Predicting Injury Severity of Angle Crashes Involving Two Vehicles at Unsignalized Intersections Using Artificial Neural Networks

Stephen A. Arhin
Howard University Transportation Research and Data Center
Washington, DC, USA

Adam Gatiba
Howard University Transportation Research and Data Center
Washington, DC, USA

***Abstract*—In 2015, about 20% of the 52,231 fatal crashes that occurred in the United States occurred at unsignalized intersections. The economic cost of these fatalities have been estimated to be in the millions of dollars. In order to mitigate the occurrence of theses crashes, it is necessary to investigate their predictability based on the pertinent factors and circumstances that might have contributed to their occurrence. This study focuses on the development of models to predict injury severity of angle crashes at unsignalized intersections using artificial neural networks (ANNs). The models were developed based on 3,307 crashes that occurred from 2008 to 2015. Twenty-five different ANN models were developed. The most accurate model predicted the severity of an injury sustained in a crash with an accuracy of 85.62%. This model has 3 hidden layers with 5, 10, and 5 neurons, respectively. The activation functions in the hidden and output layers are the rectilinear unit function and sigmoid function, respectively.**

Keywords-crashes; unsignalized intersection; artificial neural network; injury severity

I. INTRODUCTION

Even though intersections constitute a relatively low proportion of the facilities of transportation systems, a significant number of crashes occur at these locations, especially in urban areas. In California for instance, an annual average of 1.5 crashes occur at unsignalized intersections in rural locations, compared to an average of 2.5 crashes per year in urban locations [1]. Data from the World Health Organization (WHO) reveal that 1.25 million people die annually worldwide in road crashes. The economic cost of these deaths is estimated to be approximately $260 billion per year [2]. In the United States, there were a total of 37,456 fatalities in road-related crashes reported in 2016 [3]. Though most of these crashes occurred on road segments, a significant number occurred at or near intersections. Out of the total of 52,231 fatal crashes in the United States in 2015, approximately 4.4% (2,298) of the crashes occurred at STOP-controlled intersections, while 7.5% (3,917) of the crashes occurred at intersections controlled by traffic signals. Intersections without any type of traffic control device recorded the highest number of fatal crashes (4,227) [4].

Several studies have investigated the causes of these crashes. These causes are either driver-induced, or occur due to road geometry, road defects, vehicle defects and atmospheric or weather conditions. Various countermeasures have been proposed and/or implemented to reduce the occurrence of crashes at intersections, which in some instances have been successful. In order to effectively reduce the frequency and mitigate the severity of intersection related crashes, it is necessary to explore the predictability of these crashes based on the pertinent factors and circumstances that might have contributed to the occurrence of these crashes. Several studies have resulted in the development of mathematical models that predict crashes on roadways in general and, in a few instances, at unsignalized intersections in particular. These mathematical models include linear regression and machine learning methods. Given the varying characteristics of intersections, it is necessary to develop models that are focused and specific to a particular set of conditions. This study therefore focuses on the development of models to predict the severity of right-angle crashes involving two vehicles at unsignalized intersections in urban centers using ANNs.

II. LITERATURE REVIEW

A. Contributory Factors for Intersection-Related Crashes

There are many factors that determine the degree of injury sustained by people involved in crashes at unsignalized intersections. However, it is shown that only certain factors are statistically significant predictors. Authors in [5] assessed the degree of injury sustained by drivers involved in angle collisions in relation to the fault status of drivers. The results of the study showed that drivers who were not at fault tended to sustain more severe injuries than those who were at fault. It was further determined that injury severity was affected by factors including time of year, speed limit, age, gender, restraint/helmet use, and alcohol/drug use. Authors in [6] concluded that the road surface condition (wet or dry) was a significant predictor of injury severity. Additionally, female drivers are more likely to sustain severe injuries than male drivers. Crashes at urban areas were determined to result in less serious injuries than crashes at rural areas [6]. Also, traffic volume on a major road is a significant predictor of crashes at unsignalized intersections [7].

Corresponding author: Stephen A. Arhin (saarhin@howard.edu)

The geometric characteristics and features of unsignalized intersections have also been found to be potential explanatory variables in crash prediction models. Authors in [8] predicted the frequency of accidents at unsignalized intersections in urban areas using negative binomial models. It was concluded that besides traffic exposure functions such as traffic flow, which usually significantly predict crashes, intersection geometrics, absence of street lighting and dedicated left-turning lanes are positively correlated with accident frequency at intersections. Typical geometric characteristics included number of lanes on major road, width of lanes, and presence of median on intersecting roads. The study further revealed that T-intersections with Yield control had a much lower accident potential than those with Stop control.

B. Crash Prediction Models

Several modeling techniques have been employed to predict crashes at intersections.

1) Linear Regression Models

Linear regression modeling is an approach to establish a relationship between scalar responses, 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:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{in} + \xi_i \quad (1)$$

where, y_i is the ith dependent variable, β_1, $\beta_2 \ldots \beta_p$ are estimated parameters, x_{i1}, $x_{i2} \ldots x_{in}$ are the predictor variables of the ith dependent variable and ξ_i is the error term. The error term is an independent and normally distributed random variable with 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. Authors in [9] investigated the relationship between the number of injuries or property damage only (PDO) crashes that occur annually at intersections and traffic and environmental factors. The crash records (ranging from 1984 to 1987) of 2,488 intersections in California were sampled. The linear regression analysis employed in this study was conducted in two levels. In 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 were included as predictor variables. The results of the analysis showed that the accuracy of the model improved as more predictors variables were added.

