFF	205	62.4	37.6
	4	Night- Lights ON	1461	80.2	19.8
Road description	1	Straight and flat	6017	81.3	18.7
	2	Curve only	409	72.4	27.6
	3	Incline only	86	75.6	24.4
	4	Curve & Incline	314	72.6	27.4
	5	Bridge (name river)	11	36.4	63.6
Road surface type	1	Tar Good	5696	81	19
	2	Tar few Potholes	766	78.5	21.5
	3	Tar many Potholes	159	74.8	25.2
	4	Gravel	27	70.4	29.6
	5	Earth few Potholes	157	71.3	28.7
	6	Earth many Potholes	32	53.1	46.9
Shoulder condition	1	Good	3451	78.6	21.4
	2	Poor	789	77.1	22.9
	3	Overgrown	94	69.1	30.9
	4	No Shoulder	2503	83.8	16.2
Road separation	1	Median	1753	86.5	13.5
	2	No median	5084	78	22
Surface condition	1	Dry	6777	80.2	19.8
	3	Wet	55	76.4	23.6
	3	Muddy	5	80	20
Location type	1	Not at junction	5411	78	22
	2	regular-intersection	406	93.3	6.7
	3	T-intersection	784	86.4	13.6
	4	Staggered intersection	56	87.5	12.5
	5	Y-intersection	16	87.5	12.5
	6	Roundabout	101	95	5
	7	Railway crossing	2	0	100
	8	Other	61	80.3	19.7
Traffic control	1	None	4832	78.2	21.8
	2	Pedestrian crossing	32	81.3	18.8
	3	Signals	839	92.1	7.9
	4	Stop sign	10	90	10
	5	Give Way	96	90.6	9.4
	6	Other	1028	78.9	21.1
Collision type	1	Head on	697	67.7	32.3
	2	Rear end	1567	91.6	8.4
	3	Right Angle	446	90.8	9.2
	4	Side Swipe	797	92.7	7.3
	5	Ran off road /Overturn	973	80.7	19.3
	6	Hit Object on road	53	94.3	5.7
	7	Hit Object off road	186	89.2	10.8
	8	Hit parked vehicle	269	83.6	16.4
	9	Hit Pedestrian	1751	64.9	35.1
	10	Hit Animal	29	89.7	10.3
	11	Other	69	62.3	37.7
Hit and Run	1	Not Hit & Run	6584	80.8	19.2
	2	Hit & Run	253	63.6	36.4
Road works	1	Not at Roadworks	6789	80.2	19.8
	2	At Roadworks	48	75	25
Road environment	1	Urban	4125	86.4	13.6
	2	Village	59	78	22
	3	Rural	2653	70.7	29.3
Severity	0	Non-fatal Injury (nFI)	5484		
	1	Fatal injury (FI)	1353		

three widely used imbalance treatment methods. The methods are briefly described as follows:

- Synthetic Minority Oversampling Technique (SMOTE): Chawla et al. (2002) oversampled the minority class by SMOTE technique, which generated new synthetic samples along the line between the minority samples and their selected nearest neighbors. These synthetic samples are created by taking the difference between the variable vector into consideration and its closest neighbor. The difference is then multiplied by a random number between 0 and 1 and added to the variable vector. This causes a random point to be chosen along the line segment between two specified variables, thus effectively expanding the decision region of the minority class.
- Borderline-SMOTE (BL-SMOTE): BL-SMOTE only considers the borderline samples of the minority class for the up-sampling since they are more easily misclassified than the ones far away from the borderline (Han et al., 2005). The authors identify a minority class borderline sample according to the expression:

$$\frac{k}{2} \leq \alpha \leq k \tag{1}$$

Where α denotes the number of majority class samples, and k represents the number of nearest neighbors of the minority class sample under consideration.

- Majority Weighted Minority Oversampling (MWMOTE): The technique involves three main stages. Firstly, the most important and hard-to-learn minority class samples are selected from the original minority to develop a set of important minority samples. Secondly, different weights are assigned to the selected minority according to their importance. Finally, synthetic samples are generated from the weighted informative minority class samples by a clustering method. In this way, the right synthetic samples are efficiently generated even in complex data situations to ease the learning task for the entire spectrum of data (Barua et al., 2014).

2.2. Bayesian networks (BNs) classifiers

BNs are graphical models of interactions amongst a set of variables, where the variables are denoted by nodes and the relations between variables are denoted by directed links/arcs between the nodes. Any pair of unconnected nodes of the graph shows the (conditional) independence between the variables.

Let X = [x1, x2,. . ., xn, n ≥ 1] be a set of variables, where n = 18 in the present study. A BN developed on X is a Directed Acyclic Graph (DAG) network structure over X and a set of probability tables Xp = [p(xi |pa(xi), xi ∈ X)] where pa(xi) is the set of parents or antecedents of xi in

BN. A BN denotes joint probability distributions:

$$P(X) = \prod_{x \in X} P(x_i | pa(x_i)\ ) \quad (2)$$

Consider the problem of assigning a class variable V = vi (where i = 1 or 0), given a set of attribute variables X = x1, x2,. . ., xn, A classifier h: X → V is a function that assigns a sample of X to a value of V. The classifier is trained with dataset D consisting of samples over (X, V). In this study, the classification task involves finding an appropriate BN over a set of crash variable X given a data set D. To obtain a suitable BN that fits D, different score metrics are used.

