Accident Analysis and Prevention 149 (2021) 105860

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Modeling and predicting vehicle accident occurrence in Chattanooga, Tennessee

Jeremiah Roland *, Peter D. Way, Connor Firat, Thanh-Nam Doan, Mina Sartipi

University of Tennessee at Chattanooga Department of Engineering and Computer Science, United States

ARTICLE INFO

Keywords:
Machine learning
Traffic accident prediction
Neural networks

ABSTRACT

Given the ever present threat of vehicular accident occurrence endangering the lives of most people, preventative measures need to be taken to combat vehicle accident occurrence. From dangerous weather to hazardous roadway conditions, there are a high number of factors to consider when studying accident occurrence. To combat this issue, we propose a method using a multilayer perceptron model to predict where accident hotspots are for any given day in the city of Chattanooga, TN. This model analyzes accidents and their associated weather and roadway geometrics to understand the causes of accident occurrence. The model is offered as a live service to local law enforcement and emergency response services to better allocate resources and reduce response times for accident occurrence. Multiple models were made, each having different variables present, and each yielding varying results.

1. Introduction

One of the most dangerous activities that people face each day is vehicle accident occurrence. There are three main target audiences subject to vehicle accidents: drivers, pedestrians, and local governments. Drivers face countless hazards on the roadway, from weather related hazards to unstable or dangerous roadway conditions, and even other drivers facing those same weather and roadway conditions. For pedestrians, they face the more physical consequences of vehicle accidents, such as being injured or killed through being hit by a vehicle. Local governments face a more economic based consequence of vehicle accidents, such as in 2018 when the total economic cost of accidents within the United States exceeded $12.5 million (CostGuide, 2018; HamiltonCountyCensus, 2018). These economic damages range from property damage to human costs. Estimations from the US Department of Transportation have the number of injuries resulting from vehicular accident at 1.6 million in 2018, with 4.1 million accidents resulting in property damage (USDOTNHTSA, 2018). In Tennessee alone, there were 208,605 accidents reported in 2018 that caused over $400 in damages within its counties (TITAN, 2018).

Accident occurrence throughout the United States, and Tennessee alone, has been rapidly escalating throughout recent years as shown in Fig. 1. According to the Center for Disease Control and the National Center for Health Statistics, vehicular accidents are one of the two highest fatality risks across all age groups between 1999 and 2012 (CDC, 2014; Olaisen et al., 2019).

Keeping in mind the previously mentioned statistics, vehicular accidents are clearly one of the greatest threats facing the average citizen in a day to day setting. As these accidents continue to increase in frequency, the need to reduce the impact of accidents or even outright prevention of accidents, increases as well. Beyond the obvious danger to the safety of pedestrians and drivers, there exists the issues of how to counteract vehicle accident occurrence. Given the expansive amount of features surrounding traffic accidents (weather, roadway geometrics, human error, etc.), it can be difficult determining the root causes of an accident. In addition, knowing what actions to take to mitigate accident occurrence is something law enforcement and emergency services have been struggling with for decades. By analyzing spatial and temporal data surrounding accident occurrence, we set out to provide the Chattanooga Police Department with a predictive model to determine where in Chattanooga accidents are more likely to happen on a given day within a certain time frame.

1.1. Contribution

This work addresses mitigation of accident occurrence through analysis, prediction, and prevention of vehicular accidents within Chattanooga, Tennessee. Presented here is a Multilayer Perceptron

* Corresponding author.
E-mail address: fpf852@mocs.utc.edu (J. Roland).

https://doi.org/10.1016/j.aap.2020.105860
Received 15 June 2020; Received in revised form 15 September 2020; Accepted 29 October 2020
Available online 7 November 2020

(MLP) neural network model created to predict where and when accidents will occur on a provided day and time within the area of study. The model utilizes historical vehicular accident records, weather conditions, roadway geometrics, and other aggregated variables in the creation of predictions for future accident occurrence. This model provides benefits for two groups, police officers/emergency responders, and local government officials. For emergency responders, the model will be provided as a live service application to display the most dangerous areas in Chattanooga for better resource allocation, whether that may be patrol route alterations or police car placement. For local government officials, our project can provide prescriptive analysis on dangerous areas and bring to light certain features of these areas that lead to higher accident counts. This allows local governments to make alterations to these locations, such as adding in yield signs, speed bumps, traffic lights, etc. Chattanooga itself acts as a perfect test bed for accident research, as the area has a wide range of weather conditions over a comparably small area. This allows a tighter focus of study over the area without needing to traverse over an extensive amount of data that would be required for projects that cover a wider area.

2. Related works

A case study was conducted concerning the prediction of traffic accidents through utilization and comparison of results between four different classification models (Yuan et al., 2017). These methods included: linear Support Vector Machine (SVM) (Suthaharan, 2016), Decision Tree (DT) (Kotsiantis, 2013), Random Forest (RF) (Liaw and Wiener, 2002), and Deep Neural Networks (DNN) (Sze et al., 2017). Within this study a method of generating non-accident data was performed and called negative sampling. For each positive example (accident), the value of only one feature was changed among hour, day, and the unique road identification. Afterwards, the resulting samples were checked for positive (match found) or negative (no match found) results amongst the existing dataset. In the end, results dictated that the optimal model was DNN for the application in discussion.

Convolutional Long-Short Term Memory (ConvLSTM) (Sainath et al., 2015) is a subsidiary of Long Short Term Memory (Cheng et al., 2016) involving the use of convolutional operations inside of the LSTM cell. The convolutional operations allow for multi-dimensional data such as radar or satellite imagery. This ConvLSTM setup was applied to a study concerning vehicular accidents in Iowa, between 2006 and 2013. Data included crash reports from Iowa DOT, rainfall data, Roadway Weather Information System (RWIS) reports, and further data provided by Iowa DOT such as speed limits, AADT, and traffic camera counts (Yuan et al., 2018). This study mentions that no other previous work had fused such large heterogeneous datasets together before, nor had any included spatial structures of the roadway network. As such, a five kilometer square per block grid layout was constructed to cover the state for prediction forecasting. Training data included 2006 to 2012 reports, with 2013 being reserved for testing. Tests involved predicting locations for the next seven days based on data provided by the previous seven days. ConvLSTM results out performed all baselines in prediction accuracy. As well, the system correctly predicted accidents resulting from the case study of December eighth in 2013, where a significant snowstorm caused numerous accidents.

