IJTE, Vol. 9, No. 3, Winter 2022, pp. 635-652

Research Paper

Analyzing and Predicting Fatal Road Traffic Crash Severity Using Tree-Based Classification Algorithms

Saba Momeni Kho [1], Parham Pahlavani [2,*], Behnaz Bigdeli [3]

Received: 2020/11/24 *Accepted: 2021/11/20*

Abstract

Nowadays, a significant part of goods and passengers are transported on suburban highways with mainly high-speed vehicles. Hence, these highways are very prone to accidents with different injuries. Due to the high fatality or severe physical/mental injury rates caused by car crashes, analyzing these accident-prone areas and identifying the factors affecting their occurrences is crucial. The specific objective of the study was to compare Chi-square Automatic Interaction Detector (CHAID), Classification and Regression Tree (CART), C4.5 and C5.0 decision tree data mining classification algorithms in building classification models for the fatality severity of 2355 fatal crash data records during 2007-2009 occurred in the roadways of 8 states in the USA. The results were evaluated using the accuracy metrics such as overall accuracy, kappa rate, precision, recall, and F-measure. The investigations confirmed that C5.0 had the best performance with the overall accuracy, and kappa rates of 94% and 92%, respectively. Additionally, classified fatality severity levels of the crashes were proposed for each algorithm to generate risk maps on the roads, to create potential accident risk spots. Decision tree models can be used for real-time data to find invariants in the tree over a period of time, which would be beneficial for policymakers.

Keywords: Fatality Severity, Risk Map, Classification, Decision Tree algorithms

[*] Corresponding author. E-mail: pahlavani@ut.ac.ir

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran.

[2] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran.

[3] School of Civil Engineering, Shahrood University of Technology, Shahrood, Iran.

1. Introduction

According to the World Health Organization (WHO), traffic accidents are among the top eight causes of death in the world. Up to 1.2 million people are killed in accidents worldwide each year, and 20-25 times as many are seriously injured [World Health Organization, 2018]. Among the various infrastructures of a country, roads are of great importance in the transfer of goods and passengers. In order to manage and reduce accidents and increase safety in a suburban road, it is necessary to find out when and where it has happened. By modeling accident hotspots to identify the factors affecting the occurrence of accidents, it is possible to make a valuable contribution to reducing the severity of accidents and improving road safety with the identification of these points. Crash factors can be included into multiple categories such as Driver-related (e.g., physical and mental disabilities, improper driving skills, careless attention to traffic signs, alcohol/drug use, tiredness, using a cell phone, not wearing a seat belt, etc.), Vehicle-related (e.g., the vehicle model and its technical defects), Environmental-related (e.g., weather situation, light conditions and the land use of the area) and Road-related (e.g., the number of lanes, slope, curvature, surface condition, speed limit, intersection types, etc.) [Effati, Thill and Shabani, 2015]. The accumulation of several factors in one place causes an increase in the rate of accidents. In these areas, which are called critical points, accidents occur with greater intensity or rate [Thakali, Kwon and Fu, 2015]. By analyzing accidents, critical points and their relationship between various factors can be discovered [Blazquez and Celis, 2013]. As traffic accidents and their location and time are unpredictable, forecasting them is important to reduce their happening and subsequent damages. Accordingly, prediction models have been developed to improve road infrastructure safety management [Lee et al. 2020]. Data mining is referred to as the knowledge discovery in data and is one of the most widely used techniques for most engineers and business people [Chang and Wang, 2006]. Various methods such as classification, clustering, association rule mining, etc. are considered as data mining techniques. Ample research has been undertaken into the use of various methodologies including regression, statistical and, machine learning models to inspect injury severity outcomes and the fatality risk. Decision trees have been used more recently, as they provide an explanation together with an accurate, reliable and, quick response. In this study, the main objective is to compare CHAID, C4.5, C5.0 and, CART algorithms, as four popular decision trees to classify fatal accidents and assess their performance based on different accuracy metrics. Moreover, Risk map generation is applied for each method to identify high-risk areas and prevent future accidents in the related hotspots. The proposed methodology could be used to determine the best classifier in road safety management. The rest of the paper is organized as follows: Section 2 gives a summarized explanation about the most relevant research on studying crash severity. Section 3 introduces the study area and describes detailed information about the fatality severity and crash-related variables. The main concepts of classification methods and evaluation metrics are given in section 4. The results of predictive models are presented in section 5. Important findings and possible ideas for future research are discussed in section 6. Finally, the concluding remarks are illustrated in section 7.

2. Literature Review

This section expands some comparative studies related to data mining in road accidents by means of different algorithms, mainly including decision trees. A considerable number of

studies have been conducted to evaluate different methods in crash analysis in order to modulate crash severity levels, reengineer environmental, road, and vehicle factors, and of course, reduce fatality/injury rate. Related to this field, machine learning-based models have been developed to estimate the results of accidents. Delen et al. [Delen et al. 2017] used a survey to model the relationships between various levels of injury severity and crash factors. They applied numerous experimentation with four top prediction models including Neural Networks (NN), Support Vector Machines (SVM), C5.0 tree, and Logistic regression (LR) on a nationwide data collection. According to the results, SVM was the most accurate classifier with an accuracy rate of 90.41% followed by the C5.0 tree with an accuracy of 86.61%. In the final part of their research, the sensitivity analysis results revealed that factors like wearing a seat belt, manner of collision, ejection from the car, and drug use were the most important variable affecting accident occurrence. Zhang et al. [Zhang et al. 2018] compared the prediction accuracy and variable importance of statistical and machine learning methods with the Florida crash dataset in the United States at freeway diverge areas. RF as a mostly analyzed method was proved to outperform the other classifiers. They stated that machine learning methods can predict better than statistical methods, as it is unnecessary to make presumptions about the relationship between the dependent and independent variables and the data distribution. Based on the dataset of the critical expressways in Seoul, South Korea, machine learning methods including Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Random Forest (RF) were used to assess their powers and weaknesses [Lee et al. 2020]. RF was found to produce the most accurate results in terms of mean squared error (MSE) and root mean squared error (RMSE). According to [Xing et al. 2020], toll plazas with both electronic and manual toll collection increase the risk of vehicle collision because of lane-change behaviors. Thus, the authors conducted a comparative study of logistic regression (LR) and five non-parameter models to examine the vehicle collision risk. Interestingly, the results demonstrated that ANN did not outperform LR unlike the other methods, which is in contrast with [Zhang et al. 2018] results. Since the study explained here focuses on traffic accident analysis with "tree-based" algorithms, the literature in this section will be mostly specific to the relevant works in this particular area. Among different approaches for studying the injury severity of accidents, decision trees are being more extensively used; because they are easily understandable and yield more productive results [Ahmed, Rizaner and Ulusoy, 2018]. Apart from all the crash-related factors, the manner of collision affects the fatality rate. In view of this, Shanthi and Ramani [Shanthi and Ramani, 2011] classified a total number of 37259 U.S. traffic accident records in 2007 to mine vehicle collision patterns with algorithms including Naïve Bayes, C4.5, CART, ID3, cost-sensitive decision tree, and random tree. The results indicated that the random tree achieved the highest accuracy of 87% among the other algorithms. Oña et al. [Oña, López and Abellán, 2013] examined the accuracies obtained by ID3, C4.5 and, CART methods in a 19-variable dataset of rural highway accidents in Spain. They claimed that CART, followed by C4.5 and ID3 obtained accuracies of 55.87%, 54.16%, and 52.72%, respectively. Mansouri and Kargar [Mansouri and Kargar, 2014] made an analysis of 10000 accidents from 2011 to 2013 in Isfahan province, Iran with CART, C5.0, CHAID, and the quick unbiased efficient statistical (QUEST) trees. They found that out of the mentioned methods, the C5.0 tree out-performed the other decision trees with an

