Article

Analysis of Run-Off-Road Accidents by Association Rule Mining and Geographic Information System Techniques on Imbalanced Datasets

Feifeng Jiang ^{1,*}, Kwok Kit Richard Yuen ¹ , Eric Wai Ming Lee ¹ and Jun Ma ²

¹ Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China; bckkyuen@cityu.edu.hk (K.K.R.Y.); ericlee@cityu.edu.hk (E.W.M.L.)

² Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong, China; luckyma1990@gmail.com

* Correspondence: feifjiang2-c@my.cityu.edu.hk

Received: 17 May 2020; Accepted: 10 June 2020; Published: 15 June 2020

Abstract: Run-off-road (ROR) accidents cause a large proportion of fatalities on roads. Exploring key factors is an effective method to reduce fatalities and improve safety sustainability. However, some limitations exist in current studies: (1) Datasets of ROR accidents have imbalance problems, in which the samples of fatal accidents (FA) are much less than non-fatal accidents (NFA). Data mining methods on such imbalanced datasets make the results biased. (2) Few studies conducted spatial analysis of ROR accidents in visualization. Therefore, this study proposes an association rule mining (ARM)-based framework to analyze ROR accidents on imbalanced datasets. A novel method is proposed to address the imbalance problem and ARM is applied to analyze accident severity. Geographic information system (GIS) is adopted for spatial analysis of ROR accidents. The proposed framework is applied to ROR accidents in Victoria, Australia. Six FA factors and seven NFA factors are identified from two-item rules. The results of three-item rules indicate factors acting interactively increase the likelihood of FA or NFA. Hot spots of ROR accidents are presented by GIS maps. Effective measures are accordingly proposed to improve road safety. Compared with traditional data-balancing methods, the proposed framework has been validated to provide more robust and reliable results on imbalanced datasets.

Keywords: run-off-road accidents; imbalanced dataset; bootstrap-resampling-data-balancing method; association rule mining; ensemble method; geographic information system

1. Introduction

Road traffic accidents are considered as a public safety problem globally [1,2]. They cause great casualties, economic losses, and traffic congestion each year [3,4]. The World Health Organization (WHO) reported that over 1.35 million people died, and 50 million people were injured each year due to traffic crashes [5]. The economic cost of traffic crashes was estimated as 3% of gross domestic product (GDP) globally, and up to 5% for low-income and middle-income countries [5]. Among all the traffic accident types, run-off-road (ROR) accidents are an important subset of traffic accidents because they yield most of the fatalities and severe injuries on roads. However, existing studies on ROR accidents are still very limited.

ROR accidents (commonly known as roadway departure accidents) are a type of road crash that occur when a vehicle crosses an edge line or leaves the designated roadway. ROR accidents are considered one of the most dangerous accidents because they cause most fatalities on roads [6]. As reported by Roads Corporation of Victoria, ROR accidents accounted for approximately 60% of

road fatalities in 2015 in Victoria, Australia. This fatality rate increased by about 10% compared with 2014. Similar trends are presented in other countries, for example, 54% of road fatalities were caused by ROR accidents in the U.S. between 2013 to 2015. Therefore, analyzing ROR accidents is considerably important to reduce road fatalities and improve safety sustainability.

Exploring key factors associated with accident severity is an effective method to reduce ROR accidents and improve road safety [7–9]. Most existing studies applied regression methods for the severity analysis of ROR accidents. For example, Al-Bdairi and Hernandez (2017) used an ordered random parameter probit model to identify contributory factors related to injury severity in large-truck ROR crashes. Gong and Fan (2017) used a mixed logit model to investigate the factors associated with injury severity in single-vehicle ROR accidents [10]. However, the regression methods commonly have underlying assumptions (e.g., linear relationship, the type of distribution) [11,12]. If the assumptions (e.g., linear relationship) are not fulfilled in practical applications, the results can be biased. Besides, the existing studies only identified the individual factors related to accident severity, and seldom explored the interactive relationship between multiple factors. Therefore, a more appropriate method is essential for the severity analysis of ROR accidents. Association rule mining (ARM) is an emerging data mining method without predefined assumptions to discover interesting relations between variables in large datasets [13]. It can not only identify individual key factors but also interpret interesting relationships among multiple factors for accident severity analysis [13–17]. Therefore, this study tends to explore the application of ARM on the severity analysis of ROR accidents.

Datasets of ROR accidents usually have imbalance problems. The imbalance problems indicate a situation where instances of one class (minority class) are far outnumbered by the other class (majority class) [18]. For traffic accidents, the more serious the accident is, the fewer samples can be collected—there are far fewer samples of fatal accidents than non-fatal accidents. To obtain roughly balanced datasets, the existing studies usually classified the crash severity into two categories: Serious accidents (i.e., accidents with fatal or serious injuries) and non-serious accidents (i.e., accidents with slight or no injuries). However, few studies have been conducted to explore the key factors associated with fatal accidents. It is significantly important to specifically analyze fatal accidents because effective measures can be proposed directly to reduce fatalities in ROR accidents and improve road safety. To fill this research gap, this study aims to use ARM for accident severity analysis, especially to identify the contributory factors related to fatal accidents.

However, the instances of fatal accidents (FA) are much fewer than instances of non-fatal accidents (NFA), resulting in extremely imbalanced datasets. Most data mining methods have an assumption that the dataset is balanced [11,19]. Data mining methods on such extremely imbalanced datasets may make the results deteriorated or biased [11,19]. For example, because of rare cases of FA, regression methods tend to overlook those fatal cases, and their discovered relationships may be skewed to non-fatal cases [20,21]. ARM method does not have underlying assumptions; however, determining proper parameter thresholds of ARM is difficult on such imbalanced datasets. If the thresholds are set too high, no rules about minority class can be found. If the thresholds are set very low, it leads to a combinatorial explosion and most of the rules are meaningless [21–23]. It is worth noting that minority class (e.g., FA) is often the class of particular interest. Therefore, it is important to balance the class distribution to improve the performance of accident severity analysis.

To address the imbalance problem, sampling methods are widely used in existing studies to balance class distribution and are divided into three groups: Under-sampling, over-sampling, and mix-sampling methods [24,25]. Under-sampling methods randomly eliminate instances from the majority class to obtain a desirable balanced dataset. However, this method may lose much potentially useful information contained in the dropped instances [24,26]. Over-sampling methods aim to create instances of minority class (i.e., by duplication) while keeping the majority class unchanged. However, this method may increase the likelihood of overfitting in the induction process (i.e., the duplication process) [24]. Mix-sampling methods eliminate instances from the majority class and create new instances of minority class to produce a balanced dataset. This method contains the drawbacks of

under-sampling and over-sampling methods [24,26]. Another limitation of the sampling methods is randomness. The sampling methods construct one balanced dataset by randomly eliminating instances from the majority class or creating instances of the minority class. Results obtained from one balanced dataset are deficient in robustness and reliability. Therefore, proposing a method to address the imbalance problem is necessary to improve the robustness and persuasiveness of the results.

Records of ROR accidents often contain longitude and latitude information, which endows the accidents with spatial properties. However, the existing studies only display the key factors associated with accident severity, and seldom provide a spatial analysis of ROR accidents. Geographic information systems (GIS) are an efficient platform to provide graphical outputs in visualization [27–31]. Spatial analysis of GIS can identify hot spots by the calculation of density distribution. To fill the research gap, this study aims to apply GIS for spatial analysis to identify hot spots of ROR accidents related to key factors. Therefore, policymakers can refer to these GIS maps when making decisions.

To fill the above research gaps and limitations, this study proposes a framework to explore key factors associated with accident severity on imbalanced datasets of ROR accidents, especially the key factors of fatal accidents. This study proposes a bootstrap-resampling-data-balancing method (BRDB method) to address the imbalance problems in ROR accidents, which converts an imbalanced dataset into multiple balanced datasets. ARM is applied to each balanced dataset to identify the rules associated with accident severity. An ensemble method is proposed to integrate the rules to improve the robustness and reliability of the results. GIS is adopted for spatial analysis to provide hot spots of the ROR accidents related to the key factors in visualization. The proposed framework is applied to a case study of ROR accidents in Victoria, Australia. Two-item rules and three-item rules associated with accident severity are explored in this study. The hot spots of ROR accidents related to the key factors are displayed and effective measures are proposed to reduce ROR accidents in hot spots and improve road safety. The necessity of applying data-balancing methods is validated in this study. The proposed framework is compared with traditional data-balancing methods to validate its effectiveness and robustness on imbalanced datasets.

The main contributions of this study can be summarized as follows:

1. Few studies were conducted to analyze ROR accidents, which cause most fatalities on roads. This paper is one of the few studies to explore the key factors associated with the accident severity of ROR accidents, especially the key factors related to fatal accidents.
2. Datasets of ROR accidents are extremely imbalanced, in which FA are much less than NFA. Data mining methods on such extremely imbalanced datasets make the results deteriorated or biased. This study proposes a novel method to address the imbalance problem in ROR accidents. The proposed method can avoid the randomness caused by sampling methods and improve the robustness and reliability of the results on imbalanced datasets.
3. This study applies ARM for the severity analysis of ROR accidents. It can not only explore individual key factors associated with injury severity but also identify the interactive relationship between multiple factors in ROR accidents.
4. This is one of the few papers to conduct spatial analysis of ROR accidents by GIS technology. The hot spots of ROR accidents associated with key factors can be presented in GIS maps. Policymakers can refer to these maps when making decisions.

