ASEAN Engineering Journal

Full Paper

CRASH PREDICTION ON ROAD SEGMENTS USING MACHINE LEARNING METHODS

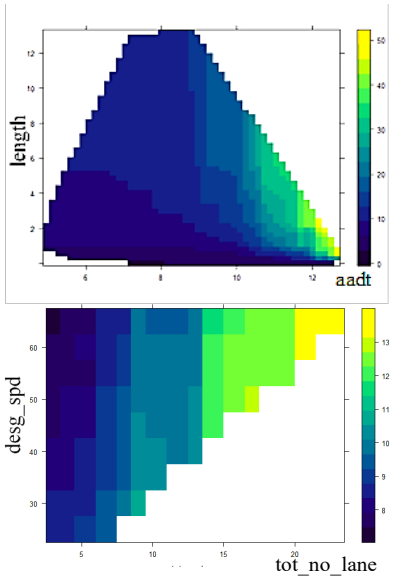
Agustin Guerra[a], Vivek Gadhiya[a], Punyaanek Srisurin[b]*

[a]Department of Civil and Coastal Engineering, University of Florida, Gainesville, FL, USA.
[b]Transportation Institute, Chulalongkorn University, Bangkok, Thailand.

Article History
Received
07 September 2021
Received in revised form
20 March 2022
Accepted
21 March 2022
Published online
31 August 2022

*Corresponding Author
punyaanek.s@chula.ac.th

Graphical Abstract



Abstract

This study adopted the Highway Safety Information System's (HSIS) data for crashes occurred on road segments to develop supervised machine learning prediction models. Five machine learning models are developed: Linear Regression (LR), Generalize Additive Model (GAM), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). A comparison among the five model was performed using the root mean square error (RMSE) and the mean absolute error (MAE) as quality model indicators. The results indicated that the RF model was found to produce the best crash prediction results. The findings suggested that the increase in Annual Average Daily Traffic (AADT) exponentially increased the number of crashes on highway segments. In addition, roadway segments with the higher design speed induced the lower number of crashes, compared to the segments with the lower design speed. For segments of shorter than 5-mile long, the number of crashes rapidly increased as the segment length increased. However, there was no substantial increase in the number of crashes as the segment length increased for segments of longer than 5 miles. Also, the greater number of lanes on a roadway segment, the greater chance for increasing the number of crashes. Finally, the moderate grades showed the highest risk for occurrences of crashes, respectively followed by flat and rolling grades. These findings are useful for transportation professionals to consider when designing highways.

Keywords: Machine Learning, Statistical Learning, Crash Prediction, Random Forest, Linear Regression, Support Vector Machine, Artificial Neural Network.

1.0 INTRODUCTION

Fatalities due to vehicle crashes in the U.S. were reported to be over 36,000 deaths per year in 2019. California alone contributed to approximately 10% of the total deaths [1]. Specifically, approximately 9.2 deaths per 100,000 population solely occurred in California. After several years of research and development on highway safety, crashes still remain one of the leading causes of death in the U.S [2]. Some of the underlying reasons include driver's behavior, climate condition, and geometric design of roadways. Crashes take place as a combination of several factors (i.e., pre/post-crash events). In other words, if the underlying factors can be identified based on the data obtained from crash incidents, the crash occurrences and crash severities can be alleviated.

There are several simulations and data-driven technics available for understanding, analyzing, and predicting crash occurrences. For instance, Pasquale et al. 2021 [3] reviewed the different levels of traffic safety simulation technics and suggested that there were very few simulation models. Also, the authors argued that, for understanding the relative causes and effects of road traffic crashes, expertise in traffic safety was required. In addition, most of the simulation tools were found to require expensive license purchasing and barely consider conflicts in overlapping trajectories between objects. In contrast, statistical models, such as regression analysis, have been widely used to estimate the number of crashes occurred. Unlike simulation tools, statistical models are license free since the prediction models are usually developed by the researchers themselves. Data driven models, such as machine leaning, have been widely

applied in several fields; however, its implementation in the crash prediction is still underexplored [4]. In an attempt to identify the leading factors and crash occurrence prediction based on machine learning concepts, this study performed an analysis on the Highway Safety Information System's (HSIS's) database of the state of California [5]. The HSIS database contained crash occurrences on roadway segments. In this data different features of roadway, such as type of roadway, the number of lanes, average annual daily traffic, and divided or undivided roadway, were taken into account [6]. We proposed five machine learning models that estimate the number of crashes occurred per year based on the given conditions of the roadway segments. The RF model was found to provide higher accuracy based on the resultant RMSE and MAE. The most important factors leading to crashes on roadway segments were identified, and interpretation of these factors was provided.

The remaining part of this study is structured as follows. Section 2 reviews the classical model prediction for crashes. Section 3 presents the data preparation for this study. Section 4 and 5 show the model results and exploratory analysis. Section 6 shows the model's overview and comparison while Section 7 provides the model interpretation and conclusion.

2.0 LITERATURE REVIEW

2.1 Classical Approaches

Several studies have been conducted to establish relationships between crashes and their explanatory variables. For several decades, most of the studies merely focused on developing regression models considering the available data. However, with the recent increase in data availability, this limitation had enormously lessened, allowing the implementations of new models for crash prediction. Most of the statistical approaches, such as the Poisson and negative binomial (NB) approaches, belong to the Gaussian family [7].Their variants in univariate and multivariate regression outlines were successfully applied in crash prediction models. Even though these models presented doubtful prediction in many cases, they helped explain the associations between the influential factors and the occurrences of road traffic crashes [8].

Most of the models were specific to the facility type, such as rural two-way highway, rural multilane highways, urban arterial highways [7, 8]. However, there were no accurate generic models for crash prediction. Moreover, one of the major concerns for these models was the prediction's errors caused by the nature of the driving task itself. This forced transportation professionals to look for various solutions to adjust the crash prediction models. One of the most practical methodologies for increasing the accuracy of these models is the process described in the Highway Safety Manual, considering the regression to the mean effect [3, 4]. In this process, the prediction is adjusted through weighting factors that increase or reduce the predicted values from the models to account for the uncertainty of the human driving task.

2.2 State-of-the-Art Approaches

According to the literatures, machine or statistical learning is referred to a set of tools adopted to summarize and perform in-depth analysis on the data for the purpose of understanding the outcomes. In general, this set of tools is categorized into two types: supervised and unsupervised learning. Supervised learning involves building statistical models to perform predictions or categorizing data regarding a set of independent variables from the data [5, 6].

Alternative to the classical approaches for crash prediction, machine learning approaches were recently adopted in this field [14]. The main advantage of this approach is that the models are created and enhanced using tested error data. This allows to create a single model with higher prediction power, compared to other statistical approaches. Recently, more flexible tools, such as deep learning and random forest approaches, are found to be implemented in this field. Deep learning models possess the advantage of permitting computational models to learn the representations of data with multiple levels of abstraction [8]. In addition, deep learning methods can deal with non-linear data; however, this type of model had a trade-off between interpretability, computation demand, and accuracy. Therefore, deep learning models are not included in this study since they are considered as a black box, and they usually require a very large training data set. Furthermore, as mentioned before interpreting the deep learning model results is very challenging.

