A FRAMEWORK FOR REAL-TIME CRASH PREDICTION: STATISTICAL APPROACH VERSUS ARTIFICIAL INTELLIGENCE*

by Moinul HOSSAIN** and Yasunori MUROMACHI***

1. Introduction

The attempts to predict crashes on freeways through statistical modeling involving capacity driven measures of traffic flow (e.g., AADT) and road geometry have spanned for more than two decades. However, success in crash prediction involving these static data has so far been limited. In recent times, some researchers made efforts to accommodate the weather conditions and seasonal effects to better predict crashes through time series analysis. So far, these models have shown their incapability to accommodate the ever complex human factors, which, to a great extent, are believed to be directly or indirectly responsible for most of the crash cases. Moreover, the prediction outcome of these models is long-term and mainly used to forecast yearly crash, their seasonal variation, identify black-spots or to assess the impact of road improvement initiatives, etc. However, these models overlook the fact that crash can happen even on geometrically correct roads with excellent weather condition due to human factors. One of the latest additions to the crash prediction modeling is the approach to predict crashes based on real-time traffic data which considers the human factors in macroscopic level from the traffic engineering perspective. The assumption in these models is, instantaneous traffic flow data are indirect representation of the human factors. The concept of real-time crash prediction modeling is gaining momentum due to its proactive nature of application and the growing implementation of ITS, which ensures the future availability of real-time traffic data. However, being at its primitive stage and due to scarcity of past real-time data, the present models are yet theoretical and largely prone to unrealistic data requirements and lack of reliability. Moreover, there has not been any well established norm for modeling approach as well. Some researchers preferred using well known statistical methods while some others opted for artificial intelligence based methods. However, the basic concept in modeling crash in real-time has remained the same, in which, the traffic flow data are separated into two groups – condition leading to crash and normal condition, based on specified assumptions, and then future traffic flow data are evaluated to estimate how likely they are to fit into each of these groups. As crash is considered as a rare event, these models need to have the capability of being frequently updated as soon as new data are available. In future, we can expect that the models will also incorporate non-traffic flow variables to attain higher prediction accuracy. Moreover, the operation of these models is highly dependent on the reliability of the detectors, and hence, is expected to predict even when some data are missing. Considering these requirements, in this paper, we have investigated the applicability of Bayesian Network (BN), a popular real-time prediction method in the field of information science, as a probable method to predict crash in real-time. Bayesian Network is a graphical probabilistic modeling approach which can be calibrated in real-time with limited effort, can handle integration of new variables in future with little effort and also suitable in predicting with missing data. Although its inherent features are suitable for real-time crash prediction, it is important to make sure that BN exhibits satisfactory level of accuracy in predicting crashes, too. For this, after developing the model with BN we developed another model with Binary Logistic Regression and used it as a baseline to investigate the prediction capability of BN.

We have organized the paper into four parts. In the first part, we conducted a literature review on real-time crash prediction models and explain the steps involved in modeling as well as making prediction in practical situation. Afterwards, we provided a short introduction to BN, clarifying the concepts important for the readers for understanding the paper. Then we have developed a prototype real-time crash prediction model with artificially generated data and compared its performance in crash prediction with Binary Logistic Regression. Lastly, we discussed the results, elaborated the limitations of the study and evaluated the possibilities of using BN for real-time crash prediction.

2. Literature Review

The study by Oh *et al.*[1)-3)] was the first of its kind to assess the possibility to classify a traffic condition as crash prone to estimate the likelihood of potential crashes. In that study, they separated traffic dynamics into two categories – disruptive (or hazardous) and normal. A hazardous traffic condition is defined as that potentially leading to a crash occurrence whereas a normal traffic condition does not instigate crashes. Normal condition in these studies was specifically defined as a 5 minute

* Keywords: real-time crash prediction, logistic regression, Bayesian Network

** Graduate Student, M. Eng., Dept. of Built Environment, Tokyo Institute of Technology
(Nagatsuta-machi 4259,Midori-ku, Yokohama, Kanagawa 226-8502, Japan, Telefax: +81-(0)45-924-5524, Email: moinul048i@yahoo.com)
***Member of JSCE, Dr. Eng., Dept. of Built Environment, Tokyo Institute of Technology

period occurring at 30 minutes prior to the crash and the disrupted condition was defined as the 5 minute time period just before the crash. A total of 52 crashes had adequate real-time traffic data to be matched with. Later on, they employed t-test on the mean and deviation of three variables – occupancy, flow and speed, to identify the crash indicators. However, there was no suggestion explaining if they have tested the data for normality as t-test is applicable only with the assumption that the data follow normal distribution. Oh *et al.*[2] identified standard deviation of speed to be the most significant variable although most of the variables came out as significant in the t-test. Later, they defined the two traffic conditions with two probability density functions (PDF) and used those to identify the posterior probability of a traffic condition to belong to either of these two traffic conditions and determine crash probability thereby. In their latest study (Oh *et al.*[3]) they also used average of occupancy as a predictor. In their second study (Oh *et al.*[2]) they employed a nonparametric Bayesian approach to identify the real-time crash likelihood. In the latest study (Oh *et al.* [3]) they applied Probabilistic Neural Network (PNN) to accomplish their estimation purpose. Lee *et al.*[4], [5] conducted two studies to predict crash risk in real time where the second study basically reduced the number of assumptions made in the first study to make it more acceptable. In their latest model they selected speed variations along a lane, traffic queue and traffic density at given road geometry, weather condition and time of the day as predictors and applied aggregated first order log-linear model to predict crash. The initial study by Abdel-Aty *et al.*[6] concentrated on classifying speed patterns to predict crashes in real-time. Their later study (Abdel-Aty *et al.*[7]) incorporated road geometry into the model using Generalized Estimating Equations for correlated data. This research was the first of its kind where a series of crash and non-crash traffic classification analysis were conducted to identify the predictors. They used Probabilistic Neural Network method to separate traffic patterns leading to crash from those not leading to crash. The variables found significant in their model were mean and variance of volume, occupancy and speed. Later on, they employed Generalized Estimating Equations to predict crashes. The study further calculated the false alarm rate, too. It was suspected that Abdel-Aty *et al.*'s [6], [7] model may generate high false alarm as they considered the traffic variables independently ignoring their interactions. The research project by Luo and Garber [8] is the latest addition in the attempt to predict crashes in real-time. They commenced their research with intensive investigation of previous research studies. Luo and Garber [8] analyzed each crash case independently to identify crash leading patterns and factors describing the patterns. Like Abdel-Aty *et al.* [6], [7], they also suggested mean and variance of occupancy, speed and volume to be suitable for predicting crashes in real-time. However, they limited their study within identifying traffic patterns distinguishing crash and non-crash situations. They applied three pattern recognition methods: K-means clustering method, Naive-Bayes method and Discriminant Analysis and compared the outcomes. Apart from Naive-Bayes method, no other method in their study reached an overall 50% prediction success. We have summarized the overall modeling approach and the hypothetical work-flow of a real-time crash prediction model in Figure1.

