DALIA SHANSHAL, M.Sc. Candidate[1, 2]
E-mail: dshansha@ryerson.ca
CENI BABAOGLU, Ph.D.[2]
(Corresponding author)
E-mail: cenibabaoglu@ryerson.ca
AYŞE BAŞAR, Ph.D.[2]
(Corresponding author)
E-mail: ayse.bener@ryerson.ca
[1] The G. Raymond Chang School, Ryerson University
350 Victoria Street, Toronto, ON M5B 2K3, Canada
[2] Data Science Laboratory, Ryerson University
44 Gerrard Street East, Toronto, ON M5B 1G3, Canada

Safety and Security in Traffic
Original Scientific Paper
Submitted: 11 Dec. 2018
Accepted: 26 Sep. 2019

PREDICTION OF FATAL AND MAJOR INJURIES OF DRIVERS, CYCLISTS, AND PEDESTRIANS IN COLLISIONS

ABSTRACT

Traffic-related deaths and severe injuries may affect every person on the roads, whether driving, cycling or walking. Toronto, the largest city in Canada and the fourth largest in North America, aims to eliminate traffic-related fatalities and serious injuries on city streets. The aim of this study is to build a prediction model using data analytics and machine learning techniques that learn from past patterns, providing additional data-driven decision support for strategic planning. A detailed exploratory analysis is presented, investigating the relationship between the variables and factors affecting collisions in Toronto. A learning-based model is proposed to predict the fatalities and severe injuries in traffic collisions through a comparison of two predictive models: Lasso Regression and Random Forest. Exploratory data analysis results reveal both spatio-temporal and behavioural patterns such as the prevalence of collisions in intersections, in the spring and summer and aggressive driving and inattentive behaviours in drivers. The prediction results show that the best predictor of injury severity for drivers, cyclists and pedestrians is Random Forest with an accuracy of 0.80, 0.89, and 0.80, respectively. The proposed methods demonstrate the effectiveness of machine learning application to traffic and collision data, both for exploratory and predictive analytics.

KEY WORDS

collision; injury severity; prediction; classification; behavioural patterns;

1. INTRODUCTION

The World Health Organization estimates that over 3,400 people die in traffic collisions on a daily basis, and tens of millions are injured and disabled on a yearly basis [1]. In 2016, Canada's collisions leading to personal injuries reached a total of 115,956, and collisions leading to fatalities reached 1,717 [2]. In that same year, Toronto traffic fatalities hit their highest number since 2002 [3]. As a result, collision prevention, analysis, and prediction have been crucial topics in the traffic and transportation discipline [4]. Collisions are studied through various angles, such as the development of Accident Prediction Models (APM), road safety measure assessment, user behaviour analysis and others [4]. Various initiatives have been developed in response to this issue. In Europe, the PRACT project (Predicting Road Accidents) was developed in 2013 with the purpose of building an accident prediction model framework applicable to different European roads and networks [4]. In 1997, Sweden launched the Vision Zero project [5], aimed at eliminating road fatalities and serious injuries. Many countries have adopted this project, such as Canada, Germany, the UK, the Netherlands, and the US [6]. The Vision Zero Canada has been implemented in Edmonton [7], Vancouver [8], Ottawa [9], and Toronto [5, 6]. Toronto, the largest city in Canada and the fourth largest city in North America, saw a recent increase in road fatalities [10]. The Toronto Vision Zero Safety Plan is a 5-year plan (2017-2021) aiming at identifying the factors contributing to this type of collisions, with an ultimate goal of reducing collision fatalities and severe injuries to as close as possible to zero.

The goal in this study is to identify the patterns in Toronto severe and fatal collisions and to build a predictive model to estimate injury severity of

individuals in a collision, that is, drivers, pedestrians and cyclists. This paper is organized as follows: in Section 2, the related literature is discussed, then in Sections 3 and 4, an overview of the dataset is presented and the methodology discussed. In Section 5, data mining is performed and rules and patterns in Toronto collisions are presented. Section 6 presents and discusses the results of the predictive models, performance and the variable implications in the models. The threats to validity are considered in Section 7 and the paper is concluded in Section 8.

2. BACKGROUND

Different types of research have been undertaken in relation to collision analysis and prevention. Outside Canada, many studies analyse the physical aspect of a collision, such as structure, weight, and velocity of a car with regards to cyclists' [11] and pedestrians' injuries [12]. Both studies proposed safety measures to dampen the severity of injuries resulting from such collisions. Additionally, the analysis of children's injuries is conducted in the literature. The analysis of children injuries is performed using data from China [13] and Norway [14]. Research in [14], for example, found that misuse of the seatbelt is a major contributor of injuries in child passengers. Driver's characteristics and behaviour have been also extensively studied. The research in [15] demonstrates that attributes such as seatbelt misuse, speed higher than 111 km/h, female drivers and older drivers increase the probability of collision fatality. Similarly, studies on the drivers' behaviour and personality traits reveal that impulsivity and aggressiveness, as well as driver fatigue, are significant contributors to traffic collision occurrence [16, 17], and may lead to severe injuries [18].

Many studies use machine learning approaches to detect patterns and factors contributing to severe collisions. Research in [19] uses decision-tree-based algorithm to extract rules from the Spanish rural highway dataset, whereas research in [20] performs an extensive analysis to explore the factors contributing to collision occurrences. They construct a Bayesian network to classify crash types. Other different prediction models have been examined with regards to traffic collisions, such as artificial neural networks and support vector machine for predicting collision duration [21]; decision trees, Naïve Bayes, KNN and AdaBoost for predicting collision occurrence [22]; binary and skewed logistic regression [23], decision trees, multilayer perceptron [24], probabilistic neural net, Random Forests [25], and Bayesian networks [26] for predicting injury severity. Sensors and vehicle-to-vehicle communication [27], as well as genetic programming [28], are also investigated in the literature in the context of real-time collision prediction. Some studies have also taken a time series approach to analyse the fatalities in traffic collisions [29, 30]. Real-time driving environmental data have been explored in [31], where data such as real-time traffic flow, weather, road design, and others were added to the Colorado State Patrol crash database. The authors in [31] build a crash prediction model using mixed logit models, and find that weather, road surface, and traffic conditions play an important role in crash prediction. Other studies focus on injury severity. In [32], the authors explore the truck drivers' severity injury characteristics in single-vehicle and multi-vehicle accidents by building mixed logit models and evaluating the corresponding independent variables. Similarly, the authors in [33] analyse injury crash versus non-injury crash by building different spatio-temporal models and by evaluating the parameter estimates.

In Canada, a study was done in 2007, analysing the age and gender patterns in relation to collision injury, using the Canadian National Population Health Survey and Transport Canada data. It was found that injury rates between males and females are not significantly different; however, fatality rates in males are twice as high in Canada [34]. The children involved in collisions are also of interest in the Canadian research. Research [35] found that children in Canada are at a much higher risk of major injuries when involved in a back-over collision. The physical aspect of a collision is studied in a few cities in Canada, such as Edmonton [36] and Ottawa [37]. These studies analyse the proximity of two vehicles and its effect on the collisions.

In Toronto, a crash potential index (CPI)-based collision prediction model is built based on the proximity, velocity, and type of vehicles using past collisions data and data from loop detectors on Gardiner Expressway [38]. Further, research [39] investigates the pedestrians' injuries in collisions, but it is limited to children and elderly pedestrians' collisions. Another study [40] focuses on cyclists' injuries and road type analysis in both Vancouver and Toronto. The injury severity prediction models

in the Canadian collision datasets have not been discussed in the literature. This study aims to build models to predict the injury severity in collisions in the city of Toronto. For this type of problem, the regression models are the most widely used algorithms, mainly logistic regression in a classification problem [20, 23]. Due to sparsity of the data, Lasso Regression is used to avoid overfitting and to compare it with a tree-based model.

3. DATA DESCRIPTION

The KSI (Killed or Seriously Injured) dataset provided by the Toronto Police Services is used in this study. The dataset is now available at data.tps.on.ca as part of the Public Safety Data Portal [41]. It includes all traffic collision events in which at least one person was killed or seriously injured and covers the years from 2007 to 2017. The dataset includes 58 variables and 12,557 observations. The variables can be categorized into individual attributes and collision attributes. Individual attributes describe the characteristics and behaviour of each individual involved in the collision. Collision attributes describe the temporal, spatial and environmental conditions. Each row in the dataset represents an involvement type, an individual involved in the collision, such as a driver, pedestrian, etc.