Although most of the crash prediction models adopted Linear Regression [15, 16], recent studies are found to prefer applying Generalize Additive Model (GAM) to predict crash frequencies [17–22]. However, very few studies were found to apply Random Forest (RT) to predict crash occurrences on highways [23, 24]. Pham et al. developed random forest models for identifying motorway rear-end crash risks by using disaggregate data [23]. Jiang et al. [24] conducted a study to investigate the feasibility of using random forest for identifying macro-level crash risk. There have been some studies in the literature that utilize Support Vector Machine (SVM) to develop crash prediction models for predicting crash frequency [25–27]. Several previous studies were found to apply artificial neural network models to quantify crash frequency [28–34]. A study applied an Artificial Neural Network (ANN) approach to model highway traffic crash frequency and found that this approach was more effective than the Poisson regression and the negative binomial regression models [33]. Furthermore, some previous studies attempted to develop real-time crash prediction models for predicting crash occurrences on highways [22, 27, 34–37].

2.3 Road-Related Crash Contributing Factors

Several studies attempted to quantify risk factors that contribute to the occurrences of crashes, as well as the magnitudes of their influences on highways. Roadway geometry was found to have the highest influence in estimating the crash occurrence rate since it affected the operational speed of the vehicles [33].

Several studies found that greater AADT led to the increase in the number of crashes occurred on roadway segments [16, 32, 33, 38, 39]. Some research concluded that length of a highway segment also had an impact on the crash frequency. Previous studies found that the crash frequency tended to increase with respect to the length of highway segments to some degree [26, 36, 38, 40]. The increase in the degree of curvature was also found to produce the higher number of crashes on curves by several previous studies, since sharp curves worsen the stability of vehicles in terms of slippage and overturns [36, 38, 41].

A study found that road segments with the higher number of lanes typically induced unsafe driving behaviors, which led to the greater chances for crashes to occur [42–44]. However, a study rebutted that roadway segments with the greater number of lanes tended to induce the smaller number of crashes [36]. In addition, speeding was pinpointed by most of the previous studies as a major cause of crash occurrences [26, 39, 45–49]. Recent studies developed crash prediction models and found that the abrupt transition of speed within the roadway section usually existed at the time crashes occurred [23, 35]. This study constructed five supervised learning models and evaluated the crash prediction performances of these models on highway segments. Although crashes were found to occur with regard to interactions between three main crash contributing factors: human-related factors, roadway-related factors, and vehicle-related factors [50], this study intended to merely focus on the influences of roadway characteristics reflecting frequencies of crashes occurred on roadway segments. We examined linear regression models, tree-based models, artificial neural networks, random forest, and support vector machine models for crash number prediction. In contrast to the Highway Safety Manual (HSM) models, where the focus is facilities with similar characteristics, the main purpose of this study was to attempt a development of a generic model to predict the number of crashes on highway segments.

3.0 METHODOLOGY

3.1 Data Overview

The data for this project was obtained from the Highway Safety Information System (HSIS) (https://www.hsisinfo.org) . The data was collected on highway segments in California. The dataset contained 17,959 datapoints of crashes occurred during the year 2003 on highway segments in California. Nevertheless, while the current data were being pre-processed, there were still some variables that could be discarded for the purpose of this analysis. For instance, as in the case of CNTYRPE variable, of which the useful information was not provided since geographic information was not considered under this analysis. Table 1 provides an overview of the collected dataset for this project.

Table 1 Data Overview

Label	Definition	Categories, Descriptive Statics
ID	A number corresponds to the ordinal event ordered from 1 to n crash reported.	ID corresponds to a numeric value from 1 to 17,959. This variable identifies a reported crash on a segment.
CNTYRTE	County route of the roadway segment	CNTYRTE corresponds to a numeric value from 1 to 17,959. This variable only identifies a reported crash on a segment.
BEGMP	Calculated begin milepost of the segment	These variables were recoded to take into account the segment length.
ENDMP	Calculated end milepost of the segment	
NO_LANE2	Number of through lanes towards increasing/decreasing mile points	Number of lanes; including through, HOV and other auxiliary lanes, of greater than 0.2 miles in length.
DIVIDED	Divided or Undivided highway	This is a dummy variable: 1 for a divided roadway and 0 for an undivided roadway.
MED_TYPE	Median type on the divided roadway segment (categorical variable)	'A' = Undivided, Not Separated or Striped 'B' = Undivided, Striped 'C' = Undivided, Reversible Peak Hour Lane(s) 'E' = Divided, Reversible Peak Hour Lane(s) 'F' = Divided, Two-Way Left Turn Lane 'G' = Divided, Continuous Left-Turn Lane 'H' = Divided, Paved Median 'J' = Divided, Unpaved Median 'K' = Divided, Separate Grades 'L' = Divided, Separate Grades with Retaining Wall 'M' = Divided, Sawtooth (Unpaved) 'N' = Divided, Sawtooth (Paved) 'P' = Divided, Ditch 'Q' = Divided, Separate Structure 'R' = Divided, Railroad or Rapid Transit 'S' = Divided, Bus Lanes 'T' = Divided, Paved Area, Occasional Traffic Lane 'U' = Divided, Railroad and Bus Lane 'V' = Divided, Contains Reversible Peak-Hour Lane(s) 'Z' = Divided, Other '-' = Invalid Data '+' = No Data

Table 1 Data Overview (Cont. 1)

Label	Definition	Categories, Descriptive Statics
NO_LANE1	Number of through lanes towards increasing/decreasing mile points.	Number of lanes; including through, HOV, and other auxiliary lanes, of greater than 0.2 miles in length.
HWY_GRP	Highway group based on the alignment and cross-section characteristics (categorical variable)	'R' = Right Independent Alignment 'L' = Left Independent Alignment 'D' = Divided Highway 'U' = Undivided Highway 'X' = Unconstructed 'Z' = Other '-' = Invalid Data '+' = No Data Other Error/Other Codes
ACCESS	Access control of the highway segment (categorical variable)	'C' = Conventional - No Access Control 'E' = Expressway - Partial Access Control 'F' = Freeway - Full Access Control 'S' = One-Way City Street - No Access Control '-' = Invalid Data '+' = No Data Other Error/Other Codes
TERRAIN	Terrain Type (categorical variable)	'M' = Mountainous 'R' = Rolling 'F' = Flat '-' = Invalid Data '+' = No Data
DESG_SPD	Design Speed (in mph)	Range: 25 to >70 mph Min: 25.0 1st Quartile: 60.0 Median: 70.0 Mean: 62.5 3rd Quartile: 70.0 Max: 70.0
AADT	Average Annual Daily Traffic (ADT within a year)	Range: 0 to > 40,000 Min: 120 1st Quartile: 13,500 Median: 45,000 Mean: 87,901 3rd Quartile: 150,000 Max: 371,317
RURURB	Zone type for identifying whether the segment is in rural or urban setting (categorical variable).	'R' = Rural 'U' = Urban '-' = Invalid Data '+' = No Data
RODWYCLS	Classification of the roadway where the crash occurred (categorical variable).	'01' = Urban Freeways '02' = Urban Freeways of greater than 4 Lanes '03' = Urban Two-Lane Roads '04' = Urban Multilane, Divided, Non-Freeways '05' = Urban Multilane Undivided, Non-Freeway Urban Multilane, Undivided Non-Freeways '06' = Rural Freeways '07' = Rural Freeways of less than 4 Lanes '08' = Rural Two-Lane Roads '09' = Rural Multilane, Divided, Non-Freeways '10' = Rural Multilane, Undivided, Non-Freeways '99' = Others
CRTOT_03	Dependent variable which is an integer that represents the number of crashes per observation.	Min: 1.000 1st Quartile: 1.000 Median: 3.000 Mean: 7.464 3rd Quartile: 8.000 Max: 162.000

