Scientific Research Publishing

Journal of Transportation Technologies, 2025, 15(2), 222-251
https://www.scirp.org/journal/jtts
ISSN Online: 2160-0481
ISSN Print: 2160-0473

Forecasting Road Incident Duration Using Machine Learning Framework

Smrithi Ajit[1], Varsha R. Mouli[2], Skylar Knickerbocker[2], Jonathan S. Wood[2*]

[1]College of Nursing, Michigan State University, East Lansing, MI, USA
[2]Department of Civil, Construction, and Environmental Engineering, Iowa State University, Ames, IA, USA
Email: *jwood2@iastate.edu

How to cite this paper: Ajit, S., Mouli, V.R., Knickerbocker, S. and Wood, J.S. (2025) Forecasting Road Incident Duration Using Machine Learning Framework. *Journal of Transportation Technologies*, **15**, 222-251.
https://doi.org/10.4236/jtts.2025.152012

Received: January 27, 2025
Accepted: March 11, 2025
Published: March 14, 2025

Abstract

Traffic congestion caused by nonrecurring incidents such as vehicle crashes and debris is a key issue for Traffic Management Centers (TMCs). Clearing incidents in a timely manner is essential to improve safety and reduce delays and emissions for the traveling public. However, TMCs and other responders face a challenge in predicting the duration of incidents (until the roadway is clear), making decisions about what resources to deploy is difficult. To address this problem, this research developed an analytical framework and end-to-end machine learning solution to predict the duration of the incident based on the information available as soon as an incident report is received. Quality predictions of incident duration can help TMCs and other responders take a proactive approach in deploying responder services such as tow trucks, and maintenance crews, or activating alternative routes. The predictions use a combination of classification and regression machine learning modules. The performance of the developed solution has been evaluated based on the Mean Absolute Error (MAE), or deviation from the actual incident duration as well as Area Under the Curve (AUC) and Mean Absolute Percentage Error (MAPE). The results showed that the framework significantly improved the prediction of incident duration compared to previous research methods.

Keywords

Traffic Incident Management, Model Blending, Model Selection, Decision Making, Transportation Forecasting

1. Introduction

According to the Traffic Incident Management Handbook, an incident is defined as a non-recurring event that results in a reduction in the capacity of the roadway

or an abnormal increase in demand [1]. These incidents include, but are not limited to, vehicle crashes, disabled vehicles, debris, and spilled cargo. Incidents not only result in traveler delay but also increase the likelihood of secondary crashes and other secondary effects due to increased opportunities for secondary events to occur [2]. Secondary events can lead to increased demand for police, fire, and emergency services, reduced air quality, and other environmental impacts.

Total incident duration is comprised of incident notification time, response time, and clearance time, as illustrated in **Figure 1**. As shown, the total incident duration is the total time from the start of the incident until the reported time of clearance for the event [3] and includes incident notification, response, and clearance times. The incident notification time is from the start of the incident until the time it is reported. The response time is from the report time until the response unit's arrival. The clearance time is the time taken to clear the incident after emergency responders have arrived on the scene.

While the incident duration time is not controlled by the responding agencies, dispatching the correct personnel and equipment can reduce the total incident duration by minimizing the sum total of response and clearance times. This requires planning, preparedness, and coordination between responders. Information on what resources should be dispatched can be improved using predictive models of the total incident duration. For instance, having accurate predictions of the total incident duration can assist Traffic Management Center (TMC) operators in selecting the appropriate actions from potential options such as the following: (1) diverting traffic to an alternate route, (2) providing a warning of a potential delay to travelers planning to take a congested route and (3) ensuring helper services, such as safety service patrol [4] or maintenance crews, arrive at the incident spot on time. For example, if an incident will be cleared within a half hour, it may not be reasonable to detour traffic on a route that increases the travel time over that amount of time. Helper services may also not be requested if the incident is cleared before the time it takes the service patrol to arrive at the incident location.

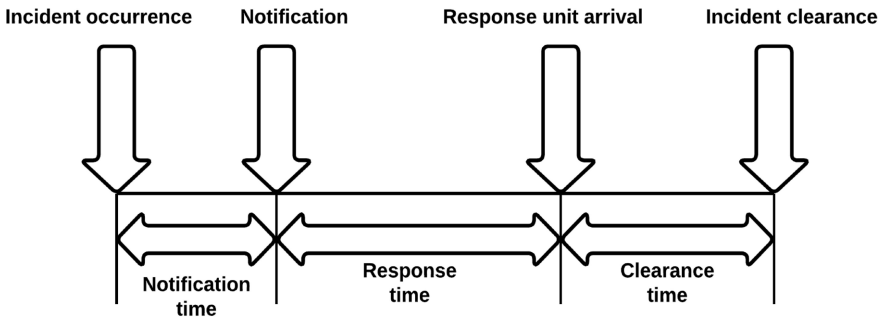


Figure 1. Illustration of traffic incident duration.

As indicated above, incident duration prediction is critically important for TMC for timely mitigation of traffic congestion, not only to forewarn people of crashes on a particular route in advance but also to reduce the likelihood of sec-

ondary crashes. Past studies have focused on understanding and identifying factors related to incident type, roadway data, time of the day, weather conditions, speed, traffic volume, blocked lane, location, environment, weather, road characteristics, temporal and spatial factors, and so on, either through associative mining or through prediction-based descriptive models that are either statistical or machine learning-based models.

1.1. Literature Review

Early work in the area of incident duration prediction used linear regression models and related statistical tests such as Analysis of Variance (ANOVA). Many of these models were limited by the number of data points [5]-[7]. The coefficient of determination, root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) were the performance metrics commonly used to report the efficiency of the ML models. The lowest MAPE that was reported with these models include that by [8], who reported 34.1% for a total of 2,512 accidents using a cluster-based log-normal distribution model and a MAPE of 37% reported by Khattak *et al.* for a total of 59,804 accidents.

Other methods that have been used in the published literature include fuzzy logic, Artificial Neural Networks (ANN), Bayesian methods, Survival or Hazard models, tree-based machine learning methods, and text mining. Among fuzzy logic-based approaches, the lowest MAPE was 36% with an average error of 0.3 minutes reported by the authors in [9] and [10], respectively. However, these studies were also limited by fewer data points. Methods using ANN have also been attempted on small and large datasets, with Pereira *et al.* [11] achieving a median error of 9.9 minutes for a dataset of size 10,139 and Lopes *et al.* [12] achieving within 10 to 20 minutes error for a dataset of 10,762 incidents.

Bayesian models are useful in applying prior knowledge of associations between variables in developing predictive models. Several authors have developed the incident duration prediction problem into a classification problem and have used Bayesian approaches for predictions achieving accuracy ranging between 74% to 80%. The authors of [13] used the Bayesian ANN model for a large dataset achieving a low MAPE of 0.18 to 0.29. Other Bayesian approaches that have been reported include Bayesian propagation neural networks and Bayesian support vector regression [14].

Survival models or Hazard based models are used to predict the length of the duration between the occurrence of the event, and its clearance while also being capable of predicting when the incident duration will end given it has proceeded for time t. This ability of the models helps in the prediction of incident duration at different stages of the incident as presented in [15] using an AFT hazard-based model. Random parameter versions of these models have been suggested [16] [17], yet predictions using these are not straightforward. Fully parametric Hazard based models using log-normal and log-logistic distributions were used in [18] and reported a MAPE value of 47%. A logistic Accelerated Failure Time model

(AFT) has been presented in [19] that reported a MAPE of 43.7%, and another Weibull distribution parametric model using gamma heterogeneity was used in [20] to account for unobserved heterogeneity. In their exploration of proactive crash prediction models the authors in [21] developed a duration-based modeling framework using a nested logit model capable of predicting the occurrence of traffic crashes and their severity. The model incorporates static and dynamic covariates demonstrating enhanced predictive capabilities in real-time. In real-time, as time progresses, the traffic conditions change, and the new set of factors influence the incident duration, due to which time sequential models are useful. Incorporating additional features as the incident progresses can be very helpful in traffic decision-making. The inverse Gaussian frailty AFT model, Multilevel mixed-effect AFT model, and FMAM models were used to predict the incident duration with sequential TIM information in [22]. Multiple studies also identified that a single method could not suit all the incident duration ranges and came up with hybrid approaches that combine two or more methods. In the study [23], Hazard Based Duration Models (HBDMs) were used as leaves of the M5P Tree, and in the work of [24], fuzzy entropy was used to select the features, and ANN was used to predict the duration based on these features. A competing mixture model was presented in [25] to analyze the influence of the clearance method and other covariates on traffic incidents.

Tree-based models have also received significant attention in the published literature for incident duration. Results from the literature have reported MAPE values ranging between 42.7% to 65% [26]-[28]. A more recent focus has been on ensemble techniques like RF, Gradient Boost method, Extreme Gradient Boost Method, and Ada Boost owing to the superiority of performance with datasets with a large number of categorical variables [29]-[31]. In [30], RF models developed for long and short-duration datasets reported an MAE of 36.652 minutes and 14.97 minutes, respectively.

Besides the approaches discussed above, text-based analysis was also used. In [11], the authors illustrated an improvement of more than 35% over non-textual models, and a knowledge-based method was used in [32] to estimate incident clearance duration on Maryland I-95. The authors of [33] utilized a combination of text mining and ensemble learning techniques to predict traffic accident durations, demonstrating that integrating unstructured and structured data significantly enhances prediction accuracy. Similarly, the authors in [34] implemented a heterogeneous deep learning framework that leverages both structured and unstructured text data to make accurate predictions of accident duration. In [35], a spatiotemporal feature learning model called TITAN was proposed that considers hidden spatiotemporal associations by considering connectivity between road segments in addition to identifying high-level features. In [36], a copula-based trivariate framework was used to generate a stochastic dependence relationship between the various variables capable of predicting incident duration.