The remainder of this paper is organized as follows: Section 2 introduces the proposed framework; Section 3 describes the case study of ROR accidents; Section 4 presents the results of the case study by implementing the proposed framework; Section 5 displays the discussions. Finally, conclusions are remarked in Section 6.

2. Methodology

This study proposes a framework to explore key factors associated with accident severity on imbalanced datasets of ROR accidents. As is shown in Figure 1, the framework can be summarized

into four parts: (1) Data balancing (BRDB method); (2) ARM; (3) ensemble method, and (4) GIS analysis. This study first uses the BRDB method to convert an imbalanced dataset into multiple balanced datasets. ARM is then applied to these balanced datasets separately to identify the rules associated with accident severity. An ensemble method is proposed to integrate the rules to improve the robustness and reliability of the results. GIS is adopted for spatial analysis to provide hot spots of the ROR accidents in visualization.

Figure 1. The proposed framework. Note: BRDB Method, bootstrap-resampling-data-balancing method; ARM, association rule mining; FA, fatal accidents; NFA: non-fatal accidents; CI: confidence interval; GIS: geographic information system.

2.1. BRDB Method

This study proposes a novel BRDB method to address the imbalance problems in ROR accidents. BRDB method transfers an imbalanced dataset into multiple balanced datasets. It firstly uses the bootstrap resampling method to produce bootstrap subsets from the majority class (NFA), in which the number of instances is the same as that of minority class (FA). This study then combines each bootstrap subset of majority class (NFA) with instances of minority class (FA) to create multiple balanced datasets. In the case that enough balanced datasets are provided, the proposed methodology can avoid the randomness caused by sampling methods. This can improve the robustness of the results obtained from imbalanced datasets.

2.1.1. Bootstrap Resampling Method

The bootstrap resampling method is an important component of the proposed BRDB method for data-balancing, which is a resampling method to obtain statistical inference of a target population [32,33].

This study uses the bootstrap resampling method to produce bootstrap subsets from the majority class (NFA), with the purpose that bootstrap subsets of NFA can represent the true distribution of NFA.

Figure 2 shows the basic scheme of the bootstrap resampling method. A bootstrap subset is taken from the observed data by randomly resampling with replacement. This resampling process is repeated n times; therefore, n bootstrap subsets can be obtained with the same size. A statistic of interest (also called bootstrap statistic) can be calculated from each bootstrap subset. It is worth noting that the extracted bootstrap subsets can be quite good approximations of the observed data in the case that enough bootstrap subsets are provided. Therefore, the distribution of the bootstrap statistic can represent the true distribution of the observed data [34].

Figure 2. Description of the bootstrap resampling method.

Confidence interval (CI) of the bootstrap statistic is often selected to present the true distribution of the observed data. The equations of CI calculation in the bootstrap resampling method are listed below.

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^n \theta_i^* \quad (1)$$

where $\bar{\theta}$ is the mean value of the bootstrap statistic of bootstrap subsets; θ_i^* is a statistical value calculated from each bootstrap subset; n is the number of bootstrap subsets.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (\theta_i^* - \bar{\theta})^2}{n-1}} \quad (2)$$

where σ is the standard deviation of the bootstrap statistic of bootstrap subsets; θ_i^* , $\bar{\theta}$, and n are interpreted in Equation (1).

$$CI^m = \bar{\theta} \pm t^* \frac{\sigma}{\sqrt{n}} \quad (3)$$

where CI is the confidence interval of the bootstrap statistic of the bootstrap subsets; m is the level of confidence interval (e.g., 90%, 95%); t^* is the upper $(1-m)/2$ critical value with t distribution and $n-1$ degrees of freedom; $\bar{\theta}$, σ , and n are interpreted in Equations (1) and (2).

2.1.2. Process of BRDB Method

The process of the BRDB method is shown in Figure 1. The BRDB method transfers an imbalanced dataset into multiple balanced datasets. It firstly uses the bootstrap resampling method to extract

bootstrap subsets from NFA. The size of each bootstrap subset is the same as that of FA. As is shown in Figure 2, these bootstrap subsets can be good representatives of NFA in the case that enough bootstrap subsets are extracted. This study then combines each bootstrap subset of NFA with instances of FA to create multiple balanced datasets. Each balanced dataset contains instances of FA and NFA with a ratio of 50%:50%. A statistic of interest (also called bootstrap statistic) can be calculated from each balanced dataset. The distribution of the bootstrap statistic of the balanced datasets can represent the true distribution of the observed data. These balanced datasets can be good representatives of the original datasets in the case that enough balanced datasets are provided.

The equations of BRDB method are listed as follows:

$$T_{FA} = \{FA_1, FA_2, \dots, FA_M\} \quad (4)$$

$$T_{NFA} = \{NFA_1, NFA_2, \dots, NFA_N\} \quad (5)$$

where T_{FA} is a set that includes M instances of FA; T_{NFA} is a set that includes N instances of NFA; $FA_1, FA_2, \dots, FA_M$ are the instances of FA; $NFA_1, NFA_2, \dots, NFA_N$ are the instances of NFA.

$$S_{NFA} = \text{Bootstrapping}_n^M(T_{NFA}) = \{S_{NFA}^1, S_{NFA}^2, \dots, S_{NFA}^n\} \quad (6)$$

where S_{NFA} is a set that includes n bootstrap subsets of NFA; *Bootstrapping* is the bootstrap resampling method to extract bootstrap subsets of NFA; M is the number of instances in each bootstrap subset of NFA, which is the same with the number of FA instances; n is the number of bootstrap subsets; $S_{NFA}^1, S_{NFA}^2, \dots, S_{NFA}^n$ are the bootstrap subsets of NFA; T_{NFA} is explained in Equation (5).

$$B = \text{Combination}(S_{NFA}^i, T_{FA}) = \{B_1, B_2, \dots, B_n\}, S_{NFA}^i \in S_{NFA} \quad (7)$$

where B is a set that includes n balanced datasets and each balanced dataset contains M instances of FA and M instances of NFA; $B_1, B_2, \dots, B_n$ are the balanced datasets in B ; S_{NFA}^i indicates a bootstrap subset of NFA; *Combination* is a calculation method to combine S_{NFA}^i and T_{FA} , which integrates each bootstrap subset of NFA with the set that includes M instances of FA; T_{FA} and S_{NFA} are interpreted in Equations (4) and (6).

2.2. ARM

ARM is a popular machine learning method for discovering insightful and interesting relations between variables in large databases [35]. It aims to identify and extract strong rules from databases using different measures. Compared with other machine learning methods, ARM has the following advantages: (1) ARM is an efficient data mining method without predefined assumptions; (2) an association rule is in the form of $X \Rightarrow Y$, which can be interpreted by 'IF-THEN' statements. If X happens, then Y will happen. The rule can explicitly explain the relationship between factors and accident severity; (3) ARM can not only extract individual key factors by two-item rules but also explore the interactive relationships between multiple factors by multiple-item rules [36,37].

The basic concepts of ARM are defined as follows:

Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of n binary attributes called items, and let $D = \{t_1, t_2, \dots, t_m\}$ be a set of instances called the database. Each instance in D has a unique ID and contains a subset of the items in I . An association rule is an inference in the form of $X \Rightarrow Y$, where $X, Y \subseteq I, X \cap Y = \Phi$. In general, X is called antecedent, and Y is the consequent.

To extract strong relationships from all possible rules, constraints on significance are needed. The best-known constraints are minimum thresholds on support, confidence, and lift [22,36–38]. The definitions of these three criteria are as follows:

Support is an indication of how frequently the itemset appears in the database. For a rule $X \Rightarrow Y$, support is defined as the fraction of the number of records that contain both X and Y to the total number of records in the database. The equation is shown below.

$$\text{Support}(X \Rightarrow Y) = P(X \cap Y) = \frac{\text{Support_Count}(X \cup Y)}{|D|} \quad (8)$$

where $\text{Support}(X \Rightarrow Y)$ indicates the support level; $\text{Support_Count}(X \cup Y)$ refers to the number of records that contain both X and Y in the database; and $|D|$ refers to the total number of records.

Confidence is an indication of how often the rule has been found to be true. It is an effective parameter to represent the strength of association rules. For a rule $X \Rightarrow Y$, confidence is defined as the fraction of the number of records that contain both X and Y to the total number of records that only contain X . The equation is shown below.

$$\text{Confidence}(X \Rightarrow Y) = P(Y|X) = \frac{\text{Support_Count}(X \cup Y)}{\text{Support_Count}(X)} \quad (9)$$

where $\text{Confidence}(X \Rightarrow Y)$ indicates the confidence level; $\text{Support_Count}(X)$ refers to the number of records that only contain X ; $\text{Support_Count}(X \cup Y)$ is interpreted in Equation (8).