3.2 Data Preparation and Processing

The variables that were not considered significant on explaining the model were dropped. The excluded variables were identification number (ID), begin milepost (BEGMP), end milepost (ENDMP), number through lanes towards increasing and decreasing mile points (No_Lane1 and No_Lane2). Note that for the last four variables, two new variables were recoded to account for the segment length (LENGTH = ENDMP - BEMP) and for the number of lanes in the segment (No_Lanes = No_Lane1 + No_Lane2). Missing data affects the error of the predictive model; therefore, an educated procedure to deal with missing data problems was considered [6–7, 51]. The multivariate imputation by change (MICE) package was used to fill the missing data [52–54]. The selected imputation method is predictive mean matching (PMM). This method does not assume a particular distribution of the data, such as the regression and other methods. In addition, this method has been applied in the previous studies [9–10]. Figure 1 shows the histogram of the missing data values per each variable. As shown, the independent variables indicating the number through lanes towards increasing/decreasing mile points (No_Lane1 and No_Lane2) possessed the greater number of missing values compared to the other variables in the dataset. Therefore, these missing data were fulfilled by using MICE package.

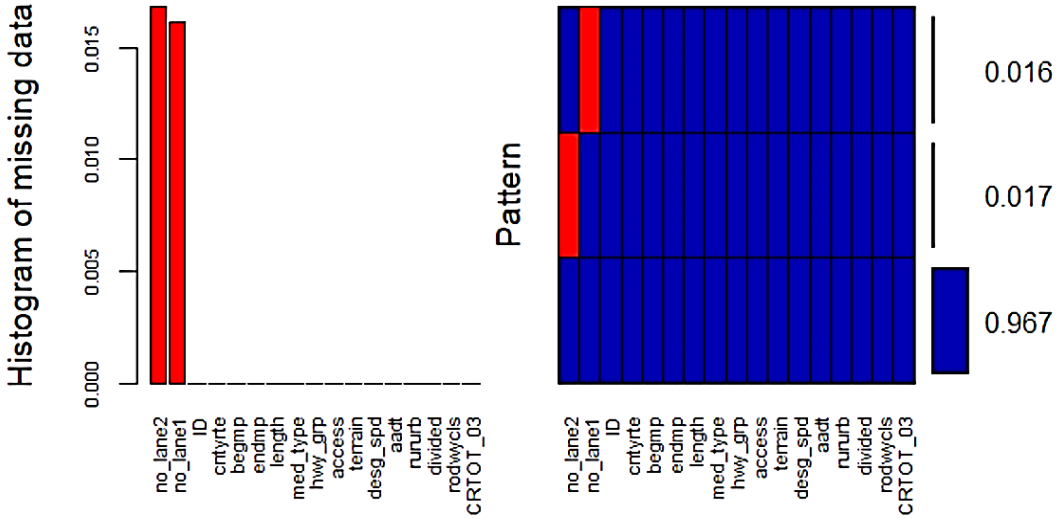


Figure 1 Histogram of the missing data

3.3 Exploratory Analysis

This section seeks to identify possible trends in the number of crashes and identify if the data is reasonable. For the exploratory analysis, a descriptive statistic was performed. Table 1 shows the descriptive statistics of the numerical variables in the dataset. It is worth mentioning that the AADT variable was transformed to logarithmic scale. Therefore, a conversion of this variable was considered for the final model. On the other hand, the number of crashes per segment variable (CRTOT_03) showed a Poisson distribution, as shown in Figure 2.

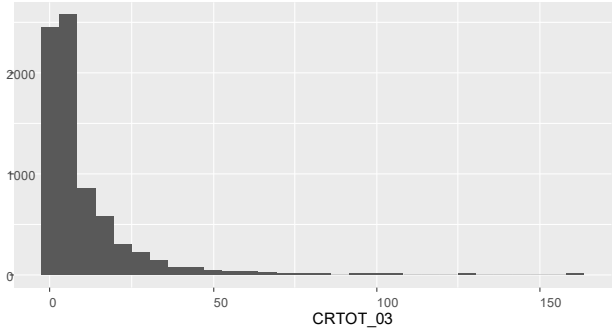


Figure 2 Distribution of crash counts (CRTOT_03)

The variation of the number of crashes (CRTOT_03) was considered in conjunction with length of the segment and AADT, to evaluate the effect of it with the number of crashes. These variables were selected based on the HSM typical predictors for crashes. It was found that the number of accidents increased with the AADT, and decreased with the length of the segment, as shown in Figure 3 and Figure 4, respectively.

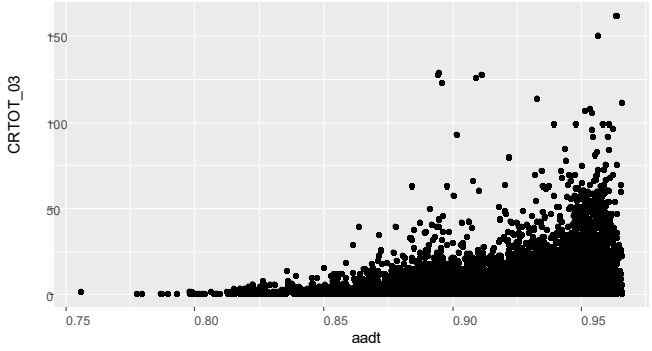


Figure 3 Number of Crashes versus AADT

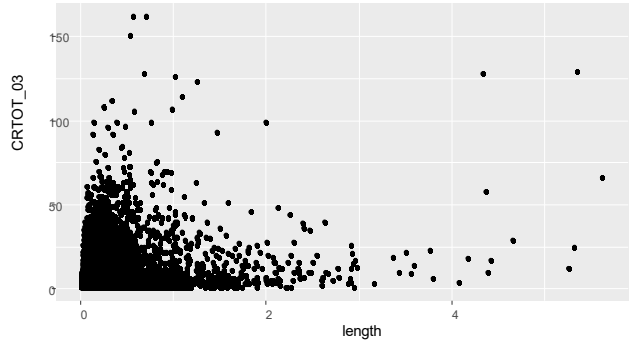


Figure 4 Number of Crashes versus segment length

4.0 MODELS AND RESULTS

This section summarizes the models that were implemented in this study. As mentioned before, this research considered five models to predict the number of crashes per year on highway segments. The first model considered was a linear regression (LR) model, followed by a general additive model (GAM). For these two models, cross-validation was implemented to tune the final models. The third and fourth models implemented were a random forest (RF) and support vector machine (SVM) which were tuned using the caret and e1071 packages, respectively [55, 56]. Finally, an artificial neural network (ANN) approach was attempted. Here, the hyper-parameter tuning was completed through the NNET package. A comparison among different models was performed using the root-mean-square error (RMSE) and the mean absolute error (MAE).