The focus is on the drivers, cyclists, and pedestrians. Drivers include automobile drivers, motorcycle drivers, or truck drivers. Cyclists include bicycle riders and moped drivers. Pedestrians include any pedestrian, in-line skater, or wheelchair user.

As for the variable selection, these are decided based on two selection measures: (1) a qualitative selection is performed to remove redundant variables; (2) a quantitative selection is performed using an analysis of Spearman correlation, in which highly correlated variables are removed including multi-collinear variables. Additionally, data engineering was performed to extract monthly information and to merge injury levels into fatal or major, and, minimal, minor or none. Prior to merging injury levels, there was a total of 542 fatal injuries, 3,598 major injuries, 465 minimal injuries, 566 minor injuries, and 3,751 none. Their definition is as follows: (1) Fatal: person sustaining bodily injuries resulting in death. This includes only cases where death occurs in less than 366 days as a result of the collision (does not include death from natural causes such as heart attack, stroke, etc. or suicide). (2) Major: a non-fatal injury that is severe enough to require the injured person to be admitted to hospital, even if only for observation at the time of the collision (includes: fracture, internal injury, severe cuts, crushing, burns, concussions, severe general shocks). (3) Minor: a non-fatal injury requiring medical treatment at a hospital emergency room, but not requiring hospitalization of the involved person at the time of the collision. (4) Minimal: a non-fatal injury at the time of the collision, including minor abrasions, bruises, and complaints of pain, which does not require the injured person to go to hospital. (5) None: uninjured person.

The final dataset has 8,922 observations and 26 variables including both collision and individual related attributes (*Table 1*).

The data are subset into four different datasets: collisions, drivers, cyclists, and pedestrians, each including their specific attributes. Each of the subsets is examined for missing values or data inconsistencies. Fifteen variables have some blank values and two variables have data inconsistencies. Inconsistent values are corrected accordingly, and each blank record is added to an existing or new category that is either called Other or Unknown. The final variables selected are described in *Table 1*, where S1-S5 represent spatial characteristics, E1-E3 represent environmental characteristics, T1-T3 represent temporal characteristics, and I1-I13 represent traffic participant characteristics, including age, actions, conditions, type of vehicle operated at the time of collision and injury levels. Only the two most frequent levels are reported due to paper space capacity.

An initial analysis was carried out and it was found that during the eleven years from 2007 to 2017, KSI collisions followed a general decreasing pattern, going from 453 collisions in the year 2007 down to 331 collisions in 2017, the lowest number of collisions since 2007 (*Figure 1*). Similarly, both fatal and major injuries, as well as minimal, minor and no injury instances were at their lowest in 2017, decreasing by 26% and 60%, respectively since 2017. Meanwhile, the data obtained for all other collisions, including less serious collision types such as property-damage-only collisions, show an increase by 15% (*Figure 1*). In that same period, KSI collisions went down by 5%.

It was also found that the most frequent type of involvement were drivers, followed by pedestrians, cyclists, motorcycle drivers and truck drivers

Table 1 – Variable description

Variable		Category
Code	Description	
ACCNUM	Collision ID	4397 Unique collision IDs
S1: District	District	1. Toronto East York (3,042); 2. Etobicoke York (2,080)
S2: LOCCOORD	Location coordinate	1. Intersection (6,031); 2. Mid-block (2,835)
S3: ROAD_CLASS	Road class	1. Major arterial (6,131); 2. Minor arterial (1,518)
S4: TRAFFCTL	Traffic control	1. No control (4,229); 2. Traffic signal (3,777)
S5: Ward_ID	Ward ID	Wards 1 to 44. 1. Ward 20 (530); 2. Ward 28 (413)
E1: VISIBILITY	Visibility	1. Clear (7,645); 2. Rain (983)
E2: LIGHT	Light	1. Daylight (5,346); 2. Dark (1,821)
E3: RDSFCOND	Road surface condition	1. Dry (7,088), 2. Wet (1,551)
T1: Hour	Hour	Hour from 0 to 23. 1. 18 (619); 2. 17 (578)
T2: YEAR	Year	Years from 2007 to 2017 1. 2007 (935); 2. 2012 (905)
T3: month	Month	1 to 12
I1: INVAGE	Age of involved individual	Ages 0 to over 95. 1. [25 to 29] (902); 2. [20 to 24] (840)
I2: VEHTYPE	Vehicle type	1. Automobile, Station wagon (5,071); 2. Other (1,390)
I3: MANOEUVER	Driver manoeuver	1.Going ahead (4129); 2. Turning left (1,199)
I4: INITDIR	Initial direction	1. East (2,198); 2. West (2,115)
I5: DRIVACT	Driver action	1.Driving properly (2,858); 2. Failed to yield right of way (1,035)
I6: DRIVCOND	Driver condition	1. Normal (3,989); 2. Inattentive (1,061)
I7: PEDTYPE	Pedestrian crash Type details	1. Pedestrian hit at mid-block (513); 2. Vehicle turns left while ped crosses with ROW at inter. (414)
I8: PEDACT	Pedestrian action	1. Crossing with right of way (642); 2. Crossing, no traffic control (471)
I9: PEDCOND	Pedestrian condition	1. Normal (1,125); 2. Inattentive (359)
I10: CYCLISTYPE	Cyclist crash Type details	1. Motorist turned left across cyclists path (90); 2. Cyclist without ROW rides into path of motorist at inter, lnwy, dwy-cyclist not turn (77)
I11: CYCACT	Cyclist action	1. Driving properly (285); 2. Disobeyed traffic control (58)
I12: CYCCOND	Cyclist condition	1. Normal (354); 2. Inattentive (80)
I13: INJURY	Injury level	1.Fatal or major (4783); 2. Minimal, minor or none (4,139)

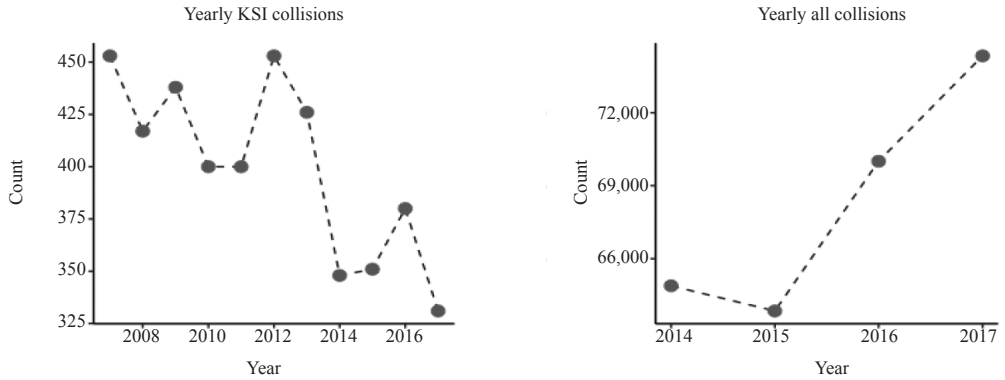

Figure 1 – Yearly collision frequency

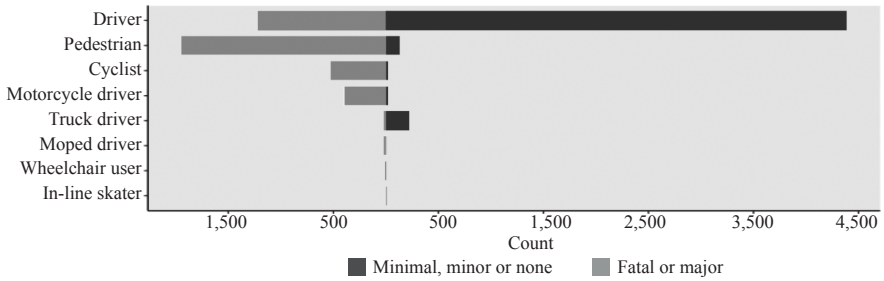

Figure 2 – Traffic participant injury severity

(sedan drivers) (*Figure 2*). Pedestrians, cyclists and motorcycle drivers are the most affected in collisions, with more than 90% of each of the involvement mentioned above type having a major or fatal injury.