Accidents within the city of Montreal were studied using three open datasets and the Balanced Random Forest algorithm (Hébert et al., 2019). Accident data was retrieved from Montreal Vehicle Collisions, weather information was provided by the Historical Climate Dataset, and roadway segment information was retrieved from the National Road Network database, provided by the Canadian government. Four different models were tested, including BRF (Balanced Random Forest), RF (Random Forest), XGB (XG Boost), and a baseline model. Negative samples (that is, examples of non-accident occurrence) were created as well. A total of two billion negative samples were possible, with the team electing to only utilize 0.1% of such. Predictions were for roadway segments by hour, considered a highly specific definition. All together, the systems were able to predict 85% of Montreal accidents, with a False Positive Rate (FPR) of only 13%. It is notable that the datasets in use were open source, implying that the study could easily be shifted to another locale relatively easily, since no data restrictions were in place.

An initiative led by the United States Department of Transportation (USDOT) means to partner crowd-sourced data provided by Waze and safety policy decisions to help predict vehicular accidents (Dan et al., 2018). Prediction is completed through the use of Classification and Regression Trees (CART) and Random Forest models. The pilot for the study included six months of accident data from Maryland, paired with the corresponding (if any) Waze alerts. Specific temporal and spatial event patterns created by the pilot model are quite similar to the actual accident records, albeit not identical. Additionally, the model tends to under predict accidents in early morning hours, while over predicting accidents for high-commute periods. This is attributed to the historical spread of accidents. Of particular note is the model's ability to predict Waze alerts for minor accidents, that is, those not serious enough to report to law enforcement but significant enough to inhibit standard traffic flow. The study continues with the partnership of state and local partners for implementation of multiple case studies of the Waze crash estimation model.

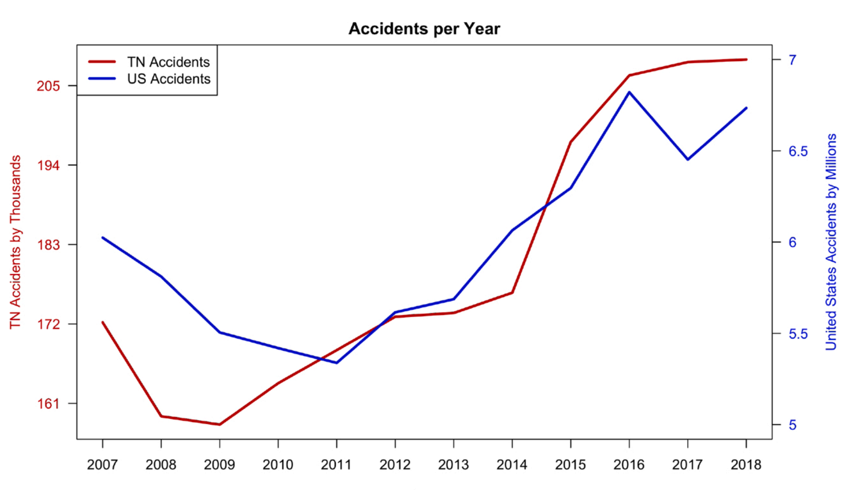

Fig. 1. Total vehicular accident counts for United States and Tennessee from 2007 to 2018 (TITAN, 2018; USDOTNHTSA, 2018). Note the different Y-axis values per data.

2.1. Existing work comparisons

There are a wide variety of different attempts at accident prediction, with many of them having different methodologies. The closest related work in terms of both execution and results to our own project would be (Hébert et al., 2019), where accidents within the city of Montreal were studied by a team from Concordia University via the use of a Balanced Random Forest algorithm. Both this study and our own focus upon a singular city, utilizing roadway geometrics and weather conditions in addition to the temporal and spatial specifics of accidents. However, the team of Montreal elected to test using a roughly 94% division of negative samples, in comparison to our three splits of data. This study aggregated accident occurrence by roadway intersections, not by a grid layout. This allowed for a more finely tuned variable list of roadway geometrics. Additionally, predictions were assessed on an hourly unit, not by particular times of day. The study of Montreal concluded with an 85% Recall score, with 13% FPR. The reason for this team's increased performance is due to two outlying differentiating factors between their project and our own.

The strongest differentiating factors of Hébert et al. (2019) are the amount of data used and the specificity of the data itself. From their study, four years-worth of data was utilized for training, testing was performed on the remaining two years of data. Our own project only had access to 3 years of data all together. Furthermore, the work utilized more detailed information, such as specific roadway segments and hours, as opposed to aggregated spatial and temporal breakdowns. Previously, attempts were made for the study present within this work using similarly high detail information for accident analysis, that being specific roadway segments and hour values for accidents. However, results were negligible at best. We believe this is due to lower count of accidents to analyze, and that employing highly specific data requires a larger pool of data to learn from. Due to our lower count of accidents, it was necessary to approach the situation with a more generalized viewpoint, thus necessitating the aggregation of spatial and temporal variables.

The study completed by Yuan et al. (2017) is the origin of our own study's negative sampling technique. This study also utilized 75% negative samples within their data, informing our usage of the division. However, we expanded upon this idea in the testing of an even rarer split (90/10) as well as testing predictions on an even split of data. The highest Recall within the four different classification techniques from Yuan et al. (2017). was listed as 86.89%, quite a bit higher than our own. Precision on the study also outperformed our own, with a very low FPR score. Accuracy in total for this study completed at 95.11%. However, like the work completed in Montreal, this study had the advantage of working on a finer roadway level of precision. This study also had the ability of integrating human factors into their models, something of which we currently do not have the ability to do. Perhaps the addition of said demographics could assist in boosting performance of our own models.

The study from Yuan et al. (2018) utilized Cross Entropy, Mean Squared Error and Root Mean Squared Error as its measure of performance. As such, this study and the one presented within this work cannot be directly compared, but can still be discussed together. Both studies did employ neural networks in their research, and did aggregate the locations of accidents. However, this study did not group accidents by a grid system, but by locational proximity. Study was also divided between rural and urban areas, which may provide interesting further research for our own work. Additionally, this study provides visual comparison results of their predictions, without providing any numerical accompaniment.

Similar to the study presented within this work, Dan et al. (2018) aggregated their spatial location into a set of grid blocks, including 0.5, 1, and 4 square mile grid blocks. Due to the study analyzing the accidents across all of Maryland, the increased spatial aggregation areas are to be expected. However, this team had a higher degree of detail for roadway related information, including Annual Average Daily Traffic. Human based data was also utilized, such as job class, sex, and other economic data. Substantially more data was available to study, nearly 2 million accident entries. This overwhelmingly higher amount of data to learn from, alongside the human economic data and higher detailed roadway data, are likely the contributing factors to the presented increase in model performance and results in contrary to the work presented here.