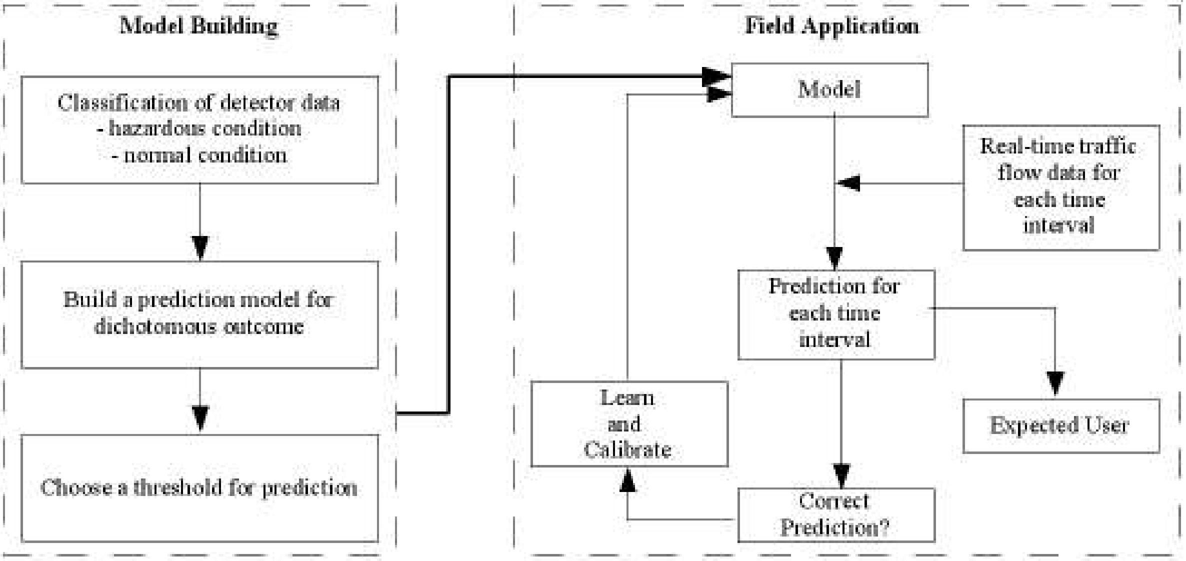

Model Building
Classification of detector data
- hazardous condition
- normal condition
Build a prediction model for dichotomous outcome
Choose a threshold for prediction
Field Application
Model
Real-time traffic flow data for each time interval
Prediction for each time interval
Expected User
Learn and Calibrate
Correct Prediction?

Figure 1: Modeling approach and work-flow of a real-time crash prediction model

In all of the aforementioned prominent studies on real-time crash prediction, they used statistical measures of one or more of the traffic flow variables – speed, flow and occupancy, in real-time. They had high similarity in overall modeling approach, too. All these studies were conducted on freeways and most of them clearly fixed the section of the road under consideration for study to eliminate the influence of varying road geometry. All the studies followed the norms of Oh *et al.*[1)-3)] by separating the traffic condition into two parts – leading to crash and normal. However, they had differences in defining, specially the normal traffic condition. In case of Oh *et al.*[1)-3)], they defined normal traffic condition as a 5 minute traffic flow data collected 30 minutes prior to crash on the same day when the crash took place. However, Abdel-Aty *et al.*'s [6], [7] argued that this classification of normal traffic condition may be not be justifiable and assumed that traffic condition for a specific time of a day remain similar to that of same day, same time for different weeks of the year. They defined the normal traffic condition as a 5 minute traffic flow data for the same time period for the same days through out the year. Apart from this, the main major difference among these proposed models have been the method of modeling. Table 1 summarizes our findings from the literature review.

