Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 8737185, 16 pages
https://doi.org/10.1155/2023/8737185

Research Article

A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data

Amirhossein Abdi,[1] **Seyedehsan Seyedabrishami**,[1] **and Steve O'Hern**[2]

[1]*Faculty of Civil and Environmental Engineering, Tarbiat Modares University, P.O. Box 14115-397, Tehran, Iran*
[2]*Sustainable Transport, Transport Research Centre VERNE, Tampere University, Tampere, Finland*

Correspondence should be addressed to Seyedehsan Seyedabrishami; seyedabrishami@modares.ac.ir

Received 17 February 2023; Revised 11 June 2023; Accepted 20 June 2023; Published 30 June 2023

Academic Editor: Tiziana Campisi

Detecting road accident impacts as promptly as possible is essential for intelligent traffic management systems. This paper presents a sequential two-stage framework for predicting the most congested traffic level that appears after an accident and the recovery time required for returning to the level of service that existed at the accident report time. As fewer accident characteristics are available at the report time, stage one models rely on real-time traffic and weather variables. With the arrival of the responders at the accident scene, more information is gained; therefore, the second stage model is activated, which updates the remaining accident duration time. We used eXtreme Gradient Boosting (XGBoost), a machine learning algorithm, and Shapley Additive exPlanations (SHAP) for making predictions and interpreting results, respectively. The results show that our framework predicts traffic levels with overall accuracies of around 80%, and duration models have high forecast accuracy with mean absolute percentage errors ranging between 7.26% and 21.59%. Overall, in the absence of accident information, SHAP values identified that weather factors, the traffic speed difference before and after an accident, traffic volume, and the percentage of heavy vehicles before the accident are the most important variables. However, accident variables, including the occurrence of injury or fatal accidents, rear-end collisions, and the number of involved vehicles, are among the most important variables in the second stage of the framework. The findings have practical implications for real-time traffic management of accident events. Road operators could manage post-accident traffic conditions more effectively, and road users could be alerted to take another route or manage their trip.

1. Introduction

Road accidents, extreme weather, and large-scale events can significantly impact traffic conditions. Traffic congestion caused by these events is known as non-recurrent congestion [1]. While accidents are not the only cause of non-recurrent congestion, they are estimated to be the principal reason for 72% of cases [2]. Research shows that a one-minute reduction in delay triggered by an accident has an average incident duration cost of 57 euros and can cost up to 1200 euros in very congested conditions [3]. In this respect, traffic accident post-impact (TAPI) models have been introduced to address this issue and guide traffic management centers and road users during post-accident periods.

One of the main objectives of the TAPI model is to predict accident duration, which is often divided into four parts: reporting time, dispatching time, response arrival time, and road clearance time [1, 4]. However, few studies to date have included traffic flow recovery time when considering the accident duration; as such, the duration analysed in this study covers the period between the reported accident time and when the traffic condition returns to the state that existed at the accident time (recovery time). Furthermore, in this study, traffic flow conditions have been divided into five service levels, and the proposed model identifies the appearance of a reduced level of service as well as accident duration.

The TAPI model should be immediately activated when an accident is reported to the traffic management center. Therefore, one of the key features of this model is to have a less time-consuming operation [1]. However, obtaining some information, such as detailed accident characteristics, depends on the police or responders' arrival at the accident site. Regarding this challenge, unlike most previous studies that rely on a single model, we proposed a novel sequential framework containing two stages. The sequential framework is developed in the first stage based only on readily available variables such as real-time traffic and weather factors. The second stage is activated to make updates as time passes and new accident details become available. This way, implementing post-impact prediction becomes more credible, and it could serve the needs of decision makers for reliable post-accident forecasts and provide real-time traffic information for other road users.

This paper investigated the potential of utilizing eXtreme Gradient Boosting (XGBoost) as an artificial intelligence model for enhancing the predictive capability of TAPI models. In addition, unlike previous TAPI studies that relied on prediction rather than drawing inferences from artificial intelligence techniques, this paper employed Shapley Additive exPlanations (SHAP) as a tool for knowledge generation and identifying factors that significantly impact the outputs.

2. Background

TAPI is usually measured by accident duration, which includes four phases. In this respect, previous studies focused on predicting the period between crash occurrence and road clearance [1]. The modelling techniques used by researchers can be classified into two groups: statistical and artificial intelligence methods.

Using statistical approaches, researchers explored the best probabilistic distribution for accident duration. Then, they adopted statistical models to shed light on the relationship between accident duration and other factors. Although the utilized dataset significantly limits the choice of the best-fitted distribution, log-logistic [5–8], log-normal [9–11], and Weibull [12–14] distributions were the most frequently reported distributions. Among statistical methods, regression models have been the primary choice [15–17]. Also, some researchers used hazard-based duration models, such as parametric accelerated failure time (AFT) models, to determine significant variables in different duration time phases [8, 18, 19]. Table 1 summarizes some of the TAPI studies that utilized statistical approaches.

With the rise in the amount and variety of data accessible via intelligent transportation systems, machine learning algorithms have been widely implemented to detect patterns behind big data with high accuracy. Table 2 presents some of the TAPI studies that used machine learning methods. Most of these studies applied the following methods.

2.1. Artificial Neural Network (ANN). This method is a non-linear, data-driven, and self-adaptive approach. Wei and Lee constructed two adaptive sequential neural networks to generate updates for the prior sequence [20]. Pereira et al. enhanced the predictive performance of the TAPI model by combining ANNs with a text analysis technique as a tool for the online extraction of accident information [21]. Lin and Li employed three artificial intelligence methods, including ANN, and proved that using crowdsourcing data from mobile apps results in better prediction of outliers [24].

2.2. Decision Tree and Tree-Based Approaches. Tree models are non-linear methods that can intrinsically identify and select the most important variables [25]. Decision trees, random forests, and gradient boosting have been chosen by researchers in TAPI studies [23, 24, 26]. Ma et al. employed a gradient boosting decision tree model and showed that this model is superior compared to conventional models and other machine learning methods, including random forest, support vector machine, and backpropagation neural network [23]. Lin and Li showed that random forest has a better prediction power than support vector machines, mainly when forecasting short-period congestions [24]. To the best of our knowledge, previous studies have not investigated applications of XGBoost models in TAPI framework prediction. We utilized this model in our framework to compare its prediction performance to prior results.

2.3. Hybrid Models. Researchers have also developed hybrid models to reach more accurate predictions. Lin et al. combined the MP5 tree model with a hazard-based duration model for predicting urban freeway traffic accident durations [22]. Shang et al. proposed the Bayesian optimization algorithm to optimize the parameters of the random forest model [27]. Zou et al. used a Bayesian averaging model to deal with uncertainties by averaging all plausible models [28].

3. Data and Proposed Framework

3.1. Data Description and Preparation. Three data sources are utilized in this research: (1) accident reports for almost 850 kilometers of rural highways of Khorasan Razavi province in Iran between 2015 and 2020 (selected highways are mostly in flat terrain with low curvature and speed limits ranging from 80 to 90 km/h); (2) real-time climatic data gathered by 20 synoptic weather stations throughout the province; and (3) real-time traffic data which are collected by 131 video detectors with an average spacing of 5.52 km between the detectors in the study area.

As shown in Figure 1(a), to combine the above-mentioned sources, each reported accident is assigned to the nearest upstream video detector and the nearest weather station. Subsequently, climatic factors (precipitation, temperature, visibility, and weather categories) and traffic variables (traffic speed, volume, headway, and traffic composition) can be extracted from the selected stations for a time window that is relevant to the accident. In our

Table 1: TAPI studies with statistical modelling.

