Predicting crash injury severity at unsignalized intersections using support vector machines and naïve Bayes classifiers

Stephen A. Arhin* and Adam Gatiba

Howard University Transportation Research Center, 2300 6th Street NW, Washington, DC 20059, United States

*Corresponding author. E-mail: saarhin@howard.edu
Stephen A. Arhin, http://orcid.org/0000-0002-1306-4017

Abstract

The Washington, DC crash statistic report for the period from 2013 to 2015 shows that the city recorded about 41,789 crashes at unsignalized intersections, which resulted in 14,168 injuries and 51 fatalities. The economic cost of these fatalities has been estimated to be in the millions of dollars. It is therefore necessary to investigate the predictability of the occurrence of these crashes, based on pertinent factors, in order to provide mitigating measures. This research focused on the development of models to predict the injury severity of crashes using support vector machines (SVMs) and Gaussian naïve Bayes classifiers (GNBCs). The models were developed based on 3307 crashes that occurred from 2008 to 2015. Eight SVM models and a GNBC model were developed. The most accurate model was the SVM with a radial basis kernel function. This model predicted the severity of an injury sustained in a crash with an accuracy of approximately 83.2%. The GNBC produced the worst-performing model with an accuracy of 48.5%. These models will enable transport officials to identify crash-prone unsignalized intersections to provide the necessary countermeasures beforehand.

Keywords: crashes; unsignalized intersection; support vector machines; Gaussian naïve Bayes classifier; injury severity

1. Introduction

The US Federal Highway Administration has reported that more than 50% of fatal and injury crashes occur at or near intersections, with angle crashes at unsignalized intersections having a higher tendency of resulting in fatalities. In order to mitigate the occurrence of these crashes, it is...
necessary to investigate their predictability based on pertinent factors and circumstances that may have contributed to their occurrence. Unsignalized intersections are known to be crash hotspot areas. About 71.4% of intersection-related fatal crashes occurred at unsignalized intersections. In addition, angle crashes accounted for about 53% of these crashes. The remaining crash types were rear-end (6%), sideswipe (2%) and pedestrian/bicycle (14%). Other crash types accounted for the remaining 17%. Furthermore, about one-third of all fatal crashes in the United States involved only two vehicles [1]. These statistics show that angle crashes at unsignalized intersections have a high tendency of resulting in fatalities. The crashes could be attributed to several reasons, including intersection geometry limitations, driver behaviour and environmental/weather conditions, among others. It is therefore necessary that the causes and predictability of these crashes be explored. There are several strategies that could be used to predict crashes, including artificial neural networks and ensembles. These strategies have been implored in several previous studies. However, this study explores the development of models to predict the injury severity of crashes using support vector machines (SVMs) and Gaussian naive Bayes classifiers (GNBCs).

2. Literature review

2.1 Summary of crash statistics

In 2017, the United States recorded 34,247 fatal crashes involving 52,643 vehicles and 84,921 persons. Of the total number of persons involved, 37,113 were fatally injured. Twenty-five per cent (8371) of these crashes occurred at or near intersections, while 25,748 fatal crashes occurred outside of intersections. Of the intersection-related fatal crashes, 53,342 occurred at four-legged intersections, 2787 at T-intersections, 219 at Y-intersections and 150 at other intersection types. Furthermore, crashes involving only two vehicles constituted the highest proportion of crashes recorded. Thus, 12,165 fatal crashes involving 36,332 persons resulted from the collision of only two vehicles [1].

The Washington, DC crash statistic report for the period 2013 to 2015 reports a total of 41,789 crashes at unsignalized intersections, which resulted in 14,168 injuries and 51 fatalities. Collisions at STOP-controlled intersections constituted 10.1% (4241) of the crashes, while crashes at intersections controlled by yield signs, ‘traffic control officers’ and ‘no control signs’ constituted the remaining 89.9% of the reported crashes [2].

2.2 Contributory factors of intersection-related crashes

In 2008, the National Highway Traffic Safety Administration submitted a report to the United States Congress that documented the results of a three-year national survey (from 2005 to 2008) that investigated the causes of intersection-related crashes [3]. The research team collaborated with local law enforcement and emergency responders who were granted timely access to crash scenes. Researchers had access to on-the-scene information regarding factors leading to each crash, as well as the opportunity to interview persons directly involved in the crash and witnesses. Most significantly, information from vehicles’ event-data recorders was available for download. Analyses of the crash events showed that 96% of the 787,236 crashes that were investigated were attributable to drivers, while about 3% were due to vehicle or environmental-related factors. Specific driver behaviours that led to crashes were identified as illegal manoeuvres, inattention, turning with obstructed view, misjudgement of gaps or speed and false assumptions about the actions of others. Also, the results revealed that in 36% of the crashes, at least one of the vehicles was turning left, crossing over or turning right at the intersection.

Crashes caused by intersection geometry have also been well studied and documented in previous research. Harwood concluded that the provision of a right turn-lane on one major approach to a stop-controlled intersection resulted in a reduction in the number of crashes at that intersection of approximately 5% over a period of three years. Furthermore, intersections with one lane per approach were determined to have higher right-angle crash rates than intersections with two or more lanes per approach [4].

Lighting conditions have also been found to contribute to the frequency of crashes at intersections. Yannis, Kondyli and Mitzalis found that night-time lighting had great potential to improve traffic safety and reduce crash severity [5].

Traffic characteristics such as speed and volume also contribute to the occurrence of crashes at intersections. Kim, Washington and Oh developed crash-prediction models for different types
of crashes at rural intersections. The findings showed that traffic volume variables affected the safety of two-lane intersections. An increase in AADT at the intersections resulted in an increase in exposure to risk of a crash [6].