Table 1: Real-time crash prediction models at a glance

Study	Motivation	Sample Size	Variable	Method	Evaluation	Outcome
Oh et al. (3)	Proactive Road Safety	52	Standard deviation of speed, average occupancy	Probabilistic Neural Network	random sampling from the model building data	Conclusive
Lee et al. (5)	Proactive Road Safety	234	standard deviation of speed, difference is speed between upstream & downstream, density	1st order log-linear Models	Model Statistics	Conclusive
Abdel-Aty et al. (6)	Proactive Road Safety	149	variance of speed and volume, mean of speed	Probabilistic Neural Network	49 random samples mutually exclusive from the model building data	Conclusive
Luo et al. (8)	Proactive Road Safety & Crash Phenomena	391	various combination of descriptive statistics of speed, flow and occupancy	K-cluster mean, Naïve Bayes, Discriminant Analysis	N/A	Inconclusive

As the idea of predicting crash in real-time is in its infancy, neither of the studies recommended their models to be directly used for practical situation. They emphasized that due to lack of data, it is difficult to develop the initial model with acceptable accuracy and these models need to be calibrated frequently with new crash and non-crash data. They also mentioned that in future the generalized models will be required to have capabilities to incorporate variables related to road and environment. In fact, Abdel-Aty *et al.*[7)] developed their latest model incorporating weather and road geometry related variables into their previously developed model. The studies also suggested that a practical real-time crash prediction model must not be susceptible to missing data as detectors may often fail to yield all the data for each time interval. However, so far the focus of the previous studies have been on developing or improving the framework and choosing the method to improve the crash prediction capability rather than the model's suitability for practical use. In this study, we bridge this gap by choosing BN as the modeling method, which is widely known for its flexibility in future calibration, present data requirements and features to incorporate new variables easily in future. However, as prediction capability is another indispensable requirement, we investigate if BN can yield satisfactory prediction accuracy by comparing its performance with Binary Logistic Regression.

3. Fundamentals of Bayesian Network (BN)

Bayesian Network (BN) can be defined as an acyclic directed graph (DAG) which defines a factorization of a joint probability distribution over the variables that are presented by the nodes of the DAG, where the factorization is given by the directed links of the DAG[9)]. BN is a graphical modeling method and it is presented with a graph and a basic equation. The graph consists of two parts as well – the nodes and the arcs. The variables in a BN are represented as nodes and their inter-relationship is represented with the arcs. For example, in Figure 2, a problem domain is represented with five variables – A, B, C, D and E, where B is caused by D, E is caused by B and C is caused by A and B together. Thus, A and B are the parent nodes of C and C is their common child node, D is the parent node of B and B is its child node and so on.

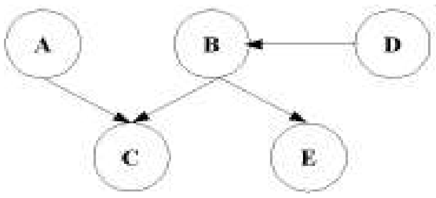


Figure 2: A Simple Bayesian Network

The inter-relationships of the variables are represented by drawing arcs from the parent nodes to the child node. If a BN contains 'n' number of variables, then we can represent the complete problem domain as Equation 1.

$$P(x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n} P(x_i|pa(x_i)) \quad (1)$$

In order to specify a BN, we need to provide the probability (or conditional probability) distributions of all the nodes. Thus, as we have five variables in our example, if each variable is dichotomous, i.e., had only two states then we are expected to need 2^5-1, or, 31 joint probabilities. However, the architecture of BN suggest that each network can be presented with the probabilities of each node who do not have any parent node and the conditional probabilities of the child nodes with respect to their immediate parent nodes only. Thus, Equation 1 can be re-written as Equation 2 for our example.

$$P(A,B,C,D,E)=P(A)P(D)P(E|B)P(B|D)P(C|A,B) \quad (2)$$

Now, in future, if we think that only the behavior of variable B and D has changed that may require the model to be calibrated, we need to collect data for only P(B), P(D) and P(B|D) to recalibrate the whole model. Thus, the same model can be built with data of different time period. Moreover, in course of time, if we find that it is necessary to incorporate a new variable F which influences both A and B, then we can update the model by modifying the graph as shown in Figure 3 and calculate the probabilities following Equation 3. Thus, we only need to re-construct the probability tables for P(F), P(A|F) and P(B|D,F).

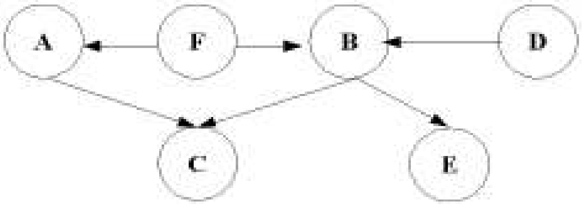


Figure 3: Incorporating New Variables in Existing Bayesian Network

$$P(A,B,C,D,E,F)=P(F)P(D)P(A|F)P(E|B)P(B|D,F)P(C|A,B) \quad (3)$$

Lastly, let us assume that each of these variables has two states and they are represented with small letters (e.g., state of A as a_1 and a_2). Now, we would like to predict the probability of C being in c_2 state when we have information about only A D and E ($A = a_1$, $D = d_2$ and $E = e_1$). This can be obtained by using Baye's theorem by marginalizing the variables as presented with Equation 4.

$$P(C=c_2|A=a_1,D=d_2,E=e_1)=\frac{\sum_{B}\sum_{F} P(A=a_1,B,C=c_2,D=d_2,E=e_1,F)}{\sum_{B}\sum_{C}\sum_{F} P(A=a_1,B,C,D=d_2,E=e_1,F)} \quad (4)$$

Thus, BN has inherent capability to update the model in real-time, revise the model with partial new data, incorporate new variables by updating the probability tables of other variables directly connected to the new variables and make inferences even when information regarding all the variables are not available; which are the major qualities required in a suitable method for real-time crash prediction. In the next section, we have investigated if BN can produce satisfactory results in case of crash prediction as compared to the well known method to predict dichotomous outcome – the Binary Logistic Regression method.