In the context of ML models, since no prior assumptions are made about the

distribution of the data, the ML models will be able to identify underlying relationships only if data challenges such as skewness and heterogeneous nature of the data are adjusted prior to training the ML models. The heterogeneous nature of the data is handled through the use of feature selection methods that help restrict the variables to a few of the most significant ones. In order to group similar data points, a general approach is to use supervised or unsupervised clustering methods. Supervised clustering of data points followed by the application of regression models can be seen in [37] while an un-supervised approach can be seen in [29]. A supervised approach that makes use of the blending of results from multiple ML models for the prediction of incident duration has not been applied in the published literature.

Ensemble learning is a method which combines multiple base learners or base models that can be classifiers or regressors which generates predictions that are learned and integrated into a single prediction by the next layer also called the meta-classifier or meta-regressor [38] [39]. Both homogenous and heterogenous ensemble methods exist with the former including methods like bagging and boosting and the latter including stacking [40] [41] and blending. Ensemble models are better suited to predict the outcome when there are complex patterns and interactions in the data.

1.2. Objectives

The general objective of this study is to develop a comprehensive incident duration prediction framework to support TMC operations using machine learning in a two-step process. Initially when an incident is reported only basic features like date and time, injuries, location and type of incident are reported. The first step of the developed process is simple and classifies the incident into one of the three incident duration zones—within a half hour (class 1), between half an hour and 2 hours (class 2), and more than 2 hours (class 3) based on [11]. The advantage of this initial prediction is that it is simple and provides a rough estimate that enables prompt decision-making, such as resource allocation or emergency response planning. The second step of the developed process involves making a refined prediction of the incident duration in minutes when more features are available from the RAMS. The availability of more features helps capture the nuances of the incident duration more accurately by helping develop a more sophisticated regression model that is able to make a more accurate prediction of the total incident duration. By providing both quick estimates and more accurate predictions over time, the two-step approach helps in optimizing resource allocation. Emergency response teams can use the initial predictions for immediate action, while more detailed predictions can inform longer-term planning and resource allocation. The two-step approach also allows for continuous learning and model improvement. As more data becomes available and more features are collected, the models can be updated and refined, leading to improved predictions over time. The ensemble techniques of blending and stacking have been have been leveraged to en-

hance the performance of base models. By harnessing the collective insights of multiple models, blending and stacking amalgamate diverse perspectives captured by individual models, thereby refining predictions and enhancing accuracy. These ensemble methods not only mitigate overfitting but also excel in capturing intricate patterns within the data, contributing to more robust and reliable predictions. The performance of the developed supervised ML framework has also been validated against an alternative framework, an unsupervised ML framework.

A few applications of the developed framework have been provided in the Discussion section. Additionally, to understand if the classification step improves predictions, the Mean Absolute Error (MAE) results from the developed framework have been compared with MAE results obtained from a framework that predicts without classifying the data into ranges.

2. Data and Methodology

The Advanced Traffic Management System (ATMS) events data is the primary source of information used in this paper. It is maintained by the Iowa Department of Transportation Traffic Management Center (TMC) and contains detailed records of traffic incidents that occurred in the state. The data are collected by TMC operators that actively manage incidents on roads and include information such as the number of lanes closed, start and end times, the presence of emergency responders, and the severity of the incident. This information is useful for creating after-action reports, which allow the Iowa DOT to review and analyze their response to incidents in order to identify areas for improvement. ATMS data are made available on a daily basis and include all incidents that have been closed the previous day [41]. The data used in this study was collected from 2017 to 2019.

2.1. Data

The Advanced Traffic Management System (ATMS) data used in this study include recurring and nonrecurring events, such as work zones, collisions, debris, stopped vehicles, and special events. However, the study focuses specifically on collisions involving 1, 2, or 3 vehicles and debris, as these types of incidents can have a significant impact on traffic flow and are within the control of TMC operators. The study also limits the analysis to incidents that occurred on rural and municipal interstates within the state of Iowa and excludes days with severe weather events as these can significantly impact incident duration. Concerning duration all incidents that recorded 0 minutes were removed along with those having duration greater than 1.5 time the inter-quartile were removed. To identify severe weather days, the study used the National Oceanic and Atmospheric Administration's (NOAA) storm database and excluded incidents that occurred in affected counties. After removing incomplete records, the study included a total of 5884 incidents that occurred over a three-year period.

To address the correlation between variables in the data, the study removed features that had a high correlation value (greater than 0.4) and used the pycaret

tool to further address this issue. This helps to ensure that the model is not influenced by correlated features, which can impact the accuracy of the model. Table 1 presents pairs of variables that exhibit strong correlation. To maintain the integrity of the analysis and mitigate the effects of multi-collinearity, only one variable from each pair while excluding the counterpart. The data was split into a train and test subset on which the machine learning models were trained and tested and a validation or holdout dataset which the models have never seen and represented 5% pf the data was used to test the efficiency and robustness of the built models. To ensure that the training dataset and validation dataset are similar and that the split of data into these subsets does not introduce systemic bias, a Levene's test for equality of variances and an independent t-test test were conducted, the results of which are summarized in Table 2 and Table 3, respectively. Levene's test results show a p-value greater than 0.05, suggesting that the variances of the two subsets are more or less equal. The t-values are close to zero at 0.973 and 0.986 (p-values of 0.330 and 0.325) for equal and not equal assumed variances at an alpha level of 0.05, indicating that there is statistically no significant difference in mean values between the two groups.

Table 1. Correlation coefficients between variable pairs.

Variable 1	Variable 2	Correlation
Severity	Total	0.817
Injuries	EMS	0.467
Injuries	Fire	0.402
Terrain	Road municipal interstate	−0.729
Terrain	Road rural interstate	0.729
Surface width	Road municipal interstate	0.411
Surface width	Road rural interstate	−0.411
Surface width	AADTnew 5	0.580
EMS	Fire	0.779
County central	County southeast	−0.626
County central	County southwest	−0.443
County southwest	AADTnew 3	0.531
Detection cameras	Detection police	−0.432
Detection HH	Detection police	−0.440
Season autumn	Season winter	−0.408
Dir E	Dir W	−0.402
Surface type 65	Surface type 74	−0.458
Road municipal interstate	Road rural interstate	−1.000
AADTnew 4	AADTnew 5	−0.672

Table 2. Levene's test for equality of variances.

Hypothesis	F	P-value
Equal variances assumed	0.543	0.461

Table 3. Comparison of means of train-test subset and validation subset. (Independent t-test)

Hypothesis	t	df	One-sided p	Two-sided p	Mean difference	Std error difference	95% CI (lower)	95% CI (upper)
Equal variances assumed	0.973	5882	0.165	0.330	2.131	2.189	−2.160	6.421
Equal variances not assumed	0.986	728.147	0.162	0.325	2.131	2.161	−2.113	6.374

For the analysis, the dependent variable will be the duration of each incident, calculated as the time between when the incident started and when the roadway was cleared. The incident duration can be separated into three groups within 30 minutes (short), between 30 minutes and 2 hours (medium) and more than 2 hours (long). These groups were tested for their variance and the results of the ANOVA test can be seen in Table 4. From the results it is evident that a high value of F and p-value of 0 indicate significant differences in the mean responses of the three duration categories. Since F-test is highly significant, we can reject the null hypothesis that there is no difference in the means across the categories. This understanding is beneficial in designing and improving machine learning models since they can train and understand the specific factors that most influence a particular category of incident duration thus tailoring the model parameters to better fit the specific data distribution and pattern within each of these categories. A post hoc analysis using Tukey's Honest Significant Difference (HSD) test was done and summarised in Table 5 in order to make multiple pairwise comparison of the data and helps control TypeI error [2] [42]. When comparing short versus medium-duration events, it is found that the medium incidents are 47.3 minutes longer than short-duration events on average. The 95% confidence interval indicates that the true mean difference is found to fall between 45.76 and 48.83 and the finding is statistically significant. When comparing short and long-duration events long incidents are on average 176.92 minutes longer than short-duration events and the 95% confidence interval falls between 173.81 and 180.02 minutes. Finally, when comparing medium and long duration events, the mean duration events are found to be 129.62 minutes longer than medium incidents and the 95% confidence interval falls between 126.48 and 132.76 minutes. All these findings were statistically significant indicating that there are statistically significant differences between the three categories of events.

Table 4. ANOVA results for incident time duration.

Source	Sum of squares	df	F-value	Pr (>F)
C (Actual class)	1.148×10^7	2	9852.797	<0.001
Residual	3.428×10^6	5881	NA	NA

The underlying behavior of each of these categories can also be understood

from the feature importance graphs provided in **Figure 2.** It is evident that a different set of factors significantly influence the total duration in each category. The prominence of factors like hourly traffic volume, event type, and day of the week reflects their practical implications in traffic incident management. For example, the high importance of hourly traffic volume across all durations implies that higher traffic density can both expedite incident detection due to more witnesses and delay clearance due to congestion. This duality demonstrates the complex role traffic volume plays in incident dynamics. The significance of event type indicates that the nature of the incident (e.g., crash vs. debris) directly affects how resources are allocated and the strategies used for clearance, impacting the duration. The day of the week factor suggests that traffic patterns, which vary between weekdays and weekends, influence incident management, possibly due to differing traffic volumes or available response services. In medium and long durations, the importance of the time of day, specifically afternoons and nights, reflects practical challenges such as reduced visibility at night or peak traffic volumes in the afternoon, which can complicate the response efforts and prolong the clearance time. Similarly, environmental conditions like road surface types and seasonal influences underline how physical and weather-related factors can complicate incident handling, especially in regions prone to adverse conditions. Thus, understanding these factors helps in strategizing more effective incident management approaches tailored to specific circumstances.