accuracy rate of 70.18%, while CART had the worst prediction on test data with an accuracy of 43.98%. Bahiru et al. [Bahiru et al. 2018] employed the J48, ID3, CART, and Naïve Bayes classification algorithms on 3000 records of UK traffic accident repository to predict the accident severity and find out the most significant factors. Based on their experimental results, J48 classifier was the most accurate prediction model with an accuracy of 96.3%. Yuan et al. [Yuan et al. 2020] established C5.0, CHAID and CART decision trees to identify high-influence factors on the severity of side right-angle collision accidents. Apart from C5.0 better performance with an accuracy of 61.9%, drunk driving was found to be the most important factor followed by weather conditions and over speeding.

3. Data Description

The Fatality Analysis Reporting System (FARS) database is a census of all fatal crashes in the U.S. which covers crashes that lead to at least one fatality within thirty consecutive days from the time of the crash. FARS is directed by the National Center for Statistics and Analysis (NCSA), which is a component of the National Highway Traffic Safety Administration (NHTSA) [Fatality Analysis Reporting System, 2019]. NHTSA has an agreement with an agency in each state to provide information on all fatal crashes in the state, which includes driver, vehicle, roadway, and environmental factors and crash characteristics to represent the crashes and their events.

FARS data are obtained from multiple states' documents, such as police crash reports, death certificates, state vehicle registration and driver licensing files, and medical service reports. The FARS data elements are coded from these documents by the analysts by means of a manual with written instructions [United States. National Highway Traffic Safety Administration, 2006]. After the data file is created, quality checks are performed to improve the accuracy of the data. The FARS data are generally reliable and complete, and they are available free online.

The FARS data for the current study is sourced from the FARS database for the year 2007. In this year, a total of 41,059 people was killed in near 6,000,000 police-reported motor vehicle traffic crashes throughout the USA. Here, the dataset contains fatal crashes that occurred in the roadways (see Figure 1) connecting eight eastern U.S. states (see Figure 2), namely Virginia, West Virginia, Kentucky, Tennessee, North Carolina, South Carolina, Georgia, and Alaska. A total of 2,355 records from 2007 to 2009 (vehicle crashes only; neither pedestrian nor bicycle) was collected. This dataset was chosen for the study, mainly due to including adequate features and accessibility for the analysis.

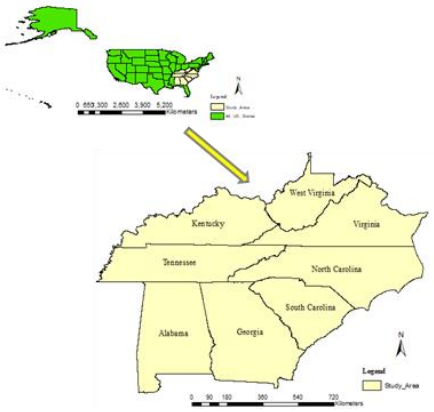


Figure 1. The Study Area of Eight Eastern US States

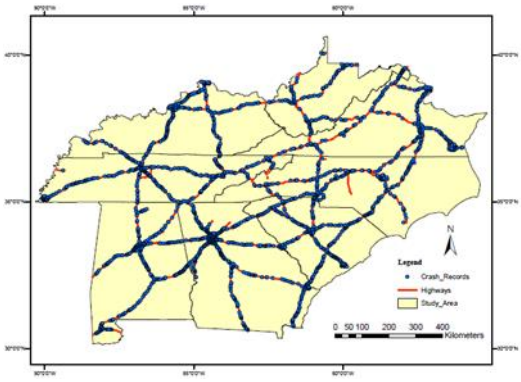


Figure 2. The Study Area of Highway Crashes, with North American 1983 Geographic Coordinate System