Lift is a measure to represent the degree to which antecedent and consequent of a rule are dependent on each other. For a rule $X \Rightarrow Y$, lift is an indication of the statistical dependence of X and Y . The equation is shown below.

$$\begin{aligned} \text{Lift}(X \Rightarrow Y) &= \frac{\text{Support}(X \Rightarrow Y)}{\text{Support}(X) * \text{Support}(Y)} \\ &= \frac{\text{Support_Count}(X \cup Y) * |D|}{\text{Support_Count}(X) * \text{Support_Count}(Y)} \end{aligned} \quad (10)$$

where $\text{Lift}(X \Rightarrow Y)$ indicates the lift level; $\text{Support_Count}(Y)$ refers to the number of records that only contain Y ; $\text{Support_Count}(X \cup Y)$, $|D|$ and $\text{Support_Count}(X)$ are interpreted in Equations (8) and (9).

As Equation (11) shows, the values of lift have different meanings and interpretations:

$$\text{Lift}(X \Rightarrow Y) \begin{cases} = 1, & \text{if } X \text{ and } Y \text{ are independent;} \\ > 1, & \text{if } X \text{ and } Y \text{ are positively correlated;} \\ < 1, & \text{if } X \text{ and } Y \text{ are negatively correlated.} \end{cases} \quad (11)$$

This study selects the Apriori algorithm to generate association rules, which is an efficient method to extract rules from a large amount of data and widely used in practical applications [38,39]. Apriori algorithm uses a 'bottom-up' structure, where frequent itemsets are extended by one item at a generation round. For example, the frequent k -length item sets are generated from the frequent $(k-1)$ -length item sets. The algorithm keeps the frequent itemsets that satisfy the thresholds of support and prunes the infrequent itemsets. It terminates when no further frequent itemsets can be created. Association rules are then extracted from the identified frequent itemsets by the thresholds of confidence and lift.

This study applies Apriori-based ARM on each balanced dataset to extract rules associated with accident severity. The consequent of each rule is the severity of ROR accidents (i.e., FA or NFA). Both two-item and three-item rules are extracted in this study. The antecedents of two-item rules indicate the individual factors associated with accident severity. The antecedents of three-item rules indicate the interactive relationship between multiple factors associated with accident severity. It is also worth noting that parameter thresholds of ARM on each balanced dataset should be very low; therefore, almost all the rules associated with accident severity can be extracted. These extracted rules can provide completed parameter distribution for ensemble rules in the next step of the ensemble method, which can improve the reliability of the results.

2.3. Ensemble Method

This study uses the BRDB method to transfer an imbalanced dataset into multiple balanced datasets. ARM is then applied to each balanced dataset to extract rules associated with accident severity. This study proposes an ensemble method to integrate the extracted rules from multiple balanced datasets, which contains two steps: (1) Ensemble rules and (2) select rules by thresholds. Figure 3 shows the process of the ensemble method.

Figure 3. The process of the ensemble method.

The first step of the ensemble method is to generate ensemble rules. Many rules can be obtained from each balanced dataset with three parameters (i.e., support, confidence, and lift). Rules with the same antecedents and consequents are integrated as one ensemble rule in this study. The ensemble method then calculates the distribution of parameters (i.e., support, confidence, and lift) for each ensemble rule with criteria of the mean ($\bar{S}$, $\bar{C}$, and $\bar{L}$) and the lower limit of $m\%$ CI ($LLCI_S^m$, $LLCI_C^m$, and $LLCI_L^m$). The equations to calculate the mean and the lower limit of $m\%$ CI are listed below.

$$\bar{S} = \frac{1}{n} \sum_{i=1}^n S_i^* \quad (12)$$

$$\bar{C} = \frac{1}{n} \sum_{i=1}^n C_i^* \quad (13)$$

$$\bar{L} = \frac{1}{n} \sum_{i=1}^n L_i^* \quad (14)$$

where $\bar{S}$, $\bar{C}$, and $\bar{L}$ are mean values of support, confidence, and lift for each ensemble rule; S_i^* , C_i^* , and L_i^* are values of support, confidence, and lift of the rule in each balanced dataset; n is the number of balanced datasets.

$$LLCI_S^m = \bar{S} - t^* \frac{\sigma_S}{\sqrt{n}} \quad (15)$$

$$LLCI_C^m = \bar{C} - t^* \frac{\sigma_C}{\sqrt{n}} \quad (16)$$

$$LLCI_L^m = \bar{L} - t^* \frac{\sigma_L}{\sqrt{n}} \quad (17)$$

where $LLCI_S^m$, $LLCI_C^m$, and $LLCI_L^m$ are the lower confidence limit of support, confidence, and lift for each ensemble rule; m is the level of confidence interval (e.g., 95%); σ_S , σ_C , and σ_L are the standard deviation of support, confidence, and lift for each ensemble rule and Equation (2) shows the detailed calculation; t^* is the upper $(1 - m)/2$ critical value with t distribution and $n - 1$ degrees of freedom; $\bar{S}$, $\bar{C}$, $\bar{L}$, and n are interpreted in Equations (12) and (14).

The second step of the proposed ensemble method is to select rules by thresholds. If one ensemble rule with $\bar{S}$, $\bar{C}$, $\bar{L}$, $LLCI_S^m$, $LLCI_C^m$, and $LLCI_L^m$ all larger than specified thresholds, then this rule would be selected; otherwise, it would be deleted. The selected rules are considered as important rules associated with accident severity. The key factors related to FA or NFA can be explored from these rules. Two kinds of criteria are adopted in this study to select ensemble rules: The mean values ($\bar{S}$, $\bar{C}$, and $\bar{L}$) and the lower limit of 95% CI ($LLCI_S^m$, $LLCI_C^m$, and $LLCI_L^m$). The mean values can reflect the central tendency of the parameters (i.e., support, confidence, and lift) of ensemble rules. CI quantifies the interval that the true value of the parameter lies. The lower limit of 95% CI can ensure that the true values of parameters (i.e., support, confidence, and lift) of ensemble rules are larger than the specified level. These criteria can ensure the ensemble rules with accepted levels of support, confidence, and lift, which can improve the reliability of the results from imbalanced datasets.

2.4. GIS Analysis

GIS is an efficient platform to store, manage, calculate, and analyze a large amount of spatial data to provide graphical outputs in visualization. Spatial analysis is an important component of GIS, which can identify hot spots by the calculation of density distribution [27,28]. Therefore, this study applies the spatial analysis of GIS to identify hot spots of ROR accidents.

Records of ROR accidents often contain longitude and latitude information, which endows the accidents with spatial properties. Therefore, spatial analysis of GIS can be applied to present the hot spots of ROR accidents related to FA factors. The hot spots are hazardous locations with a very high frequency of ROR accidents that contain factors of FA. For these hot spots, effective measures associated with FA factors should be proposed to prevent ROR accidents and improve road safety.

3. Case Study

This study collects records of ROR accidents between 2006 to 2017 in Victoria, Australia. The dataset is managed by Roads Corporation of Victoria, which is the official traffic authority in Victoria, Australia. ROR accidents are a type of road crash that occur when a vehicle crosses an edge line or leaves the designated roadway. Eight different types of ROR accidents are shown in Figure 4, which can be classified into two categories: Off-road crashes on straight and off-road crashes on curve. There are four types of accident severity in the original datasets, including fatal accident, serious injury accident,

other injury accident, and non-injury accident. Among all types of traffic accidents, ROR accidents accounted for approximately 40% of road fatalities between 2006–2017 in Victoria, Australia. It is essential to explore the key factors associated with fatal accidents; accordingly, effective measures can be proposed to prevent fatalities and improve road safety. Therefore, the accident severity in this study is classified into two groups: Fatal accident (FA) and non-fatal accident (NFA). After data preprocessing, the dataset consists of 31,940 records of ROR accidents with 1224 FA and 30,716 NFA. The dataset is extremely imbalanced with 4% of FA and 96% of NFA. The distribution of ROR accidents each year is shown in Figure 5.

Figure 4. Different types of ROR accidents.

Figure 5. The distribution of run-off-road (ROR) accidents.

To identify key factors associated with accident severity, this study integrates the variables from 10 tables in the original dataset, including Table Accident, Table Person, Table Vehicle, etc. After data cleaning and preprocessing, 22 variables are extracted for severity analysis. The variables can be classified into four types of information: (1) Road characteristics, (2) crash characteristics, (3) human characteristics, and (4) environmental information. Table 1 presents the descriptive statistics of ROR accidents.

Table 1. Descriptive statistics of ROR accidents.

No.	Variable	Categories	Number of Cases	Severity	
				FA (%)	NFA (%)
(1) Road characteristics					
1	Road Geometry	Cross intersection	1924	1.72	98.28
		T' intersection	3961	2.73	97.27
		Y' intersection	77	1.30	98.70
		Multiple intersections	350	2.57	97.43
		Not at intersection	25,542	4.20	95.80
		Others	86	1.16	98.84
2	Speed Limitation	30–50	4571	1.73	98.27
		60–75	8294	2.65	97.35
		80–90	4852	3.28	96.72
		100–110	13,100	5.79	94.21
		Unknown	1123	0.62	99.38
3	Road Surface Type	Paved	28,081	4.02	95.98
		Unpaved	3747	2.54	97.46
		Unknown	112	0.89	99.11
4	Road Type	Highways	6867	4.81	95.19
		Forest roads	46	4.35	95.65
		Tourist roads	992	4.74	95.26
		Main roads	9089	3.91	96.09
		Freeway ramps	408	1.72	98.28
		Unclassified roads	14,538	3.32	96.68
5	Road Surface Condition	Dry	23,325	4.21	95.79
		Not dry	7902	2.75	97.25
		Unknown	713	3.65	96.35
(2) Crash characteristics					
6	Time of Day	Peak time	6278	3.33	96.67
		Day time off-peak	16,179	3.28	96.72
		Night time off-peak	3619	4.28	95.72
		Late in night (23 p.m. or 0–4 a.m.)	5864	5.61	94.39
7	Day Type	Weekend	11,035	4.30	95.70
		Weekday	30,905	2.43	97.57

Table 1. Cont.