Though linear regression models are easy to use and interpret, it has been shown that they are not ideal for crash predictions. 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. Further, some factors have been determined to strongly correlate with each other, thus introducing multicollinearity thereby invalidating such linear models [10]. In overcoming the shortcomings of the linear regression models, generalized linear models (GLMs) have been used to model crashes at intersections. GLMs are a flexible generalization of the ordinary linear regression that can accommodate the non-normal distributed error terms. The most common forms of generalized linear models used in crash prediction models are the negative (NB) model and the ordered probit model (OPM)

2) Negative Binomial Model

NB models are a generalization of the Poisson regression. Unlike the 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 predictions due to the nature of the dependent variables in such analysis. Usually the response required is the number of crashes at a specific location. Such responses are nonnegative integers and generally follow the NB distribution. The distribution is given by the following Poisson-Gamma distribution:

$$Pr(Y=y_i|u_i,\alpha)=\frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{1}{(1+\alpha e^{xi\beta})}\right)^{\alpha^{-1}}\left(\frac{1}{(1+\alpha e^{xi\beta})}\right)^{y_i} \quad (2)$$

where, u is the mean of the dependent variable y, β is an estimated parameter to be estimated, α is the heterogeneity parameter, and x_i is the ith the predictor variable. Authors in [11] investigated the relationship between crash frequencies and factors such as traffic conditions, geometric and operational characteristics or roadways, and weather conditions using data of crashes that occurred from 2004 to 2010 on a motorway in Auckland. The NB regression model developed had a goodness of fit, ρ^2 of 0.119. Additionally, several individual predictors such as length of road segments, AADT, number of lanes and shoulder width were found to be significant predictors of the model.

3) Ordered Probit Models

The ordered probit model (OPM) is used in developing models which have an ordered response. This approach in modeling data employs the use of 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:

$$Pr(y_i = j|X_I) = \frac{\exp(\tau_j - X_i\beta)}{(1+\exp(\tau_j - X_i\beta))} - \frac{\exp(\tau_{j-1} - X_i\beta)}{(1+\exp(\tau_{j-1} - X_i\beta))} \quad (3)$$

where, y_i is an observable ordinal variable, X_i is a vector of exogenous variables, β is a vector of unknown parameters to be estimated and and τ_j is the threshold associated with the j^{th} ordinal partition interval which are assumed to be of ascending order. OPM has been applied in the development of several crash prediction models which seek to predict injury severity based on several factors. Authors in [12] developed an OPM that sought to relate the severity of crashes experienced at freeway exits. Crash data for 326 locations in Florida were sampled. The results of the study indicated that the factors which significantly influenced crash severity included mainline

lane number, length of ramp, difference of speed limits between mainline and ramp, light condition, weather condition, surrounding land type, alcohol/drug involvement, road surface condition, and crash type. The model developed had a goodness of fit of 0.019 and a chi-squared goodness of fit value of 95.63.

4) Empirical Bayes Refinement of the GLM

Crash estimates made with GLMs are susceptible to regression-to-the-mean. The regression to mean occurs when a randomly large number of accidents during a period is normally followed by a reduced number of accidents during a similar after period, even if no measure has been implemented. The GLMs do not account for this effect. Hence, to improve the accuracy of the predictions made with GLMs, the empirical Bayes (EB) method is usually applied. The EB method compensates for the regression-to-the-mean bias by pulling the crash count towards the mean. Thus, prior data (observed crash counts) are combined with the predicted crash frequency from the GLM to calculate a corrected value. The corrected value is expected to lie somewhere between the observed crash frequency and the predicted frequency from the GLM. This is expressed as:

$$E = Weight \times \mu + (1 - Weight) \times \text{Observed crashes frequency} \quad (4)$$

where, E is the corrected value, and μ is the average number of crashes (determined from the GLM) [13].

5) Artificial Neural Networks (ANNs)

ANNs are mathematical models inspired by the biological neural networks in the human brain. ANNs are used in engineering to perform complex tasks such as pattern recognition, forecasting, data compression and classification. The effectiveness of an ANN is based on its ability to approximate both linear and nonlinear functions to a required degree of accuracy using a learning algorithm, and to build ''piece-wise'' approximations of the functions [14]. Classification or forecasting using ANNs involves training and learning procedure, where, historical data (a set of input data with known outputs) is presented to the network. Usually large amounts of such data are required for the training of the network. The network goes through a learning process by constructing a network of inputs and outputs, and weights assigned to each mapping are adjusted at each iteration. The method by which these weights and bias levels of a network are updated is determined by the learning rule used. Thus, the learning rule helps a neural network to learn from the existing conditions and improve its performance. There are several learning rules used in training neural networks. Notable among the rules are the hebbian, perceptron (error-correction), delta, correlation and outstar learning rules [15]. However, the most common known rule is the multilayer perceptron (MLP). MLP basically consists of three layers: input layer, hidden layer, and output layer. MLP is a feed forward network in which information flows from the input layer through the hidden to the output layer to produce the outcome. These layers have interconnected nodes (neurons). The interconnections are assigned weights (representing information flow) which are computed using mathematical functions. The outputs for specific inputs are obtained by adjusting the weights to minimize the errors between the output produced and the desired output by error-back propagation. The MLP is known to be a universal approximator because of its ability to approximate continuous functions on a compact set of real numbers with little assumption made. Activation functions, also called transfer functions, are an essential component of ANNs. Activation functions are models in the output neurons of the ANN which introduce non-linearity into the network. They function by calculating the weighted sum of their inputs and adding a bias, then deciding whether a neuron should be activated or not. The three most common types of activation functions used in an ANN are the sigmoid, the hyperbolic tangent, and the rectified linear unit [16]. Authors in [17] utilized ANNs to develop a model to show the relationship between crash severity on urban highways, and traffic variables such as traffic volume, flow speed, human factors and road, vehicle and weather conditions. The study showed that MLP with feed forward back propagation networks provided the best results compared to other learning methods. Network architecture with 2 hidden layers with 17 and 7 neurons respectively were determined to be the best. Mean square errors (MSE) within acceptable range of 3% to 4% were achieved. Also, correlation coefficients of 86% to 87 % were achieved.

III. METHODOLOGY

A. Study Area

This study is based on data obtained in the District of Columbia (DC). The capital of the USA, Washington, DC is divided into four (equal) quadrants 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% [18]. The city is highly urbanized, and it's ranked the sixth most congested city in the United States with each driver spending an average of 63 hours in traffic annually [19]. It has a land area of 68.34mi^2 and a total of 1,503 miles of roadway comprised of local roads, collector roads, minor arterials, principal arterials, freeways and interstates [20]. Also, the city has about 7,700 intersections of which 1,450 are signalized [21]. The American Society of Civil Engineers' 2017 infrastructure report card reported that about 95% of the roads in DC are in poor condition [22].

