Jestr

Journal of Engineering Science and Technology Review 12 (4) (2019) 15 - 27

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Investigating Key Factors Influencing the Severity of Drivers Injuries in Car Crashes Using Supervised Machine Learning Techniques

Mohammad Fuad Aljarrah[1], Mohammad Ali Khasawneh[1,2,*] and Aslam Ali Al-Omari[1]

[1]*Civil Engineering Department, Jordan University of Science and Technology, Irbid, Jordan.*
[2]*Civil Engineering, Prince Mohamad Bin Fahd University, College of Engineering/Civil Engineering Department, Al Azeziya, Eastern Province, Kingdom of Saudi Arabia*

Received 16 April 2019; Accepted 14 September 2019

Abstract

Limiting number of fatalities and reducing injury severity of car crashes is a continuing global concern. This study investigates crash risk factors including driver, vehicle, roadway, and crash characteristics role in determining injury severity levels encountered by drivers in fatal car crashes. Three types of supervised machine learning techniques were used; Classification and Regression Tree (CART), Artificial Neural Networks (ANN) and Multinomial Logistic Regression. The CART model was used to elect the most influential factors in determining drivers' injury severity levels. The ANN model was used to predict drivers' injury severity based on crash attributes. The logistic model was used to identify the effect of different crash attributes in distinguishing drivers' injury severity levels and for comparison purposes. Consequently, CART model resulted in six significant factors, these factors are: airbag deployment, seatbelt use, driver age, vehicle rollover, collision type, and vehicle model year. It was found that airbag deployment has a strong correlation with severe injuries and even fatalities. The use of seatbelts appears to reduce injuries and fatalities. Furthermore, elderly drivers, front to front collisions, vehicle rollover and older vehicles seem to cause more mortalities and injuries. On the other hand, according to the logistic model, all the crash attributes were found significant in distinguishing between drivers' injury severity levels except for the roadway functional system. However, the ANN model outperformed the CART model in terms of accuracy and stability. Further, both models seem to outperform the logistic regression model in terms of prediction accuracy.

Keywords: Classification and Regression Tree (CART), Artificial Neural Networks (ANN), Multinomial Logistic Regression, Severity, Injury, Fatality, Crashes, Fatality Analysis Reporting System (FARS)

1. Introduction

Worldwide, around 1.3 million lives are lost, and approximately 50 million people are injured each year due to traffic crashes. According to data collected from 178 countries by the World Health Organization, traffic crashes are considered the ninth most common cause of death [1]. In 2016, around 40,000 lives were lost on U.S roads, an increase of 5.6% from the year 2015 [2].

Several factors influence crashes injury severity. These factors are mostly related to one or more of the following: driver characteristics, vehicle characteristics, roadway characteristics, crash characteristics and atmospheric factors [3,4].

Limiting the number of fatalities and reducing injury severity in car crashes is a continuing concern within the traffic safety field. Investigating key risk factors of traffic crashes helps determine the significant factors that need immediate attention by governments and transportation agencies in order to eliminate or at least minimize crashes number and severity.

Although statistical models have been largely employed in crashes injury severity detection [5-10], other superior supervised machine learning techniques such as decision trees and artificial neural networks have proven to provide higher prediction accuracy than ordinary statistical models [11-13]. Nevertheless, the use of a softmax activation function in the neural network design has not been broadly employed in the crash injury severity detection field. Moreover, neural network designers usually use either early stopping or regularization method to improve model generalization abilities. The use of both techniques in the same model has not been widely implemented either.

Therefore, this study used 7,394 car crashes occurred across the United States of America in the year 2015 to investigate accident risk factors including driver, vehicle, roadway, and crash characteristics role in determining injury severity levels encountered by drivers in fatal car crashes. Data was retrieved from the Fatality Analysis Reporting System (FARS). Three types of supervised machine learning techniques were utilized for this purpose; Classification and Regression Trees (CART), Artificial Neural Networks (ANN) and Multinomial Logistic Regression. The CART model was used to elect the most influential factors in determining drivers' injury severity levels. On the other hand, the multi-layer feedforward ANN model with a softmax activation function and both early stopping and regularization generalization techniques was used to predict drivers' injury severity based on crash attributes. The utilization of a softmax

*E-mail address: mkhasawneh@just.edu.jo
ISSN: 1791-2377
doi:10.25103/jestr.124.03

function in the ANN output layer was not presented in the literature; therefore, this study also aims to examine its role in the ANN training. The logistic model was used to identify the effect of different crash attributes in distinguishing between drivers' injury severity levels and for comparison purposes. By determining key risk factors of car crashes and building prediction models, safety procedures can be carried out and policies can be implemented in order to reduce fatalities and injuries caused by car crashes.

2. Literature Review

Several analytical techniques have been used to examine crashes attributes. One of the most famous fields in data analysis is data mining. Vehicle crashes data can be investigated by employing different data mining techniques including statistics, machine learning, high-performance computing, artificial neural networks, pattern recognition, decision trees, data visualization, image and signal processing and spatial data analysis [14,15]. Such techniques are able to detect relationships and patterns between inputs and outputs with a decent accuracy level.

2.1 Statistical Models on Crash Injury Severity

Statistical analysis has been the most popular technique to examine risk factors affecting crashes. Zhang et al. [5] studied crashes involving elderly drivers in Ontario, Canada using a multivariate unconditional logistic regression model. It was found that male drivers older than 75 years old were the most vulnerable class to face severe injuries due to their physical conditions. Keall et al. [6] used a logistic model to study the effect of drivers' alcohol consumption, their age, and the influence of other passengers on drivers' risk of being fatally injured in New Zealand. The model showed high relation between blood alcohol concentration and fatality risk. Bedard et al. [8] employed a multivariate logistic regression to investigate the contribution of driver, vehicle, and crash characteristics to driver fatality. It was found that increasing seatbelt use and reducing travel speed may reduce drivers' fatalities. It was also found that older drivers and female drivers need more attention than younger and male drivers.

A group of researchers explored some methodological barriers existing in crash data statistical analyses that need profound attention such as missing data, unobserved heterogeneity, endogeneity, risk compensation, temporal instability, spatial and temporal correlations [10,16,17]. According to these studies, it is believed that new methodological applications like random parameter models, latent class models and multi-state switching models have great potential in improving the understanding of several factors that could affect the likelihood and the injury severity level of highway crashes.

2.2 Decision Trees and Artificial Neural Networks on Crash Injury Severity

Decision Trees and Artificial Neural Networks have proven to outperform ordinary statistical modeling in terms of prediction power. These techniques can easily detect non-linear relationships between inputs and outputs.

Chang and Wang [3] employed a CART model to investigate crashes risk factors in Taiwan. It turned out that most influential factors were manner of collision, contributing circumstances, and driver behavior. However, the proposed CART model failed to identify conditions that resulted in a fatal injury, thus it was proposed to associate the proposed model with other data mining techniques such as artificial neural networks. Another Taiwanese study used CART modeling focused on truck-involved crashes was carried out by [18]. It was concluded that drunk driving is the most significant factor in these types of crashes, followed by seatbelt use, collision type, vehicle type, contributing circumstances, and drivers' actions.

Kashani and Mohaymany [19] used a CART classifier to investigate the factors influencing crash injury severity for crashes occurring on two-lane two-way rural road in Iran. It was found that not using a seatbelt is the most influential factor that affects the severity of an injury. In a recent work, [20] used a CART model to identify the most significant factors that affect car crashes severity and injury for crashes on Slovenian Roads. It was concluded that the most vital factor in predicting the crash severity is the contributing circumstances specially speeding and driving on the wrong lane.

Tong et al. [12] applied an L1 penalized logistic regression and binary decision tree to predict factors causing fatal injuries in crashes occurred in Virginia between 2010 and 2015. An L1 penalized logistic regression is a type of logistic regression which enforces a penalty (punishment) to prevent the model from having "too many variables", the L1 type refers to the Lasso regression ("Least Absolute Shrinkage and Selection Operator"); which sums the "absolute value of magnitude of coefficients" as a penalty term to the loss function. In this study, both models predicted different set of variables to be risky; factors such as lighting condition, speed limit, roadway alignment, number of vehicles and type of intersection were found significant when using the penalized logistic regression model. On the other hand, weather condition, driver drinking and drug using, lighting condition, and type of collision were found significant when the decision tree model was used.