4.1 Linear Regression Model (LR Model)

The Linear Regression model (LR model) was performed considering the important or significant variables after a preliminary evaluation [11]. For simplifying the interpretability, only five variables were considered. However, the results obtained were not superior to the results yielded by the other models.

For the LR model, a full model and a reduced model were compared. For the complete model, all variables were considered, whereas only five variables were considered in the reduced model. The error introduced for having a reduced model is barely significant, which can be proved by running an ANOVA analysis. Therefore, the selected LR model only contained five predictors, as represented by Equation (1).

$$\begin{aligned} \text{Crashes} = {} & 5.4 + 3.2(\text{segment length}) + 5.8\times10^{5}(\text{AADT}) \\ & - 390(\text{DESG\_SPD}) - 2.4\times10^{1}(\text{lanes}) \\ & + 65(\text{RODWYCLS02}) - 2.6(\text{RODWYCLS03}) \end{aligned} \quad (1)$$

According to Equation (1), the explanatory variables in the linear regression model showed that the segment length, AADT, and RODWYCLS02 (a type of less-than-4-lane urban freeway) increase the crash occurrences; while the numbers of lanes, design speed, and RODWYCLS03 (a type of urban two-lane road) decrease the occurrences of crashes. The nomenclature of these parameters is seen in Table 1.

4.2 Generalize Additive Model (GAM Model)

Generalize Additive Models (GAM) are an extension of the linear regression models, with the particularity that allows non-linearity among predictors [12]. Initially, all the variables were evaluated, and the prediction was measured. Considering that not all the variables provided good explanations for the number of crashes, the number of variables was reduced from 11 to 4 variables (i.e., AADT, design speed, length of the segment, and number of lanes). Consequently, a hyper-parameter tuning was performed using cross-validation to determine the degree of the predictors in the model.

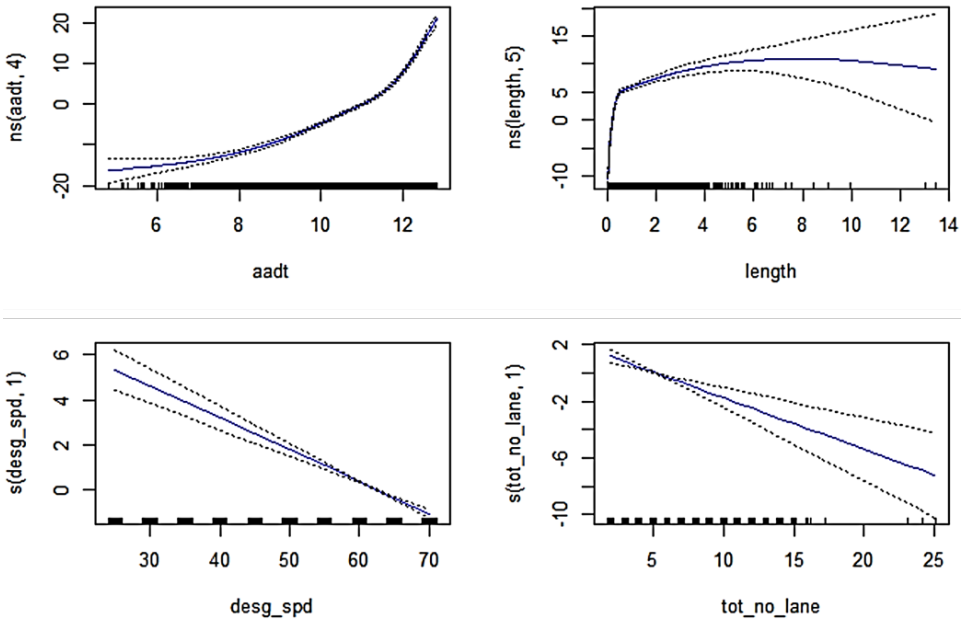


Figure 5 GAM plots for selected variables

Following the selection in the LR model, the GAM model was tuned using cross-validation testing to different degrees for the 4 selected predictors. The equation of the best GAM model produced by this approach is displayed by Equation (2). The characteristic of each variable in this model is illustrated in Figure 5.

$$\text{Crashes} = \text{AADT}^4 + \text{length}^5 + \text{DESG\_SPD} + \text{lanes} \quad (2)$$

Regarding Equation (2), the occurrences of crashes are contributed by the 4th degree polynomial of AADT, followed by the 5th degree polynomial of the segment length, and linear contributions of the design speed and number of lanes of the segment, respectively.

4.3 Random Forest Model (RF Model)

Random forest models (RF models) take into account the ideal trees candidates for bagging and capturing the complex interaction of the data structure, [13]. In this model, a grid search was implemented by varying the tuning parameters "ntree" and "mtry". In this case, the number of variables in the final model were reduced from 11 to 7 variables, based on their importance.

The RF model showed the best results, compared to the other models. Similar to the other models, the final model was a reduced version of the complete model. In this case, the RF model contained seven predictors: AADT, length of segment, design speed, total number of lanes, median type, roadway classification, and terrain, based on the importance of these variables and the errors, as depicted in Figure 6.

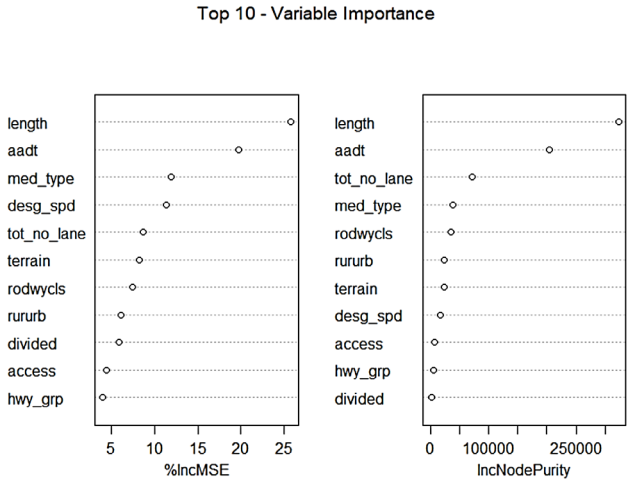


Figure 6 Top-ten variables in the RF model

4.4 Support Vector Machine Model (SVM Model)

The hyperplane-based support vector machine models (SVM) were tuned using the caret package. In this case, the tuned parameters were epsilon, gamma, cost, and the polynomial degree of the kernel. For the SVM model tuning, the considered parameters were cost (20, 50, 60, 100), degree (1, 2, 3), gamma (0.5, 1, 2), and epsilon (0, 1, 0.1). The best SVM model tune was found at epsilon = 0.3, cost = 20, gamma = 0.5, polynomial degree = 2.