B. The Crash Database System

Crash prediction models are data dependent and as a result the accuracy of the developed models depends largely on the quality of the available crash data. To ensure that a reliable model is developed, this research utilized traffic crash data from the District Department of Transportation's (DDOT's) crash database called 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 Police Department Form number 10 (PD-10) crash reporting form. The crash data is then downloaded through secure servers from MPD into DDOT's database and are then processed and made available in TARAS2, which is an Oracle-based application. TARAS2 contains data fields that can be broadly categorized

under vehicle characteristics, environmental conditions, roadway characteristics, traffic exposure characteristics, as well as crash location, date, time, crash type, crash severity and information on of persons involved.

C. Data Extraction and Encoding

Nine years of crash data (2008-2015) were queried and extracted from TARAS2. The data were then filtered to obtain angle crashes involving two vehicles at unsignalized intersections. Further, the extracted data were cleaned by identifying and removing duplicate and incomplete crash records and irrelevant data fields. In all, 3,307 data points were extracted and used for analysis. 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 condition, street lighting condition, lighting condition, weather condition, traffic condition, traffic control type, drivers' age, drivers' gender, contributing circumstances, and injury severity. Only numerical data can be analyzed by ANNs. Hence, qualitative data must be converted to quantitative data. Thus, both input and output data must be encoded into either real or integer values. Secondly, binary method (0 and 1) of encoding has been determined to yield better results since it minimizes the loss functions values with respect to the models' parameters. The loss value determines how well the model fits the data set. The lower the loss function value the better the model fits the data set. Table II presents the variables and coding scheme used in this study.

D. Types of Collision

The crash types considered for this study are angle collisions. Three types of angle collisions are specified: right-angle, right turn, and left turn collisions.

- Right-angle collision: This type of collision occurs when the side of one vehicle is impacted by the front of another vehicle which is traveling in a direction at right angle to the direction of the former vehicle. Figure 1 depicts a right-angle collision at an intersection.
- Right turn collision: This type of collision occurs when a vehicle turning right at an intersection is impacted by a vehicle from the other intersecting road. Figure 2 depicts a right turn collision.
- Left turn collision: This type of collision occurs when a left turning vehicle at an intersection is impacted by a vehicle from the oncoming traffic. Figure 3 depicts a left turn collision.

E. 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. Table I shows the levels of crashes used in the analysis.

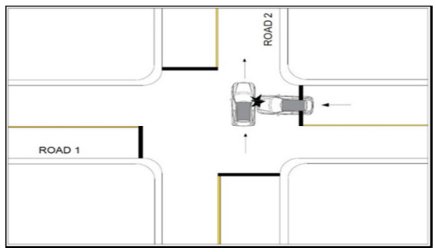

Fig. 1. Right-angle collision

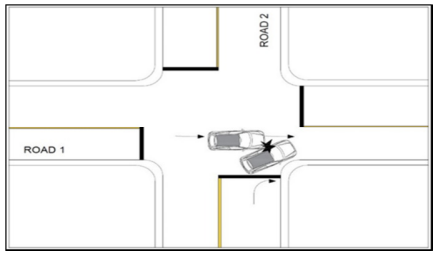

Fig. 2. Right turn collision

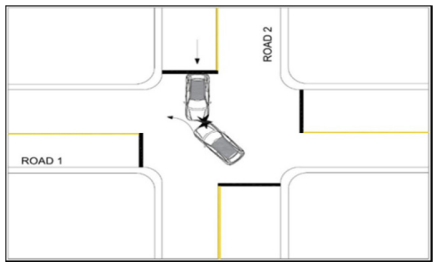

Fig. 3. Left turn collision

TABLE I. LEVELS OF INJURY SEVERITY

Injury Severity	**Level**
No Injury	Non-Injury
Complain	Injury
Non-Disabling Injury	
Disabling Injury	
Fatal	

F. Data Standardization

To achieve accurate predictions from machine learning models it is necessary that variables used in developing the models are of equal scale. Also, most optimization algorithms minimize the loss function converge faster when variables are of the same scale. The method of scaling used on this data set is standardization. The raw scores (of the encoded data) are converted to standard scores by subtracting the mean of each variable from the raw score of each observation and then dividing the difference by the standard deviation of the variable. By doing so, the variables are transformed to have a mean of zero and a unit variance. The standardized value, Z, of each score of each variable is given by (5):

$$Z = (X - \bar{X}) / \sigma \quad (5)$$

where, $\bar{X}$ is the mean of the variable, X is the encoded score of each observation of a variable and σ is its standard deviation.