2.2.1. BN learning and the scoring metrics

Learning the BN requires two-stage processes: firstly, a local search is conducted on the data to ascertain the network that is consistent with the observed dependence and independence. Secondly, a score is specified to evaluate how well the dependence or independence in a structure fits the data and searches for a network structure that optimizes the score. To conduct the local search, we apply the HillClimber (Buntine, 1996), K2 (Cooper and Herskovits, 1992), and SimulatedAnnealing (Bouckaert, 1995) algorithms. Through the hill climbing search, the BN arcs are added and deleted to modify the model with no fixed ordering of the variables. The K2 search method applies the hill climbing restricted by K2 ordering of the variables. Simulated annealing (Bouckaert, 1995) uses the general-purpose search method of simulated annealing to find a better scoring BN structure. To score the BNs, BDeu, and Minimum Description Length (MDL), Akaike Information Criterion (AIC) is specified. We choose these search algorithms and the scores for this research mainly because, apart from being widely used, they yield networks that are accurate enough to predict the target concept (Mujalli, et al., 2016). Table 2 presents the descriptions of datasets used and the models developed.

Table 2
Description of the Models Developed and Dataset Used.

Datasets	Classifier	Search method	Score	Code
Original Dataset	BayesNet	HillClimber	Bdeu	OD-BN.HC. BDeu
	BayesNet	HillClimber	MDL	OD-BN.HC.MDL
	BayesNet	HillClimber	AIC	OD-BN.HC.AIC
	BayesNet	K2	Bdeu	OD-BN.K2. Bdeu
	BayesNet	K2	MDL	OD-BN.K2.MDL
	BayesNet	K2	AIC	OD-BN.K2.AIC
	BayesNet	SimulatedAnnealing	Bdeu	OD-BN.SA. BDeu
	BayesNet	SimulatedAnnealing	MDL	OD-BN.SA.MDL
	BayesNet	SimulatedAnnealing	AIC	OD-BN.SA.AIC
SMOTE	BayesNet	HillClimber	Bdeu	SM.BN.HC. BDeu
	BayesNet	HillClimber	MDL	SMBN.HC.MDL
	BayesNet	HillClimber	AIC	SMBN.HC.AIC
	BayesNet	K2	Bdeu	SM.BN.K2. BDeu
	BayesNet	K2	MDL	SMBN.K2.MDL
	BayesNet	K2	AIC	SMBN.K2.AIC
	BayesNet	SimulatedAnnealing	Bdeu	SM.BN.SA.BDeu
	BayesNet	SimulatedAnnealing	MDL	SMBN.SA.MDL
	BayesNet	SimulatedAnnealing	AIC	SMBN.SA.AIC
BL-SMOTE	BayesNet	HillClimber	Bdeu	BL-SM.BN.HC. Bdeu
	BayesNet	HillClimber	MDL	BL-SM.BN.HC. MDL
	BayesNet	HillClimber	AIC	BL-SM.BN.HC.AIC
	BayesNet	K2	Bdeu	BL-SM.BN.K2. BDeu
	BayesNet	K2	MDL	BL-SM.BN.K2. MDL
	BayesNet	K2	AIC	BL-SM.BN.K2.AIC
	BayesNet	SimulatedAnnealing	Bdeu	BL-SM.BN.SA. BDeu
	BayesNet	SimulatedAnnealing	MDL	BL-SM.BN.SA. MDL
	BayesNet	SimulatedAnnealing	AIC	BL-SM.BN.SA.AIC
MWMOTE	BayesNet	HillClimber	Bdeu	MW-BN.HC. BDeu
	BayesNet	HillClimber	MDL	MW-BN.HC.MDL
	BayesNet	HillClimber	AIC	MW-BN.HC.AIC
	BayesNet	K2	Bdeu	MW-BN.K2. BDeu
	BayesNet	K2	MDL	MW-BN.K2.MDL
	BayesNet	K2	AIC	MW-BN.K2.AIC
	BayesNet	SimulatedAnnealing	Bdeu	MW-BN.SA.Bdeu
	BayesNet	SimulatedAnnealing	MDL	MW-BN.SA.MDL
	BayesNet	SimulatedAnnealing	AIC	MW-BN.SA.AIC

Note: OD = original data, HC = HillClimbing, SA = SimulatedAnnealing, SM = SMOTE.

2.3. Assessment metric

In machine learning domains, the quality of dataset is ascertained based on the classification performance of inductive learning algorithm (s) trained on the given dataset (Chawla et al., 2002). Accordingly, four widely-used performance metrics are considered in this paper to measure the BNs classification and prediction performance of the models developed on the original and oversampled datasets to analyze the crash severity. These include accuracy, sensitivity, Mathew Correlation Coefficient (MCC), and the receiver operating characteristic curve (ROC) area (Fawcett, 2006).

$$Accuracy = \frac{tFI + tnFI}{tFI + tnFI + fFI + fnFI} 100\% \quad (3)$$

$$Sensitivity = \frac{tFI}{tFI + fnFI} \quad (4)$$

$$MCC = \frac{(tFI * tnFI) - (fFI * fnFI)}{\sqrt{(tFI + fFI)(tFI + fnFI)(tnFI + fFI)(tnFI + fnFI)}} \quad (5)$$

Where tFI is true fatal injury instances, tnFI indicates true non-fatal injury cases, fFI is false fatal injury cases, and fnFI is false non-fatal injury cases.

Accuracy (Eq. 3) is a measure of the percentage of samples correctly identified by the classifier. The Accuracy metrics provides information on the general performance of the classifier. However, when the data are imbalanced, it becomes inappropriate to use accuracy measure to assess machine learning methods (Fan et al., 2019). The sensitivity measure refers to the proportion of correctly identified fatal injured among the entire fatal injured cases observed in this study. In binary classification, the sensitivity measure only evaluates the model performance in one class and does not provide information about the other. On the other hand, the MCC measure takes all elements of the confusion matrix into consideration to evaluate the model performance across the data (Chicco, 2017). MCC is a correlation coefficient between the observed and predicted classes. Its value ranges between −1 and +1. A coefficient of +1 depicts a perfect prediction, 0 no better than random guess, and −1 indicates total disagreement between the predicted and observed classes. The ROC curve is also considered as a useful measure of classifier overall performance when the data are imbalanced (Mujalli et al., 2016). It plots the true positive rate (sensitivity) on the y-axis against the false positive rate on the x-axis. The ROC area value summarizes the overall classifier performance with a maximum of 1 describing a perfect prediction and a value of 0.50 describing a prediction that is not better than a mere guess.