4. Model Building and Performance Evaluation

The model building process consists of four major steps. At first, we artificially generated a dataset for crash prone and normal traffic condition, then we separately built two models, one with BN and the other with Binary Logistic Regression, then we evaluated their performance separately and lastly, we tested if BN could predict crashes as well as the model with Binary Logistic Regression. The work flow followed to compare Logistic Regression and Bayesian Network in this study is presented in Figure 4. The study is a framework study by nature and the data used are statistically simulated. It is understandable that the data may not represent real world situation properly. Thus, the study is not concerned with the phenomena of crash prone and normal traffic conditions, rather, the study asks the question, 'If the data are like this, then can BN predict crashes as well as the widely accepted Binary Logistic Regression?'. Being a real-time prediction model, BN has advantages over Binary Logistic Regression in suitability for real-time crash prediction. This section investigates if its prediction capability is also acceptable considering Binary Logistic Regression as the bench mark.

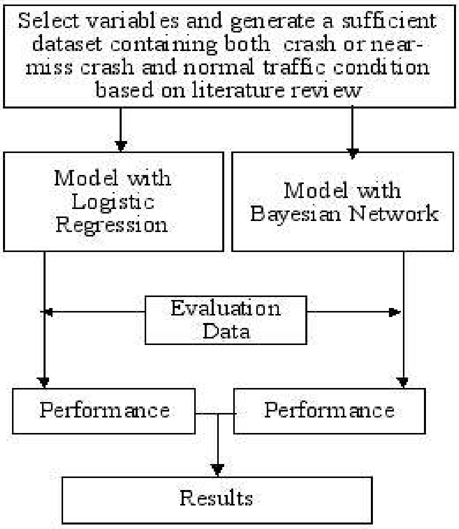

Select variables and generate a sufficient dataset containing both crash or near-miss crash and normal traffic condition based on literature review
Model with Logistic Regression
Model with Bayesian Network
Evaluation Data
Performance
Performance
Results

Figure 4: Work flow of the study

(1) Data Preparation

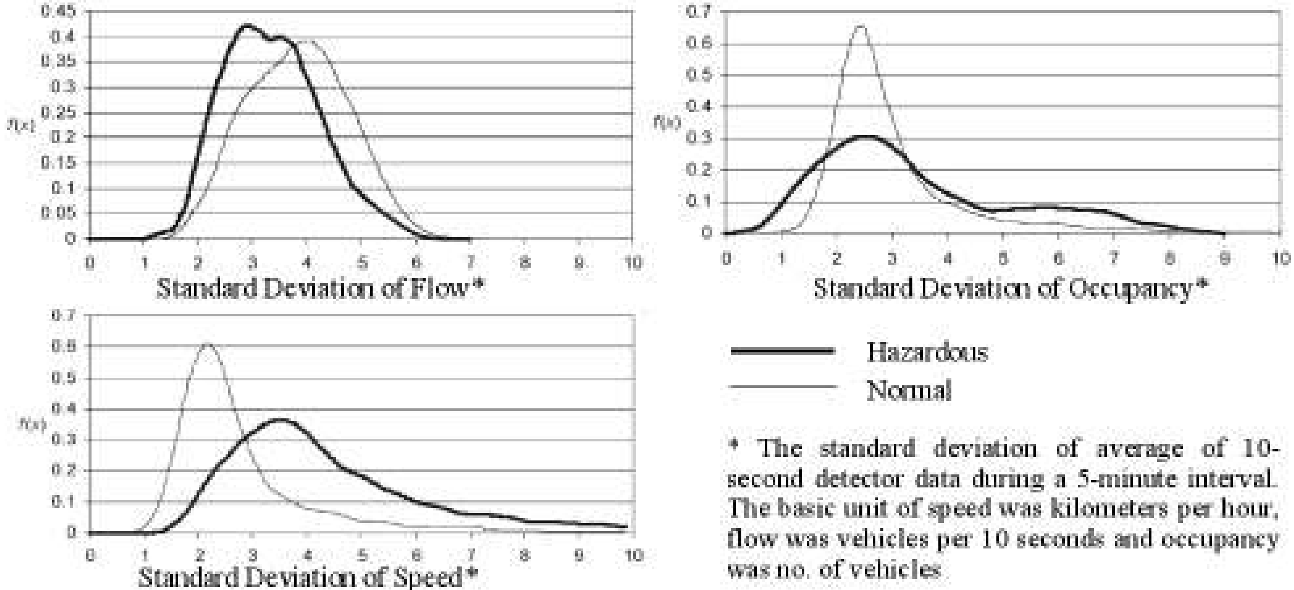
In this study, we have used artificially generated data using the descriptive statistics of real-time traffic flow data provided by Oh *et al.* [2,3]. In their studies, they obtained data from loop detectors installed on a 15.3 kilometers four to five lane straight section of the I-880 freeway in Hayward, California from February 16 through March 19, 1993 which was stored by the authority. They eliminated the impact of road geometry by choosing the section with uniform road geometry. They focused on the data during the peak period (5 am to 10 am and 2 p.m. to 7 p.m.) as both possibilities and consequences of road crashes during the peak period are the highest. For the model building purpose, the separated the data into two parts – disrupted or hazardous traffic condition and normal traffic condition. The defined normal traffic condition as a 5 minute period, 30 minute prior a crash and disrupted traffic condition as a 5 minute period just prior crash. Later, they aggregated the detector data for each 10 seconds and then calculated the mean and the standard deviation of aggregated speed, flow and occupancy for the 5 minute period to obtain the model predictors. However, the found significant pattern difference between hazardous and normal traffic conditions only in case of the standard deviation of the aggregated data. In their study, they provide these descriptive statistics of standard deviation of speed, flow and occupancy of 10 minute aggregated data for the 5 minute period accompanied with their plots of their distributions (see Figure 5) as well as the results of t-test (see Table 2).