Besides traffic and geometric characteristics, intersection-related crashes have been determined to be significantly influenced by weather conditions. Hermans, Brijs and Stiers investigated the impact of wind, temperature, sunshine, precipitation, weather image and visibility on the hourly number of crashes in the Netherlands. It was concluded from the study that increased wind speeds resulted in an increased number of crashes. Also, road safety was more affected by the duration of precipitation than the amount of precipitation. Global radiation and sunshine duration were determined to have a significant negative impact on road safety [7].

### 2.3 Crash-prediction models

Several modelling techniques have been employed to predict crashes at intersections. These include linear regression models, generalized linear regression models (such as negative binomial regression models and ordered probit models) and machine-learning models. The following sections present a discussion of these models.

#### 2.3.1 Linear regression models

Linear regression modelling is an approach to establishing a relationship between scalar response (also called dependent) variables and other explanatory (or independent) variables. Model parameters are estimated using a data set of values of the response and explanatory variables. The model is usually fitted to the observed data set using the least square approach. Linear regression models take the form:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{in} + \epsilon_i \tag{1}
\]

where \(y_i\) is the \(i\)th dependent variable, \(\beta_1, \beta_2, \ldots, \beta_p\) are estimated parameters, \(x_{i1}, x_{i2}, \ldots, x_{in}\) are the predictor variables of the \(i\)th dependent variable and \(\epsilon_i\) is the error term. The error term is an independent and normally distributed random variable with a mean of zero and a variance greater than zero. Linear regression modelling has been applied in several studies to establish various relationships between the frequency of injury crashes and other traffic characteristics. Lau and May conducted a study to investigate the relationship between the number of injury or property damage only (PDO) crashes that occur annually at intersections and traffic and environmental factors [8]. The crash records (ranging from 1984 to 1987) of 2488 intersections in California were sampled. The linear regression analyses employed in this study were conducted on two levels. On the first level, a simple linear regression model was developed with injury/PDO crashes per year as the response variable and traffic intensity, expressed in millions of vehicles entering the intersection per year from all approaches, as the predictor variable. In the second model, additional information, such as design, traffic control, proportion of cross-street traffic and environmental features of the intersections, were included as predictor variables. The results of the analysis showed that the accuracy of the model improved as more predictor variables were added.

Though linear regression models are easy to use and interpret, it has been shown that they are not ideal for crash prediction. Crashes are usually sporadic and random in nature and hence are not best fitted by linear relationships. Also, the assumption that the error term is normally distributed is not accurate for crash predictions, which are usually discrete and non-negative. Moreover, some factors have been determined to strongly correlate with each other, thus introducing multicollinearity, thereby invalidating such linear models [9]. To overcome the shortcomings of linear regression models, generalized linear models (GLMs) have been used to model crashes at intersections. GLMs are flexible generalizations of ordinary linear regression that can accommodate the non-normal distributed error terms. The most common forms of GLM used in crash-prediction models are the negative (NB) model and the ordered probit model (OPM).

#### 2.3.2 Negative binomial model

NB models are a generalization of Poisson regression. Unlike in Poisson models, where the variance of the distribution of the response variables is equal to its mean, in NB models, the variance differs from the mean. NB models have been found to be suitable for crash prediction due to the nature of the dependent variables in such analysis. Usually, the response required is the number of crashes experienced at a specific location. Such responses are non-negative integers and generally follow the NB distribution. The distribution is given by the
following Poisson–Gamma distribution:
\[
Pr(Y = y_i | u, \alpha) = \frac{r(\gamma + 1)}{r(\alpha^{-1})} \left( \frac{1}{1 + \alpha e^{\alpha \beta}} \right)^{y_i} \times \left( \frac{1}{1 + \alpha e^{\alpha \beta}} \right)^{\gamma - 1}
\]

where \(u\) is the mean of the dependent variable \(y\), \(\beta\) is an estimated parameter to be estimated, \(\alpha\) is the heterogeneity parameter and \(x_i\) is the \(i\)th predictor variable. Ackaah and Salifu also developed a crash-prediction model using NB distribution using crash data from 1996 to 1998 for 91 unsignalized intersections in two cosmopolitan cities in Ghana. The explanatory power of the model showed that the model was a good fit. Though the model had a low log-likelihood ratio value, it was still reported that it compared favourably with what has been widely reported in the literature. Also, the model met the overall goodness-of-fit test [10].

2.3.3 Ordered probit models. OPMs are used in developing models that have an ordered response. This approach in modelling data employs the probit link function. The latent continuous metric underlying the ordinal responses observed are partitioned into a series of regions corresponding to the ordinal categories. Generally, the probability of obtaining a particular outcome is given by the following:
\[
Pr(y_i = j | X_i) = \frac{\exp(r_j - X_i \beta)}{1 + \exp(r_j - X_i \beta)} - \frac{\exp(r_{j-1} - X_i \beta)}{1 + \exp(r_{j-1} - X_i \beta)}
\]

where \(y_i\) is an observable ordinal variable, \(X_i\) is a vector of exogenous variables, \(\beta\) is a vector of unknown parameters to be estimated and \(r_j\) is the threshold associated with the \(j\)th ordinal partition interval, which are assumed to be in ascending order. OPMs have been applied in the development of several crash-prediction models that seek to predict injury severity based on several factors. Rifaat and Chin investigated the various significant factors that determine the injury severity of two-vehicle crashes, single-vehicle crashes and pedestrian crashes in Singapore using an OPM. The order of severity used in the analysis was fatal, seriously injured and slightly injured. The study found that factors such as vehicle type, road type, collision type, location type, pedestrian age and time of day of accident were significantly associated with injury severity. The overall model was determined to have goodness of fit \(\rho^2\) of 0.0322 [11].