4. METHODOLOGY

We use the *a priori* association rule technique to mine the dataset and uncover patterns and rules between the variables [42]. An association rule is of the form $X \Rightarrow Y$, where $X=\{x_1,x_2,\ldots,x_n\}$, and $Y=\{y_1,y_2,\ldots,y_m\}$ are two sets of mutually exclusive observations. For an association rule to be of interest, it must satisfy two interest measures: support and confidence. Support is an indication of how often an observation or a set of observations appear in the dataset and it equals $P(X,Y)$. Confidence measures the strength of the rule and is equal to $P(Y|X)$. A rule of the form $\{\} \Rightarrow \{Y\}$ means that the observation in Y will appear with the probability given by the rule support (which equals confidence).

To predict the severity of an injury as one of the two classes: 'fatal or major' and 'minimal, minor or none,' a classification approach was used.

In such a binary context, the authors in [43] stated one possible limitation related to endogeneity of the explanatory variables: "one potential concern […] is the possibility that the explanatory variables may be endogenous with respect to injury severity". The authors explained that a possible solution for that is to add more variables, which in turn can provide a better explanation of the overall picture. The authors gave an example of airbag as an explanatory variable. They stated that drivers owning vehicles with airbags may also tend to be risk-averse [44]. As such, airbags can be coupled with risk-averseness variable to avoid endogeneity problem and over-estimation of the importance of the airbag variable.

In this study, the data used are very sparse, with high dimensionality. Due to sparsity, dimensionality reduction (quantitative and qualitative as mentioned in Section 3) was performed. Adding more variables will give our models a propensity to over-fit the data, resulting in inaccurate outcomes. Additionally, many of our variables inherently include other information; such as variable CYCLISTYPE, which has 22 levels (examples of this variable are found in *Table 1*). We wanted to keep the original levels for replicability purposes. Moreover, when later the variable importance is discussed, we list specific levels, and not just the variable itself, in order to distinguish the effect of different possible causes, and to avoid overestimating one particular variable.

Two classification algorithms are used: Lasso Regression and Random Forest. Lasso Regression is a method for the estimation in linear models that performs variable selection and regularization, which is an approach to fine-tuning model complexity. It is used to deal with the sparsity in our dataset. A sparse dataset implies high variance. As per the bias-variance trade-off, high variance in the dataset increases the model complexity and the mean squared error [42]. Lasso Regression adds a penalty (lambda) to the coefficients, and therefore reduces the model complexity [42]. Random forest is a tree-based classifier that uses an impurity measure (Gini) to decide on the best split. Each variable is considered as a candidate for a splitting node. Splits are assessed and chosen using the Gini impurity measure. A split is pure if after the split, for all branches, all the instances choosing a branch belong to the same class [42]. A low Gini measure indicates that the split variable is important for data partitioning.

The main difference between the two models is how they deal with complexity and generalizability. In Random Forest, which is an ensemble method, complexity is decreased through the training process. In Lasso Regression, complexity is decreased through regularization, where an augmented error function is used [42].

Table 2 – Confusion matrix

	Actual fatal/major	Actual minimal/minor/none
Predicted fatal/major	*TP*	*FN*
Predicted minimal/minor/none	*FP*	*TN*

The advantage of Lasso Regression is its ability to take into account the correlation among the variables; its weakness, however, is that some features' coefficients can be reduced to 0 through regularization; therefore, bias could be introduced in the model. Random Forest's advantage, on the other hand, is its ability to deal with complexity and generalization error. This is done by its training process, and also by pre-pruning the tree. Pre-pruning the tree ensures that a node is not split further if the number of observation reaching that node is smaller than a certain percentage of the training set [42].

For modelling, the dataset is divided into 80% training set and 20% test set; then a 10-fold cross validation is conducted on the training set. Because the dependent variable is imbalanced in each of the three subsets (as seen in *Figure 2*), it is treated using Synthetic Minority Oversampling Technique (SMOTE) [45]. In the drivers' subset, 'fatal or major' instances are oversampled and 'minimal, minor or none' are undersampled. The opposite is done for pedestrians and cyclists.

To assess the performance of the proposed predictor, the performance measures used in two-class problems (*Table 2*) are used. The number of true 'minimal, minor or none' estimations are denoted with *TN*, the number of false 'true minimal, minor and none' estimations with *FN*, the number of 'false fatal or major' estimations with *FP*, and the number of true 'fatal and major' estimations with *TP*.

The *accuracy* measures the rate of correct estimations (*Equation 1*). The True Positive Rate is also known as *sensitivity* (*Equation 2*). The True Negative Rate is also known as *specificity* (*Equation 3*) [42].

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$specificity = \frac{TN}{TN + FP} \quad (3)$$

The variable importance in each model is then analysed to detect which variables have the most weight on the models. For Lasso Regression, the coefficient t-test is used [42]. For Random Forest, the out-of-bag error is used [46]. The measures reported are scaled (0-100).

5. DATA MINING

The apriori algorithm was applied to collision subset. It was noticed that the majority of collisions occur on major arterials and/or intersections (Rules 1, 2, *Table 3*). Within the collisions taking place in major arterials, 72% are located in intersections (Rule 3, *Table 3*).

Collisions in Toronto mostly occur in locations where there is a traffic signal or no traffic control at all. Ninety percent of collisions in Toronto took place in locations with either of those two traffic control characteristics (traffic signal or no traffic control) (Rules 4, 5, *Table 3*).

It can be seen that the largest proportion of collisions happen under clear and dry conditions, and in daylight (Rules 6, 7, 8, *Table 3*). It is found that these three characteristics together occur 51% of the time (Rule 9, *Table 3*).

A related trend is noticed in the time patterns of collisions; that is, most collisions occur during the summer/spring season, seasons associated with dry, and clear conditions. Additionally, one can see

Table 3 – Collision subset rules

Rule 1: {} => {ROAD_CLASS=Major Arterial}; Support = 0.67; Confidence = 0.67; Count = 2,928
Rule 2: {} => {LOCCOORD=Intersection}; Support = 0.66; Confidence =0.66; Count = 2,899
Rule 3: {ROAD_CLASS=Major Arterial} => {LOCCOORD=Intersection}; Support = 0.48; Confidence = 0.72; Count = 2,089
Rule 4: {} => {TRAFFCTL= No Control}; Support = 0.50; Confidence = 0.50; Count = 2,179
Rule 5: {} => {TRAFFCTL=Traffic Signal}; Support = 0.40; Confidence = 0.40; Count = 1,763
Rule 6: {} => {VISIBILITY=CLEAR}; Support = 0.86; Confidence = 0.86; Count = 3,752
Rule 7: {} => {RDSFCOND=Dry}; Support = 0.79; Confidence = 0.79; Count = 3,470
Rule 8: {} => {LIGHT=Daylight}; Support = 0.59; Confidence = 0.59; Count = 2,578
Rule 9: {LIGHT=Daylight, RDSFCOND=Dry} => {VISIBILITY=Clear}; Support = 0.51; Confidence = 0.997; Count = 2,207

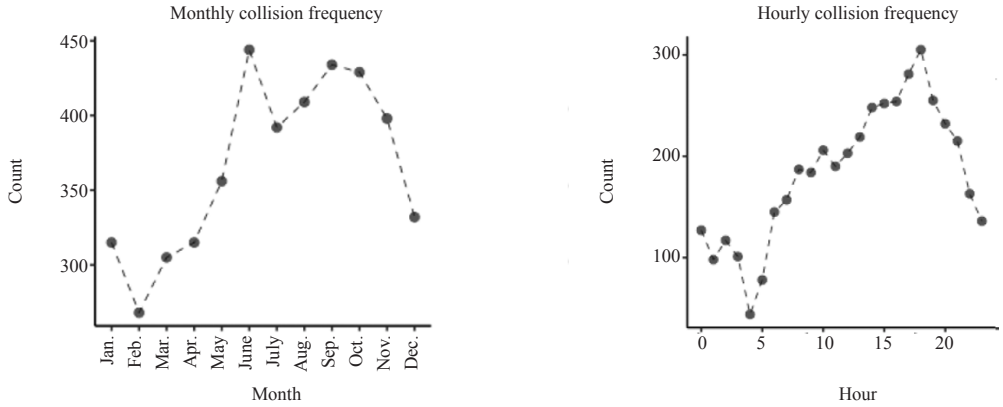

Figure 3 – Monthly and hourly collisions

within the hourly patterns of collisions (*Figure 3*) that collisions peak between 4 p.m. to 7 p.m., a period usually associated with the end of a working day, and, in the summer and spring season, associated with daylight.