Table 1. Description of Variables

Category	Abbreviation	Description	Details	
Driver factor	DRUNK_DR	Number of Drunk Drivers in Crash	Range 0-3	
Environmental factors	WEATHER	Atmospheric Condition	1	Clear/Cloudy (No Adverse Conditions)
			2	Rain
			3	Sleet (Hail)
			4	Snow or Blowing Snow
			5	Fog, Smog, Smoke
	LGT_COND	Light Condition	1	Daylight
			2	Dark - Not Lighted
			3	Dark but Lighted
			4	Dawn
			5	Dusk
Road factors	JUNCTYPE	Relation to Junction	1	Non-Junction (Non-Interchange)
			2	Intersection (Non-Interchange)
			3	Intersection Related (Non-Interchange)
			4	Driveway, Alley Access, etc. (Non-Interchange)
			5	Entrance/Exit Ramp Related (Non-Interchange)
			6	Rail Grade Crossing (Non-Interchange)
			7	Crossover-Related (Non-Interchange)
			8	Driveway-Access Related
			10	Intersection (Interchange Area)
			11	Intersection Related (Interchange Area)
			12	Driveway Access (Interchange Area)
			13	Entrance/Exit Ramp Related (Interchange Area)
			14	Crossover-Related (Interchange Area)
			15	Other location in Interchange (Interchange Area)
	NO_LANES	Number of Travel Lanes	Range 1-7	
	SPEED	Speed Limit	Range 25-70 (mph)	
	ALIGNMENT	Roadway Alignment	1	Straight
			2	Curve
	PROFILE	Roadway Profile	1	Level
			2	Grade
			3	Hillcrest
			4	Sag
	PAVE_TYPE	Roadway Surface Type	1	Concrete
			2	Blacktop, Bituminous, or Asphalt
	SUR_COND	Roadway Surface Condition	1	Dry
			2	Wet
			3	Snow or Slush
			4	Ice/Frost
			5	Sand, Dirt, Mud, Gravel
			6	Water (standing or moving)

The factors used in this study were mainly based on the opinion of experts, and they were common in similar studies. Three sets of the driver, environmental, and road factors were extracted from the data. The detailed information description on the crash factors is shown in Table 1. For the fatal occurrences, a discrete variable was created to represent the fatality severity of these crashes based on the death of road users, as seen in Table 2. The ratio of training to validating and testing was chosen as 70% to 30%, respectively.

Table 2. Fatality Severity Level of Crashes in the Dataset

Fatality Severity Level	Definition	Train		Test		Total	
		Freq.	Ratio	Freq.	Ratio	Freq.	Ratio
Level 0	No-crash	329	20%	142	20%	471	20%
Level 1	1 person killed	444	27%	184	26%	628	27%
Level 2	2-4 people killed	515	31%	216	31%	731	31%
Level 3	More than 4 people killed	360	22%	165	23%	525	22%

No-crash points were also included in the dataset to recognize safe locations with no crashes, set as Level 0 containing 20% of the whole dataset. From least severe to most severe fatal crashes with their frequencies within the dataset: Level 1 Led to the death of one person (27%), Level 2 Led to the death of between two and four people (31%), and Level 3 Led to the death of more than four people.

4. Methodology

The objective of the study is to build and compare the performance of four decision tree algorithms. At first, the classification models were applied to the training set to build the hierarchical structures for predicting the test data. Then the prediction results were evaluated based on the confusion matrices and five accuracy measures. The programming language R was used for the whole process.

4.1. Classification Models

For the classification process, CHAID, C5.0, C4.5, and CART were used. CHAID only deals with categorical data and compared to the other methods, it uses a different measure to select input variables. C5.0 handles various data types and function fast, and it is suitable for big datasets. CART is able to select the most discriminatory factors which lead to less computation.

4.1.1. CHAID Tree

CHAID stands for Chi-square Automatic Interaction Detector. It is a decision tree technique created by V.Kass [Kass, 1980]. CHAID develops a pruned non-binary tree with only categorical variables based on Chi-square variable independence test (χ^2) , as shown in Equation (1) [Lin and Fan, 2019]:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{k}\frac{(x_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

where x_{ij} is the observed frequency and E_{ij} is the expected frequency. As the test works with nominal data, it uses frequencies rather than means or variances. It is also more beneficial when dealing with large size of data, because of generating multiple splits. The tree generation can be divided into three phases: 1. A Chi-square test is performed to determine important independent variables. 2. The splitting process is started using the independent variables with the smallest p-values. If this p-value is less than or equal to the alpha4 parameter specified by the user, the node is split. Otherwise, the node is considered a terminal node. The separation of the nodes is continued until there is no independent variable with the smallest p-value and then the terminal node is reached. 3. The first phase is repeated until all the subgroups of

the tree are processed [Susanti et al. 2017] . If the p-value is larger than the alpha2 parameter for a pair of nodes with the largest p-value, the pair is merged into a single category. If this new category consists of three or more nodes, a binary split will be formed if the p-value is the smallest and not larger than the alpha3 parameter. These user-specified parameters are determined during the training process.

4.1.2. C5.0 Tree

C5.0 is an improved version of the C4.5 tree, which is also an extension of the ID3 algorithm developed by Quinlan [Quinlan, 1986]. The algorithm uses information gain as the splitting criterion and gives a binary or multi branches tree. In comparison with C4.5, C5.0 accounts for missing values, dates, times, and ordered variables. Moreover, it has a faster functionality and less memory consumption. The tree structure uses boosting, i.e. many generated decision trees are combined to improve the prediction. A variable with the highest information gain is chosen as the splitting variable of the root node. Then the algorithm is recursively applied to each branch to build the nodes and branches. In order to define information gain, entropy must be introduced first. Entropy is an impurity measure in the dataset, defined as Equation (2) [Adhatrao et al. 2013]:

$$Ent(D) = -\sum_{d=1}^{|k|} p_k \log_2 p_k \qquad (2)$$

where D refers to the train data and p_k is the probability that an item in the train data belongs to class k. If the data consists of just one class (k=1), then p_k is 1 and the entropy would be zero. Therefore, the entropy for a homogeneous dataset would be minimized. As mentioned before, to minimize the tree depth, an optimal variable is needed to split the tree node, which can be implied that the variable with the most entropy reduction is the ideal choice. According to Equation (3), the information gain for each variable can then be defined as the difference between the entropy of the dataset and the weighted sum of the entropy from the data subsets [Y. Wang et al. 2017]:

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^V|}{|V|} \times Ent(D^V) \qquad (3)$$

where $|D|$ is the number of train data, a is a variable in the dataset with V different values, D^V is the samples in every value, and $|D^V|$ is the number of these samples. Three parameters are optimized after the training process: 1. The number of boosting iterations called “trials”. 2. A logical value whether to decompose the tree into a rule-based model or not, denoted as “model”. 3. Another logical value whether to use feature selection or not, denoted as “winnow”.

