Evaluation of Injury Severity for Pedestrian-Vehicle Crashes in Jordan Using Extracted Rules

Randa Oqab Mujalli, Ph.D [#, a], Laura Garach, Ph.D [b], Griselda López, Ph.D [c], Taleb Al-Rousan, Ph.D [d]

[a] Assistant Professor at Department of Civil Engineering, The Hashemite University, 13115 Zarqa, Jordan
[b] Assistant Professor at Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada, Spain, e-mail: lgarach@ugr.es
[c] Assistant Professor at Highway Engineering Research Group, Universitat Politècnica de València, Camino de Vera, s/n; 46022-Valencia, Spain, e-mail: grilomal@tra.upv.es
[d] Associate Professor at Department of Civil Engineering, The Hashemite University, 13115 Zarqa, Jordan, e-mail: taleb@hu.edu.jo
[#] Corresponding author, The Hashemite University, 13115 Zarqa, Jordan, Phone: +96253903333-4777
e-mail: randao@hu.edu.jo

Abstract

Safety of pedestrians is a major concern all over the world since pedestrians are considered to be the most vulnerable roadway users. This paper sought to determine the main factors in pedestrian-vehicle crashes that increase the risk of a fatality or severe injury. Pedestrian-vehicle crashes which occurred in urban and suburban areas in Jordan between 2009 and 2011 were investigated. Extracted rules from Bayesian networks were used to identify factors related to severity of pedestrian-vehicle crashes. To obtain as much information as possible about these factors, three subsets were used. The first and second subsets contain all types of collisions (pedestrians and non-pedestrians), where the first subset used collision type as a class variable and the second subset used injury severity. The third subset contains pedestrian collisions only and used injury severity as the class variable. The results indicate that when using collision type as the class variable, better performance was obtained and that the following variables increase the risk of fatality or severe injury: roadway type, number of lanes, speed limit, lighting, and adverse weather condition.

Author Keywords: pedestrians; Bayesian networks; rules extraction; imbalanced dataset; collisions; urban areas.

Introduction

Pedestrians are considered vulnerable road users who are more likely to suffer from severe injuries or even get killed than other road users if involved in a crash. In fact, the largest group of road fatalities worldwide is pedestrians hit by motorized vehicles (Peden et al. 2004; Naci et al. 2009; Rosén et al. 2011). In 2015, pedestrian deaths accounted for 22% of the traffic collision deaths worldwide (WHO 2015).

Economic level of countries highly affects the places or areas where pedestrians' collisions occur (WHO 2013). In high income countries pedestrians have high probability of being involved in collisions at urban areas, while pedestrians in some low and middle-income countries are expected be involved in collisions that occur in rural areas (WHO 2013). In Jordan, an increased vehicle ownership accompanied the rapid growth in population. According to the 2014 Jordan Statistics there are 6.675 million inhabitants (excluding visting labor and refugees) with 1,329,888 licensed vehicles (DOS 2014). 82.6% of the Jordanian population lives in urban or suburban areas, with 44% of this population living in Amman, the capital city (In Jordan, areas are defined as urban if the population is more than 10,000 inhabitants and as suburban for populations between 5,000 to 10,000 inhabitants) (Ababsa, 2013). Thus, rapid growth in both population and vehicle ownership has contributed, amongst other factors, to increasing the number of collisions. Data for 2014 indicate that in Jordan 102,441 traffic crashes occurred, of which 3,839 were pedestrian collisions. Although the number of pedestrian collisions does not seem to be high relative to the total number of crashes which accounts for 3.7% of total traffic crashes, however, the number of fatalities which resulted from pedestrian-vehicle collisions was 255 fatalities, which makes about (37%) of the total fatalities number which was (688) from traffic collisions (PTD 2014) as shown in Figure 1.

[Insert Figure 1]

Jordanian Traffic law imposes penalties on pedestrians crossing roads from illegal points i.e. not using defined crosswalks, pedestrian underpasses or overpasses. However, this penalty is not applied in practice because crosswalks sometimes do not provide the best crossing (MOI 2008). Often pedestrians are forced to cross at illegal locations because of bad driver behavior, for example stopping on crosswalks. Moreover, most crosswalks throughout Jordan are not clearly marked and pedestrians are forced to walk on the roads because of missing or inadequate sidewalks or because the sidewalks are obstructed by trees or merchants (JTI 2012). When pedestrian overpasses are available, about 60% of pedestrians do not use them because of high stairs, health related issues, extra walking distance, inconvenience and safety fears (Abojaradeh 2013).

Studying risk factors for pedestrian-vehicle collisions will help develop preventive measures and hence reducing pedestrian-vehicle collisions rate (Zhang et al. 2014). Researchers have explored collisions involving pedestrians using several statistical methods, such as logit or probit models (Zajac and Ivan 2003; Lee and Abdel-Aty 2005; Xie et al. 2009; Tay et al. 2011; Islam and Hernandez 2013; Mohamed et al. 2013; Obeng and Rokonuzzaman 2013;Olszewski et al. 2015). These models rely on a pre-assumption of the data distribution and the relationship between explanatory variables and response variable (Chang and Wang 2006).