3. Our project and data

3.1. Project synopsis

The goal of the project presented herein is the creation of a live service application for the use of local law enforcement, tasked with displaying the likelihood of accident "hotspots" throughout the city of Chattanooga for a given day and time of day. Through the use of this app, local law enforcement can more efficiently allocate resources throughout the city to deter accident occurrence. This allocation can be in the form of altering where police station themselves for paperwork completion, altering patrol routes, or placement of temporary speed deterrents. This type of allocation allows officers to perform the simplest form of accident prevention, simply being present in high risk areas. Many drivers will slow down and drive more appropriate whenever a police car monitoring the roadway for hazardous drivers is present.

An additional application of this project is prescriptive analysis, where the local government officials would be informed of dangerous parts of the city and given analytical data displaying what makes a certain area of the city an accident hotspot. Through this, the city can take preventative measures to increase roadway safety, such as adding in speed bumps, yield signs, or even altering the structure of the road (e.g., add in or take out a lane of traffic).

3.2. The data

Data utilized consists of all reported accidents in Chattanooga, provided by the Hamilton County Emergency Services District. The data covers all reported accidents from late 2016 up to the present day, where daily records are received, ready for appending to the main dataset. These records included the time of the call, the time of response, the date, the GPS location, and injury level (no injury, injury, entrapment, mass casualty). In this study, records are included based upon accident reports from 2017, 2018, and 2019. Any duplicate calls in the dataset, where a single vehicle accident is called in by multiple people, were dropped.

Additional spatial and temporal information was added to the accidents, including weather and roadway geometrics. Weather information was provided by a Python API library called DarkSky, which compiles weather from multiple different weather stations and provides the best suited weather report for the given time and location information. Roadway geometric information was provided through a combination of ETRIMS and ArcGIS. ETRIMS is database of Tennessee roadway information and ArcGIS is a mapping program used for manipulating spatial information into desired formats.

Through DarkSky, weather variables were added to accident records based upon the temporal and spatial details of the occurrence. Through ETRIMS, spatial variables were added including road count, land use, pavement type, and other roadway specific information. Lastly through ArcGIS, the testing area was divided into a 'fishnet' made of 694 hexagons, each covering a 0.2 square mile area. This aggregation was done to lower the specificity required by the model, as previous attempts of model creation with highly detailed roadway data yielded poor results. Due to this aggregation process, specific roadway information was required to also be aggregated based on the fishnet, simplifying the prediction process by reducing the number of areas to predict for. Table 1 summarizes the variables used in testing. Fig. 2 presents a visual

Table 1
Variables Used in Study

	Explanation
911 variables	
Accident	No accident (0) or accident (1)
Hour	The hour of the day accident occurred
WeekDay	If accident was on weekend or weekday (binary)
DayOfWeek	Day of the week (0–6, Monday–Sunday)
Unix	Timestamp of the accident in seconds
DayFrame	Aggregated hour times of the day (see Table 2)
Weather variables	
Rain/cloudy/foggy/snow/clear	Precipitation conditions (binary)
Rain before	Rain in previous hour (binary)
Temperature	Temperature at time of record
Dewpoint	Air temp required for water vapor saturation
Humidity	Amount of water vapor in the air (0 to 1)
Cloud Coverage	Percentage of the sky covered by clouds (0 to 1)
Precipitation intensity	Intensity of precipitation at time of record
Wind speed	Speed of the Wind (mph)
Road Variables	
Grid number	Position in aggregated spatial hex layout
Type of terrain	The type of land terrain (rolling, flat, etc.)
Number of lanes	Number of lanes
Function class	Function Class (municipal highway agency, etc.)
Join count	Accident count for that grid number

Note: All variables (excluding join count) listed under road variables are average values based on Grid Num.

representation of the fishnet layout across the study area.

3.3. Negative sampling

Once the appropriate spatial and temporal information was appended to accident data, negative samples were generated to allow a more thorough understanding behind accident occurrence. While there exists a generous amount of accident data within the dataset, it proved difficult to extract meaningful results and predictions from solely using accidents. For a model to accurately predict an accident versus a non-accident, it must learn non-accidents, which led to the use of negative sampling. Tests performed by Yuan et al. (2017) introduced the concept of negative example creation from positive records, or, creating non-accidents from accidents. Their process involved shifting the value of a singular variable within an accident record from the available options of hour, date, or location. After the change, the newly altered record is compared to the dataset for a matching accident event. If no match exists, it is considered a non-accident. For hour changes, a new hour is chosen between 0 and 23, excluding the hour of the current accident record. For Date, a new date is chosen for the same year as the accident record. For location, a new roadway segment was chosen for the accident record. This process resulted in roughly three times as many negatives than positives in the dataset.

This process of negative data generation informed the usage of negative sampling employed here. A more "random" method of negative data generation was considered, where the date, hour, and grid number of the accident record were changed simultaneously. This process was repeated 9 times per accident record in the database, resulting in roughly a 90%/10% split of data. When conducting model runs, different versions of the model were created with differing negative/positive data ratios to examine how data should be split between positive and negative samples. This examination of data splitting originated

Table 2
DayFrame breakdown.

DayFrame	Hours covered
DayFrame 1	0–4 and 19–23 (overnight)
DayFrame 2	5–9 (morning rush)
DayFrame 3	10–13 (lunch hours)
DayFrame 4	14–18 (evening rush)