Study	Author(s)	Study area	Methodology	Variables	Accuracy
[17]	Garib et al.	I-880, California, USA	Regression model	Number of lanes affected, number of vehicles involved, truck involvement, time of day, police response time, and weather condition	R^2: 0.83 Adjusted R^2: 0.81
[8]	Chung	24 major freeways, in Korea	Hazard-based duration model	Number of vehicles involved, number of injuries, occurrence of a fatal accident, involved vehicle type, accident type, accident location, and accident time	MAPE (mean absolute percentage error): 47%
[18]	Araghi et al.	London (the region within M25), UK	Hazard-based duration model	Accident type, accident severity, and network performance	MAPE: 43.7%
[16]	Weng et al.	I-95, Maryland, USA	Regression model (cluster-based log-normal distribution)	Accident severity, number of notifications sent out after the accident, number of responders on the scene, number of vehicles involved, number of lanes closed, time of day, day of the week, weather condition, and average traffic flow and speed	Mean absolute percentage error (MAPE): 34.1%
[15]	Khattak et al.	Hampton Roads, Virginia, USA	Regression model (quantile regression)	Roadway type, accident type, time of day, day of the week, number of vehicles involved, work zone involved, and accident detection source	Each quantile's pseudo-R^2: (0.4, 0.5, 0.1, 0.41)

Table 2: TAPI studies with machine learning techniques.

Study	Author(s)	Study area	Methodology	Variables	Accuracy
[20]	Wei and Lee	Freeways, Taiwan	ANN models	Accident characteristics, traffic data, time gap (between the accident occurrence and the recording time of detectors), space gap (between the accident and detector location), and geometric characteristics	MAPE is mostly under 40%
[21]	Pereira et al.	Expressways, Singapore	ANN models	Day of the week, time of day, blocked lanes, queue length, congestion status, and textual information	A median error of 9.9 min in the best model
[22]	Lin et al.	I-64 and I-190, USA	Hybrid model	Season, weekday, the hour of the day, weather condition, accident type, the fire involved, number of vehicles involved, rollover, number of blocked lanes, congestion level before the accident, and injury involved	MAPE: 36.2% for I-64 and 31.87% for I-190
[23]	Ma et al.	I-5, Washington, USA	ANN, support vector machine, random forest, and gradient boosting models	Accident type, lane closure type, injury involved, the fire involved, work zone involved, heavy truck involved, time of day, day of the week, the month of the year, weather condition, peak or off-peak, and hourly volume	Best model: gradient boosting MAPE for clearance time less than 15 min: 16.63% MAPE for clearance time larger than 15 min: 33.15%
[24]	Lin and Li	Urban and suburban areas of Beijing, China	ANN, support vector machine, and random forest models	Classification of roads, workday or not, peak hour or not, lane location, weather condition, pollution level, accident type, and congestion level before the accident	MAPE range: 5.5–53.8% RMSE range: 5.50, 34.74 min

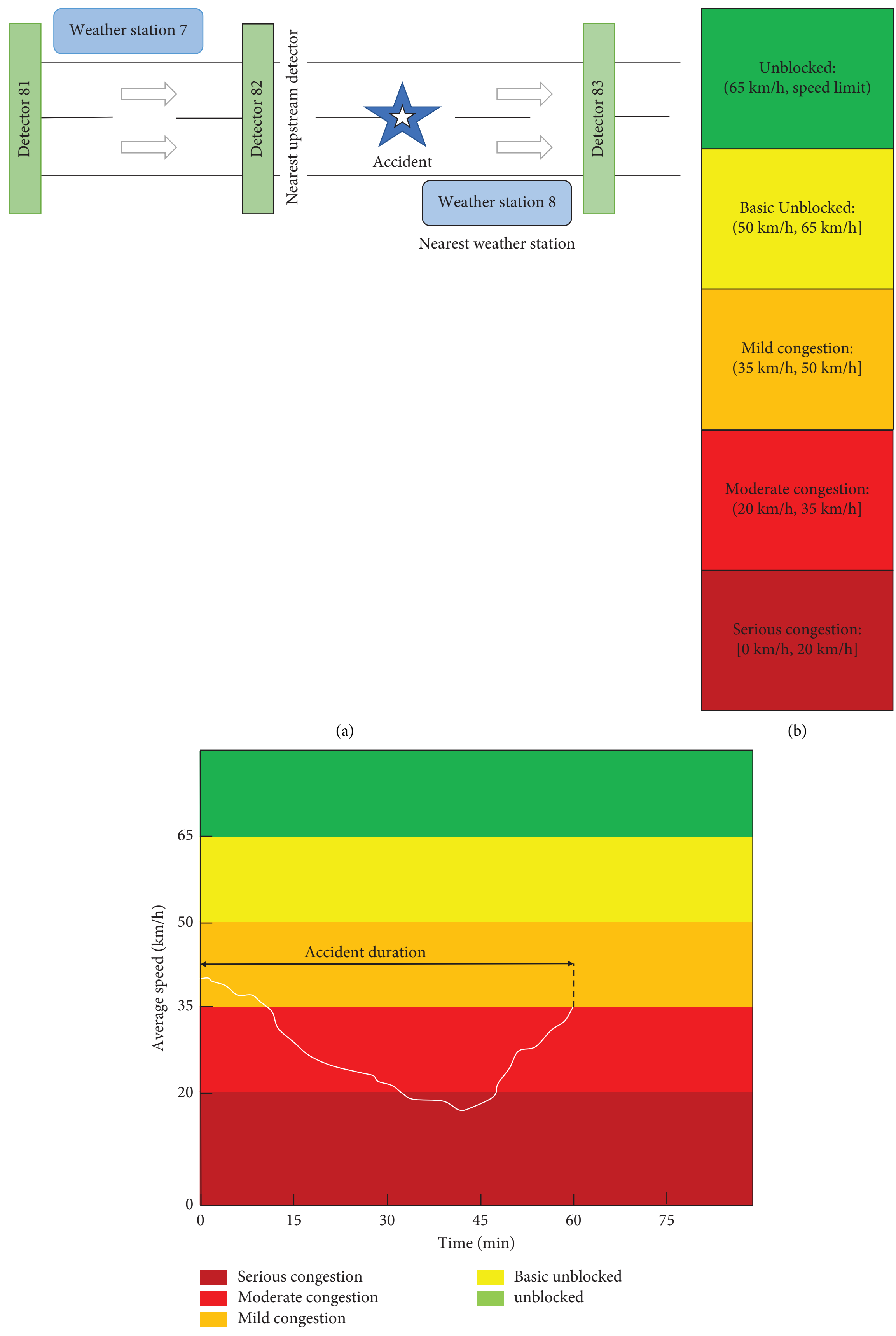

Figure 1: (a) Assigning accidents to detectors and weather stations. (b) Representation of highway traffic conditions. (c) Illustration of accident duration.

previous study, we showed that varying detector-to-accident distances affect the importance of traffic variables. It was demonstrated that the 1,000 m threshold has the highest Variable Importance Measure (VIM) for traffic factors [29]. Concerning this point, we only included accidents that occurred less than 1,000 m away from the nearest upstream detectors. The resulting dataset contains 6381 incidents with the associated accident, real-time traffic, and weather factors. As this dataset contains almost 120 explanatory variables, we only reported our models' five most important variables in the results and discussion section. Table 3 summarizes descriptions and statistics of these significant variables used in this paper.

We employed the average speed of the rural highway to introduce traffic levels and measure how an accident impacts these levels. Figure 1(b) shows five highway traffic levels with associated speed intervals and color indicators [30]. In this study, traffic recovery time is defined as when the traffic condition returns to a similar level to the accident time. Take the example of Figure 1(c); traffic was at a mild congestion level at the accident time. However, the condition deteriorates over time due to the accident and reaches a serious congestion state. Next, recovery starts, and the condition returns to a mild level. As shown in Figure 1(c), this period between two subsequent similar levels is the accident duration.