TABLE II. VARIABLE ENCODING

Variable	Variable Name	Code	Variable	Variable Name	Code
	Day of Crash			**Lighting condition**	
X_1	Monday	1-Present, 0-Otherwise	X_{26}	Dark	1-Present, 0-Otherwise
X_2	Tuesday	1-Present, 0-Otherwise	X_{27}	Dark Lighted	1-Present, 0-Otherwise
X_3	Wednesday	1-Present, 0-Otherwise	X_{28}	Daylight	1-Present, 0-Otherwise
X_4	Thursday	1-Present, 0-Otherwise		**Weather Condition**	
X_5	Friday	1-Present, 0-Otherwise	X_{29}	Clear	1-Present, 0-Otherwise
X_6	Saturday	1-Present, 0-Otherwise	X_{30}	Rain	1-Present, 0-Otherwise
X_7	Sunday	1-Present, 0-Otherwise	X_{31}	Snow	1-Present, 0-Otherwise
	Time of Day		X_{32}	**Traffic Condition**	0-Low, 1-Medium, 2-High
X_8	A.M. Peak (06:00 – 10:00)	1-Present, 0-Otherwise		**Traffic Control Type**	
X_9	Off Peak (10:00 – 15:00)	1-Present, 0-Otherwise	X_{33}	Stop	1-Present, 0-Otherwise
X_{10}	P.M. Peak (15:00 – 19:00)	1-Present, 0-Otherwise	X_{34}	Yield	1-Present, 0-Otherwise
X_{11}	Evening (19:00 – 00:00)	1-Present, 0-Otherwise	X_{35}	None	1-Present, 0-Otherwise
X_{12}	Night (0000 – 0600)	1-Present, 0-Otherwise		**Contributing Circumstances of Driver 1**	
	Quadrant		X_{36}	No Violation D1	1-Present, 0-Otherwise
X_{13}	NW	1-Present, 0-Otherwise	X_{37}	Alcohol/ Drug Use D1	1-Present, 0-Otherwise
X_{14}	SW	1-Present, 0-Otherwise	X_{38}	Speeding D1	1-Present, 0-Otherwise
X_{15}	NE	1-Present, 0-Otherwise	X_{39}	STOP/ YIELD Sign Violation D1	1-Present, 0-Otherwise
X_{16}	SE	1-Present, 0-Otherwise	X_{40}	Improper Maneuvering D1	1-Present, 0-Otherwise
X_{17}	BN	1-Present, 0-Otherwise		**Contributing Circumstances of Driver 2**	
	Type of Collision		X_{42}	No Violation D2	1-Present, 0-Otherwise
X_{18}	Right Angle	1-Present, 0-Otherwise	X_{43}	Alcohol/ Drug Use D2	1-Present, 0-Otherwise
X_{19}	Left Turn	1-Present, 0-Otherwise	X_{44}	Speeding D2	1-Present, 0-Otherwise
X_{20}	Right Turn	1-Present, 0-Otherwise	X_{46}	Improper Maneuvering D2	1-Present, 0-Otherwise
	Road Surface Condition		X_{47}	Distraction D2	1-Present, 0-Otherwise
X_{21}	Wet	1-Present, 0-Otherwise			
X_{22}	Dry	1-Present, 0-Otherwise	X_{48}	Age of Driver 1	1-Present, 0-Otherwise
	Street Lighting		X_{49}	Age of Driver 2	1-Present, 0-Otherwise
X_{23}	Light Off	1-Present, 0-Otherwise	X_{50}	Gender of Driver 1	0-Female, 1-Male
X_{24}	Light On	1-Present, 0-Otherwise	X_{51}	Gender of Driver 2	0-Female, 1-Male
X_{25}	None	1-Present, 0-Otherwise	Y_1	Injury Severity	0-No Injury, 1-Injury

G. Development of Models

The process of classification by ANN is an iterative process of weight adjustments based on information flow that mimics the functioning of neurons in the human brain. The steps below describe in detail how models for crash injury severity classification were developed using ANN:

- Selection of network architecture.
- Training of neural network.
- Testing and evaluation of model.

1) Selection of Network Architecture

The network architecture was first set up. A multi-layer perceptron (MLP) feedforward ANN was adopted to develop classification models. An MLP consists of at least three layers: an input layer, hidden layer(s) and an output layer. Each layer consists of nodes or neurons. The neurons of each layer are interconnected with those of the succeeding layer. Also, the neurons of the hidden and output layers are embedded with nonlinear activation functions. The MLP ANN architecture used in this research consists of an input layer with 44 neurons (each neuron represents each of the input variables, X_i in Table II) and an output layer with 1 neuron, which is the target or dependent variable, Y. The number of hidden layers and neurons varied for several iterations until the optimal numbers of hidden layers and neurons which produced the best model were obtained. Figure 4 shows the MLP ANN architecture used in developing the model.

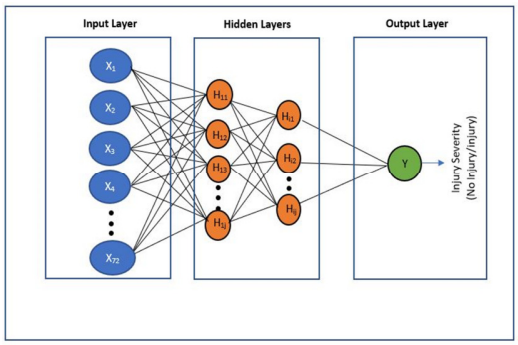

Fig. 4. MLP ANN

2) Training of Neural Network

Training of the neural network by backward propagation was carried out in the following sequence:

- Presentation of training dataset to the network: The training dataset was imported into the network to commence training. The vector of independent variables was fed into each input neuron connected to neurons of the first hidden layer. The training process was initialized by randomly

selecting weights for all interconnections between the neurons of the input and hidden layers.

- Forward Computation: The forward propagation was then implemented by multiplying the weights with the values of the input neurons and the sum products are stored in the corresponding neurons of the hidden layer. The weighted sums are subsequently transferred into an activation function and based on the output of the functions, the neuron is either activated or not. Mathematically this can be expressed as:

$$V_j^l = \sum_{i=0}^{m} w_{ji}^l \, x_i^{(l-1)} \quad (6)$$

$$y_j^l = \phi_i(V_j) \quad (7)$$

where, V_j^l is the weighted sum in j^{th} neuron of the l^{th} hidden layer, w_{ji}^l is the weight coefficient of the j^{th} neuron of the l^{th} layer that is fed from the i^{th} neuron in layer l-1, $x_i^{(l-1)}$ the output of th i-th neuron in the previous layer l-1, y_j is the output of the of the j^{th} neuron in layer l-1, ϕ_i is the activation function which is a rectilinear unit function in the hidden layers and a sigmoid function in the output layer. Hence for the last layer (output layer) l=L,

$$y_j^L(n) = O_J \quad (8)$$

where, O_J is the output of the n-th iteration.

- Computation of error: The error of the j^{th} neuron of the n^{th} iteration is then computed as

$$e_j(n) = t_j(n) - O_J(n) \quad (9)$$

where, t_j is the target output.

