Fall 12-15-2021

Traffic Crash Prediction Using Machine Learning Models

Yifeng Chen
ychen7@my.harrisburgu.edu

Traffic Crash Prediction Using Machine Learning Models

Chen, Yifeng

ANLY 699

Harrisburg University of Science and Technology

Abstract

Traffic crashes account for most of casualties and injuries worldwide, and there has been growing concerns and studies regarding the contributing factors of traffic crashes. There are many factors causing or related to an occurrence of traffic crash, e.g., land use, traffic flow conditions, driver behavior and weather condition. This paper studied the spatial and temporal distribution of crashes on highway and developed real-time prediction models for crash occurrence. Traffic flow data, weather data, and crash data from multiple data sources were collected and processed to develop the model. Multiple machine learning models, such as SVM model and Decision Tree model, were used as the candidate models. It was found that weather, crash time, and traffic flow shortly prior to the crash occurrence are critical impacting factors for real-time crash prediction. The candidate models have low to moderate sensitivity to predict the crash occurrences due to limited sample size. To use the models in a traffic operations environment, a prediction tool with interactive map could be developed to proactively monitor crash hot spots, and prepare staffing and resources for the potential crash occurrences.

Keywords: Spatial-temporal Analysis, Traffic Flow, Machine Learning Models, Crash Prediction

Introduction

As a result of rapid increase of vehicle ownership and traffic demand, traffic crashes account for most casualties and injuries worldwide, especially for those of young people between 15 and 29 years (Jia, Khadka, & Kim, 2018). Annually, about 1.35 million people per year lost their lives due to traffic crashes, and 20 to 50 million people are injured in traffic crashes (World Health Organization, 2018). The occurrence of traffic crashes seems random, because oftentimes their causes are complicated, and are related to multiple factors, such as bad driving behavior (e.g., alcohol impacted, distraction from texting and drowsy driver), severe weather, icy road surface, low visibility, or a combination of these factors. Which factors contribute most to the traffic crashes? How can we reduce the risk of traffic crashes and improve road safety?

To address these questions, comprehensive and quantitative/qualitative analysis with multiple data sources is needed. Therefore, data mining and statistical models are critical to quantify the contribution of these factors to traffic crashes in the studies of traffic crashes and safety. The objective of this study is to study the spatial and temporal distribution patterns of traffic crashes, if any, and their correlations to potential impacting factors. With this knowledge, effective countermeasures, such as street lighting improvement, speed warning systems, and road geometry improvement, can be carried out to lower or eliminate the risk of crashes, or reduce the severity of crashes, at specific locations. Intuitively, more traffic crashes are expected during severe weather like thunderstorm or snowstorm, and more crashes usually occur during weekdays than weekend due to higher traffic volumes.

Previous studies (Chung, Abdel-Aty, & Lee, 2018; Aghajani, Dezfoulian, Arjroody, & Rezaei, 2017) have demonstrated that weather and traffic volumes are two major factors contributing to traffic crash occurrence. Other contributing factors found in the studies (Rolison,

Regev, Moutari, & Feeney, 2018) include human factors (drowsiness, distraction etc.), pavement surface conditions, and visibility. In the study of Jia et al. (2018), the spatial distribution of traffic crashes is found to be highly related to land use properties, especially the land use of public services, such as hospitals and banks. Many studies have focused on the relationship between crash hot spots and different contributing factors from mid-term or long-term perspective. Limited studies look into the prediction of crash occurrence in short-term period or real-time. This study is trying to contribute to this gap.

In this study, a highway corridor in Michigan is selected and multiple data sources pertaining to the study corridor, containing weather data, traffic crash data, traffic volume data, and corresponding geographical data etc., were used to analyze the spatial-temporal distributions of crashes, and develop models for predicting the probability of crash occurrence. Geographical Information System (GIS) is a powerful tool to process, visualize and analyze geographically related data in this study. The final product of the study is a short-term, or even real-time prediction tool, which is based on the prediction models developed in the study. The prediction tool can predict probabilities of traffic crashes cross the study road network/corridor at different time of day, which can be displayed as a dynamic heat map.

Literature Review

An intuitive question arising after the occurrence of a traffic crash would be: what is the cause of the crash? This remains to be a big topic in the research area of traffic safety and accident prevention. Many studies found it highly related to drivers' characteristics and behavior. For example, Rolison et al. (2018) conducted a comprehensive survey study on significant contributing factors of driver's behavior to traffic crashes in England, UK. It was found that impaired driving (drugs or alcohol), speeding, inexperience, distraction, medical condition, and

poor eyesight, are the six most significant factors to the crashes in the hypothetical scenarios in the survey.

Besides drivers' behavior, environmental and geographical factors such as weather, visibility, land use properties, and road layout are also related to traffic crashes. A crash hot spot is identified as locations, where crashes, especially crashes with injuries or casualties, are more likely to occur than its neighboring area (Hakkert & Mahalel, 1978). To identify crash hot spots and the contribution of environmental and geographical factors, spatial analysis is conducted in many studies. For example, Anderson's (2009) studied the identification of traffic crash hot spots using clustering method, and in addition, associates the hot spots with environment and land use data to further understand the complexities of traffic crashes and their spatial dependence. Jia et al. (2018) conducted spatial analysis of traffic crashes in Suzhou Industrial Park (SIP) with datasets of clustered point of interest (POI) information. The results show that both spatial lag model (SLM) and spatial error model (SEM) outperform the Ordinary least squares (OLS) model, and the SEM has the best performance in terms of coefficient of correlation and p-value. Vaz, Techranchi, and Cusimano (2017) specifically studied the spatial distribution of traffic crashes in the greater Toronto area (GTA). The results of the study show that there are statistically significant clustering patterns in the traffic crashes of vehicles with pedestrians, vehicles with trains, and vehicles with vehicles. The hot spots of crashes are located at areas with higher population density or where railroad crosses roadways. For the factors contributing to the number of crashes, the results of the GWR model show that the percentage of seniors and education level of residence were found to have the strongest correlation. Chung et al. 's (2018) study focused on fatal traffic crashes, and analyzed the spatial relationship between adverse weather, specifically in the format of weather station coverage, and fatal crashes. The

scope of the study covers all states of America, which was divided into nine climate regions in the study, while most studies in the past cover only a part of the states. The study concludes that adverse weather conditions are significantly related to the increase of fatal crashes, and that most weather-related fatal crashes occurred during rainy, snowy, or foggy days. In addition, traffic volume, in the format of vehicle miles traveled (VMT), was found to be a significant positive independent variable in the models, too.

The above studies focus on identifying crash hot spots and estimating crash frequency at mid-term or long-term period (e.g. monthly or annually) and at macro level (e.g. county or nationwide). With the advance of Intelligence Transportation System (ITS) and real-time traffic data collection techniques, there has been increasing interest in short-term or real-time crash prediction at local highway corridor level. For example, Abdel-aty and Pemmanaboina (2006) developed a methodology to predict the probability of crashes along a freeway corridor, using historical weather data and ITS traffic flow data collected at loop detectors installed along the freeway corridor. The paper aimed at real-time crash prediction at highway corridor level. A matched case-control logit regression (LR) model was applied to predict the odds of a crash occurrence at a location associated with a loop detector station on the corridor. It was found that high variation in the traffic flow, in terms of traffic speed, volume and occupancy, 5-10 minutes prior to the crash occurrence and heavy rain weather increase the odds of crash occurrence most significantly and indicate crash-prone conditions. Similar results were found by Lee et al. (2002) that the variation in traffic speed and density is significantly related to the likelihood of crash occurrence, and thus are crash precursors.

From the literature review, it has been shown that contributing factors to traffic crashes include severe weather, traffic flow characteristics such as traffic volume and speed, land use

properties, human factors, and pavement conditions etc. Many previous studies focused on the impacting factors of crashes, and their relationship with hot spot from regional or national level, for mid-term to long-term period. However, not many studies investigated the impacts of these factors at corridor level for a short-term period or in real time. In addition, most studies have used clustering method or traditional statistical models such as regression models. However, advanced machine learning models, such as random forest model and dynamic neural network (DNN) model, have not been used extensively in the field of crash analysis.

Research Questions

The purpose of this study is to answer the following questions: 1) can we develop a short-term, or a real-time prediction model to detect or predict potential crashes at specific locations? 2) which contributing factors can we use to predict crashes? 3) Can we use machine learning models to predict crash occurrence? It is hypothesized that 1) a crash is highly associated with the traffic flow in proximity to the crash locations and weather conditions when the crash occurs, 2) the proposed machine learning models can identify the relationship between potential crashes and external factors such as traffic volumes and weather conditions in a short-term period and predict the probability of crashes at highway corridor level.