Table 2: Descriptive statistics of standard deviation of speed, flow and occupancy

Category	10 second detector data aggregated for 5 minutes					
	Std. Dev. of Speed*		Std. Dev. of Flow*		Std. Dev. of Occupancy*	
	Disrupted	Normal	Disrupted	Normal	Disrupted	Normal
Mean	3.66	2.77	3.84	3.37	3.46	3.08
Std. Dev.	1.52	1.3	0.9	0.83	1.46	1.72
t-stat.	4.94		-3.68		1.87	

* see the explanation section within Figure 5 for the units of the variables

Considering the other previous research studies on real-time crash prediction, it can be understood that the most significant variables had been the aggregated descriptive statistics (hear mean and standard deviation) form of 'Flow', 'Speed' and 'Occupancy' – which can be easily obtained from the basic outputs of most of the real-time traffic data collection equipments. Moreover, these three variables are considered sufficient to explain the traffic condition on roads as well. From Table 2 and Figure 5, we can understand that the standard deviation of the aggregated speed, flow and occupancy data roughly follow a bell shaped curve for both disrupted and normal conditions. Thus, for our study, we used these statistical values to generate our datasets.



* The standard deviation of average of 10-second detector data during a 5-minute interval. The basic unit of speed was kilometers per hour, flow was vehicles per 10 seconds and occupancy was no. of vehicles

Figure 5: Distribution of Standard Deviation of Speed, Flow and Occupancy of Oh *et al.* [1-3]'s Data
Source: Oh *et al.* [3]

We used built in function of the R-Package (Dalgaard[10]) to simulate the data and assuming normal distributions using the parameters in Table 2. The simulation time period was one year – 12 hours a day, 5 days a week and 52 weeks a year. The data were then aggregated for each 5 minute period (following Oh *et al.* [2,3]) and we assumed that disrupted traffic conditions are best defined by the 5 minute traffic condition prior to a crash and the rest of the time periods are normal traffic conditions. In case of crashes, we assumed that 10 crashes occur every week day in the simulated study section (can be assumed as a relatively long section) through out the year. Thus, we obtained 2600 (10 X 5 X 52 = 2600), 5 minute aggregated data points of standard deviation of speed, flow and occupancy for traffic conditions causing crashes or near miss crashes. Similarly, we simulated 34840 (12 X 12 X 5 X 52 – 2600 = 34840) data points for the normal traffic condition. Lastly, we assigned a value of '1' with each data point of the crash or near miss crash dataset and '0' for the normal traffic condition dataset. Figure 6 presents the box plots of the three variables (explanation provided within the figure): 5 minute standard deviation of speed, flow and occupancy. Here '1' represents disrupted traffic condition and '2' represents normal traffic condition. The boxes in the figure represents the inter quartile distance and the middle bold horizontal like represents the median and the cross represents the mean value of the generated data. Comparing Table 2 and Figure 6, we can understand that the artificially generated dataset follows normal distribution with means close to the corresponding values mentioned in Table 2.

(2) Modeling with Bayesian Network

The first step in BN modeling involves identifying a hypothesis variable where each of its states (mutually exclusive and collectively exhaustive) will be defined. The next step involves identification of the predictors for the hypothesis variable, normally referred as information variable. In this study, the hypothesis variable is 'Crash' with two states – 'Yes' and 'No', where ‘Yes’ represents both crash and near miss crash situations. This is a dichotomous variable providing information regarding the instantaneous crash risk based on predictor data. Here, the previously mentioned standard deviation of 10 second aggregated values of speed, flow and occupancy for the duration of 5 minutes are chosen as the information variables and are represented as SSD5, FSD5 and OSD5. The next step requires finalizing the graph where the information variables are linked with the hypothesis variables – directly or indirectly. In our case, there can be two potential possibilities for the Bayesian Network as presented in Figure 7 (a) and (b). We did not find high correlation among the information variables and therefore, did not provide any direct links among them. In Figure 7, SSD5, FSD5 and OSD5 are called the parent nodes of ‘Crash’ connected with a converging connection (a) or child node of 'Crash' connected with a diverging connection (b). The converging connection in this situation has potential advantage over the diverging connection as it facilitates flow of inference among the information variables when hard or soft evidence about ‘Crash’ is available. It means that in case of Figure 7(b), if information regarding whether a crash or near-miss crash has taken place is known, then information regarding SSD5 does not have any impact on our belief regarding FSD5 or OSD5 and vise versa. However, the model in Figure 7(a) can be used to understand the inter-relationship among all these variables during crash. To elaborate, any information about ‘Crash’ and any of the information variables will have influence on the belief regarding the other two information variables. This characteristic of Bayesian Network is termed as the ‘explaining away effect’ or sometimes the ‘bi-directional inference’.

