Assessment of freeway traffic parameters leading to lane-change related collisions

Anurag Pande, Mohamed Abdel-Aty

Abstract

This study aims at 'predicting' the occurrence of lane-change related freeway crashes using the traffic surveillance data collected from a pair of dual loop detectors. The approach adopted here involves developing classification models using the historical crash data and corresponding information on real-time traffic parameters obtained from loop detectors. The historical crash and loop detector data to calibrate the neural network models (corresponding to crash and non-crash cases to set up a binary classification problem) were collected from the Interstate-4 corridor in Orlando (FL) metropolitan area. Through a careful examination of crash data, it was concluded that all sideswipe collisions and the angle crashes that occur on the inner lanes (left most and center lanes) of the freeway may be attributed to lane-changing maneuvers. These crashes are referred to as lane-change related crashes in this study. The factors explored as independent variables include the parameters formulated to capture the overall measure of lane-changing and between-lane variations of speed, volume and occupancy at the station located upstream of crash locations. Classification tree based variable selection procedure showed that average speeds upstream and downstream of crash location, difference in occupancy on adjacent lanes and standard deviation of volume and speed downstream of the crash location were found to be significantly associated with the binary variable (crash versus non-crash). The classification models based on data mining approach achieved satisfactory classification accuracy over the validation dataset. The results indicate that these models may be applied for identifying real-time traffic conditions prone to lane-change related crashes.

1. Background

Real-time assessment of crash risk on the freeways has recently received much attention. This is a diversion from the past when the research in traffic management was focused on incident detection algorithms. Recent technological advances have brought about this change. Not only have the increased usage of cell-phones rendered the incident detection somewhat irrelevant, the enhancements in data collection, storing, and analysis capabilities have encouraged the traffic management authorities to look into proactive safety strategies. Real-time identification of crash prone conditions on freeways would be the first step towards proactive traffic management. It requires establishing relationship(s) between historical crash occurrences and the loop data recorded at stations surrounding the crash locations (just prior to crash occurrence). The basic premise is that these relationships may be used to 'predict' crashes by monitoring the surveillance data in real-time.

Such relationships have so far been explored to develop generic crash 'prediction' models, i.e., single generic model was adopted to identify all crashes (such as rear-end, sideswipe, or angle). These models were proposed by Abdel-Aty et al. (2004), Abdel-Aty and Pande (2005), Lee et al. (2002, 2003), and Oh et al. (2001). These studies employed interesting methodologies for analyzing the crash and loop detector data, i.e., matched case–control logistic regression, probabilistic neural network (PNN), log–linear model, and Bayesian classifier, respectively. However, the conditions preceding crashes are expected to differ by type of crash and therefore the approach towards proactive traffic management should be type (of crash) specific in nature.

The only conceivable reason for the generic nature of the models developed in these studies was that the crashes are rare events and until sufficient effort has been devoted to data collection and preparation, the sample size would not be sufficient

for disaggregating crash data by type. This is especially true for the crashes that are not as frequent as the rear-end crashes. The majority of crashes on freeways are rear-end collisions and tend to dominate the sample of the crashes used for developing the generic models. Thus, the real-time traffic parameters identified as indicative of crash prone conditions on freeways through these generic models can by in large be associated with rear-end crashes. In fact, the list of traffic parameters found significantly associated with rear-end crash occurrence (Pande and Abdel-Aty, 2006) included the variables constituting the generic logistic regression model developed by Abdel-Aty et al. (2004).

While the cause for this 'bias' (i.e., high frequency of rear-end crashes) in the generic models may also be the justification for it; other types of crashes (e.g., sideswipe or angle crashes) also occur on the freeway in significant numbers. To identify the real-time traffic conditions associated with crashes other than rear-ends, the data must be segregated by type of crash. The other advantage of the models developed using segregated crash data would be that the outcome of these models may help with the application of specific countermeasures, e.g., the application of variable speed limits for rear-end crashes or a temporary "no lane-changing" sign to avoid an impending sideswipe crash.