4.5 Artificial Neural Network Model (ANN Model)

The artificial neural network model (ANN model), which is a more flexible machine learning approach, did not show favorable results. In this case, only one layer was considered. The parameters tuned were the decay and the net size. All variables were considered; however, reducing the number of variables from 11 to 7 variables showed the same results when using the considered indicators.

For the ANN, similar to the LR model, the complete and the reduced models were evaluated. In this case, the reduced and the complete models showed the same margins of error for both RMSE and MAE. Therefore, the reduced model was selected for its convenience in interpretation. For the hyper-parameter tuning of the model, the decay in the range of 0.0–01 - 0.1 and the size (1, 3, 5) was evaluated. Similar results were found for each model combination in terms of RMSE and MAE. Therefore, the model selection was executed by the R2 value, which was better for the decay of 0.1 and the size of 5. Figure 7 shows the best ANN model.

4.6 Models Comparison

Model comparison was performed for all models using the RMSE and MAE. The following sections show the in-depth results for each model.

Table 2 shows the resultant RMSE and MAE of all five models, as sorted from the worst to the best results. The statistical indicators indicated that the RF model was the model that performed the best in predicting the number of crashes based on both indicators. The SVM also showed the slightly worse results compared to the RF model, followed by the GAM and LR models, which showed the similar results. Finally, the worst model appeared to be the ANN model, as seen in Table 2.