3.2. Proposed Framework. Figure 2 presents the proposed framework of the sequential post-impact prediction models. As shown in Figure 2, the framework includes two major stages.

The first prediction stage starts when an accident is reported to the traffic management center (TMC). At this point, there is a low chance of having detailed information about the accident, and since the TAPI models should react promptly, only readily available variables (real-time traffic and weather factors) are fed into the first-stage models to predict the impact of the accident on traffic levels. The first-stage models are presented as follows:

(i) Model 1.1: The first model in the sequence is Model 1.1. This model is a binary prediction tool for forecasting whether the reported accident will aggravate traffic congestion (scenarios 6–14 in Figure 3) or whether the traffic level will remain unchanged at the same level as existed before the accident (scenarios 1–5 in Figure 3(a)).

(ii) Models 1.2, 1.3, and 1.4: As shown in Figure 2, these models will be activated only if Model 1.1 predicts that a poorer traffic level may appear after the accident. These models predict the worst traffic level of service that could be experienced during the post-impact period. The respective presentations of these models are as follows:

  (1) Model 1.2 is used for accidents occurring at mild congestion levels (Figure 3(b)) and predicts the poorest level among moderate (scenario 6 in Figure 3(b)) and serious congestion (scenario 7 in Figure 3(b)).

  (2) Model 1.3 is used for accidents occurring at the basic unblocked level (Figure 3(c)) and predicts the poorest level among mild (scenario 8 in Figure 3(c)), moderate (scenario 9 in Figure 3(c)), and serious congestion (scenario 10 in Figure 3(c)).

  (3) Model 1.4 is used for accidents occurring at the unblocked level (Figure 3(d)) and predicts the poorest level among basic unblocked (scenario 11 in Figure 3(d)), mild (scenario 12 in Figure 3(d)), moderate (scenario 13 in Figure 3(d)), and serious congestion levels (scenario 14 in Figure 3(d)).

(iii) Model 1.5: The final model in this sequence is Model 1.5. Unlike previous models in this stage, Model 1.5 has a continuous outcome and predicts the duration (recovery time) until the traffic level returns to a similar level to the accident time.

The second stage starts when responders arrive at the accident scene and report additional information about the accident to the TMC. It should be noted that in a real-world accident scenario, the information reported by responders to the TMC may be insufficient or incomplete. In this study, one of the advantages of the stage two model is its flexibility as it adopts not only reported accident information but also employs updated real-time traffic and weather factors at the officers' arrival as well. Therefore, in cases where there is a lack of accident information, real-time factors will be used to make updates.

Figure 2 represents how Model 2 works. This model aims to predict the remaining time required for reaching traffic recovery and enhance the performance utilizing updated variables (accident factors, real-time traffic, and weather variables). For this purpose, as shown in Figure 2, the traffic condition is first assessed by the model upon the arrival of the responders. If there is still a poorer level, Model 2 calculates the remaining time required for traffic recovery. Otherwise, no calculations are needed since the traffic status has returned to the before-accident condition.

Table 4 summarizes the characteristics of models proposed in our framework with associated objectives, data, and parameters necessary to be employed. To create these models, the dataset was firstly split into the train (80%) and test (20%) sets. Then, four artificial intelligence techniques including ANN (multilayer perceptron), random forest, support vector machine, and XGBoost were trained on the train set based on the necessary datasets and parameters for each model mentioned in Table 4. Subsequently, all the trained models were tested on the test set and various performance measures were calculated. Finally, the most successful modelling technique was selected to be discussed and interpreted in Section 5 of this study (results and discussion).

4. Methodology

4.1. Extreme Gradient Boosting (XGBoost). XGBoost, an ensemble-based tree model, is applied as a machine learning technique to model the TAPI framework

Table 3: Summary of descriptions and statistics of variables.

Variables	Description (unit)	Mean (std. dev.)
Traffic		
PctHV_B	Percentage of heavy vehicles for a 5 min interval before an accident	4.63 (9.11)
AvgSpdDiff_BA	Average speed difference between a 1 min interval before and 1 min interval after an accident (km/h)	22.71 (13.38)
AvgSpdArrival_A	Average speed for a 1 min interval after the arrival of officers at the accident scene (km/h)	58.46 (18.44)
StdHdw_A	Standard deviation of headway for a 1 min interval after an accident (seconds)	5.62 (3.23)
Vol_B	Traffic volume of a 5 min interval before an accident (vehicles)	19.84 (11.04)
Accident		
Injury or fatal	1 if either an injury or fatal accident occurred (35%); 0 otherwise (65%)	0.35 (0.47)
No Inv_Veh	Number of involved vehicles in the accident	2.28 (1.04)
Rear end	1 if a rear-end accident occurred (48%); 0 otherwise (52%)	0.48 (0.50)
Weather		
Precipitation	Amount of precipitation at the accident time (mm)	1.84 (3.49)
Weather	Ordinal variable for weather condition {1: sunny (56%), 2: cloudy (24%), 3: light rain (12%), 4: heavy rain or snow (8%)}	1.72 (0.96)

presented in the previous section. XGBoost is an effective tool for numerous reasons: it can handle missing values, has high flexibility, and is insensitive to multicollinearity similar to decision trees [31–33]. Chen introduced this model in 2016 as an efficient implementation of gradient boosting models [34]. The model improves its operation by utilizing a set of decision trees. Each tree can learn from the previous tree to support the subsequent tree and hence provide stronger results [35]. Each single decision tree has a tree-like structure starting from a root node and ending at terminal nodes, with the nodes between the root node and leaf nodes being internal nodes. The decision tree algorithm splits nodes into further subnodes based on conditional statements for variables. The splitting process ends when the greatest possible homogeneity is met [36].

Regarding the equations behind XGBoost, assume a dataset with $i = \{1, 2, \ldots, n\}$ samples, X_i is the vector of independent variables, and y_i is the dependent variable. Equation (1) predicts the dependent variable $\hat{y}_i$ using the independent variables X_i and K decision tree models:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in F, \tag{1}$$

where F refers to the tree space and f_k is the additive function of each tree in this space. The best set of functions for equation (1) is found by minimizing the objective function expressed as

$$\text{Objective} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2}$$

where $l(\hat{y}_i, y_i)$ is a loss function measuring the performance of the model on a dataset and $\Omega(f_k)$ represents the regularization term penalizing the model complexity to harness overfitting. $\Omega(f)$ for each tree is defined as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, \tag{3}$$

where T is the number of leaves in the tree; ω represents the weight of the leaf; and γ and λ are regularization parameters. Further detailed descriptions of the XGBoost equations are available in the study by Chen and Guestrin [34].

4.2. XGBoost Hyperparameters. Hyperparameter tuning is necessary to maximize model prediction performance and control overfitting [32]. We used the XGBoost [37] and scikit-learn [38] packages for model training and parameter tuning in Python 3.10.5. Several parameters that are selected for tuning are as follows:

(i) "The number of iterations" represents the number of decision trees that are trained in the ensemble model.

(ii) "Maximum depth" is the maximum allowed depth for each fitted tree. Large values of this parameter result in overly large trees and overfitting.

(iii) "Subsample" represents the proportion of observations randomly chosen for each tree. Lower values can prevent overfitting. However, too small parameters cause underfitting.

(iv) "Colsample bytree" denotes the fraction of columns randomly selected for each tree. This parameter can be used to prevent overfitting.

(v) "Learning rate" or "Eta" shrinks the weights and provides a more robust model.

(vi) "Alpha" and "Lambda" are regularization terms to make the model more conservative.