Methodology

Data Description and Processing

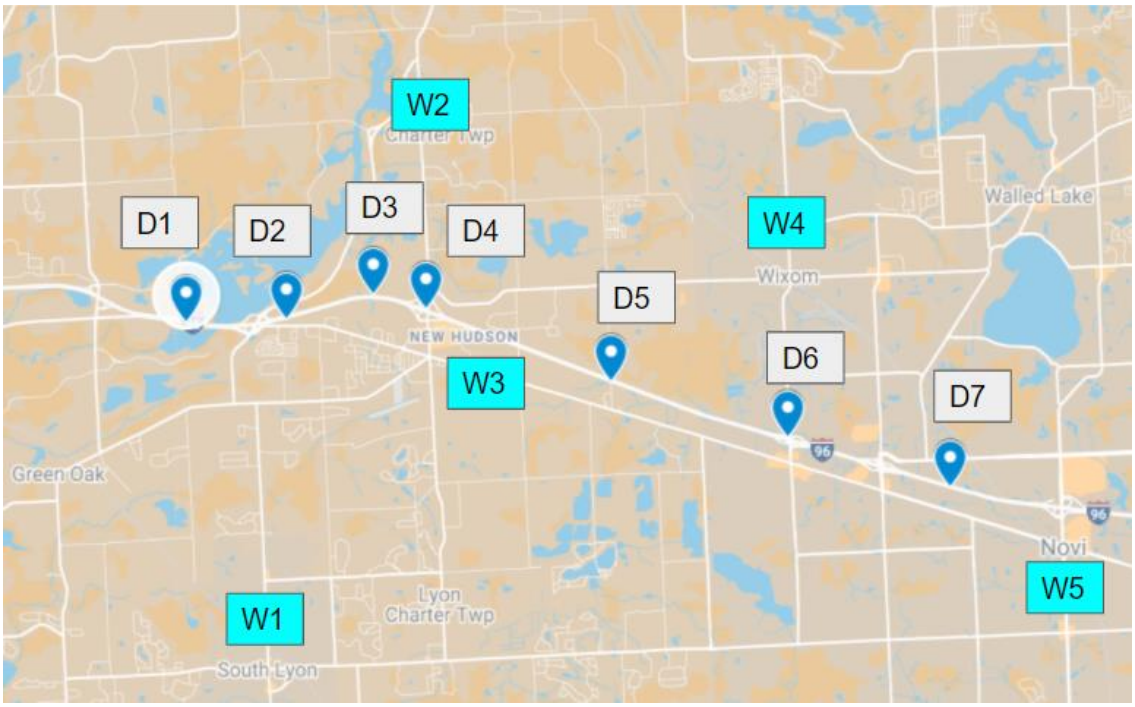
Traffic Data.

The traffic data for this study was collected from the radar detector stations, owned by Michigan Department of Transportation (DOT). The radar detectors along the highways in Michigan measure three parameters of traffic flow: vehicle speed in mile per hour (mph), traffic volume in vehicles per hour (vph) and detector occupancy rate in percent. Detector occupancy

rate represents the density of traffic flow and is defined as the percentage of time a radar detector is occupied by vehicles within the detection interval, e.g., 30 seconds. For the study, I selected a 9-mile-long segment on I-96 freeway in City of Wixom, Michigan as the study corridor. And I used traffic flow data between January 2017 and September 2019 from Michigan DOT's traffic database for the study. The freeway segment has three lanes in each direction on its mainline. There is a total of 7 radar stations and 14 radar detectors on this segment of the freeway, each of which is about 1 to 2 miles apart. Appendix1 from the Appendix shows the information regarding radar detection station and radar detectors, including their names, directions, and crossing road. It can be seen from Appendix1 that there are two radar detectors at each radar detection stations, each detecting one direction of traffic on the freeway. The map of the study corridor and the locations of the radar detection stations are illustrated in Figure 1.

Figure 1

Study Corridor



Note. The map shows the location of the detection stations, as well as their associated weather stations (see Appendix 1 and Appendix 6).

The original data were collected in 30 seconds intervals. A program developed in MATLAB was used to aggregate the data to a specific interval, e.g., 2 min, 5 min, or 15 min. In this study, 2 min interval was used to gain the highest resolution of the traffic flow. The original data are measured by each lane of the freeway. The MATLAB program averages the speeds and occupancies of all lanes to obtain the average speed and average occupancy, entitled approach speed and approach occupancy, respectively. The program also sums up the lane volumes to obtain approach traffic volume. Then the program removes contradicting data points in the dataset, e.g., zero speeds but positive traffic volumes, or negative speeds but positive volumes. It also removes unrealistic data points, e.g., speeds higher than 120 mph, or occupancy greater than 100%. The original datasets were collected monthly, and the MATLAB program combines

monthly datasets into one big dataset, using the 'merge' function in MATLAB. There is a total of 76 million rows in the dataset, with a size of 1.5 gigabytes. This is the largest dataset in the study. Appendix 2 illustrates an excerpt of cleaned traffic data. For simplicity, I used approach traffic data rather than lane-based traffic data, with the assumption that most crash events will impact at least one lane of traffic, and thus will impact the average traffic flow on the approach.

I also used the function describe() from Pandas in python to calculate the summary statistics, and listed it in Appendix 3. The statistics table shows that there are outliers in the traffic volume data because the maximum value of volume approach is more than 55 million vph, and standard deviation is 19 thousand vph. According to the highway capacity manual (HCM, 2010), the practical traffic volume on a single freeway lane is mostly below 2200 vph, and this makes a total of 6600 vph for a three-lane freeway. In addition, there are outliers in the occupancy data, because the mean occupancy is 1245%, and the maximum occupancy is 65535%. Logically, the occupancy cannot exceed 100%. Therefore, the outliers in the approach volume and occupancy data are false data and need to be removed from the original dataset. The speed data statistics look reasonable with a mean speed of 70 mph, and a maximum of 120 mph. Therefore, a filtering was conducted in python to remove the measurements with traffic volume greater than 6600 vph, and occupancy greater than 100 percent. The other issue I identified in Appendix 3 is that there is missing data in the speed data because the count of speed data measurements is less than that of the other data measurements. Then I ran a summary of null data in python (see in Appendix 4), and there are around 419 thousand of null speed data, and the corresponding volume and occupancy data are all zeros. Therefore, these null speed data should have zero speeds, since there are no vehicles detected during these measurements. I then converted the null speed data (i.e. 'NaN' in the data) to zero speed. A summary of the cleaned

dataset is shown in Appendix 5. After removing the false outliers, the mean and maximum value of approach volume and approach occupancy are within the expected range. Meanwhile, the number of speed data measurements is the same as that of the other measurements.

Weather Data.

The historic weather data is downloaded from a third-party weather forecast data website named Darksky (Dark Sky Team, 2015), which provides Application Programming Interface (API) data feed and can be accessed through Python or R after I registered an account on it. Darksky contains historical and forecast weather data from weather stations worldwide. In this study, weather stations near the study corridor were selected. A python program was developed to access the API and download hourly weather data between 2017 and 2019 from Darksky. The hourly weather data provides information including timestamp, temperature, precipitation, weather conditions (e.g., clear, rain, snow, fog), wind speeds etc. The Darksky website uses the weather data from the closest weather stations at the nearby town or city. Five weather stations of weather data were used for the study corridor, as shown in Appendix 6. The locations of these weather stations are illustrated in the study corridor map in Figure 1. Appendix 7 provides an excerpt of the weather data downloaded through the Darksky API. The datasets were downloaded by year and station, which makes a total of 15 data files (3 years by 5 stations).

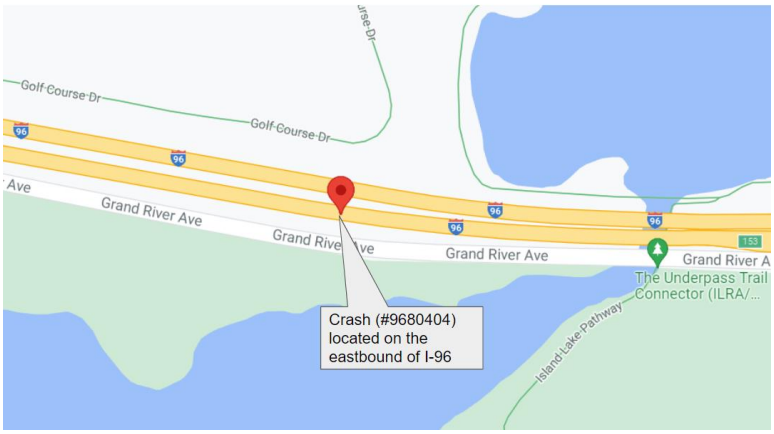
Crash Data.

The crash data was obtained from Michigan State Police's UD-10 traffic crash report database (MSP, 2021), which contains detailed crash information such as crash ID, timestamp of crash, GPS coordinates of crash locations, number of lanes impacted by the crashes, fatal or injury, weather conditions, number of vehicles involved, and information of involved vehicles etc. Appendix 8 (a), (b) and (c) show an excerpt of the crash data in three parts.

I applied a combination of R program and manual cleaning to clean up the crash data. The R program first deleted empty rows and rows with 'NA' values. The original crash data were saved as a text file, and some of the columns shifted to the neighboring columns after imported into R in csv format due to inconsistent spacing in the original text files. The R program filtered out those columns with the wrong data, and then deleted them and shifted the rest of the columns to the left to align the data columns. The datasets were then output to csv file again, and I manually checked if there were other anomalies in the datasets. I found that some crashes occurred on the eastbound of the freeway, however, the direction was marked as 'West' in the data table. The last two rows in Appendix 8 represent a crash (#9680404) occurred on the eastbound of I-96. Figure 2 shows the location of the crash on Google map, and verified it occurred on the eastbound approach. However, the direction was marked as 'West' in the raw data. Therefore, I manually corrected the direction to 'East' in the table and corrected the other rows with the similar errors.