In a recent study [13], several prediction models using decision trees and logistic regression classifiers were developed. These models aimed to describe factors that contribute the most to non-fatal crashes in Malaysia. Five different types of decision trees were used namely Gini, Entropy, CART, Logworth, and CHAID. Multinomial logistic regression was employed to compare the results. It was shown that the decision tree with CART algorithm outperformed the other types of decision tree and the logistic regression model. It was also concluded that crash cause, road geometry, vehicle type, age, and collision type are factors that significantly contribute to these types of crashes.

Chen et al. [21] used classification and regression tree along with Support Vector Machine (SVM) model to investigate driver injury severity in rollover crashes in New Mexico. SVMs are a non-parametric kernel based classifiers that are used to investigate patterns in classification problems. The CART model was used to detect the significant variables causing injuries while the SVM classifier was utilized to investigate the injury severity patterns. It was concluded that wearing a seatbelt, driving in a comfortable environment, alcohol and drug use are highly associated with incapacitating and fatal injuries.

Feedforward ANNs are vastly used for prediction. Several studies used this particular type of ANN to investigate crashes risk factors. To name a few, [22] used a feedforward backpropagation ANN model to predict the occurrence of injuries in vehicle collision crashes in Florida. It was found that number of lanes and road surface conditions contribute the most to the injury occurrence. A study by [23] considered vehicle crashes occurred in Turkey using feedforward

backpropagation ANNs. The model revealed that the degree of vertical curvature is the most important factor that influences the number of crashes followed by the degree of the horizontal curve.

Other types of ANN have also been used for risk factor investigation. Abdelwahab and Abdel-Aty [11] utilized multilayer perceptron and fuzzy adaptive resonance theory ANN models. It was found that female drivers are more prone to severe injuries than males. Further, increased speed and not wearing a seatbelt increase the probability of severe injuries. Delen et al. [24] utilized a series of ANNs to model the relationship between risk factors causing injuries in car crashes and the injury severity levels. It was found that the use of seatbelts, drivers' age and gender and their alcohol and drug consumption seem to have large contribution to injuries severity.

Sohn and Shin [25] used three data-mining techniques; ANNs, logistic regression, and classification trees to determine the influential factors for traffic crashes occurred in South Korea and to build prediction models for injury severity. It was found that there is no significant difference in the accuracy between the three techniques used. It was also found that the use of a protective device is the most influential factor. Chong et al. [26] employed ANNs and decision trees to study traffic crashes data. Decision tree model was found to outperform neural networks in terms of prediction accuracy. It was also shown that the most significant factors in fatal injuries are drivers' seatbelt use, alcohol intake and roadway light condition.

Many other studies employed other machine learning techniques to investigate crashes risk factors and predict injury severity levels. Some of these tools include Bayesian networks [27] modified neural network pruning algorithm N2PFA [28], recurrent neural networks [29], data clustering [30,31], and non-dominated Sorting Genetic Algorithm with a neural network classifier [32].

3. Data Compilation and Reduction

The data used in this study was retrieved from the Fatality Analysis Reporting System (FARS). This database system is maintained by the National Highway Traffic Safety Administration (NHTSA). FARS is a census tool used in the United States which provides data regarding fatal injuries in motor vehicle traffic crashes.

To consider a crash as fatal, at least one fatal injury shall be reported. Fatal injuries include all deaths within 30 days of the crash date [33]. FARS database contains comprehensive information on persons and vehicles involved in the crash besides crash circumstances. These data points are collected from police reports, state administrative files, and medical records by specialists. Moreover, automated error examination and data monitoring guarantee that data fall within logical ranges [34].

In this study, sixteen accident attributes were utilized as input variables to assess driver injury severity levels in fatal traffic crashes occurred across the United States of America (USA) in the year 2015. Studied variables include driver characteristics such as age, gender, use of seatbelt and alcohol consumption. In addition, vehicle characteristics such as model year and air bag deployment were included, and roadway features such as the type of roadway, roadway alignment, surface conditions, lighting condition, speed limit and whether the crash occurred on an intersection or not. Finally, crash features like travel speed, manner of collision, ejection and vehicle rollover were also considered as input variables. On the other hand, injury severity level was used as the target variable.

The 16 input attributes were chosen based on previous studies such as [3,11,13,18,21,24,35] and based on the availability of data records on FARS website. Fortunately, choosing this many attributes would make comparisons with other studies that focused on only one or more of these attributes easier and more comprehensible. Further, other attributes are missing some values, thus making the utilization of these attributes somewhat cumbersome. These missing data points cannot be easily predicted or interpolated, thus, choosing these 16 variables led to a dataset that presented sufficient attribute values.

It should be noted that this study focused on drivers' injury solely. The rationale behind this is to reduce noise in dataset. In some cases, different injury severities resulted for the same vehicle occupants; where all occupants share at least ten crashes attribute values, this could lead to misleading conclusions. Thus, limiting the scope of this research to drivers only reduced noise in dataset and yielded better accuracy in prediction.

Data had 7,394 crash records occurred in the US in 2015. These records are originally coded into a predefined coding system developed by FARS database. However, using data as it is could lead to incorrect conclusions, and could affect models' development. Therefore, all attributes were re-coded in a way that is easily comprehended by users and consistent. Table 1 shows the categorical variables used in this study, their coding and some descriptive statistics.

Another point worth mentioning is that for the neural network model, the target variable was coded using the "One Hot Encoding" system which converts numerical values into a binary system (zeros and ones). Hence, the number of bits in the binary number is equivalent to the number of categories of the injury severity variable which is three. Therefore, the No Injury Category will be coded as [1 0 0], the Injury Category is coded as [0 1 0] and the Fatality Category is coded as [0 0 1].

The following Figures, 1 through 3, illustrate crashes distribution according to the continuous variables used in the study. Driver age (years), speed limit (mph), and travel speed (mph) were rounded to the nearest 5.

Figure 1 shows drivers' distribution by age. The Figure indicates a decreasing trend as age increases, which means that the majority of roadway users are young drivers. In a study done by the AA insurance company, it was found that around 25% of young drivers' crashes occur in the first six months of issuing their driving license, and 40% of them crash by the age of 23 [36]. Figure 2, on the other hand, shows that most crashes occurred on sections where speed limits are 45 and 55 mph. This does not reflect the danger of these speed limits; it rather indicates that a great portion of American roads are assigned these speed limits. In addition, Figure 3 reveals that 1,335 out of the 7,394 crashes include stationary vehicles such as in cases where drivers tend to drive at low speeds close to zero mph in urban areas, in traffic jams, or when standing in a queue at signalized intersections for instance. Finally, 1,867 out of the 7,394 crashes happened at speeds of 45 and 55 mph, which is consistent with the speed limit distribution in Figure 2.

4. Methodology

4.1 Classification and Regression Trees

Decision trees are one type of supervised machine learning techniques that are commonly used to construct classification

models. A Classification and Regression Tree (CART) is an example of decision trees which are used when the dependent variable takes a finite number of distinctive values [37].

A decision tree is a hierarchal structure that grows down in the shape of connected nodes. Each node represents an input variable where this input node splits down into its attribute values. Each attribute value will be connected to a new node that represents another variable. The new node will also be split into its attribute values. The first node (variable) to be chosen is called a root (parent) node. Any following nodes will be called an internal (child) nodes. At each splitting process, the resulting child nodes become parents their successor child nodes. The splitting process continues until a class (category) is determined. The last node showing the class of the outputs is called the terminal (leaf) node.