2.3.4 Support vector machines. SVMs are supervised learning models that utilize associated learning algorithms for classification and regression analysis. They are non-probabilistic binary linear classifiers. However, SVMs can perform non-linear classification using the kernel approach, which implicitly maps input data into a high-dimensional feature space. SVMs generally classify data by constructing a hyperplane that separates the data into two sets. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class, since in general, the larger the margin, the lower the generalization error of the classifier. The set of data points closest to the hyperplane are referred to as the support vectors.

Li, Lord and Zhang evaluated the application of SVMs for predicting vehicle crashes and compared the evaluated models with a negative binomial regression model. The results showed that SVMs predicted crash data more effectively and accurately than conventional models [12].

2.3.5 Naïve Bayes classifiers. A naïve Bayesian classifier (NBC) is a simple probabilistic classifier based on applying Bayes’ theorem with naïve independence assumptions. It is one of the methods used for supervised learning. It provides an efficient way of handling any number of attributes or classes that is based purely on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge for observed data. Khera analysed crashes to develop a model to predict injury severity using an NBC. The results showed that the NBC performed better than the random tree classifier, with approximately 92% accuracy [13].

3. Methodology

3.1 Description of study jurisdiction

This study is based on data obtained in the District of Columbia (DC). The capital city of the United States, Washington, DC, is divided into four (equal) quadrant areas—Northwest (NW), Northeast (NE), Southeast (SE) and Southwest (SW)—which are further divided into eight wards. As of July 2018,
the population of DC was about 702,455, with a growth rate of approximately 1.41% [14]. The city is highly urbanized and is ranked the sixth most congested city in the United States, with each driver spending an average of 63 hours in traffic annually [15]. It has a total of 1503 miles of roadway comprising local roads, collector roads, minor arterials, principal arterials, freeways and interstates [16]. Also, the city has about 7700 intersections, of which 1450 are signalized [17]. The American Society of Civil Engineers’ 2017 infrastructure report card reported that about 95% of the roads in DC are in poor condition [18].

3.2 The crash database system
Crash-prediction models are data-dependent and as a result, the accuracy of the models developed depends largely on the quality of available crash data. To ensure that a reliable model was developed, this study utilized traffic crash data from the District Department of Transportation’s (DDOT) crash database, the Traffic Accident Reporting and Analysis Systems Version 2.0 (TARAS2). The District of Columbia Metropolitan Police Department (MPD) records traffic crash information at the scene of crashes electronically on a PD-10 crash-reporting form. The crash data is then downloaded through secure servers from the MPD into the DDOT database and is then processed and made available in TARAS2, which is an Oracle-based application. TARAS2 contains data fields that can broadly be categorized under vehicle characteristics, environmental conditions, roadway characteristics, traffic exposure characteristics, crash location, date, time, crash type, crash severity and information on persons involved.

3.3 Data extraction and encoding
Nine years of crash data (2008 to 2015) was queried and extracted from TARAS2. The data was then filtered to obtain angle crashes involving two vehicles at unsignalized intersections. The extracted data was then cleaned by identifying and removing duplicate and incomplete crash records as well as irrelevant data fields. In all, 3307 data points were extracted and used for the analyses. The extracted data set contained the following fields: accident complaint number, main street name, side street name, year of accident, month of accident, time of accident, day of week, quadrant of accident occurrence, type of collision, road surface conditions, street lighting conditions, lighting conditions, weather conditions, traffic conditions, traffic control type, age of drivers, gender of drivers, contributing circumstances and injury severity. Only numerical data can be analysed using SVMs and GNBCs. Hence, qualitative data needed to be converted to quantitative data. Thus, both input and output data had to be encoded into either real or integer values. Secondly, the binary method (0 and 1) of encoding has been determined to yield better results, since it minimizes the loss function values with respect to the model parameters. The loss value determines how poorly or well the model fits the data set. The lower the loss function value, the better the model fits the data set. Table 1 presents the variables and coding scheme used in this study.

3.4 Injury severity
The outcome variable describes the degree of injury severity sustained by persons involved in a crash. The crash database specifies five degrees of injury severity: no injury, complain, non-disabling injury, disabling injury and fatal. Due to the insignificant percentage of fatal and disabling injury crashes in the data set, all complain, injury and fatal crashes were categorized as injury crashes. Thus, two levels of crash injury severity were created and used in the analysis, as presented in Table 2.

3.5 Development of models
Though all machine-learning techniques have unique procedures and algorithms used in developing models for classification of data, generally their development follows a generic procedure. The following steps constitute the generic procedure of machine-learning techniques:

(i) Selection of training and testing data sets
(ii) Training of model and optimization of parameters
(iii) Validation of model using testing data set

Selection of training data and testing set: At this stage of model development, the data is divided into training and testing sets. Each data set is a matrix of the independent and dependent variables. The training set is used to train the model, while the testing set is used to evaluate the performance of the model after training. The percentage of the data set designated the testing set usually ranges from 10% to 35%, with the data randomly
### Table 1: Variables and coding scheme used in this study