No.	Variable	Categories	Number of Cases	Severity	
				FA (%)	NFA (%)
(2) Crash characteristics					
8	Accident Type	Collision with vehicle	519	1.35	98.65
		Struck pedestrian	0	–	–
		Struck animal	7	0.00	100
		Collision with a fixed object	25,001	4.39	95.61
		Collision with some other object	174	2.87	97.13
		Vehicle overturned (no collision)	4139	2.46	97.54
		Fall from or in moving vehicle	10	10.00	90.00
		No collision and no object struck	2086	0.58	99.42
		Other accident	4	0.00	100
9	Types of ROR Accidents	1 Off carriageway to Left	2281	1.62	98.38
		2 Left off carriageway into object	10,594	3.40	96.60
		3 Off carriageway to right	1292	2.24	97.76
		4 Right off carriageway into object	7676	4.46	95.54
		5 Off carriageway right bend	1662	1.68	98.32
		6 Off right bend into object	4348	5.20	94.80
		7 Off carriageway left bend	993	2.01	97.99
		8 Off left bend into object	3094	5.88	94.12
10	Number of Vehicles involved	1	29,609	3.89	96.11
		2	1974	2.99	97.01
		>=3	357	3.36	96.64
11	Number of Persons involved	1	21,719	3.61	96.39
		2	6368	3.85	96.15
		>=3	3853	5.04	94.96
12	Motorcycle/Bicycle Involved	Yes	4373	3.52	96.48
		No	27,567	3.88	96.12
13	Trucks Involved	Yes	1301	4.30	95.70
		No	30,639	3.81	96.19
14	Pedestrian Involved	Yes	146	3.42	96.58
		No	31,794	3.83	96.17
15	Vehicle Used for Years	>=5	25,116	3.88	96.12
		<5	6824	3.66	96.34
(3) Human characteristics					
16	Driver Sex	Male	21,677	4.46	95.54
		Not male	10,263	2.51	97.49
17	Driver Age	>=65	2460	5.53	94.47
		<65	29,480	3.69	96.31
18	Helmet/Belt Worn	Yes	29,564	3.14	96.86
		No	2376	12.50	87.50

Table 1. Cont.

No.	Variable	Categories	Number of Cases	Severity	
				FA (%)	NFA (%)
(4) Environmental information					
19	Light Condition	Dark street with no lights	5230	6.37	93.63
		Dark street with lights on	5838	3.55	96.45
		Day light	20,397	3.26	96.74
		Unknown	475	4.00	96.00
20	Traffic Control	Yes	3241	2.07	97.93
		No	28,699	4.03	95.97
21	Atmospheric Condition	Clear	24,840	4.03	95.97
		Not clear	5961	3.07	96.93
		Unknown	1139	3.42	96.58
22	Urbanization Class	Large provincial cities	1157	2.25	97.75
		Melbourne urban	11,881	2.42	97.58
		Melbourne CBD	53	0.00	100
		Rural Victoria	15,910	5.15	94.85
		Small cities	1285	2.80	97.20
		Small towns	436	4.36	95.64
		Towns	967	3.62	96.38
		Unknown	251	0.00	100

According to Table 1, 22 variables related to accident severity are collected and analyzed in this study, which are as follows:

- (1) Road characteristics: Five variables are collected to describe the roads on which ROR accidents occur, including “Road Geometry,” “Speed Limitation,” “Road Surface Type,” “Road Type,” and “Road Surface Condition.”
- (2) Crash characteristics: Ten variables are collected to indicate the details of ROR accidents, including “Time of Day,” “Day Type,” “Accident Type,” “Types of ROR Accidents,” “Number of Vehicles involved,” “Number of Persons involved,” “Motorcycle/Bicycle Involved,” “Trucks Involved,” “Pedestrian Involved,” and “Vehicle Used for Years.”
- (3) Human characteristics: Three variables are collected to describe the details of drivers, including “Driver Sex,” “Driver Age,” and “Helmet/Belt Worn.”
- (4) Environmental information: Four variables are collected to indicate the environmental conditions when ROR accidents occur, including “Light Condition,” “Traffic Control,” “Atmospheric Condition,” and “Urbanization Class.”

It is worth noting that the categories of the variables in Table 1 are classified based on the original dataset or previous studies with similar objectives [40–48]. For example, the categories of some variables are the same as they were in the original dataset, such as “Road Type,” “Accident Type,” “Types of ROR Accidents,” “Driver Sex,” “Helmet/Belt Worn,” and “Urbanization Class.” However, other variables are re-coded with reduced categories to improve analysis performance according to previous studies. For example, “Day Type” contains seven categories (Monday to Sunday) in the original dataset, and it is recoded into two categories by similar properties, including “Weekend” and “Weekday” [49].

Table 1 presents the detailed categories of the features and their frequency distribution between classes of severity (i.e., FA and NFA). The frequency of FA is much less than that of NFA for all the categories of the features. Especially, the frequency of FA for “Accident Type = Struck animal,” “Accident Type = Other accident,” “Urbanization Class = Melbourne CBD,” and “Urbanization Class = Unknown” is zero, which means no fatality occurs in the relevant ROR accidents. This imbalanced frequency distribution may be caused by the extremely imbalanced dataset with 4% of FA and 96% of NFA. Therefore, this study proposes a novel method to balance class distribution and identify key factors associated with accident severity on such an imbalanced dataset. The processes and results are presented in the following sections.

4. Results

This study applied the proposed framework in the case study to explore the key factors associated with accident severity. The dataset of ROR accidents is extremely imbalanced with 4% of FA and 96% of NFA. The BRDB method is first applied to convert the imbalanced dataset into multiple balanced datasets. Each balanced dataset in this study contains 1224 FA and 1224 NFA. ARM is then applied to each balanced dataset to extract rules associated with FA or NFA. The extracted rules are integrated and selected by the proposed ensemble method. The ensemble rules are significantly associated with accident severity.

Support, confidence, and lift are user-specified parameters in ARM, which can be determined by the significance level of the results and specific research purposes [37,50]. Nine parameters associated with support, confidence, and lift are set as follows: $Support_{min} \geq 0.01$, $Confidence_{min} \geq 0.50$, $Lift_{min} \geq 1$, $\bar{S} \geq 5\%$, $\bar{C} \geq 59\%$, $\bar{L} \geq 1.15$, $LLCI_S^m \geq 4\%$, $LLCI_C^m \geq 58\%$, and $LLCI_L^m \geq 1.10$. $Support_{min}$, $Confidence_{min}$, and $Lift_{min}$ are used to extract rules associated with accident severity on each balanced dataset by ARM. $\bar{S}$, $\bar{C}$, $\bar{L}$, $LLCI_S^m$, $LLCI_C^m$, and $LLCI_L^m$ are used to select ensemble rules in the ensemble method.

4.1. Parameter Optimization: The Number of Balanced Datasets

To ensure the robustness and reliability of the results, one important consideration is to determine the number of balanced datasets in this study. This study tries different numbers of balanced datasets (i.e., 1, 2, ..., 25). It is impractical to compare all the rules from these 25 tests; therefore, this study compares the results of two-item rules (i.e., individual key factors) to determine the proper number of balanced datasets. Three criteria are adopted for determination: (1) The number of two-item rules keeps stable; (2) 95% CIs of two-item rules keep stable; (3) the more the balanced datasets, the better the results.

Figure 6 shows the number of two-item rules in different numbers of balanced datasets. The results indicate that if the balanced datasets are not enough (i.e., 1, 2, ..., 19), the number of two-item rules would greatly fluctuate. However, when the number of balanced datasets is larger than 20, the number of two-item rules keep constant and stable. In total there are 13 two-item rules related to accident severity, with six factors of FA and seven factors of NFA. The reason is that when balanced datasets are not enough, the results are significantly affected by sampling randomness. However, increasing the number of balanced datasets can reduce this kind of effect and improve the robustness of the results.

Figure 7 shows 95% CIs of two-item rules in 20–25 balanced datasets. The x-axis indicates the index of two-item rules and the y-axis indicates the 95% CI of confidence for two-item rules. The results indicate that 95% CIs of two-item rules keep stable in 20–25 balanced datasets. For each two-item rule, CI only fluctuates slightly in 20–25 balanced datasets. Besides, the differences between the upper limit and lower limit of CIs are smaller than 3% for all the two-item rules. This implies that the confidence levels of two-item rules are very concentrated and stable, and 20–25 balanced datasets are enough to provide reliable results. There is no need to test more numbers of balanced datasets (e.g., 30, 50, 100).

Figure 6. Number of two-item rules in different numbers of balanced datasets.

Figure 7. Confidence interval (CI) for two-item rules in 20–25 balanced datasets.