Figure 2

Crash (#9680404) located on the eastbound of the freeway



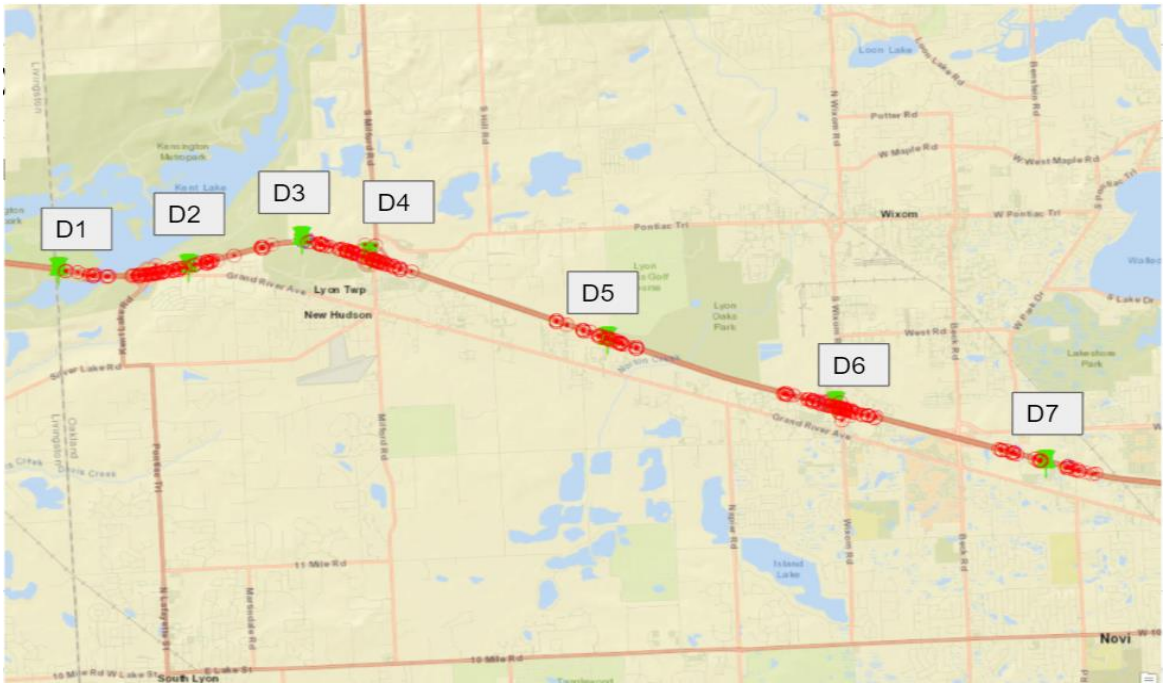
Note. Figure 2 shows the location of a crash (#9680404) occurred on the eastbound of I-96 on Google map. However, the direction was marked as 'West' in the raw data and was corrected to 'East' in the last two rows of Appendix 8.

The crash data information is used as the target feature in the training model, to label whether a crash occurred at locations near one of the radar detectors. For this purpose, each crash location is assigned to its closest radar detector location. A python program was developed to calculate the distance between all crash locations and each of the detector locations and output the results to excel spreadsheet. Then I identified the crashes within 0.5 mile of each of the detector stations using the filter function in excel and add the columns for the closest radar detector to each crash, their distances to the radar detector, and the GPS coordinates of the radar directions, respectively.

It can be observed from the Appendix 8 that there is more than one row for some of the crash IDs, each row representing an involved vehicle in the crash. For the modeling purpose, I do not need the vehicle information. Therefore, the python program deleted the columns for the vehicle information and removed the duplicate rows for the same crash IDs. After the cleanup, there are a total of 540 crashes from 2017 to 2019 along the study corridor. Appendix 9 illustrates an excerpt of the cleaned crash data. Figure 3 is a GIS map illustrating the crash locations along the radar detection stations on I-96. It can be seen from Figure 3 that all the selected crashes occurred in proximity to the seven crash locations.

Figure 3

A GIS Map of Crashes and Radar Detection Stations



Note. The GIS map shows the crash locations as red circle along the freeway, against the 7 radar detection stations as green pushpins, to verify all the crashes occurred on the I-96 freeway, and in proximity to the radar detection stations.

Model Selection

In this study, the target variable of the model is whether a crash will occur or not. In another word, the target feature is a binary variable of 0 and 1, with 0 representing non-crash event and 1 representing crash event. Initially, three machine learning models, including decision tree model, random forest model (Kelleher, Namee & D'Arcy, 2015), and Dynamic Neural Network (DNN) model (Goodfellow, Bengio & Courville, 2016), were selected as the candidate model for analyzing the relationship between crash probability and its contributing factors, such as traffic volume and weather condition. These machine learning models can be used for binary

classification problems. In addition, random forest model can identify the feature importance for each input feature the model.

I also observed from the data processing that the number of crashes (i.e., 540 crashes) is only 2.3 percent of the number of the hours between Jan 2017 and Sep 2019 (23,760 hours). of the data comparing to the traffic data. Therefore, most of the target features would be zeros in the training data. With this consideration, support vector machine (SVM) model (Kelleher, Namee & D'Arcy, 2015) was also selected as for comparison, as it can better handle the data with unbalanced binary outputs (Kelleher, Namee & D'Arcy, 2015).

Data Merging

Each crash event was assigned to the nearest weather station and radar detection station listed in Appendix 1 and Appendix 6, depending on their geographical locations. In this way, I associated the crashes with local weather conditions and traffic flow conditions prior to, during and after the occurrence of crashes. These are the key steps before building up the models in python.

As shown in the data description and processing section, Python and R were used as the major programming tools for the data processing, model development, model calibration and validation. The weather data, traffic volume data, and crash data are all time-series data. However, they have different time intervals. For example, traffic volume data are recorded at 2 min intervals, while the weather data are usually in one-hour intervals. After all sources of data were collected and imported into Python, the data manipulation and analysis library in Python named Pandas (Tutorialspoint, 2006) was used to process, combine, and align different sources of data to the same time intervals.

After importing the different datasets mentioned above into python, I applied the data merging function ‘pandas.merge’ in python pandas to align traffic data, crash data, and weather data horizontally, so that they were combined to one big dataset for the modeling. The traffic data are aggregated in 2-minute intervals, while the weather data are collected every hour. Therefore, I merged weather data with traffic data by the time variables of year, month, day, hour, and by the assigned weather station. The merged traffic and weather data are the intermediate dataset. I then merged the crash data with the intermediate data to form the full dataset for the model training. The crash data are timestamped to the accuracy of a minute. The key variables I selected to merge the crash data and the intermediate data are year, month, day, hour, minute, and radar detector names, so that I can include all the traffic and weather data within same hour of the crash occurrence. I also randomly selected non-crash hours, i.e., hours with no crash occurrences, from the intermediate dataset and combine them with the dataset for crash hours as the model training dataset. In this way, the model will learn the traffic and weather conditions for both crash and normal traffic situation.

Model Evaluation

After the model is built up and trained with the training dataset, a comparison between the four candidate models was performed using the testing dataset. Statistical measures for binary classification model, such as confusion matrix, precision/ accuracy, sensitivity, specificity, and Cohen’s kappa (Sim & Wright, 2005), can be used to evaluate performance of the models.

Model Development

Preliminary Data Analysis

Crash Data Analysis.

Table 1 lists the number of crashes for each year of the study period. It can be seemed that there are around 50% less crashes at the study corridor in 2015 comparing to 2016 and 2017.

Table 1

Number of Crashes by Year

Year	Number of Crashes
2015	101
2016	201
2017	238
Total	540

Note. The table shows a breakdown of the number of crashes by year, and the total number crashes used for the model training and testing.

A spatial-temporal analysis for the crash data was conducted and the results were illustrated in Appendix 10-Appendix 14. From the analysis, the winter months from October through January have higher number of crashes than the rest of the months. Within a month, the number of crashes has a pattern that peaks in around every 7 days. Within a day, the number of crashes has a morning peak between 7 AM and 10 AM and a afternoon peak between 4 PM and 7 PM. There is not significant difference in the number of crashes between the traffic directions of eastbound and westbound. The breakdown of number of crashes by the seven radar detection stations shows that the stations R2, R4 and R6 have significantly higher number of crashes than R1, R3 R5, and R7.

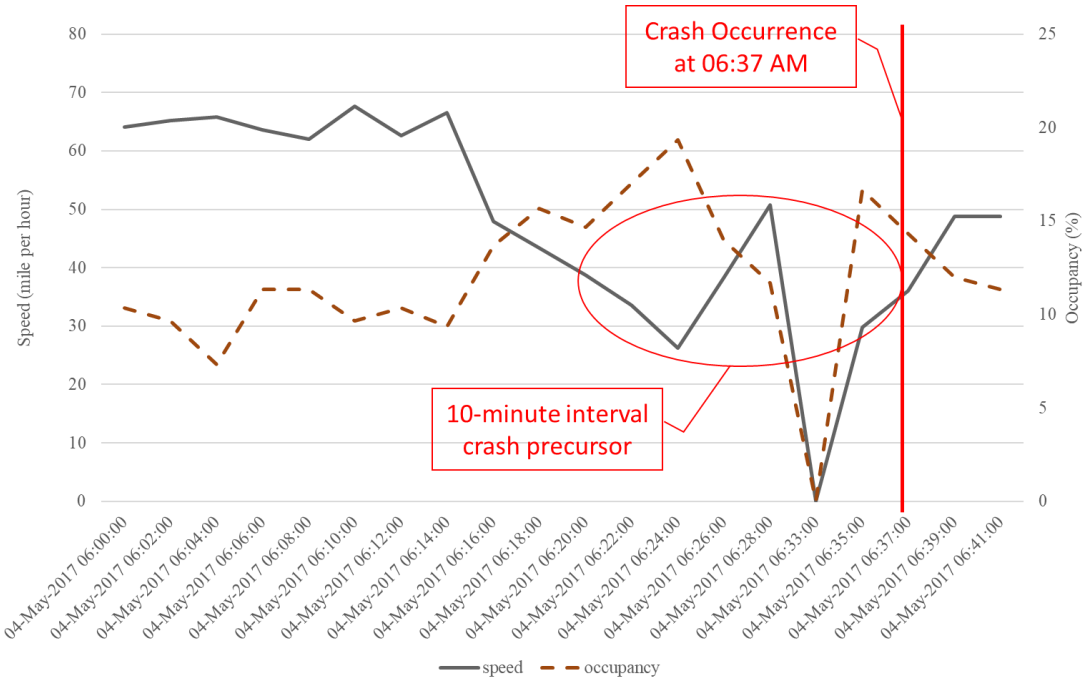
Traffic Data Analysis.