We apply Wilcoxon signed-rank test to evaluate the statistical significance of the different imbalance treatment methods by a paired comparison (Garcia and Herrera, 2008). Assuming λ pairs of results are tested. First, the difference for each pair is estimated. Let δi be the difference for i = 1…. λ. All the λ differences are then arranged in an ascending order according to the absolute value and a rank assigned to each, starting from rank 1 for the smallest to rank λ for the largest. Where more than one difference is equal, Garcia and Herrera (2008) proposed an averaged rank for each of them. The ranks are converted into a signed rank by assigning signs to the differences. All the + and – ranks are independently summed up and indicated as R + and R-,

respectively. At a specified significance level, the minimum difference found between R + and R- is compared with a critical value obtained from the standard statistical table to confirm the validity or otherwise of the null hypothesis.

2.4. Goodness of fit measure

Akaike Information Criterion (AIC) as an estimator of out-of-sample prediction error is calculated as: AIC = 2[K − LL(β)], where, K is the number of model parameters. Through the simulation studies, it has been shown that AIC has a practical advantage in the ability to select better models compared to BIC (Burnham et al., 2011; Vrieze, 2012). To obtain the best fitted BN for the injury severity analysis, we compute the Akaike Information Criterion (AIC) goodness-of-fit measures where lower values point to a better statistical fit.

3. Results

3.1. Bayes classifiers and data sets

To deal with the imbalanced issue in the original dataset, three new balanced datasets were developed by using SMOTE, BL-SMOTE and MWMOTE oversampling techniques as explained in section 2 above. In resampling, we follow 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. Table 3 presents the data distribution amongst the classes in the original imbalanced and balanced versions. Fig. 2 demonstrates the performance results of 9 models through the experimental process shown in Fig. 1. The results displayed in Fig. 2a–d capture the performance of each Bayesian classifier in terms of ROC area, MCC, Accuracy, and Sensitivity measures. It can be seen that the models developed on the balanced datasets record higher ROC area and MCC values relative to those on the original imbalanced dataset (Fig. 2a and b). Among the balanced datasets, we find that the classifiers obtain higher values when MWMOTE technique is applied for the data imbalance treatment, followed by BL-SMOTE. SMOTE oversampling results in classifiers with abysmal results based on the ROC area and MCC measures. With the accuracy metrics (Fig. 2c), however, we find that the classifiers obtained by using the original imbalanced data record higher values relative to the models based on BL-SMOTE and SMOTE. Ordinarily, such a result is expected since the accuracy measure is highly biased towards the majority class (nFI). Surprisingly, the models developed with the MWMOTE data show better results than those based on the original imbalanced data. Using the sensitivity measures (Fig. 2d), we find that the models obtained by the original imbalanced data are incapable of predicting the FI samples that form the minority group. It is noted that the classifiers developed on BL-SMOTE balanced data produce the highest sensitivity value, followed by MWMOTE and SMOTE.

Table 4 shows the Wilcoxon signed-rank test results which demonstrate that the Bayesian classifiers obtain significantly better results with the balanced data compared to the original imbalanced one. However, when the accuracy metric is used to assess the models, it tends out that the original imbalanced data significantly outperform all the balanced datasets except MWMOTE. For the paired comparison amongst the imbalance treatment methods, we find that the models with MWMOTE show a significantly better performance than those based on SMOTE and BL-SMOTE. However, when the sensitivity metric is applied, BL-SMOTE appears to be the better choice for the imbalance data treatment. Table 5 shows the AIC results. It can be seen that BN.SA.AIC has superior statistical power over the others.

Table 3
Number of Crash Instances and Severity Distribution in the Original and Oversampled Datasets.

Dataset	Total samples	FI samples	nFI samples
Original Data	6837	1353	5484
SMOTE	10,968	5484	5484
BL-SMOTE	10,968	5484	5484
MWMOTE	10,968	5484	5484

3.2. Injury severity analysis

Formally, the direct dependence relationships between the injury severity variable and the rest of the variables as well as the interdependence amongst the different variables are easily established based on the theory of BN. However, since we are interested in the factors contributing to fatal injury severity crashes, we focus on the variables that exhibit direct relationships with the injury severity and ignore the interdependence between predictors as practiced in the literature (de Oña et al., 2011; Mujalli et al., 2016).

Accordingly, Table 6 presents only the variables that have direct dependence with the CIS in the BNs. When MWMOTE is applied for the data balancing, more variables are found to be directly linked to the injury severity compared to the other data balance techniques. As expected, with the use of imbalanced training data, very few variables are apparently associated with the CIS. It suggests that, the data skew can obscure important relation between crash factors and the injury severity, causing erroneous inferences and misleading study conclusions. The result also means that suitable data balance technique must be employed to reduce the influence of imabalance. The models developed by using MWMOTE and BL-SMOTE techniques to treat the traning data imbalance are able to classify and predict the injury severity, suggesting that the variables that are common to these models have strong dependence with the injury severity. These variables include: day of the week (D_WK), time (TME), light condition (LT_C), road description (RDES), location type (LOC_T), collision type (COL_T), (WTHR), road environment (RD_EVMT), road surface type (RS_T), shoulder type (SLD_T), and the number of vehicles involved (N_VEH).