Figures 6 and 7 show that 20–25 balanced datasets can provide reliable results. However, increasing the number of samples in bootstrapping can reduce the risk of random sampling errors [32,33]. This means the more balanced datasets, the better the results. Therefore, this study determines 25 balanced datasets to identify key factors associated with the accident severity of ROR accidents.

4.2. Analysis of Two-Item Rules

This study extracts two-item rules from 25 balanced datasets on the case study. The results are shown in Table 2. The antecedents of two-item rules are individual key factors associated with accident severity, and the consequents are accident severity, namely FA or NFA.

Table 2. Item rules (individual key factors associated with fatal accidents (FA) and non-fatal accidents (NFA)).

No.	Antecedents	Consequents	Support		Confidence		Lift	
			Mean %	95% CI	Mean %	95% CI	Mean %	95% CI
#1	Helmet/Belt Worn = No	FA	12.13	–	78.41	(77.70,79.11)	1.57	(1.55,1.58)
#2	Light Condition = Dark street with no lights	FA	13.60	–	63.77	(63.25,64.29)	1.28	(1.27, 1.29)
#3	Types of ROR Accidents = 8	FA	7.43	–	61.81	(60.88,62.74)	1.24	(1.22,1.25)
#4	Speed Limit = 100/110	FA	31.00	–	61.05	(60.67,61.43)	1.22	(1.21,1.23)
#5	Time of Day = Late in night (23 p.m. or 0–4 a.m.)	FA	13.44	–	59.57	(58.87,60.27)	1.19	(1.18,1.21)
#6	Driver Age >=65	FA	5.56	–	59.51	(58.37,60.64)	1.19	(1.17,1.21)
#7	Speed Limit = 30/40/50	NFA	7.28	(7.08, 7.48)	69.22	(68.63,69.8)	1.38	(1.37,1.40)
#8	Traffic Control = Yes	NFA	5.26	(5.12, 5.41)	65.72	(65.11,66.32)	1.31	(1.30,1.33)
#9	Urbanization Class = Melbourne urban	NFA	19.00	(18.72,19.28)	61.73	(61.38,62.09)	1.23	(1.23,1.24)
#10	ACCIDENT TYPE = Vehicle overturned (no collision)	NFA	6.58	(6.41, 6.75)	61.16	(60.55,61.77)	1.22	(1.21,1.24)
#11	Driver Sex = Not male	NFA	16.37	(16.10,16.65)	60.81	(60.41,61.21)	1.22	(1.21,1.22)
#12	Road Surface Type = Unpaved	NFA	5.94	(5.76, 6.11)	60.39	(59.68,61.10)	1.21	(1.19,1.22)
#13	Speed Limitation = 60/75	NFA	13.36	(13.16,13.56)	59.76	(59.40,60.13)	1.20	(1.19,1.21)

The results indicate that 13 factors are related to the accident severity of ROR accidents. Six factors are associated with FA and seven factors are associated with NFA. These thirteen factors contain different types of information. For example, factors of #1, #6, and #11 belong to human characteristics. Factors of #2, #8, #9 are environment information. Factors of #3, #5, and #10 belong to crash characteristics. And factors of #4, #7, #12, and #13 are road characteristics. The support level of these factors is between 5.26–31.00%, the confidence level is between 59.51–78.41% and the lift level is between 1.19–1.57. ‘Speed Limit = 100/110’ has the highest support level of 31.00%. This indicates that ‘Speed Limit = 100/110’ has a very high frequency in fatal ROR accidents. ‘Helmet/Belt Worn = No’ has the highest confidence and lift level. This implies that ‘Helmet/Belt Worn = No’ has a very strong and positive relationship with fatal ROR accidents. Therefore, it is essential to pay more attention to these factors with high support, confidence, or lift level.

The confidence level indicates the probability of FA or NFA related to the individual key factors. According to the confidence level, the probability of the same factor to FA and NFA can be calculated, which is shown in Figure 8. The results indicate that although the confidence level of most factors is not very high, concentrated between 60–70%, the probability difference of the same factor to FA and NFA is significantly large. For example, ‘Helmet/Belt Worn = No’ is a key factor associated with FA. For all the ROR accidents including ‘Helmet/Belt Worn = No,’ the probability of FA is 78.41%, while the probability of NFA is only 21.59%. ‘Light Condition = Dark street with no lights’ is another individual key factor of FA. For all the ROR accidents including ‘Light Condition = Dark street with no lights,’ the probability of FA is 63.77%, while the probability of NFA is only 36.23%. Therefore, it is important to analyze the key factors of FA. Effective measures are accordingly proposed to reduce fatalities and improve road safety. The individual key factors will be analyzed by GIS in Section 5.1.

Figure 8. Probability of FA and NFA related to the individual key factors.

4.3. Analysis of Three-Item Rules

Individual key factors associated with accident severity can be identified from two-item rules in ARM. However, the occurrence of traffic accidents is not only related to individual key factors; they are likely to be interactive results of multiple factors. For example, driving without a seat belt and driving on roads with a 100/110 speed limit are two individual key factors associated with FA in ROR accidents. However, if these two factors act interactively, the accident is more likely to be a fatal ROR accident. In other words, individual key factors cannot provide complete explanations for all the crashes. It is essential to explore the interactive relationship between multiple factors in ROR accidents. The rules with many factors in the antecedents are too complex for analysis; therefore, this study extracts three-item rules associated with accident severity on the case study.

4.3.1. An Overall Analysis of Three-Item Rules

There are 218 three-item rules related to accident severity, with 123 rules of FA and 95 rules of NFA. Figure 9 shows the scatter plot of three-item rules, in which the x-axis indicates support level,

the y-axis indicates confidence level and the color bar indicates lift level. Each three-item rule is plotted by a scatter. The results show that the support, confidence, and lift level of FI is between 5.07–30.96%, 59.15–86.80%, and 1.18–1.74, while the support, confidence, and lift level of NFI is between 5.00–20.99%, 59.04–80.22%, and 1.18–1.60. This indicates that three-item rules of FA have a wide range of support, confidence, and lift. Especially, the rules of FA at the top-left corner have very high confidence and lift level, and rules of FA at the bottom-right corner have very high support level. Therefore, this study mainly analyzes high-confidence-lift rules and high-support rules of FA.

Figure 9. Scatter plot of three-item rules.

To identify the high-confidence-lift rules and high-support rules, Figure 10 shows the number of rules with different support and confidence level. It is worth noting that the lift level is not considered in Figure 10 because Figure 9 indicates that the higher the confidence is, the higher the lift. Therefore, rules with a high confidence level in Figure 10 also have a high lift level. The results indicate that six rules of FA with a red border are high-confidence-lift rules with confidence higher than 80% and lift higher than 1.61, while no such rules of NFA can be identified. Sixteen rules of FA with a green border are high-support rules with support higher than 25%, while no such rules of NFA are extracted.

Figure 11 shows the network plot of six high-confidence-lift rules of FA. The network plot shows how the three-item rules are constructed and how the factors interact with each other. Figure 11 indicates that most high-confidence-lift rules center on ‘Helmet/Belt Worn = No.’ If ‘Helmet/Belt

Worn = No' is interactive with other factors (i.e., 'Speed Limit = 100/110,' 'Urbanization Class = Rural Victoria,' 'Day Type = Weekend,' 'Road Geometry = Not at intersection,' 'Number of Persons involved = 1,' and 'Driver Sex = Male'), the probability of ROR accident to be FA is higher than 80%. Besides, these interactive factors have a very strong and positive relationship with fatal ROR accidents. Therefore, traffic authorities should pay enough attention to these high-confidence-lift rules.

Figure 10. The number of rules with different support and confidence level.

Figure 11. Network plot of high-confidence-lift rules of FA.

Figure 12 shows the network plot of 16 high-support rules of FA. Figure 12 indicates that most high-support rules center on 'Speed Limit = 100/110' or 'Urbanization Class = Rural Victoria.' If 'Speed Limit = 100/110' or 'Urbanization Class = Rural Victoria' is interactive with other factors (e.g., 'Accident Type = Collision with a fixed object,' 'Traffic Control = No'), the appearing frequency of these interactive factors is higher than 25%, and the probability of ROR accident to be FA is between 59–70%. Therefore, traffic authorities should pay enough attention to high-support rules with very high frequency, especially the rules including 'Speed Limit = 100/110' or 'Urbanization Class = Rural Victoria'.

Figure 12. Network plot of high-support rules of FA.

4.3.2. Comparison between Two-Item and Three-Item Rules

Two-item rules of ARM provide individual key factors related to accident severity, while three-item rules indicate the interactive relationship of three factors. Table 2 presents the two-item rules and Table 3 shows the three-item rules with the highest confidence level that contain individual key factors in two-item rules. Figure 13 shows the comparison between two-item and three-item rules of FA and Figure 14 shows the comparison between two-item and three-item rules of NFA.

The results indicate that three-item rules have higher confidence and lift levels than two-item rules, but lower support level. The higher confidence and lift levels imply that factors acting interactively increase the likelihood of FA or NFA. For example, as a two-item rule of #1 shows, if occupants do not use helmet or belt in ROR accidents, the probability of FA is 78.41%. However, the relative three-item rule of *1 shows, if occupants do not use helmet or belt and the ROR accident occurs on a road with a 100/110 limit, the probability of FA is as high as 86.80%, increased by 8.39%. As a two-item rule of #5 shows, if the ROR accident occurs in late night, the probability of FA is 59.57%. However, the relative three-item rule of *5 shows, if the ROR accident occurs in late night and in rural Victoria, the probability of FA is 69.75%, increased by 10%. Therefore, the interactive relationship between different factors should be considered because they significantly increase the risk of fatalities in ROR accidents.