In this regard, Golob and Recker (2004) did assemble data for more than 1000 crashes from five instrumented corridors of California freeways and associated traffic flow characteristics with different types of crashes. In one of their earlier studies (Golob and Recker, 2001), they also demonstrated that collision type is the best-explained crash characteristic and is related to the median speed and the left and interior lane variations in speed. It was also pointed out that some collision types are more common under certain traffic conditions. However, no non-crash data were used in their studies. Therefore, while they were able to establish traffic conditions which precede certain types of crashes it was without any measure of 'exposure' for such conditions. Therefore, their findings albeit insightful, are not applicable in the framework of a proactive system capable of separating real-time 'crash prone' conditions from 'normal' freeway traffic.

A comprehensive analysis of rear-end crash occurrence and its relationship with the freeway loop detector data was conducted by Pande and Abdel-Aty (2006). The crash and loop detector data from instrumented corridor of Interstate-4 (Orlando, FL) were used in the study. Crashes most commonly observed after rear-ends on the aforementioned corridor are sideswipe, angle, and single vehicle crashes, respectively. Angle crashes on freeways are classified as such by the enforcement officers but are in fact a slight variant of the sideswipe crashes. These crashes should not be confused with the right-angle collisions on intersections which are commonly referred to as "angle crashes" in the traffic safety literature.

Sideswipe crashes along with the angle crashes that occur on the inner lanes of the freeway are the focus of this study. It was found that these two groups of crashes tend to occur while drivers attempt lane-changing maneuvers. Traffic data from loop detector stations located immediate upstream and downstream of the location of these crashes are compared to a sample of randomly selected non-crash cases to set up a binary classification problem. Neural network based classification models have been trained and validated using the historical crash data collected over a period of 5 years (1999–2003). The multilayer perceptron (MLP) and normalized radial basis function (NRBF) based neural network architectures are explored for classification. The outputs of the best models within both architectures were combined in order to examine the performance of the resulting 'hybrid' model. The input variables to the neural networks were finalized based on the variable importance measure (VIM) estimated through a classification tree based variable selection procedure. The step-by-step approach to modeling adopted in this study is sometimes referred to as the data mining process.

2. Modeling methodologies: components of the data mining process

Data mining is the analysis of large "observational" datasets to find unsuspected relationships potentially useful to the data owner (Hand et al., 2001). It typically involves analysis where objectives of the data analysis have no bearing on the data collection strategy. Freeway traffic surveillance data, collected through loop detectors, is one such "observational" database maintained for various Intelligent Transportation Systems (ITS) applications, such as travel time prediction, etc. In this research, data mining process is used to relate the surrogate measures of traffic conditions (data from freeway loop detectors) with the occurrence of lane-change related crashes on freeways. Note that data mining based analysis is preferred here since techniques from traditional statistics are more suitable for handling the data obtained through an experimental design, which is clearly not the case here. The data mining process has two key components, namely, variable selection procedure based on classification tree and neural network based modeling procedure with parameters identified through the preceding classification tree as inputs. These components of the data mining process are described in the ensuing section.

2.1. Decision tree based classification and its application for variable selection

A classification tree represents segmentation of data created by applying a series of simple rules. Each rule assigns an observation to a group based on the value of an input. One rule is applied after another, resulting in a hierarchy of groups within groups. The hierarchy is called a tree, and each group is called a node. The final or terminal nodes are called leaves. For each leaf, a decision is made and applied to all observations in that leaf. Decision trees are the most widely utilized tools in data mining applications. Classification trees can be used to automatically rank the input variables based on the strength of their contribution to the tree. This ranking may act as the basis for variable selection for subsequent modeling procedures such as the neural networks. In the following subsection theoretical details of the classification tree are described along with its application for variable selection. Since target variable of interest is binary in

nature (crash versus non-crash) the details of the methodology are provided in the context of a binary target.