Some of the very interesting recent methods used in crashes analysis are data mining (DM), the methods used in DM do not rely on pre-assumptions of the data distribution nor the relationship between variables, and they help in discovering some new relationships (Prato et al. 2010). In addition, patterns of collisions can be discovered using DM techniques. One of the aims of Data Mining is identifying valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad et al. 1996). Different data mining methods have been used to analyze road crashes, such as association rules (Pande and Abdel-Aty 2009; Montella et al. 2012), decision trees (Prato et al. 2010; Abellán et al. 2013; De Oña et al. 2013; Jung et al. 2016), clustering analysis (Mohamed et al. 2013; Thompson et al. 2013), neural networks (Abdel-Aty and Abdelwahab, 2004), Bayesian Networks (BNs) (De Oña et al. 2011; Mujalli and De Oña 2011; Kwon et al. 2015; Mujalli et al. 2016). One of the main advantages of BNs is that inference can be used to find out key variables which significantly affect the dependent variable (class variable) in the resulting models (De Oña el al. 2012, Mujalli et al. 2016).

Traffic collision datasets are usually imbalanced, which means that the dataset has one or some of the classes (categories) with much more samples in comparison to the others. The most prevalent class is called the majority class, while the rarest class is called minority class (Yijing et al. 2016). The minority class usually represents a concept with greater interest than the majority class. However, it is often outnumbered by the majority class, and sometimes this scenario may be very severe (Sun et al. 2015). One of the most popular methods for solving the class imbalance problems is sampling. The two most used sampling techniques are undersampling and oversampling. In undersampling, instances of the minority and majority classes are selected randomly in order to achieve a balanced stratified sample with equal class distributions, often using all instances of the minority class and only a subset of the majority class, or undersampling both classes for even smaller subsets with equal class sizes. In oversampling, the class distribution is approximately equalized between minority and majority classes, where minority cases are replicated (Crone and Finlay 2012). Accordingly, Mujalli et al. (2016) found that when using balanced datasets by oversampling, better classification of collisions is achieved and less misclassification of the minority class occurs.

The main aim of this research is the exploratory analysis of pedestrian-vehicle collisions in Jordan using DM techniques. DM allows knowledge extraction of previously hidden patterns (Fayyad et al. 1996). According to Zhang et al (2004), using a single database limits the possibility of discovering specific relationships due to the fact that knowledge might be hidden in multiple databases. In order to explore all possible patterns associated with pedestrian-vehicle collisions and their severity, 3 subsets are developed and explored.Two subsets contain all types of collisions, with one of them using collision type as a class variable and the other using injury severity as the class variable, whereas the third subset is pedestrian-specific which only contains pedestrian collisions. Both subsets were used in order to find out if different patterns of variables affecting severity could be extracted when using a general subset or a pedestrian specific subset. To take into account the problem of imbalanced data, sampling techniques will be applied prior to developing BN models and analysis.

The remainder of the paper is organized as follows: First, the method and data used are presented, followed by results and their discussion. Finally, the main conclusions of this study and recommendations are stated.

Material and Methods

The procedure used in this study has been the following:

1. Dataset for collisions which occurred in urban and suburban areas in Jordan was obtained from Jordanian Police Traffic Department (PTD), this dataset included 14 variables.
2. All types of collisions were extracted from the dataset excluding property damage only collisions (PDO); this subset is called hereafter general subset (see Table 1).

[Insert Table 1]

3. Subset Col-ACT and subset Col-SEV contain the same number of records as general subset; the only difference between these subsets and the general subset is that the collision type variable (ACT) was categorized into two categories (values): Pedestrian (PED) and others (OT) (see Table 1). Also, in subset Col-ACT, the class variable was ACT and in subset Col-SEV the class variable was SEV. More specifically, these subsets contain all records (pedestrian-vehicle collisions and non- pedestrian collisions.
4. Subset Ped-SEV is obtained from subset Col-ACT. In subset PED-SEV, only records belonging to (ACT=PED) were used and thus all records belonging to other collisions types (OT) were eliminated. As a result, variable ACT was eliminated, and only 13 variables were used (see Table 1). More specifically, this subset contains only pedestrian-vehicle collisions records.
5. All subsets are now characterized as follows:

    a. Subsets Col-ACT and Col-SEV, in which all types of collisions were included: collisions with vehicles and collisions with pedestrians.
    b. Subset PED-SEV, in which only pedestrian-vehicle collisions were included.

6. In both subsets Col-ACT and PED-SEV, collisions were classified according to severity (SEV), whereas in subset Col-ACT, collisions were classified according to collision type (ACT).
7. All subsets were balanced using oversampling balancing technique.
8. All balanced subsets were used to develop BNs using different search algorithms (K2 and simulated Annealing) and scores (BDeu, MDL, AIC) (De Oña et al. 2011 and Mujalli et al. 2016).
9. Six Developed BNs were then compared (using 10-folds cross validation method) based on their performance as obtained using performance measures.
10. The BN with the best performance using each subset was further analyzed using the conditional probability tables to extract significant rules.
11. The extracted rules that obtained the highest probability of fatality or severe injury (FS) in each BN from each subset were presented and analyzed.