4.3. XGBoost Evaluation Metrics. In our study, the XGBoost model is randomly trained on 80% of the dataset, and the remaining 20% is the test set. Models 1.1, 1.2, 1.3, and 1.4 are used as classifiers to predict the most congested traffic level after a specific accident. The overall performance of the classifiers is assessed by calculating the overall accuracy defined in the following equation:

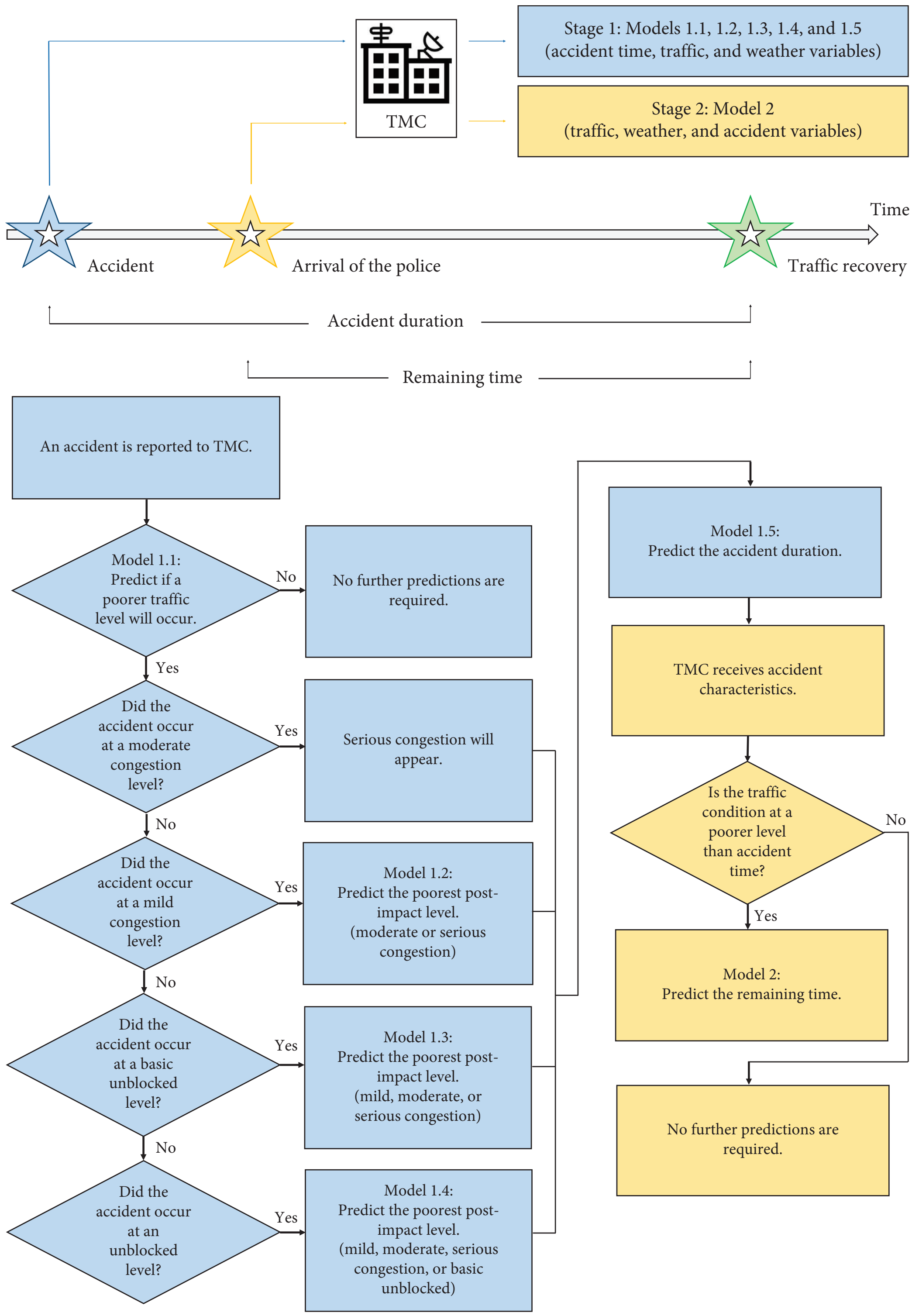

FIGURE 2: Schematic diagram of the TAPI models.

$$\text{Accuracy}_{\text{overall}} = \frac{\text{number of true predictions}_{\text{overall}}}{\text{total number of cases}_{\text{overall}}}. \quad (4)$$

The accuracy of a specific class is also computed to evaluate the model's performance with respect to each traffic condition. Per-class accuracy for class label m can be defined as

$$\text{Accuracy}_{m} = \frac{\text{number of true predictions}_{m}}{\text{total number of cases}_{m}}. \quad (5)$$

Models 1.5 and 2 are quantitative models that predict accident duration. Two metrics were used for these models: mean absolute percentage error (MAPE), as shown in

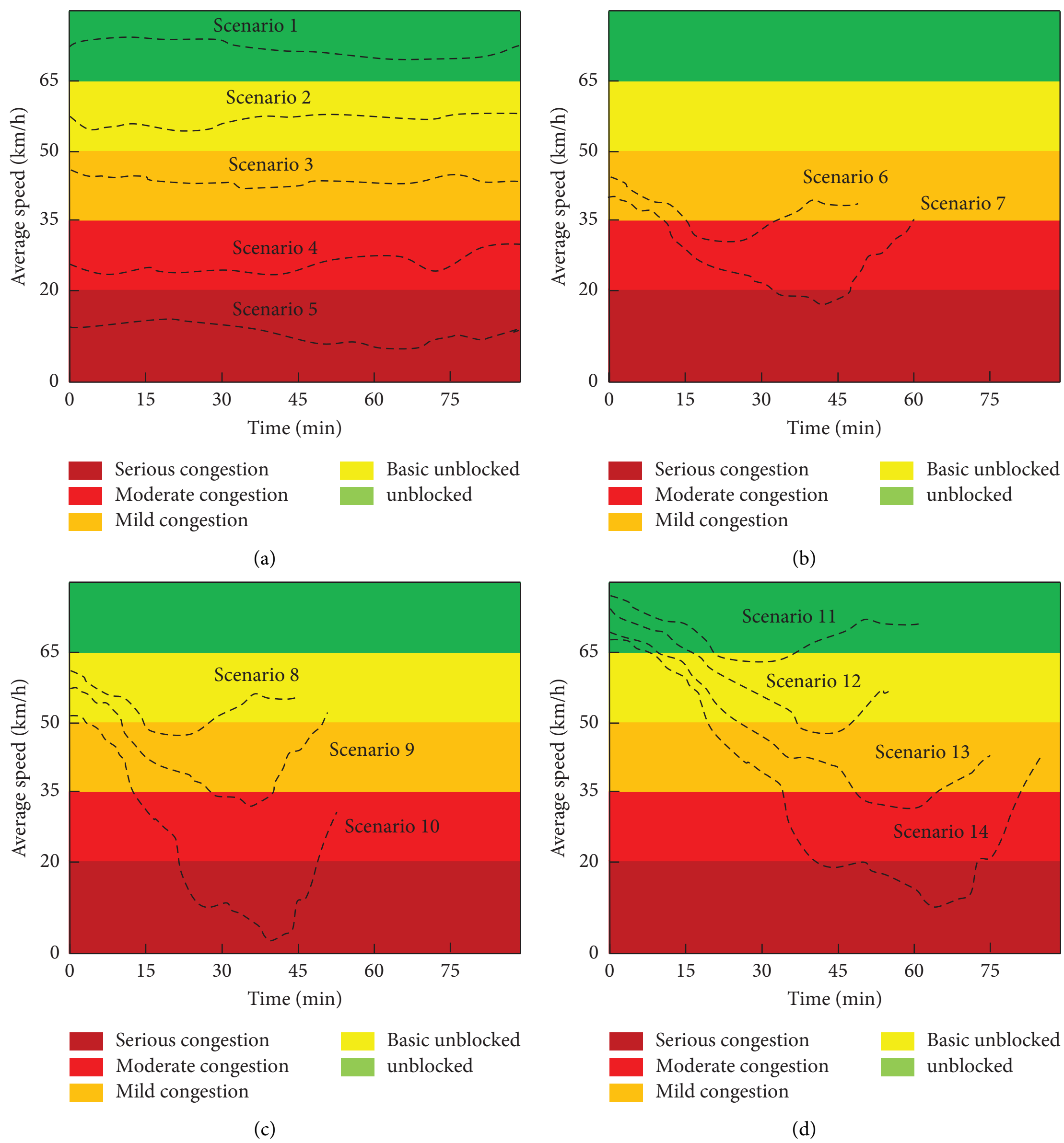

FIGURE 3: (a) Example scenarios of the traffic levels which have remained unchanged after the accident. (b) Possible scenarios predicted by Model 1.2. (c) Possible scenarios predicted by Model 1.3. (d) Possible scenarios predicted by Model 1.4.

equation (6), is introduced to measure the overall performance, and root mean square error (RMSE), as shown in equation (7), is utilized to explore the absolute value of deviation for these models [24].