| Variable | Variable name                                | Code                              | Variable | Variable name                                | Code                              |
|----------|----------------------------------------------|-----------------------------------|----------|----------------------------------------------|-----------------------------------|
| X1       | Day of crash                                 | Monday 1: present, 0: otherwise   | X16      | Lighting conditions                          | Dark 1: present, 0: otherwise     |
| X2       |                                              | Tuesday 1: present, 0: otherwise  | X17      |                                              | Dark (lighted) 1: present, 0: otherwise |
| X3       |                                              | Wednesday 1: present, 0: otherwise| X18      |                                              | Daylight 1: present, 0: otherwise  |
| X4       |                                              | Thursday 1: present, 0: otherwise | X19      |                                              |                                   |
| X5       |                                              | Friday 1: present, 0: otherwise   | X20      |                                              |                                   |
| X6       |                                              | Saturday 1: present, 0: otherwise | X21      |                                              |                                   |
| X7       |                                              | Sunday 1: present, 0: otherwise   | X22      | Time of day                                  | AM peak (06:00–10:00) 1: present, 0: otherwise |
|          |                                              |                                   |          |                                              | Off-peak (10:00–15:00) 1: present, 0: otherwise |
|          |                                              |                                   |          |                                              | PM peak (15:00–19:00) 1: present, 0: otherwise |
|          |                                              |                                   |          |                                              | Evening (19:00–00:00) 1: present, 0: otherwise |
|          |                                              |                                   |          |                                              | Night (0000–0600) 1: present, 0: otherwise |
|          |                                              |                                   |          |                                              |                                   |
| X33      | Quadrant                                     | NW 1: present, 0: otherwise       | X34      |                                              | SW 1: present, 0: otherwise       |
| X35      |                                              | NE 1: present, 0: otherwise       | X36      |                                              | NE 1: present, 0: otherwise       |
| X37      |                                              | SE 1: present, 0: otherwise       | X38      |                                              | SE 1: present, 0: otherwise       |
| X13      | Collision type                               | Right-angle 1: present, 0: otherwise| X18      |                                              | Distracted D2 1: present, 0: otherwise |
| X19      |                                              | Left-turn 1: present, 0: otherwise| X19      |                                              | Distinguished D2 1: present, 0: otherwise |
| X20      |                                              | Right-turn 1: present, 0: otherwise| X20      |                                              | Distinguished D2 1: present, 0: otherwise |
|          | Road surface conditions                      |                                   |          |                                              |                                   |
| X21      |                                              | Wet 1: present, 0: otherwise      | X21      |                                              | Wet 1: present, 0: otherwise      |
| X22      |                                              | Dry 1: present, 0: otherwise      | X22      |                                              | Dry 1: present, 0: otherwise      |
| X23      |                                              | Lights off 1: present, 0: otherwise| X23      |                                              | Lights off 1: present, 0: otherwise |
| X24      |                                              | Lights on 1: present, 0: otherwise| X24      |                                              | Lights on 1: present, 0: otherwise |
| X25      |                                              | None 1: present, 0: otherwise     | X25      |                                              | None 1: present, 0: otherwise     |
| X26      |                                              |                                   | X26      |                                              |                                   |
| X27      |                                              |                                   | X27      |                                              |                                   |
| X28      |                                              |                                   | X28      |                                              |                                   |
| X29      |                                              |                                   | X29      |                                              |                                   |
| X30      |                                              |                                   | X30      |                                              |                                   |
| X31      |                                              |                                   | X31      |                                              |                                   |
| X32      |                                              |                                   | X32      |                                              |                                   |
| X33      |                                              |                                   | X33      |                                              |                                   |
| X34      |                                              |                                   | X34      |                                              |                                   |
| X35      |                                              |                                   | X35      |                                              |                                   |
| X36      |                                              |                                   | X36      |                                              |                                   |
| X37      |                                              |                                   | X37      |                                              |                                   |
| X38      |                                              |                                   | X38      |                                              |                                   |
| X39      |                                              |                                   | X39      |                                              |                                   |
| X40      |                                              |                                   | X40      |                                              |                                   |
| X41      |                                              |                                   | X41      |                                              |                                   |
| X42      |                                              |                                   | X42      |                                              |                                   |
| X43      |                                              |                                   | X43      |                                              |                                   |
| X44      |                                              |                                   | X44      |                                              |                                   |
| X45      |                                              |                                   | X45      |                                              |                                   |
| X46      |                                              |                                   | X46      |                                              |                                   |
| X47      |                                              |                                   | X47      |                                              |                                   |
| X48      |                                              |                                   | X48      |                                              |                                   |
| X49      |                                              |                                   | X49      |                                              |                                   |
| X50      |                                              |                                   | X50      |                                              |                                   |
| X51      |                                              |                                   | X51      |                                              |                                   |

### Table 2: Injury severities

| Injury severity | Level       |
|-----------------|-------------|
| No injury       | Non-injury  |
| Complain        | Injury      |
| Non-disabling   |            |
| Disabling       |            |
| Fatal           |            |

Selected from the dataset. In this study, 30% of the data was used for testing.

**Training of model and optimization of parameters:** The model is initially fitted on the training data set using a supervised learning method. This is accomplished by presenting the training algorithm with the training data set. The outputs after each iteration are then compared with the desired outputs (vector of the dependent variables). Based on the results of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted and a new set of output are produced. This iterative process continues until the least error (the difference between the calculated output and the targeted output) is achieved. The specific training algorithms for the machine-learning techniques used in this study are discussed in the next section.

**Validation of model with testing data set:** The accuracy of the model attained after training is validated using the testing data set. The matrix of independent variables of the testing data set are used as input in the trained model. The predicted classifications are compared with the actual classifications. The performance of the model is then assessed using a confusion matrix. Fig. 1 presents a schematic diagram of the machine-learning process.

### 3.6 Support vector machines

SVMs classify data by determining a hyperplane that maximizes the margin between the classes.
of data. The hyperplane is defined by data points that are closest to the decision surface. These data points are referred to as the support vectors. SVMs perform classification through an iterative process that requires tuning of parameters until the best model is obtained. The training data set is generally of the form

\[ \{ \vec{x}_k, \vec{y} \}, k = 1, \ldots, n \]  

where \( \vec{x}_k \) is a real vector of the \( k \)th independent variable, \( n \) is the number of dependent variables and \( \vec{y} \) is a vector of the dependent variables. The vector \( \vec{y} \) is a binary variable with values of -1 and 1, corresponding to the results of each observation. Thus, \( y = 1 \) when a crash results in an injury and \( y = -1 \) when no injury occurs. Given that the training set is linearly separable after being mapped into higher-dimensional feature space by non-linear function, the hyperplane, \( H_0 \), separating the two classes of data can be expressed as

\[ \vec{w} \cdot \vec{x} + b = 0 \] (5)

where \( \vec{w} \) is the normal vector to the hyperplane, \( H_0 \). Also, the two parallel hyperplanes, \( H_1 \) and \( H_2 \), which maximize the region of separation between the two classes of data, are constructed as

\[ H_1 : \vec{w} \cdot \vec{x} + b \geq 1 \]  
\[ H_2 : \vec{w} \cdot \vec{x} + b \leq -1 \] (6) (7)

The distance between \( H_1 \) and \( H_2 \) is given as

\[ \frac{2}{||\vec{w}||} \] (8)

Thus, the goal of the classifier is to maximize the margin distance by minimizing \( ||\vec{w}|| \). The classification problem is actually an optimization problem, which can be formulated as

Minimize \( ||\vec{w}|| \).