2.1.1. Decision tree methodology for binary classification

The basic idea in classification tree construction is to split each (non-terminal) node such that the descendent nodes are 'purer' than the parent node. To achieve this, a set of candidate split rules is created, which consists of all possible splits for all variables included in the analysis. These splits are then evaluated and ranked based on one of three criteria, namely Chi-square test, entropy reduction, or Gini reduction, to choose amongst the available splits at every non-terminal node. According to Chi-square test criterion, the split resulting in the cross-frequency table with maximum $-\log(p\text{-value})$ (i.e., minimum p-value) is selected. Note that the selection of the split with minimum p-value would ensure that Child nodes resulting from the selected split are more homogeneous in nature. Entropy reduction and the Gini reduction criteria measure the "worth" of each split in terms of its contribution toward maximizing the homogeneity through the resulting split. If a split results in splitting of one parent node into B branches, the "worth" of that split may be measured as follows:

$$\text{Worth} = \text{Impurity(Parent node)} - \sum_{b=1}^{B} P(b) \times \text{Impurity}(b) \quad (1)$$

where Impurity(Parent node) denotes the entropy or Gini measure for the impurity (i.e., non-homogeneity) of the parent node and $P(b)$ denotes the proportion of observations in the node assigned to branch b. The impurity measure, Impurity(node), may be defined as follows:

According to the entropy criteria:

$$\begin{aligned}\text{Impurity(node)} &= -\sum_{\text{all classes}} p_{\text{class}} \log_2 p_{\text{class}} \\ &= -(p_{\text{crash}} \times \log_2 p_{\text{crash}} + p_{\text{non-crash}} \times \log_2 p_{\text{non-crash}}) \quad (2)\end{aligned}$$

where $\log_2$ represents log to the base 2, p_{crash} represents the proportion of crash cases in the node and $p_{\text{non-crash}}$ represents the proportion of non-crash cases in the node.

According to the Gini measure:

$$\begin{aligned}\text{Impurity(node)} &= 1 - \sum_{i}^{\text{classes}} \left(\frac{\text{number of class } i \text{ cases}}{\text{all cases in the node}} \right)^2 \\ &= 1 - [(p_{\text{crash}})^2 + (p_{\text{non-crash}})^2] \quad (3)\end{aligned}$$

If a node is 'pure', i.e., consists of only crash or only non-crash cases than these measures (Eqs. (2) and (3)) will have minimum values, and their values will be higher for less homogeneous nodes. If one considers the definition of "worth" according to Eq. (1), a split resulting in more homogeneous branches (Child nodes) will have more "worth".

While developing a classification tree, one of these criteria is applied recursively to the descendents, to achieve Child nodes having maximum worth, which in turn become the parents to successive splits, and so on. The splitting process is continued until there is no (or less than a pre-specified minimum) reduction in impurity and/or the limit for minimum number of observation in a leaf is reached (SAS Institute, 2001).

2.1.2. Application of classification trees for variable selection

Breiman et al. (1984) devised a variable importance measure (VIM) for trees. This measure may be applied as a criterion to select a promising subset of variables for other modeling tools, especially for flexible tools such as neural network.

In a classification tree with T total nodes, let $S(x_j, k)$ be the split at the kth internal node using the variable x_j. The variable importance measure for variable x_j is the weighted average of the reduction in the Gini impurity measure (defined in Eq. (3)) achieved by all splits using the variable x_j across all internal nodes of the tree and the weight is the node size. If N is the total number of observations in the training sample, then the formula for the importance for variable x_j may be given by the following:

$$\text{VIM}(x_j) = \sum_{t=1}^{T} \frac{n_t}{N} \Delta\text{Gini}(S(x_j, t)) \quad (4)$$

where $\Delta\text{Gini}(S(x_j, t))$ is the reduction in Gini measure of impurity (defined in Eq. (3)) achieved by splitting the variable x_j at node t, and n_t/N represents the proportion of the observations in the dataset that belong to node t.