Table 1. Categorical Variables Descriptive Statistics

Variable	Code	Descriptive Statistics	Variable	Code	Descriptive Statistics
Driver Injury Severity			**Horizontal Alignment**		
No Injury	0, [1 0 0]	29%	Straight	0	87%
Injury	1, [0 1 0]	41%	Curved	1	13%
Fatality	2, [0 0 1]	30%	**Surface Conditions**		
Driver Gender			Dry	0	89%
Male	1	65%	Wet/Icy	1	11%
Female	0	35%	**Lighting Conditions**		
Seatbelt Use			Daylight	1	66%
Used	1	85%	Dark, Lighted	2	14%
Not Used	0	15%	Dark, Not Lighted	3	20%
Drunk Driving			**At Intersection**		
Drunk	1	9%	Yes	1	28%
Not Drunk	0	91%	No	0	72%
Model Year			**Manner of Collision**		
1980-2000	1	16%	Front to Rear	1	26%
2001-2005	2	25%	Front to Front	2	24%
2006-2010	3	25%	Angle	3	41%
2011-2017	4	34%	Sideswipe	4	9%
Airbag Deployment			**Driver Ejection**		
Deployed	1	59%	Ejected	1	4%
Not Deployed	0	41%	Not Ejected	0	96%
Roadway Type			**Vehicle Rollover**		
Local Road	1	6%	Rollover	1	8%
Collector	2	13%	No Rollover	0	92%
Minor Arterial	3	18%			
Major Arterial	4	47%			
Interstate	5	16%			

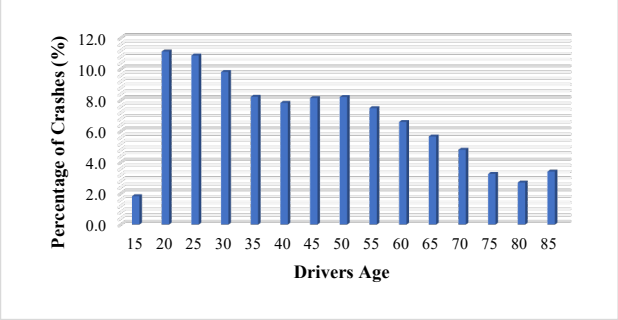

Fig. 1. Crashes Distribution According to Driver Age

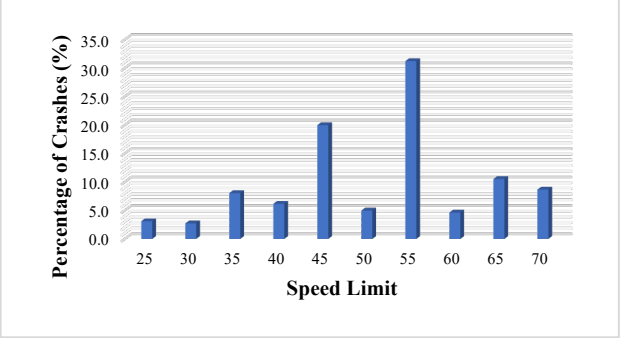

Fig. 2. Crashes Distribution According to Speed Limit

The CART model was developed using the Statistical Package for Social Sciences (SPSS) software. The aim of using such technique is to investigate the contribution of crash risk factors in determining drivers' injury severity in fatal car crashes. The Gini impurity measure (the default splitting criterion in CART) was chosen in this study. For a given node (t), the Gini index is calculated using Equation 1:

$$\text{Gini (t)} = \sum_{j=1}^{n-1} p(j|t))\times(1- p(j|t))=1- \sum_{j=1}^{n-1} [p(j|t)^2] \quad (1)$$

where $p(j|t)$ is the class distribution or the relative frequency of class j at node t, and n is total number of classes.

To determine the efficiency of the Gini impurity measure, the difference between the degree of impurity of the parent nodes and their child nodes will be calculated. The larger the difference the better the test condition, thus, a better candidate for the next split. This can be expressed numerically by calculating the gain value Δ as in Equation 2:

$$\Delta = I_{parent} - \sum_{i=1}^{k} \frac{N(v_i)}{N} I(v_i) \quad (2)$$

where I is the impurity measure for any given node, $N(v_i)$ is the number of records in the child node i for all different target classes, N is the number of records in the parent node for all

different target classes and k is the number of attributes classes.

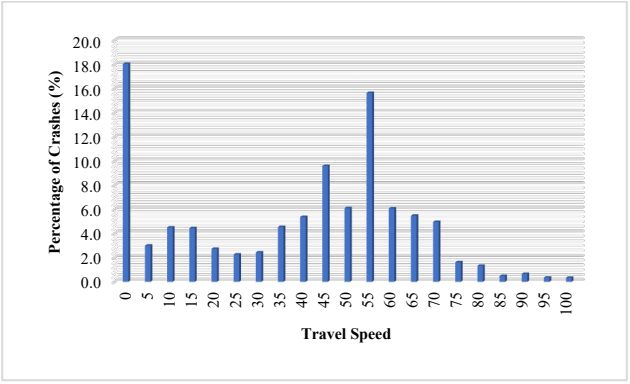

Fig. 3. Crashes Distribution According to Travel Speed

Decision trees algorithm usually chooses a test condition that maximizes the value of the gain ratio Δ. Since $I_{(parent)}$ is the same for all child nodes, maximizing gain value is equivalent to minimizing the impurity measure for the child nodes.

The Gini index tends to favor inputs variables with more attribute values (e.g. a variable that has ten values is preferred over a binary variable) [38]. Moreover, variables with larger number of attribute values usually have a smaller number of records for each class label, which will lead to unreliable predictions. To overcome this problem, CART algorithm adopts binary splitting for the attributes.

For tree growth termination, a minimum number of 50 observations in parent nodes and 25 observations in child nodes were chosen to stop tree expansion, i.e. if less than 25 observations are seen in a child node the tree will stop growing and the node will become a terminal node. The dataset was divided into 70% training dataset and 30% testing datasets to estimate the generalization error. These values were determined by trial and error corresponding to a minimum generalization error and maximum model prediction. Moreover, to overcome overfitting, a maximum tree depth of 5 branches was used, plus a post-pruning technique; where trimming of a fully-grown tree moves upward was implemented. Post-pruning is believed to be preferred over pre-pruning, mainly because post-pruning relies on the fully-grown tree to decide whether to trim or not, while pre-pruning can result in premature termination of the tree growing operation [38].

Generally, in a classification problem, a confusion matrix showing the relationship between the observed and the predicted outputs is used in order to fully realize the performance of the classifier. A confusion matrix delivers detailed information on how output data are classified by the model. The matrix consists of rows and columns for each category of the target variable class. The categories represented in the columns are the predicted categories of the target variable while the categories shown in the rows are the actual categories. The numbers in each cell represent the number of predictions for each class. Therefore, the diagonal cells represent the correctly classified cases (where the predicted category matches the actual category provided to the network). Conversely, the other cells represent the incorrectly classified classes.

4.2 Artificial Neural Networks

ANNs are combination of parallel artificial neurons connected by weighted links. Neurons are arranged in layers, these layers are: the input layer, the hidden layer(s), and output layer. ANNs algorithm starts by multiplying the input values stored in the input layer neurons by the weights associated to these neurons, next, the sum of the previous values will pass through an activation function which will produce a value that is an input for the following layer [39]. The output of any given layer will be the input of the following layer. The process continues until the output layer is reached. This process is called feedforward. The mathematical operation of the feedforward process is shown in Equation 3:

$$Y = f(\vec{X} \,.\, \vec{w} + \vec{b}) \tag{3}$$

where $\vec{X}$ refers to input vector and $\vec{w}$ refers to weight vector, $\vec{b}$ is the bias vector at each neuron, f is the activation function at that layer and Y is the output of the neuron. The weight matrix $\vec{w}$ in Equation 3 has the following form shown in Equation 4:

$$\vec{w} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \tag{4}$$

where (n) is the number of inputs, and (k) is the destination neuron associated to the weights.