4.1.3. C4.5 Tree

This algorithm is another decision tree generator developed by Quinlan [Quinlan, 1993]. C4.5 was proposed as an extension of Quinlan’s ID3 algorithm, as it is sensitive to variables with so many values [Hssina et al. 2014]. The variable with the most effective data splits into classes is chosen by C4.5 to build the parent node. The process is then repeated for each branch of the tree until having the same class for all samples in the branch. The algorithm determines the gain ratio for variable splitting. It normalizes the information gain by the information needed to determine the class for an instance in the dataset, which is called the split entropy [Mienye, Sun and Wang, 2019]. This can be calculated as follows:

$$SplitInfo(D,a) = -\sum_{v=1}^{V} \frac{|D^V|}{D} . \log_2 \frac{|D^V|}{D} \qquad (4)$$

According to Equation (4), the gain ratio for each variable can then be defined as follows:

$$GainRatio(D,a) = \frac{Gain(D,a)}{SplitInfo(D,a)} \qquad (5)$$

During the training process of the tree, two parameters are tuned: pruning confidence (C) and a minimum number of instances (M). The

former calculates an upper bound on error rate at the leaf, which is used to enhance the performance and prevent overfitting by removing unnecessary nodes. The latter enforces a minimum size of instances per node in order to fit the tree.

4.1.4. Classification and Regression Tree

Classification and Regression Tree (CART) is one of the widely used non-parametric data mining techniques which can analyze data with various independent quantitative or qualitative variables. CART can be a sturdy model to analyze complex tasks in a simple hierarchical form and discover rules [Choi et al. 2020]. Problems like the way of splitting each node, determining the completeness of a tree, and giving the terminal nodes a class label are noticed in the algorithm. CART uses a top-down partitioning with selecting the most suitable variable to split the data into two groups at the root node (the parent node), such that the class labels in each group are as homogeneous as possible. Then, splitting is recursively applied to each group [Rovšek, Batista and Bogunović, 2017]. The Gini Index (GI) is used for CART implementation in R as the splitting criterion. It measures the degree of a particular randomly chosen item of data being wrongly classified, calculated in Equation (6) [Wang and Li, 2019]:

$$GI = 1 - \sum_{i=1}^{n} p_i^2 \quad (6)$$

where p_i is the relative frequency of class i in the dataset. During the training process, an optimal value for Complexity Parameter (CP) is determined. This time-saver parameter prunes off the unimportant splits. In other words, any split which does not improve the fit by CP will be pruned off by cross-validation.

4.2. Classification Accuracy Metrics

In an attempt to recognize the best performance when comparing multiple classifiers, a confusion matrix is used. This multi-class confusion matrix identifies how many of the *N*-predicted samples are correctly or incorrectly classified in each class [Tallón-Ballesteros and Riquelme, 2014], as shown in Table 3. The "Predicted" classifications section contains four subsections for each of the classes that want to be classified into and the "Actual" classifications section which has four subsections for each of the classes. In other words, every column of the confusion matrix represents the instances of that predicted class and each row of the confusion matrix represents the instances of the actual class. Moreover, the sum of actual and predicted instances for each class, namely Ci_{Act} and Ci_{Pred} (i is the class identifier) has been calculated and specified in the table. Hence, CM_{ii} or the elements in the major diagonal (**a**, **f**, **k,** and **p**) are the elements correctly classified, while the elements out of this diagonal are misclassified. In a classification problem with more than two classes, the "one versus all" approach is used for calculating accuracy metrics [Rhys, 2020]. One class should be considered as the positive class and the other classes should be considered as the negative class.

Table 3. An Example of a Confusion Matrix with Four Classes

		Predicted				
		Level 0	**Level 1**	**Level 2**	**Level 3**	**Total**
Actual	**Level 0**	**a**	**b**	**c**	**d**	$C0_{Act}$
	Level 1	**e**	**f**	**g**	**h**	$C1_{Act}$
	Level 2	**i**	**j**	**k**	**l**	$C2_{Act}$
	Level 3	**m**	**n**	**o**	**p**	$C3_{Act}$
	Total	$C0_{Pred}$	$C1_{Pred}$	$C2_{Pred}$	$C3_{Pred}$	N

Overall accuracy is the proportion of correctly classified samples among all N predicted samples. It indicates the classifier quality to correctly identify samples, as shown in

Equation (7) [Tallón-Ballesteros and Riquelme, 2014]:

$$OA = \frac{a + f + k + p}{N} \quad (7)$$

Kappa is an agreement measure between observed and predicted classes for cases in the test set, ranging from -1 to 1. It can be calculated from the following equation [Tallón-Ballesteros and Riquelme, 2014]:

$$Kappa = \frac{N \times \sum_{i=1}^{4} CM_{ii} - \sum_{i=1}^{4} Ci_{Act} Ci_{Pred}}{N^2 - \sum_{i=1}^{4} Ci_{Act} Ci_{Pred}} \quad (8)$$

where CM_{ii} are the major diagonal elements of the confusion matrix. False-positive (FP) representing the type I error for a particular class can be calculated by taking the sum of all the values in the column corresponding to that class except the value in the major diagonal [Diez, 2018] :

$$FP_i = Ci_{\mathrm{Pred}} - CM_{ii} \quad (9)$$

False-negative (FN) representing the type II error for a particular class can be calculated by taking the sum of all the values in the row corresponding to that class except the value in the major diagonal [Diez, 2018] :

$$FN_i = Ci_{Act} - CM_{ii} \quad (10)$$

Given all the predicted labels for Class i, precision or positive predictive value of the class determines the number of correctly classified samples divided by the sum of the class predicted samples [Kumar and Toshniwal, 2017]:

$$Precision_{Ci} = \frac{CM_{ii}}{Ci_{Pred}} \quad (11)$$

Moreover, class recall, also called True Positive Rate (TPR) or sensitivity is the ratio of correctly classified samples divided by the number of samples in the actual class. The formula is given as follows [Kumar and Toshniwal, 2017]:

$$Recall_{Ci} = \frac{CM_{ii}}{Ci_{Act}} \quad (12)$$

which can also be shown in percentage for each class, which can be called prediction percentage.