Figure 4 illustrates an example of crash impact on traffic speed, volume and occupancy at a radar detection station. From the crash data, the recorded crash occurrence time was at 6:37 AM, however, the profiles of the speed, occupancy, and volume suddenly dropped to zero exactly at 6:34 AM. This indicates that the traffic flow stopped harshly due to the crash, and the impact of the crash hit the traffic flow at the radar station four minutes earlier than the recorded crash time. The example also shows that the dramatical variation in the traffic volume, speed and occupancy around 10 min prior to the crash occurrence is a precursor of a crash. This is consistent with the conclusions from Abdel-aty & Pemmanaboina (2006) that the traffic flow (i.e., speed, volume, and occupancy) 5 to 10 minutes prior to a crash occurrence has the most significant influence in predicting the crash events. Therefore, I decided to use both the average of and change in traffic flow (i.e., volume, speed, and occupancy) 10 minutes prior to the crash occurrence as the input features of the model.

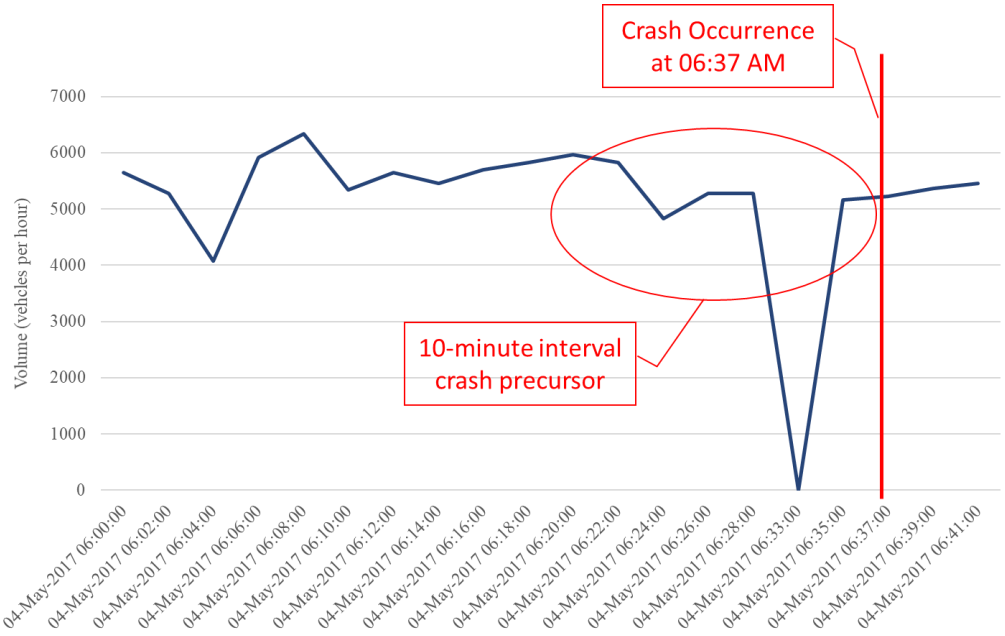
Figure 4

Traffic speed, occupancy, and volume profiles, before, during and after a crash occurrence

(a) Speed and occupancy profile for a crash event on May 04, 2017, 6 AM



(b) Volume profile for a crash event on May 04, 2017, 6 AM



Note. Table 4 (a) illustrates the traffic speed and occupancy profile, prior to, during and after a crash event on May 04, 2017, 6 AM. Table 4 (b) illustrates the traffic volume profile, prior to, during and after the same crash event. The red oval in the plots highlights the dramatic variation in the profiles of the traffic speed, occupancy, and volume around 10 minutes prior to the crash occurrence. The variation of traffic flow in the 10-minute interval indicates a precursor for the crash occurrence.

Input Features and Target Features

Based on the spatial-temporal analysis of the crash events, the number of crashes has a pattern along month of year, day of month and time of day. Therefore, I include these three time related variables in the input features, so that the model can learn the seasonality trend of the crashes. There is no significant difference between the eastbound and westbound traffic directions, and the model is not calibrated for specific locations. Therefore, the traffic direction and the radar detection stations are not included in the input features.

Based on the above analysis, I selected traffic volume, volume change, traffic speed, speed change, occupancy, occupancy change, 10 minutes prior to the crash occurrence as the traffic flow related input features for the model. Volume change, speed change and occupancy change are calculated as the difference between the current and the previous timesteps. Since the traffic data are measured in 2-minute intervals, the 10 minutes interval is equivalent to 5 timesteps back from the time of the crash occurrence. In the crash data, the time stamp is recorded to the accuracy of one minute, while the traffic data are measured in 2-minute intervals. In addition, the crash location is within 0.5 mile distant to the radar detector location. Due to the temporal and spatial differences between the crash occurrence and the traffic data measure, I define the closest timestamp of traffic data prior to the crash timestamp as the timestep of the

crash occurrence and keep the traffic data 5 timesteps prior to the crash occurrence timestep as traffic flow related input feature data.

I also selected weather condition, temperature, and precipitation amount as the weather-related input features. There are 14 types of weather conditions in the original weather data. I categorized the weather conditions from the weather data into the two classes: good weather and bad weather. Thus, the weather condition feature in the models is a binary variable, where '0' represents 'good weather' and '1' represents 'bad weather'. For example, snow and heavy rain were categorized to 'bad weather', and clear and cloudy weather were categorized to 'good weather'. Appendix 15 lists the weather conditions from the original weather data, and their categories. It should be noted that the weather data were aggregated hourly, and therefore, were identical for the five timesteps prior to the crash occurrence.

Table 2 lists the continuous input features and their summary statistics. It can be seen that the values of these input features are within reasonable ranges.

Table 2

Summary table for the input features in the model

(a) Summary statistics for the traffic related input features

	n	mean	std	min.	25%	50%	75%	max.
volume (vehicles/h)	14076	2198.7	1590.5	0	860	1950	3330	6600
volume change (vehicles/h)	14076	465	623.4	0	120	270	570	6110
occupancy (percent)	14076	6.2	6.9	0	2	4.3	8	69.3
occupancy change (percent)	14076	1.6	3.2	0	0.3	0.7	1.7	67.3
speed (mile/h)	14076	63.9	20.9	0	60	70.7	75.9	120
speed change (mile/h)	14076	6.7	13	0	1.1	2.9	6.1	120

(b) Summary statistics for the weather-related input features

	n	mean	std	min.	25%	50%	75%	max.
Temperature (Fahrenheit)	14076	46.7	22	0.4	28.4	46.5	64.2	89.3
Precipitation Accumulation (inches)	14076	0	0	0	0	0	0	0.2

Note. Table 2 lists input features for the training model and their summary statistics, including count of the data points (n), mean, standard deviation (std), minimum, maximum, 25 percentile, 50 percentile (median) and 75 percentile of the data. It can be seen that the values of all features are within their reasonable ranges.

Table 3 shows the correlation values between the input features. It can be seen that the correlation value between traffic volume and traffic occupancy is 0.55. This indicates moderate positive correlation between the traffic volume and the occupancy. This is reasonable, because higher traffic volume means more vehicles occupy the radar detector, and thus results in higher

occupancy rate. The rest pairs of the input features have correlations lower than 0.5, and indifferent correlation.

Table 3

Correlation Table for Input Features

	volume (vehicles/hour)	occupancy (percent)	speed (mile/hour)	Temperature (Fahrenheit)	Precipitation Amount (inch)
volume (vehicles/hour)	1.00	0.55	0.32	0.18	0.05
occupancy (percent)	0.55	1.00	0.24	0.14	0.04
speed (mile/hour)	0.32	0.24	1.00	0.00	0.00
Temperature (Fahrenheit)	0.18	0.14	0.00	1.00	0.11
Precipitation Amount (inch)	0.05	0.04	0.00	0.11	1.00

Note. The table shows the correlation values between the major input features pairs. There are a total of five input features in the table.

The target feature is a binary variable named ‘Crash Occurrence’ with two values ‘0’ and ‘1’, where ‘1’ represents each crash event in the merged dataset, and ‘0’ represents no crash occurrence. I used python random number function to randomly select non-crash events out of the hours without crash events. To be consistent with crash events, I used the first ten minutes (i.e. five timesteps) of traffic flow data and the corresponding weather data in the non-crash hours randomly selected, and combined them with the crash related traffic data and weather data.

Data Table Transformation

Because the input features are time-series variables with 2-minute intervals, the original input feature data table is a tall table with 13,142 rows, Since I related each crash and non-crash

event to the 5 timesteps prior to their occurrence timestep, I used the 'pivot_table' function from python to reshape the tall table to a wide table, so that each crash or non-crash event and their related input feature data are in the same row. The transformed dataset table has a total of 1177 rows, where 867 of them (i.e., 74%) are non-crash events and 310 of them (i.e., 26%) are crash events.

Model Training and Parameter Tuning

To calibrate and validate the models, I used the sample split function from python to split the transformed dataset into 80% for training data and 20% for testing data, and make sure the portion of crash events in the training dataset is about the same as that in the testing dataset.

SVM, Decision Tree and Random Forest Model.

The functions from python sklearn library (Pedregosa et al., 2011) were used to build up the SVM, decision tree and random forest models. For the SVM model, three kernels, linear kernel, polynomial kernel, and radial basis function (RBF) kernel, were applied to the training and testing datasets, to find out which kernel fits the data best. For the random forest model, a built-in function named 'feature_importance' from python sklearn (Pedregosa et al., 2011) was used to evaluate the importance score of each input feature in the model. The evaluation metrics for model performance include confusion matrix, accuracy, precision, sensitivity (recall), and f1-score.

Dynamic Neural Network (DNN) Model.

I used the 'Sequential' function from python tensorflow to build up the DNN models. Two hidden layers with 50 neurons in each layer were built initially. Then one more hidden layer was added to improve the complexity of the model and overcome the overfitting issue. Sigmoid function was used as the activation function. Binary-cross entropy function was used as the loss

function of the model. The optimizer of the model was ‘adam’. The evaluation metric for model training was accuracy, which is defined as the ratio of the correct predictions, including true positive (TP) and true negative (TN), to the total number of predictions.