Subsequently, the outputs of the neural network are compared to the target data which was fed to the network along with the inputs. The difference between the resulting output and the fed target represents error. Errors at each output neuron are computed and utilized to adjust the weights connected to the neurons. This process is called backpropagation. The feedforward backpropagation process will continue until the error between the targets and the network outputs is optimized (minimized). The error at the output of neuron k at iteration p is defined in Equation 5:

$$e_k(p) = d_k(p) - O_k(p) \tag{5}$$

where $e_k(p)$ is performance function or error function, $d_k(p)$ is desired or target value for the neuron k at iteration p, and $O_k(p)$ is the output of the neuron k at iteration p. The algorithm for updating the weights at the output layer is the perceptron learning rule shown in Equation 6:

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \tag{6}$$

where $w_{jk}(p+1)$ is the modified weight at iteration $(p+1)$, $w_{jk}(p)$ is the weight at epoch p and $\Delta w_{jk}(p)$ is the weight correction term at epoch p. The weight adjustment in the multilayer network is computed using Equation 7:

$$\Delta w_{jk}(p) = \alpha \times O_j(p) \times \delta_k(p) \tag{7}$$

where $O_j(p)$ is output of the hidden layer at iteration p which is equal to $X_k(p)$; the input to the output layer (α) is the learning rate, and $\delta_k(p)$ is the error gradient at neuron k in the output layer at iteration p. The learning rate represents how quickly a neural network modifies its weights and biases. The error gradient is the derivative of the activation function multiplied by the error at the neuron output which is shown in Equation 8:

$$\delta_k(p) = \frac{\partial O_k(p)}{\partial X_k(p)} \times e_k(p) \tag{8}$$

where e_k (p) is the error function output in neuron k in the output layer at iteration p, $X_k(p)$ is an input to the k neuron in the output layer at iteration p, and $O_k(p)$ is the output of k neuron in the output layer at iteration p. For a sigmoid activation function in the hidden layer and a softmax function in the output layer, the sigmoid function and the softmax function are shown in Equations 9 and 10, respectively:

$$O_j(p) = \frac{1}{1 + e^{-X_j(p)}}, \text{ at the hidden layer neurons} \quad (9)$$

$$O_k(p) = \frac{e^{X_k(p)}}{\sum_{n=1}^{N} e^{X_n(p)}}, \text{ at the output layer neurons} \quad (10)$$

Where $O_j(p)$ and $O_k(p)$ are outputs of the hidden and output neurons, respectively. $X_j(p)$ and $X_k(p)$ are inputs to the hidden and output neurons, respectively.

It is also crucial to normalize the input and output values to the same order of magnitude [40]. If inputs and outputs have diverse ranges, some variables may seem more significant than others which will lead to faulty conclusions. In this study, all the data used were scaled from (0 to 1) for both input and output values using the linear interpolation formula that was suggested by [41] and shown in Equation 11:

$$Y_i = \frac{(Y_{max} - Y_{min})(X_i - X_{min})}{(X_{max} - X_{\min})} + Y_{min} \quad (11)$$

where $Y_{min} = 0$, $Y_{max} = 1$, and X is the range of data to be scaled. This equation is already stored in MATLAB library as a function called "*mapminmax*" that will be used in this study to scale the dataset.

Finding the best formulation of the neural network model was an iterative process which involved modifying the number of hidden neurons until the best performance is reached. After several set of iterations, the final neural network model consisted of 50 neurons in the hidden layer. Biases were connected to both hidden and output layer. In this study, both early stopping and regularization generalization techniques were used to achieve the best generalization performance and to overcome overfitting. The reason behind using both techniques is because regularization alone resulted in an over-fitted model. Ten validation checks were used to terminate the learning algorithm using the early stopping technique. Moreover, the dataset was divided into three categories: 70% training dataset, 15% validation dataset, and 15% testing dataset. The training set is used to compute errors and gradients and to adjust weights of iterations. The validation set is used as an indicator of what is happening to the network function in between the training points. The testing dataset is set aside and not introduced to the training process and used after the training is finished to check the generalization ability of the trained network. Figure 4 shows the final network design with the activation function of each layer.

Nevertheless, the prediction accuracy using a confusion matrix alone does not entirely describe the efficiency of the neural network model. Thus, other means of assessing the model's efficiency are necessary. The receiver operating characteristics (ROC) curve, which is also known as the relative operating characteristic curve, is a graphical plot that demonstrates the classifying ability of a classifier as its discrimination threshold is varied. It is represented by plotting the fraction of true positives (TPR = True Positive Rate) versus the fraction of false positives (FPR = False Positive Rate). The area under the ROC curve measures the overall classification ability of a test. An entirely random test has an area under curve of 0.50, while a perfect test has an area under curve of 1.00.

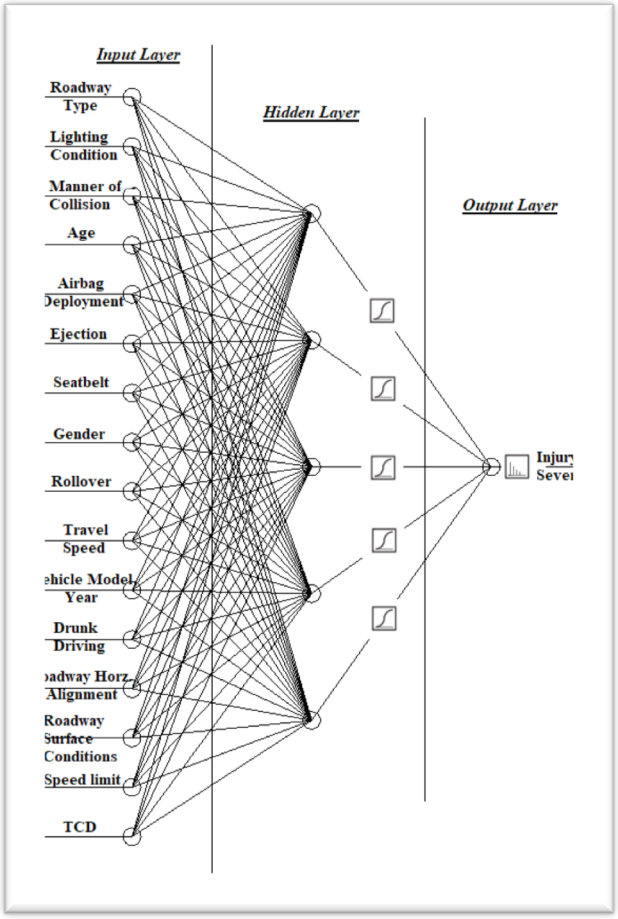

Fig. 4. The Complete Architecture of The Neural Network Model

4.3 Multinomial Logistic Regression

Multinomial Logistic Regression is a statistical tool used to predict a categorical dependent variable that has more than two categories using a set of independent variables. These variables can be categorical, continuous, or a mix of both. The assumptions of linearity, normality and homoscedasticity are not considered in logistic regression [13,42]. The p-value of likelihood ratio chi-square test is used to determine model significance. The Cox and Snell and Nagelkerke Pseudo R-Square measures are also used to calculate the significance of the model. For the variable to be significant, the p-value of the Wald Chi-Square test must be less than 0.05. The odds ratio is defined as the odds of a certain outcome relative to the odds of another outcome. It is used to assess the risk of a certain output if another factor is present. The odds of a certain event occurring are the probability of that event happening divided by the probability of that event not happening, as in Equations 12 and 13:

$$\text{Odds ratio for Event 1} = \frac{P(E_1)}{1 - P(E_1)} \quad (12)$$

$$\text{Odds ratio for Event 2} = \frac{P(E_2)}{1 - P(E_2)} \quad (13)$$

where $P(E_1)$ is the probability of event 1 occurring and $P(E_2)$ is the probability of event 2 occurring. Then using Equation 14 one can calculate the odds ratio between event 1 and 2:

odds ratio for event 1 / odds ratio for event 2 (14)

5 Results and Discussion

5.1 Classification and Regression Trees

In order to fully comprehend the performance of the CART model in predicting injury severity, a confusion matrix showing the relation between the observed injury severity and predicted injury severity for the training and testing datasets is summarized in Table 2. The model showed good classification abilities with training and testing prediction accuracies of 61.4% and 59.5%, respectively. For the individual severity categories, Table 2 shows prediction accuracy for each class for no injury, injury and fatality classes for both training and testing models. For the training model, 75.1% of the no injury class, 59.6% of the injury class and 51.1% of the fatality class were correctly classified. In the testing model, 74.4% of the no injury class, 56.5% of the injury class and 48.8% of the fatality class were correctly classified. Both models show good prediction abilities compared to literature. It is also noted that the training and the testing models have close prediction accuracies which indicates high generalization abilities and no overfitting.