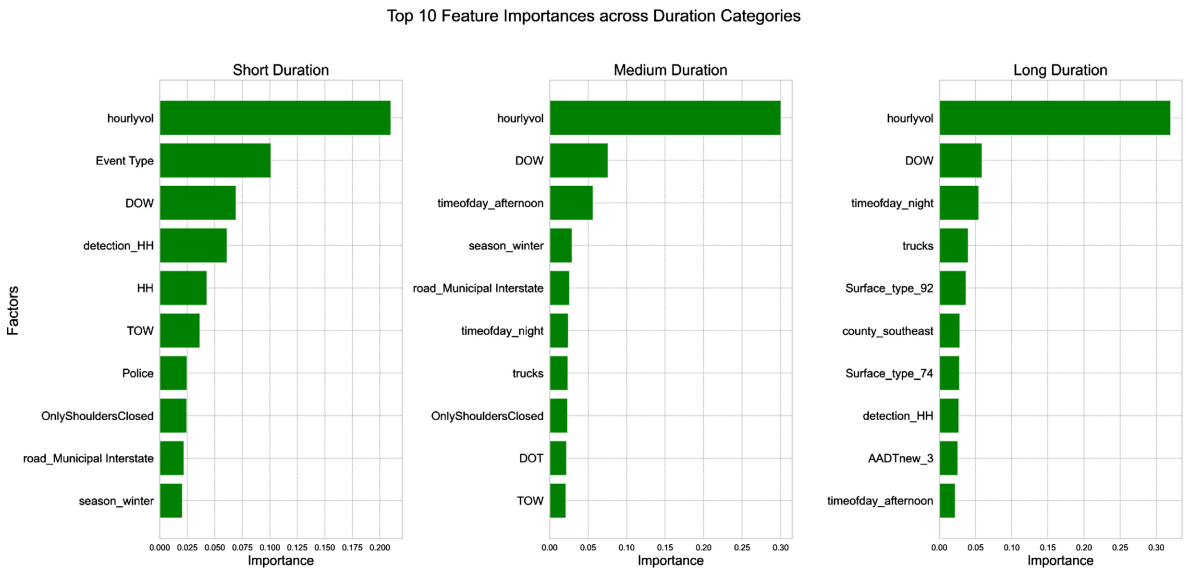


Figure 2. Feature importance in the three duration categories.

Table 5. Tukey HSD post hoc comparison of means for incident duration categories.

G1	G2	Mean Diff.	Adj. P-Val	Lower CI	Upper CI	Reject H_0
1	2	47.3	<0.001	45.76	48.83	Yes
1	3	176.92	<0.001	173.81	180.02	Yes
2	3	129.62	<0.001	126.48	132.76	Yes

The dataset includes several variables, which are listed in Table 6, along with their data types and any transformations that were applied. Columns with text data, such as information about responders, were converted to categorical variables. Binary variables were used to encode information about the presence of specific responders (e.g., police, firefighters, etc.) at the incident, with a value of 1 indicating presence and 0 indicating absence. This responder information is described in more detail in Table 7, which also includes statistics about the duration of incidents for each type of responder.

Table 6. Explanation of variables in the dataset, their datatypes and transformations.

Feature group	Variable name	Derived	Datatype	Categories
Lane information	Number of lanes(lanes)	N	Integer	NA
	Only shoulders closed	Y	Binary	1-responded, 0-did not respond
Location information	Direction	N	Binary	1: North, 2: South, 3: West, 4: East, North East, North West, Central, South East, South West
	County region city number	Y N	Factor Factor	Numerical binning
Incident	Event type	Y	Factor	Crash (1vehicle, 2 vehicle and 3 vehicle crash), Debris
Vehicles involved	Vehicles	N	Factor	0,1,2 and more than 3
	Trucks	N	Factor	0,1,2 and more than 3
Severity	Injuries	N	Factor	injured, 0-not injured
	Fatalities	N	Factor	or more fatality, 0- no fatality
Responder information	Responder type: Police, TOW, DOT, DPS	Y	Factor	1-responded, 0-did not respond
Detection	Detection method	N	Factor	Police, Highway Helper, Automated, DOT, Cameras, Others
Temporal information	Time of day (TOD)	Y	Factor	Morning: 7am to 9am, Early Afternoon: 10am to 12pm, Afternoon: 1pm to 3pm, Evening rush:4 pm to 6pm, Evening: 7pm to 9pm, Night: 10pm to 6 am Integer encoding 0 to 6 represent Monday to Sunday
	Day of week (DOW)	Y	Factor	
	Season		Integer	1 = Winter (December to February), 2 = Spring (March to May), 3 = Summer (June to August), 4 = Autumn (September to November)
	12 year	Y	Integer	2017, 2018, 2019
Traffic information	Hourly traffic volume	Y	Integer	NA
	AADT	Y	Factor	1: 8,000, 2:8,000 - 12,000, 3:12,000 - 24,000, 4:24,000 - 48,000, 5: 48,000
Road characteristics	Surface width surface type	N Y	Float Factor	NA
	Terrain	Y	Factor	Grade and drained earth, gravel or stone, bituminous over gravel or stone, etc flat, rolly, hilly

Table 7. Responder information.

Responder (presence)	% of total incidents	Incident duration mean (min)	Incident duration median (min)
DOT-No	91%	39.3	26
DOT-Yes	9%	88.7	61

Continued

EMS-No	88%	38.9	23
EMS-Yes	12%	76.2	57
HH-No	51%	40.0	22
HH-Yes	49%	47.7	33
Police-No	67%	33.6	17
Police-Yes	33%	65.5	47
TOW-No	77%	76.3	19
TOW-Yes	23%	23.7	51

Categorical variables that were not ordered were one-hot encoded. To avoid the dummy variable trap in the dataset, one category was excluded from each set of one-hot encoded variables, preventing multicollinearity among the predictors. This adjustment is made to enhance the stability and interpretability of the models. Missing values for numerical variables were imputed with the mean value while missing values for categorical variables were imputed with a constant based on the mode. For the variables vehicles, trucks, injuries, fatalities, and AADT, the variables were converted to categorical variables as described in Table 6. The statistics for the Average Annual Daily Traffic (AADT) information in the dataset are provided in Table 8. It can be seen from these statistics that the majority of incidents have a duration in short to medium range.

Table 8. AADT versus incident duration.

AADT	Avg. total	Median total	No. of records
16,000	47.8	28	153
27,700	28.8	4	131
45,300	54.6	43	151
57,000	28.4	14	266
100,500	30.4	20	194

The basic statistics related to the dataset have been provided in Table 9. These descriptive statistics show the wide variance in roadway clearance times with values ranging from 1 to 542 minutes, along with a positive skew of 3.06 minutes. This is addressed by removing the skew in the data using the Yeo-Johnson power transformation, a default transformation method found in Pycaret module which is more flexible than the box-cox transformation and capable of handling positive and negative values. Applying the best lambda value of 0.123 will reduce the skew of the dataset to −0.0279.

Table 9. Statistical information on the total incident duration.

Metric	Value (minutes)
Standard deviation	53.73
Mean	45.23
Median	31
Minimum	1

Continued

25th percentile	10
75th percentile	59
Maximum	542

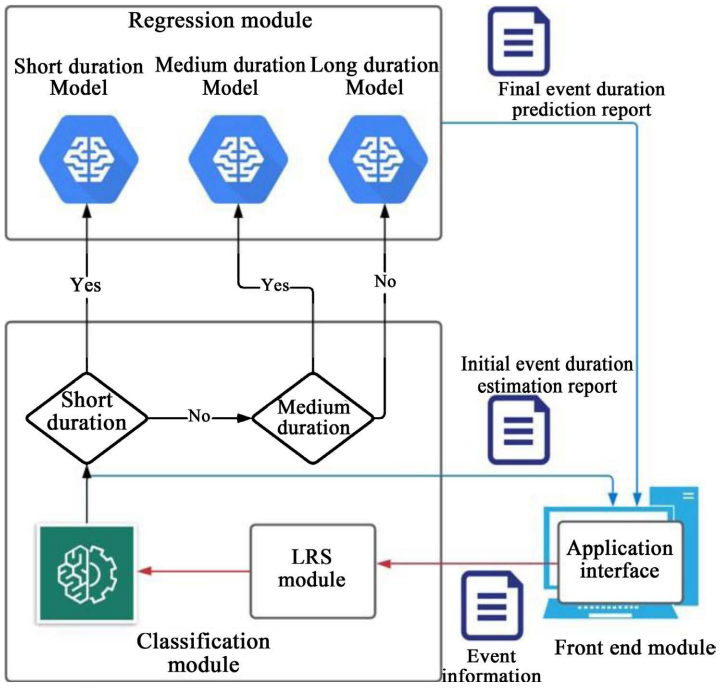


Figure 3. Overall workflow of developed framework when additional features are available.

2.2. Methodology

The framework for estimating event duration shown in Figure 3 can be separated into two core components, including the classification and regression modules. The goal of the first part of the framework in the classification module is to classify an event as being of short, medium, or long duration based on its total incident duration. Events with a duration of less than 30 minutes are classified as short duration, those between 30 minutes and 2 hours are classified as medium duration, and those lasting more than 2 hours are classified as long duration. The pre-classification module is built to direct the data flow toward specific predictive regression models that are custom-built for each of these duration categories. Moreover in the real-world situation knowing the possible duration of each of the reported incidents can help traffic management agencies allocate resources more efficiently.

The initial classification is based on only limited information available about an event, including its location, date, and time of day, since not all details about the event are available when the incident occurs. This minimal information is used to assign the event to one of the three duration categories. After the initial classification, the data is then processed through a linear referencing system (LRS) module, which removes the spatial dependencies (latitude and longitude coordinates) of the data by adding two additional features: a Route ID and a Measure. The Route

ID represents the route on which the event occurred, while the Measure provides the linear distance of the event's location from a reference point. These parameters are used to extract additional features, such as Average Annual Daily Traffic (AADT), terrain, surface width, and surface type, from the Roadway Asset Management System (RAMS).

With the initial classification and additional attributes from the asset management system, the event then moves to the regression module. The regression module runs a regression model that is fine-tuned to predict the exact duration of the event based on the subset to which the data point belongs. Each model is trained on the training data, tested on holdout or test data, and then retrained on both the train and holdout data before saving the model.

To develop a model with the best efficiency, the solution employs a process called blending, which combines the strengths of multiple models to create a single model with hybrid features.

In the classification module, it was found that the Random Forest model, the Extra Tree Classifier model, and the Light GBM model achieved the lowest mean absolute error (MAE). The predictions of these three models are combined by a meta-learner to give a blended final prediction, as shown in **Figure 4**. Similarly, for the regression module, the predictions from the combinations of Random Forest model in combination with Catboost model are fed into meta-learner for short-duration prediction. Predictions from Random Forest model and Huber model are fed into meta-learner for mediumduration prediction. For long-duration event prediction, XG Boost model alone is used since it performed better than blended models. A Tobit model is used as a reference to evaluate the MAE value and determine how the methodology is improving the results [43].