- Backward computation: The weights in the network are adjusted based on a local gradient, σ, which is a function of the error, e, and computed as follows:

$$\sigma_j^l(n) = e_j^L(n)\, \phi' \left(V_j^L(n)\right) \quad (10a)$$

for neuron j in the output layer L, and

$$\sigma_j^l(n) = e_j^L \phi' \left(V_j^l(n)\right) \sum_k \sigma_k^{(l+1)}(n) w_{kj}^{(l+1)}(n) \quad (10b)$$

for neuron j in the hidden layer L, where, k is the succeeding neuron in layer l+1 and $\phi'(\cdot)$ is the derivative of the function $\phi(\cdot)$. The weights in the network are then adjusted by the given relation:

$$w_{ji}^l(n+1) = w_{ji}^l(n)_J + \alpha\left[w_{ji}^l(n)_J(n-1)\right] + \eta\sigma_j^l(n) y_i^{(l-1)}(n) \quad (11)$$

where η is the learning-rate parameter and α is the momentum constant.

- Iteration: The procedures in the three previous steps are repeated for batches of 3 observations per iteration until the stopping criteria of 100 epochs is met. Figure 5 illustrates the training process.

3) Model Testing and Evaluationl

After the training of the network for the required number of epochs (100), the model was tested using the test dataset. The accuracy of the model was evaluated by the confusion matrix. The number of hidden layers and neurons in the network architecture was varied and the training process was repeated. This iterative process was done until the model with the best performance was achieved.

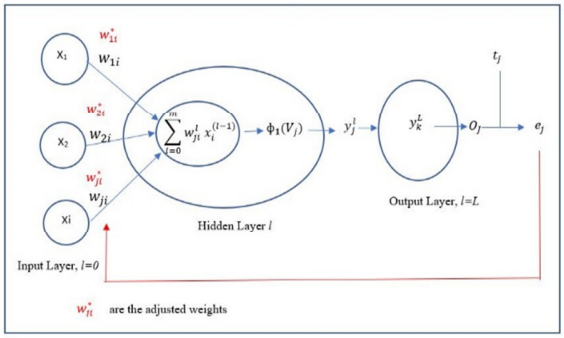

Fig. 5. ANN training process

4) Model Evaluation

The performance of each of the models was assessed using the test dataset. The results were then evaluated by 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 III shows the confusion matrix for a two-class classifier.

TABLE III. CONFUSION MATRIX

Total No. of Observations		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

The entries of the CM are defined as follows: True Positive (TP) instances are positive and correctly classified as positive, True Negative (TN) instances are negative and correctly classified as negative, False Positive (FP) instances are negative but wrongly classified as positive, and False Negative (FN) instances are positive but wrongly classified as negative. Based on the CM, the following measures were computed to evaluate the models developed.

- Accuracy (AC): The accuracy is the proportion of the total number of predictions that were correctly classified. It is computed as:

$$AC=(TN+TP)/(TN+FP+FN+TP) \quad (12)$$

- Error Rate (ER): The error rate is the rate at which predictions will be misclassified:

$$ER=1-AC \quad (13)$$

- Sensitivity (S): It is the proportion of positive cases that were correctly identified:

$$S=TP/(FN+TP) \quad (14)$$

- Precision (P): It is the proportion of the predicted positive cases that were correct:

$$P=TP/(FP+TP) \quad (15)$$

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

$$F=2(S\cdot P)/(S+P) \quad (16)$$

H. Analysis Software

The classification models of all three machine learning techniques were developed by using the high-level general-purpose programming language Python. Especifically, the Anaconda Python distribution was used. This is an open source distribution 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, Tensorflow and Keras libraries were imported to develop the ANN models. In addition, the descriptive statistics of the data were obtained using IBM Statistical Software for Social Scientist (SPSS).

IV. RESULTS

A. Descriptive Statistics

Tables IV and V present the descriptive statistics of the data set. The frequencies of categorical variables are presented in Table IV, while Table V presents the mean and standard deviation of the continuous variable Age. It can be observed from Table IV that the highest number of crashes (1,252) occurred during the off-peak period, from 10:00A.M. to 3:00P.M., while the least number of crashes (176) occurred at night, between 12:00AM to 6:00AM. Most of the crashes occurred on Tuesdays, Wednesdays and Thursdays while Sundays recorded the least number of crashes. The Northwest quadrant of Washington D.C. recorded the highest number of crashes (1,167). Right-angle collision was the most frequent occurring crash type. Most of the crashes occurred under daylight, clear weather and light level traffic conditions. Though most crashes were as a result of no violation on the part of one or both drivers, distracted driving and Stop/Yield-sign violation were also reported as comparatively high contributing circumstances. Among the drivers involved, 3,936 were male and 2,678 were female. Of the 3,307 recorded crashes, 1,274 resulted in injury. It is observed that the rate of injury crashes was highest during the night (41.24%), on Fridays (41%), and in the northeast quadrant (40.44%). Most were right turn collisions (40.69%), absent street lights (39.52%), rainy weather (50.57%), under light traffic conditions (54.78%). 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 a result of at least one driver's failure to comply with a Stop/Yield sign. Thus, the contributing circumstance which resulted in the highest rate (69.94%) is Stop/Yield sign violation. Crashes in which at least on driver was a female recorded the highest rate of injury crashes. A correlation analysis was conducted to investigate the relations between age and injury severity. The results are presented in Table VI. The Spearman's Rho of -0.52 was found to be statistically significant (p=0.03). This implies that, the severity of a crash increased with decreasing age of drivers involved in the crash.