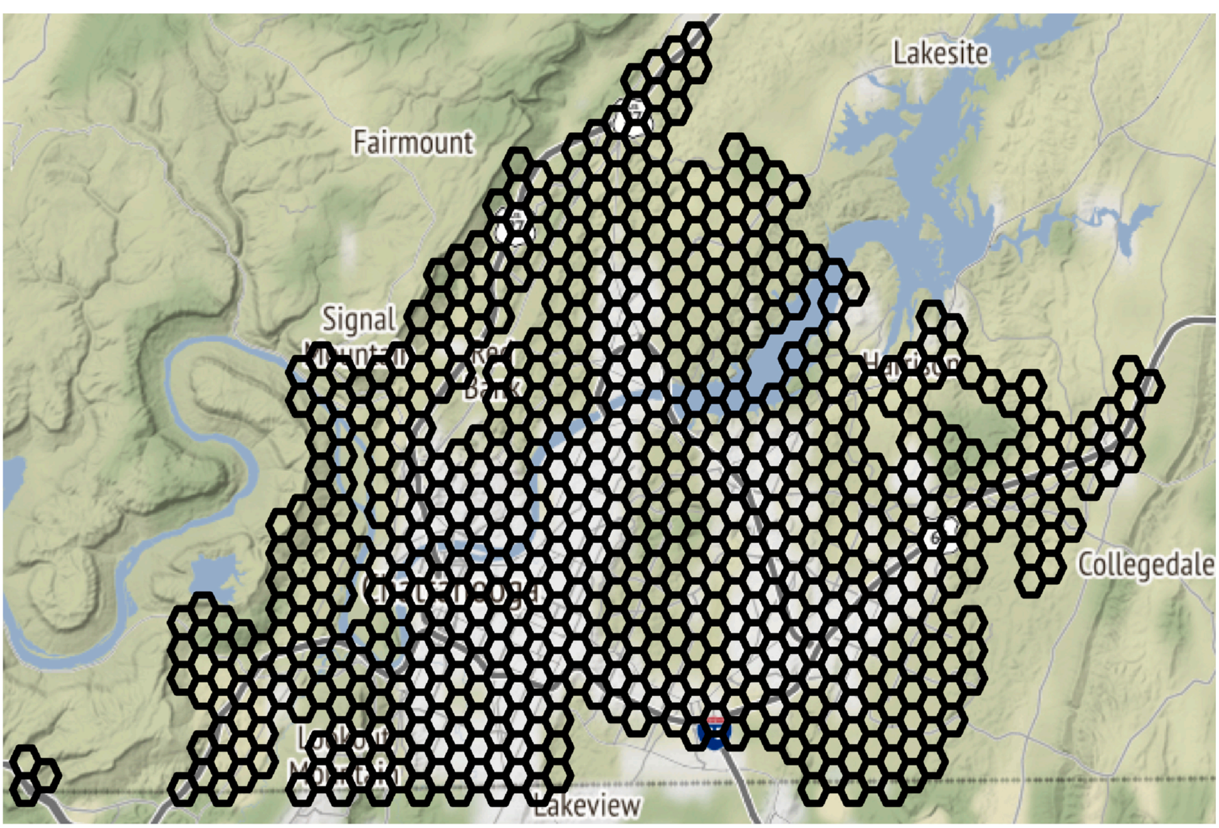

Fig. 2. Hex layout view of Chattanooga, TN.

within an article by Ranjan (2019), which covered the importance of having a higher amount of negative samples for an event class when the positive samples of that event are naturally rare. Given the inherently rare nature of accident occurrence, it was decided to follow this methodology to ensure accidents maintained their rare status.

4. Geo-spatial model design

Visually presenting historical accident data related to time presented some challenges. The Geographical Information System (GIS) has many solutions for time enabled data. One solution was the creation of a 3D time enabled Geo-spatial Model. This model essentially is an aggregation of accidents counts extruded into 3D, within predefined hexagon grids based on a 24-h period, animated by frame by frame in Fig. 3. To create this Geo-spatial model the data was categorized based on hour of the day from a value ranging from 0 - 23. In order to get a count of accidents a frequency count of how many times the Hour field and Grid ID field in a grid must be assessed. To normalize the data the count of accidents in our grids is divided by the amount of years that have passed since the day of the first recorded accident in the dataset to the most recent accident. This presents a reference of what grids have hotspot spikes depending on what time of day (Table 3).

5. Methods and analysis

5.1. Multilayer perceptron review

A multilayer perceptron (MLP) is a type of artificial neural network that is composed of one or more input layers, one output layer, and a number of hidden layers in between the input and output. Neural networks themselves are modelled after the human brain in an attempt to mimic the natural learning process. They are constructed of individual nodes which are then clustered into groups called layers. The hidden layers (layers between input and output) are where all computations happen, and can span as long as the designer decides. What separates an MLP network from other artificial neural networks is its multiple layers of hidden nodes. Multilayer perceptron networks are useful for predicting both continuous and classification problems, however they are more often used for the latter. Fig. 4 shows the basic layout of a multilayered neural network with one hidden layer, where:

Table 3
Variables used in geo-spatial model.

Geo-spatial model variables	Explanation
Hour	The hour of the day
Grid ID	Position in aggregated spatial hex layout
Accidents per Year	Calculated ratio of accidents per time

- $a_i^{(h)}$ is the ith node in layer h
- $w_{i,j}^{(h)}$ represents the weight of node i going into node j in layer h
- g is the activation function at each layer
- b^h is the bias for layer h
- $\hat{y}$ is the output of the prediction

The mathematical representation of Fig. 4, when $h = 1$, is shown in Eq. (1):

$$\hat{y}_k = g(b^{(h)} + \sum_j g(b^{(\text{in})} + \sum_i a_i^{(\text{in})} w_{ij}^{(h)}) w_{jk}^{(\text{out})}) \quad (1)$$

Multilayer perceptron networks are applied for supervised learning problems and operate using the back-propagation algorithm. This algorithm has two steps – the feed forward pass and the backward pass. The forward pass moves from the input through the hidden layers into the output, and the prediction is measured. The backward pass uses partial derivatives of the error function and back-propagates them through the network. This gives a gradient of error that can be adjusted in order to find the minimum error rate. The back-propagation algorithm continues trying to optimize the error rate of the predictions, until it can no longer be optimized. This is known as convergence, where the network can no longer improve in its current configuration. The training of multilayer perceptron networks involves using a set of input and output pairs and learning to model the correlation between them. This