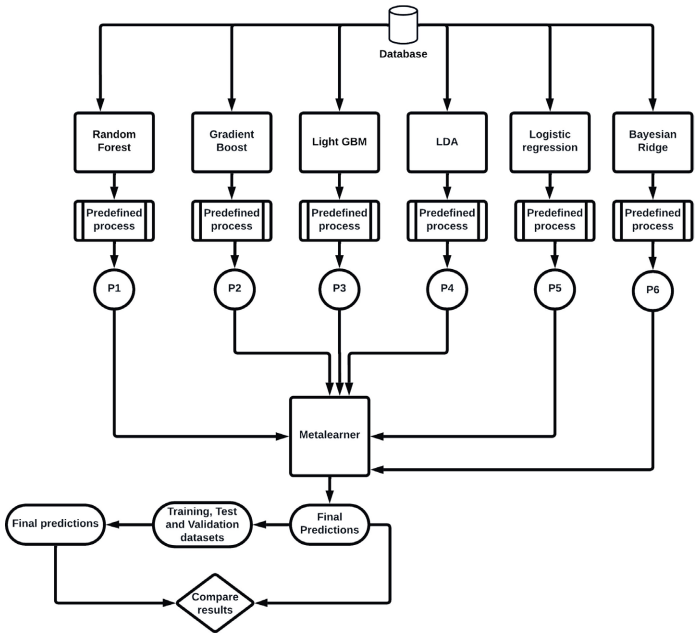


Figure 4. Illustration of the concept of blending.

2.3. Model Evaluation Metrics

One commonly used performance metric for evaluating classification models is the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) plot. The ROC curve plots the true positive rate against the false positive rate at different classification thresholds, and the AUC is calculated by finding the area under the ROC curve. A higher AUC value indicates that the model is better at distinguishing between the classes. In the problem at hand, accurate classification of an event into the correct duration category is crucial because it ensures that the appropriate regression model, which has been specifically trained for that category, is applied.

$$\text{TPR} = \frac{\text{TP}}{\text{FP} + \text{TN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Additionally, precision, accuracy, and recall are useful parameters for evaluating model performance. While accuracy indicates the performance of the model in predicting the correct class or category, precision indicates the ratio of what the model predicted correctly to what it predicted overall, and recall indicates what the model predicted correctly to what the true classifications are. Along with high AUC values, the best possible value for recall is also desired because the model needs to be sensitive toward the higher severity incidents that are associated with longer durations. In the case of regression models, commonly used evaluation metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

$$\text{MAE} = \left(\frac{1}{n}\right)_{i=1}^{n} \left|t_{pi} - t_{oi}\right|$$

2.4. Model Description

The best-performing machine learning model for the classification and regression modules was selected based on the highest AUC value for the classification module and the lowest mean absolute error (MAE) for the regression module. Tree-based models generally performed better. To further improve model performance, ensemble methods of stacking and blending were applied. The following section provides a general overview of the top-performing models used in the classification and regression modules.

Random Forest Classifier and Regressor

Random Forest is a robust ensemble learning method that leverages the concept

of bootstrap aggregation, or bagging, to train multiple decision trees on various subsamples of the dataset, each selected with replacement. The final prediction of the model integrates the outputs from all these individual trees. Two critical hyperparameters that significantly influence the efficacy of a Random Forest model are max depth and n estimators. The max depth parameter controls the maximum depth of each tree, affecting the complexity and the level of detail the trees can capture. On the other hand, n estimators define the total number of trees constructed in the forest. Fine-tuning these parameters can markedly enhance the model's accuracy and overall performance.

One of the primary objectives of Random Forest is to mitigate variance. Each tree in the ensemble is independently constructed using a random subset of the dataset, which helps to average out the high variance associated with individual trees, thereby enhancing the stability and accuracy of the model. Random Forest aggregates the predictions from multiple decision trees to form a final output, whether classifying objects into categories or predicting continuous values. Adjusting the depth of the trees and the number of trees in the ensemble allows for a balance between complexity and computational efficiency, optimizing the performance for a variety of tasks.

2.5. Gradient Boost Classifier and Regressor

The Gradient Boosting algorithm constructs a robust predictive model by integrating multiple simpler models. It progressively adds predictors to the ensemble, with each new predictor aiming to amend the errors of its predecessor. Specifically, each additional tree focuses on the residual errors left by earlier trees, thereby refining the overall predictions. This method is highly effective in handling various feature types and is designed to prevent overfitting.

The Gradient Boosting regressor operates similarly to the classifier. Initially, the model predicts the mean of the target values, and subsequent trees are added to predict and minimize residuals. Unlike the Random Forest model, which employs bagging to generate an ensemble, Gradient Boosting uses a boosting approach. In this approach, trees are constructed sequentially, and each new tree is tasked with correcting errors made by the trees that preceded it.

Light Gradient Boost Classifier and Regressor (Light GBM)

Light Gradient Boosting Machine (Light GBM) is an advanced implementation of gradient boosting algorithms, specifically optimized for efficiency and performance. Unlike traditional gradient boosting techniques that build trees level-wise, Light GBM constructs trees leaf-wise. This distinctive strategy allows for faster learning and better adaptation to data, especially on large datasets.

The fundamental operation of Light GBM can be described by the update rule where the model iteratively enhances its predictions. At each step, the algorithm selects the leaf that minimizes the loss when a new tree is added

Key advantages of Light GBM include its high computational speed and lower

memory usage compared to other similar methods. It efficiently handles large-scale data and focuses on achieving high accuracy in predictive tasks. This efficiency is partly due to its leaf-wise tree growth and effective handling of sparse data, which reduces both the number of splits required and the overall computational complexity.

Additionally, Light GBM incorporates several innovative techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which further enhance its performance and scalability. These features make Light GBM particularly well-suited for scenarios requiring rapid processing of extensive data without compromising on model accuracy.

2.6. Logistic Regression

Logistic Regression is a widely-used statistical method for binary classification. It models the probability of a binary response based on one or more predictor variables. The probability that an outcome belongs to a particular class is modeled as a function of the predictors, using the logistic function, which is expressed mathematically as:

$$p\left(y=1|\vec{x}\right)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)'}}$$

where $\vec{x}=(x_1,\cdots,x_n)$ is the vector of predictor variables, $\beta_0,\beta_1,\cdots,\beta_n$ are the coefficients of the model, and $p\left(y=1|\vec{x}\right)$ is the probability that the target variable y is 1 given $\vec{x}$. This model uses the logistic (or sigmoid) function to ensure that the output lies between 0 and 1, thus interpreting it as a probability.

2.7. Bayesian Ridge Classifier

Bayesian Ridge Classifier is an adaptation of ridge regression within a probabilistic framework that incorporates Bayesian inference for parameter estimation. It assumes a prior distribution over the coefficients, usually Gaussian, and updates this prior in light of the observed data to produce a posterior distribution.

2.8. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) aims to project data onto a lower dimensional space while maintaining excellent class separability to mitigate overfitting and reduce computational costs. LDA achieves this by identifying axes, known as linear discriminants, which maximize the separation between classes.

LDA is highly efficient and effective under the assumptions of normally distributed classes and equal covariance matrices, although it remains vulnerable to outliers and deviations from these assumptions.

2.9. Ensemble Techniques: Stacking and Blending

Stacking, also known as stacked generalization, is an advanced ensemble technique that combines multiple classification or regression models through a meta-

classifier or meta-regressor. In stacking, individual models, referred to as base learners, are first trained on the full training dataset. The predictions from these base learners are then used as inputs for a second-level model, known as the meta-learner. The meta-learner aims to synthesize the predictions of the base models to improve the overall predictive accuracy.

This model structure allows stacking to exploit the strengths of each base model, potentially leading to superior performance compared to any single model alone. Blending is similar to stacking but simplifies the approach by using a hold-out set from the training data to train the meta-learner.

The main dataset is split into a training set and a validation set. First, the base models are trained on the training set. Their predictions for the validation set are then blended together using simple functions such as weighted averages, and this blended result is used to train the meta-learner.

Tobit Model

Tobit model is a class of regression models developed in 1958 with the intention of mitigating the problem truncated data. The Tobit model can also be considered a special case of a censored linear regression model that can be represented by the equation below:

$$y_i^* \to if \to y_L < y_i^* < y_U$$
$$y_L \to if \to y_i^* < y_L$$
$$y_U \to if \to y_i^* \geq y_U$$

Here, yi* is a latent variable that is not always observable and, y_L and y_U are the lower and upper limits to which the model is censored.

3. Results and Discussion

During the training phase of the analysis, several models, including Random Forest (RF), CatBoost Regressor, Gradient Boost, Ada Boost, XGBoost, Linear Discriminant Anlsysis (LDA), Light GBM etc were trained and tested on the datasets, along with well as other base models such as SVM, Decision Tree, Logistic regression model and KNN. The models were then used to make predictions on the test dataset and were ranked in descending order of AUC values for classification models and ascending order of MAE values for regression models. The classification model (s) with the highest AUC values and the regression model(s) with the lowest MAE values were identified as the best models. Additionally, stacking and blending were performed on the top 2, top 3 (BM3), top 4 (BM4), and top 5 (BM5) best machine learning models to create four new models. These models were also evaluated in the same way and compared against the individual models to evaluate the improvement resulting from stacking and blending.

The remainder of this section discusses the training, testing and validation results provided in **Tables 10 and 11**.

Table 10. Results from classification module (based on FS1 and FS2).

Data	Features	Best model	AUC	Precision	Accuracy	Recall
Train	Basic	Blended model	0.7758	0.671	0.6639	0.6639
Test	Basic	Blended model	0.7756	0.6497	0.6646	0.6646
Validation	Basic	Blended model	0.7511	0.67	0.6429	0.6429
Train	All	RF	0.812	0.6926	0.6933	0.6933
Test	All	RF	0.81	0.691	0.697	0.697
Validation	All	RF	0.78	0.699	0.69	0.69

3.1. Regression Module Results

The results were obtained through the process of deciding the ML components of the developed framework under the availability of both the feature sets explained in Table 12.