Table 3. Item rules including key factors.

No.	Antecedents	Consequents	Support		Confidence		Lift	
			Mean %	95% CI	Mean %	95% CI	Mean %	95% CI
*1	Helmet/Belt Worn = No and Speed Limit = 100–110	FA	7.11	–	86.80	(85.69,87.90)	1.74	(1.71,1.76)
*2	Light Condition = Dark street with no lights and Speed Limit = 100–110	FA	10.09	–	68.09	(67.43,68.76)	1.36	(1.35,1.38)
*3	Types of Off-path Accidents = 8 and Driver Sex = Male	FA	5.96	–	64.24	(63.46,65.03)	1.28	(1.27,1.30)
*4	Speed Limit = 100–110 and Helmet/Belt Worn = No	FA	7.11	–	86.80	(85.69,87.90)	1.74	(1.71,1.76)
*5	Time of Day = Late in Night and Urbanization Class = Rural Victoria	FA	6.82	–	69.75	(68.92,70.58)	1.40	(1.38,1.41)
*6	Driver Age >=65 and Accident Type = Collision with a fixed object	FA	5.11	–	62.81	(61.86,63.76)	1.26	(1.24,1.28)
*7	Speed Limit = 30/40/50 and Helmet/Belt Worn = Yes	NFA	6.57	(6.40,6.74)	75.50	(75.01,75.99)	1.51	(1.50,1.52)
*8	Traffic Control = Yes and Pedestrian Involved = No	NFA	5.17	(5.01,5.33)	65.64	(64.95,66.33)	1.31	(1.30,1.33)
*9	Urbanization Class = Melbourne Urban and Driver Sex = Not Male	NFA	5.66	(5.50,5.81)	80.22	(79.78,80.66)	1.60	(1.59,1.61)
*10	Accident Type = Vehicle overturned (no collision) and Helmet/Belt Worn = Yes	NFA	6.17	(5.96,6.37)	72.82	(72.12,73.53)	1.46	(1.44,1.47)
*11	Driver Sex = Not Male and Urbanization Class = Melbourne Urban	NFA	5.66	(5.50,5.81)	80.22	(79.78,80.66)	1.60	(1.59,1.61)
*12	Road Surface Type = Unpaved and Helmet/Belt Worn = Yes	NFA	5.57	(5.38,5.75)	70.03	(69.33,70.74)	1.40	(1.39,1.41)
*13	Speed Limitation = 60–75 and Light Condition = Day light	NFA	7.17	(6.95,7.40)	71.11	(70.46,71.77)	1.42	(1.41,1.44)

Figure 13. Comparison between two-item and three-item rules of FA.

Figure 14. Comparison between two-item and three-item rules of NFA.

5. Discussions

5.1. GIS Analysis

5.1.1. GIS Analysis of Overall Density Distribution

GIS is an efficient platform to store, manage, calculate, and analyze a large amount of spatial data to provide graphical outputs in visualization. This study applies GIS for spatial analysis of ROR accidents. Figure 15 shows the density distribution of ROR accidents in Victoria, Australia. Comparing Figure 15A,B, “Melbourne Metropolitan Area” is the hot spot of ROR accidents. It is a hazardous region that has a higher concentration of accidents and requires higher attention. For better spatial analysis, this study mainly applies GIS on the “Melbourne Metropolitan Area” for spatial analysis in the following discussions. Figure 16 shows the density distribution of ROR accidents in the “Melbourne Metropolitan Area.”

Figure 15. Spatial analysis of ROR accidents in Victoria, Australia.

Figure 16. Spatial analysis of ROR accidents in “Melbourne Metropolitan Area.”

5.1.2. GIS Analysis of ROR Accidents Related to Individual Key Factors

Six individual key factors are identified to be associated with FA in Section 4.2. This study applies GIS for spatial analysis to investigate the hot spots of ROR accidents related to the key factors of FA in the “Melbourne Metropolitan Area.” Therefore, effective measures can be proposed to reduce ROR accidents in hot spots and improve road safety. Figure 17 shows the results.

Figure 17. Spatial analysis of ROR accidents related to individual key factors.

The density distribution of ROR accidents related to the key factors of FA is shown in Figure 17. The hot spots are hazardous locations with a very high frequency of ROR accidents. Comparing Figure 16 with Figure 17, region names of hot spots can be identified. Accordingly, effective measures can be proposed and implemented in hot spots to reduce ROR accidents and improve road safety. The six key factors of FA and relative precaution measures are discussed as follows:

- ‘Helmet/Belt Worn = No’: This factor has the highest confidence and lift level among all the factors of FA. This is because the use of helmets and seat belts can significantly reduce impact force to victims in traffic accidents. Without this protection, victims are more likely to get a fatal injury. Effective measures for hot spots in Figure 17A: (1) Traffic authority needs to enforce the use of helmet and seat belt (e.g., policy publicity, violation penalty); (2) drivers and occupants should consciously use helmets and seat belts for safety consideration.
- ‘Light Condition = Dark street with no lights’: Dark condition reduces drivers’ visibility range. Drivers in darkness are more likely to encounter unexpected situations and need more reaction time to control vehicles. These reasons increase the impact force on victims and also increase the likelihood of fatalities in ROR accidents. Effective measures for hot spots in Figure 17B: (1) Traffic authority should install more streetlights for drivers in the case of sufficient funds; (2) drivers should drive more cautiously in dark conditions.
- ‘Types of ROR Accidents = 8 (Off Left Bend into Object)’: This type of accident is left-curve driving that collides with fixed objects, which is shown in Figure 4. Curve driving reduced drivers’ visibility and maneuverability to control vehicles [51]. Besides, drivers sit on the right side of vehicles in Victoria. This obstructs drivers’ visibility when driving on a left curve. Also, vehicles colliding into fixed objects yield huge impact force to victims. These reasons increase the likelihood of fatalities in ROR accidents. Effective measures for hot spots in Figure 17C: (1) Traffic authority should install more signboards and warning lights on curved roads; (2) more rumble strips and monitoring cameras need to be installed on curved roads to control vehicle speeds; (3) drivers should drive more carefully on left-curve roads.
- ‘Speed Limit = 100–110’: The reason is that drivers drive very fast on roads with a high-speed limit. They have less time to control vehicles and therefore have a huge impact force in an emergency. This increases the likelihood of fatalities in ROR accidents. Effective measures for hot spots in Figure 17D: (1) Traffic authority should reduce the speed limit of some roads with high frequency of FA; (2) drivers should drive more cautiously on high-speed roads.
- ‘Time of Day = Late in night (23 p.m. or 0–4 a.m.)’: Drivers drive very fast due to low traffic flow late at night. Collisions with high speed impose huge impact force on victims. Besides, tiredness and poor visibility also increase the possibility of fatalities in ROR accidents. Effective measures for hot spots in Figure 17E: (1) Traffic authority should install more signboards and warning lights to remind drivers to be vigilant late at night; (2) drivers try not to drive late at night. If they must drive, they should beware of driving fatigue.
- ‘Driver Age ≥ 65 ’: The reason may be that old drivers need more reaction time in an emergency. Besides, their physical conditions increase the probability of fatalities in ROR accidents. Effective measures for hot spots in Figure 17F: (1) Traffic authority may consider setting an upper age limit for drivers; (2) old drivers should drive slowly and keep particularly cautious on roads.

5.2. The Necessity to Balance Data Distribution

The collected dataset is significantly imbalanced, with 4% of FA and 96% of NFA. To validate the necessity of balancing data distribution, this study applies ARM on the original imbalanced dataset. The results are shown in Table 4.

The results indicate that more factors of NFA can be identified than factors of FA. If support is set as 5%, no rules of FA can be identified. If support is less than 1% (i.e., 1%, 0.5%, or 0.1%), confidence is less than 10% (i.e., 10%, 5%, or 1%) and lift is set as 1, a few factors of FA can be identified with more than 40 factors of NFA. However, these rules with such low parameters are meaningless because it is hard to describe the importance of the rules and propose effective measures according to these rules. In summary, due to the imbalanced class distribution, the ARM method tends to extract enormous meaningless rules of NFA with few valuable rules of FA on imbalanced datasets. Therefore, it is necessary to apply a data-balancing method on imbalanced datasets to extract meaningful rules associated with accident severity.

Table 4. Results of association rule mining (ARM) on the imbalanced dataset.