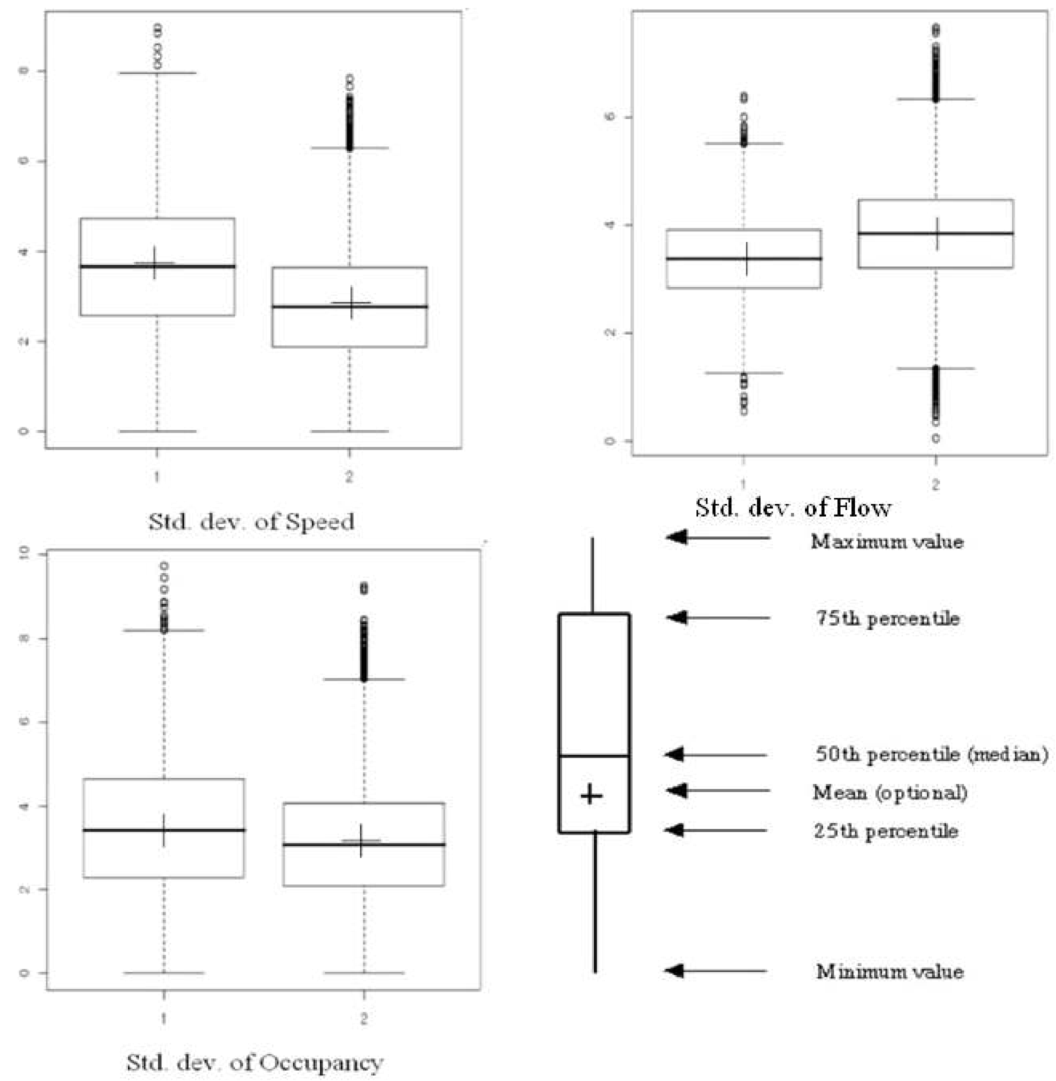
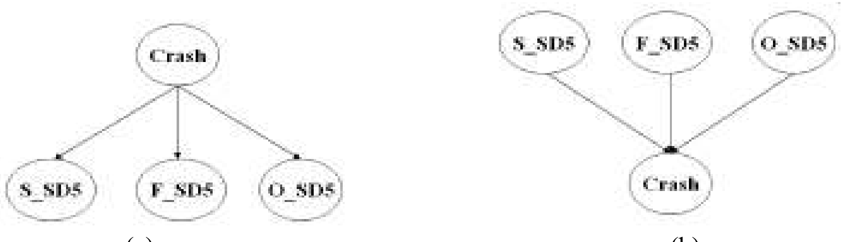


Figure 6: Box plot of statistically simulated data (1=Disrupted, 2=Normal traffic condition)
(Units are as explained in Figure 5)



(a) (b)

Figure 7: Potential Bayesian Networks

Now, using the universal probability distribution of Bayesian Network (see Equation 1), the probability of any state of any variable can be calculated from the universe of probability distribution P(U) = P(Crash, SSD5, FSD5, OSD5) as presented in Equation (5).

$$P(Crash, SSD5, FSD5, OSD5) = P(SSD5)P(FSD5)P(OSD5)P(Crash|SSD5, FSD5, OSD5) \quad (5)$$

In the next step, we create the probability tables for each node in the BN. In our model, data for each information variable (standard deviation of speed, flow and occupancy) were classified into smaller groups and their frequency distribution was calculated as presented in Table 3. This was done to facilitate producing the probability distribution tables for the information variables. Now a days, BN can handle continuous data, however, it requires a substantial amount of time for modeling, which can be saved through careful discritization of the continuous data into smaller groups. The classification process went through rigorous trials in order to find separate suitable class intervals for each of the information variables which will not compromise with the accuracy of the model but reduce the number of categories for each variable and thus reduce the calculation time.

Table 3: Prior probability distribution

5 minute aggregated standard deviation of					
Speed		**Flow**		**Occupancy**	
Category	**P(Speed)**	**Category**	**P(Flow)**	**Category**	**P(Occupancy)**
<=1	0.08	<=2.5	0.08	<=1	0.08
1.5	0.08	3	0.12	1.5	0.06
2	0.11	3.5	0.18	2	0.09
2.5	0.14	4	0.21	2.5	0.11
3	0.15	4.5	0.19	3	0.13
3.5	0.14	5	0.13	3.5	0.13
4	0.12	5+	0.1	4	0.12
4.5	0.08			4.5	0.1
5	0.05			5	0.07
5+	0.05			5+	0.1

Afterwards, the process in BN updates the conditional probability tables of each child node by calculating their posterior probabilities based on the information on the prior probabilities of the parents and the node itself using the well known Baye's theorem. This process is complex and the calculation is time consuming. For this, we used limited version of Netica (http://www.norsys.com/), a popular commercial package to develop models with Bayesian Network, to build our prototype the real-time crash prediction model. For this, we connected the three information variables using a converging connection to the hypothesis variable named 'Crash'. This process proceeded by entering the prior probabilities of the information variables and the conditional probability tables for the variable 'Crash' (i.e., probability of crash based on different combinations of standard deviation of speed, flow and occupancy; P(Crash|SSD5, FSD5, OSD5) and building of the model. The prior probability of each variable when no prior information regarding any variable is available is presented in Figure 8. The outcome of the Bayesian Network suggests that based on the data, the average risk of crash or near miss crash for the road section is 6.89%, i.e., when no other information is available, there is a 6.89% chance that a crash or a near miss crash can occur at the road section during peak hour of a weekday.