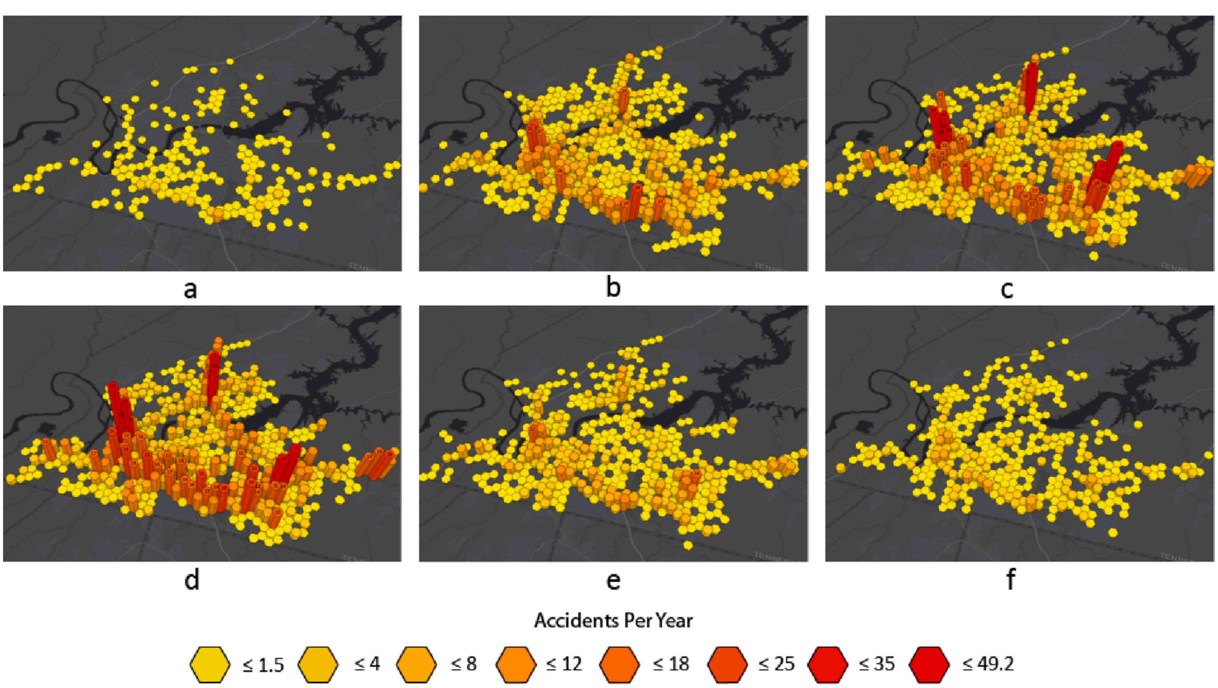

Fig. 3. Geo-spatial model of Chattanooga, TN for all accidents between 2017 and 2019. Displays 6 frames of the 24 h period. (a) 4 am. (b) 8 am. (c) 12 pm. (d) 5 pm. (e) 8 pm. (f) 12 am. The color and height of the pillars in the images reflect the number of accidents present at the grid coordinate, where the darker the color and higher the tower, the higher number of accidents. For the proper interactive model, visit this link: https://connorfirat.github.io/Hexagon_24hr_Grids/index.html.

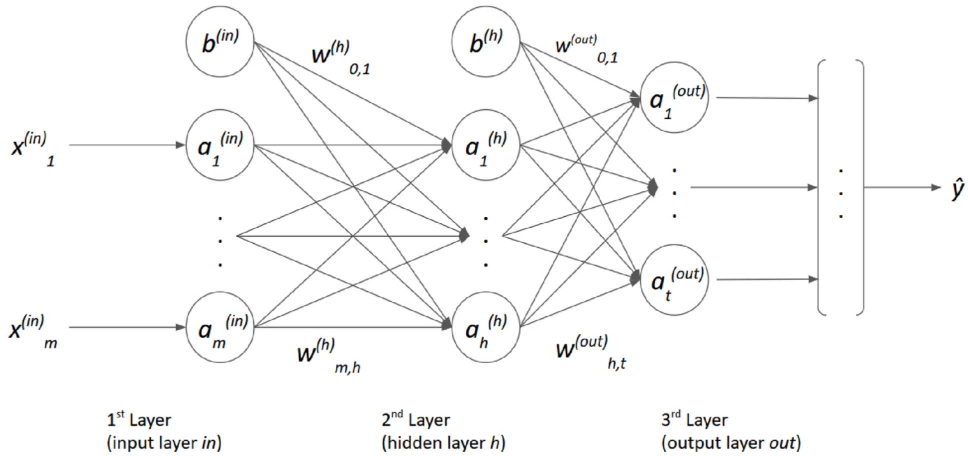

Fig. 4. A basic multilayer perceptron network with one hidden layer.

involves adjusting weights and biases of the model in order to minimize error. To illustrate the process, it is beneficial to think of an MLP network like a game of tennis. There is a constant back and forth with the network working in both ways, estimating prediction calculations and then receiving feedback.

5.2. Model selection and architecture

Several different types of testing were conducted to find the model that would best fit the data and project. These tests consisted of conventional regression and machine learning methods. Their results were overall lackluster. Before the implementation of an MLP network, select K Best testing (Bisong, 2019) was applied for possible dimension reduction. When compared to the standard results of the MLP, the results of the various Select K Best tests under performed in both accuracy and area under the curve, with K ranging from 5 to 25. Additional tests consisted of Naive Bayes and a standard accuracy score test provided by Sklearn, both of which provided worse performance scores than a standard MLP model. Due to the lackluster testing results mentioned above, a standard MLP model (Ramchoun et al., 2016) was chosen for our study's machine learning technique. Labelled inputs are used for classification prediction, which MLPs are suitable for. Additionally, MLPs are flexible with the use of data, which is beneficial to our study as our dataset is very complex and intricate. MLP networks consist of one input layer, one output layer, and one or more layers between the two. The details of the architecture used in this project are displayed in Table 4.

Our project's model is a standard Keras Sequential MLP offered through a Python module through Pycharm. Several different versions of the MLP model were tested using our dataset, with a variety of different parameter combinations. Initially, compilation was provided by binary cross-entropy, which is particularly useful for binary results and classification. However, mean squared error was chosen to provide compilation as it provided a significantly lower loss score with only a 2% cost in accuracy. Tested activation functions identity, tanh, and relu all under performed when compared to sigmoid, due to it being particularly useful for probability predictions due to it limiting a prediction model's output to a range of 0–1. Additionally, Nadam (2015) provided superior performance when compared to alternative optimizers, such as sgd and adam. Furthermore, a combination of a different number of hidden layers and dropout layers of varying sizes were tested, with our current layout yielding the best performance. The best results were achieved by using one dropout layer set to 0.1 and 2 dense hidden layers whose node counts were X-5 and X-10, with X being the number of variables used by the data. A summarization of the model's layout and parameters can be seen in Table 4.

Table 4
MLP neural network architecture.

Layer	Location	Type	Node	Activation
1	Input	Dense	X	Sigmoid
2	Hidden	Dense	X-5	Sigmoid
3	Hidden	Dropout	–	–
4	Hidden	Dense	X-10	Sigmoid
5	Output	Dense	1	–

Note: X in the Node column refers to the number of variables in the data used to create the model.

In the Node column, a formula is used to determine the number of nodes used per layer. A simple subtraction equation was put in place to set the number of nodes per layer based on the number of variables supplied to the model. Note that this method requires there to be no less than 10 variables present for the model's uses. This was implemented due to the Test Types described in Section 5.4, as each test has a different number of variables used.

There have been several attempts to use machine learning models to predict accidents, each having different setups and, consequently, varying results (Theofilatos and Yannis, 2014). Poisson distribution was also utilized in accident prediction by Abdel-Aty and Radwan (2000), showing better performance over traditional linear regression based models in terms of modelling vehicle accident frequency. Additionally, Khattak et al. (2017) used Negative Binomial models and ordered logit/probit models to explore crash severity. There have also been studies conducted to apply binary logistic modeling for studying injury severity (Weng and Meng, 2011; See, 2008; Li and Bai, 2008).