Specificity or True Negative Rate for each class is calculated as the ratio of the true negatives of a specific class to the sum of its true negatives and false positives. Based on the confusion matrix introduced in Table 3, this measure can be calculated as [Janney et al., 2020]:

$$Specificity_{Ci} = \frac{N - FP_i - Ci_{Act}}{N - Ci_{Act}} \quad (13)$$

F-measure or F-score is the harmonic mean of precision and recall, ranging from 0 to the optimal value 1 [Hossin and Sulaiman, 2015]:

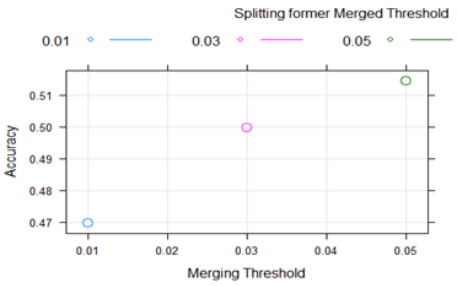
$$F - score_{Ci} = 2 \times \frac{Precision_{Ci} \times Recall_{Ci}}{Precision_{Ci} + Recall_{Ci}} \quad (14)$$

The receiver operating characteristic (ROC) curve graph is a technique for classifier visualization and organization based on their performance [Okasha & Abu-Saada, 2014]. The area under the ROC curve (AUC) is a widely used measure of the performance of supervised classification. The simple form is only applicable to the case of two classes. But here, the definition is extended to the case of four classes by averaging pairwise comparisons defined by [Hand & Till, 2001]. In this study, ROC curves and their corresponding AUCs are calculated for all possible combinations for pairs of classes. Then, the final ROC curve and AUC value for each method is calculated by averaging those pairwise combinations. The ROC plots show the sensitivity (or TPR) and specificity as the output threshold is moved over the range of all possible values [Robin et al., 2011]. The AUC provides a single measure of a classifier's performance for evaluating which model is better on average. A random classifier has an AUC of 0.5, while a perfect classifier has 1. The AUC measure for each pair of classes is computed by obtaining the area of the graphic [López et al., 2013]:

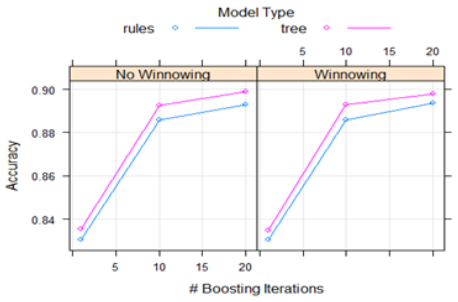
$$AUC = \frac{Sensitivity + Specificity}{2} \quad (15)$$

5. Experimental Results

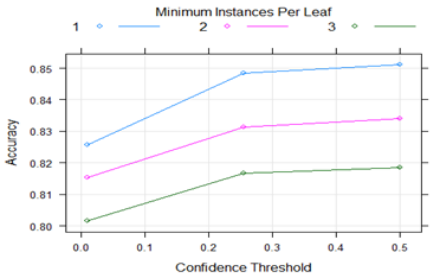
In this section, the decision tree algorithms were compared to each other in the prediction process. Classification evaluation metrics explained in section 4.2 were used for the comparisons. The computer specifications used in this study are Intel® Pentium® CPU B970 @ 3.30 GHz with 8 GB RAM. The fatality severity of the crashes was predicted with four decision tree classification models. For better performance, all the variables were normalized. All of the models were trained by the 10-fold cross-validation method [Rhys, 2020]. Figure 3(a) shows the result of the CHAID tree training process. According to the figure, the ideal values for alpha2 and alpha4 parameters are 0.05. The parameter alpha3 was determined to be -1, which is not shown in the figure. The package 'CHAID' was used to run the model in R. Training the C5.0 tree also showed that the model is superior with the tree structure and 20 trials without winnowing, as shown in Figure 3(b). The package 'C5.0' was used to run the model in R. According to Figure 3(c), C=0.5 and M=1 were determined as the best values for the parameters in order to reach the highest accuracy for the C4.5 tree. To fit the tree in R, the package 'RWeka' was used. According to Figure 3(d), a value of 0.26 was chosen for the complexity parameter in the CART training process. The package 'rpart' was used to run the model in R. Figure 4 shows the diagram of each decision tree resulted from the classification process. Table 4 and Table 5 represent the resulting confusion matrices and classification accuracy metrics for each model on the test set, respectively.



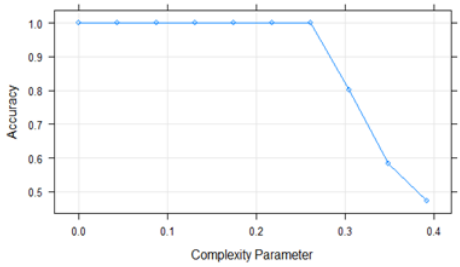
(a) CHAID



(b) C5.0

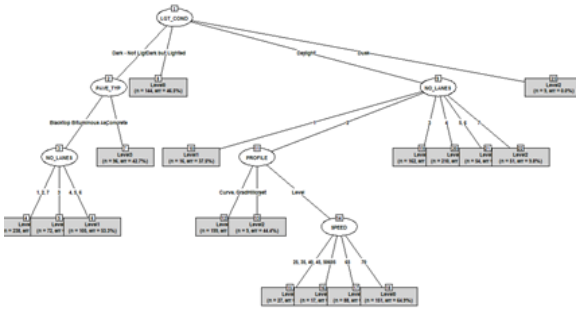


(c) C4.5

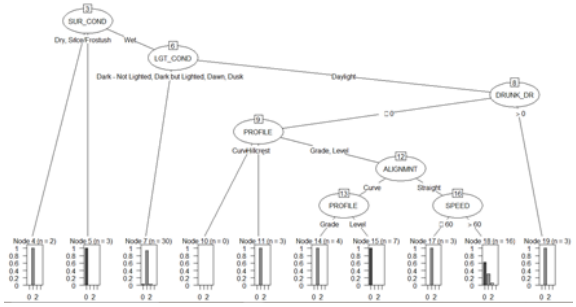


(d) CART

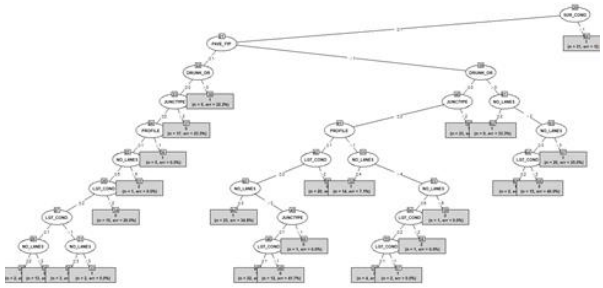
Figure 3. Parameter Tuning of Classification Models