The procedure is shown in Figure 2.

[Insert Figure 2]

Data

Collision data were obtained from PTD with records for traffic collisions that occurred on urban and suburban roads in Jordan from year 2009 to 2011.Unrealistic data was removed from the dataset, as well as, PDO collisions. The final dataset contained 21,852 records (general subset). Table 1 includes detailed explanation of the subsets used.

Fourteen (14) variables were used to find out which factors significantly affect the occurrence of a severe or fatal outcome in pedestrian-vehicle collisions. These variables were related to road characteristics, prevailing conditions, prevailing weather conditions, collision type and the resulting injury severity of the worst injured (Chang and Wang 2006; Abellán et al. 2013, De Oña et al. 2013; López et al. 2014).

Table 2 includes the classification of records between the classes of classification, where in subset ACT and subset 1: there are 17,769 records with slight injuries and 4,083records with fatal and severe injuries. In subset Col-ACT the class variable ACT is imbalanced with (94.4%) cases classified as OT and 5.6% cases classified as PED. In subsets Col-SEV the class variable SEV is imbalanced, with 81.32% cases classified as SI, and 18.68%

cases classified as FS. Also, in subset 2 the class variable SEV was imbalanced with 82.9% cases classified as slight injuries and 17.1% cases classified as fatality or severe injuries.

[Insert Table 2]

Balancing subsets

Datasets in which records belonging to categories of the class variable do not have almost the same distribution between the categories of the class variable are called imbalanced datasets (Crone and Finlay 2012).The problem of imbalanced datasets is more apparent in cases where the main interest of the classification task is to distinguish cases belonging to the class having rare occurrence in nature, such as collisions resulting in severe or fatal outcomes. The possible problem is misclassifying a collision as minor when in reality it should have been classified as severe injury or fatal. To deal with the problem of imbalanced datasets, sampling prior to classification is most commonly used, in which class distribution of records is altered so that the minority class (rare cases) records in the dataset are increased in size (Thammasiri et al. 2014).

In this research work, and based on the results obtained by Mujalli et al. (2016), oversampling balancing technique in Weka software was used (Witten and Frank 2005).

Oversampling

The oversampling technique used herein is Synthetic Minority Oversampling Technique (SMOTE), in which a subset of the original dataset is created by creating synthetic minority examples; each minority class sample is taken and synthetic examples along the line of the segments joining any/all of the (k) minority class nearest neighbors are introduced. Neighbors from the (k) nearest neighbors are randomly chosen and one sample is generated in the direction of each, based on the amount of oversampling prescribed (Chawla et al. 2002). For further details on sampling techniques used in the analysis of traffic crashes, interested readers are referred to Mujalli et al. (2016)

Bayesian networks

A BN is a Directed Acyclic Graph (DAG) over a set of variables: $U=\{x_1, x_2, \ldots, x_n\}, n\geq 1$ and a set of probability tables $B_p= p(x_i|pa\ (x_i)),x_i \in U$ where $pa\ (x_i)$ is the set of parents of x_i in BN and $i= (1,2,3,\ldots.,n)$. A BN represents joint probability distributions $P(U)=\prod_{x_i \in U} p\left(x_i \middle| pa(x_i)\right)$.The arcs are interpreted as direct dependence relationships between the linked variables, however indirect dependence relationships between variables could exist (Acid et al., 2004).

The classification task consists in classifying a variable $y=x_0$ called the class variable, given a set of variables $U=x_1 . . . x_n$, called attribute variables. A classifier $h : U \to y$ is a function that classify an instance of U to a value of y. A dataset D consisting of samples over $(U, y\ )$is used to learn a classifier where the learning task consists of finding an appropriate BN given a data set D over U.

In this study, BNs are used in order to develop different models and to compare their results in terms of their ability to correctly classify collisions according to either of their injury severity or collision type and also to extract significant rules. Two search methods were used: hill climber algorithm restricted by an order on the variables (K2) and simulated annealing search algorithm. Also, three different score functions were used: BDe score metric (BDeu); Minimum Description Length (MDL); and the Akaike Information Criterion (AIC). The search algorithms and the scores were applied in this study because they are widely used, relatively quick, and they produce good results in terms of network overall performance (Madden 2009; Mujalli et al. 2016).

Performance evaluation measures

To compare the different BNs, the following performance measures were used: accuracy, sensitivity, specificity, and area under receiver operating characteristic curve (AUC). Their equations and explanation of AUC are given below:

$$Accuracy = \frac{TSI + TFS}{TSI + FSI + TFS + FFS} \qquad (1)$$

$$Sensitivity = \frac{TSI}{TSI + FFS} \qquad (2)$$

$$Specificity = \frac{TFS}{TFS + FSI} \qquad (3)$$

Where,

TSI: number of (SI) instances correctly classified,

TFS: number of (FS) instances correctly classified,

FSI: number of (SI) instances incorrectly classified, and

FFS: number of (FS) instances incorrectly classified.

AUC is the area below a curve where sensitivity (plotted on y-axis) and 1-specificity (Bradley 1997).