Dataset	Support	Confidence	Lift	Factors (FA)	Factors (NFA)	Total Factors
Balanced dataset with the proposed methodology	5.00%	59%	1.15	6	7	13
Original imbalanced dataset (FA: NFA = 4%: 96%)	5.00%	59%	1.15	0	0	0
	5.00%	10%	1	0	29	29
	5.00%	5%	1	0	29	29
	5.00%	1%	1	0	29	29
	1.00%	10%	1	0	41	41
	1.00%	5%	1	4	41	45
	1.00%	1%	1	19	41	60
	0.50%	10%	1	1	43	44
	0.50%	5%	1	8	43	51
	0.50%	1%	1	23	43	66
	0.10%	10%	1	1	48	49
	0.10%	5%	1	9	48	57
	0.10%	1%	1	28	48	76

5.3. Comparison with Traditional Data-Balancing Methods

To validate the effectiveness of the proposed methodology on imbalanced datasets, this study compares the results of the proposed methodology with traditional data-balancing methods. Under-sampling, over-sampling, and mix-sampling are commonly used data-balancing methods and they are selected in this study for comparison. Ten tests are conducted to present the robustness and reliability of the results. Table 5 shows the description of datasets with different data-balancing methods. Figure 18 shows the comparison results.

Table 5. Description of datasets with different data-balancing methods.

Data-Balancing Methods	Tests	Number of Balanced Datasets in Each Test	Number of Total Accidents in Each Test	Number of FA in Each Test	Number of NFA in Each Test
Original dataset	/	/	31,940	1224	30,716
Under-sampling	10	1	2448	1224	1224
Over-sampling	10	1	61,432	30,716	30,716
Mix-sampling	10	1	24,480	12,240	12,240
The proposed methodology	10	25	31,824	1224	30,600

The results indicate that the factors related to accident severity identified by the proposed methodology are consistent and stable with six factors of FA and seven factors of NFA. All 13 factors are unchanged in different tests. However, the factors obtained from traditional data-balancing methods dramatically fluctuate in different tests. For example, the maximum and minimum numbers of total factors is 16 and 10 in under-sampling tests, while the number of FA factors ranges from four to nine in different tests. Similar results are also presented in over-sampling and mix-sampling tests. Besides, the factors are not consistent in different tests by traditional data-balancing methods. For example, both Test 4 and Test 7 in under-sampling tests can identify six factors of FA; however, the

identified factors are not the same. Therefore, compared with traditional data-balancing methods, the proposed methodology can provide more robust and reliable results on imbalanced datasets. Moreover, the proposed methodology can identify more factors of FA and fewer factors of NFA than traditional data-balancing methods. It is worth noting that the factors of FA are more important because targeted measures can be proposed to reduce fatalities and improve road safety. Therefore, the proposed methodology is better than traditional data-balancing methods.

Figure 18. Comparison results of traditional data-balancing methods and the proposed methodology.

The reasons for the superiority of the proposed methodology are summarized as follows:

- (1) Unlike traditional data-balancing methods with one randomly sampling dataset, the proposed methodology transfers an imbalanced dataset into multiple balanced datasets. If enough balanced datasets are provided, the randomness caused by sampling methods can be avoided in the proposed methodology. Therefore, the proposed methodology can improve the robustness and reliability of the results on imbalanced datasets.
- (2) Two kinds of criteria are adopted to select ensemble rules: The mean values ($\bar{S}$, $\bar{C}$, and $\bar{L}$) and the lower limit of 95% CI ($LLCI_S^m$, $LLCI_C^m$, and $LLCI_L^m$). The mean values can reflect the central tendency of the parameters (i.e., support, confidence, and lift) of ensemble rules. The lower limit

of 95% CI can ensure that the true values of parameters (i.e., support, confidence, and lift) of ensemble rules are larger than the specified level. These criteria can ensure the quality of the ensemble rules, which can improve the reliability of the results from imbalanced datasets.

6. Conclusions

This study proposes an ARM-based framework to explore key factors associated with accident severity on imbalanced datasets of ROR accidents. BRDB method is proposed to address the imbalance problem, which converts an imbalanced dataset into multiple balanced datasets. ARM is applied on each balanced dataset to identify the rules associated with accident severity. An ensemble method is proposed to integrate the rules to improve the robustness and reliability of the results. GIS is adopted for spatial analysis to present hot spots of ROR accidents related to key factors in visualization. The proposed framework is applied to a case study of ROR accidents in Victoria, Australia. The findings and contributions are summarized as follows:

- (1) Six individual key factors are identified to be closely associated with fatal ROR accidents, including 'Helmet/Belt Worn = No,' 'Light Condition = Dark street with no lights,' 'Types of Off-path Accidents = 8,' 'Speed Limit = 100/110,' 'Time of Day = Late in night,' and 'Driver Age ≥ 65 .' Hot spots of ROR accidents related to these factors are presented by GIS technology. Effective measures are accordingly proposed to reduce ROR accidents in hot spots and improve road safety.
- (2) The results indicate that three-item rules have higher confidence and lift levels than two-item rules, but lower support level. The higher confidence and lift levels imply that factors acting interactively increase the likelihood of FA or NFA.
- (3) ARM method tends to extract enormous meaningless rules of NFA with few valuable rules of FA on imbalanced datasets. Therefore, it is necessary to apply a data-balancing method on imbalanced datasets to extract meaningful rules associated with accident severity.
- (4) Compared with traditional data-balancing methods, the proposed framework has been validated to provide more robust and reliable results on imbalanced datasets. It is worth noting that the proposed framework can identify more factors of FA; therefore, more effective measures can be proposed to reduce fatalities and improve road safety.
- (5) Imbalance problems exist in various fields, for example, traffic accidents, credit scoring, machinery fault diagnosis, occupational accidents in construction industry, and diagnosis of rare diseases [24, 52–56]. The proposed framework can be applied to address the imbalance problem in various applications and improve the analysis performance of the results.

However, the limitation of this study is that this study determines the parameter thresholds of ARM by trial-and-error. This is because support, confidence, and lift are user-specified parameters in ARM, which can be determined by the significance level of the results and specific research purposes. Further study needs to propose a more objective method to determine the parameter thresholds in ARM. Besides, more factors related to accident severity will be considered in future work, such as traffic flow conditions, population density, economic conditions. However, these factors may emerge from different data sources with different data structures. Therefore, more advanced data fusing technologies are needed in the proposed framework to process and integrate the data.

Author Contributions: Conceptualization, F.J., K.K.R.Y. and E.W.M.L.; methodology, F.J. and J.M.; software, F.J.; validation, F.J.; data curation, F.J.; writing, F.J.; supervision, K.K.R.Y. and E.W.M.L.; project administration, K.K.R.Y. and E.W.M.L.; Funding Acquisition, K.K.R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11301015 and Project No. T32-101/15-R).

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Hong, J.; Tamakloe, R.; Park, D. A Comprehensive Analysis of Multi-Vehicle Crashes on Expressways: A Double Hurdle Approach. *Sustainability* **2019**, *11*, 2782. [CrossRef]
2. Casado-Sanz, N.; Guirao, B.; Attard, M. Analysis of the Risk Factors Affecting the Severity of Traffic Accidents on Spanish Crosstown Roads: The Driver's Perspective. *Sustainability* **2020**, *12*, 2237. [CrossRef]
3. Jou, R.-C.; Chen, T.-Y. External Costs to Parties Involved in Highway Traffic Accidents: The Perspective of Highway Users. *Sustainability* **2015**, *7*, 7310–7332. [CrossRef]
4. Wang, J.; Lu, H.; Sun, Z.; Wang, T.; Wang, K. Investigating the Impact of Various Risk Factors on Victims of Traffic Accidents. *Sustainability* **2020**, *12*, 3934. [CrossRef]
5. WHO. Global Status Report on Road Safety 2018 (World Health Organization (WHO)). 2018. Available online: http://www.who.int/violence_injury_prevention/road_safety_status/2018/en/ (accessed on 13 June 2020).
6. Al-Bdairi, N.S.S.; Hernandez, S. An empirical analysis of run-off-road injury severity crashes involving large trucks. *Accid. Anal. Prev.* **2017**, *102*, 93–100. [CrossRef]
7. Dirnbach, I.; Kubjatko, T.; Kolla, E.; Ondruš, J.; Šarić, Ž. Methodology Designed to Evaluate Accidents at Intersection Crossings with Respect to Forensic Purposes and Transport Sustainability. *Sustainability* **2020**, *12*, 1972. [CrossRef]
8. Griselda, L.; Juan, D.O.; Joaquín, A. Using Decision Trees to Extract Decision Rules from Police Reports on Road Accidents. *Procedia—Soc. Behav. Sci.* **2012**, *53*, 106–114. [CrossRef]
9. Eboli, L.; Forciniti, C. The Severity of Traffic Crashes in Italy: An Explorative Analysis among Different Driving Circumstances. *Sustainability* **2020**, *12*, 856. [CrossRef]
10. Gong, L.; Fan, W. (David). Modeling single-vehicle run-off-road crash severity in rural areas: Accounting for unobserved heterogeneity and age difference. *Accid. Anal. Prev.* **2017**, *101*, 124–134. [CrossRef]
11. Cheng, J.C.P.; Ma, L.J. A data-driven study of important climate factors on the achievement of LEED-EB credits. *Build. Environ.* **2015**, *90*, 232–244. [CrossRef]
12. Cheng, J.C.P.; Ma, L.J. A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects. *Build. Environ.* **2015**, *93*, 349–361. [CrossRef]
13. Ma, J.; Cheng, J.C.P. Data-driven study on the achievement of LEED credits using percentage of average score and association rule analysis. *Build. Environ.* **2016**, *98*, 121–132. [CrossRef]
14. Lee, S.; Cha, Y.; Han, S.; Hyun, C. Application of Association Rule Mining and Social Network Analysis for Understanding Causality of Construction Defects. *Sustainability* **2019**, *11*, 618. [CrossRef]
15. Arreeras, T.; Arimura, M.; Asada, T.; Arreeras, S. Association Rule Mining Tourist-Attractive Destinations for the Sustainable Development of a Large Tourism Area in Hokkaido Using Wi-Fi Tracking Data. *Sustainability* **2019**, *11*, 3967. [CrossRef]
16. Park, J.; Cha, Y.; Al Jassmi, H.; Han, S.; Hyun, C. Identification of Defect Generation Rules among Defects in Construction Projects Using Association Rule Mining. *Sustainability* **2020**, *12*, 3875. [CrossRef]
17. Ma, J.; Ding, Y.; Cheng, J.C.P.; Jiang, F.; Wan, Z. A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5. *J. Clean. Prod.* **2019**, *237*, 117729. [CrossRef]
18. Lee, H.K.; Kim, S.B. An overlap-sensitive margin classifier for imbalanced and overlapping data. *Expert Syst. Appl.* **2018**, *98*, 72–83. [CrossRef]
19. Ma, J.; Ding, Y.; Cheng, J.C.P.; Jiang, F.; Xu, Z. Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Res.* **2020**, *170*, 115350. [CrossRef]
20. Taamneh, M. Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm. *J. Saf. Res.* **2018**. [CrossRef]
21. Wang, Y.; Cao, J.; Li, W.; Gu, T.; Shi, W. Exploring traffic congestion correlation from multiple data sources. *Pervasive Mob. Comput.* **2017**, *41*, 470–483. [CrossRef]
22. Thabtah, F. A review of associative classification mining. *Knowl. Eng. Rev.* **2007**, *22*, 37–65. [CrossRef]
23. Liu, B.; Ma, Y.; Wong, C.-K. Classification Using Association Rules: Weaknesses and Enhancements. In *Data Mining for Scientific and Engineering Applications*; Massive Computing; Springer: Boston, MA, USA, 2001; pp. 591–605. ISBN 978-1-4020-0114-7.
24. Mujalli, R.O.; López, G.; Garach, L. Bayes classifiers for imbalanced traffic accidents datasets. *Accid. Anal. Prev.* **2016**, *88*, 37–51. [CrossRef] [PubMed]