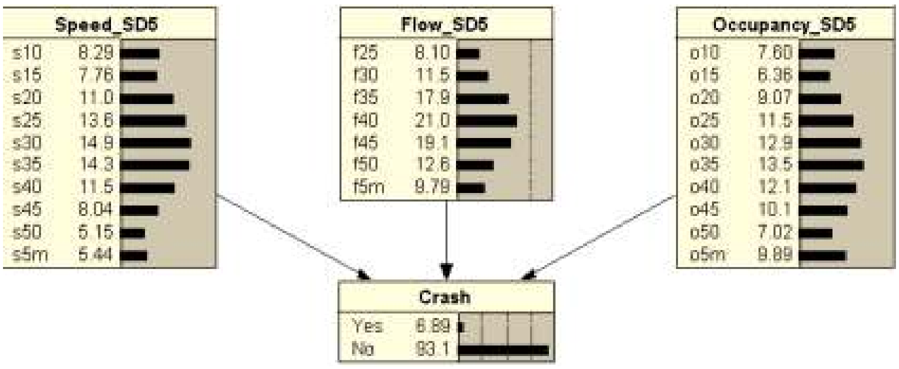


Figure 8: Bayesian Network Model

As mentioned in Section 3, the probability of the universe in a Bayesian Network can be represented only with the parent nodes and the conditional probability of the child nodes given their immediate parent nodes. Thus, when we obtain new findings e = {S_SD5, F_SD5, O_SD5} in real-time through traffic data sensing equipment, we can calculate the probability of crash using Bayes rule as shown in Equation (6).

$$P(U|e)=P(U,e)/P(e) \tag{6}$$

Here, we can calculate P(e) by marginalization of the universal probability distribution P(U) (see Equation 4, too). This empowers BN with the ability to coordinate bi-directional inferences, which is difficult to achieve with conventional statistical approaches. For example, in this prototype model we have used the BN to predict the probability of crashes based on available data. However, the same model can be used to predict the probability of SSD5 or FSD5 or OSD5 when data about crash and one of these three information variables become available. Thus, the same model can be used to understand what kinds of combinations of standard deviations of speed, flow and occupancy have high likelihood to cause road crashes.

(3) Modeling with Logistic Regression

As we have already separated the artificially generated data into two categories (crash or near miss crash condition and normal condition), we could build a Binary Logistic Regression model based on that. We used R-package for this purpose, too. The results are as presented in Table 4.

Table 4: Results of the logistic regression model

	(intercept)	**Std. Dev. of Occupancy**	**Std. dev. of Flow**	**Std. dev. of Speed**
Coefficient	-2.89958	0.1797	-0.53475	0.51098
Pr(>|z|)	<2e - 16 ***	<2e - 16 ***	<2e - 16 ***	<2e - 16 ***
Sig. Code	*** 0.001, ** 0.01, * 0.05			
Deviance	Null: 18885 on 37439 degrees of freedom		Residual: 17073 on 37436 degrees of freedom	

(4) Performance Evaluation

The last step of the study we conducted a random sampling of 50 data points, each from both the simulated data points of normal and disrupted traffic condition and tested the results with the newly developed models following previous studies that compared Logistic Regression and Bayesian Network [11)]. For this, we calculated the probability of crash for each data point and classified it as crash if the value is over 6.98%, which is the average crash risk for the simulated data (see Figure 6). In real life scenario, the choice of the threshold will be a more complex activity as measures involved with reducing the crash risk is cost dependent and the authorities need to judge the threshold value after carefully conducting the cost-benefit analysis. However, in our case, the concern is to assess if BN can predict crash properly by taking the prediction capability of the Binary Logistic Regression model as the standard. The outcome of the validation process is presented in Table 5. From this, it can be realized that both the developed models show similar level of accuracy in predicting non-crash situation whereas the BN based model was 18% more successful (Binary Logistic Regression and BN successfully predicted 37 and 28 crash cases respectively out of 50 evaluation samples) in identifying crash or near-miss crash situations than the Logistic Regression based model. We did not conduct any further statistical test as our interest was to investigate if BN can predict as closely as the Binary Logistic Regression model or not.

Table 5: Performance evaluation

		Predicted			
		Bayesian Network		**Logistic Regression**	
		Crash	**Non-Crash**	**Crash**	**Non-Crash**
Actual	**Crash**	37	13	28	22
	Non-Crash	14	36	16	34