Eq. (4) depicts the variable importance measure as proposed by Breiman et al. (1984). In this study, however, the VIM used has been scaled by maximum importance for the tree so that it lies between 0 and 1. One may conveniently use a threshold of 0.05 on VIM to separate variables critically associated with the binary target from the variables that are not. These critical variables can then be used as inputs to the classification models in subsequent step(s) of the data mining process. Moreover, a closer examination of the resulting classification tree structure, based on which the VIM is calculated, may also provide insight into crash precursors and their relationships with crash occurrence. The variables selected through this procedure would be used to develop classification models belonging to MLP and NRBF neural network architectures.

2.2. MLP neural network architectures and training procedure

2.2.1. MLP neural network architecture

A neural network may be defined as a massively parallel-distributed processor made up of simple processing units having natural propensity for storing experimental knowledge and making it available to use (Christodoulou and Georgiopoulos, 2001). The ability to learn and generalize provides neural networks with the computing power it possesses. Generalization refers to the ability of a "trained" network to provide satisfactory responses even for the inputs that it has not seen during the training process. Neural network models may usually be specified by three entities, namely, model of processing elements them-

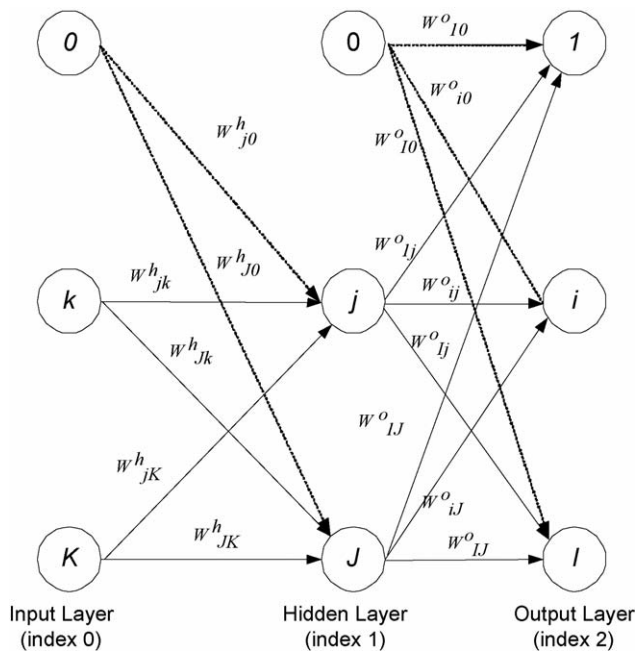

Fig. 1. MLP neural network architecture with feed-forward connections (Christodoulou and Georgiopoulos, 2001).

selves, model of interconnections and structures (i.e., network topology), and the learning rules. In this section, we describe multi-layer perceptron (MLP) network with feed-forward connections. It is one of the most commonly used neural network architectures.

An MLP neural network shown in Fig. 1 has input layer of size K (Index 0), a hidden layer of size J (Index 1) and output layer of size I (Index 2) along with input and output bias. In the MLP architecture shown here, the connections are of feed-forward type; it means that the only connections allowed between nodes are from a layer of a certain index to the next layer with higher index. The net input to hidden layer neurons is determined through inner product between the vector of connection weights and the inputs. The activation function is applied to this net input of hidden neurons. The weights from the hidden to output layer are then used to estimate the output of the network. These weights are the parameters recursively estimated during the supervised learning (i.e., training) process and are used to 'score' unseen observations following calibration. The activation function of hidden neurons is nonlinear in nature and is critical in the functioning of the neural network. It allows the network to 'learn' any underlying relationship of interest between inputs and outputs. The procedure adopted for training is also crucial in the performance of a neural network.