3.3. BN inference

Table 7 depicts the conditions in which the likelihood of a FI is greater than that of a nFI. The table shows that, given evidence for location type to be 'not at intersection' (Location type = 1), light condition to be 'night-light off' (Light condition = 3), road environment to be 'no settlement area' (Road environment = 3) the probability that the injury severity will be fatal is 0.946 and 0.980 in BN.SA.AIC based on MWMOTE and BL-SMOTE, respectively. For a more reliable study conclusion, we consider the significant variables that are common to both models as the most authentic determinants of FI severity crashes, which include off-peak time (Time = 0), non-intersections (Location type = 1), collision involved pedestrian (Collision type = 9), rural road environment (Road environment = 3), good tarred road (Road surface type = 1), roads without shoulders (Shoulder condition), and more than one vehicle involved (No. of vehicles involved = 1). It is noted that some variables which are not significant determinants of FI crashes tend to produce a notable effect when combined with others. For instance, even though lighting condition is not found to be associated with the injury severity, driving under night light off in rural mid block sections is significantly related to FI crashes for both MW.BN.SA.AIC and BL-SM. BN.SA.AIC models with probabilities of 0.946 and 0.980, respectively.

4. Discussion and conclusion

In this study, different Bayesian classifiers are developed based on the treated and untreated datasets and four measures are employed for the performance evaluation. The empirical results reveal that the Bayes classifiers developed on the treated datasets by oversampling can classify the injury severity better than those developed on the original imbalanced data. However, when accuracy measure is employed for the

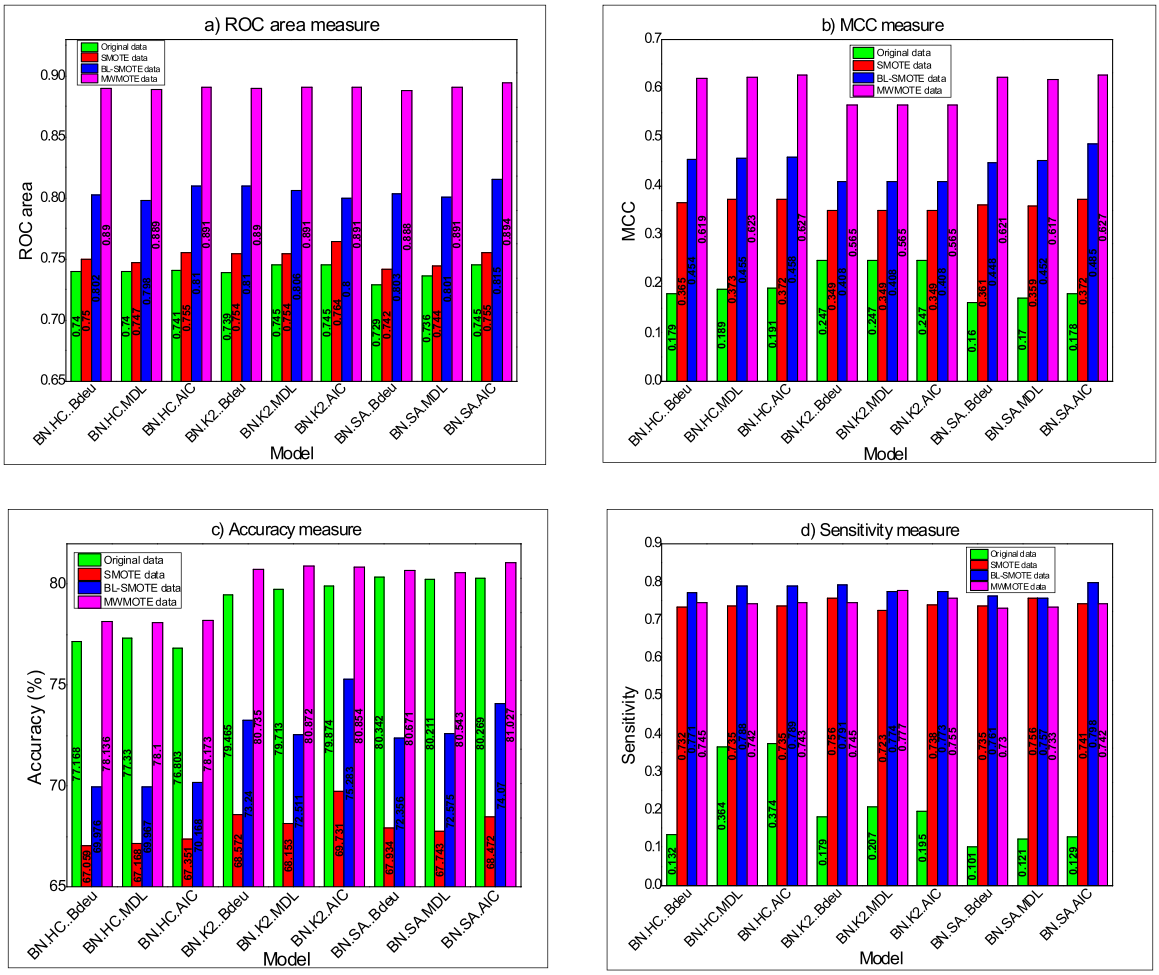

Fig. 2. Model comparison for the different data using (a) ROC area, (b) MCC, (c) Accuracy, and (d) Sensitivity metrics.

Table 4
Wilcoxon Test for the Model Performance in Imbalance and balanced Datasets.

Dataset	ROC area			MCC			Accuracy			Sensitivity		
	R+	R-	P-value	R+	R-	P-value	R+	R-	P-value	R+	R-	P-value
S^a vs. OD^d	45	0	0.006	45	0	0.004	0	45	0.006	45	0	0.006
BL^b vs. OD^d	45	0	0.005	45	0	0.004	0	45	0.006	45	0	0.006
MW^c vs. OD^d	45	0	0.006	45	0	0.004	45	0	0.006	45	0	0.006
BL^b vs. S^a	45	0	0.004	45	0	0.004	45	0	0.006	45	0	0.006
MW^c vs. BL^b	45	0	0.003	45	0	0.003	45	0	0.006	1	44	0.006
MW^c vs. S^a	45	0	0.001	45	0	0.003	45	0	0.006	30	15	0.343

Note: S^a, BL^b, MW^c and OD^d represent SMOTE, BL-SMOTE, MWMOTE and Original data respectively.