Table 11. Results from regression module (based on all features).

Duration	Ensemble type	Model	MAE (train)	MAE (test)	MAE (valid)
Short	Blend	GBR + LGBM + BR	5.892	6.11	15.78
Medium	Blend	RF + LGBM + GBR + XGBoost	14.815	15.63	31.42
Long	Stacked	GBR + LGBM + RF	44.25	47.26	36.45

Table 12. Feature set description.

Feature group	Description
Basic features (FS1)	Basic temporal and spatial information, vehicles, trucks, injuries
Full features (FS2)	AADT, hourly volume, road and environment factors, Responder information

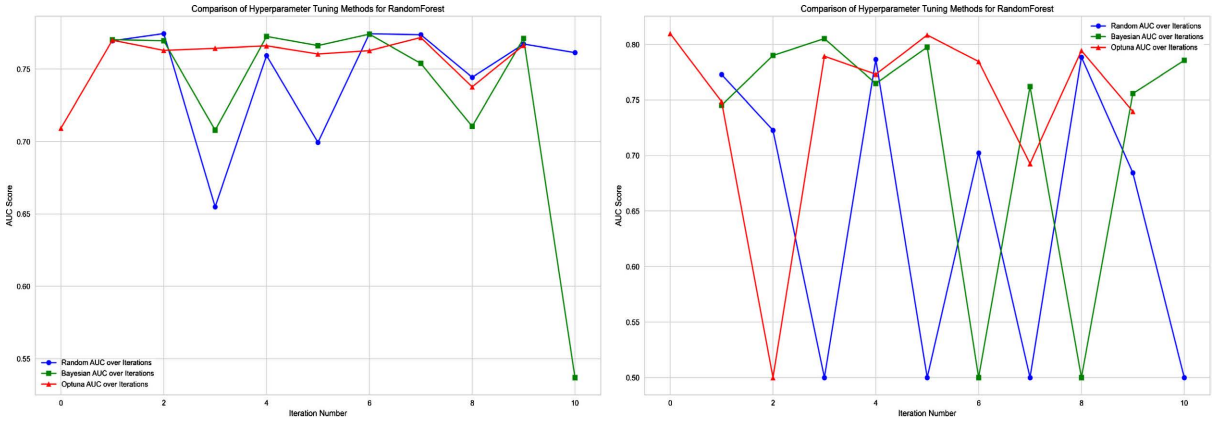
The basic feature set includes information collected from the site of the incident by the reporting policeman. This includes information about the time and location of crash, type of crash, vehicle information, involvement of trucks in the incident and the number of injured occupants and so on.

This information can be used to make the first prediction. In the next phase, more information relating to the incident is obtained like the AADT, hourly volume, road and environmental factors related to the location of the incident from RAMS, and whether a responder arrived or not at the incident site. This additional information along with the basic features (full features) helps make a more refined prediction of the time required to clear the incident.

3.2. Classification Module Results

A gradient Boost classifier model exhibited the best performance. However, To improve the performance of the model, hyperparameter tuning was done using 3 popular libraries-Scikit-learn, Scikit-Optimize, and Optuna. The algorithms run included Randomized Search CV, Bayesian optimization and Optuna is a newer open-source hyperparameter optimization framework based on TPE that is both

flexible and has a lightweight architecture helping define complex search spaces. Bayesian hypertuning improved the AUC to about 0.7742 and Random search produced an AUC of 0.7746 as shown in **Figure 5(a)**. A Random Forest classifier was also tested, but the performance was worse than for the gradient Boost classifier as shown in **Figure 5(b)**.

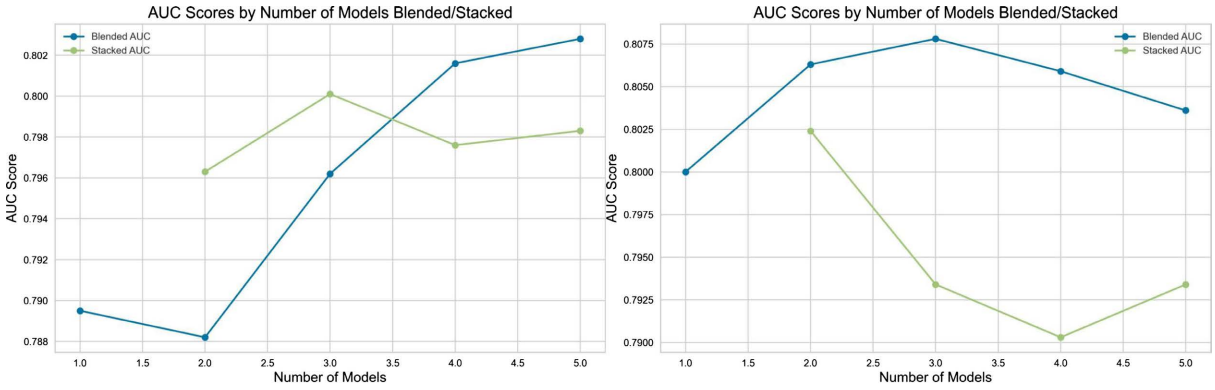


(a) (b)

Figure 5. Comparison of hyper-tuning performance for the best performing classifiers. (a) Gradient boost classifier (Basic feature set) (b) Random forest classifier. (All feature set)

Blending and stacking were also tested on the top performing model and as evident in **Figure 6(a)** blending the top 5 models improved AUC to 0.7758 and accuracy to 0.6639, a slight improvement of 1.5% and 0.65% respectively compared to the gradient boost classifier model. When more features are made available, the best-performing model is found to be the Random Forest model with an AUC of 0.812. Though hypertuning, blending and stacking were tested to improve performance of the Random Forest model, these models did not perform better than the base model as shown in **Figure 6(b)**. The parameter ccp_alpha is set to zero implying no minimal cost-complexity pruning (a technique that helps prevent overfitting) is applied. Gini impurity is used as a measure of the impurity or uncertainty to evaluate the quality of the split at each node. The maximum depth has been set to None implying that nodes are expanded until all leaves are pure or until all leaves contain less than min samples split samples. The maximum depth is set to "sqrt" implying that the square root of the number of features is considered to prevent over-fitting. The parameter max leaf nodes is set to None indicating an unlimited number of leaf nodes. The parameter min samples leaf or minimum number of samples required to be at a leaf node is set to 1 and min samples split or the minimum number of samples required to split an internal node is set to 2. The min weight fraction leaf set to 0 implies that any leaf node can contain samples with any weight including small weights.

The AUC graphs illustrating the performance of the models can be found in



(a) (b)

Figure 6. Comparison of hyper-tuning performance for the best performing classifier across different feature sets with stacking and blending. (a) Comparison of AUC for gradient boost classifier; (b) Comparison of AUC for random forest classifier.

Table 13. Configuration summary of the random forest classifier (All features).

Parameter	Value
Bootstrap	True
Ccp alpha	0.0
Criterion	"gini"
Max depth	None
Max features	"sqrt"
Max leaf nodes	None
Max samples	None
Min impurity decrease	0.0
Min samples leaf	1
Min samples split	2
Min weight fraction leaf	0.0
n estimators	100
n jobs	−1
Oob score	False
Random state	4867
Verbose	0
Warm start	False

Figure 7 and illustrate the training performance of the models. The categories 0, 1, 2 in the graph correspond to short, medium and long duration events. These models are trained, saved and used to predict on the validation dataset. The prediction results are shown in Figure 8. It is evident from comparing the results in Table 10 that the classification accuracy of shortduration incidents improved by 5% and medium duration events improved by 4% when additional features are available beyond the basic features. The 5 machine learning models that were blended to get a classificaton model for the basic feature set has been summarised in Table 10. The parameters of the random forest classification model used when additional feature are available is indicated in Table 13.

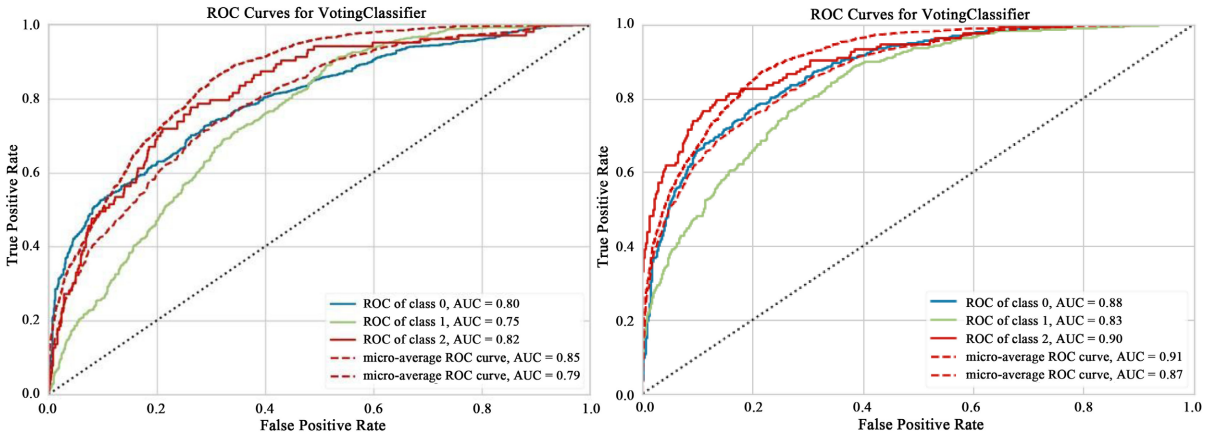


Figure 7. Comparison of AUC (basic feature set(left), all features (right) (validation dataset).