25. Thammasiri, D.; Delen, D.; Meesad, P.; Kasap, N. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Syst. Appl.* **2014**, *41*, 321–330. [\[CrossRef\]](#)
26. Longadge, R.; Dongre, S.S.; Malik, L. Class Imbalance Problem in Data Mining: Review. *Int. J. Comput. Sci. Netw.* **2013**, *2*, 6.
27. Ma, J.; Ding, Y.; Cheng, J.C.P.; Tan, Y.; Gan, V.J.L.; Zhang, J. Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective. *IEEE Access* **2019**, *7*, 148059–148072. [\[CrossRef\]](#)
28. Ma, J.; Cheng, J.C.P. Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. *Appl. Energy* **2016**, *183*, 182–192. [\[CrossRef\]](#)
29. Ma, J.; Ding, Y.; Cheng, J.C.P.; Jiang, F.; Tan, Y.; Gan, V.J.L.; Wan, Z. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *J. Clean. Prod.* **2020**, *244*, 118955. [\[CrossRef\]](#)
30. Macharia, D.; Kaijage, E.; Kindberg, L.; Koech, G.; Ndungu, L.; Wahome, A.; Mugo, R. Mapping Climate Vulnerability of River Basin Communities in Tanzania to Inform Resilience Interventions. *Sustainability* **2020**, *12*, 4102. [\[CrossRef\]](#)
31. Wang, S.W.; Gebru, B.M.; Lamchin, M.; Kayastha, R.B.; Lee, W.-K. Land Use and Land Cover Change Detection and Prediction in the Kathmandu District of Nepal Using Remote Sensing and GIS. *Sustainability* **2020**, *12*, 3925. [\[CrossRef\]](#)
32. Li, K.; Wang, R.; Lei, H.; Zhang, T.; Liu, Y.; Zheng, X. Interval prediction of solar power using an Improved Bootstrap method. *Sol. Energy* **2018**, *159*, 97–112. [\[CrossRef\]](#)
33. Matsuyama, T. An application of bootstrap method for analysis of particle size distribution. *Adv. Powder Technol.* **2018**, *29*, 1404–1408. [\[CrossRef\]](#)
34. Beyaztas, U.; Bickici Arıkan, B.; Beyaztas, B.H.; Kahya, E. Construction of prediction intervals for Palmer Drought Severity Index using bootstrap. *J. Hydrol.* **2018**, *559*, 461–470. [\[CrossRef\]](#)
35. Noh, B.; Son, J.; Park, H.; Chang, S. In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining. *Sustainability* **2017**, *9*, 2119. [\[CrossRef\]](#)
36. Li, Y.; Yamamoto, T.; Zhang, G. Understanding factors associated with misclassification of fatigue-related accidents in police record. *J. Saf. Res.* **2018**, *64*, 155–162. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Montella, A. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. *Accid. Anal. Prev.* **2011**, *43*, 1451–1463. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Xu, C.; Bao, J.; Wang, C.; Liu, P. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. *J. Saf. Res.* **2018**, *67*, 65–75. [\[CrossRef\]](#)
39. Verma, A.; Khan, S.D.; Maiti, J.; Krishna, O.B. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. *Saf. Sci.* **2014**, *70*, 89–98. [\[CrossRef\]](#)
40. Pai, C.-W.; Saleh, W. Modelling motorcyclist injury severity by various crash types at T-junctions in the UK. *Saf. Sci.* **2008**, *46*, 1234–1247. [\[CrossRef\]](#)
41. Abrari Vajari, M.; Aghabayk, K.; Sadeghian, M.; Shiwakoti, N. A multinomial logit model of motorcycle crash severity at Australian intersections. *J. Saf. Res.* **2020**, *73*, 17–24. [\[CrossRef\]](#)
42. Yannis, G.; Laiou, A.; Papantoniou, P.; Christoforou, C. Impact of texting on young drivers' behavior and safety on urban and rural roads through a simulation experiment. *J. Saf. Res.* **2014**, *49*, 25.e1–31. [\[CrossRef\]](#)
43. Waseem, M.; Ahmed, A.; Saeed, T.U. Factors affecting motorcyclists' injury severities: An empirical assessment using random parameters logit model with heterogeneity in means and variances. *Accid. Anal. Prev.* **2019**, *123*, 12–19. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Kim, H.S.; Kim, H.J.; Son, B. Factors associated with automobile accidents and survival. *Accid. Anal. Prev.* **2006**, *38*, 981–987. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Morgan, A.; Mannering, F.L. The effects of road-surface conditions, age, and gender on driver-injury severities. *Accid. Anal. Prev.* **2011**, *43*, 1852–1863. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Yau, K.K.W.; Lo, H.P.; Fung, S.H.H. Multiple-vehicle traffic accidents in Hong Kong. *Accid. Anal. Prev.* **2006**, *38*, 1157–1161. [\[CrossRef\]](#)
47. Weng, J.; Zhu, J.-Z.; Yan, X.; Liu, Z. Investigation of work zone crash casualty patterns using association rules. *Accid. Anal. Prev.* **2016**, *92*, 43–52. [\[CrossRef\]](#)

48. Kumar, S.; Toshniwal, D. A data mining approach to characterize road accident locations. *J. Mod. Transp.* **2016**, *24*, 62–72. [[CrossRef](#)]
49. Lee, J.-Y.; Chung, J.-H.; Son, B. Analysis of traffic accident size for Korean highway using structural equation models. *Accid. Anal. Prev.* **2008**, *40*, 1955–1963. [[CrossRef](#)]
50. Pande, A.; Abdel-Aty, M. Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Saf. Sci.* **2009**, *47*, 145–154. [[CrossRef](#)]
51. Kim, J.-K.; Kim, S.; Ulfarsson, G.F.; Porrello, L.A. Bicyclist injury severities in bicycle–motor vehicle accidents. *Accid. Anal. Prev.* **2007**, *39*, 238–251. [[CrossRef](#)]
52. Brown, I.; Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.* **2012**, *39*, 3446–3453. [[CrossRef](#)]
53. Zhang, X.; Jiang, D.; Han, T.; Wang, N.; Yang, W.; Yang, Y. Rotating Machinery Fault Diagnosis for Imbalanced Data Based on Fast Clustering Algorithm and Support Vector Machine. *J. Sens.* **2017**, *2017*, 8092691. [[CrossRef](#)]
54. Cheng, C.-W.; Lin, C.-C.; Leu, S.-S. Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. *Saf. Sci.* **2010**, *48*, 436–444. [[CrossRef](#)]
55. Dong, Y.; Wang, X. A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets. In *Proceedings of the Knowledge Science, Engineering and Management*, Irvine, CA, USA, 12–14 December 2011; Xiong, H., Lee, W.B., Eds.; Springer: Berlin, Heidelberg, Germany, 2011; pp. 343–352.
56. Jiang, F.; Yuen, K.K.R.; Lee, E.W.M. A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. *Accid. Anal. Prev.* **2020**, *141*, 105520. [[CrossRef](#)] [[PubMed](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).