Results

The aim herein is to investigate the main factors that affect the occurrence of a specific outcome in pedestrian-vehicle collisions. To achieve this aim three subsets have been analyzed:

- Subset Col-ACT: contains all records (pedestrian-vehicle collisions and non- pedestrian collisions.

- Subset Col-SEV: contains all records (pedestrian-vehicle collisions and non- pedestrian collisions).
- Subset Ped-SEV: contains only pedestrian-vehicle collisions records.

In subset Col-ACT the class variable was the collision type (ACT) and the collisions are classified according to ACT: PED, OT; whereas in subsets Col-SEV and Ped-SEV, the class variable was the severity and thus collisions were classified according to SEV: FS or SI.

The following sections include a presentation of the subsets used in analysis and the BNs developed using these subsets. Next, a detailed analysis of the developed BNs along with the findings of this research work is shown.

Subsets used

As previously indicated subsets were imbalanced as shown in Table 2. To solve this issue, oversampling was used to develop balanced subsets. Details of the balanced subsets are shown in Table 2.

It is illustrated in Table 2 that the number of the instances in the resulting subsets was increased to the size of the majority class (20,636 instances for PED in subset Col-ACT; 17,761 instances for FS in subset Col-SEV and 1,008 for FS in subset Ped-SEV).

Bayesian networks and subsets

For each balanced subset, six models using six different algorithms (as shown in Figure 1) were developed. Subsets were first divided into ten subsamples, training subsamples and testing subsamples. In each run, nine subsamples were used to train the model and the last was used to test the model. The process was repeated for ten runs for each subsample (ten-fold cross-validation).

To compare the resulting models, performance evaluation measures were used. The results shown in Table 3 are the averages of the ten runs of the test subsamples. It should be mentioned that the statistical significance of each model obtained was tested using a corrected paired t-test.

[Insert Table 3]

Results extracted from the developed models using the balanced subsets indicated that accuracy results for both subset Col-SEV and subset Ped-SEV were in range of 59% to 61% and 51% to 61%, respectively, where accuracy for subset Col-ACT was 85.7% to 86.8% indicating a better overall performance obtained by subset ACT.

Regarding sensitivity, the values ranges in subset Col-SEV were between 58% and 61.6%; whereas in subset Ped-SEV, the ranges increased from 1.9% to 58.7% indicating that subset COl-SEV overpowered subset Ped-SEV in classifying instances belonging to SI class. In subset Col-ACT sensitivity results ranged from 75.9% to 76.4%.

In terms of specificity, in subset Col-SEV the values ranged from 56.3% to 64.2%, where in subset Ped-SEV the values ranged from 64.2% to 99.7%. It is evident here that all classifiers in both subsets were remarkably capable of classifying the minority class (FS) correctly. In subset Col-ACT, the values ranged from 95.5% to 97.3% indicating that all algorithms used were capable of classifying PED correctly.

Finally, with respect to AUC, results for subset Col-SEV ranged from 63% to 65.6%, for subset Ped-SEV from 50.8% to 64.7% and for subset Col-ACT the results ranged from 90.4% to 91.6%.

Based on the aforementioned results, subset Col-ACT had more wins (21 wins) than subset Col-SEV with no wins and subset Ped-SEV with 2 wins only. Regarding the best algorithms used, results in Table 4 shows the averages for the test subsamples using the ten-fold cross-validation and testing the statistical significance of each algorithm using a corrected paired t-test. It is shown than the best algorithm used K2+AIC with 11 wins.

[Insert Table 4]

As shown in Table 4, algorithm K2+AIC was the best in the balanced subsets. In subsets Col-ACT and Col-SEV it obtained the highest values in accuracy, specificity and AUC; whereas in subset Ped-SEV it achieved the highest values in accuracy, sensitivity and AUC. To this end, BNs obtained using K2+AIC will be used to interpret the contributing factors into pedestrian-vehicle collisions.

Interpretation of Bayesian networks

BNs developed using K2+AIC algorithm were used for further analysis and to determine which variables are responsible for FS injuries in pedestrian-vehicle collisions as shown in Table 5.

[Insert Table 5]

As illustrated in Table 5, the developed BN using Col-ACT subset had the following variables directly connected to the variable severity (SEV) which is the variable of interest: collision type (ACT), road type (DIR), number of lanes (LANE), traffic control (CONT), and speed (SPE). The severity presents a direct relationship with both pedestrian and vehicle crashes, with road type, number of lanes and traffic control and speed limits in all types of crashes included in subset ACT. When using subset Col-SEV it was found that all the independent variables were connected to severity (SEV) except for traffic control (CONT), whereas in subset Ped-SEV (only pedestrian collisions) seven variables were connected to severity (SEV): vehicles involved (VEH), collision pattern (PAT), road type (DIR), number of lanes (LANE), road grade (GRADE), lighting condition (LIG) and speed limit (SPE). On the other hand, the following variables were found to be common in all BNs using all subsets and with arcs connecting them to severity (SEV): number of lanes (LANE), road type (DIR) and speed limit (SPE); indicating the significance of these variables in pedestrian-vehicle collisions.