2.2.2. *Normalized radial basis function neural network*

In feed-forward neural network architectures, the activation function of hidden neurons is applied to a net single value that is obtained by combining input vectors with the vector of connection weights between input layer and the hidden layer. The function that combines the inputs with the weights may be referred to as the 'combination function'. In the MLP neural network architecture, the combination function was simply the inner product of the inputs and weights. A radial basis function (RBF) network is a feed-forward network with a single hidden layer for which the 'combination function' is more complex and is based on a distance function (referred to as width) between the input and the weight vector. Ordinary RBF (ORBF) networks using radial combination function and exponential activation function are universal approximators in theory (Powell, 1987), but in practice they are often ineffective in estimating multivariate functions. To avoid the pitfalls of ORBF networks, softmax activation function may be used. It essentially normalizes the exponential activations of all hidden units to sum to one. The network with softmax activation functions is called a "normalized RBF" or NRBF network. The distinction and advantages of NRBF networks (over the ORBFs) are discussed in detail by Tao (1993). It was argued by Tao (1993) that the normalization not only is a desirable option but is in fact imperative.

In NRBF networks, one may add another term "altitude" to the Gaussian combination function. It determines the maximum height of the Gaussian curve over the horizontal axis. Based on the two parameters (width and altitude) defining the shape of combination function, the NRBF networks may be categorized into five different types:

(1) NRBFUN: Normalized RBF network with unequal widths and heights.
(2) NRBFEV: Normalized RBF network with equal volumes ($a_i = w_i$).
(3) NRBFEH: Normalized RBF network with equal heights (and unequal widths) ($a_i = a_j$).
(4) NRBFEW: Normalized RBF network with equal widths (and unequal heights) ($w_i = w_j$).
(5) NRBFEQ: Normalized RBF network with equal widths and heights ($a_i = a_j$) and ($w_i = w_j$).

where w_i and a_i represent the widths and altitudes, respectively, of the neurons in the hidden layer. Note that the last four categories of networks are special cases of the first and are more parsimonious in nature. It essentially means that with certain assumptions about the shape of the combination functions they reduce the number of parameters that need to be estimated. In this study, the networks belonging to the first category would be used. NRBFUN networks are preferred over other architectures because no assumptions regarding the form of combination functions are needed.

The NRBF networks may be trained by "hybrid" methods, in which the hidden weights (centers) are first obtained by unsupervised learning and then the output weights are obtained by supervised learning. However, according to Tarassenko and Roberts (1994), the supervised training will often let one use fewer hidden units (with fewer training cases) for a given accuracy level of the approximation than the hybrid training. Hence, fully supervised training is adopted for NRBF and MLP neural networks. Supervised training for these networks can be accomplished using Levenberg–Marquardt algorithm.

2.2.3. Training of MLP-NN: Levenberg–Marquardt (LM) algorithm

Training a neural network essentially involves numerical optimization of a non-linear function. Error back-propagation (EBP) algorithm proposed by Rumelhart et al. (1986) still remains the most widely used supervised training algorithm. It, however, has been known to have a poor convergence rate for more complex problems (Wilamowski et al., 2001). A significant improvement in the performance of the network may be achieved by using second-order approaches such as the Levenberg–Marquardt (LM) optimization technique. For LM algorithm, the objective function takes the following form (Wilamowski et al., 2001):

$$F(w) = \sum_{p=1}^{P} \left[\sum_{k=1}^{K} (d_{kp} - o_{kp})^2 \right] \tag{5}$$

where $w = [\, w_1 \quad w_2 \quad \ldots \quad w_N \,]^{\mathrm{T}}$ consists of the interconnection weights in the network, d_{kp} and o_{kp} are the desired and actual values of the target, respectively, for kth output and pth pattern. N is the total number of weights, P the number of patterns, and K is the number of network outputs. The above equation may be rewritten as:

$$F(w) = E^{\mathrm{T}} E \tag{6}$$

$$\mathbf{E} = [e_{11} \ldots e_{K1}\, e_{12} \ldots e_{K2} \ldots e_{1P} \ldots e_{KP}]^{\mathrm{T}},$$

$$e_{kp} = d_{kp} - o_{kp}, \quad k = 1, \ldots, K, \quad p = 1, \ldots, P$$

where $\mathbf{E}$ is the cumulative error vector. Based on Eq. (6), the Jacobian matrix of the output errors with respect to the N interconnection weights will be:

$$J = \begin{bmatrix} \frac{\partial e_{11}}{\partial w_1} & \frac{\partial e_{11}}{\partial w_2} & \cdots & \frac{\partial e_{11}}{\partial w_N} \\ \frac{\partial e_{21}}{\partial w_1} & \frac{\partial e_{21}}{\partial w_2} & \cdots & \frac{\partial e_{21}}{\partial w_N} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial e_{KP}}{\partial w_1} & \frac{\partial e_{KP}}{\partial w_2} & \cdots & \frac{\partial e_{KP}}{\partial w_N} \end{bmatrix} \tag{7}$$