Table 2 Model Comparison

Model	RMSE	MAE
ANN	14.000	6.671
LR	9.477	5.428
GAM	9.294	5.444
SVM	8.932	4.631
RF	8.284	4.281

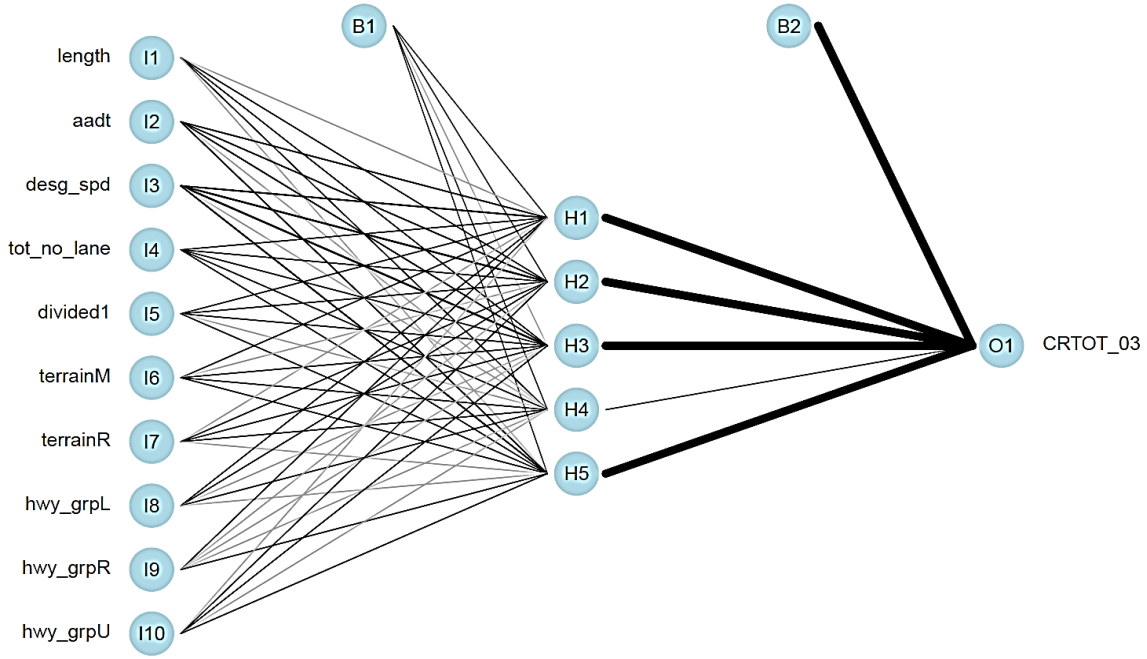


Figure 7 Best ANN model

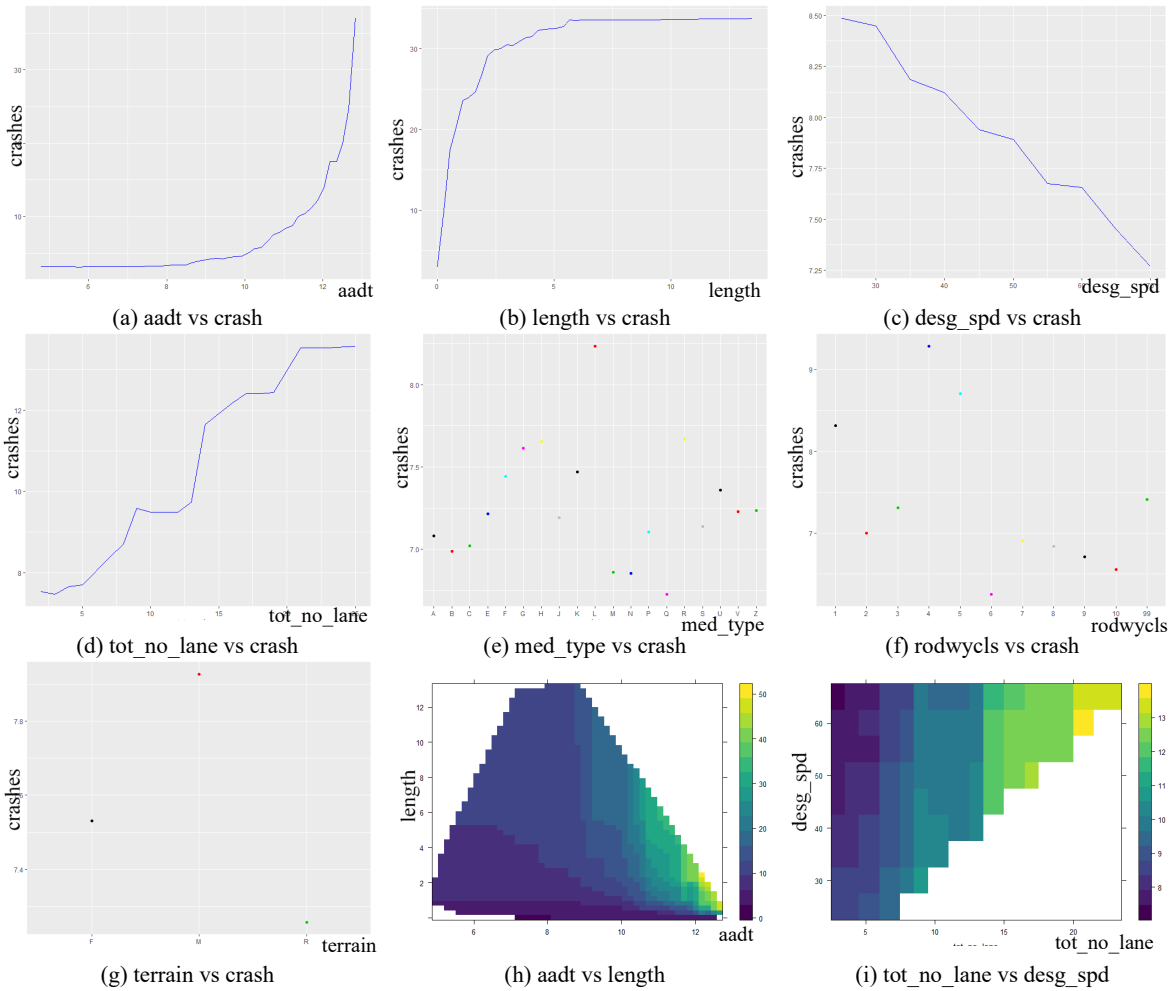


Figure 8 PDP results of the RF model

4.7 Results Interpretation

According to the variable importance considered in the RF model. A partial dependence plot (PDP) was performed on the selected variables to provide some interpretability of our best model. As shown in Figure 8.a, it can be seen that the effect of the increased AADT exponentially increased the number of crashes on highway segments. For AADT of less than 22,000 vehicles, the number of crashes did not marginally increase (1-5 crashes/year). However, for greater AADT, the number crashes exponentially increased (6-35 crashes/year). For shorter segments of less than 5 miles, the number of crashes rapidly increased as the segment length increased. On the other hand, it was found that for longer segments, i.e., above 5 miles, there was no substantial increase in the number of crashes as the segment length increased. In addition, roadway segments with high-speed design were found to induce the lower number of crashes, compared to segments with lower design speed. In other words, the number of crashes was found negatively correlated with design speed. This may be due to the underlying reason that drivers may trade-off driving carefully at higher speeds under higher risk with being less carefully at lower speeds, given that the fatal risk is perceived by drivers to be lower at the lower speed [57]. Another reason could be the roadway segments with the greater design speeds were designed for accommodating vehicles traversing the segment safely at high speeds. However, speeding could be an issue for roadway segments designed to merely accommodate lower speed choices of vehicles since there were relatively greater number of crashes predicted to occur as the design speed of the roadway segments decreased. As expected, the greater number of lanes on a roadway segment, the greater chance for increasing the number of crashes. For the median type, the findings indicated that the type-L medians contributed to the higher risk of crash occurrence among the other median types. In addition, the type-L median belongs to the retaining wall type. For the roadway type, categories 4 and 5, which are the urban multilane highway, presented the higher number of crashes predicted. Finally, the moderate grades showed the highest risk for occurrences of crashes, respectively followed by flat and rolling grades, which complied with the results found by Vanderbilt [57]

5.0 DISCUSSION AND CONCLUSIONS

This study evaluated different machine learning models to predict the average number of vehicle crash occurred during a year using data from the HSIS database. After a comprehensive evaluation, the RF model was found to predict the number of crashes more accurately, compared to the other machine learning models. The final model had the advantage of providing higher accuracy using a maximum number of seven predictors. Furthermore, the RF model had the advantage of providing variable importance which may help safety professionals to make the better decisions regarding what features of highway safety to improve. For instance, the findings showed that the type-L median had the significantly greater influence on contributing to occurrences of crashes, compared to the other variables. Furthermore, this demonstrated the potential of modeling predictive models for crash occurrence using machine learning technics.

The results suggested that the increase in AADT exponentially increased the number of crashes on highway segments. This finding was consistent with the previous studies in the literature [16, 32, 33, 38, 39]. The underlying reason could be the increase in vehicle counts led to the higher chance that the number of crashes surged. Also, as the AADT increased, there were more opportunities that the greater number of vehicles interacted with the surrounding vehicles in traffic, which could lead to the rise in the number of crashes. This study also found that road segments with the higher number of lanes tended to increase chances for crashes to occur, which was complied with the majority of the previous studies [42–44]. As the number of lanes on the roadway segments increased, there were more opportunities for vehicles to perform more times of lane-changing and overtaking maneuvers, which could be one of the underlying reasons that led to the increase in the number of crashes on roadways [33].

Interestingly, the shorter length of roadway segments appeared to induce the smaller number of crashes. This finding was consistent with the results suggested by the previous studies in literature [26, 36, 38, 40]. The underlying reason could be the corresponding speed limit of the shorter roadway segments tended to be lower than the longer roadway segments. In addition, most of the times, there were conflict points at each end of each roadway segment; therefore, acceleration rates of vehicles on shorter roadway segments tended to be bounded by this limitation, which led to the reluctantly less aggressive drivers' behavior, compared to the roadway segments of longer lengths.

This study found that roadway segments with the higher design speed induced the lower number of crashes, compared to the segments with the lower design speed. This could be because drivers may trade-off driving carefully at higher speeds under higher risk with being less carefully at lower speeds, given that the fatal risk is perceived by drivers to be lower at low speed, as suggested by Vanderbilt [57]. Another underlying reason could be the roadway segments with the greater design speeds were designed for accommodating vehicles traversing the segment safely at high speeds. In contrast, speeding could be an issue for roadway segments designed to merely accommodate lower speed choices of vehicles since there were relatively greater number of crashes predicted to occur as the design speed of the roadway segments decreased. Nevertheless, in terms of crash severity, please note that previous studies pointed out that speeding was found to be one of the main causes of death in crashes [48, 58, 59].

Finally, the limitation of this study is the testing of the model data for evaluation of the error prediction. Future studies should evaluate the models with a test dataset [60]. Other types of machine learning models, such as deep neural networks and convolutional neural networks, can be explored to identify the best possible model for predicting crashes. However, the results from such models are difficult to interpret. A similar study can be conducted for other states to see if there are any differences in terms of the risk factors contributing to crashes. In addition, future predictions may also

consider adjustments taking into account the effect of regression to the mean.

Acknowledgement

We would like to thank the Highway Safety Information System (HSIS) for facilitating the dissemination of their crash database, which was adopted as an input for conducting crash prediction models in this study (https://www.hsisinfo.org).

References

[1] United States Department of Transportation, Fatality Analysis Reporting System (FARS). 2020. Available: https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/ [Accessed: May 2021]

[2] R.M. Cunningham, M.A. Walton, and P.M. Carter, 2018. "The Major Causes of Death in Children and Adolescents in the United States," *New England Journal of Medicine*, 379(25): 2468-2475, doi: 10.1056/nejmsr1804754.

[3] V. Pasquale, G. Guido, V. Astarita, V. P. Giofrè, G. Guido, and A. Vitale, 2021. "Review of the Use of Traffic Simulation for the Evaluation of Traffic Safety Levels: Can We Use Simulation to Predict Crashes?" *Transportation Research Procedia*, 52: 244–251, doi: 10.1016/j.trpro.2021.01.028.