To find the underlying issue of these time and location trends, the behavioural patterns within the most common collision dynamics are investigated. These are one driver and one pedestrian collision, which represent 40% of all collisions in the dataset (1,689 collisions), and two drivers' collisions, which represent 25% of collisions in the dataset (1,113 collisions).

In the one driver and one pedestrian collisions, the intersections were found to be the most frequent collision locations (70% of all such collisions) (Rule 1, *Table 4*). It was also noticed that collisions that occur while a pedestrian is crossing with right of way at an intersection, is almost always associated with a driver failing to yield right of way; this happens 85% of the time (Rule 2, *Table 4*).

Failing to yield right of way is the most common aggressive driving behaviour. Aggressive driving is defined as any of the following actions [47]: exceeding speed limit, speeding too fast for the conditions, following too close, disobeying traffic control, failing to yield right of way, passing improperly. One-third of the drivers in our dataset exhibited aggressive driving behaviour (31%). Amongst these drivers, 55% failed to yield right of way, and 17% disobeyed traffic control; these are the two most common aggressive driving behaviours.

The data show that failing to yield right of way is a common action in case of inattentive drivers. The vast majority (85%) of inattentive drivers failed to yield right of way (Rule 3, *Table 4*). It was also observed that when a vehicle is turning right while a pedestrian is crossing with right of way, 88% of the time that driver failed to yield right of way while turning (Rule 4, *Table 4*). Similarly, when a driver is turning left, 78% of the time that driver failed to yield right of way (Rule 5, *Table 4*).

It can be seen that the turning left manoeuver occurs in almost a third (27%) of the one driver and one pedestrian collisions (Rule 6, *Table 4*). In the one driver and one pedestrian collisions, the likelihood

Table 4 – One driver - One pedestrian subset rules

Rule 1: {} => {LOCCOORD=Intersection}; Support = 0.70; Confidence = 0.70; Count = 1,176
Rule 2: {LOCCOORD=Intersection, PEDACT=Crossing with right of way} => {DRIVACT= Failed to Yield Right of Way}; Support = 0.27; Confidence = 0.85; Count = 443
Rule 3: {DRIVCOND=Inattentive} => {DRIVACT=Failed to Yield Right of Way}; Support = 0.27; Confidence = 0.85; Count = 235
Rule 4: {PEDTYPE=Vehicle turns right while ped. crosses with ROW at inter.} => {DRIVACT=Failed to Yield Right of Way}; Support = 0.14; Confidence = 0.69; Count = 345
Rule 5: {MANOEUVER=Turning Left} => {DRIVACT=Failed to Yield Right of Way}; Support = 0.21; Confidence = 0.78; Count = 456
Rule 6: {} => {MANOEUVER=Turning Left}; Support = 0.27; Confidence = 0.27; Count = 354
Rule 7: {PEDACT=Crossing with right of way, LOCCOORD=Intersection} => {PEDTYPE=Vehicle turns left while ped. crosses with ROW at inter.}; Support = 0.21; Confidence = 0.66; Count = 87

of a driver turning left given that a pedestrian is crossing with right of way at an intersection is 66% (Rule 7, *Table 4*).

In collisions between two drivers, 64% of the time, one driver simply goes ahead (Rule 1, *Table 5*). Whenever one driver fails to yield right of way on a left turn, the driver almost always collides with the driver going ahead (Rule 2, *Table 5*). It is also observed that when a driver makes an improper turn (Rule 3, *Table 5*), or turns left inattentively (Rule 4, *Table 5*), the other driver almost always goes ahead. However, it is also seen that a small portion of drivers going ahead disobey traffic control. Most of the time, these drivers collide with another driver who is, in turn, driving properly (Rule 5, *Table 5*). Similarly, drivers who follow too close or drive inattentively have a 90% probability and more of colliding with a driver who drives properly (Rule 6, 7, *Table 5*).

In the first type of collision, which is one driver and one pedestrian collision, 1,695 individuals get fatally or majorly injured. In the second type; the collision between two drivers, there are 948 fatally or majorly injured. The patterns leading to major or fatal injuries amongst each subgroup of drivers, pedestrians and cyclists, are analysed.

Amongst drivers, the majority of fatal or major injuries occur as a consequence of losing control of the vehicle amounting to 422 collisions (Rule 1, *Table 6*). Particularly on mid-blocks where there is no traffic control. In fact, losing control of a vehicle in such a location is associated with a 94% probability of fatal or major injury (Rule 2, *Table 6*).

On the other hand, the drivers' subset presents a new finding regarding motorcyclists. Motorcyclists have a 94% probability of a fatal or major injury (Rule 3, *Table 6*). More specifically, motorcyclists going ahead in an intersection where a traffic signal is located, and, either on the major arterial or in normal condition, have a probability of 97% or more of a fatal or major injury (Rules 4, 5, 6, *Table 6*). Motorcyclists driving in Toronto East York during daylight also have a similar probability of fatal or major injury (Rule 7, *Table 6*).

Table 5 – Two drivers subset rules

Rule 1: {} => {MANOEUVER=Going Ahead}; Support = 0.64; Confidence = 0.64; Count = 707
Rule 2: {DRIVACT_2=Failed to Yield Right of Way, MANOEUVER_2=Turning left} => {MANOEUVER=Going Ahead}; Support = 0.10; Confidence = 0.99; Count = 111
Rule 3: {DRIVACT_2=Improper Turn} => {MANOEUVER=Going Ahead}; Support = 0.11; Confidence = 0.98; Count = 118
Rule 4: {MANOEUVER_2=Turning Left, DRIVCOND_2=Inattentive} => {MANOEUVER=Going Ahead}; Support = 0.08; Confidence = 0.99; Count = 90
Rule 5: {MANOEUVER=Going Ahead, DRIVACT=Disobeyed Traffic Control} => {DRIVACT_2=Driving Properly}; Support = 0.05; Confidence = 0.91; Count = 50
Rule 6: {DRIVACT_2=Following too Close} => {DRIVACT=Driving Properly}; Support = 0.07; Confidence = 0.95; Count = 71
Rule 7: {DRIVCOND_2=Inattentive} => {DRIVACT=Driving Properly}; Support = 0.20; Confidence = 0.93; Count = 221

Table 6 – Drivers subset rules

Rule 1: {DRIVACT=Lost control} => {INJURY=Fatal or Major} Support = 0.07; Confidence = 0.68; Count = 422
Rule 2: {LOCCOORD=Mid-Block, TRAFFCTL=No Control, DRIVACT=Lost control} => {INJURY=Fatal or Major}}; Support = 0.04; Confidence = 0.94; Count = 255
Rule 3: {VEHTYPE=Motorcycle Driver} => {INJURY=Fatal or Major}; Support = 0.06; Confidence = 0.94; Count = 391
Rule 4: {LOCCOORD=Intersection, TRAFFCTL=Traffic Signal, VEHTYPE=Motorcycle, MANOEUVER=Going Ahead} => {INJURY=Fatal or Major}; Support = 0.02; Confidence = 0.98; Count = 102
Rule 5: {LOCCOORD=Intersection, TRAFFCTL=Traffic Signal, ROAD_CLASS=Major Arterial, VEHTYPE=Motorcycle, MANOEUVER=Going Ahead} => {INJURY=Fatal or Major}; Support = 0.01; Confidence = 0.98; Count = 83
Rule 6: {LOCCOORD=Intersection, TRAFFCTL=Traffic Signal, VEHTYPE=Motorcycle, MANOEUVER=Going Ahead, DRIVCOND=Normal} => {INJURY=Fatal or Major}; Support = 0.01; Confidence = 0.98; Count = 81
Rule 7: {District=Toronto East York, Light=Daylight, VEHTYPE=Motorcycle => {INJURY=Fatal or Major}; Support = 0.02; Confidence = 0.98; Count = 102
Rule8: {VEHTYPE=Automobile, Station Wagon, DRIVCOND=Medical or Physical Disability} => {INJURY=Fatal or Major}; Support = 0.02; Confidence = 0.90; Count = 103

It can be noticed that drivers with medical or physical disability are also more prone to fatal or major injuries, particularly those driving an automobile or a station wagon (Rule 8, *Table 6*).