Table 5
Goodness-Of-Fit Measure for The BNs Developed on the BL-SMOTE and MWMOTE Datasets.

Data/model	BL-SMOTE	MWMOTE
Perf. Measure	AIC	
BN.HC. Bdeu	−113568.532	−122947.671
BN.HC.MDL	−113592.376	−122984.187
BN.HC.AIC	−113564.806	−122947.671
BN.K2. Bdeu	−113568.532	−122947.671
BN.K2.MDL	−113592.376	−122984.187
BN.K2.AIC	−113564.806	−122947.671
BN.SA. Bdeu	−112245.706	−121992.703
BN.SA.MDL	−111629.535	−121635.059
BN.SA.AIC	**−113712.801**	**−123653.485**

model evaluation, it tends out that using the original imbalanced data is preferable to the preprocessed ones obtained by BL-SMOTE and SMOTE techniques. At the same time, the models developed by using the imbalanced data fail to identify the FI samples that form the minority. This finding confirms that it is improper to use the accuracy metric to evaluate a model's classification performance in the imbalanced data (Leevy et al., 2018).

When the entire probability threshold is considered in the performance measure (i.e., ROC area, accuracy, and MCC metrics), MWMOTE oversampling treatment seems to be more effective, followed by BL-SMOTE with SMOTE being the least productive method. The blind oversampling by SMOTE might have led to the creation of too many needless samples that tend to disrupt the class boundary and exacerbate the learning task. This may explain why the application of SMOTE technique for the data imbalance treatment leads to classifiers that

Table 6
Relationship between the Crash Severity and Other Variables in the BN.SA.AIC Model Developed Using the Original Imbalanced Data and the Balanced Counterparts.

dataset	Original data			SMOTE data		
Dependence relationship between CIS and other variables	Severity	→	Day of week	–	–	–
	Severity	→	Time	Severity	→	Time
	Severity	→	Light condition	Severity	→	Light condition
	Severity	→	Road description	Severity	→	Road description
	Severity	→	Location type	Severity	→	Location type
	Severity	→	Collision type	Severity	→	Collision type
	Severity	→	Hit and Run	Severity	→	Hit and Run
	–	–	–	Severity	→	Weather
	–	–	–	Severity	→	Season
	–	–	–	Severity	→	Road environment
	–	–	–	Severity	→	Traffic control
	–	–	–	–	–	–
	–	–	–	–	–	–
	–	–	–	–	–	–
	–	–	–	–	–	–
	–	–	–	–	–	–
No. of dependence[a]	7			10		

dataset	BL-SMOTE data			MWMOTE data		
Dependence relationship between CIS and other variables	Severity	→	**Day of week**	Severity	→	**Day of week**
	Severity	→	**Time**	Severity	→	**Time**
	Severity	→	**Light condition**	Severity	→	**Light condition**
	Severity	→	**Road description**	Severity	→	**Road description**
	Severity	→	**Location type**	Severity	→	**Location type**
	Severity	→	**Collision type**	Severity	→	**Collision type**
	Severity	→	**Hit and Run**	–	–	–
	Severity	→	**Weather**	Severity	→	**Weather**
	–	–	–	Severity	→	Season
	Severity	→	**Road environment**	Severity	→	**Road environment**
	–	–	–	Severity	→	Traffic control
	Severity	→	**Road surface type**	Severity	→	**Road surface type**
	Severity	→	**Shoulder condition**	Severity	→	**Shoulder condition**
	Severity	→	**No. of Vehicles involved**	Severity	→	**No. of Vehicles involved**
	–	–	–	Severity	→	Road works
	–	–	–	Severity	→	Surface condition
No. of dependence[a]	12			15		

Note: The symbol "→" indicates direct dependence relationship.
Bold letters indicate the variables that had direct dependence with CIS and common to the BN.SA.AIC model in all the datasets.
[a] This row denotes the number of dependence relationships established with the CIS in each classifier.

Table 7
Inference Results for Variables That Are Significantly Associated with FI Crashes in The BNs Developed.

No. Var	Condition satisfied	Prob. of FI in MW.BN.SA. AIC	Prob. of FI in BL-SM.BN.SA. AIC
1	P(Severity = FI|Time = 0)	0.899	0.994
2	P(Severity = FI|Location type = 1, Light condition = 3, RD_EVMT = 3)	0.946	0.98
3	P(Severity = FI|Collision type = 9, Location type = 2, Road environment = 1)	0.415	0.548
4	P(Severity = FI| Road environment = 3, Road description = 3, Light condition = 2)	0.791	0.95
5	P(Severity = FI|Road surface type = 1, Road description = 1, Light condition = 4)	0.921	0.949
6	P(Severity = FI|Shoulder condition = 4, Road environment = 3)	0.766	0.947
7	P(Severity = FI|No. of vehicles involved = 1, Collision type = 8)	0.997	0.996

produce suboptimal results in predicting the injury severity. In the case of BL-SMOTE, the KNN-based criterion set to identify the borderline minority samples may fail to detect them in some complex data scenarios. In such an instance, the synthetic minority samples may be erroneously located deep inside the region of the other class and make the learning task challenging (Barua, et al., 2014). It interprets why the classifiers developed using the BL-SMOTE data produce a lower performance compared to those founded on MWMOTE in this category.