Table 2. CART Model Confusion Matrix

Sample	Observed	Predicted			
		No Injury	Injury	Fatality	Percent Correct
Training	**No Injury**	**1108**	294	74	75.1%
	Injury	494	**1299**	387	59.6%
	Fatality	170	597	**800**	51.1%
					61.40%
Testing	**No Injury**	**486**	140	27	74.4%
	Injury	217	**478**	151	56.5%
	Fatality	80	264	**328**	48.8%
					59.5%

Figures 5 and 6 below show the output of the CART model using training and testing datasets, respectively. It is obvious that only six attributes out of the 16 inputs were identified as influential factors in determining the injury severity. These factors are: airbag deployment, seatbelt use, driver age, vehicle rollover, collision type, and vehicle model year. These findings are consistent with [3,11-13,18,19,21,24,26,32,35]. At least one of the factors which were found significant in this study was also significant in these studies.

Referring to Figures 5 and 6, the first factor which was chosen as the primary splitter in the CART model is airbag deployment. Surprisingly, data of current study showed that most fatalities and injuries have occurred in crashes where airbag was deployed (node 1 and 2). These results are consistent with [43], where in their study, a great correlation was found between airbag deployment and minor/major injuries. This finding warrants comprehensive studies regarding airbags adverse effect on car crashes in order to reduce injuries, prevent fatalities, and protect occupants from being harmed.

The use of seatbelts was found essential in reducing fatalities and injuries; it is clearly seen in nodes 3, 4, 5 and 6 that not wearing a seatbelt resulted in higher number of fatalities and injuries especially in crashes where airbag was deployed. This was also addressed by [44], where it was found that drivers not wearing seatbelts faced flexion injuries in the spine, and some severe fractions in the sternum and facial bones. Another interesting finding is shown in node 4; it appears that use of seatbelts with no airbags significantly reduced fatalities and injuries. This reflects the great role of seatbelts in protecting occupants.

Drivers' age is as of similar importance; where it was found that elderly drivers (aged 72 and above), had faced fatal and severe injuries compared to younger drivers. Although elderly drivers comprise only 14% of our data sample, their association to fatal and severe injuries was strong. This is shown in nodes 9 and 10 and nodes 11 and 12. This finding requires a move from legislatures and decision makers; where it should be recommended that there would be a maximum driving age to prevent such injuries to happen. This also forces car manufacturers to increase safety precautions and measures to be more suitable for elderly people.

Vehicle rollover was also found significant contributor to fatalities and severe injuries. Despite the fact that only 9% of traffic crashes data included vehicle rollovers, it was found that rollover (node 8) increases the risk of severe and fatal injuries even when seatbelt is used. Consequently, 94% of the cases where rollover occurred had either a severe injury of a fatality. This factor is associated with striking the occupant's body with the internal parts of the vehicle. It is also considered as an impediment to rescue teams when extracting the victims.

The manner of collision was also found significant; where front to front crashes (node 13) led to the highest rates of fatalities and severe injuries compared to front to back, angled, and sideswipe crashes (node 14). This seems logical because in front to front crashes especially in high speeds, the driver's face strikes the steering wheel causing high pressure on the brain leading to internal trauma. This also is related to basic physics; where the momentum due to front to front impact is the highest compared to the other type of crashes

The final factor that was found to be significant is the vehicle model year. It was found that vehicles of model year higher than 2000 (node 16) had lower fatality and injury rates. This makes great sense since newer models usually have higher safety measures with higher danger sensing accuracy and better restrains system and airbag technologies.

It is also likely that some types of crashes are not adequately represented in the dataset. This can be obvious when analyzing factors such as driving under the influence of alcohol, roadway surface conditions, roadway horizontal alignment, drivers' ejection and whether the crash occurred at an intersection or not. If we take drunk driving for example, this factor is considered one of the most significant factors in causing severe and fatal injuries. However, the CART model

could not distinguish its sole effect and thus, it was not considered as a significant factor. This finding is considered surprising. Nevertheless, by analyzing the dataset, only 668 crashes (9% of the data) included cases where driver was under influence.

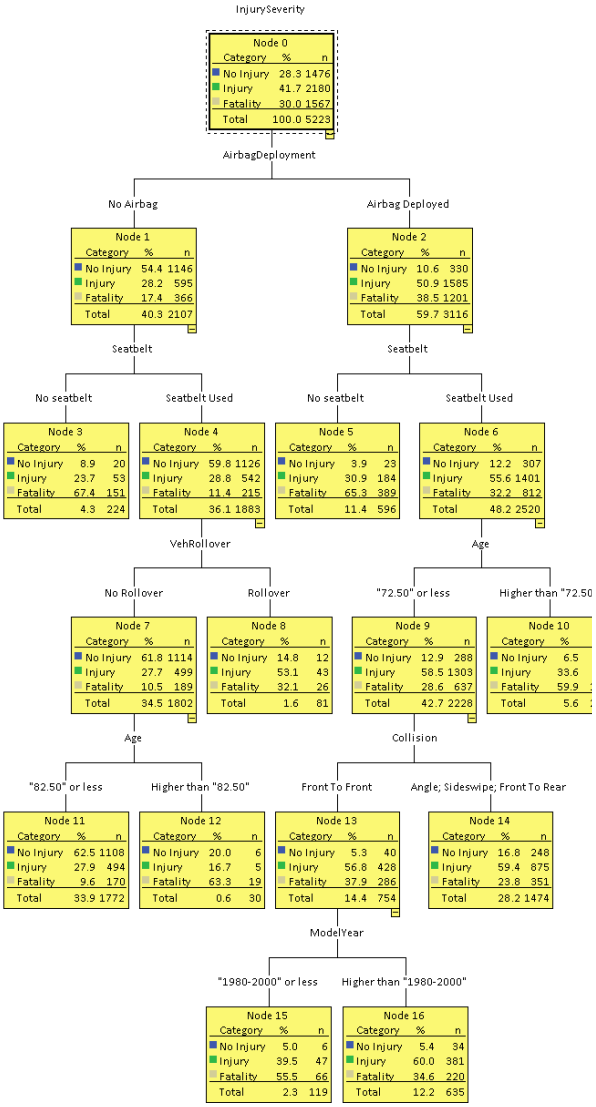

Fig. 5. The Output of the CART Model Using Training Dataset

5.2 Artificial Neural Networks

The ANN training algorithm showed convergence after reaching a constant number of effective parameters (number of used weights and biases) of 270. The sum squared parameter (sum squared of weights and biases values) also converged to a constant value of 99.5. Further, the gradient (δ) and the learning rate (α) parameters have converged to constant values of 0.0119 and 24, respectively, and the final recorded validation mean square error was 0.2868. This indicates stability in the performance of the neural network model.

Comparing the CART model, the ANN model showed very good classification abilities with training, validation and testing prediction accuracies of 64.8%, 61.1% and 63.5%, respectively. For the individual severity categories, Table 3 shows prediction accuracy for each class of no injury, injury and fatality for both training and testing models. For the training model, 72.6% of the No Injury class, 64.8% of the Injury class and 57.7% of the Fatality class were correctly classified. The validation model showed good prediction accuracy as well; where 69.3% of the No injury class, 60.5% of the Injury class and 54.0% of the Fatality class were also correctly classified. The testing model also showed great generalization abilities, where 69.8% of the No Injury class, 63.7% of the Injury class and 56.3% of the Fatality class were correctly classified. The model also showed good generalization abilities; where the training, validation and testing prediction accuracy were close. The matrix also reveals that prediction accuracy for the ANN classifier is higher than that of the CART model.