The risk of cyclists' fatal or major injury in the months of June and July exceeds 95% (Rules 1, 2, *Table 7*). As noted earlier, these months have a very high collision frequency (*Figure 3*).

Consistent with our previous findings, it was noticed that the cyclists' fatality or major injuries occur primarily on major arterials or intersections (Rules 3, 4, *Table 7*).

Many rules were found in which 100% of injuries were fatal or major. For example, all cyclists' collisions in ward 18 and ward 28 resulted in such severe injuries (Rules 5, 6, *Table 7*). Also, it appears that cyclists aged 50 to 54, although driving properly, are also greatly affected in collisions (Rule 7, *Table 7*).

Similarly, three types of collisions were detected that always result in a fatal or major injury. These are collisions that involve a cyclist and a driver travelling in the same direction where one vehicle sideswipes the other, a motorist turning left across the cyclists' path, and cyclists struck by the opened vehicle door (Rules 8, 9, 10, *Table 7*).

When it comes to pedestrians, we see that one-fourth of pedestrians are fatally or majorly injured on mid-blocks (Rule 1, *Table 8*), particularly on major arterial (Rule 2, *Table 8*).

Table 7 – Cyclists subset rules

Rule 1: {month=06} => {INJURY=Fatal or Major}; Support = 0.16; Confidence = 0.96; Count = 91
Rule 2: {month-07} => {INJURY=Fatal or Major}; Support = 0.12; Confidence = 0.98; Count = 67
Rule 3: {ROAD_CLASS=Major Arterial} => {INJURY=Fatal or Major}; Support = 0.60; Confidence = 0.95; Count = 342
Rule 4: {LOCCOORD=Intersection, MANOEUVER=Going Ahead} => {INJURY=Fatal or Major}; Support = 0.55; Confidence = 0.96; Count = 315
Rule 5: {Ward_ID=18, LOCCOORD=Intersection} => {INJURY=Fatal or Major}; Support = 0.06; Confidence = 1; Count = 34
Rule 6: {Ward_ID=28, District=Toronto East York} => {INJURY=Fatal or Major}; Support = 0.06; Confidence = 1; Count = 34
Rule 7: {INVAGE=50 to 54, CYCACT=Driving Properly} => {INJURY=Fatal or Major}; Support = 0.06; Confidence = 1; Count = 33
Rule 8: {TRAFFCTL=No Control, CYCLISTYPE=Cyclist and Driver travelling in same direction. One vehicle sideswipes the other} => {INJURY=Fatal or Major}; Support = 0.08; Confidence = 1; Count = 44
Rule 9: {VISIBILITY=Clear, RDSFDCOND=Dry, CYCLISTYPE=Motorist turned left across cyclists path., CYCCOND=Normal} => {INJURY=Fatal or Major}; Support = 0.12; Confidence = 1; Count = 66
Rule 10: {District=Toronto East York, CYCLISTYPE=Cyclist struck opened vehicle door} => {INJURY=Fatal or Major}; Support = 0.08; Confidence = 1; Count = 42

Table 8 – Pedestrian subset rules

Rule 1: {PEDTYPE=Pedestrian hit at mid-block} => {INJURY=Fatal or Major} Support = 0.24; Confidence = 0.96; Count = 494
Rule 2: {PEDTYPE=Pedestrian hit at mid-block, ROAD_CLASS=Major Arterial} => {INJURY=Fatal or Major}; Support = 0.16; Confidence = 0.96; Count = 337
Rule 3: {TRAFFCTL=No Control, PEDTYPE=Pedestrian hit at mid-block, VISIBILITY=Clear, RDSFCOND=Dry} => {INJURY=Fatal or Major}; Support = 0.17; Confidence = 0.96; Count = 357
Rule 4: {TRAFFCTL=No Control, VISIBILITY=Clear, PEDACT=Crossing, no Traffic Control} => {INJURY=Fatal or Major}; Support = 0.16; Confidence = 0.96; Count = 338
Rule 5: {District=Toronto East York, LOCCOORD=Intersection, PEDTYPE=Vehicle turns left while ped. crosses with ROW at inter.} => {INJURY=Fatal or Major}; Support = 0.06; Confidence = 0.96; Count = 120
Rule 6: {LOCCOORD=Intersection, RDSFCOND=Wet} => {INJURY=Major or Fatal}; Support = 0.16; Confidence = 0.95; Count = 333
Rule 7: {LOCCOORD=Intersection, ROAD_CLASS=Major Arterial, VISIBILITY=Rain} => {INJURY=Fatal or Major}; Support = 0.09; Confidence = 0.97; Count = 184
Rule 8: {LIGHT=Dark, RDSFCOND=Wet} => {INJURY=Fatal or Major}; Support = 0.08; Confidence= 0.96; Count = 170

Areas with no traffic control also result in high probability of pedestrian fatal or major injury, particularly in cases where pedestrians are hit at mid-block or when pedestrians are crossing in areas with no traffic control (Rules 3, 4, *Table 8*).

Intersections are also risky areas when it comes to pedestrian injuries. In Toronto East York, for example, a vehicle turning left at an intersection while a pedestrian is crossing with right of way is associated with 96% probability of fatal or major injury (Rule 5, *Table 8*). This finding is consistent with the rules discovered earlier regarding one driver and one pedestrian collision type, where it was found that many drivers fail to yield right of way on a left turn.

It was noticed that pedestrians' injury level is affected by the weather. At an intersection, a rainy day and wet surface condition result in a major or fatal injury 95% of the time or more (Rules 6, 7, *Table 8*). In general, a wet road surface condition and a dark lighting condition (the time between sunset and sunrise) is associated with a 96% probability of major or fatal injury (Rule 8, *Table 8*).

6. RESULTS OF PREDICTION MODELS

The performance measures were used to assess how well each algorithm predicts the injury severity, and the analysis of variance to test for statistical difference between the models. The test showed that the two models are statistically different for each subset (p-value<0.05). Both Random Forest and logistic regression resulted in a good prediction with a minimum of 76% accuracy and maximum accuracy of 89%. However, it is observed that Random Forest algorithm is consistently generating higher overall accuracy for all the subsets (*Table 9*). Random Forest, as a non-linear model, uses the mean decrease Gini statistics as the basis for deciding on the splitting node. In this way, Random Forest captures the importance of each variable in classification.

To understand which variables affect the models the most, the top 20 most important variables in the models are listed.

Within the driver model, motorcycle has the most weight importance in both Random Forest and Lasso Regression models. There exist other common variables between the two models; these are medical or physical driver disability, losing control of the vehicle, and failing to yield right of way (*Figure 4*).

As for cyclists and pedestrians, it can be seen that Random Forest captured behavioural variables, whereas Lasso Regression captured mostly locations and hours. The common variable between the two models in the cyclists' subset is age-related; it is cyclists aged 50 to 54. Within the pedestrian subset, there are no common variables (*Figures 5* and *6*).