In view of the above results, it is of interest for safety researchers to find out what differences exist between collisions if pedestrian only subset was used and those if all types of collisions subsets were used, as well as, what differences exist if the root node (class variable) was altered. In light of this, the developed BNs need to be further analyzed using more deep knowledge extraction analysis. The following section describes the main differences and similarities between the developed BNs using the different subsets.

Rules extraction using BNs

In order to identify the most significant values (categories) of variables that affect the occurrence of fatality or severe injury in a pedestrian-vehicle collision, the most significant rules were extracted from the conditional probability tables of the developed BNs using each subset. The extracted rules allow finding the particular variables and their respective values which are associated with the highest probability (confidence) of occurrence for fatality or severe pedestrian-vehicle collision. The results for the extracted rules are shown in Table 6.

[Insert Table 6]

Based on results in Table 6, it would be possible to identify the factors that affect the occurrence of fatal or severe pedestrian-vehicle collisions. In which the combinations of factors that define a pedestrian collision to be classified as FS were determined. To this end the rules were divided into three groups:

1. High confidence rules group: the rules which achieved a confidence of at least 90% are listed.

2. Moderate confidence rules group: the rules which achieved a confidence between 70-89% are listed.

3. Low confidence rules group: the rules which achieved a confidence between 50-69% are listed.

It should be noted that the categorization into these three categories was based on the number of rules and the confidence level obtained by these rules where it was found in literature that researchers used similar thresholds to extract knowledge from used datasets. For example both Abellán et al. (2013) and López et al. (2014) used a confidence threshold of 60% in order to extract significant rules when using decision trees.

In the high confidence group, the results indicated that there were only two rules belonging to this group: the first most important rule having a confidence of 92.1% was obtained when using subset Col-ACT. This rule has the parent node severity (SEV) with its state (FS), the rule strongly suggests that the occurrence of FS pedestrian-vehicle collision is almost certain if the following conditions exist: the collision occurring on roads with speed limit of 30 km/hr on a 4 lanes divided roadway. This points out that as the width of the crossed roads increases the risk of fatal or severe pedestrian collisions increase even if the posted speed was as low as 30 km/hr.

The second rule which had a confidence of 92.1% of fatal or severe collision occurrence was obtained when using subset Ped-SEV. The extracted rule has the parent node number of lanes (LANE) with its state (1) where it involved the existence of the following factors at time of collision: the collision occurs on 2 lanes one way roads, in which the rule basically states that the risk of crossing wider roads is associated with almost a certain fatal or severe pedestrian-vehicle collision occurrence.

In the moderate confidence rules group, there were six extracted rules belonging to all subsets, the first rule with the highest confidence in this group was extracted from subset Col-SEV. This rule has the parent node lighting condition (LIG) with its state (DARK) and it indicates that the occurrence of fatal or severe pedestrian collisions is associated with the existence of the following conditions at time of collision: if the collision occurs at night without lighting on 2 ways undivided roadway. This rule does not clearly identify the prevailing weather condition, even though the atmospheric weather exists as a contributory factor but it says (other). If we refer to the original data, (other) refers to one of the following states: snow, storm, wind, fog, or dust which means that the prevailing conditions were not clear. This rule ascertains the importance of the type of roadway and adds two other factors: dark and unclear weather.

The second highest confidence rule was extracted from subset Ped-SEV with a 76.1% confidence. This rule has the parent node number of lanes (LANE) with its state (2) and it states that if the collision occurs on 2 lanes undivided roads, the probability of a pedestrian being involved in severe or fatal collision is moderate; highlighting the significance of the type of roadway in pedestrian collisions.

The third highest confidence rule was extracted from subset Col-SEV with a 74% confidence, where a new factor was added to the contributory factors in pedestrian-vehicle collisions which is the number of involved vehicles. The rule has the parent node speed (SPE) with its state (40) and it states that if collision occurs on one way road with a speed limit of 40 km/hr and there was only one vehicle involved in this pedestrian-vehicle collision, then there is a moderate probability that the collision will be classified as fatal or severe collision. The last rule in this group was extracted from subset Ped-SEV with a confidence of 73.9%, in which the rule has the parent node number of lanes (LANE) with its state (2) and it states that the following combination of factors have a moderate confidence that the collision would be classified as fatal or severe injury: if the collision occurs on a four lanes-tow-2 way divided road.

In the low confidence group, there were 5 extracted rules belonging to this group. The first highest confidence rule was extracted from subset Ped-SEV, in which the rule has the parent node speed (SPE) with its state (40) and it states that if collision occurs on two-ways undivided road with a speed limit of 40 km/hr, then the probability is relatively low that the collision will be classified as fatal or severe injury. It should be mentioned that the absence of the number of vehicles factor (as shown from rule no. 2 from subset Col-SEV) and the change of the type of roadway from one way (as shown in rule no. 2 from subset Col-SEV) to two-ways has decreased the confidence from 74% to 66% indicating the significance of the factors which were present in rule Ped-SEV from subset Col-SEV in fatal or severe injury collisions.