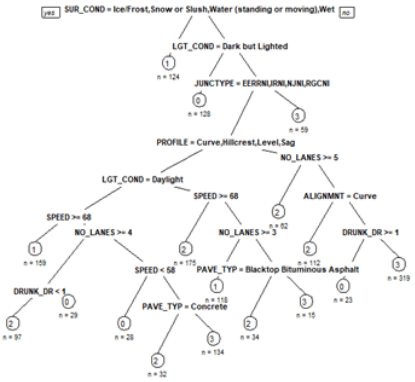
(a) CHAID



(b) C5.0



(c) C4.5



(d) CART

Figure 4. Diagrams of the Decision Tree Classification Models

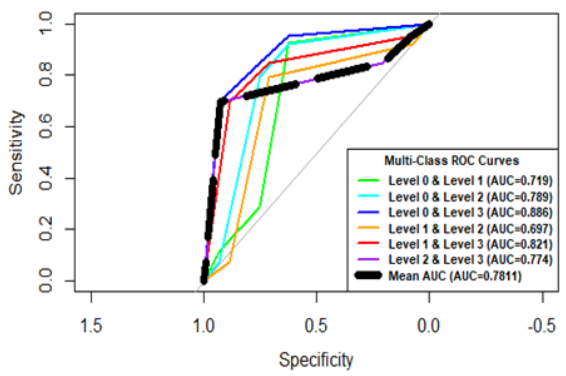
Table 4. Confusion Matrices of Classification Models on the Test Set

Model	Actual	Predicted				
		Level 0	Level 1	Level 2	Level 3	Total
CHAID	Level 0	111	13	6	15	145
	Level 1	37	114	15	30	196
	Level 2	15	42	130	23	210
	Level 3	9	9	8	130	156
	Total	172	178	159	198	707
C5.0	Level 0	127	9	3	3	142
	Level 1	8	167	8	1	184
	Level 2	1	4	208	3	216
	Level 3	1	2	0	162	165
	Total	137	182	219	169	707
C4.5	Level 0	123	14	3	2	142
	Level 1	18	154	12	0	184
	Level 2	1	4	210	1	216
	Level 3	3	1	1	160	165
	Total	135	173	226	163	707
CART	Level 0	81	39	30	4	154
	Level 1	24	81	18	6	129
	Level 2	31	54	147	34	266
	Level 3	6	10	21	121	158
	Total	142	184	216	165	707

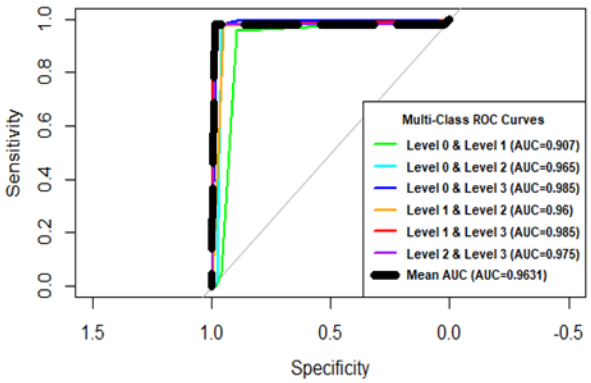
Table 5. Accuracy Evaluation Metrics of Classification Models

Model	Fatality Severity Level	FP	FN	Precision (%)	Sensitivity/ Recall (%)	Specificity (%)	F-measure (%)	Overall Accuracy (%)	Kappa (%)
CHAID	Level 0	61	34	77	65	94	69	67	58
	Level 1	64	82	58	64	84	61		
	Level 2	29	80	62	82	85	70		
	Level 3	68	26	83	66	95	73		
C5.0	Level 0	10	15	92	91	97	92	94	92
	Level 1	15	17	89	93	97	91		
	Level 2	11	8	96	94	98	95		
	Level 3	7	3	98	97	99	97		
C4.5	Level 0	22	19	87	85	97	86	92	89
	Level 1	19	30	84	89	94	86		
	Level 2	16	6	97	93	99	95		
	Level 3	3	5	97	98	99	98		
CART	Level 0	61	73	52	57	93	54	60	47
	Level 1	103	48	62	44	85	51		
	Level 2	69	119	55	68	78	60		
	Level 3	44	37	76	73	87	74		

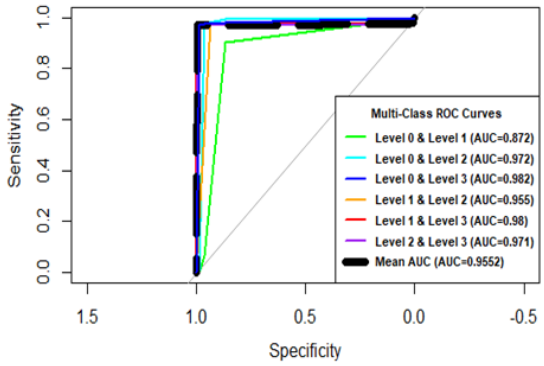
From the non-diagonal elements of the confusion matrices in Table 4 it can be seen that C5.0 had the least wrong predictions and according to Table 5, the tree obtained the best results in comparison with the other methods with an overall accuracy of 94% and a kappa of 92%. Figure 5 indicates the ROC plots of the methods containing the curves and AUCs for the pairwise classes and their average representing each method. Figure 6 shows the risk maps of the classifiers produced by the whole dataset prediction, to inspect the fatality severity distribution throughout the study area.