5.3. Feature selection

For the purposes of creating a simpler and more comprehensive model, Feature Selection using an ExtraTreesClassifier algorithm was employed to reduce the input dimensions down to the top 15 most important variables, presented in Table 6. ExtraTreesClassifier (Pedregosa et al., 2011) is a decision tree based ensemble method that randomizes decisions and data subsets to minimize over-fitting and over-learning. Of the top fifteen variables, three are considered as temporal variables (Hour, DayFrame, and Unix), four are spatial variables (Join_Count, Latitude, Longitude, and Grid_Num), and the remainder of the variables are weather related.

5.4. Variable combination tests

To gain a better understanding of individual variable importance, as well as variable category importance (traffic variables, weather variables, roadway variables, etc.), different combinations of variables were used during model creation. These variable combinations include:

- Test A (TA): all available variables are used
- Test B (TB): all redundant variables are removed
- Test C (TC): dropped all weather variables
- Test D (TD): dropped most location variables (kept Grid_Num)

Regarding Test B, the types of variables dropped were any variables covered by an alternate aggregated version. For example, DayFrame acts as an aggregated Hour so the hour variable was not used in model creation.

6. Results

As a preface to this section, in Eqs. (2)–(4), TP is True Positive, FN is False Negative, FP is False Positive, and TN is True Negative. When considering the results of a rare event predictor, rating performance solely upon the Accuracy metric is not a suitable manner to evaluate performance. Accuracy considers both the number of correct negative and correct positive events predicted. This skews the actual performance rating of the predictor where positive events are uncommon since negative predictions should vastly outweigh positive predictions. More fitting performance metrics for rare event predictors are the Recall, Precision, and F1 Score values. Recall, shown in Eq. (2) (Shung, 2018), refers to the percentage of correctly predicted accidents amongst all actual accidents. Precision, shown in Eq. (3) (Shung, 2018), is the ratio of correctly predicted accidents to all of the predicted accidents. F1 score, shown in Equation (4) (Shung, 2018) is the weighted average of recall and precision, and the higher the value the better.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{3}$$

$$F1\text{Score} = 2 * \frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \tag{4}$$

6.1. Model prediction results

Contradictory to a previous statement from Ranjan (2019) discussing the importance of class balance, the overall best performing model split was the even 50–50 split. The other model splits of 90–10 and 75–25 both had inferior performance when compared to the 50–50 models. Furthermore, it was found that models with a 90–10 split were more likely to have a higher prediction count for non-accidents and a lower prediction count for accidents. Table 5 shows the model's predictions, otherwise known as the testing accuracy. Note that 50–50 refers to the ratio of negative to positive data for the model, and FS refers to the implementation of Feature Selection in model creation.

For a more concrete understanding of how a model is performing, the recall, precision, and F1 score values are used to evaluate a model's performance. However, the results in Table 5 are not indicative of how a model would perform in the real world when attempting to predict accidents. Therefore, the next section will review the results of how these best performing models were able to perform when creating actual predictions.

Table 5
Model training and testing results.

Negative sample type	Train Acc	Test Acc	AUC	Recall	Precision	F1 Score
Total Shift 50–50 TA	81.58	81.91	81.88	83.16	81.75	82.45
Total Shift 50–50 FS TA	80.54	80.78	80.76	81.95	80.61	81.28
Spatial Shift 50–50 TA	80.16	79.64	79.54	82.82	78.77	80.74
Spatial Shift 50–50 FS TA	78.92	79.10	78.98	83.59	77.46	80.41
Total Shift 50–50 TB	79.78	79.64	79.62	80.42	79.78	80.10

Note: FS refers to the implementation of Feature Selection, and 50–50 means the data consisted of roughly a 50–50 split between positive and negative data samples.

6.2. Prediction implementation results

Extensive prediction testing was performed for the entire month of January 2020 using the years of 2017–2019 as training data. Contrary to initial impressions, the best performing model from Table 5 did not have the best predictive capability. Indeed, the best performing model for real world predictions is the Total Shift 50–50 FS TA, which is the same model as the best performing model from Table 5 but with feature selection applied. An extended prediction period covering the remainder of January 2020 can be seen in Fig. 6, where the TS 5050 FS TA model remains the best performing model.

7. Discussion

7.1. Interpreting results

Feature importance analysis was performed before model creation to restrict the dimensions of the data to the top 15 most important variables. Table 6 shows the variables ranked as most important for the model Total Shift 50-50 TA, which in turn act as the input variables for the model Total Shift 50–50 FS TA. Of interest is the somewhat repetitive nature of highly ranking variables. Hour ranks higher than DayFrame while at the same time representing a finer time division, yet both are present in the top five variables. Latitude and Longitude are both present as well, despite relative location being represented by Grid_Num. Finally, the inclusion of Join_Count at the top of the variable list presents the importance of previously occurring accidents when considering future accidents. For the real world prediction results shown in Figs. 5 and 6, it is seen that by implementing feature selection, the F1 Score consistently sees improvements across the prediction time frame. PCA testing was also performed on the variables used for TS 50–50 TA, which showed that the more significant variables used in the model were variables pertaining to location and time, similar to the ExtraTreesClassifier results.

Regarding Test A yielding the most viable results, the inclusion of potentially redundant variables (e.g., Lat/Long and Grid_Num or Hour and DayFrame) is not detrimental to the model's usability, as those

Table 6
Variable importance testing results.

Rank	TS 50–50 FS	Score
1	Join_Count	0.2258
2	Hour	0.0800
3	DayFrame	0.0732
4	Latitude	0.0632
5	Longitude	0.0590
6	Grid_Num	0.0480
7	Unix	0.0447
8	humidity	0.0392
9	windSpeed	0.0375
10	uvIndex	0.0371
11	temperature	0.0364
12	dewPoint	0.0358
13	pressure	0.0347
14	visibility	0.0324
15	cloudCover	0.0300

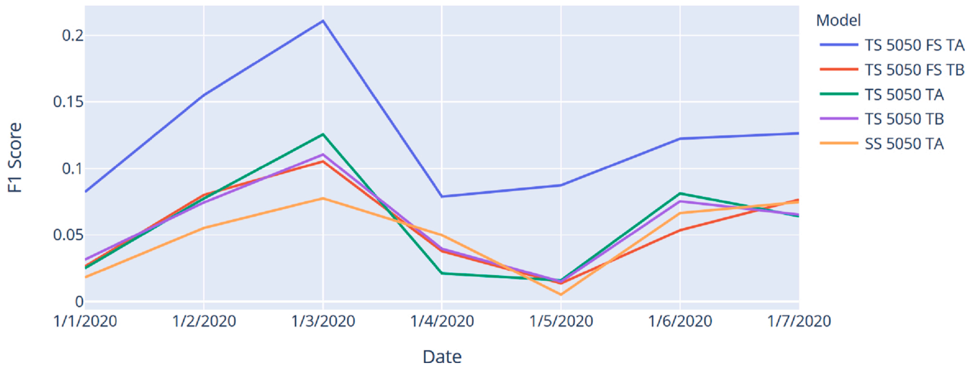

Fig. 5. Performance of model prediction for the first week in January 2020 using models from Table 5.