TABLE IV. CRASH FREQUENCIES

No	Factor	Level	Crashes			
			Total	Injury	Non-Injury	Injury Rate (%)
1	Period of Day	A.M. Peak	730	296	435	40.49
		Off Peak	1252	466	785	37.25
		P.M. Peak	776	298	478	38.4
		Evening	373	142	230	38.17
		Night	176	73	104	41.24
2	Day of Week	Monday	265	102	163	38.49
		Tuesday	566	228	338	40.28
		Wednesday	957	371	586	38.77
		Thursday	657	243	414	36.99
		Friday	400	160	240	40
		Saturday	261	90	170	34.62
		Sunday	201	80	122	39.6
3	Quadrant	Northwest	1,167	442	725	37.87
		Northeast	858	347	511	40.44
		Southwest	226	76	150	33.62
		Southeast	984	382	602	38.82
		Boundary	72	27	45	39.13
4	Type of Collision	Right Angle	1,338	530	808	39.61
		Left Turn	1,217	438	779	39.61
		Right Turn	752	306	446	40.69
5	Street Lighting Condition	Lights Off	2,503	967	1,536	38.63
		Lights On	680	258	422	37.94
		None	124	49	75	39.52
6	Lighting Condition	Dark	757	15	727	2.02
		Dark (Lighted)	581	193	388	33.22
		Day Light	1,967	1,063	906	53.99
7	Weather Condition	Clear	2,350	921	1,429	39.19
		Rain	609	308	301	50.57
		Snow	348	45	303	12.93
8	Traffic Condition	Light	2,178	1,193	985	54.78
		Medium	808	71	737	8.79
		Heavy	321	71	737	8.79
9	Traffic Control Type	STOP Sign	2,504	1,066	1,450	42.37
		YIELD Sign	604	132	55	70.59
		None	187	76	528	12.58
10	Gender of Driver 1	Male	1,621	419	1,202	25.85
		Female	1,686	855	831	50.71
11	Gender of Driver 2	Male	2,315	1,026	1,289	44.32
		Female	992	248	744	25
12	Contri. Circum. of Driver 1	No Viloation	1,700	869	831	51.12
		Alcohol	159	0	159	0
		Distracted	682	122	560	17.89
		Speed	430	134	296	31.16
		STOP/YIELD Sign Violation	310	148	162	47.74
		Improper Maneuver	24	2	22	8.33
13	Contri. Circum. of Driver 2	No Viloation	1,041	7	764	0.91
		Alcohol	160	0	160	0
		Distracted	996	408	588	40.96
		Speed	276	7	269	2.54
		STOP/YIELD Sign Violation	672	470	202	69.94
		Improper Maneuver	161	112	49	69.57
14	Injury Severity		3,307	1,274	2,033	38.52

TABLE V. DRIVER AGE STATISTICS

Factor	Mean	Standard Deviation	Min.	Max
Drivers Age	42.56	15.73	14	86

TABLE VI. AGE-INJURY SEVERITY CORELATION ANALYSIS

Factor	Test Statistic (Spearman's Rho)	P-value
Age of Driver	-0.52	0.03

B. Spatial Distribution of Crashes

This section presents the results of the spatial analysis of the crashes using ArcGIS Pro software program. The spatial analysis performed included the spatial distribution of crashes based on injury severity and a kernel density analysis for injury crashes. The spatial distribution and density of crashes are shown in Figures 6 and 7, respectively. Figure 7 shows that most of the crashes were located in the NW quadrant. This covers the downtown and central business district of Washington DC. Figure 7 also shows that higher densities of injury crashes are in the same region of Washington DC.

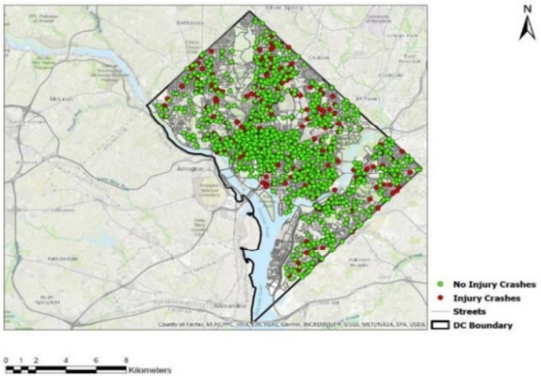

Fig. 6. Spatial distribution of crashes [source: ArcGISPro]

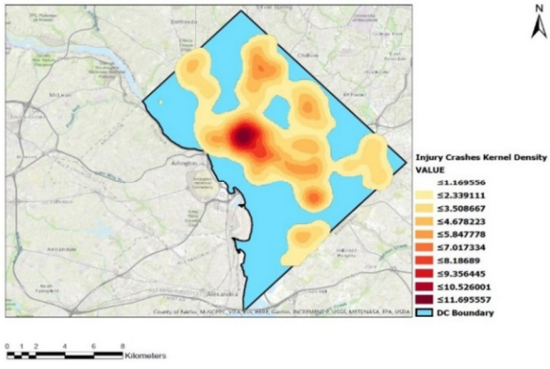

Fig. 7. Kernel density of injury crashes [source: ArcGISPro]

C. Results of Classification of Crashes

Twenty-five distinct ANN models were developed using the training dataset. Each model was trained with batches of 3 observations per iteration until the stopping criteria of 100 epochs was met. The performance of each model was then evaluated using the test data set (which constitutes of 25% of the total dataset). The performance of the models after training and testing are presented in Tables VII and VIII respectively. The Tables show the number of models explored and the structure of the neural network. The performance measures (accuracy, error rate, sensitivity, precision and F-measure) of each model were computed and are also presented.

TABLE VII. RESULTS OF TRAINING ANN

Model	Network Arch.		AC	ER	S	P	F
	Hidden Layer No.	No. of Neurons					
1	1	20	0.9181	0.0819	0.8995	0.8892	0.8943
2	1	15	0.9032	0.0968	0.9162	0.8454	0.8794
3	1	5	0.8649	0.1351	0.8366	0.8170	0.8267
4	1	3	0.8573	0.1427	0.8461	0.7961	0.8203
5	2	25-20	0.9585	0.0415	0.9455	0.9465	0.9460
6	2	20-25	0.9472	0.0528	0.9435	0.9213	0.9322
7	2	20-15	0.9512	0.0488	0.9874	0.8964	0.9397
8	2	15-20	0.9258	0.0742	0.9539	0.8668	0.9083
9	2	10-15	0.9157	0.0843	0.9529	0.8473	0.8970
10	2	15-10	0.9302	0.0698	0.9445	0.8826	0.9125
11	2	5-10	0.8722	0.1278	0.8785	0.8067	0.8411
12	2	10-5	0.9060	0.0940	0.8953	0.8654	0.8801
13	2	6-3	0.8685	0.1315	0.8628	0.8086	0.8349
14	2	3-6	0.8597	0.1403	0.8440	0.8020	0.8224
15	2	2-2	0.8427	0.1573	0.8304	0.7767	0.8026
16	3	30-20-25	0.9516	0.0484	0.9832	0.9170	0.9490
17	3	25-30-20	0.9689	0.0311	0.9204	0.9565	0.9381
18	3	20-15-20	0.9402	0.0598	0.9916	0.8926	0.9395
19	3	15-20-15	0.9404	0.0596	0.9644	0.8916	0.9266
20	3	15-10-15	0.9293	0.0707	0.9738	0.8692	0.9185
21	3	10-15-10	0.9310	0.0690	0.8995	0.8677	0.8833
22	3	5-10-5	0.9115	0.0885	0.8859	0.8270	0.8554
23	3	10-5-10	0.9102	0.0898	0.9414	0.8293	0.8818
24	3	6-4-2	0.9159	0.0841	0.9058	0.8374	0.8702
25	3	6-2-6	0.9237	0.0763	0.9058	0.8547	0.8795