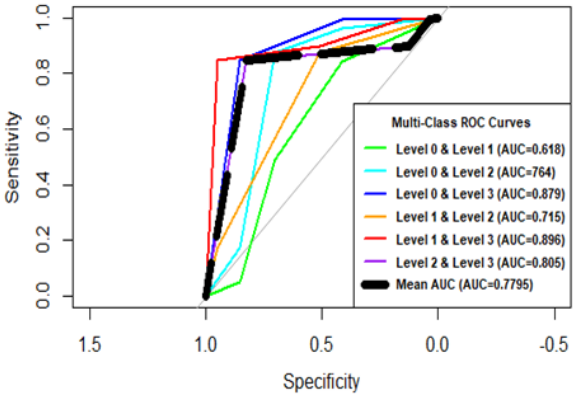
(a) CHAID



(b) C5.0



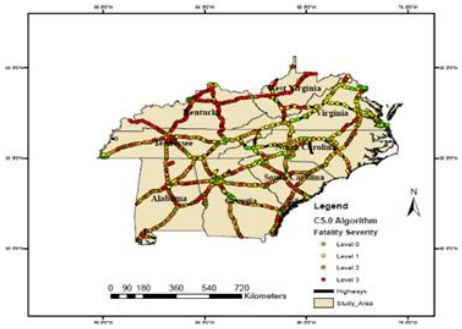
(c) C4.5



(d) CART

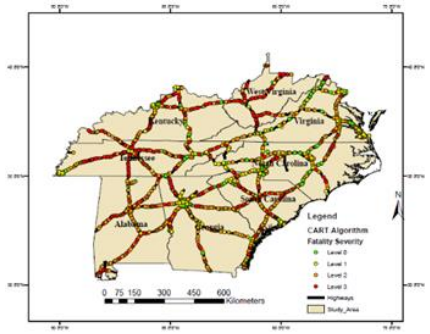
Figure 5. Multi-Class ROC Curves of Classification Models

(a) CHAID



(b) C5.0

(c) C4.5



(d) CART

Figure 6. Risk Maps of Road Fatal Accidents along the Study Area by the Decision Tree Algorithms: (a) CHAID, (b) C5.0, (c) C4.5, and (d) CART

6. Discussion

The study set out to train the classifiers with 70% of the original dataset to make them optimal for the prediction process by parameter tuning, as expressed in Figure 3. After examining the models with the remaining 30% of the dataset as the test set, C5.0 gained an overall accuracy of 94% and kappa of 92%, which were the highest followed by 92% and 89% respectively to C4.5 then by 67% and 58% respectively to the CHAID tree. Therefore, a slight distinction in the performance of C5.0 and C4.5 trees was observed, since the former outperformed 2% and 3% in terms of overall accuracy, and 3% in terms of kappa, respectively. Also, it performed slightly better in other accuracy metrics, but some of the differences are almost insignificant. As mentioned in section 4.1.2, this can be due to the boosting technique applied to the C5.0 tree to improve its performance. CART's worst performance stands out in Table 5 in comparison with the other three methods, as the tree provided the lowest values for all the accuracy evaluation metrics. A possible explanation for the CART's weak performance in the classification process can be that the outcomes are restricted by the smaller size of the data. Therefore, this issue may raise the importance of a suitable dataset selection for prediction models. For more efficiency, the model can be trained with different methods. Specifically, the variables could be extended to more driver/time-related crash factors. High values for these two measures may be due to the appropriate proportion of instances in the data set. In general, these two metrics cannot be always reliable and the model performance should not be based on them. As a result, other metrics like precision, recall, specificity, and F1-measure should be taken into consideration. Accordingly, these four metrics were achieved on the test set, calculated via Table 4, and presented in Table 5. One can see that C5.0 obtained superior values for these measures. To further compare these metrics for each class, the highest values for precision were observed in level 3 of the crashes in all the models.

On the other hand, CHAID, C5.0, and C4.5 gained the lowest precision on Level 1 of the crashes, while CART had the lowest precision in Level 0 of the crashes. Considering recall, the highest values in all the classifiers also belonged to Level 3 of the crashes, except the CHAID tree which had the highest recall in Level 2 of the crashes. This comparison conveys the notion that the models had superior predictions in these levels rather than the remaining two. This measure was also the lowest in Level 0 and Level 1 of the crashes. The highest and lowest values for the specificity in the CHAID, C5.0, and C4.5 algorithms were observed in Level 3 and Level 1 of the crashes, respectively; However, CART appeared differently, as it gained the highest and lowest specificity values in Level 0 and Level 2 of the crashes, respectively. The F-measure values in Table 5 proved that the relatively balanced data highly affected the model overall accuracy in predicting class labels of the test set. The highest F-measure values were also observed in Level 3 of the crashes in each classifier, which is owing to the fact that the classifiers have better predicted this level.

The ROC curves of the classification models with their AUC values are shown in Figure 5. The x-axis represents sensitivity or recall values, while the y-axis represents specificity values. The multi-class AUC was calculated based [Hand & Till, 2001] method. In this study as a classification problem with more than two classes, this multi-class AUC was obtained by averaging. The classes were paired and then the ROC of each pair of classes was plotted with separate colors and their AUCs were calculated and given, as shown in the figure. Based on this method, the average AUC for each model was

calculated from the AUC values of the pairwise classes, where the average ROC is shown as a black dashed line graph. The more inclined the curve is toward the upper left corner, the better is the classifier's power to discriminate between the classes. As well as similarities found in the performance of C5.0 and C4.5 trees as shown in Table 5, it was also visualized in the resulting ROC curves as represented in Figure 5. Where there can be difficulties encountered in the comparison of ROC curves, the AUC can sort models by their overall performance. As a result, the AUC is more a determining factor in the assessment of the models rather than ROC. Thus, it can be concluded that by comparing the mean AUC values achieved in the models, the C5.0 tree is considered as the best-performing classifier in this study with the highest AUC value of 0.9631. Regarding the risk maps in Figure 6, red spots indicating Level 3 of the crashes were mainly observed in Tennessee, Kentucky, and West Virginia states. The risk maps resulted from C4.5 and C5.0 algorithms seem similar, however, the minor differences can be clearly seen in Tennessee, Virginia, and North Carolina states. It is apparent that C5.0 classified Level 2 crashes more than the other methods. Moreover, the fatality severity dispersion is most diverse in CART risk map.