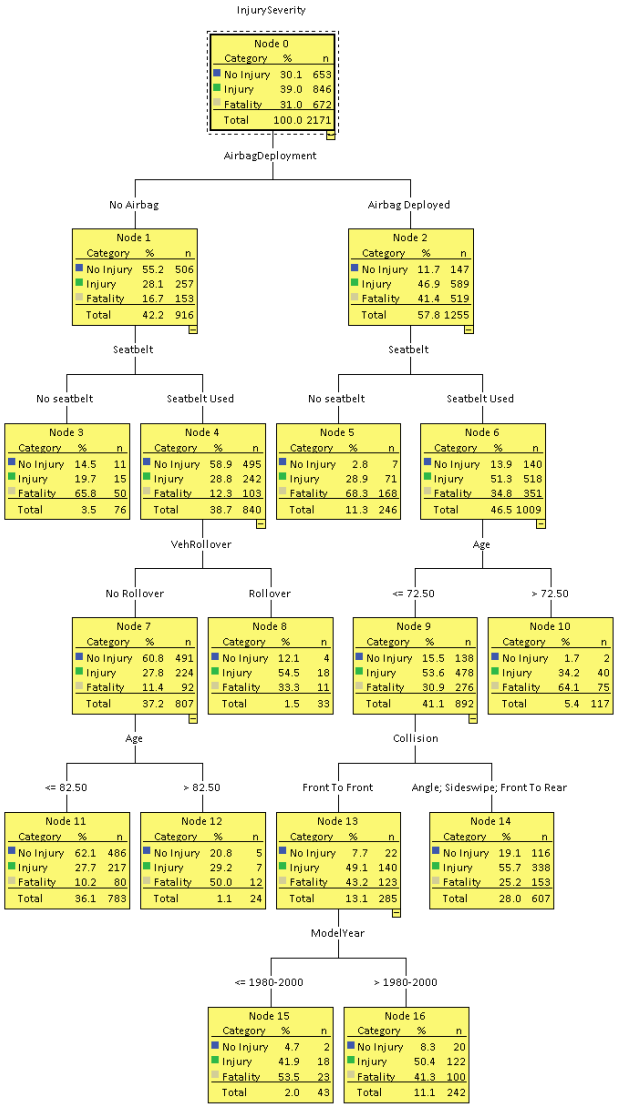

Fig. 6. The Output of the CART Model Using Testing Dataset

Since the accuracy of the model was 63.5%, the ROC curves were used to further evaluate their performance. As shown in Figures 7 through 9, the area under the curves of the training, validation and testing datasets were greater than 0.5. These results indicate satisfactory prediction abilities for new unseen data. However, the area under curves for ROC plots shows very close prediction accuracy for class 1 and 3 (no injury and fatality), and less accuracy for class 2 (injury), this could be due to the fact that minor and major injuries were used undistinguishably with small dataset size for major injury cases. However, the overall prediction accuracy is considered acceptable and promising compared with previous literature such as [11,22,28,45]. In these studies, the ANN prediction accuracy ranged between 55% and 65%. On the other hand, the ANN model by [46] resulted in overall

prediction accuracy for the testing data of 74.6%. However, although their model could predict moderate and minor injuries with 81.6% and 74.6%, respectively, the model could not predict fatal and severe injuries at all.

Table 3. ANN Model Confusion Matrix

Sample	Observed	Predicted			
		No Injury	Injury	Fatality	Percent Correct
Training	**No Injury**	**1059**	332	68	72.6%
	Injury	419	**1388**	335	64.8%
	Fatality	143	524	**908**	57.7%
					64.8%
Validation	**No Injury**	**221**	88	10	69.3%
	Injury	89	**273**	89	60.5%
	Fatality	30	126	**183**	54.0%
					61.1%
Testing	**No Injury**	**245**	81	25	69.8%
	Injury	86	**276**	71	63.7%
	Fatality	25	117	**183**	56.3%
					63.5%

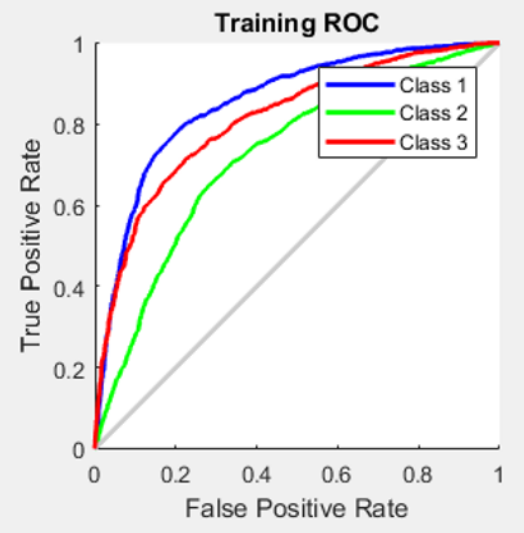

Fig. 7. ANN Model ROC Curve for the Training Dataset

Fig. 8. ANN Model ROC Curve for the Validation Dataset

It was also found that using the softmax function in the ANN output layer does not largely enhance the prediction accuracy of the model compared to other activation functions. However, the use of a softmax function in classification problems assists the network in distinguishing different categories of the target variable. This will eventually lead to more reliable model. Furthermore, the network showed more stable training performance and better generalization abilities compared to other transfer functions.

Figure 9: ANN Model ROC Curve for the Testing Dataset

Additionally, it was clearly noted that the ANN model outperformed the CART model in terms of accuracy and stability, with testing prediction accuracy of 63.5% compared to 59.5%. These findings coincide with [28]. This indicates that ANNs are powerful tools that can be utilized to detect more complex relationships and reveal hidden non-linear effects of multiple crash attributes and resulting in better prediction abilities. This is why ANNs are considered one of the most robust tools used in data mining.

5.3 Multinomial Logistic Regression

The logistic model resulted in a p-value < 0.001 for the likelihood ratio chi-square test. The model also resulted in Cox and Snell R-Square value of 0.385 and Nagelkerke R-Square value of 0.434. Although both values are close to each other, they are lower than the prediction accuracy produced by both CART and ANN models.

According to Table 4, driver's age was significant in determining injury severity. A 5-year increment in driver's age led to 0.9% more odds in facing an injury and 3.7% more odds in facing a fatality. Travel speed was also found significant in distinguishing between no injury and injury

levels. It was found that a 5 mph increment in speed led to 0.5% higher odds in facing an injury compared to no injury. However, speed was found insignificant when comparing fatality to no injury level. Moreover, results revealed that an increment of 5 mph in a roadway speed limit led to 1% more odds of facing an injury and 2.5% more odds of facing a fatality. The roadway functional system was found insignificant.

Daylight driving led to 1.2 times more odds in facing an injury and 1.3 times more odds in facing a fatality compared to driving lighted roadways at night. On the other hand, when roadways are not lighted, facing a fatality will increase by 1.3 times when compared to lighted roadways. Front to front collision was found to have 2.8 times more odds in causing an injury and 3.4 times more odds in causing a fatality when compared to sideswipe collisions. This reflects the hazardous effect of this type of collision.

Table 4 also shows that airbag deployment is 8 times more likely to cause an injury and 11.3 times more likely to cause a fatality. The ejection of the driver's body will also lead to 11 times more odds in causing an injury and 59 times more odds in causing a fatality.

Seatbelt seems to save lives and reduce injuries; where an injury is 67.7% less likely to happen and a fatality is 91.2% less likely to happen when seatbelt is fastened. Further, Female drivers are 1.8 times more likely to have an injury and around 2.2 times more likely to have a fatality. Vehicle rollover increases the odds of having an injury by 4 times and increases the odds of having an injury by 4.2 times.