However, when a single probability threshold (i.e., sensitivity metrics) is chosen for the model evaluation, BL-SMOTE appears to be the better imbalance treatment choice compared to MWMOTE. The technique adopted in BL-SMOTE can lead to wrong synthetic samples and expand the region of the minority samples erroneously (Barua, et al., 2014). The consequence is that the region of minority class will swell into the majority class region to enhance the classification accuracy on the minority samples, i.e., sensitivity. At the same time, due to the wrong expansion of the minority class region, the classifiers developed on BL-SMOTE data misclassify many of the majority class samples (nFI) as minority (FI). This explains why BL-SMOTE appears superior with the sensitivity measure but fails to maintain its position when the other performance metrics are applied. The sensitivity metric only accounts for a single to evaluate the classifiers. Therefore, it might not be sufficient to judge the overall performance of the methods used. However, since imbalance treatment methods seek to enhance the prediction performance of the minority class, we consider it a relevant measure of model performance. Hence, BL-SMOTE can also be a useful imbalance treatment method for the crash data.

The results clearly indicate the classifiers developed on the balanced datasets are more useful to classify crashes according to the injury severity than the original imbalanced data. Among the data balancing techniques, it appears that MWMOTE is the best treatment method if the objective is to improve the overall performance of the classifier. On the other hand, if the focus is to improve the performance for the minority class without regarding for the other(s), BL-SMOTE may be the optimal way to proceed.

Therefore, by applying MWMOTE and BL-SMOTE together with the BN.SA.AIC model, eleven variables are found to be strongly correlated to the injury severity. The inference results of the BNs suggest that seven out of eleven variables are significantly linked to FI severity crashes. These include off-peak time, non-intersections, collision involved pedestrian, rural road environment, good tarred road, roads without shoulders, and multiple vehicles involved crash.

The findings herein suggest that off-peak time and night light off are of the influencing factors of FI severity crashes, which is consistent with the literatures (Haleem and Gan, 2011; Michalaki et al., 2015). In Ghana, many long-distance passenger vehicles (buses and mini-buses) are often scheduled to travel during the night. It is to evade the deadlock traffic and reduce travel time and the surface roughness coupled with the heat easily cause the tire bust for day time travel schedule. However, driving continuously throughout the night often causes fatigue, which is believed to be one major contributing factor to the gory accidents in recent times in Ghana. Thus, this finding is intuitively linked to several reasons such as fatigue-driving and the propensity towards the excessive speeding that may lead to FI severity crashes (Moral-García et al., 2019; Mujalli et al., 2016). The observation indicates that reducing the fatigue-related crashes may require measures to ensure more than one driver for each of such passenger transport vehicles. Alternatively, rest areas could be provided at the suitable locations along the Ghanian highways for drivers to take intermediate break that is necessary for sound driving.

In terms of location, we identify that crashes occurring at non-intersection areas of rural highways in night light-off conditions are significantly associated with FI severity. It is shown that crashes occurred in the rural settings, especially on road segments without lighting are highly associated with FI (Rab et al., 2018). In fact, rural highways account for the majority of traffic related fatalities in Ghana. Arguably, these roads are narrow for the excessive speed induced by the low traffic volume often observed in the night. Rural highways in Ghana are notably two-lane bidirectional with an average width of 7.5 m which might not be wide enough for an errant vehicle to recover. Providing sufficiently wide driving lanes would allow for safer overtaking maneuvers which may reduce crash frequency and severity even when speed increases (Frost and Morrall, 1998). It is also noted that the increase in average speed resulting from the upgrading of these rural roads increases the risk of pedestrians. Unfortunately, safety awareness programs which are necessary to cause change in behavior of these pedestrians are often missing during the construction stage or even after. Therefore, fatalities amongst pedestrians tend to rise in rural settings when the road suddenly becomes motorable. Similarly, at horizontal curve areas of rural highways, crashes which take place during night with no-light conditions are found to be significantly linked to fatal injury severity. The horizontal curve alignment coupled with no lighting may cause driver's visibility impairments that eventually can cause FI severity in the event of a crash (Abdel-Aty et al., 2011; Uddin and Huynh, 2018).

In terms of the crash characteristics, the findings show that good tarred roads in urban settings appear to influence the occurrence of FI crashes. The results demonstrate that a collision involved with a pedestrian at urban intersections has a higher probability of a FI severity (Abdel-Aty and Keller, 2005; Fountas and Anastasopoulos, 2018; Haleem and Abdel-Aty, 2010; Jiang et al., 2013). This result can be explained by many factors. First, the favorable roadway condition is likely to induce unsafe driving behaviors such as overspeeding, which may in turn lead to FI severity in the crash occurrence (Mujalli et al., 2016). Formally, driving beyond the maximum speed limit of 50 km/hr in an urban setting in Ghana constitutes a traffic offence punishable by law. Actually, the likelihood of a fatality involving vehicle(s) under the stipulated speed limit is extremely low. The fact that urban intersection crashes are significantly associated with FI, suggests that speed limit violations are common on the Ghanaian urban roads. This might be so, because the violation of speed limit is considered a minor traffic offence with a paltry amount as penalty that is not punitive enough to deter prospective offenders. Besides, the necessary technology to monitor the speeding on urban roads in Ghana is not available. The consequence is that law enforcement agencies fall below the average in implementing the speed related regulations. Also, due to a lack of safety awareness, many pedestrians in Ghana cross the road at non-designated points and without necessary precautionary measures, which consequently lead to a FI crash. To this effect, it is important to develop rigorous safety campaigns and provide adequate pedestrian crossing facilities and speed calming measures at regular intersections in the Ghanaian cities.

We also find that crashes occurring on the roadways where there are no shoulders are closely associated with FI. Road shoulders serve to accommodate stopped/distressed vehicles, and provide traversable area for the recovery of errant vehicles. Without shoulders, distressed vehicles left on the carriageway increase the likelihood of FI collisions, particularly so, in areas of poor visibility (Abdel-Aty, et al., 2011; Ahmed, 2013). This finding reinforces our earlier claim of insufficient width for the rural roads in Ghana and the devastating safety impact. For a two-lane two-way road, which is common in Ghana, the importance of supplying the road shoulders becomes more acute.