The accuracy, sensitivity, precision and F-measure (F) performance measures range from 0 to 1, with values closer to 1 showing models with better performance measures and conversely values closer to 0 showing worse performance measures. In contrast, models with error rates (ER) closer to 0 are better than models with error rate closer to 1. The results of the analysis in Table VII show that after the training of the models, the accuracy ranged from 84.87% to 96.89%. Model 17 produced the best classification accuracy (96.89%) with a corresponding error rate of 3.11%, while Model 15 produced the worse accuracy (84.87%) with a corresponding error rate of 15.73%. Model 7 had the highest sensitivity (S) measure, while Model 15 had the least sensitivity measure. With regards to the precision measure, Model 17 was the most precise (P) model with a precision of 0.9565, while Model 15 was the least precise one. Model 16 recorded the highest F-measure of 0.9490, while the lowest F-measure was recorded by Model 6. The variation of performance measures with varying models is shown in Figure 8. Table VIII presents the results of evaluation of the trained models using the test data set. The results show that the accuracy (after testing) of the models ranged from 76.54% to 85.62%. Model 22 produced the best classification accuracy (85.62%) with a corresponding error rate of 14.38%, while Model 6 produced the worse accuracy. Model 14 had the highest sensitivity measure, while Model 16 had the least sensitivity measure. With regards to the precision measure, Model 15 was the most precise model with a precision of 0.7850, while Model 18 was the least precise model with a

precision of 0.6882. Model 15 recorded the highest F-measure of 0.7875, while the lowest F-measure was recorded by Model 6. The variation of performance measures with varying models is shown in Figure 9.

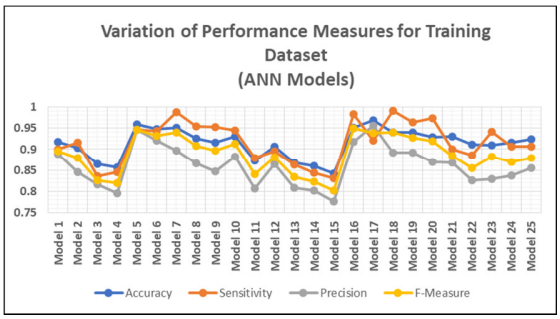

Fig. 8. Variation of performance measures for training dataset using ANN

TABLE VIII. RESULTS OF TESTING ANN

Model	Network Arch.		AC	ER	S	P	F
	Hidden Layer No.	No. of Neurons					
1	1	20	0.8005	0.1995	0.7900	0.7200	0.7534
2	1	15	0.8114	0.1886	0.7492	0.7587	0.7539
3	1	5	0.7896	0.2104	0.7524	0.7164	0.7339
4	1	3	0.8295	0.1705	0.7806	0.7781	0.7793
5	2	25-20	0.7872	0.2128	0.7210	0.7256	0.7233
6	2	20-25	0.7654	0.2346	0.7116	0.6900	0.7006
7	2	20-15	0.7836	0.2164	0.7304	0.7147	0.7225
8	2	15-20	0.7787	0.2213	0.7179	0.7112	0.7145
9	2	10-15	0.7944	0.2056	0.7586	0.7224	0.7401
10	2	15-10	0.7715	0.2285	0.7429	0.6890	0.7149
11	2	5-10	0.8198	0.1802	0.7680	0.7656	0.7668
12	2	10-5	0.7993	0.2007	0.7524	0.7339	0.7430
13	2	6-3	0.8174	0.1826	0.7774	0.7561	0.7666
14	2	3-6	0.8114	0.1886	0.8276	0.7233	0.7719
15	2	2-2	0.8356	0.1644	0.7900	0.7850	0.7875
16	3	30-20-25	0.8440	0.1560	0.6865	0.7252	0.7053
17	3	25-30-20	0.8256	0.1744	0.7398	0.7024	0.7206
18	3	20-15-20	0.8100	0.1900	0.8025	0.6882	0.7410
19	3	15-20-15	0.8300	0.1700	0.7837	0.7163	0.7485
20	3	15-10-15	0.8511	0.1489	0.7367	0.7460	0.7413
21	3	10-15-10	0.8532	0.1468	0.7712	0.7546	0.7628
22	3	5-10-5	0.8562	0.1438	0.7586	0.7586	0.7586
23	3	10-5-10	0.8457	0.1543	0.7524	0.7385	0.7453
24	3	6-4-2	0.8406	0.1594	0.8119	0.7379	0.7731
25	3	6-2-6	0.8340	0.1660	0.7868	0.7233	0.7538

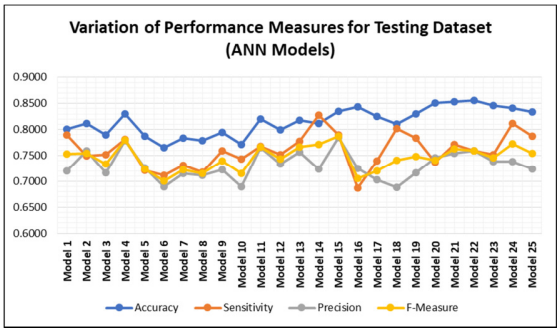

Fig. 9. Variation of performance measures for testing dataset using ANN

V. DISCUSSION

The study sought to develop classification models to predict injury severity of angle crashes involving two vehicles at unsignalized intersections using ANNs. A total of 3,307 reported crashes from 2008 to 2015 were extracted from a crash database and used in the analysis. Of the total number of crashes, 1,272 resulted in injury and/or fatality, while the remaining 2,035 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 which result in injury were Stop/Yield sign violation by one driver, and no violation on the part of the other driver.