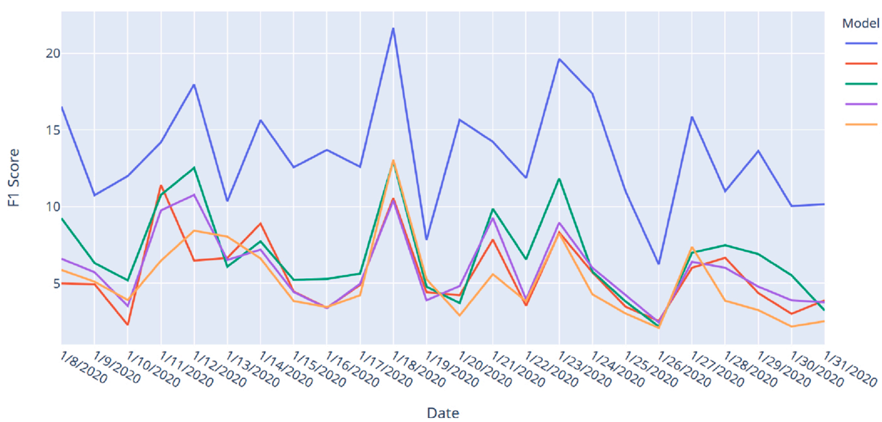

Fig. 6. Performance of model prediction across January 2020 using TS 5050 FS TA model. Best presented in color. FS is Feature Selection, where the colors reflect if feature selection was applied.

potentially redundant variables use different scales. For example, latitude and longitude use the standard GPS coordinate scale, while Grid_Num is an integer from 1 to 694. In other words, while some of the variables may reflect similar information, they are presented in inherently different ways.

7.2. Variable type significance testing

Regarding the implementation of ExtraTreesClassifier as a variable significance identification tool, as mentioned in Section 5.3, the most important variables returned for the Test A model were mostly location and time based. To investigate this, additional tests were performed using only the top 7 most significant variables from Table 6, from here referred to as Weather-Exclusion. Overall, the results for the Weather-Exclusion model were worse than the original model using Test A, as shown in Table 7.

While the TP and FN values for the Weather-Exclusion model are more preferable, these differences are statistically outweighed by the larger differences in TN and FP between the two models. With the TS 50–50 FS TA model having 240 more True Negatives and 241 fewer False Positives, at the cost of 9 fewer True Positives and 10 more False Negatives, the overall performance of the TS 50–50 FS TA model is preferable to the Weather-Exclusion model. The key take away from this is that by including weather information, the model is less inclined to have a higher prediction count for accidents, leading to a more balanced model in terms of prediction accuracy.

Table 7
Model prediction averages for January 1 to January 7, 2020.

Model name	TP	FN	TN	FP	Recall	Precision	F1 Score
TS 50–50 FS TA	49	35	3753	668	60.18	7.04	12.33
Weather-Exclusion	58	25	3513	909	70.92	6.03	10.99

7.3. Significance testing

As a final insight into the performance of the different models, *t*-tests (Kim, 2015) were performed on the prediction outputs for each model. In this experiment, the hypotheses deal with the similarity of the two models' performance, with the null hypothesis stating that the two models being compared are similar, while the alternative hypothesis stating that the performances of the two models being compared are significantly different. Each model was directly tested against its reduced version; as Weather-Exclusion is to TS 5050 FS TA, TS 5050 TC is to TS 5050 TA, where the former is a version of the latter without weather. As seen in Table 8, the results for each prediction day are different on the whole, with the majority of the *p*-values being less than 0.05. From the result, we can conclude that the performances of the two compared models are significantly different.

8. Conclusion

Vehicular accidents are a common threat for most civilians. With ever present hazards increasing the likelihood of an accident happening, accident mitigation must attempt to prevent these hazards. Our contribution is the creation of a predictive Multilayer Perceptron Model to inform local law enforcement officers of high likelihood accident hotspots for any given day. This issue was approached by analyzing different spatial attributes, such as roadway aggregation and historical accident counts, and temporal attributes, such as weather, associated with accident occurrence. This leads to several different predictive

Table 8
t-Test results for model prediction outputs.

TS 5050 TA and TS 5050 TC	T Stat	DoF	Crit Val	P Val
1-1-2020	1.30	7666	1.65	0.19
1-2-2020	4.23	8944	1.65	0.00
1-3-2020	3.79	8944	1.65	0.00
1-4-2020	14.93	8944	1.65	0.00
1-5-2020	−4.91	8944	1.65	0.00
1-6-2020	−1.59	8944	1.65	0.11
1-7-2020	2.71	8944	1.65	0.01
TS 5050 FS TA and Weather-Exclusion	**T Stat**	**DoF**	**Crit Val**	**P Val**
1-1-2020	−5.12	7882	1.65	0.00
1-2-2020	−13.53	9196	1.65	0.00
1-3-2020	−8.79	9196	1.65	0.00
1-4-2020	−1.68	9196	1.65	0.09
1-5-2020	−8.88	9196	1.65	0.00
1-6-2020	−7.12	9196	1.65	0.00
1-7-2020	−3.30	9196	1.65	0.00

models being created, each yielding varying results. However, the application of feature selection allowed for a marked increase of Recall, Precision, and F1 Score values. The best performing accident prediction model resulted from changing the hour, date, and location values of an accident entry when creating negative samples, having an even split of negative to positive data, providing all available variables for analysis, and applying feature selection, referred to as Total Shift 50–50 FS TA. The variety of results presented across the dates of study we believe stem from the inherent chaotically random nature of accident occurrence. In addition to the prediction capabilities, our project can provide local government officials prescriptive analysis on dangerous areas and bring to light certain features of these areas that lead to higher accident counts. With this information, local governments can take the necessary measures to reduce the hazardous nature of these locations, such as adding in yield signs, speed bumps, traffic lights, etc.