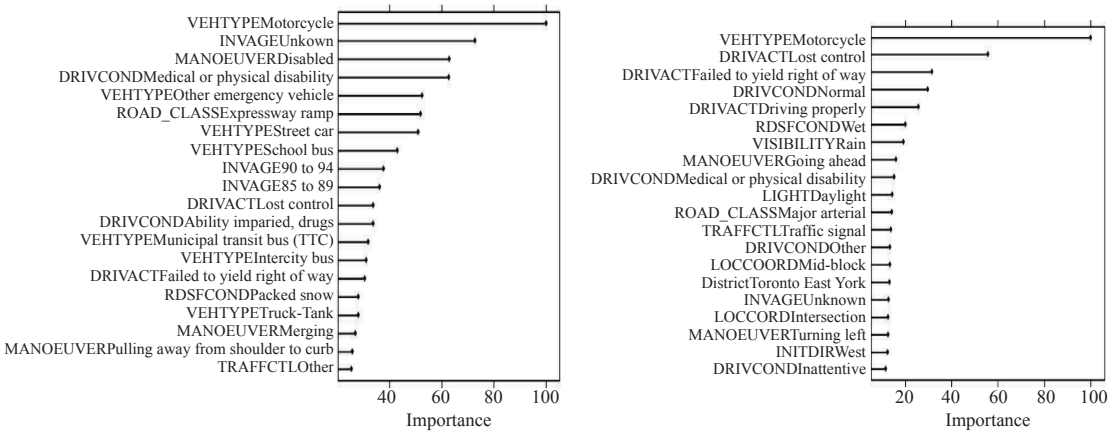

Figure 4 – Drivers top 20 GLM and RF variable importance

Table 9 – Performance metrics

Metrics	Drivers lasso	Drivers random forest	Cyclists lasso	Cyclists random forest	Pedestrians lasso	Pedestrians random forest
Training set accuracy	0.78	0.81	0.95	0.89	0.78	0.77
Test set accuracy	0.76	0.80	0.87	0.89	0.76	0.80
Test set sensitivity	0.62	0.64	0.91	0.91	0.78	0.83
Test set specificity	0.81	0.84	0.34	0.5	0.5	0.4

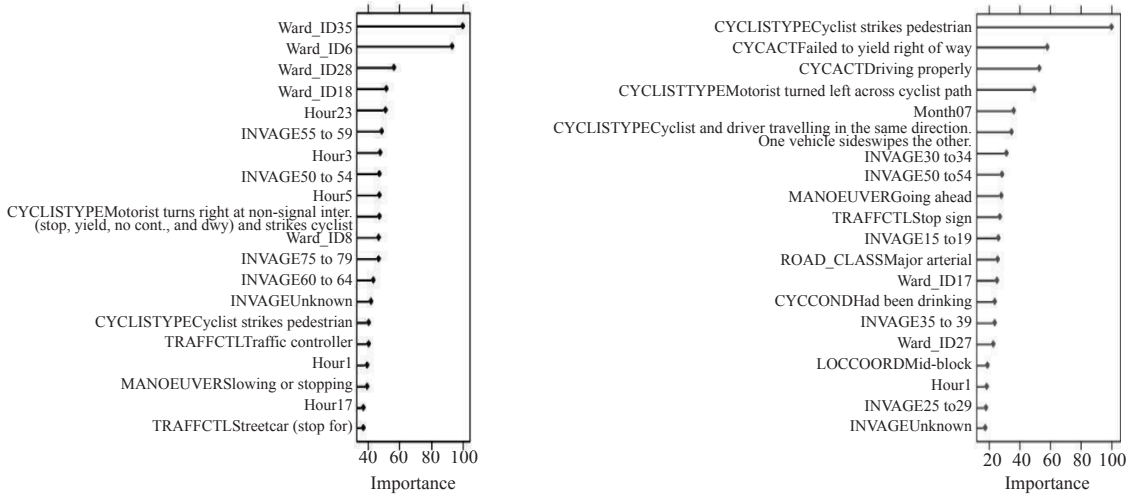

Figure 5 – Cyclists top 20 GLM and RF variable importance

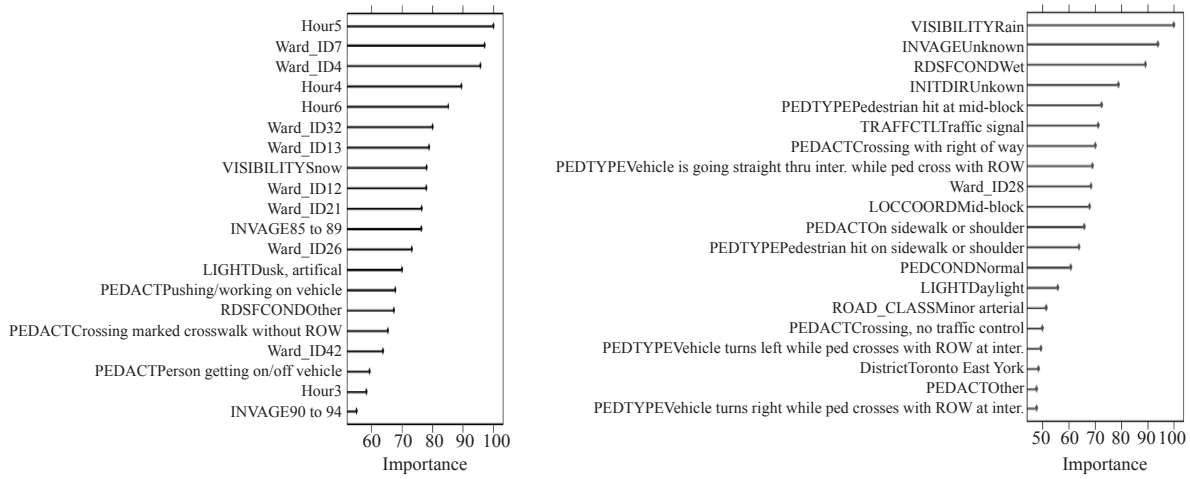

Figure 6 – Pedestrians top 20 GLM and RF variable importance

It was also noticed that Random Forest captures many of the patterns presented in the data mining section. Within the drivers' subset, ten out of the 20 variables listed in Random Forest are discussed in Section 4, such as motorcycle, losing control of the vehicle and intersection. Logistic regression only includes three out of 20.

Within cyclists, eight out of the 20 variables listed in Random Forest model are discussed, such as major arterial, intersection and motorist turning left across the cyclist's path. Logistic regression only includes three of the 20 variables listed.

Within pedestrians, logistic regression does not list any of the variables discussed in Section 4, whereas Random Forest lists seven, such as pedestrian hit at mid-block, crossing at no traffic control area and rain.

However, when analysing the 20 most important variables in logistic regression, it can be seen that many of the variables are associated with 100% fatal or major injury probability. For example, in case of pedestrians, the following variables are associated with 100% fatal or major injury: ward 4, ward 12, ward 26, hour 5, snow, age 90 to 94, and a person getting on/off a vehicle. Another example is the wards discussed in the cyclists' association rules, where it was found that wards 28 and 18 are associated with 100% fatal or major injury. These instances, however, represent less than 35 cases within the pedestrians and cyclists' subsets.

7. THREATS TO VALIDITY

Internal Validity. The dataset under consideration in this study had some missing values. We were informed by the Toronto Police Service that there could be some cases where police officers may have skipped some items in the questionnaire especially when the conditions were normal. The removal of the missing records or their implementation by means or

mode could cause concern for internal validity. In order to mitigate this effect, we performed a very detailed exploratory analysis and used information within the dataset to impute the missing values.

External Validity. To analyse the whole dataset without any sub-setting of drivers, cyclists and pedestrians could result in an external threat to validity since our model would not be generalizable. To ensure generalizability of our results, we ensured that each of the involved types was treated separately.

Construct Validity. In a binary classification setting, in our case fatal/major vs. non-fatal/minor outcome, a class imbalance affects the impact that a given exploratory variable has on the outcome, which can cause construct validity. To overcome this problem, we treated our data for imbalance prior to applying the models.

Statistical Conclusion Validity. The association rules have been qualitatively selected due to the high number of rules (exceeding 10,000). As such, the findings presented are not exhaustive of all possible rules. However, we ensured that the rules selected were based on both the highest support and confidence, and a lift greater than 1.

8. DISCUSSION AND CONCLUSION

This paper analyses and predicts the collision injury severity in Toronto using both data mining techniques (association rules), and classification algorithms (Lasso regression and Random Forest).

Severe collision prevention measures can be tackled by spreading more awareness among the drivers, pedestrians and cyclists. We found that drivers tend to get involved in severe collisions when the following characteristics are exhibited: aggressive driving, particularly failing to yield right of way and improper turns, and inattentiveness. We found that pedestrians are at a much higher risk of severe injuries when crossing at mid-blocks, whereas cyclists are at high risk of severe injuries particularly when colliding with motorcyclists.

The prediction of such injuries through Lasso regression model and a Random Forest tree-based model is promising. We found that Random Forest's accuracy consistently exceeded Lasso regression's accuracy for all three subsets: drivers, cyclists and pedestrians.