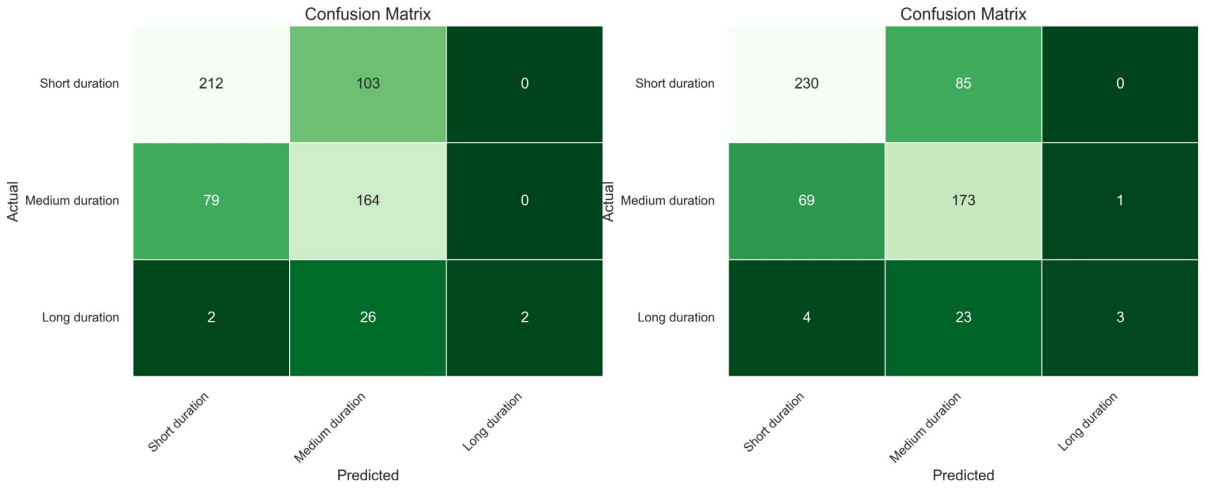


Figure 8: Confusion matrices (basic feature set(left), all features(right) (validation dataset) with Short (S), Medium (M), and Long (L) classifications. Observed classifications are on the x-axis and predicted classifications are on the y-axis.

From Table 11, it was found that for short and medium duration incidents, the RF model had the best performance, while a blend model consisting of the RF model and XGBoost performed the best in the case of long duration incidents. The best MAEs reported on the test dataset were 5.66, 14.07, and 41.9 minutes, respectively. When all the variables were considered, the MAE values improved to 5.76, 15.73, and 33.27 minutes, respectively. When the performance of the models on the validation dataset was evaluated, it was found that the MAE for the validation dataset was significantly higher than the MAE of the test dataset. This deviation from the expected values could be due to the large variance in the data. When additional features were available, an improvement in MAE values was observed for long duration incidents, as shown in Table 11. In addition to blending, hyperparameter tuning of the best model, as reported by the pycaret package, was also performed using the Optuna library. However, hyper-tuning did not help improve the performance of the model in any of the cases considered.

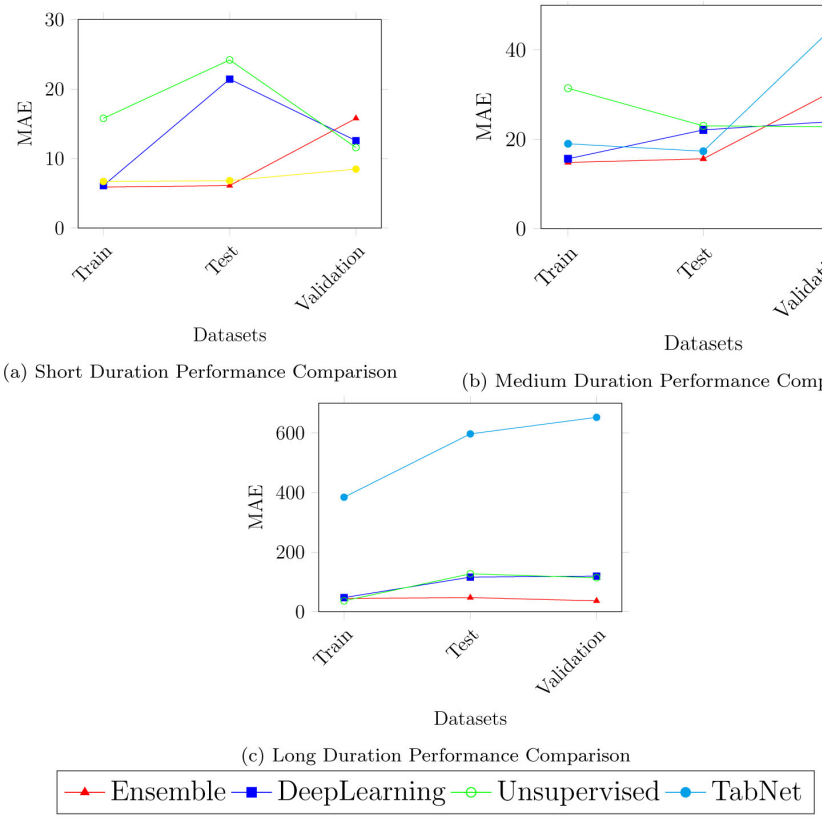


Figure 9. Performance comparison for different models across short, medium, and long durations.

3.3. Comparison between Blending and Stacking Approaches

A comparison of performance between stacked and blended model can be seen in Table 14 when only basic feature like location, day and time, detection method variables are available. When 2 or 3 models are stacked they perform slightly better than blending. However, as the number of base models are increase blending produces a better AUC with blending of 5 base models performing best. When more features are available it is evident from Table 15, even with 2 to 3 base models, stacking and blending produce highest AUC scores and increasing the number of base models most likely leads to overfitting. However, when more features are available, the Random Forest which by itself is an ensemble model is able to exhibit optimal performance.

Table 14. Results from classification module (based on basic features).

Model	Number of base models	AUC (train dataset)
Blended	2	0.7998
Blended	3	0.8065
Blended	4	0.8053
Stacked	2	0.8039
Stacked	3	0.8023
Stacked	4	0.8025

Table 15. Results from classification module (Based on all features).

Model	Number of base models	AUC (train dataset)
Blended	2	0.8636
Blended	3	0.8702
Blended	4	0.8674
Stacked	2	0.6983
Stacked	3	0.7011
Stacked	4	-

As the number of base models were increased from 3 to 4, the time taken to train the stacked base models also increased. With 3 base models blending was able to achieve an AUC of 0.8065 while for stacking adding more base models deproved the performance. The best stacking performance was achieved with stacking the top two models with an AUC of 0.8039.

A comparison of performance between stacked and blended model when all the features are considered together can be seen in Table 15. From the results it may be seen that the stacked model performs worse than the blended model. While the blended model achieved an AUC of 0.87 with 3 base models, the stacked model could only achieve an AUC of 0.7011. The amount of time required for training the stacked model is also exponentially greater than the blended model. The stacked model is unable to predict on the test dataset producing an AUC value of 0.

4. Discussion

In this section, the results of the classification and regression models that were developed and tested in the previous section are discussed. The motivation for choosing certain models or combinations of models over others is also explained.

A variety of models were tested for the classification module and ranked based on the highest AUC for both the basic and full feature sets. The highest AUC produced by ensemble models are 0.7756 and 0.81 on test dataset and 0.7511 and 0.78 on validation dataset. The error propagated through the classification module is reflected in the MAE value of the regression module. When the framework is used to predict on the validation dataset the model has never seen, the MAE produced is nearly 10 minutes higher than the test dataset for short duration events lasting less than 30 minutes. Using a Tabnet classifier instead of an ensemble classifier helps to achieve MAE close to the value obtained fo test dataset. For medium duration events the ensemble framework produces an MAE of 15.63 minutes for test dataset and 31.42 minutes for validation dataset. For the long duration dataset, we see that MAE is 47.26 minutes for test dataset and 36.45 minutes for the validation dataset. However Tabnet classifier does not do a good job of distinguishing medium and long duration events in the validation dataset and therefore does not help improve MAE for medium and long duration events producing MAE of 44.82 and 189.56 respectively.

Ultimately, the model that combines a random forest, an extra tree, and a light gradient boosting machine had the highest AUC and was selected for the classification module. The classification module provides a good estimation of the incident duration even with basic features. For the regression module, more features are available and the best performing models were selected based on the lowest MAE.

The combination of the classification and regression modules combine the best models for estimating the incident duration based on the pre-trained dataset. For validating the results against other models, the performance of the developed supervised framework was compared to two additional frameworks that included an unsupervised clustering, Deep learning model and a Tabnet model. The mean absolute error (MAE) was used as the evaluation metric for comparison.

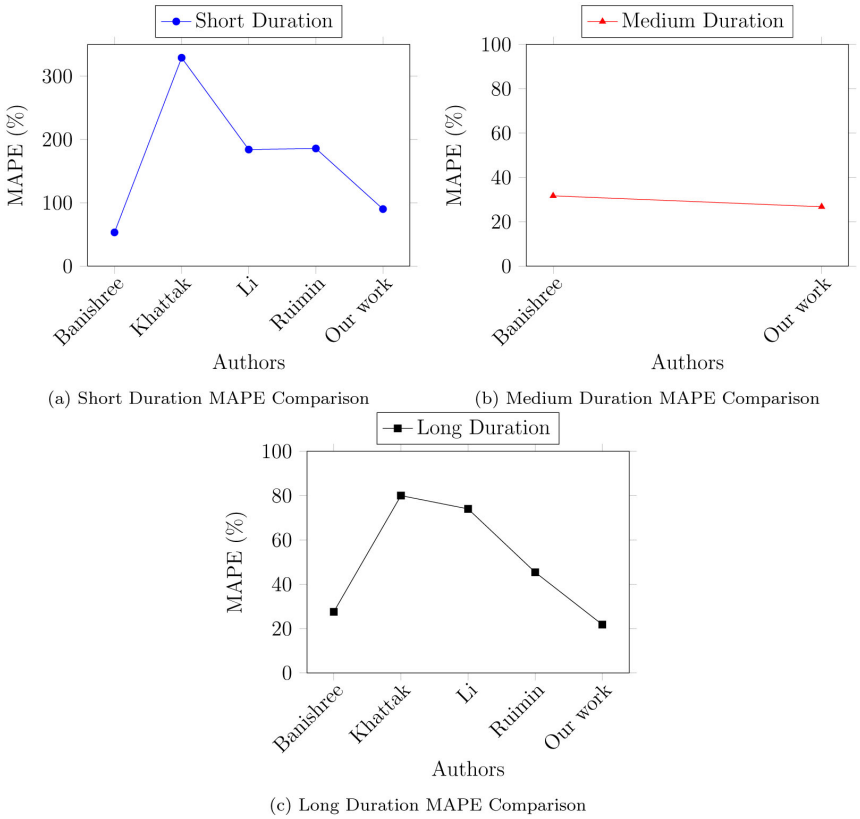
It was also observed that for short and medium-duration data, all three approaches produced similar MAE values. However, the pre-classification module improved the results of long-duration events more than the other two categories.