The hyperparameters include the number of hidden layers, number of neurons in the hidden layers, number of iteration epochs and batch size. I set the number of epochs to be 200, and batch size to be 50. Table 4 lists the hyper parameters of the DNN models.

Table 4

Hyper parameters of the DNN model

Hyper Parameter	Settings
Number of hidden layers	3
Number of Neurons in the hidden layer	50
Epochs	200
Batch Size	50

Note. The table lists the values of the hyperparameter used in the DNN model

Model Results and Comparison

SVM Model Results

The confusion matrix of the SVM models can be viewed in Appendix 16. Table 5 shows the performance metrics of the SVM models with three types of kernels. They are calculated from the confusion matrix. From Table 5, the model with RBF kernel has the highest model accuracy of 83%, followed by that with linear kernel with an accuracy of 82%. The model with polynomial kernel has the best sensitivity of 62%, which means the model can predict 62 percent of the crash events in the testing dataset. Although the F1 score of the model with the polynomial

kernel is 3% higher than that with the RBF kernel, its precision is 30% lower than that with the RBF model. Therefore, the SVM model with the RBF kernel has the best overall performance among the three kernels.

Table 5

Model Performance for the SVM models

Model	Model (Crash as Positive Case)				Model Accuracy
	Precision	Sensitivity (Recall)	Specificity	F1-score	
SVM (RBF Kernel)	93%	43%	99%	59%	83%
SVM (Linear Kernel)	92%	37%	99%	52%	82%
SVM (Polynomial Kernel)	63%	62%	86%	62%	79%

Note. These are the results of the model testing data.

Decision Tree Model Results

The confusion matrix of the Decision Tree can be viewed in Appendix 17. Table 7 lists the performance metrics for the decision tree model. It can be seen that the decision tree model has an overall accuracy of 87% and a sensitivity of 67%. The precision is 84%, which means out of all the predicted crash events 84% of them are actual crashes, and 16% are non-crashes. The specificity is 99%, which means it can predict 95% of the non-crash events correctly. The F1-score is 74%, which shows that the model has a good balance between precision and sensitivity.

Random Forest Model Results

The confusion matrix of the random forest model can be viewed in Appendix 18. Table 7 lists the performance metrics for the random forest model. It has an accuracy of 86%, and a sensitivity of 52%. The specificity is 99%, which means it can predict most of the non-crash events correctly. The F1-score is 67%, which indicates a good balance between precision and sensitivity.

Table 6 lists the top 10 impacting factors for predicting the crash with the random forest model. It can be observed that weather is the top one factor, followed by time of day, temperature, and day of month. The other factors include the speed and speed change at the previous time step of crash occurrence (i.e., timestep 1), speed at the third timestep prior to the crash occurrence (i.e., speed_timestep3), speed at the last timestep prior to the crash occurrence (i.e., speed_timestep5), volume change at the last timestep prior to the crash occurrence (i.e., volume change_timestep5), and the occupancy at the previous timestep (i.e., occupancy_timestep1)

Table 6

Importance Score of the Top 10 Input Features

Input Feature	Importance Score
Weather	0.06
Time of Day	0.05
Temperature	0.05
Day of Month	0.04
Speed_timestep1	0.04
Speed change_timestep1	0.03
speed_timestep3	0.03
speed_timestep5	0.03
volume change_timestep5	0.03
occupancy_timestep1	0.03

Note. The table lists the importance scores of the top 10 input features in descending order.

DNN Model Results

The confusion table of the DNN model can be viewed in Appendix 19 and Table 8 shows the performance metrics for the DNN model. It can be seen from Appendix 19 that no crash events were correctly predicted. Therefore, the sensitivity of the model was 0%, as shown in

Table 8. Although the specificity of the model was 100%, the models did not have the capability to predict the crash events, maybe due to the small portion of crash events comparing to the total events.

The learning curve of the DNN model for model accuracy and loss through the 200 iterations is plotted in Appendix 19. It can be observed that the accuracy curve of the testing data goes below that of the training data at the beginning of the iterations, and throughout the iterations. This indicates overfit of the model in the model training. The model was initially designed with two hidden layers, and I have added one more layer with 50 neurons to the DNN model after the initial model run. However, the model was still overfit due to insufficient sample size.

Model Comparison

Table 7 shows a comparison of the testing results of the four models. It can be seen that all the models have high specificity, which means they are capable of predicting most of the non-crash events. The decision tree and random forest models have higher accuracy and sensitivity than the SVM model, which means decision tree and random forest model can predict both crash and non-crash events better than the SVM model. The decision model has the highest accuracy, as well as the sensitivity and F1-score, which means the model has the best overall performance among the four models.

The DNN model has a 100% specificity, which means it predicted all the non-crash events. However, its sensitivity is zero, which indicates that it could not predict any of the crash events, and thus cannot be used for the prediction.

Table 7

Comparison of Models Performance

Model	Model (Crash as Positive Case)				Model Accuracy
	Precision	Sensitivity (Recall)	Specificity	F1-score	
SVM (RBF Kernel)	0.93	0.43	0.99	0.59	0.83
Decision Tree	0.84	0.67	0.95	0.74	0.87
Random Forest	0.94	0.52	0.99	0.67	0.86
DNN	0.00	0.00	1.00	0.00	0.73

Note. These are the results of the model testing data.

Discussion and Future Work

In this study, multiple data sources, including traffic data, weather data and crash data, were collected to analyze the impacting factors on crash occurrences. Spatiotemporal analysis on crashes was conducted for a highway corridor in Metro Detroit region for crash data. Crash data combined with geographical information, weather data, and traffic flow data were used to develop four types of machine learning models to predict whether a crash occurs at specific locations of the study corridor for the next two-minute interval. It was found that weather and traffic flow status around 10 minutes prior to the crash occurrence are precursor for the crash and can be used as input features in the crash prediction models. Overall, decision tree and random forest models perform better than SVM model and DNN model, with moderate sample size for model training. All four models were able to predict most of the non-crash events with a specificity of above 95%. However, the models could not predict crash events with high sensitivity. due to unbalanced crash and non-crash events in the target features. The decision tree model yields the highest sensitivity of 67% among the four models. From the importance scores

of the input features in the random forest model, weather condition (i.e., good or bad weather), time-related factors (e.g., time of day, day of month), and traffic flow related factors (e.g., traffic speed, speed change, volume and occupancy within 10 minutes prior to the crash occurrence) are the most important impacting factors of potential crashes.

From the literature review, crash occurrence on highways could be related to external factors, including traffic speed, traffic volume, road condition, weather, as well as driver behavior factors, such as alcohol, age, and education. In this study, only several external factors are investigated. However, there may be hidden factors contributing to individual crashes that were not included in the models. For example, the crashes occurred between 10 pm and 3 am might be related more to impaired driving or low visibility, rather than speed drop or bad weather. This could have limited the model's capability of identifying the potential crashes.

For future studies, a more comprehensive model including more environmental factors, such as road surface condition, visibility, and speed limit, are of interest. In this study, average traffic data across the highway lanes were used. However, the slowdown on one lane might not be caught with the average traffic data. Therefore, lane-by-lane traffic data with higher resolution can be used to increase the sensitivity of the model. In the study, the data from one corridor was used to develop the models. More data from different corridors can be collected to increase the sample size, so that we can calibrate a more generalized model and validate the transferability models.

With the fine-tuned prediction model, a prediction tool with interactive map could be developed. The tool can generate alerts for locations on a highway corridor with high probability of crashes for the next 2 to 5 minutes. Transportation agencies, first responders and state police patrol can use the tool to proactively monitor crash hot spots, and prepare staffing and resources

for the potential crash occurrences in advance. The information of risky areas can be disseminated to the public through dynamic message signs (DMS) on highways and social media like Twitter, especially during severe weather events, to provide situational awareness and advisory to the drivers.

Acknowledgments

The author would like to thank Michigan Department of Transportation (MDOT) for assisting with the data collection and sharing the data for this study. The contents of this paper reflect the views of the author, who is responsible for the facts and accuracy of the information presented herein and is not necessarily representative of the MDOT.

References

Abdel-aty, M. A., & Pemmanaboina, R. (2006). Calibrating a Real-Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data. *IEEE Transactions on Intelligent Transportation Systems*, *7*(2), 167–174.

Anselin, L. (2010). Local Indicators of Spatial Association-LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Al-Deek, H. M., Venkata, C., & Chandra, S. R. (2004). New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I-4 Data Warehouse. *Transportation Research Record: Journal of the Transportation Research Board*, *1867*(1), 116–126. https://doi.org/10.3141/1867-14

Aghajani, M. A., Dezfoulian, R. S., Arjroody, A. R., & Rezaei, M. (2017). Applying GIS to Identify the Spatial and Temporal Patterns of Road Accidents Using Spatial Statistics (case study: Ilam Province, Iran). *Transportation Research Procedia*, *25*, 2126–2138. https://doi.org/10.1016/j.trpro.2017.05.409

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention*, *41*, 359–364. https://doi.org/10.1016/j.aap.2008.12.014

https://doi.org/10.1016/j.apgeog.2019.04.008

Chung, W., Abdel-Aty, M., & Lee, J. (2018). Spatial analysis of the effective coverage of land-based weather stations for traffic crashes. *Applied Geography*, *90*, 17–27. https://doi.org/10.1016/J.APGEOG.2017.11.010

Dark Sky Team. (2015). *Dark Sky Weather Forecast*. darksky.

https://darksky.net/forecast/40.7127,-74.0059/us12/en

Getis, A., & Ord, J. K. (2010). The Analysis of spatial association by use of distance statistics. *Geographical Analysis*, *24*(3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press, Cambridge, Massachusetts. https://www.deeplearningbook.org/

Hakkert, A. S., & Mahalel, D. (1978). Estimating the number of accidents at intersections from a knowledge of the traffic flows on the approaches. *Accident Analysis & Prevention*, *10*(1), 69–79. https://doi.org/10.1016/0001-4575(78)90009-X

Highway Capacity Manual (HCM) (2010). Transportation Research Board of the National Academies, Washington, D.C..