Moreover, we noticed that Random Forest was able to generalize better as observed in Section 6. As mentioned in Section 4, the two algorithms differ in how they deal with complexity and generalizability. We observed that Lasso Regression model gave much importance to features that are associated with 100% fatal or major injuries such as specific wards or specific manoeuvers. For example, there are only three observations of disabled manoeuver in our drivers subset, yet, our Lasso Regression model considered this feature as one of the top five most important features; that is likely due to the fact that all three observations are associated with fatal or major injury. In that sense, we can say that Lasso Regression's feature importance selection is very precise in terms of selecting the features that best distinguish the fatal or major instances versus minimal, minor or none instances. However, overall, Random Forest generalizes better, with the most important features reflecting 'big patterns' in the dataset as highlighted in Section 5; that is due to the Random Forest training process.

Based on the findings in our data mining section and the prediction results section, the following summary conclusion is drawn: (a) The temporal and environmental characteristics of severe collisions can be summarized as follows: as shown in *Figure 3*, severe collisions in Toronto occur most frequently in the summer and spring, particularly in clear and dry conditions. Cyclists sustain major and fatal injuries particularly during the months of June and July, whereas pedestrians' risk of fatal or major injury increases in rainy conditions, in case of wet surfaces and dark light as shown in both the data mining section and feature analysis in the prediction model selection; (b) The spatial characteristics can be summarized as follows: in both data mining and prediction model we see severe collisions recurring in major arterials and intersections. Intersections are particularly high-risk locations for the pedestrians. These, along with mid-block, traffic signal and no traffic control represent the riskiest spatial features of severe collision occurrences for all the traffic participants (drivers, cyclists, pedestrians). Pedestrians are highly at risk of severe injuries in collisions taking place at mid-blocks and in no traffic control areas, whereas motorcyclists are at high risk of severe injuries at intersections where traffic signal is present; (c) Behavioural characteristics, including drivers' action and condition are summarized as follows: we see a recurrent pattern of aggressive and inattentive driving behaviours. In aggressive driving, the most common behaviour is failing to yield right of way, mostly at left turns, but also at right turns. Together with inattentive

driving at intersections, these characteristics constitute the majority of severe collisions in Toronto. This is applicable to both collisions where one driver and one pedestrian are involved, and where two drivers are involved. Another aggressive driving behaviour appears in two drivers' collisions, that is, following too close and disobeying traffic control. Drivers also seem to be at high risk of severe injuries in collisions when they lose control of the vehicle or when they have a medical or physical disability. Although medical and physical disability observations are low in our data, we were informed by TPS that these may be much higher due to the fact that not all drivers disclose that information to the police officer. As for collisions where cyclists suffer major or fatal injuries, we noticed that these are mostly associated with the following actions and conditions: driver sideswipes cyclists while driving in the same direction, motorist turning left across the cyclist's path and cyclists struck by the opened vehicle door.

The goal of such a comprehensive study of different risk factors affecting drivers, cyclists and pedestrians including temporal, environmental, spatial and behavioural characteristics, is to highlight the different features involved in severe collisions in order to facilitate the decision making of effective traffic safety and injury prevention measures. These can be translated into decisions such as: the decision to dispatch more officers on the roads given specific temporal, environmental and spatial characteristics, the design of traffic safety campaigns run by the Toronto Police Services, including strategic messaging, and the spread of more awareness about aggressive and inattentive driving.

Moving forward, we aim to include more datasets from the Toronto Police Service and the City of Toronto to make the results more generalizable.

ACKNOWLEDGEMENTS

We would like to show our gratitude to Ian Williams, Daphne Choi, Debbie Verduga and Meghan Fotak from the Toronto Police Service who provided insight and expertise that greatly assisted the research.

REFERENCES

[1] World Health Organization. *Violence and Injury Prevention, Road traffic injuries*. Available from: http://www.who.int/violence_injury_prevention/road_traffic/en/ [Accessed June 2nd 2018].

[2] Government of Canada, Transport Canada. *Canadian Motor Vehicle Traffic Collision Statistics: 2016.* Available from: https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2016.html [Accessed June 2nd 2018].

[3] McGillivray K. Toronto traffic fatalities hit 14-year high. *CBC News*. December 5 2016. Available from: http://www.cbc.ca/news/canada/toronto/traffic-fatalities-high-1.3882140 [Accessed June 18th 2018].

[4] Yannis G, Dragomanovits A, Laiou A, Richter T, Ruhl S, La Torre F, Domenichini L, Graham D, Karathodorou N, and Li H. Use of Accident Prediction Models in Road Safety Management An International Inquiry. *Transportation Research Procedia.* 2016;14: 4257-4266. Available from: doi:10.1016/j.trpro.2016.05.397 [Accessed June 18th 2018].

[5] Toronto Police Service. *Vision Zero Plan Overview*. Available from: https://www.toronto.ca/services-payments/streets-parking-transportation/road-safety/vision-zero/vision-zero-plan-overview/ [Accessed June 2nd 2018].

[6] Vision Zero Network. *European Cities Lead the Way Toward Vision Zero*. Available from: https://visionzeronetwork.org/european-cities-lead-the-way-toward-vision-zero/ [Accessed June 2nd 2018].

[7] Edmonton Traffic Safety: Vision Zero. *About Vision Zero*. Available from: https://www.edmonton.ca/transportation/traffic_safety/vision-zero.aspx [Accessed June 2nd 2018].

[8] Amit D, Arason N, Mussell L, and Woolsey D. *Moving to Vision Zero: Road Safety Strategy Update and Showcase of Innovation in British Columbia*. Ministry of Public Safety and Solicitor General RoadSafetyBC. 2016. Available from: https://www2.gov.bc.ca/assets/gov/driving-and-transportation/driving/publications/road-safety-strategy-update-vision-zero.pdf [Accessed June 2nd 2018].

[9] Minutes of Ottawa Transportation Committee, July 5, 2017. Available from: http://app05.ottawa.ca/sirepub/mtgviewer.aspx?meetid=6997&doctype=SUMMARY [Accessed: August 1st 2018].

[10] City of Toronto. *Toronto's Road Safety Plan Vision Zero*. Available from: https://www.toronto.ca/wp-content/uploads/2017/11/990f-2017-Vision-Zero-Road-Safety-Plan_June1.pdf [Accessed June 2nd 2018].

[11] Raslavičius L, Bazaras L, Keršys R. Accident Reconstruction and Assessment of Cyclist's Injuries Sustained in Car-to-bicycle Collision. *Procedia Engineering*. 2017;187: 562-569. Available from: doi:10.1016/j.proeng.2017.04.415 [Accessed June 13th 2018].

[12] Shi L, Han Y, Huang H, Li Q, Wang B, Mizuno K. Analysis of pedestrian-to-ground impact injury risk in vehicle-to-pedestrian collisions based on rotation angles. *Journal of Safety Research*. 2018;64: 37-47. Available from: doi:10.1016/j.jsr.2017.12.004 [Accessed June 13th 2018].

[13] Sun Y, Zhou X, Cuiping J, Yan C, Huang M, Xiang H. Childhood injuries from motor vehicle-pedestrian collisions in Wuhan, The People's Republic of China. *International Journal of the Care of the Injured.* 2006;37: 416-422. Available from: doi:10.1016/j.injury.2005.12.002 [Accessed June 2nd 2018].

[14] Skjerven-Martinsen M, Aksel Naess P, Bond Hansen T, Gaarder C, Lereim I, Stray-Pedersen A. A prospective

study of children aged <16 years in motor vehicle collisions in Norway: Severe injuries are observed predominantly in older children and are associated with restraint misuse. *Accident Analysis & Prevention.* 2014;73: 151-162. Available from: doi:10.1016/j.aap.2014.09.004 [Accessed June 2nd 2018].

[15] Bedard M, Guyatt G-H, Stones M-J, Hirdes J-P. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention.* 2002;34(6): 717-727. Available from: doi:10.1016/S0001-4575(01)00072-0 [Accessed 18th June 2018].

[16] Wickens CM, Mann RE, Ialomiteanu AR, Stoduto G. Do driver anger and aggression contribute to the odds of a crash? A population-level analysis. *Transportation Research Part F: Traffic Psychology and Behaviour.* 2016;42: 389-399. Available from: doi:10.1016/j.trf.2016.03.003 [Accessed June 2nd 2018].