Vehicle model year was also found significant, except for models between 2006 and 2010 and models newer than 2010, in distinguishing between no injury and injury levels. It was also found that as vehicle age increases it is more likely to have severer injuries.

Driving under influence also increased the odds of having injury by around 1.45 times and increased the odds of having fatality by 2.7 times. Curved roadways increased the odds of having injury or fatality 1.3 times. Further, wet and icy roadways increased the odds of having injury or fatality by 1.31 times. Finally, crashes occurred at intersections had 1.2 times odds having an injury and 1.4 times having a fatality. All comparisons are made with the no injury level.

Table 4. Logistic Model Parameters Estimates (Reference Category: No Injury)

Injury Severity[a]		B	Std. Error	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
						Lower Bound	Upper Bound
Injury	Intercept	3.831	1.108	.001			
	Age	.009	.002	.000	1.009	1.005	1.013
	Travel Speed	.005	.002	.003	1.005	1.002	1.008
	Speed Limit	.010	.004	.010	1.010	1.002	1.018
	[Functional System=local]	-.068	.182	.707	.934	.654	1.334
	[Functional System=collector]	-.060	.146	.683	.942	.708	1.254
	[Functional System=minor arterial]	.055	.134	.680	1.057	.812	1.375
	[Functional System=major arterial]	-.025	.109	.822	.976	.788	1.208
	[Functional System=Interstate]	0[b]	.	.	.	.	.
	[Light Condition=Daylight]	.195	.096	.042	1.215	1.007	1.466
	[Light Condition=Dark, Not Lighted]	.063	.121	.602	1.065	.841	1.349
	[Light Condition=Dark, But Lighted]	0[b]	.	.	.	.	.
	[Collision=Front to Rear]	.775	.126	.000	2.170	1.694	2.779
	[Collision=Front to Front]	1.021	.139	.000	2.777	2.115	3.646
	[Collision=Angle]	.631	.124	.000	1.879	1.475	2.394
	[Collision=Sideswipe]	0[b]	.	.	.	.	.
	[Airbag Deployment=Not Deployed]	-2.090	.074	.000	.124	.107	.143
	[Airbag Deployment=Deployed]	0[b]	.	.	.	.	.
	[Ejection=No Ejection]	-2.395	1.039	.021	.091	.012	.698
	[Ejection=Ejection]	0[b]	.	.	.	.	.
	[Seatbelt=Not Used]	1.129	.155	.000	3.091	2.283	4.185
	[Seatbelt=Used]	0[b]	.	.	.	.	.
	[Gender=Female]	.579	.071	.000	1.783	1.551	2.050
	[Gender=Male]	0[b]	.	.	.	.	.
	[Vehicle Rollover=No Rollover]	-1.389	.186	.000	.249	.173	.359
	[Vehicle Rollover=Rollover]	0[b]	.	.	.	.	.
	[Model Year=<2000]	.481	.105	.000	1.618	1.316	1.989
	[Model Year=2001-2005]	.383	.088	.000	1.466	1.234	1.742
	[Model Year=2006-2010]	.106	.084	.210	1.111	.942	1.311
	[Model Year=>2010]	0[b]	.	.	.	.	.
	[Drunk Driving=Not Drunk]	-.378	.161	.019	.685	.500	.940
	[Drunk Driving =Drunk]	0[b]	.	.	.	.	.
	[Road Alignment=Straight]	-.255	.115	.026	.775	.619	.970
	[Road Alignment=Curved]	0[b]	.	.	.	.	.
	[Surface Condition=Dry]	-.267	.108	.013	.766	.620	.946
	[Surface Condition=Wet/Icy]	0[b]	.	.	.	.	.

Injury Severity[a]		B	Std. Error	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
						Lower Bound	Upper Bound
	[Intersection=No]	-.181	.085	.032	.834	.707	.985
	[Intersection=Yes]	0[b]	.	.	.	.	.
Fatality	Intercept	3.327	1.120	.003			
	Age	.037	.002	.000	1.037	1.033	1.042
	Travel Speed	.003	.002	.125	1.003	.999	1.007
	Speed Limit	.024	.005	.000	1.025	1.016	1.034
	[Functional System=local]	-.260	.216	.229	.771	.505	1.178
	[Functional System=collector]	-.021	.169	.899	.979	.703	1.363
	[Functional System=minor arterial]	.099	.157	.525	1.105	.813	1.501
	[Functional System=major arterial]	-.021	.129	.873	.980	.761	1.261
	[Functional System=Interstate]	0[b]	.	.	.	.	.
	[Light Condition=Daylight]	.248	.117	.034	1.282	1.019	1.613
	[Light Condition=Dark, Not Lighted]	.280	.141	.047	1.322	1.003	1.743
	[Light Condition=Dark, But Lighted]	0[b]	.	.	.	.	.
	[Collision=Front to Rear]	.507	.151	.001	1.661	1.237	2.231
	[Collision=Front to Front]	1.222	.158	.000	3.395	2.491	4.628
	[Collision=Angle]	.340	.146	.020	1.404	1.054	1.871
	[Collision=Sideswipe]	0[b]	.	.	.	.	.
	[Airbag Deployment=Not Deployed]	-2.425	.089	.000	.089	.074	.105
	[Airbag Deployment=Deployed]	0[b]	.	.	.	.	.
	[Ejection=No Ejection]	-4.087	1.031	.000	.017	.002	.127
	[Ejection=Ejection]	0[b]	.	.	.	.	.
	[Seatbelt=Not Used]	2.423	.155	.000	11.281	8.323	15.290
	[Seatbelt=Used]	0[b]	.	.	.	.	.
	[Gender=Female]	.779	.082	.000	2.179	1.855	2.560
	[Gender=Male]	0[b]	.	.	.	.	.
	[Vehicle Rollover=No Rollover]	-1.435	.197	.000	.238	.162	.351
	[Vehicle Rollover=Rollover]	0[b]	.	.	.	.	.
	[Model Year=<2000]	1.310	.119	.000	3.708	2.937	4.681
	[Model Year=2001-2005]	.973	.103	.000	2.646	2.162	3.238
	[Model Year=2006-2010]	.427	.101	.000	1.533	1.257	1.869
	[Model Year=>2010]	0[b]	.	.	.	.	.
	[Drunk Driving =Not Drunk]	-.989	.168	.000	.372	.268	.516
	[Drunk Driving =Drunk]	0[b]	.	.	.	.	.
	[Road Alignment=Straight]	-.262	.128	.040	.770	.599	.989
	[Road Alignment=Curved]	0[b]	.	.	.	.	.
	[Surface Condition=Dry]	-.269	.123	.029	.764	.600	.973
	[Surface Condition=Wet/Icy]	0[b]	.	.	.	.	.
	[Intersection=No]	-.353	.100	.000	.703	.578	.854
	[Intersection=Yes]	0[b]	.	.	.	.	.

a: The reference Category is: No Injury. b: This parameter is set to zero because it is redundant.

6. Conclusions

This study presented a thorough investigation to identify drivers, vehicles, roadway, and crash characteristics that are influential in determining injury severity levels sustained by drivers in traffic crashes. Sixteen crash attributes based on 7,394 traffic car crashes occurred across the United States of America in 2015 were used in this study. The target variable (drivers' injury severity) was divided into three categories: no injury, injury and fatality. Classification and Regression Trees, feedforward backpropagation Artificial Neural Networks and Multinomial Logistic Regression models were used for this purpose.

The CART model showed reasonable classification abilities with training and testing prediction accuracies of 61.4% and 59.5%, respectively. Further, out of the sixteen crash attributes, six factors showed significant contribution in determining drivers' injury severity levels, these factors are: airbag deployment, seatbelt use, drivers' age, vehicle rollover, collision type, and vehicle model year. It was shown that most fatalities and injuries are correlated to airbag deployment. On another level, the use of seatbelts was found essential in reducing fatalities and injuries. It was also found that airbag deployment associated with no seatbelt use leads to higher fatality and injury rates. Elderly drivers (aged 72 and above) were found to be more prone to fatal and severe injuries compared to younger ones. It was found that vehicles rollover increases the risk of severe and fatal injuries even when seatbelt is used. Furthermore, front to front collisions resulted in higher rates of fatalities and severe injuries compared to front to back, angled, and sideswipe collisions. Finally, it was shown that vehicles of model year higher than 2000 had lower fatality and injury rates.