The performance of the models in the classification and regression modules was evaluated using the area under the curve (AUC) and mean absolute error (MAE), respectively. The highest AUC achieved was 86% when additional variables, such as road characteristics and responder information, were included. In terms of MAE, the lowest values achieved were 5.66 minutes for short-duration events, 14.07 minutes for medium-duration events, and an average of 41.9 minutes for long-duration events on the test dataset when using only basic features. When additional features were available, the MAE increased to 5.76 and 15.73 for short and medium incidents, respectively, but decreased to 33.27 minutes for long-duration incidents.

To compare the results to those obtained in previous research and to understand if the findings were consistent with those reported in the literature, the mean absolute percentage error (MAPE) values were also calculated for each of the models, as shown in **Figure 10**. The MAPE values achieved with the basic dataset were 99.16% for short duration, 24.45% for medium duration, and 18.34% for long duration. The inclusion of additional variables, such as road characteristics and highway helper information, resulted in improved MAPE values of 96.88% and 16.21% for short and long incidents, respectively.

In a study involving the dynamic prediction of incident duration using an adaptive feature set, the MAPE values reported for short-duration events were 100.9% for incidents lasting 5 - 15 minutes and between 75% and 96% for incidents lasting 16 - 35 minutes [14]. For medium-duration events lasting 36 - 200 minutes, a MAPE between 20% and 50% was reported. In comparison, the methods used in this paper achieved MAPE values in the range of 16% to 26% for events of a wide range of durations lasting from 30 minutes up to a day when all variables were considered, including variables that are likely not available at the initial stages of responding to incidents (included to ensure comparability with previous research). The MAPE, MAE, and mean absolute relative error (MARE) results re-

ported by other researchers working in this area have been provided in Table 16 for comparison.



(a) Short Duration MAPE Comparison

(b) Medium Duration MAPE Comparison

(c) Long Duration MAPE Comparison

Figure 10. Comparison of MAPE across different durations by authors.

Table 16. Comparison of our results with existing studies.

Reference	Results
Khattak *et al.* [44]	MAPE 5 - 15 min: 329%, >120 min: 80%
Valenti *et al.* [45]	MAPE ANN: 44%, SVR: 36%
Perera *et al.* [11]	MAPE between 100% and 40%
Qing *et al.* [28]	KNN: 59.2%, CART: 57.1% Quantile Reg: 49.1%
Tang *et al.* [29]	MAPE: 34.8%
Hamad *et al.* [30]	MAE: 36.52 min
Park *et al.* [13]	MAE: 0.18 to 0.29
Ma *et al.* [46]	MARE: 16.44% (<15 min), 33.13% (>15 min)
Ghosh *et al.* [31]	MAPE: 100.9% (5 - 15 min), 75 - 96% (16 - 35 min), 20 - 50% (>200 min), 61% and 27.58% (overall)
Our performance (basic features), MAPE	Short: 99.16% (<30min), Medium: 24.45% (30 - 120 min), Long: 18.34% (>120 min)
Our performance (all features), MAPE	Short: 96.88%, Medium: 26.4%, Long: 16.21%
Our performance (basic features), MAE	Short: 5.66 min, Medium: 14.07 min, Long: 41.9 min
Our performance (all features), MAE	Short: 5.76 min, Medium: 15.73 min, Long: 33.27 min

5. Conclusions and Recommendations

In this research, an ML-based classification and regression framework was proposed for incident duration prediction. The developed framework is capable of generating real-time predictions of incident duration based on the information provided to the response agency in the initial incident call. It additionally integrates other factors including road type, surface type, AADT, hourly volume information, etc., that are not available directly from the incident report enhancing the accuracy of prediction of the incident duration. The predicted incident duration can be used to develop a simple priority-based ranking system that helps traffic operators know which incidents to prioritize and what resources will likely be required. The incident duration information can also be combined with other information, such as the Level of Service (LOS) information, to generate additional actionable insights.

The framework utilized in this research involved two modules: a classification module followed by a regression module. The classification module has two functions: 1) immediately after an event has happened and is reported when only basic details like the location, date and time, and type of incident are available, this module provides a quick estimation of the incident duration using a supervised classification approach 2) when more details about the incident are available including road and environment conditions and helper facilities and so on the classification model makes a more accurate prediction of whether the incident will be a short, medium or long duration incident. Once the event is classified, the regression module provides a more accurate prediction in minutes, as shown in **Figure 3**. This can be very helpful for medium and long duration incidents to help deploy helper services faster or make traffic management decisions.

The success of the incident duration prediction is strongly dependent on the performance of the machine learning models. Among the various ML models trained, tested, and validated, the Random Forest model exhibited high efficiency and consistent performance across all the case scenarios tested. The ensembling technique of blending models exhibiting the lowest MAE values further improved the predictions.

Overall, the results indicate that the prediction of incident duration is more accurate when there is a pre-classification module that helps determine the class of incident duration. Though an unsupervised classification of events is an option to consider, it is not performing consistently across the three classes. The authors feel that the predictions benefit from the presence of the supervised classification module. In the regression module, however, the Tobit model performs almost on par with other single and blended regression models, as discussed earlier.

One of the challenges faced in developing and implementing the analytical framework was the skewness of the incident duration data. This was addressed by implementing a box-cox transformation on the target variable. The range of values of incident duration was large; therefore, applying a single model for the whole

dataset was not effective. This challenge was addressed by including a classification module to assign incidents to 3 ranges, short (<30 minutes), medium (30 minutes - 2 hours), and long (>2 hours). Categorical factors were hot encoded, and textual data (e.g., responder information) was extracted and converted to categorical variables for predictive model development.

A limitation of this research is the use of data for a single state (Iowa). While the results are validated for Iowa, other states may have different training and TIM resources and plans—leading to the results obtained in this research potentially not being applicable to other locations. Thus, future research should apply the framework and methods used in this paper to additional datasets from other geographical areas. Additionally, future research could evaluate the temporal transferability and stability of prediction models developed using this framework and methods.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Farradyne, P. (2000) Traffic Incident Management Handbook. Federal High-Way Administration, Office of Travel Management. https://transops.s3.amazonaws.com/uploaded_files/FHWA-HOP-10-013-Traffic-Incident-Management-Handbook.pdf

[2] Park, H., Haghani, A., Samuel, S. and Knodler, M.A. (2018) Real-Time Prediction and Avoidance of Secondary Crashes under Unexpected Traffic Congestion. *Accident Analysis & Prevention*, **112**, 39-49. https://doi.org/10.1016/j.aap.2017.11.025

[3] Cong, H., Chen, C., Lin, P., Zhang, G., Milton, J. and Zhi, Y. (2018) Traffic Incident Duration Estimation Based on a Dual-Learning Bayesian Network Model. *Transportation Research Record: Journal of the Transportation Research Board*, **2672**, 196-209. https://doi.org/10.1177/0361198118796938

[4] Oh, M., Wood, J. and Dong-O'Brien, J. (2023) The Integrated Analysis of Primary and Secondary Incident Characteristics: Focusing on the Impact and Scope of the Safety Service Patrol Program in Iowa. *Heliyon*, **9**, e17759. https://doi.org/10.1016/j.heliyon.2023.e17759

[5] Garib, A., Radwan, A.E. and Al-Deek, H. (1997) Estimating Magnitude and Duration of Incident Delays. *Journal of Transportation Engineering*, **123**, 459-466. https://doi.org/10.1061/(asce)0733-947x(1997)123:6(459)

[6] Peeta, S., Ramos, J.L. and Gedela, S. (2000) Providing Real-Time Traffic Advisory and Route Guidance to Manage Borman Incidents Online Using the Hoosier Helper Program. https://doi.org/10.5703/1288284313298

[7] Yu, B. and Xia, Z. (2012) A Methodology for Freeway Incident Duration Prediction Using Computerized Historical Database. *CICTP* 2012, Beijing, 3-6 August 2012, 3463-3474. https://doi.org/10.1061/9780784412442.351

[8] Weng, J., Qiao, W., Qu, X. and Yan, X. (2015) Cluster-Based Lognormal Distribution Model for Accident Duration. *Transportmetrica A: Transport Science*, **11**, 345-363. https://doi.org/10.1080/23249935.2014.994687

[9] Dimitriou, L. and Vlahogianni, E.I. (2015) Fuzzy Modeling of Freeway Accident Du-

ration with Rainfall and Traffic Flow Interactions. *Analytic Methods in Accident Research*, **5**, 59-71. https://doi.org/10.1016/j.amar.2015.04.001

[10] Kim, H.J. and Choi, H. (2001) A Comparative Analysis of Incident Service Time on Urban Freeways. *IATSS Research*, **25**, 62-72. https://doi.org/10.1016/s0386-1112(14)60007-8

[11] Pereira, F.C., Rodrigues, F. and Ben-Akiva, M. (2013) Text Analysis in Incident Duration Prediction. *Transportation Research Part C*: *Emerging Technologies*, **37**, 177-192. https://doi.org/10.1016/j.trc.2013.10.002

[12] Lopes, J., Bento, J., Pereira, F.C. and Ben-Akiva, M. (2013) Dynamic Forecast of Incident Clearance Time Using Adaptive Artificial Neural Network Models. https://trid.trb.org/View/1242255