The second two highest confidence rules were extracted from subsets Col-ACT and Col-SEV. The extracted rule from subset Col-ACT has the parent node severity (SEV) and it state (FS) and it states that if the collision occurs on four lanes -two- ways undivided road with a speed limit of 90 km/hr, then there is a low probability that the collision will be classified as fatal or severe injury. This result might indicate the degree of alert that the pedestrian pay when crossing roads with high speed limits, in which the probability of the pedestrians being involved in such collisions becomes relatively low. The extracted rule from subset Col-SEV has the parent node road type (DIR) and its state (2_DIV) and it states that: if the collision occurs on two ways divided road and two vehicles were involved in the collision, then the probability of having fatal or severe injury is relatively low.

The third highest rule in this group was extracted from subset Ped-SEV, in which the rule has the parent node speed (SPE) and its state (40) and it states that: if the collision occurs on a one way road having a speed limit of 40 km/hr then the probability is relatively low that the collision will be classified as fatal or severe injury. It should be mentioned that the change of road type from two ways undivided (as shown in rule no. 4 from subset Ped-SEV) into one way in the current rule has decreased the probability of fatal or severe injury collisions having the speed limit the same in both rules.

The last extracted rule in this group was obtained from subset Ped-SEV, this rule has the parent node lighting conditions (LIG) and its state (DARK) and it states that: if the collision occurs at night under dark lighting conditions when the prevailing weather condition is other (snow, storm, wind, fog, or dust) then the probability is relatively low that the collision will be classified as fatal or severe injury. It should be mentioned that the absence of the road type factor (as shown in rule 1 in subset Col-ACT) has decreased the obtained probability in the current rule, indicating the significance of road type in such collisions once more.

Absence of control type from all significant extracted rules raised some concerns since it is believed to have some effect on severity outcome, consequently inference was performed in order to caputure the effect of control type (if any) on fatality or severe injury outcome. For more details on BN inference, readers are referred to (De Oña et al., 2011).

The evidence was set to certainty for the following variables/subset combination as follows: (Subsets Col-ACT and Col-SEV: Probability (ACT=PED, SEV=FS); Subset PED-SEV: Probability (SEV=FS)). The probabilities of control type (CONT) were then calculated and the results indicated that in subset Col-ACT the probability of CONT=NO_CONT was found to be 0.8950, in subset Col-SEV the probability of CONT=NO-CONT was found to be 0.8758, and in subset PED_SEV the probability of CONT=NO-CONT was found to be 0.8765. The inference results indicat that fatal or severe pedestrian-vehicle crashes are significantly associated with locations where no control exists; this result should be interpreted with caution because this category of the variable CONT includes both uncontrolled intersections and non-intersections. In a study prepared by Chong (2018) risk of fatality was found to increase at non-intersections as compared to intersections. On the other hand, intersections without signals were found to have increased risk of fatal or injury pedestrian crashes (Moudon et al., 2011).

Discussion

The research work presented in this paper analyzed the severity of pedestrian-vehicle collisions in urban and sub-urban areas in Jordan using BNs. In this study, in order to explore all the main contributory factors affecting these collisions, three subsets were used. Two subsets included all types of collisions, pedestrian and no pedestrian collisions (subsets Col-ACT and Col-SEV), whereas the third subset included only pedestrian collisions (subset Ped-SEV). It was found that using collision type as the class variable in subset Col-ACT enhanced the performance of the developed BNs in terms of better classification of cases as compared to the other two subsets. On the other hand, it was found that using subset Ped-SEV which included only pedestrian collisions enhanced the performance with respect to subset Col-SEV which included all collision types. The enhanced performance of subset Col-ACT increases its reliability in terms of defining the factors that are highly associated with pedestrian involved collisions.

Table 7 shows a summary of the number of times each factor was present in the BNs resulting from using each subset. Road type was present 10 times in the resulting rules indicating the importance of this factor. Speed and number of lanes were present 5 times each in the resulting extracted rules. On the other hand, number of vehicles, lighting and weather were the least present.

[Insert Table 7]

Regarding to which states of factors caused FS pedestrian-vehicle collisions, higher risks of FS pedestrian collision were associated with collisions involving single vehicle as shown in rule number 2 from subset Col-SEV. This result contradicts with the results obtained by Verzosa and Miles (2016) where they found that the odds of fatal collisions involving pedestrians are three times higher when there are heavy vehicles or multiple vehicles involved than when there are light vehicles. The results herein indicated that single vehicle crashes rather than multiple vehicles, as found by Verzosa and Miles (2016), have high risk of resulting in fatality when the speed limit is 40km/hr and the road is one way. However, both studies are not completely comparable since in this paper the type of vehicle involved was missing and thus the results obtained for single vehicle crashes might be for single heavy vehicles crashes. In absence of any reserach work on injury severity of vehicle-pedestrian crashes that included number of vehicles as a predictor without reference to type of vehicle, authors of this work recommend that the result obtained herein should be interpreted with caution.