Subjectto \( y(\vec{w} \cdot \vec{x}_k - b) \geq 1 \), for \( k = 1, \ldots, n \) (9)

However, since it is usually difficult to obtain a hyperplane that perfectly separates the two classes of data, a soft-margin SVM is introduced, which reduces the effect of misclassification. The soft-margin SVM employs a parameter, \( \lambda \), to control the cost of misclassification. The optimization problem is thus reconstructed, taking into consideration the cost control parameter, \( \lambda \), and is
where \( K(x_i,x_j) \) is a kernel function used in this study.

### Table 3: SVM kernel functions

| Kernel type       | Kernel function            |
|-------------------|----------------------------|
| Linear            | \( K(x_i,x_j) = x_i^T x_j \) |
| Polynomial        | \( K(x_i,x_j) = (\gamma x_i^T x_j + d)^\alpha \) |
| Radial basis      | \( K(x_i,x_j) = \exp(-\gamma |x_i-x_j|^2) \) |
| Sigmoid           | \( K(x_i,x_j) = \tanh(\gamma x_i^T x_j + r) \) |

where \( \gamma, \alpha, d \) and \( r \) are kernel parameters expressed as

\[
\sum_{i=1}^{n} \max(0, 1 - y(\vec{w} \cdot \vec{x}_i - b)) + \lambda ||\vec{w}||^2
\tag{10}
\]

Using LaGrange multipliers, the dual of the optimization problem is expressed as

\[
\min L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i,x_j)
\tag{11}
\]

subject to \( \sum_{i=1}^{n} y \alpha_i = 0 \), and \( 0 \leq \alpha_i \leq \frac{1}{2\lambda} \) for all \( i \)

where \( K(x_i,x_j) \equiv \phi(x_i)^T \phi(x_j) \) is the kernel function. The kernel function maps the data into a space where a linear hyperplane cannot separate the two classes of data. This allows a non-linear function to be learned by a linear learning machine in a higher-dimensional space. After solving the dual problem and substituting \( w = \sum_{i=1}^{n} y \alpha_i \phi(x_i) \) into the original classification problem, the following classifier is obtained:

\[
y(x) = \text{sign} (w^T \phi(x) - b)
\]

\[
y(x) = \text{sign} \left( \sum_{k=1}^{n} \alpha_i y_i K(\vec{x}_i, \vec{x}_j) - b \right)
\tag{12}
\]

#### 3.6.1 Parameter tuning.
By tuning the kernel function in the classifier, the performance of the model can be improved. This study employs the use of four different kernel functions to improve the accuracy of the model. Table 3 shows a list of the kernel functions used in this study.

### 3.7 Gaussian Naïve Bayes classifier

A naïve Bayesian classifier is a probabilistic classifier based on Bayes’ theorem with strong (naïve) independence assumptions. The classifier predicts the probability that a given data point belongs to a class. The class with the highest probability is considered the most likely class of the data point. The predicted probabilities are the posteriori probabilities (conditional probabilities) that an observation belongs to a class.

Given a training data set \( X = \{\vec{x}_1, \ldots \vec{x}_n, \vec{y}\}, i = 1, \ldots, n \) and \( \vec{x}_i \in \mathbb{R}^n \)

where \( \vec{x}_i \) is a real vector of the \( i \)th independent variable, \( n \) is the number of independent variables and \( \vec{y} \) is a vector of the dependent variable. The vector \( \vec{y} \) is a binary variable with values of 1 and 0, corresponding to the results of each observation. Thus, \( y = 1 \) when a crash results in an injury and \( y = 0 \) when no injury occurs. The probability that a new data set belongs to a class \( y_k \) is given by

\[
P(y_k|\vec{x}_1, \ldots \vec{x}_n) = \frac{P(x_1|y_k)P(x_2|y_k) \ldots P(x_n|y_k)P(y_k)}{P(x_1)P(x_2) \ldots P(x_n)}
\tag{13}
\]

where \( P(y_k|\vec{x}_1, \ldots \vec{x}_n) \) is the probability that an observation belongs to class given the training data set, \( P(x|y_k) \) is the conditional probability of the \( i \)th independent variable of an observation given \( y \) and \( P(y_k) \) is the class probability.

However, for any given input, \( P(x_1)P(x_2) \ldots P(x_n) \) is constant. Thus, Bayes’ theorem is simplified as

\[
P(y_k|\vec{x}_1, \ldots \vec{x}_n) \propto P(y_k) \prod_{i=1}^{n} P(x_i|y_k)
\tag{14}
\]

Based on Bayes’ theorem above, the classifier model is formulated as

\[
y_k = \arg \max_y \prod_{i=1}^{n} P(x_i|y)
\tag{15}
\]

where the conditional probability \( P(x_i|y) \) is assumed to follow the Gaussian distribution and calculated by the following expression:

\[
P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left( -\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right)
\tag{16}
\]

### 3.8 Model evaluation

The performance of each model was assessed using the testing data set. The results were then evaluated using the data generated by a confusion matrix (CM). A CM contains information about actual and predicted classifications done by a classification system. Each row of the CM represents the instances of an actual class and each column represents the instances of a predicted class. Table 4 shows the confusion matrix for a two-class classifier.