[4] L. Wahab, and H. Jiang, 2019. "A Comparative Study on Machine Learning Based Algorithms for Prediction of Motorcycle Crash Severity," *PLoS ONE*, 14(4): 1–17, doi: 10.1371/journal.pone.0214966.

[5] FHWA. HSIS - Highway Safety Information System. Fhwa-Hrt-11-031.

[6] V.R. Duddu, S.S. Pulugurtha, and V.M. Kukkapalli, 2020. "Variable Categories Influencing Single-Vehicle Run-off-Road Crashes and Their Severity," *Transportation Engineering*, 2, October, doi: 10.1016/j.treng.2020.100038.

[7] K. Wang, T. Bhowmik, S. Zhao, N. Eluru, and E. Jackson, 2021. "Highway Safety Assessment and Improvement through Crash Prediction by Injury Severity and Vehicle Damage Using Multivariate Poisson-Lognormal Model and Joint Negative Binomial-Generalized Ordered Probit Fractional Split Model," *Journal of Safety Research*, 76: 44-55, doi: 10.1016/j.jsr.2020.11.005.

[8] C. Dong, C. Shao, J. Li, and Z. Xiong, 2018. "An Improved Deep Learning Model for Traffic Crash Prediction," *Journal of Advanced Transportation*, 2018, doi: 10.1155/2018/3869106.

[9] S.P. Washington, M.G. Karlaftis, F. Mannering, and P. Anastasopoulos, *Statistical and Econometric Methods for Transportation Data Analysis*, 2nd Edition, CRC Press, New York, NY, USA, 2013.

[10] S. Das, X. Sun, and M. Sun, 2021. "Rule-Based Safety Prediction Models for Rural Two-Lane Run-off-Road Crashes," *International Journal of Transportation Science and Technology*, 10(3): 235-244, doi: 10.1016/j.ijtst.2020.08.001.

[11] E. Hauer, 2014. *The Art of Regression Modeling in Road Safety*, Springer, New York, USA,

[12] G. Casella, S. Fienberg, and I. Olkin, ed., 2006. *Modern Mathematical Statistics with Applications*, 2nd Edition, Springer, New York, NY, USA,

[13] T. Hastie, R. Tibshirani, and J. Friedman, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2th Edition, Springer, New York, NY, USA,

[14] P.B. Silva, M. Andrade, and S. Ferreira, 2020. "Machine Learning Applied to Road Safety Modeling: A Systematic Literature Review," *Journal of Traffic and Transportation Engineering (English Edition)*, 7(6): 775–790, doi: 10.1016/j.jtte.2020.07.004.

[15] C. Lyon, and B. Persaud, 2002 "Pedestrian Collision Prediction Models for Urban Intersections," *Transportation Research Record*, 1818(1): 102-107. doi: 10.3141/1818-16.

[16] H. Rakha, M. Arafeh, A.G. Abdel-Salam, F. Guo, and A.M. Flintsch, 2010. Linear Regression Crash Prediction Models: Issues and Proposed Solutions, VT-2008-02, Virginia Tech Transportation Institute, Blacksburg, Virginia, USA,

[17] Y. Xie, and Y. Zhang, 2008. "Crash Frequency Analysis with Generalized Additive Models," *Transportation Research Record*, 2061(1): 39-45, doi: 10.3141/2061-05.

[18] S. Sittikariya, V. Shankar, and N. Venkataraman, 2009. "*Modeling Heterogeneity: Traffic Accidents*," VDM-Verlag, Riga, Latvia,

[19] F. Guo, X. Wang, and M.A. Abdel-Aty, 2010. "Modeling Signalized Intersection Safety with Corridor-Level Spatial Correlations," *Accident Analysis and Prevention*, 42(1): 84-92, doi: 10.1016/j.aap.2009.07.005.

[20] Y. Zhang, Y. Xie, and L. Li, 2012. "Crash Frequency Analysis of Different Types of Urban Roadway Segments Using Generalized Additive Model," *Journal of Safety Research*, 43(2): 107-114, doi: 10.1016/j.jsr.2012.01.003.

[21] M. Machsus, R. Basuki, and A.F. Mawardi, 2015. "Generalized Additive Models for Estimating Motorcycle Collisions on Collector Roads," *Procedia Engineering*, 125: 411-416, doi: 10.1016/j.proeng.2015.11.105.

[22] A. Khoda Bakhshi, and M.M. Ahmed, 2021 "Real-Time Crash Prediction for a Long Low-Traffic Volume Corridor Using Corrected-Impurity Importance and Semi-Parametric Generalized Additive Model," *Journal of Transportation Safety and Security*, 1-35. doi: 10.1080/19439962.2021.1898069.

[23] M.H. Pham, A. Bhaskar, E. Chung, and A.G. Dumont, 2010, "Random Forest Models for Identifying Motorway Rear-End Crash Risks Using Disaggregate Data," Paper presented at *The 13th International IEEE Conference on Intelligent Transportation Systems, IEEE*, Funchal, Madeira Island, Portugal, 468-473, doi: 10.1109/ITSC.2010.5625003

[24] X. Jiang, M. Abdel-Aty, J. Hu, and J. Lee, 2016. "Investigating Macro-Level Hotzone Identification and Variable Importance Using Big Data: A Random Forest Models Approach," *Neurocomputing*, 181: 53-63, doi: 10.1016/j.neucom.2015.08.097.

[25] X. Li, D. Lord, Y. Zhang, and Y. Xie, 2008. "Predicting Motor Vehicle Crashes Using Support Vector Machine Models," *Accident Analysis and Prevention*, 40(4): 1611-1618, doi: 10.1016/j.aap.2008.04.010.

[26] N. Dong, H. Huang, and L. Zheng, 2015. "Support Vector Machine in Crash Prediction at the Level of Traffic Analysis Zones: Assessing the Spatial Proximity Effects," *Accident Analysis and Prevention*, 82: 192-198, doi: 10.1016/j.aap.2015.05.018.

[27] J. Sun, and J. Sun, 2016. "Real-Time Crash Prediction on Urban Expressways: Identification of Key Variables and a Hybrid Support Vector Machine Model," *IET Intelligent Transport Systems*, 10(5): 331-337, doi: 10.1049/iet-its.2014.0288.

[28] H.T. Abdelwahab, and M.A. Abdel-Aty, 2002. "Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas," *Transportation Research Record*, 1784(1): 115-125, doi: 10.3141/1784-15

[29] L. Y. Chang, 2005. "Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network," *Safety Science*, 43(8): 541-557, doi: 10.1016/j.ssci.2005.04.004.

[30] Y.C. Chiou, 2006. "An Artificial Neural Network-Based Expert System for the Appraisal of Two-Car Crash Accidents," *Accident Analysis and Prevention*, 38(4): 777-785, doi: 10.1016/j.aap.2006.02.006.