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{t}_i - t_i}{t_i}\right| \times 100, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{t}_i - t_i\right)^2}, \tag{7}$$

where $\hat{t}_i$ and $t_i$are forecast and actual values for accident duration, respectively, and n is the total number of observations.

4.4. XGBoost Interpretation. Many researchers have preferred machine learning algorithms with great potential in forecasting. Nevertheless, interpreting the results has been a barrier to this adoption. Lundberg and Lee proposed Shapley Additive exPlanations (SHAP) as a game theoretic approach to tackle this challenge [39]. SHAP employs an additive feature attribution technique with the explanatory function g as a linear function of the feature attribution values. Assume a model with M input features $\mathrm{x} = (x_1, x_2, \ldots, x_M)$. The explanatory function $g(\mathrm{x}')$ with simplified input x' for an original model $f(\mathrm{x})$ is defined as

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M}\phi_i x_i', \tag{8}$$

Table 4: Summary of the models' characteristics used in the framework.

Model	Problem type	Objective	Activation time	Data used to define the model	Parameters necessary for the model
1.1	Binary classification	To predict whether the accident brings about congested traffic levels or not	Upon an accident being reported to the TMC	(i) Real-time traffic data captured by detectors and aggregated to 1 and 5 min intervals (ii) Real-time weather data gathered by weather stations	(i) Descriptive statistics (mean, standard deviation, and coefficient of variation) for traffic flow parameters including speed, volume, headway, density, and traffic composition for 1–5 minutes before and after the occurrence of the accident (ii) The weather condition in different categories (sunny, cloudy, light rain, heavy rain, and snow), amount of precipitation, temperature, visibility, and lighting condition at the accident time
1.2	Binary classification	To predict whether moderate or serious congestion will appear after accidents occurring at mild congestion traffic levels			
1.3	Multiclass classification	To predict whether mild, moderate, or serious congestion will appear after accidents occurring at basic unblocked traffic level			
1.4	Multiclass classification	To predict whether basic unblocked, mild, moderate, or serious congestion will appear after accidents occurring at unblocked traffic level			
1.5	Regression	To predict the recovery time required for returning to the same level that existed at the accident time			
2	Regression	To predict the remaining recovery time required for returning to the same level that existed at the accident time (this model provides an update for recovery time predicted by Model 1.5 upon arrival of responders)	Upon the responder's report of accident characteristics to the TMC	(i) Accident characteristics reported by responders at the accident scene (ii) Real-time traffic data captured by detectors and aggregated to 1 and 5 min intervals (iii) Real-time weather data gathered by weather stations	(i) Accident severity, number of injuries, collision type, type and number of involved vehicles, and occupants' characteristics (gender and age) (ii) Descriptive statistics (mean, standard deviation, and coefficient of variation) for traffic speed, volume, headway, density, and traffic composition at the arrival of responders (iii) The weather condition in different categories (sunny, cloudy, light rain, heavy rain, and snow), amount of precipitation, temperature, visibility, and lighting condition at the arrival of responders

where ϕ_0 is a constant value and x' and x in functions $g(x')$ and $f(x)$ are connected through a mapping function. Also, x_i' is 1 when a feature is observed; otherwise, it is 0. ϕ_i represents the feature attribution value for the feature i. Equation (9) calculates ϕ_i based on Shapley values:

$$\phi_i = \sum_{z' \subseteq x'} \frac{|z'|!\left(M - |z'| - 1\right)!}{M!}\left[f_x\left(z'\right) - f_x\left(\frac{z'}{i}\right)\right], \quad (9)$$

where z' is a set containing non-zero indexes in x' and $|z'|$ denotes the number of non-zeros in the set z'.

SHAP values have two outstanding merits over other measures, such as feature (variable) importance measure; first, the explanatory function in equation (8) enables them to interpret outcomes locally (for a single prediction) and globally (for the entire model). Second, SHAP can examine the negative or positive contribution of each feature along with its importance [40]. For further detailed explanations about SHAP value, readers are referred to the study by Lundberg and Lee [39].

5. Results and Discussion

5.1. XGBoost Model Hyperparameter Tuning. All the hyperparameters presented in this paper are tuned by employing a grid search with a 5-fold cross-validation process which means that every combination of the defined hyperparameters is assessed by five random subsamples of the train data, and then the best set of hyperparameters is selected. Table 5 shows the optimal XGBoost hyperparameters of the models.

5.2. XGBoost Model Evaluation and Interpretation. The first stage of the XGBoost models consists of immediately predicting the lowest traffic level that might occur after the accident (Models 1.1 to 1.4). Two metrics, including overall accuracy and per-class accuracies, are computed to investigate the predictive performance of classifiers. These criteria are shown in Table 6. It can be seen that Model 1.1 has the highest accuracies and can predict the outbreak of inferior traffic levels better than those that stayed at a specific level without deterioration. This result makes a good start for the TAPI framework because Model 1.1 is a keystone for the next classifiers, and Models 1.2, 1.3, and 1.4 are activated if Model 1.1 predicts the appearance of poorer traffic levels. These models forecast the poorest upcoming level on the basis of the traffic level at which an accident occurred. The less congested the related accident condition is, the more traffic levels there will be as possible outcomes. The number of possible outcomes may affect the predictive performance of the models. As shown in Table 6, overall performance declined below 80% for Models 1.3 and 1.4 with three and four outcomes, respectively. Moreover, serious congestion rarely occurs, so Models 1.3 and 1.4 have lower accuracy compared to other levels. However, our framework enhanced the TAPI model's predictive power compared to the previous study by Lin and Li, which reported 28.97%–32.63% accuracies for congested levels [24].

Figure 4 illustrates the SHAP summary plot with the top five important variables for Model 1.1. As shown in the figure, most of these variables are traffic-related factors that emphasize these factors' importance in the model.

The difference in average speed before and after an accident is the most important variable. Higher values of this variable (red dots) result in higher SHAP values on the horizontal axis, which means a higher probability of poorer post-accident traffic levels. This finding is also in line with previous findings that showed that upstream shockwaves can be experienced by a sudden decline in speed [41].

The second most important variable is traffic volume before an accident. Higher traffic volume values increase the appearance of poorer traffic levels. Similarly, Lin and Li concluded that higher congestion levels before an accident increases the probability of severely congested post-accident periods [24].

The third important variable is the standard deviation of headway after an accident. As shown in Figure 4, the lowest variations in headway (light blue dots) are greatly related to a low likelihood of higher congestion levels. Moreover, the highest variations (light red dots) showed a positive impact on the more congested situations, though their related SHAP values are not as large as lower variations.