The entries of the CM are defined as follows:

(i) True positive (TP): instances that are positive and correctly classified as positive,
**Table 4: Confusion matrix**

| Total no. of observations | Predicted |
|---------------------------|-----------|
|                           | Negative  | Positive   |
| Actual                    | True      | False      |
|                           | negative  | positive    |
|                           | (TN)      | (FP)       |
| Positive                  | False     | True       |
|                           | negative  | positive    |
|                           | (FN)      | (TP)       |

(iii) True negative (TN): instances that are negative and correctly classified as negative,

(iii) False positive (FP): instances that are negative but wrongly classified as positive, and

(iv) False negative (FN): instances that are positive but wrongly classified as negative.

Based on the CM, the following measures were computed to evaluate the models developed:

(i) Accuracy (AC): the proportion of the total number of predictions that were correctly classified. This is computed as

\[
AC = \frac{TN + TP}{TN + FP + FN + TP}.
\]

(ii) Error rate (ER): the rate at which predictions will be misclassified. This is computed as

\[
ER = 1 - AC
\]

(iii) Sensitivity (S): the proportion of positive cases that were correctly identified, computed as

\[
S = \frac{TP}{FN + TP}
\]

(iv) Precision (P): the proportion of the predicted positive cases that were correct, computed as

\[
P = \frac{TP}{FP + TP}
\]

(v) F-measure (F): a measure of the accuracy of the test model computed using sensitivity and precision. The value of F ranges from 0 to 1, where 1 shows an excellent model and 0 shows a bad model. The F-measure is calculated as

\[
F = \frac{2(S \cdot P)}{(S + P)}
\]

3.9 Analysis software

The models were developed using a high-level general-purpose programming language, Python, with standard and robust libraries for data processing, analysis and machine-learning applications. The NumPy and Pandas libraries were imported to facilitate data preprocessing. Also, the ScikitLearn library, which has SVM and GNBC learning algorithms, was imported for model development. In addition, the descriptive statistics of the data were obtained using IBM’s SPSS Statistics package.

4. Results

4.1 Descriptive statistics

The descriptive statistics of the data set are presented in Tables 5 and 6. The crash frequencies of the categorical variables are presented in Table 5. The highest number of crashes (1252) occurred during the off-peak period (10:00 AM to 3:00 PM), while the lowest number of crashes (176) occurred at night (between 12:00 AM and 6:00 AM). The high crash frequency in the off-peak period could be attributed to the high traffic-volume exposure during that period. Tuesdays, Wednesdays and Thursdays recorded the highest number of crashes. The lowest number of crashes occurred on Sundays. The Northwest quadrant of Washington, DC recorded the highest number of crashes (1167), while the Southeast quadrant recorded the fewest crashes. A right-angle collision was the most frequent occurring crash type.

In addition, most of the crashes occurred under daylight, clear weather and light traffic conditions. Distracted driving and stop/yield sign violation contributed to the occurrence of a high number of crashes. However, most crashes were the result of no violation on the part of one or both drivers. A total of 3936 drivers involved in the crashes were male, while 2678 were female. Furthermore, of the 3307 recorded crashes, 1274 resulted in injury, while the remaining resulted in no injury. Table 5 also shows the rates of injury crashes. It can be observed that the rate of injury crashes was highest during the night period (41.24%), on Fridays (41%) and in the Northeast quadrant (40.44%). Moreover, right-turn collisions had the highest injury-crash rate (40.69%). Crashes that occurred where street lights were absent had the highest rate (39.52%) of injury crashes. In addition, crashes that occurred under rainy weather conditions had the highest rate of injury crashes (50.57%). Light traffic conditions recorded the highest rate (54.78%) of injury crashes. Intersections controlled by yield signs also recorded the highest rate (70.59%) of injury crashes. This is complemented by the fact that the highest rates of injury crashes were the result of at least...
Table 5: Crash frequencies of categorical variables

| No. | Factor                  | Level     | Total crashes | Injury crash rate (%) |
|-----|-------------------------|-----------|---------------|-----------------------|
| 1   | Period of day           | AM peak   | 730           | 40.49                 |
|     |                         | Off-peak  | 1252          | 37.25                 |
|     |                         | Evening   | 373           | 38.17                 |
|     |                         | Night     | 176           | 41.24                 |
|     |                         | Monday    | 265           | 38.49                 |
|     |                         | Tuesday   | 566           | 40.28                 |
|     |                         | Wednesday | 957           | 38.77                 |
| 2   | Day of week             | Thursday  | 657           | 36.99                 |
|     |                         | Friday    | 400           | 40                    |
|     |                         | Saturday  | 261           | 34.62                 |
|     |                         | Sunday    | 201           | 39.6                  |
|     |                         | Northwest | 1167          | 37.87                 |
|     |                         | Northeast | 858           | 40.44                 |
|     |                         | Southwest | 226           | 33.62                 |
|     |                         | Southeast | 984           | 38.82                 |
|     |                         | Boundary  | 72            | 39.13                 |
| 3   | Quadrant                | Right-angle | 1338        | 39.61                 |
|     |                         | Left-turn | 1217          | 39.61                 |
|     |                         | Right-turn| 752           | 40.69                 |
| 4   | Collision type          | Lights off| 2503          | 38.63                 |
|     |                         | Lights on | 680           | 37.94                 |
|     |                         | None      | 124           | 39.52                 |
|     |                         | Dark      | 757           | 2.02                  |
| 5   | Street lighting conditions | Dark (lighted) | 581        | 33.22                 |
|     |                         | Daylight  | 1967          | 53.99                 |
| 6   | Lighting conditions     | Clear     | 2350          | 39.19                 |
|     |                         | Rain      | 609           | 50.57                 |
|     |                         | Snow      | 348           | 12.93                 |
| 7   | Weather conditions      | Light     | 2178          | 54.78                 |
|     |                         | Medium    | 808           | 8.79                  |
|     |                         | Heavy     | 321           | 8.79                  |
| 8   | Traffic conditions      | Stop sign | 2504          | 42.37                 |
|     |                         | Yield sign| 604           | 70.59                 |
|     |                         | None      | 187           | 12.58                 |
| 9   | Traffic control type    | Male      | 1621          | 25.85                 |
|     |                         | Female    | 1686          | 50.71                 |
| 10  | Gender of driver 1      | Male      | 2315          | 44.32                 |
|     |                         | Female    | 992           | 25                    |
| 11  | Gender of driver 2      | Male      | 2315          | 44.32                 |
|     |                         | Female    | 992           | 25                    |
| 12  | Contributing circum. of driver 1 | No violation | 1700    | 51.12                 |
|     |                         | Alcohol use| 159          | 0                     |
|     |                         | Distracted | 622           | 17.89                 |
|     |                         | Speeding  | 430           | 31.16                 |
|     |                         | Stop/yield sign violation| 310      | 47.74                 |
|     |                         | Improper manoeuvring | 24       | 8.33                  |
| 13  | Contributing circum. of driver 1 | No violation | 1041    | 0.91                  |
|     |                         | Alcohol use| 160           | 0                     |
|     |                         | Distracted | 996           | 40.96                 |
|     |                         | Speeding  | 276           | 2.54                  |
|     |                         | Stop/yield sign violation| 672      | 69.94                 |
|     |                         | Improper manoeuvring | 161      | 69.57                 |
| 14  | Injury severity         |           | 3307          | 38.52                 |