[31] C. Riviere, P. Lauret, J.F.M. Ramsamy, and Y. Page, 2006. "A Bayesian Neural Network Approach to Estimating the Energy Equivalent Speed," *Accident Analysis and Prevention*, 38(2): 248-259, doi: 10.1016/j.aap.2005.08.008.

[32] J. Kononov, B. Bailey, and B. K. Allery, 2008. "Relationships between Safety and Both Congestion and Number of Lanes on Urban Freeways," *Transportation Research Record*, 2083(1): 26-39. doi: 10.3141/2083-04.

[33] A. Abdulhafedh, "Crash Frequency Analysis," 2016, *Journal of Transportation Technologies*, 6(4): 169–180, doi: 10.4236/jtts.2016.64017

[34] J. Yuan, M. Abdel-Aty, Y. Gong, and Q. Cai, 2019. "Real-Time Crash Risk Prediction Using Long Short-Term Memory Recurrent Neural Network," *Transportation Research Record*, 2673(4): 314-326, doi: 10.1177/0361198119840611.

[35] C. Lee, B. Hellinga, and F. Saccomanno, 2003. "Real-Time Crash Prediction Model for Application to Crash Prevention in Freeway Traffic," *Transportation Research Record*, 1840(1): 67-77, doi: 10.3141/1840-08

[36] F. Chen, S. Chen, and X. Ma, 2016. "Crash Frequency Modeling Using Real-Time Environmental and Traffic Data and Unbalanced Panel

Data Models," *International Journal of Environmental Research and Public Health*, 13(6): 1-16, doi: 10.3390/ijerph13060609.

[37] Q. Cai, M. Abdel-Aty, J. Yuan, J. Lee, and Y. Wu, 2020. "Real-Time Crash Prediction on Expressways Using Deep Generative Models," *Transportation Research Part C: Emerging Technologies*, 117: (1-14) doi: 10.1016/j.trc.2020.102697.

[38] C. Caliendo, M. Guida, and A. Parisi, 2007. "A Crash-Prediction Model for Multilane Roads," *Accident Analysis and Prevention*, 39(4): 657-670, doi: 10.1016/j.aap.2006.10.012.

[39] T. Chen, C. Zhang, and L. Xu, 2016. "Factor Analysis of Fatal Road Traffic Crashes with Massive Casualties in China," *Advances in Mechanical Engineering*, 8(4): 1-11, doi: 10.1177/1687814016642712.

[40] P.C. Anastasopoulos, and F.L. Mannering, 2011. "An Empirical Assessment of Fixed and Random Parameter Logit Models Using Crash- and Non-Crash-Specific Injury Data," *Accident Analysis and Prevention*, 43(3): 1140-1147, doi: 10.1016/j.aap.2010.12.024.

[41] M.H. Islam, L. Teik Hua, H. Hamid, and A. Azarkerdar, 2019. "*Relationship of Accident Rates and Road Geometric Design*," In: IOP Conference Series: Earth and Environmental Science, IOP Publishing, Kuala Lumpur, Malaysia,

[42] M.A. Abdel-Aty, and A.E. Radwan, 2000. "Modeling Traffic Accident Occurrence and Involvement," *Accident Analysis and Prevention*, 32(5): 633-642, doi: 10.1016/S0001-4575(99)00094-9.

[43] R.B. Noland, and L. Oh, 2004. "The Effect of Infrastructure and Demographic Change on Traffic-Related Fatalities and Crashes: A Case Study of Illinois County-Level Data," *Accident Analysis and Prevention*, 36(4): 525-532, doi: 10.1016/S0001-4575(03)00058-7.

[44] A.J. Anarkooli, M. Hosseinpour, and A. Kardar, 2017. "Investigation of Factors Affecting the Injury Severity of Single-Vehicle Rollover Crashes: A Random-Effects Generalized Ordered Probit Model," *Accident Analysis and Prevention*, 106: 399-410, doi: 10.1016/j.aap.2017.07.008.

[45] D.D. Clarke, P. Ward, C. Bartle, and W. Truman, 2010. "Killer Crashes: Fatal Road Traffic Accidents in the UK," *Accident Analysis and Prevention*, 42(2):764-770, doi: 10.1016/j.aap.2009.11.008.

[46] A. Tavakoli Kashani, A. Shariat Mohaymany, and A. Ranjbari, 2012. "Analysis of Factors Associated with Traffic Injury Severity on Rural Roads in Iran," *Journal of Injury and Violence Research*, 4(1): 36-41, doi: 10.5249/jivr.v4i1.67.

[47] C. Siddiqui, M. Abdel-Aty, and K. Choi, "Macroscopic Spatial Analysis of Pedestrian and Bicycle Crashes," *Accident Analysis and Prevention*, Vol. 45, pp.382-391, 2012. doi: 10.1016/j.aap.2011.08.003.

[48] V. Ratanavaraha, and S. Suangka, 2014. "Impacts of Accident Severity Factors and Loss Values of Crashes on Expressways in Thailand," *IATSS Research*, 37(2): 130-136, doi: 10.1016/j.iatssr.2013.07.001.

[49] Y. Wang, and W. Zhang, 2017. "Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities," *Transportation research procedia*, 25: 2119-2125, doi: 10.1016/j.trpro.2017.05.407

[50] American Association of State Highway and Transportation Officials, 2010. *Highway Safety Manual*, 1st Edition, AASHTO, Washington D.C., USA,

[51] P. Royston, 2005. "Multiple Imputation of Missing Values: Update of Ice," *The Stata Journal*, 5(4): 527–536, doi: 10.1177/1536867X0900900308

[52] F. Noghrehchi, Missing Data with MICE, 2016. Available: https://web.maths.unsw.edu.au/~dwarton/missingDataLab.html [Accessed: Mar 2021]

[53] UCLA. UCLA IDRE Statistical Consulting. https://stats.idre.ucla.edu/ [Accessed: Mar 2021]

[54] MICE. Data from: Data Management in R Imputing Missing Data with R [dataset], MICE Package. [Accessed: Mar 2021]

[55] R.A. Irizarry, 2019. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*, CRC Press, Boca Raton, Florida, USA,

[56] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.C. Chang, and C.C. Lin. 2020. CRAN Package "E1071." CRAN - Package, [Accessed: Feb 2021]

[57] T. Vanderbilt, 2009. *Traffic: Why We Drive the Way We Do (and What It Says About Us)*, Penguin Group, New York, New York, USA,

[58] P. Srisurin, and S. Chalermpong, 2021 "Analyzing Human, Roadway, Vehicular and Environmental Factors Contributing to Fatal Road Traffic Crashes in Thailand," *Engineering Journal*, 25(10): 27–38,. doi: 10.4186/ej.2021.25.10.27

[59] S.A. Sarm, and K. Kanitpong, 2016. "Analysis of factors affecting the severity of motorcycle casualties in Phnom Penh using a Bayesian approach," *Asian transport studies*, 4(2): 430-443, doi: 10.11175/eastsats.4.430

[60] A. Iranitalab, and A. Khattak, 2017. "Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction," *Accident Analysis and Prevention*, 108: 27–36, doi: 10.1016/j.aap.2017.08.008.