The greatest limitation of this project is the lack of available data. Specifically, the manner in which accidents are reported for our area of study provides no driver specific information or vehicle specific information. Additionally, traffic volume and velocity data are not viable data to be gathered from ETRIMS. While the data does exist in the ETRIMS database, it is very sparse and inconsistent for the majority of the roadways in Chattanooga, leading to most of the roadway entries in the database containing missing volume and velocity data. However, this severe limitation provides our project with a unique feature. The creation of a model that uses widely and easily available data increases the use case potential of the project to different counties and cities which may not have highly specific roadway or driver specific data. With the creation of a reliable traffic accident prediction model that uses easily obtainable data, a wider group of individuals can benefit from its implementation.

As the project proceeds, adjustments to the model and its input features will continue to provide the optimal output. One such future branch of this study includes further investigation into the creation of a singular adverse weather variable, as presented by Hébert et al. (2019). The current individual weather binaries presented within this study could be complicating the models unnecessarily, although this is subject to be determined by further testing. Additionally, demographic data as presented in Dan et al. (2018), Yuan et al. (2017) could be accessed from Geographical Information System (GIS) and incorporated into future modelling, providing some of the missing human factors sought by this team. Lastly, regarding the limitation of a lack of data, any future implementations of this project in different cities/counties could potentially benefit from additional roadway or driver specific data should that city/county have access to said data.

Authors' contribution

Jeremiah Roland and Peter Way: project researcher, overall paper writing. Connor Firat: project data manipulation assistant, geo-spatial model design. Thanh-Nam Doan: project assistant, assistance with revisions. Mina Sartipi: project head, assistance with paper writing and revisions.

Conflict of interest

None declared.

Acknowledgment

We would like to thank the City of Chattanooga, Hamilton County Emergency Communications District, Tennessee Department of Transportation, and Chattanooga Department of Transportation for supplying data for this research and valuable discussions. Furthermore, we would like to acknowledge the contributions to this project by Dr. Eric LaFlamme from Plymouth University for his help in different analysis techniques for our data and negative sampling. We also extend gratitude to the NSF-US Ignite-1647161 for partially supporting this project.

References

Abdel-Aty, M.A., Radwan, A., 2000. Modeling traffic accident occurrence and involvement. Accid. Anal. Prev. 32, 633–642. https://doi.org/10.1016/S0001-4575(99)00094-9.
Bisong, E., 2019. More supervised machine learning techniques with scikit-learn. Building Machine Learning and Deep Learning Models on Google Cloud Platform 287–308.
10 Leading Causes of Death By Age Group, United States, 2017. https://www.cdc.gov/injury/wisqars/pdf/leading_causes_of_death_by_age_group_2017-508.pdf.
Cheng, J., Dong, L., Lapata, M., 2016. Long Short-Term Memory-Networks for Machine Reading. arXiv:1601.06733.
Guide to Calculating Costs, 2018. https://injuryfacts.nsc.org/all-injuries/costs/guide-to-calculating-costs/data-details/.
Dan, F., Michelle, G., Erika, S., 2018. Estimating Traffic Crash Counts Using Crowdsourced Data: Pilot Analysis of 2017 Waze Data and Police Accident Reports in Maryland.
Hébert, A., Guédon, T., Glatard, T., Jaumard, B., 2019. High-Resolution Road Vehicle Collision Prediction for the City of Montreal. arXiv:1905.08770.
Hamilton County Quick Facts, 2018. https://www.census.gov/quickfacts/fact/table/US/PST045219.
Khattak, A., Jun, L., Meng, Z., 2017. Highway Safety Manual: Enhancing the Work Zone Analysis Procedure Southeastern Transportation Center. Southeastern Transportation Center, Knoxville TN.
Kim, T.K., 2015. T test as a parametric statistic. Korean J. Anesthesiol. 68, 540.
Kotsiantis, S., 2013. Decision trees: a recent overview. Artif. Intell. Rev. 39, 261–283. https://doi.org/10.1007/s10462-011-9272-4.
Li, Y., Bai, Y., 2008. Development of crash-severity-index models for the measurement of work zone risk levels. Accid. Anal. Prev. 40, 1724–1731. https://doi.org/10.1016/j.aap.2008.06.012.
Liaw, A., Wiener, M., 2002. Classification and Regression by Randomforest. R News 2.3, pp. 18–22.
tf.keras.optimizers.nadam | tensorflow core v2.3.0, 2015. https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Nadam.
Olaisen, H., Rossen, L., Warner, M., Anderson, R., 2019. Unintentional Injury Death Rates in Rural and Urban Areas: United States, 1999 to 2017. https://www.cdc.gov/nchs/data/databriefs/db343-h.pdf.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Ramchoun, H., Idrissi, M.A.J., Ghanou, Y., Ettaouil, M., 2016. Multilayer perceptron: architecture optimization and training. IJIMAI 4, 26–30.
Ranjan, C., 2019. Extreme rare event classification using autoencoders in keras. Proceedings of Towards Data Science.
Sainath, T.N., Vinyals, O., Senior, A., Sak, H., 2015. Convolutional, long short-term memory, fully connected deep neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 4580–4584.
See, C., 2008. Crash Analysis of Work Zone Lane Closures With Left-Hand Merge and Downstream Lane Shift.
Shung, K.P., 2018. Accuracy, Precision, Recall, or F1?.
Suthaharan, S., 2016. Support Vector Machine, Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, pp. 207–235.

Sze, V., Chen, Y., Yang, T., Emer, J.S., 2017. Efficient processing of deep neural networks: a tutorial and survey. Proc. IEEE 105, 2295–2329.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256. https://doi.org/10.1016/j.aap.2014.06.017.

Tennessee Traffic Crashes by Year, Time of Day and County, Titan, 2007–2019, 2018. https://www.tn.gov/content/dam/tn/safety/documents/TimeOfDay.pdf.

Traffic Deaths, 2009-2018, 2018. NHTSA. https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812749.

Weng, J., Meng, Q., 2011. Analysis of driver casualty risk for different work zone types. Accid. Anal. Prev. 43, 1811–1817. https://doi.org/10.1016/j.aap.2011.04.016.

Yuan, Z., Zhou, X., Yang, T., Tamerius, J., Mantialla, R., 2017. Predicting traffic accidents through heterogeneous urban data: a case study. Proceedings of 6th International Workshop on Urban Computing.

Yuan, Z., Zhou, X., Yang, T., 2018. Hetero-convlstm: a deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 984–992. https://doi.org/10.1145/3219819.3219922.