Table 6: Driver age statistics

| Factor     | Mean | Standard deviation | Min. | Max. |
|------------|------|--------------------|------|------|
| Driver age | 42.56| 15.73              | 14   | 86   |

one driver’s failure to comply with a stop/yield sign. Thus, the contributing circumstance that resulted in the highest rate (69.94%) was stop/yield sign violation. Lastly, crashes in which at least one driver was female recorded the highest rate (37.85%) of injury crashes. Additionally, it can be observed from Table 6 that the highest age of drivers involved in crashes was 86, while the minimum age was 14. The mean age was approximately 42 years, with a standard deviation of 15.73.

4.2 Results of classification of crashes using SVMs

This section presents the results of the crash classification using SVMs. Eight models were developed using different kernel types. Table 7 tabulates the results of the SVM classification models using the testing data set. Model 2, with a radial basis kernel, had the highest accuracy (83.19%), while the worst accuracy rate (70.25%) was recorded by Model 8, with a polynomial kernel of degree 6. Model 1 (linear kernel) had the highest sensitivity measure (0.8150), while Model 8 (polynomial degree 6) had the lowest sensitivity measure (0.3009). With regard to the precision measure, Model 7 (polynomial degree 5) was the most precise model, with a precision of 0.8306, while Model 3 (sigmoid kernel) was the least precise model, with a precision of 0.6925. Lastly, Model 1 recorded the highest F-measure (0.7784), while the lowest F-measure (0.4384) was recorded by Model 8 (polynomial degree 6). The variation of performance measures with the different models is shown in Fig. 2.

4.3 Results of classification of crashes using a GNBC

This section presents the results of the classification of crashes using a GNBC. The results of the classification are shown in Table 8. The model produced a low accuracy (0.486) but a high sensitivity (0.98).
Table 7: Results of SVM classification

| Model no. | Kernel type     | Accuracy | Error rate | Sensitivity | Precision | F-measure |
|-----------|-----------------|----------|------------|-------------|-----------|-----------|
| 1         | Linear          | 0.8210   | 0.1790     | 0.8150      | 0.7450    | 0.7784    |
| 2         | Radial basis    | 0.8319   | 0.1681     | 0.7586      | 0.7961    | 0.7769    |
| 3         | Sigmoid         | 0.7763   | 0.2237     | 0.7555      | 0.6925    | 0.7226    |
| 4         | Polynomial degree 2 | 0.8271  | 0.1729     | 0.7461      | 0.7933    | 0.7690    |
| 5         | Polynomial degree 3 | 0.8259  | 0.1741     | 0.7461      | 0.7907    | 0.7677    |
| 6         | Polynomial degree 4 | 0.8041  | 0.1959     | 0.6458      | 0.8078    | 0.7178    |
| 7         | Polynomial degree 5 | 0.7606  | 0.2394     | 0.4765      | 0.8306    | 0.6056    |
| 8         | Polynomial degree 6 | 0.7025  | 0.2975     | 0.3009      | 0.8067    | 0.4384    |

Fig. 2: Variation of performance measures for the SVMs

Table 8: Results of GNBC classification

| Model | Accuracy | Error rate | Sensitivity | Precision | F-measure |
|-------|----------|------------|-------------|-----------|-----------|
| 1     | 0.4859   | 0.5151     | 0.98        | 0.43      | 0.59      |

5. Discussion

The study sought to develop classification models to predict the injury severity of angle crashes involving two vehicles at unsignalized intersections using SVMs and a GNBC. A total of 3307 reported crashes from 2008 to 2015 were extracted from a crash database and used in the analysis. Of the total number of crashes, 1272 resulted in injury and/or fatality, while the remaining 2035 crashes were non-injury crashes. The spatial distribution of the crashes showed that the downtown area of Washington, DC experienced the highest frequency of crashes. Also, most of the crashes occurred during off-peak periods and under light traffic conditions. Right-angle collisions were the most frequent collision type. The combination of driver-contributing circumstances that resulted in the highest injury rate was a stop/yield sign violation by one driver and no violation on the part of the other driver.