Precipitation is the most important non-traffic factor. While previous studies mainly introduced weather variables as dummy variables [16, 22, 24], we included continuous climatic factors as well as categorical variables. SHAP value results show that large amounts of precipitation at the accident time increased the probability of confronting inferior conditions. However, as expected, low amounts of rainfall were not a contributor to accident duration.

The final important variable is the percentage of heavy vehicles. It was found that a higher presence of heavy vehicles increases the likelihood of experiencing worsening traffic levels following an accident. This is likely due to the increased impact that heavy vehicles have on traffic flow, particularly in congested conditions. This association was confirmed through a correlation coefficient of −0.22 between the percentage of heavy vehicles and average traffic speed.

Models 1.5 and 2 are quantitative duration models predicting accident duration and the remaining recovery time, respectively. Table 7 summarizes measured RMSE and MAPE for these models. Compared to MAPE values in Tables 1 and 2, our proposed framework shows a considerable improvement. Also, Lin and Li calculated RMSE values at a range between 5.50 and 34.74 min [24]; in our study, these values stayed at a range from 6.05 to 16.24 min, which is a more reliable range with a lower upper bound.

To further analyse the predictive power of our framework, we computed the absolute difference (AD) between prediction and observed values for each instance in the train and test set. As shown in Figure 5, Models 1.5 and 2 accurately predicted 72% and 82% of the test set cases with absolute differences of less than 10 minutes, respectively. This result is valuable since a 10-minute error range is more tolerable for road users than the longer ranges shown in Figure 5, and our duration prediction framework, in most

Table 5: Hyperparameter tuning results.

Hyperparameter	The first-stage models					The second-stage model
	1.1	1.2	1.3	1.4	1.5	2
The number of iterations	500	125	150	300	550	350
Maximum depth	6	5	5	6	7	6
Subsample	0.65	0.8	0.75	0.7	0.6	0.65
Colsample bytree	0.4	0.45	0.45	0.4	0.4	0.4
Learning rate	0.01	0.05	0.03	0.01	0.01	0.01
Alpha	0.25	0.25	0.25	0.2	0.25	0.2
Lambda	1.3	1.5	1.5	1.4	1.35	1.4

Table 6: Results of predicting traffic levels.

Model	Traffic level statement	Train size	Test size	Per-class accuracy (train)	Per-class accuracy (test)	Overall accuracy
1.1	Poorer traffic levels will appear	3216	804	0.8781	0.8483	Train: 0.8643
	Poorer traffic levels will not appear	1889	472	0.8407	0.8114	Test: 0.8346
1.2	Serious congestion will appear	196	49	0.8214	0.7755	Train: 0.8576
	Moderate congestion will appear	415	104	0.8747	0.8365	Test: 0.8170
1.3	Serious congestion will appear	102	26	0.7647	0.7308	Train: 0.7942
	Moderate congestion will appear	309	77	0.8123	0.7922	Test: 0.7811
	Mild congestion will appear	522	130	0.8582	0.7846	
1.4	Serious congestion will appear	165	41	0.7152	0.6585	
	Moderate congestion will appear	250	63	0.7480	0.7143	Train: 0.8026
	Mild congestion will appear	421	105	0.7981	0.7429	Test: 0.7392
	Basic unblocked will appear	836	209	0.8385	0.7608	

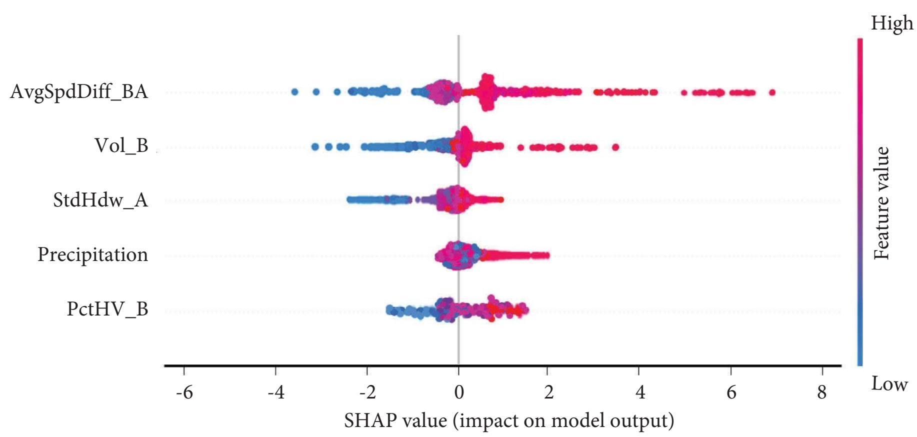

Figure 4: SHAP summary plot of the top five important variables in Model 1.1.

Table 7: MAPE and RMSE measures for Models 1.5 and 2.

Model	Model 1.5		Model 2	
	Train (3216)	Test (804)	Train (1478)	Test (370)
MAPE (%)	17.35	21.59	7.26	10.44
RMSE (min)	13.63	16.24	6.05	8.86

cases, has an error range of fewer than 10 minutes, especially after updating durations by Model 2.

Figure 6 shows SHAP values for the top five important variables in the duration models. While traffic variables play a crucial role in Model 1.5, accident-related variables become the most important variables in Model 2. With the better performance of Model 2, this result highlights how employing detailed descriptions of the accident provides more accurate outcomes. However, these factors are unlikely to be available during the first minutes following an accident, necessitating using Model 1.5.

Regarding the SHAP values of Model 1.5, the predicted level is the first important variable. This variable is the outcome of Models 1.2 to 1.4 and represents the most congested level that appears during a post-accident period. The more congested the predicted level is, the longer the accident duration will be. Similar to SHAP values in Figure 4, the difference in speed before and after an accident, traffic volume before an accident, and percentage of heavy vehicles before an accident are important variables with the same direction of effect on the outcome. Another variable

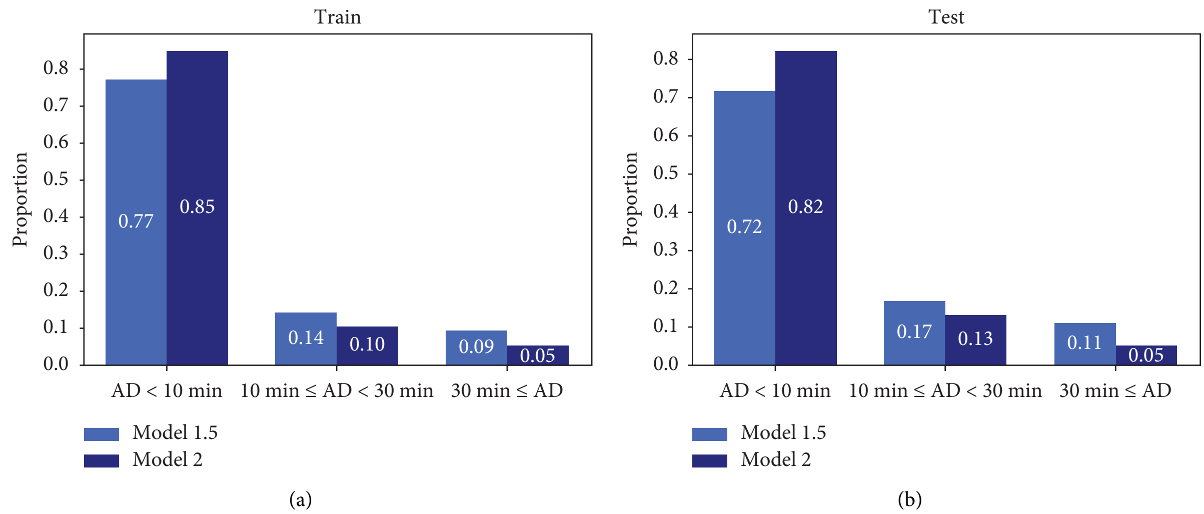

Figure 5: Results of absolute difference (AD) for (a) the train and (b) the test set.