5. Discussion and Conclusion

The idea of predicting crash in real-time is very new and it is still in the conceptual level. Its modeling methods have some additional special requirements due to resource constrains. For example, an applicable model requires to be developed based on sufficient amount of data, however, crash is a rare event and in many occasions, it is difficult to obtain crash data of good quality and quantity. Moreover, real-time crash prediction models require reliable real-time traffic flow data, which were hard to obtain until recently. At times, the model can get erroneous if the reported crash time does not match with the actual crash time. For this, these models should be built with high flexibility where the model can be calibrated in regular interval as new data become available in course of time. Moreover, most of the basic real-time models developed so far do not consider the impact of weather and eliminate the impact of road geometry by choosing homogeneous road sections. For a practical model, in future, it is expected to have capabilities to incorporate new predictors into the model without needing to collect data for all the predictors. In many cases, the time of data collection for different predictors may be different, too. Apart from these, the model needs to be capable of making inference with partial data as many times detectors fail to provide reliable data for all the variables. The model also needs to predict fast to provide buffer time to take necessary actions to acerbate the risk of crash. Thus, the requirements of a real-time crash prediction modeling method can be classified into three broad groups – a) flexibility in model calibration (both with the availability of new data and new predictors) b) speed in prediction c) accuracy of the model. However, the current attempts on real-time crash prediction are highly concentrated on how to increase the prediction capability, ignoring the former two requirements. In this paper, we have emphasized that an actionable real-time crash prediction model needs to concentrate on settling for a method that satisfies all these three needs rather than the accuracy of the model only. In this paper, after explaining the current developments and requirements for a real-time crash prediction model (see Section 2), we have explained the inherent capability of BN in updating the model in real-time, incorporating new variables, making inferences with minimum calculation; which makes it a highly suitable method for real-time prediction. Apart from these, BN has some other extra advantages over many of its competing statistical models. For example, BN is a non-parametric modeling approach and we do not need to store previous data for each time we update a BN, rather, we just store the probability tables of each node and update the probabilities with the new data using Baye's theorem. However, in case of statistical models like Binary Logistic Regression, we either need to store the previous data or need to apply complex Bayesian statistics based approaches to update the models regularly, which can be substantially resource hungry.

However, even after having these capabilities, it was important to ensure that BN can also conduct predictions with acceptable level of accuracy. For this, we have chosen the prediction capability of Binary Logistic Regression, a well known method for predicting dichotomous outcomes in the field of statistics as well as transportation engineering, as the bench mark and evaluated the prediction capability of BN thereby. For this, we used simulated data from previous studies (Oh *et al.* [2,3)]). It can be understood that the simulated data may not represent an actual situation properly; however, our interest was confined within evaluating the prediction performance of BN with respected Binary Logistic Regression, assuming that the simulated data are a representation of actual real-time traffic flow data. The outcome of the study suggests that Bayesian Network based model could predict crash situations with 18% higher accuracy than the Binary Logistic Regression based model for the simulated data, although, the performances to predict non-crash prone situations were relatively identical (68% for Logistic Regression and 72% for BN). Real-time crash prediction is a new branch of proactive road safety management systems with high promise. So far, the research development in this field is limited within the process of establishing a framework and none of the previous models have been implemented in real-life situation. This paper is a part of an ongoing research study on the development of a real-time crash prediction model for urban expressways of Japan, which is being carried out in Tokyo Institute of Technology, Japan. At present, using the findings of this paper, we have developed a basic real-time crash prediction model using Bayesian Network for Japanese urban expressways using real-time detector data and expect to excel the study to develop a model that can be implemented in practical situation in near future.

References

1) Oh, C., Oh, J., Ritchie, S., and Chang, M. Real time estimation of freeway accident likelihood. 80th Annual Meeting of Transportation Research Board, 2001.

2) Oh, J., Oh, C., Ritchie, S., and Chang, M. Real time estimation of accident likelihood for safety enhancement. ASCE Journal of Transportation Engineering, Vol. 131, No. 5, pp 358-363, 2005.

3) Oh, C., Oh, J., Ritchie, S., and Chang, M. Real time hazardous traffic condition warning system: framework and evaluation, IEEE Journal of Intelligent Transportation Systems, Vol.6. No.3, pp 265-272,2005.

4) Lee. C., Saccomanno, F., and Hellinga, B. Analysis of crash precursors on instrumented freeways. Journal of Transportation Research Board, No. 1784, 2002.

5) Lee. C., Saccomanno, F., and Hellinga, B. Real-time crash prediction model for the application to crash prevention in freeway traffic. Journal of Transportation Research Board, No.1840, 2003.

6) Abdel-Aty, M. and Pande, A. classification of real-time traffic speed patterns to prediction crashes on freeways. Preprint No. TRB 04-2635 83rd Annual Meeting of Transportation Research Board, 2003.

7) Abdel-Aty, M., and Abdalla, M. F. Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes using generalized estimating equations for correlated data. Journal of Transportation Research Board, No.1897, 2003.

8) Luo, L., and Garber, N. J. Freeway Crash Prediction Based on Real-Time Pattern Changes in Traffic Flow Characteristics. Project Report for the ITS Implementation Center, UVA Center for Transportation Studies. Research Report No. UVACTS-15-0-101, 2006.

9) Jensen, F. V. and Nielsen, T. D. Bayesian network and decision graphs.2nd Ed., Springer, New York, USA, 2007.

10) Dalgaard, P. Introductory statistics with R. 2nd Ed., Springer, New York, USA, 2008.

11) Chhatwal, J., Alagoz, O., Kahn, C. E. Jr., and Burnside, E. S. Bayesian network versus logistic regression model for computer aided diagnosis of breast cancer. Proceedings of the 28th Annual Meeting of the Society for Medical Decision Making, 2006, USA.

A Framework for Real-time Crash Prediction: Statistical Approach versus Artificial Intelligence*

by Moinul HOSSAIN** and Yasunori MUROMACHI**

This paper evaluates the possibility of Bayesian Network to predict road crash risks in real-time. Two data sets, one for normal traffic conditions and another for conditions leading to crash were statistically generated based on previous studies. Two separate crash prediction models were developed, (logistic regression and Bayesian Network) followed by their performance evaluation by randomly drawing 50 samples and comparing their computed and actual outcome. The results suggest similar prediction success for non-crash situations, whereas, the model based on Bayesian Network predicted crash-prone conditions 18% more accurately than the Logistic Regression based model.