Jia, R., Khadka, A., & Kim, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. *Accident Analysis and Prevention*, *121*(September), 223–230. https://doi.org/10.1016/j.aap.2018.09.018

Kelleher, D. J., Namee, M.B., & D'arcy A. (2015). Fundamentals of machine learning for predictive data analytics. MIT Press, Cambridge, Massachusetts.

Lee, C., Saccomanno, F., & Hellinga, B. (2002). Analysis of crash precursors on instrumented freeways. *Transportation Research Record*, *1784*, 1–8.

Michigan State Police (MSP). (2021). *UD-10 Traffic Crash Report*. michigan. https://www.michigan.gov/msp/0,4643,7-123-72297_24055_67691---,00.html

Rolison, J. J., Regev, S., Moutari, S., & Feeney, A. (2018). What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis & Prevention*, *115*, 11–24.

https://doi.org/10.1016/J.AAP.2018.02.025

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

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. *Journal of Machine Learning Research 12*, 2825-2830.

Tutorialspoint (2006). Python Pandas Tutorial. Retrieved August 5, 2019, from https://www.tutorialspoint.com/python_pandas/

Vaz, E., Techranchi, S., & Cusimano, M. (2017). Spatial assessment of road traffic injuries in the Greater Toronton Area (GTA): spatial analysis framework. *Journal of Spatial and Organizational Dynamics*, *5*(1), 37–55. Retrieved from https://pdfs.semanticscholar.org/dab4/139c9d27cbf454765f084926fb7ac3b2886a.pdf

World Health Organization. (2018). *Road traffic injuries*. who. https://www.who.int/en/news-room/fact-sheets/detail/road-traffic-injuries

Appendix

Appendix 1

Radar Detector Stations and Radar

Detection Station	Detector Name	Highway	Direction	Cross Road
D1	D1_1	I-96	East	Kensington
D1	D1_2	I-96	West	Kensington
D2	D2_1	I-96	East	Kent Lake
D2	D2_2	I-96	West	Kent Lake
D3	D3_1	I-96	East	W of Milford
D3	D3_2	I-96	West	W of Milford
D4	D4_1	I-96	East	Milford
D4	D4_2	I-96	West	Milford
D5	D5_1	I-96	East	Old Plank
D5	D5_2	I-96	West	Old Plank
D6	D6_1	I-96	East	Wixom
D6	D6_2	I-96	West	Wixom
D7	D7_1	I-96	East	E of Beck
D7	D7_2	I-96	West	E of Beck

Note. The table includes information on radar detector name, which direction of the traffic they measure, and the local crossroads of the detector location. The data comes from Michigan Department of Road.

Appendix 2

An Excerpt of Traffic Data

(a) Head of the Dataset

Date and Time	Detector Name	volume (vehicles/hour)	occupancy (percent)	speed (mile/hour)
12/1/2015 0:02	D-I96E MM1522 Kensington	600	1.00	60.00
12/1/2015 0:04	D-I96E MM1522 Kensington	490	1.67	70.00
12/1/2015 0:06	D-I96E MM1522 Kensington	460	2.33	51.11
12/1/2015 0:08	D-I96E MM1522 Kensington	160	0.33	53.33
12/1/2015 0:10	D-I96E MM1522 Kensington	510	1.67	51.00

(b) Tail of the Dataset

Date and Time	Detector Name	volume (vehicles/hour)	occupancy (percent)	speed (mile/hour)
9/19/2017 23:52	D-I96W MM1611 E of Beck	1470	4.00	73.50
9/19/2017 23:54	D-I96W MM1611 E of Beck	1280	3.67	71.11
9/19/2017 23:56	D-I96W MM1611 E of Beck	2070	5.67	73.93
9/19/2017 23:58	D-I96W MM1611 E of Beck	1560	4.33	74.29
9/19/2017 23:59	D-I96W MM1611 E of Beck	2120	6.33	73.10

Note. The excerpt shows the head and tail rows of the traffic flow data, including timestamps of measurements, detector name, volume in vehicles per hour (vph) for the approach, approach occupancy in percent, and approach speed in miles per hour.

Appendix 3

Preliminary Summary Statistics of Traffic Data

	n	mean	std	min.	25%	50%	75%	max.
volume (vehicles/hour)	7626659	2026	19999	0	650	1650	3150	55054560
occupancy (percent)	7607918	1245.33	8929.84	0	1.67	3.67	7.33	65535
speed (mile/hour)	7188111	70.46	11.3	1	6607 %	7214 %	7714 %	120

Note. Summary statistics for the traffic data, including count of the data points (n), mean, standard deviation (std), minimum, maximum, 25 percentile, 50 percentile (median) and 75 percentile of the data.

Appendix 4

Summary of Null Data

	n	mean	std	min.	25%	50%	75%	max.
volume (vehicles/hour)	419807	0	0	0	0	0	0	0
occupancy (percent)	419807	0	0	0	0	0	0	0
speed (mile/hour)	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Note. The table shows the number of null data ('NAN') in the speed data is 419,807. The corresponding occupancy and volume are zeros, which indicates that the null speeds should also be zero speeds.

Appendix 5

Summary Statistics of Traffic Data with Outliers and Null Data Removed

	n	mean	std	min.	25%	50%	75%	max.
volume (vehicles/hour)	7455198	2016	1589	0	650	1650	3140	7200
occupancy (percent)	7455198	4.96	4.99	0	1.67	3.33	7	74
speed (mile/hour)	7455198	66.49	19.58	0	64.07	71.74	76.73	120

Note. The summary statistics of the data show reasonable ranges after converting null data to zeros in the speed data.

Appendix 6

Weather Station

Detection Station	Weather Station	Location
D1	W1	South Lyon, MI
D1	W1	South Lyon, MI
D2	W2	Milford, MI
D2	W2	Milford, MI
D3	W2	Milford, MI
D3	W2	Milford, MI
D4	W2	Milford, MI
D4	W2	Milford, MI
D5	W3	New Hudson, MI
D5	W3	New Hudson, MI
D6	W4	Wixom, MI
D6	W4	Wixom, MI
D7	W5	Novi, MI
D7	W5	Novi, MI

Note. The table relates detection stations to their nearest weather stations, and the locations of the weather stations. There are a total of five weather stations, shown in the map in Figure 1.

Appendix 7

Excerpt of Weather Data

Time	Temperature (Fahrenheit)	Conditions	Precipitation Probability	Precipitation Accumulation (Inches)	Wind (mile/hour)	Weather Station
1/1/2017 0:00	29.20	Mostly Cloudy	0	0	4.73	W4
1/1/2017 1:00	27.93	Mostly Cloudy	0	0	3.00	W4
1/1/2017 2:00	27.02	Partly Cloudy	0	0	1.73	W4
1/1/2017 3:00	25.67	Clear	0	0	1.20	W4
1/1/2017 4:00	24.00	Clear	0	0	1.11	W4

Note. This is an excerpt of the weather data downloaded through the Darksky API, including timestamp, temperature (Fahrenheit), precipitation (Inches), weather conditions (e.g. clear, rain, snow, fog), and wind speeds etc

Appendix 8

An Excerpt of Crash Data with Errors

(a)

Crash ID	Date_time	Latitude	Longitude	Primary Road	Direction	Intersecting Road
9606569	1/18/2016 22:20	42.52	-83.66	I-96	West	KENT LAKE
9606569	1/18/2016 22:20	42.52	-83.66	I-96	West	KENT LAKE
9606570	1/18/2016 23:25	42.52	-83.66	I-96	West	KENT LAKE
9621026	2/9/2016 8:00	42.52	-83.66	I 96	West	KENT LAKE
9621026	2/9/2016 8:00	42.52	-83.66	I 96	West	KENT LAKE
9642465	3/5/2016 4:45	42.52	-83.66	I 96	West	KENT LAKE
9642465	3/5/2016 4:45	42.52	-83.66	I 96	West	KENT LAKE
9659681	3/5/2016 8:51	42.52	-83.66	I96	West	KENT LAKE
9659681	3/5/2016 8:51	42.52	-83.66	I96	West	KENT LAKE
9680404	4/21/2016 7:30	42.52	-83.66	E I 96	East	RAMP 007A
9680404	4/21/2016 7:30	42.52	-83.66	E I 96	East	RAMP 007A

(b)

Crash ID	Fatalities	Injuries	Crash Type	Weather	Light Condition	Road Condition
9606569	0	0	Sideswipe-Same	Snow	Dark-Unlighted	Snow
9606569	0	0	Sideswipe-Same	Snow	Dark-Unlighted	Snow
9606570	0	0	Single Motor Vehicle	Snow	Dark-Lighted	Snow
9621026	0	0	Sideswipe-Same	Snow	Daylight	Snow
9621026	0	0	Sideswipe-Same	Snow	Daylight	Snow
9642465	0	0	Sideswipe-Same	Snow	Dark-Unlighted	Snow
9642465	0	0	Sideswipe-Same	Snow	Dark-Unlighted	Snow
9659681	0	1	Head On	Snow	Daylight	Snow
9659681	0	1	Head On	Snow	Daylight	Snow
9680404	0	0	Sideswipe-Same	Rain	Daylight	Wet
9680404	0	0	Sideswipe-Same	Rain	Daylight	Wet

(c)

Crash ID	Number of Vehicles Involved	Involved Vehicle	Vehicle Type	Vehicle Year	Vehicle Make
9606569	2	1	Motor Vehicle	2003	SATURN
9606569	2	2	Motor Vehicle	2013	FORD
9606570	1	1	Motor Vehicle	2006	FORD
9621026	2	1	Motor Vehicle	2013	CHEVROLET
9621026	2	2	Motor Vehicle	2010	TOYOTA
9642465	2	1	Motor Vehicle	2008	CHEVROLET
9642465	2	2	Motor Vehicle	2011	IHC
9659681	2	1	Motor Vehicle	2008	CHEVROLET
9659681	2	2	Motor Vehicle	2015	FRHT
9680404	2	1	Motor Vehicle	2015	FORD
9680404	2	2	Motor Vehicle	2015	JEEP

Note. The table is an excerpt of the crash data, including crash ID, timestamp of crash, GPS coordinates of crash locations, direction of the traffic where the crash occurred, information of the vehicles involved etc.