The accuracy of classification models developed using ANN generally tends to increase as the number of hidden layers increases. Models with higher accuracies were attained with three hidden layers. Model 22 was the most accurate (85.62%) for predicting injury severity of angle crashes at unsignalized intersections. This model has 3 hidden layers with 5, 10, and 5 neurons respectively. The activation function in the hidden layers is the rectilinear unit function and the activation function in the output layer is the sigmoid function. The confusion matrix of this model is presented in Table IX. We can see that 51.5% of the crashes were correctly classified as non-injury crashes, while 10.3% were wrongly classified as injury crashes. Similarly, 29% of the crashes were correctly classified as injury crashes while 9.2% were wrongly classified as non-injury crashes. F-measure, is a combined measure for both precision and sensitivity. F-measures of the ANN models generally ranged between 0.7 and 0.8, and the higher values of F-measure were achieved with two hidden layers. Models 15 and 22 are the most accurate ANN models for predicting injury severity of angle crashes at unsignalized intersections.

TABLE IX. CONFUSION MATRIX OF MODEL 22

Total No. of Observations		Predicted	
		Negative	Positive
Actual	Negative	431	77
	Positive	77	242

VI. CONCLUSION AND RECOMMENDATION

In conclusion, the most accurate ANN model for predicting the severity of an injury sustained in a crash is a model with 3 hidden layers with 5, 10, and 5 neurons. The activation functions in the hidden and output layers are the rectilinear unit function and sigmoid function. This research explored the ANN machine learning technique. Future research can explore other techniques such as decision trees, K-nearest neighbors and linear discriminants. Also, other types of crashes can be explored at unsignalized intersections. Further, these analyses could be extended to signalized intersections.

REFERENCES

[1] T. R. Neuman, R. Pfefer, K. L. Slack, K. K. Hardy, D. W. Harwood, I. B. Potts, D. J. Torbic, E. R. K. Rabbani, National Cooperative Highway

Research Program: Guidance for Implementation of the AASHTO Strategic Highway Safety Plan, Transportation Research Board, 2003

[2] World Health Organization, Global Status Report on Toad Safety 2015, WHO, 2015

[3] National Highway Traffic Safety Administration, "USDOT Releases 2016 Fatal Traffic Crash Data", available at: https://www.nhtsa.gov/press-releases/usdot-releases-2016-fatal-traffic-crash-data, 2017

[4] National Highway Traffic Safety Administration, Traffic Safety Facts 2015, US Department of Transportation-National Highway Traffic Safety Administration, 2015

[5] B. J. Russo, P. T. Savolainen, W. H. Schneider, P. C. Anastasopoulos, "Comparison of factors affecting injury severity in angle collisions by fault status using a random parameter bivariate ordered probit model", Analytic Methods in Accident Research, Vol. 2, pp. 21-29, 2014

[6] R. Garrido, A. Bastos, A. de Almeida, J. P. Elvas, "Prediction of Road Accident Severity Using the Ordered Probit Model", Transport Research. Procedia, Vol. 3, pp. 214-223, 2014

[7] T. Sayed, F. Rodriguez, "Accident Prediction Models for Urban Unsignalized Intersections in British Columbia", Transportation Research Record Journal of the Transportation Research Board, Vol. 1665, No. 1, pp. 93-99, 1999

[8] W. Ackaah, M. Salifu, "Crash prediction model for two-lane rural highways in the Ashanti region of Ghana", International Association of Traffic and Safety Sciences Research, Vol. 35, No. 1, pp. 34-40, 2011

[9] M. Y. Lau, A. D. May, Accident Prediction Model Development: Signalized Intersections, Institute of Transportation Studies, University of California-Berkeley, 1988

[10] A. Kamer-Ainur, M. Marioara, "Errors And Limitations Associated With Regression And Correlation Analysis", Statistics and Economic Informatics, pp. 710-712, 2007

[11] P. Chengye, P. Ranjitkar, "Modelling Motorway Accidents using Negative Binomial Regression", Journal of the Eastern Asia Society for Transportation Studies, Vol. 10, pp. 1946-1963, 2013

[12] Z. Yang, L. Zhibin, L. Pan, Z. Liteng, "Exploring contributing factors to crash injury severity at freeway diverge areas using ordered probit model", Procedia Engineering, Vol. 21, pp. 178-185, 2011

[13] Federal Highway Administration, "Highway Safety Improvement Program Manual–Safety", available at: https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec6.cfm, 2011

[14] G. Dutta, P. Jha, A. K. Laha, N. Mohan, "Artificial Neural Network Models for Forecasting Stock Price Index in the Bombay Stock Exchange", Journal of Emerging Market Finance, Vol. 5, No. 3, pp. 283-295, 2006

[15] M. H. Hassoun, Fundamentals of Artificial Neural Networks, MIT Press, 1995

[16] S. Sharma, "Activation Functions in Neural Networks", available at: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6, 2017

[17] F. R. Moghaddam, S. Afandizadeh, M. Ziyadi, "Prediction of accident severity using artificial neural networks", International Journal of Civil Engineering, Vol. 9, No. 1, pp. 41-49, 2011

[18] K. S. Jadaan, M. Al-Fayyad, H. F. Gammoh, "Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN)", Journal of Traffic Logistics Engineering, Vol. 2, No. 2, pp. 92-94, 2014

[19] Office of the State Superintendent of Education, "New U.S. Census Bureau Numbers Officially Put DC's Population Over 700,000", available at: https://osse.dc.gov/release/new-us-census-bureau-numbers-officially-put-dc%E2%80%99s-population-over-700000, 2018

[20] T. Winship, "The 10 US cities with the worst traffic", available at: https://www.businessinsider.com/the-10-us-cities-with-the-worst-traffic-2018-2, 2018

[21] District Department of Transportation, "DDOT by the Numbers", available at: https://ddot.dc.gov/page/ddot-numbers

[22] American Society of Civil Engineers, Repord Card for D.C.'s Infrastructure, ASCE, 2016