Furthermore, multiple vehicle crashes in which a parked vehicle is hit show a higher probability to be fatal than non-fatal. The result is in line with the study by (Kitali et al., 2018) who investigated the factors contributing to injury severity of crashes in Florida and found that multi-vehicle collisions involving three or more automobiles were more severe than others. This finding may be attributed to many factors including over speeding which imposes severe impacts on the occupants, resulting in fatalities during crashes (Bahouth and Digges, 2005). In Ghana, it is observed that broken down/disabled vehicles on the highways are not disposed in a timely manner. Such vehicles left on the roadway without necessary warning signs are easily run into, especially in poor driver visibility conditions. To improve the road safety, it is imperative for the authorities to ensure that broken down vehicles on the highways are removed promptly. This study has also shed lights on the factors contributing to FI severity crashes on the Ghanaian highways. Since fatal crashes are socio-economically onerous and unintended, they can be prevented if the situations leading to the crash are proactively identified and acted upon.

In this study, BNs models built on the original imbalanced dataset failed woefully in determining the influential factors of FI crashes on the Ghanaian highways. On the other hand, the oversampling methods have been shown useful for the imbalance CIS data analysis. The general conclusion that can be deduced from this research is that, in oversampling CIS data for the analysis, different methods must be investigated to ascertain a more suitable one. Accordingly, it is demonstrated herein that MWMOTE oversampling method can improve the ability of BNs to correctly classify fatal injury samples without compromising the classification accuracy of the other class (i.e., non-fatal). Thus, controlling for features known to be associated with crash severity, such as roadway, environmental, vehicular and driver characteristics, we can accurately predict the specific severity outcome of an individual in the event of a crash. The approach provides advanced knowledge of the risk factors and helps road safety experts and city authorities to be proactive in the implementation of functional countermeasures for improving the road safety. We have also demonstrated that the accuracy measure is not appropriate to assess the CIS models, particularly in the imbalance data. Thus, researchers must use it cautiously to avoid the deceptive conclusions.

Due to the difficulty of reporting a traffic crash in a developing country like Ghana, this study comes with some limitations. Key limitations are: 1) information about driver and vehicle characteristics, which have potential effects on crash severity are missing in the database. Meanwhile, the strategic document aiming to address the alarming rate of fatal injury crashes in Ghana cites the vehicle and human factors as thematic areas of focus (http://www.nrsc.gov.gh, 2011). Thus, their absence in the crash database may not allow for the systematic study to

reveal their influences on fatal injury crashes; and 2) possible mislabels and errors in attribute data are not accounted for. Regardless of the quality of oversampling method, the data noise may still pose limitations to the classifier performance (Barua, et al., 2014). Our future work will devise a framework to simultaneously handle crash data imbalance and noise, since their presence is counterproductive for prediction and data analysis (Drummond and Holte, 2005; Saez et al., 2015).

CRediT authorship contribution statement

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

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

This study is funded by the National Nature Science Foundation of China (NSFC 71771191 and 51608455) and Sichuan Provincial Science and Technology Innovation Talents Fund (2019JDRC0023).