Appendix 9

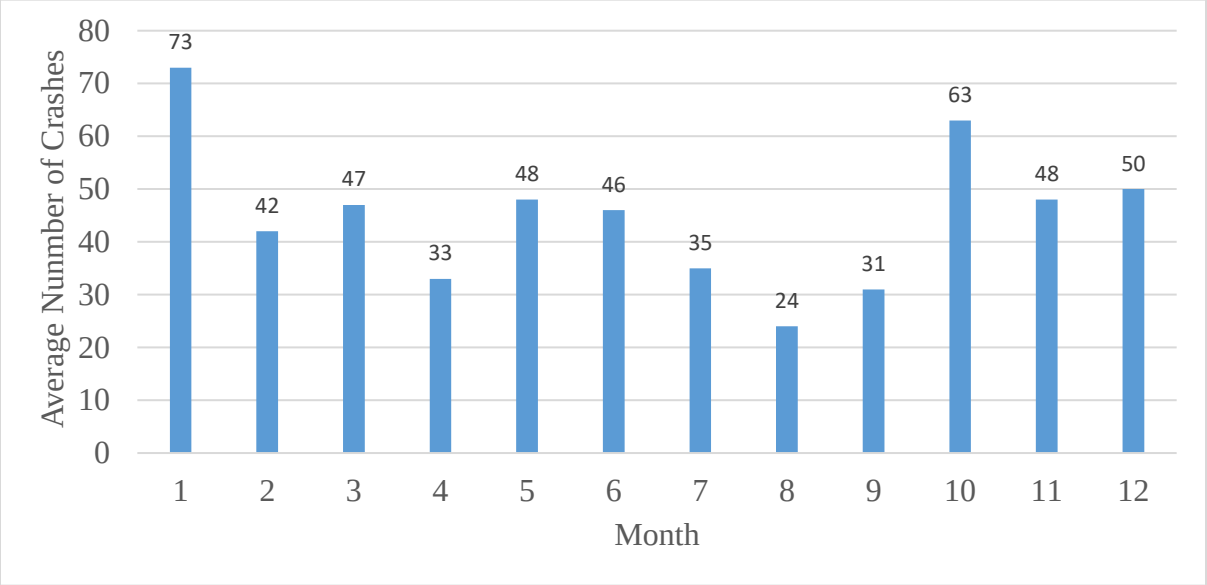
Cleaned Crash Data

Crash ID	Time	Latitude	Longitude	Primary Road	Direction	Intersecting Road
9606569	1/18/2016 22:20	42.52	-83.66	I-96	West	KENT LAKE
9606570	1/18/2016 23:25	42.52	-83.66	I-96	West	KENT LAKE
9621026	2/9/2016 8:00	42.52	-83.66	I 96	West	KENT LAKE
9642465	3/5/2016 4:45	42.52	-83.66	I 96	West	KENT LAKE
9659681	3/5/2016 8:51	42.52	-83.66	I96	West	KENT LAKE
9680404	4/21/2016 7:30	42.52	-83.66	E I 96	West	RAMP

Note. The table shows an excerpt of the traffic data, after cleaning up the direction information of the crashes, and removing unnecessary information columns, such as county and city of the crash locations.

Appendix 10

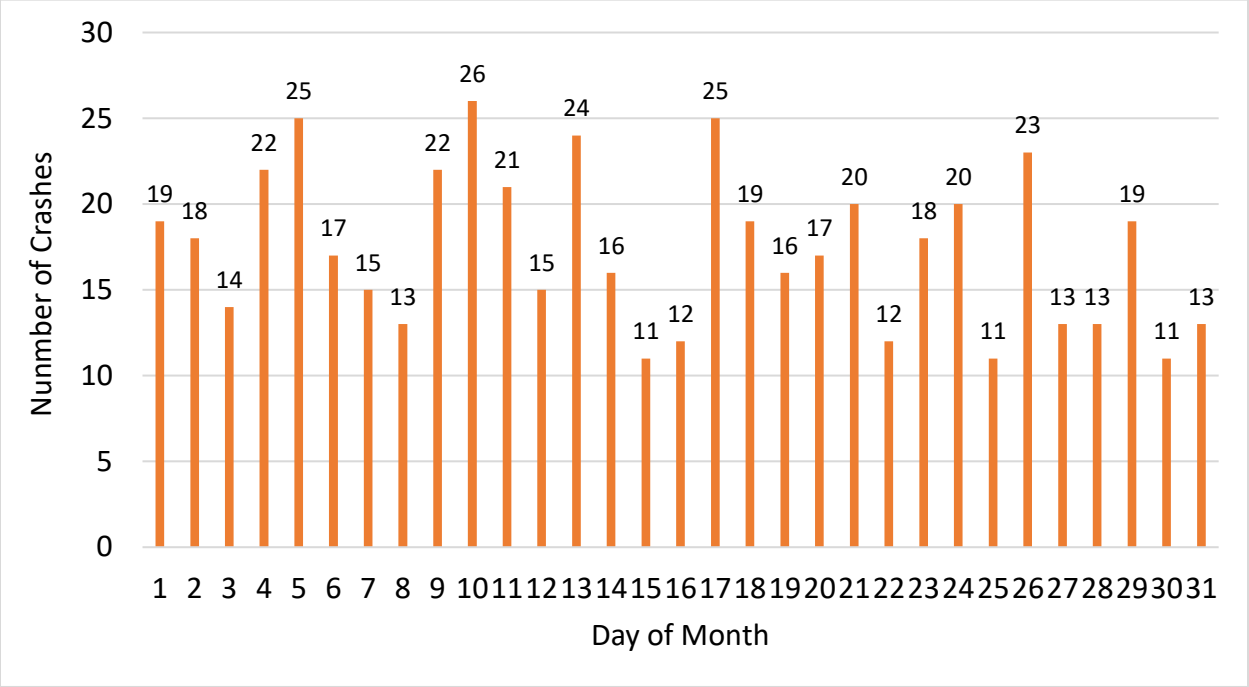
Number of Crashes by Month



Note. The table shows the number of crashes broken down by month of year. It can be seen that the months from October through January have higher number of crashes than the rest of months.

Appendix 11

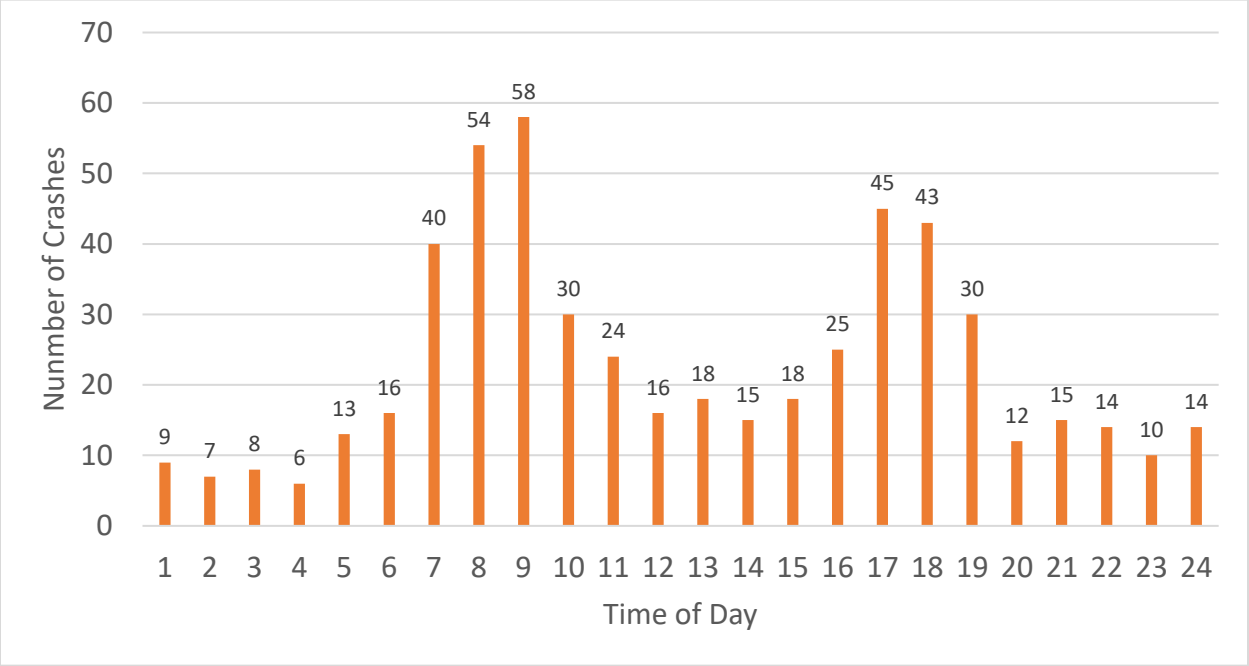
Number of Crashes by Day of Month



Note. The table shows the number of crashes broken down by day of month. It can be seen that the number of crashes has a pattern that peaks in around every 7 days.