Both CART and ANN models seemed to outperform the logistic regression model. Based on the logistic model, all crash attributes were found significant in distinguishing between drivers' injury severity levels except for the roadway functional system. The ANN model outperformed both models with prediction accuracies for the training, validation and testing datasets of 64.8%, 61.1% and 63.5%, respectively. This can be related to the capability of ANNs in detecting more complex relationships and reveal hidden non-linear effects of multiple crash attributes at the same time. Finally, although the use of the softmax function in the ANN output layer led to more stable training performance and better generalization abilities compared to other transfer functions, it did not really enhance the prediction accuracy of the model.

References

[1] United Nations Road Safety Collaboration (2011) Global plan for the decade of action for road safety 2011-2020. World Health Organization, Geneva.
[2] National Highway Traffic Safety Administration (2017) Traffic safety facts, 2014 data: occupant protection. Washington, DC: US Department of Transportation, National Highway Traffic Safety Administration; 2016.
[3] Chang LY, Wang HW (2006) Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention 38 (5):pp.1019-1027.
[4] Kopelias P, Papadimitriou F, Papandreou K, Prevedouros P (2007) Urban Freeway Crash Analysis: Geometric, Operational, and Weather Effects on Crash Number and Severity. Transportation Research Record: Journal of the Transportation Research Board 2015:pp.123-131.
[5] Zhang J, Lindsay J, Clarke K, Robbins G, Mao Y (2000) Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accident Analysis & Prevention 32 (1):pp.117-125.
[6] Keall MD, Frith WJ, Patterson TL (2004) The influence of alcohol, age and number of passengers on the night-time risk of driver fatal injury in New Zealand. Accident Analysis & Prevention 36 (1): pp.49-61.
[7] Zajac SS, Ivan JN (2003) Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut. Accident Analysis & Prevention 35 (3):pp.369-379.
[8] Bedard M, Guyatt GH, Stones MJ, Hirdes JP (2002) The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis & Prevention 34 (6):pp.717-727.
[9] Tefft BC (2013) Impact speed and a pedestrian's risk of severe injury or death. Accident Analysis & Prevention 50:pp.871-878.
[10] Behnood A, Mannering F (2017) Determinants of bicyclist injury severities in bicycle-vehicle crashes: a random parameters approach with heterogeneity in means and variances. Analytic methods in accident research 16:pp.35-47.
[11] Abdelwahab H, Abdel-Aty M (2001) Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board 1746:pp.6-13.
[12] Tong W, Cherian P, Liu J, Li H, Gu Q (2016) Statistical analysis of DMV crash data. Paper presented at the Systems and Information Engineering Design Symposium (SIEDS).
[13] Sapri FE, Nordin NS, Hasan SM, Wan Yaacob WF, Nasir M, Azlin S (2017) Decision tree model for non-fatal road accident injury. International Journal on Advanced Science, Engineering and Information Technology 7 (1):pp.63-70.
[14] Shanthi S, Ramani RG (2011) Classification of vehicle collision patterns in road accidents using data mining algorithms. International Journal of Computer Applications 35 (12):pp.30-37.
[15] Baluni P, Raiwani YP (2014) Vehicular accident analysis using neural network. International Journal of Emerging Technology and Advanced Engineering 4 (9):pp.161-164.
[16] Mannering FL, Bhat CR (2014) Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research 1:pp.1-22.
[17] Mannering F (2018) Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17:pp.1-13.
[18] Chang LY, Chien JT (2013) Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety science 51 (2):pp.17-22.
[19] Kashani AT, Mohaymany AS (2011) Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science 49 (10):pp.1314-1320.
[20] Rovšek V, Batista M, Bogunović B (2017) Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree. Transportation Research Record: Journal of the Transportation Research Board 32 (3):pp.272-281.
[21] Chen C, Zhang G, Qian Z, Tarefder RA, Tian Z (2016) Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accident Analysis & Prevention 90:pp.128-139.
[22] Liu D, Solomon D, Hardy L (2015) Investigating injury occurrence in motor vehicle collision using artificial neural networks. Human Factors and Ergonomics in Manufacturing & Service Industries 25 (3):pp.294-303.
[23] Yasin Çodur M, Tortum A (2015) An Artificial Neural Network Model for Highway Accident Prediction: A Case Study of Erzurum, Turkey. PROMET-Traffic&Transportation 27 (3):pp.217-225.
[24] Delen D, Sharda R, Bessonov M (2006) Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis & Prevention 38 (3):pp.434-444.
[25] Sohn SY, Shin H (2001) Pattern recognition for road traffic accident severity in Korea. Ergonomics 44 (1):pp.107-117.
[26] Chong MM, Abraham A, Paprzycki M (2004) Traffic accident analysis using decision trees and neural networks. arXiv preprint cs/0405050.
[27] Castro Y, Kim YJ (2016) Data mining on road safety: factor assessment on vehicle accidents using classification models. International journal of crashworthiness 21 (2):pp.104-111.
[28] Zeng Q, Huang H (2014) A stable and optimized neural network model for crash injury severity prediction. Accident Analysis & Prevention 73:pp.351-358.
[29] Sameen MI, Pradhan B (2017) Severity Prediction of Traffic Accidents with Recurrent Neural Networks. Applied Sciences 7 (6):p.476.
[30] Sohn SY, Lee SH (2003) Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Safety Science 41 (1):pp.1-14.
[31] Mohamed MG, Saunier N, Miranda-Moreno LF, Ukkusuri SV (2013) A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Safety science 54: pp.27-37.
[32] Li Y, Ma D, Zhu M, Zeng Z, Wang Y (2018) Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. Accident Analysis & Prevention, 111:pp.354-363.
[33] National Highway Traffic Safety Administration (2010) Drug involvement of fatally injured drivers. US Department of Transportation. vol Report No. DOT HS, 811.
[34] Berning A, Smither DD (2014) Understanding the limitations of drug test information, reporting, and testing practices in fatal crashes. vol 812-072. DOT HS.
[35] Durbin DR, Jermakian JS, Kallan MJ, McCartt AT, Arbogast KB, Zonfrillo MR, Myers RK (2015) Rear seat safety: variation in protection by occupant, crash and vehicle characteristics. Accident Analysis & Prevention 80:pp.185-192.
[36] Knapman C (2012) Young drivers are still the most vulnerable. The Telegraph. https://www.telegraph.co.uk/motoring/road-safety/9412798/Young-drivers-are-still-the-most-vulnerable.html.

[37] Loh WY (2011) Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (1):pp.14-23.
[38] Tan PN (2006) Introduction to data mining. Pearson Education, India.
[39] Demuth HB, Beale MH, De Jess O, Hagan MT (2014) Neural network design. Martin Hagan.
[40] Chaturvedi DK (2008) Soft computing: techniques and its applications in electrical engineering, vol 103. Springer.
[41] Jayalakshmi T, Santhakumaran A (2011) Statistical normalization and back propagationfor classification. International Journal of Computer Theory and Engineering 3 (1):p,89.
[42] Myers RH (1990) Classical and modern regression with applications, vol 2. Duxbury press, Belmont, CA.
[43] Wallis LA, Greaves I (2002) Injuries associated with airbag deployment. Emergency medicine journal 19 (6):pp.490-493.
[44] Blacksin MF (1993) Patterns of fracture after air bag deployment. The Journal of trauma 35 (6):pp.840-843.
[45] Chong M, Abraham A, Paprzycki M (2005) Traffic accident analysis using machine learning paradigms. Informatica 29 (1):pp.89-98.
[46] Alkheder S, Taamneh M, Taamneh S (2017) Severity prediction of traffic accident using an artificial neural network. Journal of Forecasting 36 (1):pp. 100-108.