7. Conclusion

Analysis of road fatal accidents is of great significance; because the irrecoverable costs of these accidents have a profound impact on society. Given the rising fatality rate worldwide and causing life and property damages, the occurrence prediction of these accidents, and identifying the high-risk areas should be highly emphasized. Therefore, we can prevent these accidents as much as possible by taking these measures, educating the public, enacting effective management laws and policies, and more oversight to deal with the accident factors increasing fatality rate. In this paper, a comparative study was conducted with the purpose of investigating the prediction power of tree-based algorithms for fatal crashes. The study used 2355 fatal crash records that occurred on the roadways of the US National Highway System from 2007 to 2009. In order to achieve the objectives of this study, four decision tree classification methods such as CHAID, C4.5, C5.0, and CART were proposed. The hyper parameters of the models were optimally tuned by training the models through 70% of the whole data as the training set to achieve their best performance. Through validation with the remaining 30% of the data as the test set and the metrics discussed in section 4.2, the C5.0 tree proved to be an outperforming model in the study that can be applied to make effective predictions with multiple factors associated with these fatal crashes. This task is an application of data mining, which has proven to be reliable and yield productive results. The proposed methodology is suitable for discovering meaningful information. However, the results are quite general as the data lacks a wide variety of variables, such as various driver-related factors. More information can be revealed by having more complementary variables in accidents such as age, gender, type of license, and driver education level. The methodology of this paper can be a field for extensive experiments and improvements. In further analyses, a cost-based crash approach can be applied to use overall crash costs in order to train the models.

8. References

- Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting students' performance using ID3 and C4. 5 classification algorithms. ArXiv Preprint ArXiv:1310.2071.

- Ahmed, A. M., Rizaner, A., & Ulusoy, A. H. (2018). A novel decision tree classification

based on post-pruning with Bayes minimum risk. Plos One, 13(4), e0194168.

- Bahiru, T. K., Singh, D. K., & Tessfaw, E. A. (2018). Comparative study on data mining classification algorithms for predicting road traffic accident severity. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 1655–1660.

- Blazquez, C. A., & Celis, M. S. (2013). A spatial and temporal analysis of child pedestrian crashes in Santiago, Chile. Accident Analysis & Prevention, 50, 304–311.

- Chang, L.-Y., & Wang, H.-W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention, 38(5), 1019–1027.

- Choi, J., Gu, B., Chin, S., & Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, 102974.

- de Oña, J., López, G., & Abellán, J. (2013). Extracting decision rules from police accident reports through decision trees. Accident Analysis & Prevention, 50, 1151–1160.

- Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health, 4, 118–131.

- Diez, P. (2018). Chapter 1—Introduction. In P. Diez (Ed.), Smart Wheelchairs and Brain-Computer Interfaces (pp. 1–21). Academic Press.

- Effati, M., Thill, J.-C., & Shabani, S. (2015). Geospatial and machine learning techniques for wicked social science problems: Analysis of crash severity on a regional highway corridor. Journal of Geographical Systems, 17, 107–135.

- Fatality Analysis Reporting System. (2019, July17). NHTSA.https://www.nhtsa.gov/crash-data-systems/fatality-analysis-reporting-system

- Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2), 171–186.

- Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1.

- Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. International Journal of Advanced Computer Science and Applications, 4(2), 13–19.

- Janney, J. B., Roslin, S. E., & Kumar, S. K. (2020). 6—Analysis of skin lesions using machine learning techniques. In J. K. Verma, S. Paul, & P. Johri (Eds.), Computational Intelligence and Its Applications in Healthcare (pp. 73–90). Academic Press.

- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical

data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(2), 119–127.

- Kumar, S., & Toshniwal, D. (2017). Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India. European Transport Research Review, 9(2), 24.

- Lee, J., Yoon, T., Kwon, S., & Lee, J. (2020). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study. Applied Sciences, 10(1), 129.

- Lin, C.-L., & Fan, C.-L. (2019). Evaluation of CART, CHAID, and QUEST algorithms: A case study of construction defects in Taiwan. Journal of Asian Architecture and Building Engineering, 18(6), 539–553.

- López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141.

- Mansouri, M., & Javad Kargar, M. (2014). Analysis and monitoring of the traffic suburban road accidents using data mining techniques; a case study of Isfahan Province in Iran. The Open Transportation Journal, 8(1).

- Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: A review. Procedia Manufacturing, 35, 698–703.

- Okasha, M. K., & Abu-Saada, A. H. (2014). Modeling violence against women in Palestinian society. American International Journal of Contemporary Research, 4(1), 209–220.

- Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

- Quinlan, J. R. (1993). C4. 5: Programming for machine learning. Morgan Kauffmann, 38, 48.

- Rhys, H. (2020). Machine Learning with R, the tidyverse, and mlr. Manning Publications.

- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). PROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12(1), 1–8.

- 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. Transport, 32(3), 272–281.

- Shanthi, S., & Ramani, R. G. (2011). Classification of vehicle collision patterns in road accidents using data mining algorithms. International Journal of Computer Applications, 35(12), 30–37.

- Susanti, Y., Zukhronah, E., Pratiwi, H., & Sri Sulistijowati, H. (2017). Analysis of Chi-square Automatic Interaction Detection (CHAID) and Classification and Regression Tree (CRT) for Classification of Corn Production. JPhCS, 909(1), 012041.

- Tallón-Ballesteros, A. J., & Riquelme, J. C. (2014). Data mining methods applied to a

digital forensics task for supervised machine learning. In Computational Intelligence in Digital Forensics: Forensic Investigation and Applications (pp. 413–428). Springer.

- Thakali, L., Kwon, T. J., & Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: A comparison. Journal of Modern Transportation, 23(2), 93–106.

- United States. National Highway Traffic Safety Administration. (2006). This is NHTSA : people saving people. Washington, D.C. : U.S. Dept. of Transportation, National Highway Traffic Safety Administration, 2006.

- Wang, S., & Li, Z. (2019). Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. PloS One, 14(3), e0214550.

- Wang, Y., Li, Y., Song, Y., Rong, X., & Zhang, S. (2017). Improvement of ID3 algorithm based on simplified information entropy and coordination degree. Algorithms, 10(4), 124.

- World Health Organization, W. H. (2018). Global status report on road safety 2018: Summary. World Health Organization.

- Xing, L., He, J., Li, Y., Wu, Y., Yuan, J., & Gu, X. (2020). Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza. Accident Analysis & Prevention, 135, 105343.

- Yuan, Y., Wang, S., Liu, Z., Cui, G., & Wang, Y. (2020). Influencing factors analysis of side right-angle collisions severity at intersections based on decision tree. International Journal of Crashworthiness, 1–11.

- Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access, 6, 60079–60087.