Appendix 12

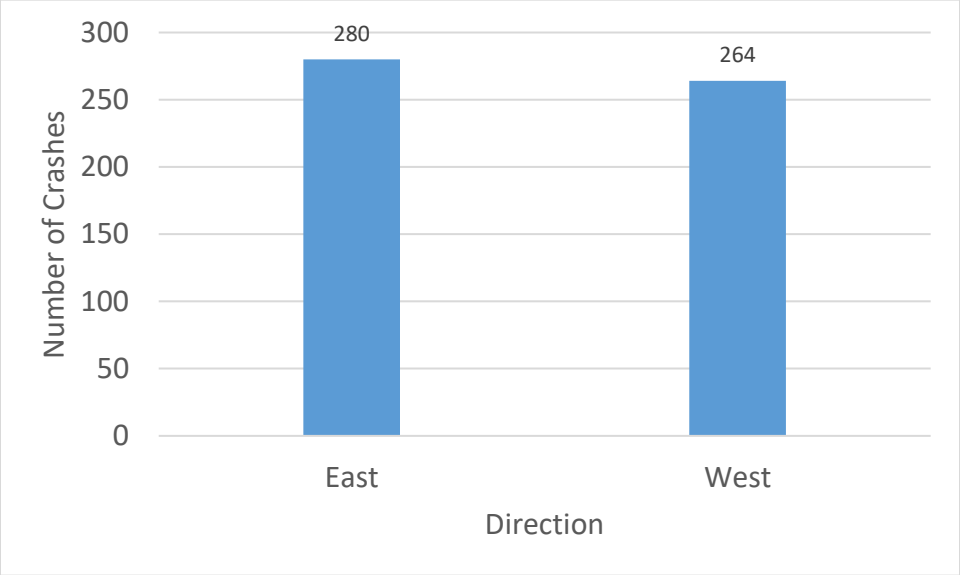
Number of Crashes by Time of Day



Note. The table shows the number of crashes broken down by time of day. It can be seen that the number of crashes has a morning peak between 7 AM and 10 AM, and a afternoon peak between 4 PM and 7 PM.

Appendix 13

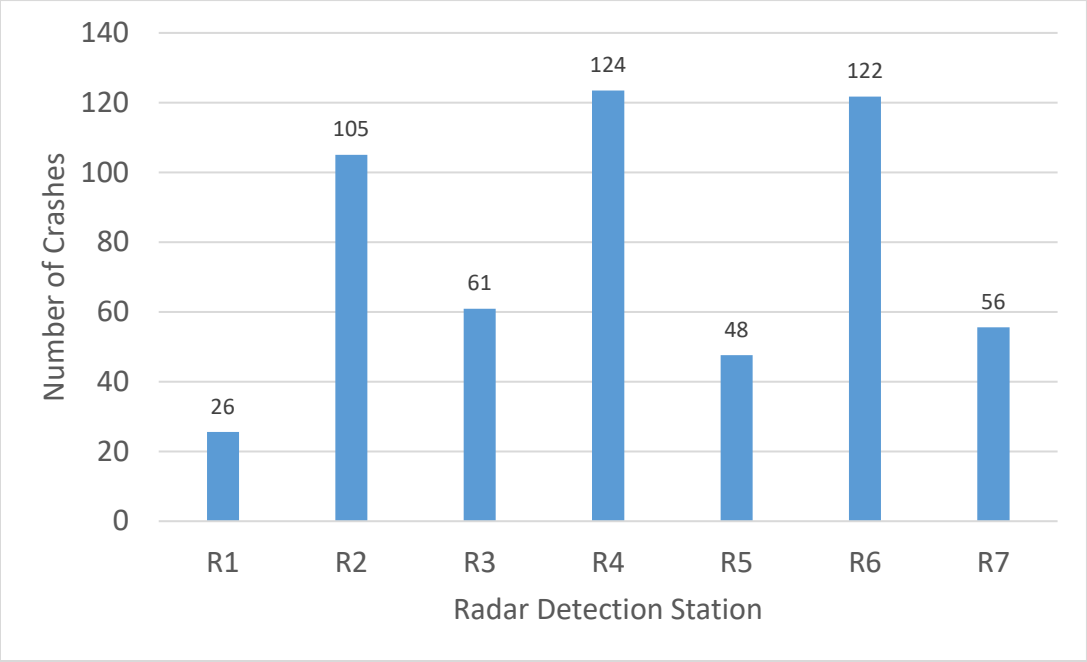
Number of Crashes by Traffic Direction



Note. The table shows the number of crashes broken down by two traffic directions: Eastbound and Westbound on the highway. It can be seen that the number of crashes are close between the two traffic directions.

Appendix 14

Number of Crashes by Radar Detection Station



Note. The table shows the number of crashes broken down by radar detection stations. It can be seen that the station of R2, R4 and R6 have significantly higher number of crashes than the stations of R1, R3 R5, and R7.

Appendix 15

Weather Conditions Category

	Weather Category	
	Good Weather (0)	Bad Weather (1)
Weather Conditions	Overcast Partly Cloudy Clear Possible Drizzle Mostly Cloudy Humid and Mostly Cloudy Possible Light Rain Cloudy	Rain Fog Blowing Snow Snow Sleet / Hail

Note. The weather conditions are classified into two categories: good weather and bad weather.

Appendix 16

Confusion Matrix of the SVM Models

(a) Linear Kernel

		Actual	
		0(noncrash)	1(crash)
Prediction	0(noncrash)	164	2
	1(crash)	40	23

(b) RBF Kernel

		Actual	
		0(noncrash)	1(crash)
Prediction	0(noncrash)	164	2
	1(crash)	36	27

(c) Linear Kernel

		Actual	
		0(noncrash)	1(crash)
Prediction	0(noncrash)	143	23
	1(crash)	24	39

Note. These are the results of model testing data.

Appendix 17

Confusion Matrix of the Decision Model

		Actual	
		0(noncrash)	1(crash)
Prediction	0(noncrash)	158	8
	1(crash)	21	42

Note. These are the results of model testing data.

Appendix 18

Confusion Matrix of the Random Forest Model

		Actual	
		0(noncrash)	1(crash)
Prediction	0(noncrash)	164	2
	1(crash)	30	33

Note. These are the results of model testing data.

Appendix 19

Confusion Matrix of the DNN Model

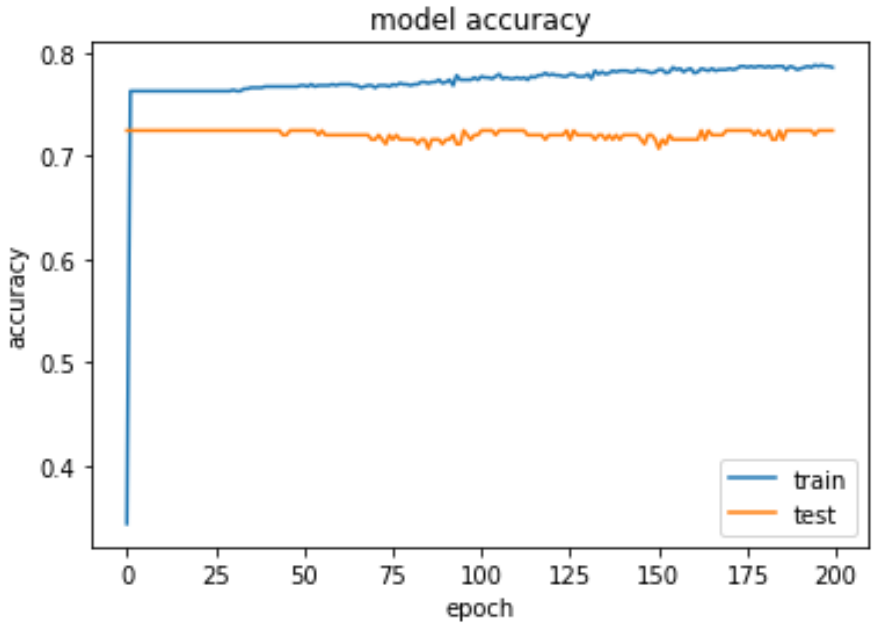
		Actual	
		0(noncrash)	1(crash)
Prediction	0(noncrash)	166	0
	1(crash)	63	0

Note. These are the results of model testing data.

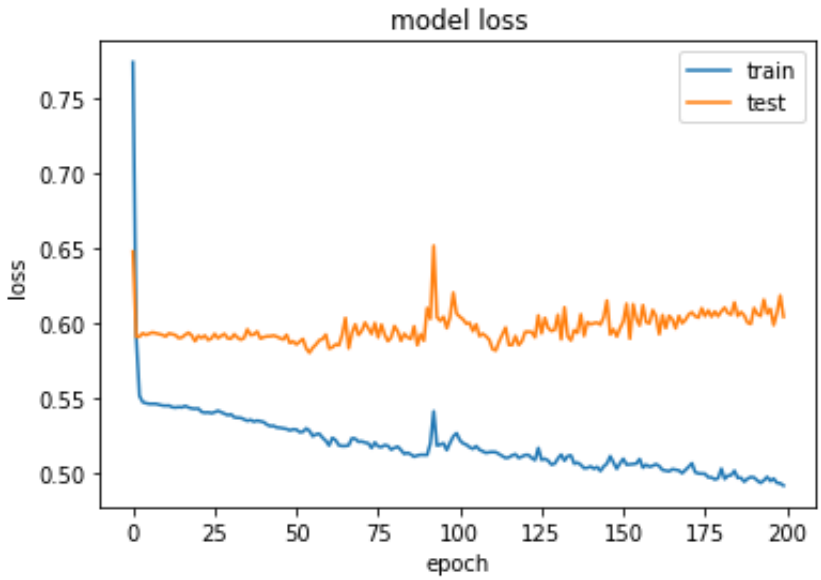
Appendix 20

Learning Curve of DNN model

(a) Model Accuracy for 200 Iterations



(b) Model Loss for 200 Iterations



Note. The gaps between training and testing data accuracy and loss indicate overfit of the model