References

Abdel-Aty, M.A., Abdelwahab, H.T., 2004. Predicting injury severity levels in traffic crashes: a modeling comparison. J. Transp. Eng. 130 (2), 204–210.
Abdel-Aty, M., Ekram, A., Huang, H., Choi, K., 2011. A study of visibility obstruction related crashes due to fog and smoke. Accident Anal. Prev 43, 1730–1737.
Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accid. Anal. Prev. 37 (3), 417–425.
Ahmed, I., 2013. Road infrastructure and road safety. Transp. Commun. Bull. Asia Pac. 83, 19–25.
Bahouth, J., Digges, K., 2005. Characteristics of multiple impact crashes that produce serious injuries. In: Proceedings of the 19th International Technical Conference on the Enhanced Safety of Vehicles. Washington DC, USA.
Barua, S., Islam, M.M., Yao, X., Murase, K., 2014. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans. Knowl. Data Eng. 26 (2), 405–425.
Bouckaert, R.R., 1995. Bayesian Belief Networks: From Construction to Inference.
Buntine, W., 1996. A guide to the literature on learning probabilistic networks from data. IEEE Trans. Knowl. Data Eng. 8 (2), 195–210.
Burnham, K.P., Anderson, D.R., Huyvaert, K.P., 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65 (1), 23–35.
Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid. Anal. Prev. 38 (5), 1019–1027.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Min. 10 (1), 35.
Chong, M.M., Abraham, A., Paprzycki, M., 2004. Traffic Accident Analysis Using Decision Trees and Neural Networks. arXiv preprint cs/0405050.
Chong, M., Abraham, A., Paprzycki, M., 2005. Traffic accident analysis using machine learning paradigms. Informatica 29 (1).
Cigdem, A., OZDEN, C., 2018. Predicting the severity of motor vehicle accident injuries in adana-turkey using machine learning methods and detailed meteorological data. Int. J. Intell. Syst. Appl. Eng. 6 (1), 72–79.
Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. 9 (4), 309–347.
de Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid. Anal. Prev. 43 (1), 402–411.
Drummond, C., Holte, R.C., 2005. Severe class imbalance: why better algorithms aren't the answer. European Conference on Machine Learning.
Fan, Y., Zhang, G., Ma, J., Lee, J., Meng, T., Zhang, X., Jiang, X., 2019. Comprehensive evaluation of signal-coordinated arterials on traffic safety. Anal. Methods Accid. Res.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.
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.
Fountas, G., Anastasopoulos, P.C., Abdel-Aty, M., 2018. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. Anal. Methods Accid. Res. 18, 57–68.
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 34, 1–15.
Garcia, S., Herrera, F., 2008. An extension on``statistical comparisons of classifiers over multiple data sets''for all pairwise comparisons. J. Mach. Learn. Res. 9 (Dec), 2677–2694.
García, S., Luengo, J., Herrera, F., 2015. A Data Mining Software Package Including Data Preparation and Reduction: KEEL.
Gregoriades, A., 2007. Towards a user-centred road safety management method based on road traffic simulation. 2007 Winter Simulation Conference.
Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. J. Safety Res. 41 (4), 347–357.
Haleem, K., Gan, A., 2011. Identifying traditional and nontraditional predictors of crash injury severity on major urban roadways. J. Crash Prev. Inj. Control. 12 (3), 223–234.
Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International Conference on Intelligent Computing.
Harris, J., Brunner, G., Faber, B., 2014. Statistical Software Package. World Environmental & Water Resources Congress.
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284.
He, H., Ma, Y., 2013. Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons.
Holland, P.W., 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81 (396), 945–960.
http://www.nrsc.gov.gh, 2011. National Road Safety Strategy III (2011–2020). Retrieved Date Accessed, 11.
Jeong, H., Jang, Y., Bowman, P.J., Masoud, N., 2018. Classification of motor vehicle crash injury severity: a hybrid approach for imbalanced data. Accid. Anal. Prev. 120, 250–261.
Jiang, X., Huang, B., Zaretzki, R.L., Richards, S., Yan, X., Zhang, H., 2013. Investigating the influence of curbs on single-vehicle crash injury severity utilizing zero-inflated ordered probit models. Accid. Anal. Prev. 57 (3), 55–66.
Kamal, B., Li, T., Chubato, W.Y., Yahaya, M., 2017. Enhancing software defect prediction using supervised-learning based framework. International Conference on Intelligent Systems and Knowledge Engineering (ISKE).
Kitali, A.E., Kidando, E., Martz, P., Alluri, P., Sando, T., Moses, R., Lentz, R., 2018. Evaluating factors influencing the severity of three-plus multiple-vehicle crashes using real-time traffic data. Transp. Res. Rec. 2672 (38), 128–137.
Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A., Seliya, N., 2018. A survey on addressing high-class imbalance in big data. J. Big Data 5 (1), 42.
Li, H., Sun, J., 2012. Forecasting business failure: the use of nearest-neighbour support vectors and correcting imbalanced samples–evidence from the Chinese hotel industry. Tour. Manag. 33 (3), 622–634.
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.
Mathew, J., Luo, M., Pang, C.K., Chan, H.L., 2015. Kernel-based SMOTE for SVM classification of imbalanced datasets. Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE.
Michalaki, P., Quddus, M.A., Pitfield, D., Huetson, A., 2015. Exploring the factors affecting motorway accident severity in England using the generalised ordered logistic regression model. J. Safety Res. 55, 89–97.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72.
Moral-García, S., Castellano, J.G., Mantas, C.J., Montella, A., Abellán, J., 2019. Decision tree ensemble method for analyzing traffic accidents of novice drivers in urban areas. Entropy 21 (4), 360.
Mujalli, R.O., López, G., Garach, L., 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accid. Anal. Prev. 88, 37–51.
Mussone, L., Bassani, M., Masci, P., 2017. Analysis of factors affecting the severity of crashes in urban road intersections. Accid. Anal. Prev. 103, 112–122.
Oña, J.D., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50 (2), 1151–1160.
Rab, M.A., Qi, Y., Fries, R.N., 2018. Comparison of Contributing Factors to Pedestrian Crossing Crash Severity at Locations with Different Controls in Illinois.
Saez, J.A., Luengo, J., Stefanowski, J., Herrera, F., 2015. SMOTEIPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf. Sci. 291 (291), 184–203.
Sáez, J.A., Luengo, J., Stefanowski, J., Herrera, F., 2015. SMOTE–IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf. Sci. 291, 184–203.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 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., Stütz, R., Laaha, G., Melcher, M., 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.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Safety Res. 27 (3), 183–194.
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.
Thammasiri, D., Delen, D., Meesad, P., Kasap, N., 2014. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Expert Syst. Appl. 41 (2), 321–330.
Theofilatos, A., Graham, D., Yannis, G., 2012. Factors affecting accident severity inside and outside urban areas in Greece. Traffic Inj. Prev. 13 (5), 458–467.
Triguero, I., del Río, S., López, V., Bacardit, J., Benítez, J.M., Herrera, F., 2015. ROSEFW-RF: the winner algorithm for the ECBDL'14 big data competition: an extremely imbalanced big data bioinformatics problem. Knowledge Based Syst. 87, 69–79.
Uddin, M., Huynh, N., 2018. Factors influencing injury severity of crashes involving HAZMAT trucks. Int. J. Transp. Sci. Technol. 7 (1), 1–9.
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.
Vrieze, S.I., 2012. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol. Methods 17 (2), 228.
Wahab, L., Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS One 14 (4), e0214966.
Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
Yahaya, M., Jiang, X., Fu, C., Bashir, K., Fan, W., 2019. Enhancing crash injury severity prediction on imbalanced crash data by sampling technique with variable selection. 2019 IEEE Intelligent Transportation Systems Conference (ITSC).
Yahaya, M., Fan, W., Fu, C., Li, X., Su, Y., Jiang, X., 2020. A machine-learning method for improving crash injury severity analysis: a case study of work zone crashes in Cairo, Egypt. Int. J. Inj. Contr. Saf. Promot. 1–10.
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., Tan, K.C., Li, H., Hong, G.S., 2018a. A cost-sensitive deep belief network for imbalanced classification. IEEE Trans. Neural Netw. Learn. Syst. (99), 1–14.
Zhang, K., Hassan, M., Yahaya, M., Yang, S., 2018b. Analysis of work-zone crashes using the ordered probit model with factor analysis in Egypt. J. Adv. Transp. 2018.
Zong, F., Xu, H., Zhang, H., 2013. Prediction for traffic accident severity: comparing the Bayesian network and regression models. Math. Probl. Eng. 2013.