[17] Olutayo V-A, Eludire A-A. Traffic Accident Analysis Using Decision Trees and Neural Networks. *I.J. Information Technology and Computer Science*. 2014;6(2): 22-28. Available from: 10.5815/ijitcs.2014.02.03 [Accessed 2nd June 2018].

[18] Bener A, Yildirim E, Özkan T, Lajunen T. Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population-based case and control study. *Journal of Traffic and Transportation Engineering (English Edition).* 2017;4(5): 496-502. Available from: doi:10.1016/j.jtte.2017.07.005 [Accessed August 3rd 2018].

[19] Abellán J, López G, Oña J. Analysis of traffic accident severity using Decision Rules via Decision Trees. *Experts Systems with Applications*. 2013;40(15): 6047-6054. Available from: doi:10.1016/j.eswa.2013.05.027 [Accessed June 2nd 2018].

[20] Xiong X, Chen L, Liang J. Analysis of roadway traffic accidents based on rough sets and Bayesian networks. *Promet – Traffic&Transportation*. 2018;30(1): 71-81. Available from: doi:10.7307/ptt.v30i1.2502 [Accessed 22nd June 2018].

[21] Yu B, Wang YT, Yao JB, Wang JY. A Comparison of the Performance of ANN and SVM for the Prediction of Traffic Accident Duration. *International Journal on Non-Standard Computing and Artificial Intelligence.* 2016;26: 271-287. Available from: doi:10.14311/NNW.2016.26.015 [Accessed 2nd June 2018].

[22] Bülbül HI, Kaya T, Tulgar Y. Analysis for Status of the Road Accident Occurrence and Determination of the Risk of Accident by Machine Learning in Istanbul. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 18-20 December 2016, Anaheim, CA, USA*. IEEE; 2016. p. 426-430.

[23] Tay R. Comparison of the binary logistic and skewed logistic (Scobit) models if injury severity in motor vehicle collisions. *Accident Analysis & Prevention*. 2016;88: 52-55. Available from: doi:10.1016/j.aap.2015.12.009 [Accessed June 2nd 2018].

[24] Taamneh M, Alkheder S, Taamneh S. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. *Journal of Transportation Safety & Security.* 2017;9(2): 146-166. Available from: doi:10.1080/19439962.2016.1152338 [Accessed June 2nd 2018].

[25] Tambouratzis T, Dora S, Miltiadis C, Andreas G. Maximising Accuracy and Efficiency of Traffic Accident Prediction Combining Information Mining with Computational Intelligence Approaches and Decision Trees. *Journal of Artificial Intelligence and Soft Computing Research.* 2014;4(1): 31-42. Available from: doi:10.2478/jaiscr-2014-0023 [Accessed 18th June 2017] .

[26] Oña J, Mujalli RO, Calvo FJ. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention.* 2011;43(1): 402-411. Available from: doi:10.1016/j.aap.2010.09.010 [Accessed 2nd June 2018].

[27] Böehmlaender D, Hasirlioglu S, Yano V. Advantages in Crash Severity Prediction Using Vehicle to Vehicle Communication. *015 IEEE International Conference on Dependable Systems and Networks Workshops, Rio de Janeiro, 2015*. p. 112-117. Available from: doi:10.1109/DSN-W.2015.23 [Accessed: June 2nd 2018].

[28] Xu C, Wang W, Liu P. A Genetic Programming Model for Real-Time Crash Prediction on Freeways. *IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 2*; June 2013. p. 574-586.

[29] Abdel-Aty M, Abdelwahab H. Analysis and prediction of traffic fatalities resulting from angle collisions including the effect of vehicles configuration and compatibility. *Accident Analysis & Prevention.* 2004;36(3): 457-469. Available from: doi:10.1016/S0001-4575(03)00041-1 [Accessed 18th June 2018].

[30] Chen Z, Gao Z, Yu R, Wang M, Sun P. Macro-level accident fatality prediction using a combined model based on ARIMA and multivariable linear regression. *Proceedings of the 2016 International Conference on Progress in Informatics and Computing (PIC), 23 – 25 December 2016, Shanghai, China.* IEEE Xplore; 2017. p. 133-137.

[31] Chen F, Chen S, Ma X. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. *Journal of Safety Research.* 2018 Jun 1;65: 153-9. Available from: doi:10.1016/j.jsr.2018.02.010 [Accessed June 5th 2019].

[32] Chen F, Chen S. Injury severities of truck drivers in single-and multi-vehicle accidents on rural highways. *Accident Analysis & Prevention*. 2011 Sep 1;43(5): 1677-88. Available from: doi:10.1016/j.aap.2011.03.026 [Accessed June 5th 2019].

[33] Ma X, Chen S, Chen F. Multivariate space-time modeling of crash frequencies by injury severity levels. *Analytic Methods in Accident Research*. 2017 Sep 1;15: 29-40. Available from: doi:10.1016/j.amar.2017.06.001 [Accessed June 5th 2019].

[34] Roberts SE, Vingilis E, Wilk P, Seeley J. A comparison of self-reported motor vehicle collision injuries compared with official collision data: an analysis of age and sex trends using the Canadian National Population Health Survey and Transport Canada data. *Accident Analysis & Prevention.* 2008;40(2): 559-566. Available from: doi:10.1016/j.aap.2007.08.017 [Accessed June 2nd 2018].

[35] Nhan C, Rothman L, Staler M, Howard A. Back-over Collisions in Child Pedestrians from the Canadian Hospitals Injury Reporting and Prevention Program. *Traffic*

Injury Prevention. 2009;10(4): 350-353. Available from: doi:10.1080/15389580902995166 [Accessed June 13th 2018].

[36] Cui G, Wang X, Kwon DW. A framework of boundary collision data aggregation into neighbourhoods. *Accident Analysis & Prevention*. 2015;83: 1-17. Available from: https://doi.org/10.1016/j.aap.2015.06.003 [Accessed June 2nd 2018]

[37] Aljeri N, Boukerche A. A predictive collision detection protocol using vehicular networks. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC*; 2017. p. 1-5.

[38] Zhao P, Lee C. Assessing rear-end collision risk of cars and heavy vehicles on freeways using a surrogate safety measure. *Accident Analysis & Prevention*. 2018;113: 149-158. Available from: doi:10.1016/j.aap.2018.01.033 [Accessed August 2nd 2018].

[39] Grisé E, Buliung R, Rothman L, Howard A. A geography of child and elderly pedestrian injury in the City of Toronto, Canada. *Journal of Transport Geography*. 2018;66: 321-329. Available from: doi:10.1016/j.jtrangeo.2017.10.003 [Accessed June 2nd 2018].

[40] Teschke K, Frendo T, Shen H, Harris MA, Reynolds CC, Cripton AP, Brubacher J, Cusimano MD, Friedman SM, Hunte G, Monro M, Vernich L, Babul S, Chipman M, Winters M. Bicycling crash circumstances vary by route type: a cross-sectional analysis. *BMC Public Health.* 2014;14: 1205. Available from: doi:10.1186/1471-2458-14-1205 [Accessed June 2nd 2018].

[41] Toronto Police Service. *Public Safety Data Portal*. Available from: http://data.torontopolice.on.ca/datasets/ksi

[42] Ethem Alpaydin. *Introduction to Machine Learning.* 3rd ed. Cambridge, Massachusetts: MIT Press; 2014.

[43] Savolainen PT, Mannering FL, Lord D, Quddus MA. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention.* 2011 Sep 1;43(5): 1666-76. Available from: doi:10.1016/j.aap.2011.03.025 [Accessed June 5th 2019].

[44] Winston C, Maheshri V, Mannering F. An exploration of the offset hypothesis using disaggregate data: The case of airbags and antilock brakes. *Journal of Risk and Uncertainty.* 2006 Mar 1;32(2): 83-99. Available from: doi:10.1007/s11166-006-8288-7 [Accessed June 6th 2019].

[45] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16: 321-57. Available from: doi:10.1613/jair.953 [Accessed June 3rd 2018] .

[46] Breiman L. Random forests. *Machine learning*. 2001 Oct 1;45(1): 5-32. Available from: doi:10.1023/A:1010933404324 [Accessed June 1st 2018].

[47] Toronto Police Service. *Public Safety Data Portal: KSI Glossary*. Available from: http://data.torontopolice.on.ca/pages/ksi [Accessed June 11th 2018].