Type of roadway; whether it was divided or one-way were found to have the highest risk in FS collisions as seen in rule no. 1 in subset Col-ACT and rule no. 1 in subset Ped-SEV. Olszewski et al. (2015) found that divided roadways were associated with increased risk of pedestrian death at un-signalized crosswalks which is consistent with the results herein. It should be mentioned that divided roadways are riskier when the speed limit is 30 km/h and the number of lanes per direction is two lanes as shown in rule no. 1 in subset Col-ACT, this might give an indication of the impaired decision of pedestrians when crossing wide divided roadways with low speed limit and their false perception of safety. On the other hand, rule no. 2 in subset Col-ACT has the same combination of factors as that in rule no. 1, but speed limit was 90 km/hr which decreased the risk of FS collisions, which might also indicate the limited accessibility the pedestrians have on higher speed roads, where the accessibility is provided through pedestrian facilities; and even if there were some violations made by pedestrians, they will be probably more careful when crossing roadways with higher speed limits. Another result is that single lane one-way roads have the same risk as four lanes divided roadways as shown in rule no. 1 in subset Ped-SEV which may indicate once again the false perception of safety that encourage pedestrians to cross roads without taking necessary precautions.

Number of lanes, whether one or two lanes, was found to be associated with FS collisions in both rules in subset Col-ACT and in rules 1 to 3 in subset Ped-SEV. Sze and Wong (2007) found that more lanes are significantly associated with the higher likelihood of a pedestrian fatality. Also, Sherony and Zhang (2015) found that in light vehicle collisions which occurred in the United Sates, more pedestrian crashes occurred on principle arterial than any other type of roads. In addition, Amoh-Gyimah et al. (2016) found that wider roads are significantly associated with fatalities in pedestrian-vehicle crashes. The interpretation of such results would be that as number of lanes in roadways increases the ability of the driver to violate speed limits increases and hence increases her/his odds of being involved in collisions.

As for speed, it was found that when the posted speed is 30 km/hr, the probability of FS collisions will be as high as 92.1% with less probability of FS collisions with higher speeds (rule no. 2 from subset Col-ACT, rule no. 2 from subset Col-SEV, rules no. 3 and 4 from subset Ped-SEV). According to Johansson et al. (2004), the 90th-percentile speed should not exceed 30 km/h for higher safety. Other researchers found that the risk of a pedestrian dying of a collision with a vehicle is estimated to jump from 5% to 45% when the speed of the striking vehicle increases from 20 (32 km/hr) to 30 mph (48 km/hr); it is 85% when vehicular speed reaches 40 mph (64 km/hr) (Limpert 1984; Leaf and Preusser 1999; Evaluating Safety and Health Impacts 2014). In addition, results reported by Al-Omari and Obaidat (2013) supported the results found herein; they indicated that 65% of pedestrian collisions which occurred at Irbid city in Jordan were at speed limit of 40 km/hr. Moreover, Knowles et al. (2012) found that around 90% of fatal pedestrian collisions on London’s roads occurred on roads with speed limits of 48 km/hour or lower.

Higher levels of severity of pedestrian collisions were also found to be highly associated with night and dark conditions as shown in rule 6 from subset 2. Sherony and Zhang (2015) found that in the United States for light vehicles collisions involving pedestrians, most of the injury crashes occurred under daylight condition (62%) while most of the fatal crashes occurred in dark condition (36% for dark but lighted, 34% for dark). In addition, many other researchers reported that the majority of pedestrian collisions and majority of severe injuries have occurred during nighttime (Campbell et al. 2004; Kim et al. 2008 and 2010; Verzosa and Miles 2016 and Al-Ghamedi 2002, Amoh-Gyimah et al. 2016).

According to the results obtained herein, a snowy, stormy, windy, fogy, or dusty day was found to be mostly associated with FS. DiMaggio and Durkin (2002) found that weather conditions have not been directly related to the risk of injury. Kim et al. (2010) found that inclement weather (rain, snow, fog, and smog) decreased the probability of fatal injury (−37%), but increased the probability of incapacitating injury (5%), non-incapacitating injury (5%), and possible or no injury (5%). Moreover, Sherony and Zhang (2015) indicated that the majority of collisions in the United States occurred under clear / cloudy condition (no adverse conditions), in which 84.6% pedestrian injury crashes, 85.4% pedestrian fatal crashes occurred during no adverse weather condition. However, Amoh-Gyimah et al. (2016) found that there is a significantly higher probability of pedestrian-vehicle crashes to occur during unclear weather conditions which is consistent with the result found herein.

Limitations of the study

First of all, this study was aimed at finding out variables that increase the probability of having a fatality or severe injury resulting from a vehicle-pedestrian crash, and should not be confused with other studies that aim at finding predictor variables that are associated with frequency of pedestrian crashes. That is said, it should be mentioned that the crash data collection is based on the standard Jordanian Police Traffic Department reporting system, where crashes were not geocoded by using milepost data. Consequently, some variables that might have significant effect on vehicle-pedestrian crashes were either unavailable or could not be linked to the data obtained and hence were not included herein.