[13] Park, H., Haghani, A. and Zhang, X. (2015) Interpretation of Bayesian Neural Networks for Predicting the Duration of Detected Incidents. *Journal of Intelligent Transportation Systems*, **20**, 385-400. https://doi.org/10.1080/15472450.2015.1082428

[14] Ghosh, B., Asif, M.T. and Dauwels, J. (2016) Bayesian Prediction of the Duration of Non-Recurring Road Incidents. 2016 *IEEE Region* 10 *Conference* (*TENCON*), Singapore, 22-25 November 2016, 87-90. https://doi.org/10.1109/tencon.2016.7847964

[15] Li, R. (2015) Traffic Incident Duration Analysis and Prediction Models Based on the Survival Analysis Approach. *IET Intelligent Transport Systems*, **9**, 351-358. https://doi.org/10.1049/iet-its.2014.0036

[16] Yang, D., Ozbay, K., Xie, K. and Yang, H. (2023) A Survival Analysis with Random Parameter Approach for Assessing Temporal Instability in Treatment Effect. *Safety Science*, **164**, Article ID: 106182. https://doi.org/10.1016/j.ssci.2023.106182

[17] Islam, N., Adanu, E.K., Hainen, A.M., Burdette, S., Smith, R. and Jones, S. (2021) A Comparative Analysis of Freeway Crash Incident Clearance Time Using Random Parameter and Latent Class Hazard-Based Duration Model. *Accident Analysis & Prevention*, **160**, Article ID: 106303. https://doi.org/10.1016/j.aap.2021.106303

[18] Chung, Y. (2010) Development of an Accident Duration Prediction Model on the Korean Freeway Systems. *Accident Analysis & Prevention*, **42**, 282-289. https://doi.org/10.1016/j.aap.2009.08.005

[19] Hu, J., Krishnan, R. and Bell, M. (2011) Incident Duration Prediction for In-Vehicle Navigation Systems. *Transportation Research Board* 90*th Annual Meeting*, Washington DC, 23-27 January 2011, 1-19.

[20] Kang, G. and Fang, S. (2011) Applying Survival Analysis Approach to Traffic Incident Duration Prediction. *ICTIS* 2011, Wuhan, 30 June-2 July 2011, 1523-1531. https://doi.org/10.1061/41177(415)193

[21] Thapa, D., Mishra, S., Velaga, N.R. and Patil, G.R. (2024) Advancing Proactive Crash Prediction: A Discretized Duration Approach for Predicting Crashes and Severity. *Accident Analysis & Prevention*, **195**, Article ID: 107407. https://doi.org/10.1016/j.aap.2023.107407

[22] Li, X., Liu, J., Khattak, A. and Nambisan, S. (2020) Sequential Prediction for Large-Scale Traffic Incident Duration: Application and Comparison of Survival Models. *Transportation Research Record*: *Journal of the Transportation Research Board*, **2674**, 79-93. https://doi.org/10.1177/0361198119899041

[23] Lin, L., Wang, Q. and Sadek, A.W. (2016) A Combined M5P Tree and Hazard-Based Duration Model for Predicting Urban Freeway Traffic Accident Durations. *Accident*

Analysis & Prevention, **91**, 114-126. https://doi.org/10.1016/j.aap.2016.03.001

[24] Lin, P.W., Zou, N. and Chang, G.L. (2004) Integration of a Discrete Choice Model and a Rule-Based System for Estimation of Incident Duration: A Case Study in Mayland. *CD-ROM of Proceedings of the* 83*rd TRB Annual Meeting*, Washington DC, 11-15 January 2004, 1-25.

[25] Li, R., Pereira, F.C. and Ben-Akiva, M.E. (2015) Competing Risks Mixture Model for Traffic Incident Duration Prediction. *Accident Analysis & Prevention*, **75**, 192-201. https://doi.org/10.1016/j.aap.2014.11.023

[26] Zhan, C., Gan, A. and Hadi, M. (2011) Prediction of Lane Clearance Time of Freeway Incidents Using the M5P Tree Algorithm. *IEEE Transactions on Intelligent Transportation Systems*, **12**, 1549-1557. https://doi.org/10.1109/tits.2011.2161634

[27] Knibbe, W.J.J., Alkim, T.P., Otten, J.F.W. and Aidoo, M.Y. (2006) Automated Estimation of Incident Duration on Dutch Highways. 2006 *IEEE Intelligent Transportation Systems Conference*, Toronto, 17-20 September 2006, 870-874. https://doi.org/10.1109/itsc.2006.1706853

[28] He, Q., Kamarianakis, Y., Jintanakul, K. and Wynter, L. (2013) Incident Duration Prediction with Hybrid Tree-Based Quantile Regression. In: Ukkusuri, S. and Ozbay, K., Eds., *Advances in Dynamic Network Modeling in Complex Transportation Systems*, Springer, 287-305. https://doi.org/10.1007/978-1-4614-6243-9_12

[29] Tang, J., Zheng, L., Han, C., Liu, F. and Cai, J. (2020) Traffic Incident Clearance Time Prediction and Influencing Factor Analysis Using Extreme Gradient Boosting Model. *Journal of Advanced Transportation*, **2020**, Article ID: 6401082. https://doi.org/10.1155/2020/6401082

[30] Hamad, K., Al-Ruzouq, R., Zeiada, W., Abu Dabous, S. and Khalil, M.A. (2020) Predicting Incident Duration Using Random Forests. *Transportmetrica A*: *Transport Science*, **16**, 1269-1293. https://doi.org/10.1080/23249935.2020.1733132

[31] Ghosh, B., Asif, M.T., Dauwels, J., Fastenrath, U. and Guo, H. (2019) Dynamic Prediction of the Incident Duration Using Adaptive Feature Set. *IEEE Transactions on Intelligent Transportation Systems*, **20**, 4019-4031. https://doi.org/10.1109/tits.2018.2878637

[32] Won, M., Kim, H. and Chang, G. (2018) Knowledge-Based System for Estimating Incident Clearance Duration for Maryland I-95. *Transportation Research Record*: *Journal of the Transportation Research Board*, **2672**, 61-72. https://doi.org/10.1177/0361198118792119

[33] Chen, J. and Tao, W. (2022) Traffic Accident Duration Prediction Using Text Mining and Ensemble Learning on Expressways. *Scientific Reports*, **12**, Article No. 21478. https://doi.org/10.1038/s41598-022-25988-4

[34] Chen, J., Tao, W., Jing, Z., Wang, P. and Jin, Y. (2024) Traffic Accident Duration Prediction Using Multi-Mode Data and Ensemble Deep Learning. *Heliyon*, **10**, e25957. https://doi.org/10.1016/j.heliyon.2024.e25957

[35] Fu, K., Ji, T., Zhao, L. and Lu, C. (2019) TITAN: A Spatiotemporal Feature Learning Framework for Traffic Incident Duration Prediction. *Proceedings of the* 27*th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Chicago, 5-8 November 2019, 329-338. https://doi.org/10.1145/3347146.3359381

[36] Laman, H., Yasmin, S. and Eluru, N. (2018) Joint Modeling of Traffic Incident Duration Components (Reporting, Response, and Clearance Time): A Copula-Based Approach. *Transportation Research Record*: *Journal of the Transportation Research*

Board, **2672**, 76-89. https://doi.org/10.1177/0361198118801355

[37] Mihaita, A.S., Liu, Z., Cai, C. and Rizoiu, M.A. (2019) Arterial Incident Duration Prediction Using a Bi-Level Framework of Extreme Gradient-Tree Boosting. arXiv: 1905.12254.

[38] Dong, X., Yu, Z., Cao, W., Shi, Y. and Ma, Q. (2019) A Survey on Ensemble Learning. *Frontiers of Computer Science*, **14**, 241-258. https://doi.org/10.1007/s11704-019-8208-z

[39] Sagi, O. and Rokach, L. (2018) Ensemble Learning: A Survey. *WIREs Data Mining and Knowledge Discovery*, **8**, e1249. https://doi.org/10.1002/widm.1249

[40] Wang, Z., Jiao, P., Wang, J., Luo, W. and Lu, H. (2023) Improved Two-Layer Stacking Model for Prediction of the Level of Delay Caused by Crashes: An Empirical Analysis of Texas. *Journal of Transportation Engineering*, *Part A*: *Systems*, **149**, Article ID: 05022008. https://doi.org/10.1061/jtepbs.teeng-7577

[41] Knickerbocker, S., Jagarlamudi, V.K., Hawkins, N. and Sharma, A. (2018) Iowa Dot Traffic Operations Open Data Service: User Guide and Software Requirements Specification. https://www.iowasudas.org/wp-content/uploads/2018/07/Iowa_DOT_traffic_ops_open_data_svc_guide_w_cvr.pdf

[42] Park, H. and Haghani, A. (2016) Real-time Prediction of Secondary Incident Occurrences Using Vehicle Probe Data. *Transportation Research Part C*: *Emerging Technologies*, **70**, 69-85. https://doi.org/10.1016/j.trc.2015.03.018

[43] Mumtarin, M., Knickerbocker, S., Litteral, T. and Wood, J.S. (2023) Traffic Incident Management Performance Measures: Ranking Agencies on Roadway Clearance Time. *Journal of Transportation Technologies*, **13**, 353-368. https://doi.org/10.4236/jtts.2023.133017

[44] Khattak, A., Wang, X. and Zhang, H. (2012) Incident Management Integration Tool: Dynamically Predicting Incident Durations, Secondary Incident Occurrence and Incident Delays. *IET Intelligent Transport Systems*, **6**, 204-214. https://doi.org/10.1049/iet-its.2011.0013

[45] Valenti, G., Lelli, M. and Cucina, D. (2010) A Comparative Study of Models for the Incident Duration Prediction. *European Transport Research Review*, **2**, 103-111. https://doi.org/10.1007/s12544-010-0031-4

[46] Ma, X., Ding, C., Luan, S., Wang, Y. and Wang, Y. (2017) Prioritizing Influential Factors for Freeway Incident Clearance Time Prediction Using the Gradient Boosting Decision Trees Method. *IEEE Transactions on Intelligent Transportation Systems*, **18**, 2303-2310. https://doi.org/10.1109/tits.2016.2635719