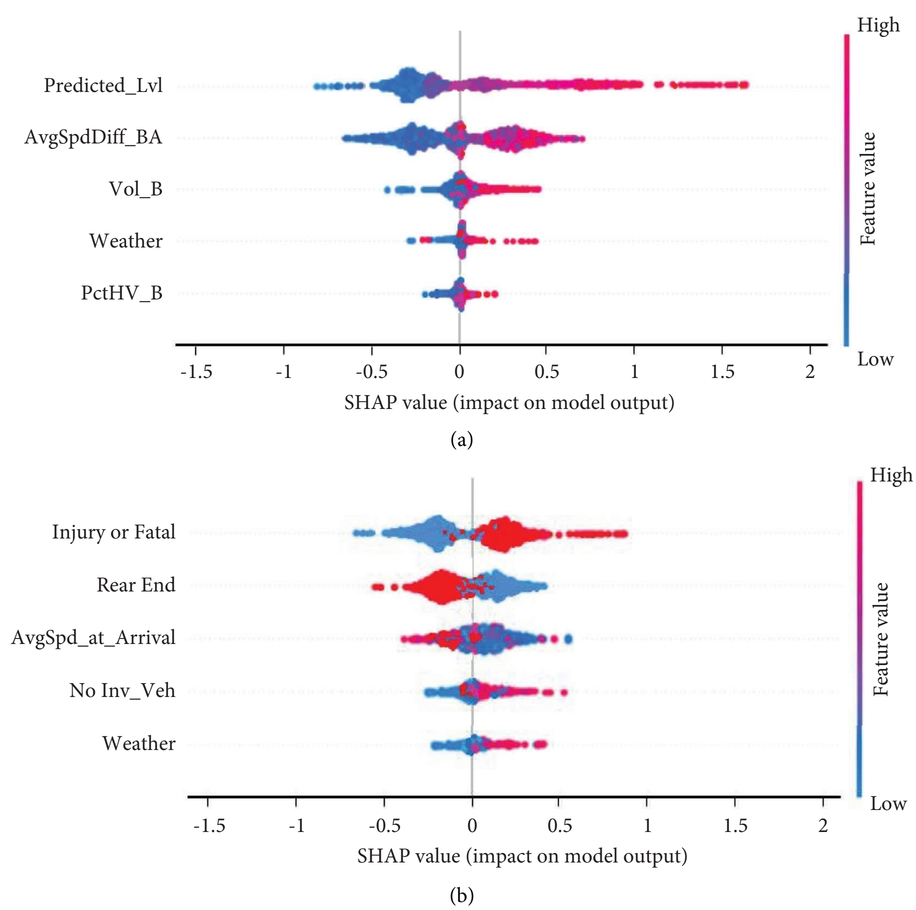

Figure 6: SHAP summary plot of the top five important factors for (a) Model 1.5 and (b) Model 2.

that showed its importance is the ordinal variable of weather. According to the corresponding SHAP values, adverse weather conditions increase the accident duration.

Regarding the SHAP values of Model 2, the occurrence of injury or fatal accidents is the first important variable with a positive contribution to the increase of predicted duration. This finding aligns with previous research [8, 18, 22, 23]. The dummy variable for rear-end accidents is the second and showed that the duration of congestion is short in the case of rear-end collisions. Analysis of our dataset reveals that 83% of rear-end accidents had the severity of property damage only, resulting in faster recovery times than other collision types. The

Table 8: Prediction results of benchmark models.

Method	Dataset	Overall accuracy				[MAPE and RMSE]	
		Model 1.1	Model 1.2	Model 1.3	Model 1.4	Model 1.5	Model 2
Random forest	Train	0.82	0.80	0.75	0.74	[25.18 and 21.08]	[12.49 and 11.26]
	Test	0.77	0.74	0.72	0.67	[30.12 and 26.38]	[18.53 and 13.74]
ANN (multilayer perceptron)	Train	0.79	0.76	0.70	0.68	[26.23 and 25.75]	[16.23 and 14.80]
	Test	0.71	0.70	0.65	0.62	[33.51 and 29.14]	[20.15 and 17.68]
Support vector machine	Train	0.78	0.77	0.71	0.66	[30.44 and 25.62]	[16.44 and 14.39]
	Test	0.71	0.71	0.66	0.61	[35.01 and 30.11]	[20.42 and 17.00]

third important variable is the average speed at the arrival time of officers. Higher values of traffic speed imply that the recovery will be reached sooner. The next variable is the number of vehicles involved in the accident, which has previously been shown to lead to longer accident durations [8, 15, 16].

6. Conclusion and Limitations

Traffic accidents are the leading causes of non-recurrent congestion and can bring economic and environmental losses [1, 2]. Timely forecasting the congestion that could emerge after a crash is of paramount importance. One of the hurdles in developing a post-impact prediction tool is the lack of precise accident information before the arrival of responders and the time for responders to report information once they arrive. This paper presents a sequential modelling framework that promptly predicts accident duration and the most congested traffic level after an accident using real-time traffic and weather variables. Also, with updates in accident-related factors, the accident duration is again estimated. The XGBoost model is applied in the presented framework with SHAP values computed for making interpretations.

Assessment of the models on the test set shows that XGBoost accurately predicted 83% of accident cases that lead to poorer traffic levels. Also, the framework successfully forecasted the exact upcoming traffic level, especially the severely congested level, with reasonable accuracy between 66% and 78%. SHAP value analysis introduces the difference in average speed before and after an accident, traffic volume before an accident, standard deviation of headway after an accident, amount of precipitation, and the percentage of heavy vehicles before an accident as the most important features with higher values positively impacting the probability of having poorer traffic levels. As for the duration prediction models, it is found that receiving accident information boosts MAPE from 22% to 10% and RMSE from 16 min to 9 min on the test set. Nonetheless, the presence of the duration model not employing accident variables is necessary for the sake of timely forecasting. Moreover, SHAP summary plots demonstrated that upon acquiring accident information, traffic variables are replaced by accident factors, and the dummy variable of injury and fatal accident becomes the most important feature with a positive effect on the increase of accident duration. In addition, variables representing rear-end collision type and the number of involved vehicles in the accident are among the top five important features with a negative and positive impact on the accident duration, respectively.

This study has practical implications for both road operators and users who will benefit from the presented framework. The output of models would allow road operators to deploy emergency responses to manage upcoming traffic levels more effectively. Additionally, road users would be informed of any possible changes in the traffic through variable message signs or smartphone applications and consequently take alternate routes.

Notwithstanding, this study has several limitations, and there are suggestions for future research. (1) Because of vast differences in traffic and environmental conditions at different times and regions, the results of this study are limited to rural highways and are not applicable to all roadway types; similarly, the findings are specific to the dataset from Khorasan Razavi province in Iran, and different performance statistics may be computed if the models were rerun in different jurisdictions. Future studies could introduce a universal framework that showcases its transferability to other regions. (2) Data on roadway characteristics and geometric design could enhance the predictive power of the models but were not available for this study. (3) As the average traffic speed is a perceptible variable for road users, it is employed for introducing five traffic levels. However, a possible future avenue is to utilize other criteria, such as speed-flow diagrams. Also, the number of introduced modelling sequences could be increased in future studies. (4) Although real-time factors are deemed readily available in most studies, any malfunction in detectors, which is not a concerning issue in this research, would make them incomplete and unreliable. In these cases, simulated datasets could compensate lack of practical datasets.

Appendix

A. Results of Benchmark Models

In order to show the suitability of XGBoost in our proposed framework, several machine learning methods have been selected as comparison benchmarks. The results are shown in Table 8. Compared to the results of XGBoost in Tables 6 and 7, it can be seen that XGBoost outperforms all these benchmarks in predictive power.