This study did not include studying the effect of Average Annual Daily Traffic (AADT) on injury severity of vehicle-pedestrian crashes since it was not available at any transportation agency. Studies found in literature used AADT as a predictor variable in frequency of vehicle pedestrian crashes, and it was found to be a significant predictor in many of these studies (Brugge et al., 2002; LaScala et al., 2000; Lee and Abdel-Aty, 2005; Loukaitou-Sideris et al., 2007). However, a study on injury severity of vehicle-pedestrian crashes found that AADT was a relatively a weak predictor; they referred the reason of this result to the fact that state routes with high AADT have lower speeds (Moudona et al. 2011). The results of this study did not differ than the results obtained by Ewing (2006) and NHTSA (1999) in which they found that injury severity is strongly associated with vehicle speed.

Population density which is usually used to define area type (urban or rural) was also found by many presvious research works to have a significant effect on pedestrians' crashes. However, its effect on higher injury severity levels is reduced as population density increases (Goel et al. 2018). The reason behind this result is referred to the fact that high numbers of pedestrians are expected to be in highly populated areas, and in the absence of dedicated facilities for pedestrians, pedestrians tend to use lanes next to curb-side which slows down vehicles' speeds and decreases the risk of high severity of pedestrian crashes, this phenomenon is called safety-in-numbers (Jacobsen, 2003;Elvik and Bjørnskau, 2017).

Vehicle type was also missing from the dataset, eventhough it is considered as an important predictor that affects the severity outcome of a vehicle-pedestrian crash. Ballesteros et al. (2004) found that size, weight, and design of vehicles involved in vehicle-pedestrian crashes are related to injuries of some types.

Conclusions

Most research on traffic crashes injury classification use injury severity as the class variable without taking into account the possibility of having better performance if another class variable was chosen for the same purpose. In order to find out if this assumption is correct, a comparison was made between Bayesian networks developed using injury severity as a class variable and that using collision type as a class variable. The results in this research work indicated that the developed Bayesian networks using collision type as a class variable obtained better performance (more reliable results).

On the other hand, one of the main challenges encountered in safety research that affect the classification of crashes is imbalanced datasets, as a rule of thumb, traffic crashes datasets are usually imbalanced and thus the results will be biased towards the class with higher number of records (majority class i.e. slight injuries class) (Mujalli et al. 2016). In the present work, a dataset was obtained from Jordanian Police Traffic Department, where the data that was obtained was imbalanced and hence needed to be balanced prior to developing models in order to enhance the classification of collisions.

For analysis and modeling purposes, three subsets were extracted from the original dataset; the first subset included all crash types but used collision type as a class variable, the second subset also included all crash types but used injury severity as a class variable, and the third subset included only pedestrian crash type and used injury severity as a class variable. All subsets were first balanced, and Bayesian networks were developed combining different search algorithms (K2 and simulated Annealing) and scores (BDeu, MDL, AIC). This process resulted in developing a total of six Bayesian networks per subset.

In order to measure the reliability of each developed Bayesian network using each subset, their classification capabilities were tested and the results indicated that the most reliable Bayesian networks were those which used collision type as the class variable.

The Bayesian network with the most reliable results from each subset was used to extract significant rules which if existed in the same crash, the crash will end up in a fatal or a severe injury. The extracted rules were grouped into three levels of confidence, as a result the following variables were found to significantly increase the probability of a fatal or severe injury outcome in pedestrian-vehicle collisions: speed limit, road type, and number of lanes where all these variables were present in the rules belonging to high confidence level group. Other variables that were found to be significant in fatal or severe injury collisions included: number of involved vehicles in a crash, lighting and adverse weather conditions. Consequently, using all subsets enhanced the extraction of rules and possibly using more variables (if available) would help extracting more valuable knowledge when more than one subset is employed. These results, which were consistent with previous studies, could be of a great interest in case they were used to enhance the existing shortcomings in the applied norms that govern the pedestrian safety on roads.

Recommendations

Authors recommend that some prevention strategies should be followed by both governmental and private agencies to improve road safety through actions aimed at decreasing the occurrence of pedestrian-vehicle collisions and resulting injuries and fatalities. Actions should be directed towards monitoring both attitude and behavior of drivers and pedestrians especially in urban areas where roads are wide and divided, without proper lighting, and with speed limits of 30-40 km/hr since these locations were found to be more hazardous and awareness must be raised to help reduce both fatalities and severity of collisions at these locations. Also, pedestrian sidewalks should be maintained and be only used by pedestrians, where crosswalks should be clearly marked and penalties should be imposed on drivers stopping on them so that pedestrians' penalties crossing from illegal points are accordingly activated. Speed limit compliance enforcement should be done using speed cameras especially at pedestrian "black spots", and traffic calming measures as well should be used more frequently at these locations to lessen the problem.

Authors believe that important knowledge could be extraced if the flow rates or average anual daily traffic (AADT) on travelled roadways where crashes occured were available, and hence a recommendation for Traffic Police Department is given to collaborate with the ministry of public Works and Housing and local muncipalities in order to collect and maintain this type of data.

Future work in this field might be directed towards using other data mining techniques in order to compare the results obtained using Bayesian networks. Moreover, to the knowledge of authors, no research is currently available about run-off-road crashes and property damage only crashes that occur in Jordan, and hence this might be a future line of investigation.

Acknowledgements

The authors are grateful to the Police Traffic Department in Jordan for providing the necessary data for this research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.