Comparatively, the SVM algorithm classified crashes more accurately than the GNBC. The SVM with a radial basis kernel function predicted injury severity with the highest accuracy (83.19%), while the GNBC predicted injury severity with the lowest accuracy (48.49%). The accuracy of an SVM model is dictated by the type of kernel function used. The linear, radial basis and sigmoid kernel types produced similar accuracies, though the best model was achieved by the SVM with the radial basis kernel. Also, it was observed that higher degrees of the polynomial kernels resulted in poor accuracies. This can be explained by the fact that...
hyperplanes of higher-degree polynomials tend to over-classify data sets, which leads to misclassifications. The GNBC produced the worst-performing model. This confirms that this machine-learning technique is not suitable for the classification of injury severity. Crashes are generally indeterministic, and their occurrence does not usually follow the Gaussian distribution.

Similarly, the SVM with polynomial degree 5 predicted injury severity with the highest precision (83.06%), and the lowest precision was provided by the GNBC (42.66%). This implies that the highest proportion of cases that were correctly classified as positive (injury) was provided by the SVM. However, the GNBC was determined to be the most sensitive model, and was able to identify the highest proportion of positive cases (injury). The F-measure, which is a function of precision and sensitivity, is a combined measure for both precision and sensitivity. The highest F-measure was provided by the SVM with a linear kernel function. The confusion matrix of the most accurate model (SVM with a radial basis kernel) is presented in Table 9.

The table shows that 53.9% of the crashes were correctly classified as non-injury crashes, while 7.5% were wrongly classified as non-injury crashes. Similarly, 29.3% of the crashes were correctly classified as injury crashes, while 9.3% were wrongly classified as injury crashes.

6. Conclusion and recommendation

In conclusion, SVMs are better predictors or classifiers of injury severity than GNBCs. However, SVMs require a higher level of parameter tuning to achieve the best model.

This study explored the SVM and GNBC machine-learning techniques. Future research might explore other techniques, such as artificial neural networks, decision trees, k-nearest neighbours and linear discriminants. Other types of crash could also be explored at unsignalized intersections. Furthermore, these analyses could be extended to signalized intersections.

Conflict of interest statement. None declared.

References

1. National Highway Traffic Safety Administration. FARS Encyclopedia. https://www-fars.nhtsa.dot.gov/Main/index.aspx (8 June 2020, date last accessed).
2. District Department of Transportation. 2013-2015 annual DC crash analysis report. Technical report, https://ddot.dc.gov/sites/default/files/dc/sites/ddot/publication/attachments/2013-2015%20Annual%20DC%20Crash%20Analysis%20Report.pdf (14 June 2020, date last accessed).
3. National Highway Traffic Safety Administration. National motor vehicle crash causation survey: report to Congress. Technical report, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059 (14 June 2020, date last accessed).
4. Harwood DW. NCHRP Synthesis of Highway Practice 191: Use of Rumble Strips to Enhance Safety. Washington, DC: Transportation Research Board, 1993, No. 178, May 1995, pp. 12–16.
5. Yannis G, Kondyli A, Mitzalis N. Effect of road lighting conditions on the frequency and severity of road accidents. ICE Proceedings Transport 2013; 166:271–81.
6. Kim D-G, Washington S, Oh J. Modeling crash types: new insights into the effects of covariates on crashes at rural intersections. J Transp Eng 2006; 132:282–92.
7. Hermans E, Brijs T, Stiers T. The impact of weather conditions on road safety investigated on an hourly basis, TRB Annual Meeting 2006. Paper No. 06-1120.
8. Lau MY, May AD. Accident prediction model development: signalized intersections. Technical report. Institute of Transportation Studies, University of California, Berkeley 1988.
9. Kamer-Ainur A, Marioara M. Errors and limitations associated with regression and correlation analysis. Statistics and Economic Informatics 709, 2007, http://steconomie.utoroadea.ro/analize/volume/2007/v2-statistics-and-economic-informatics1.pdf (14 June 2020, date last accessed).
10. Ackaaah W, Salifu M. Crash prediction model for two-lane rural highways in the Ashanti region of Ghana. Int Assoc Traffic Safety Sci Res 2011; 35:34–40.
11. Rifaat SM, Chin HC. Accident severity analysis using ordered probit model. J Adv Transp 2007; 41:91–114.
12. Li X, Lord D, Zhang Y. Predicting motor vehicle crashes using support vector machine models. Accid Anal Prev 2008; 40:1611–18.
13. Khera D. Prediction and analysis of injury severity in traffic system using data mining techniques. Int J Comput Applic 2015; 7: 1–7.
14. DC.gov. New U.S. Census Bureau Numbers Officially Put DC’s Population Over 700,000. 2018. https://osse.dc.gov/release/new-us-census-bureau-numbers-officially-put-dc%E2%80%99s-population-over-700000 (8 June 2020, date last accessed).
15. Winship T. The 10 US Cities With The Worst Traffic. 2018. https://www.businessinsider.com/the-10-us-cities-with-the-worst-traffic-2018-2 (8 June 2020, date last accessed).

Table 9: Results of classification using the SVM with a radial basis kernel

| Total no. of observations | Predicted |
|---------------------------|-----------|
|                           | Negative  | Positive |
| Actual                    | 446       | 77        |
| Positive                  | 242       |           |

| Conflict of interest statement. None declared.

12 | Arhin and Gatiba
16. DC.gov. District Department of Transportation. https://ddot.dc.gov/ (8 June 2020, date last accessed).
17. District Department of Transportation. DDOT by the Numbers. https://ddot.dc.gov/page/ddot-numbers (8 June 2020, date last accessed).
18. American Society of Civil Engineers. Report Card For District of Columbia’s Infrastructure, Technical report 2016, https://www.infrastructurereportcard.org/wp-content/uploads/2017/02/ASCE-DC-Report-Card-2016-Report-1.pdf (14 June 2020, date last accessed).