Data Availability

The sample dataset analysed during the current study is available in the GitHub repository at https://github.com/amirhosseinabdi96/post-crash-model-xgboost-shap-. The complete dataset is available from the corresponding author upon reasonable request.

Conflicts of Interest

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

References

[1] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Overview of traffic incident duration analysis and prediction," *European Transport Research Review*, vol. 10, no. 2, p. 22, 2018.
[2] A. Skabardonis, P. Varaiya, and K. F. Petty, "Measuring recurrent and nonrecurrent traffic congestion," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1856, no. 1, pp. 118–124, 2003.
[3] M. W. Adler, J. V. Ommeren, and P. Rietveld, "Road congestion and incident duration," *Economics of Transportation*, vol. 2, no. 4, pp. 109–118, 2013.
[4] H. Al-Najada and I. Mahgoub, "Real-time incident clearance time prediction using traffic data from internet of mobility sensors," in *Proceedings of the 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd intl conf on big data intelligence and computing and cyber science and technology congress(DASC/PiCom/DataCom/CyberSciTech)*, pp. 728–735, Orlando, FL, USA, November 2017.
[5] W. Junhua, C. Haozhe, and Q. Shi, "Estimating freeway incident duration using accelerated failure time modeling," *Safety Science*, vol. 54, pp. 43–50, 2013.
[6] B. Jones, L. Janssen, and F. Mannering, "Analysis of the frequency and duration of freeway accidents in Seattle," *Accident Analysis & Prevention*, vol. 23, no. 4, pp. 239–255, 1991.
[7] D. Nam and F. Mannering, "An exploratory hazard-based analysis of highway incident duration," *Transportation Research Part A: Policy and Practice*, vol. 34, no. 2, pp. 85–102, 2000.
[8] Y. Chung, "Development of an accident duration prediction model on the Korean Freeway Systems," *Accident Analysis & Prevention*, vol. 42, no. 1, pp. 282–289, 2010.
[9] T. F. Golob, W. W. Recker, and J. D. Leonard, "An analysis of the severity and incident duration of truck-involved freeway accidents," *Accident Analysis & Prevention*, vol. 19, no. 5, pp. 375–395, 1987.
[10] Y. Chung and B.-J. Yoon, "Analytical method to estimate accident duration using archived speed profile and its statistical analysis," *KSCE Journal of Civil Engineering*, vol. 16, no. 6, pp. 1064–1070, 2012.
[11] G. Giuliano, "Incident characteristics, frequency, and duration on a high volume urban freeway," *Transportation Research Part A: General*, vol. 23, no. 5, pp. 387–396, 1989.
[12] A. Tavassoli Hojati, L. Ferreira, S. Washington, and P. Charles, "Hazard based models for freeway traffic incident duration," *Accident Analysis & Prevention*, vol. 52, pp. 171–181, 2013.
[13] A. Tavassoli Hojati, L. Ferreira, S. Washington, P. Charles, and A. Shobeirinejad, "Modelling total duration of traffic incidents including incident detection and recovery time," *Accident Analysis & Prevention*, vol. 71, pp. 296–305, 2014.
[14] A. M. S. Alkaabi, D. Dissanayake, and R. Bird, "Analyzing clearance time of urban traffic accidents in abu dhabi, United Arab Emirates, with hazard-based duration modeling method," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2229, no. 1, pp. 46–54, 2011.
[15] A. J. Khattak, J. Liu, B. Wali, X. Li, and M. Ng, "Modeling traffic incident duration using quantile regression," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2554, no. 1, pp. 139–148, 2016.
[16] J. Weng, W. Qiao, X. Qu, and X. Yan, "Cluster-based lognormal distribution model for accident duration," *Transportmetrica: Transportation Science*, vol. 11, no. 4, pp. 345–363, 2015.
[17] A. Garib, A. E. Radwan, and H. Al-Deek, "Estimating magnitude and duration of incident delays," *Journal of Transportation Engineering*, vol. 123, no. 6, pp. 459–466, 1997.
[18] B. N. Araghi, S. Hu, R. Krishnan, M. Bell, and W. Ochieng, "A comparative study of k-NN and hazard-based models for incident duration prediction," in *Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1608–1613, Qingdao, China, October 2014.
[19] R. Li, F. C. Pereira, and M. E. Ben-Akiva, "Competing risk mixture model and text analysis for sequential incident duration prediction," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 74–85, 2015.
[20] C.-H. Wei and Y. Lee, "Sequential forecast of incident duration using Artificial Neural Network models," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 944–954, 2007.
[21] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Text analysis in incident duration prediction," *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 177–192, 2013.
[22] L. Lin, Q. Wang, and A. W. Sadek, "A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations," *Accident Analysis & Prevention*, vol. 91, pp. 114–126, 2016.
[23] X. Ma, C. Ding, S. Luan, Y. Wang, and Y. Wang, "Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 9, pp. 2303–2310, 2017.
[24] Y. Lin and R. Li, "Real-time traffic accidents post-impact prediction: Based on crowdsourcing data," *Accident Analysis and Prevention*, vol. 145, Article ID 105696, 2020.
[25] M. Hossain, M. Abdel-Aty, M. A. Quddus, Y. Muromachi, and S. N. Sadeek, "Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements," *Accident Analysis and Prevention*, vol. 124, pp. 66–84, 2018.
[26] C. Zhan, A. Gan, and M. Hadi, "Prediction of lane clearance time of freeway incidents using the M5P tree algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1549–1557, 2011.
[27] Q. Shang, D. Tan, S. Gao, and L. Feng, "A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis," *Journal of Advanced Transportation*, vol. 2019, p. 11, Article ID 4202735, 2019.
[28] Y. Zou, B. Lin, X. Yang, L. Wu, M. Muneeb Abid, and J. Tang, "Application of the bayesian model averaging in analyzing freeway traffic incident clearance time for emergency management," *Journal of Advanced Transportation*, vol. 2021, p. 9, Article ID 6671983, 2021.
[29] A. Abdi, S. Seyedabrishami, C. Llorca, and A. T. Moreno, "Exploring the effects of stationary camera spots on inferences drawn from real-time crash severity models," *Scientific Reports*, vol. 12, no. 1, Article ID 20321, 2022.
[30] Y. Xing, X. Ban, X. Liu, and Q. Shen, "Large-scale traffic congestion prediction based on the symmetric extreme

learning machine cluster fast learning method," *Symmetry*, vol. 11, no. 6, p. 730, 2019.

[31] S. Wu, Q. Yuan, Z. Yan, and Q. Xu, "Analyzing accident injury severity via an extreme gradient boosting (XGBoost) model," *Journal of Advanced Transportation*, vol. 2021, pp. 1–11, 2021.

[32] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, A. K. Mohammadian, and Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Analysis and Prevention*, vol. 136, Article ID 105405, 2020.

[33] D. A. Rusdah and H. Murfi, "XGBoost in handling missing values for life insurance risk prediction," *SN Applied Sciences*, vol. 2, no. 8, p. 1336, 2020.

[34] T. Chen and C. Guestrin, "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, New York, NY, USA, July 2016.

[35] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, 2001.

[36] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Routledge, Oxfordshire, UK, 2017.

[37] XGBoost, "XGBoost," 2018, https://xgboost.ai/.

[38] V. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in {P}ython," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] S. M. Lundberg and S.-I. Lee, "Unified approach to interpreting model predictions," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017.

[40] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, "Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach," *Engineering Structures*, vol. 219, Article ID 110927, 2020.

[41] J. Wang, W. Xie, B. Liu, S. Fang, and D. R. Ragland, "Identification of freeway secondary accidents with traffic shock wave detected by loop detectors," *Safety Science*, vol. 87, pp. 195–201, 2016.