The interconnection weights are adjusted after each iteration using the following equation:

$$w_{t+1} = w_t - (\boldsymbol{J}_t^{\mathrm{T}} \boldsymbol{J}_t - \lambda_t \boldsymbol{I})^{-1} \boldsymbol{J}_t^{\mathrm{T}} E_t \tag{8}$$

where $\boldsymbol{I}$ is the identity unit matrix, λ the learning parameter, and $\boldsymbol{J}$ is the Jacobian of the output errors with respect to the weights of the neural network (Eq. (6)). It should be noted that if $\lambda = 0$, then the above equation becomes the Gaussian–Newton method while for very large λ, the algorithm is equivalent to the error back-propagation algorithm. The learning parameter is automatically adjusted after every iteration in order to secure convergence.

Obviously the algorithm requires computation of Jacobian matrix and inversion of the $\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}$ matrix at each iteration step. Since the dimension of the matrix to be inverted is $N \times N$, the LM algorithm becomes computationally impractical for large size neural networks. According to Wilamowski et al. (2001), with increase in number of independent variables the computational complexity of the algorithm grows exponentially.

To overcome this limitation of the training algorithm, a reliable classification tree based variable selection algorithm has been employed in this study. It will ensure that a limited number of the most significant variables are used as inputs to the neural networks, thereby controlling the size of the network.

3. Data description and preparation

3.1. Study area and crash data composition

The Orlando area Interstate-4 (I-4) corridor under consideration is 36.25 miles long and has a total of 69 loop detector stations (numbered 2–71, with no station numbered as 39) in each direction. Distance between the two consecutive stations is approximately 0.5 miles. Each of these stations consists of dual loops and measures average speed, occupancy, and volume over 30 s period on each of the three through travel lanes in both directions. The loop detector data were continuously transmitted and archived by the UCF data warehouse. The source of crash and geometric characteristics data for the freeway is Florida Department of Transportation (FDOT) intranet server.

According to the database maintained by Florida Department of Transportation, there were 4189 mainline crashes reported on the Interstate-4 corridor under consideration over the 5-year period (1999–2003). However, out of these, only 3124 had any corresponding loop data available. Among these, about 11% were identified as sideswipes while 10% of them were classified as angle crashes. Based on the study by Wang and Knipling (1994), it could be safely assumed that the crashes classified as sideswipe crashes occur when one vehicle intentionally changes lane and sideswipes or is sideswiped by a vehicle in the adjacent lane. This postulation was verified by examining the actual reports filed by law enforcement officers at the scene of these historical crashes. Among the angle crashes, those on the inner through lanes (the center and left-most lane) of the freeway were hypothesized to be lane changing related because of the rare interaction of the vehicles on these lanes with the vehicles approaching from other directions. A closer examination of the reports for angle crashes led to the conclusion that such crashes on the center and left through lanes, although reported as angle crashes, in fact show more resemblance to sideswipe crashes in their mechanism and can be associated with lane changing (Lee et al., 2006). Hence, the crashes that are intended to be identified by the models developed in this paper include crashes that can be attributed to lane changing, i.e., all sideswipe crashes and the angle crashes on center and left lane. These crashes make up about 16% of the 3124 crashes with some corresponding loop data available and are referred to as lane-change related crashes in this study.

Variables explored as potential inputs include differences in traffic flow parameters between the three through lanes at

the station immediately upstream of the locations of historical crashes. The reason for including measures of between-lane variations is that the interaction between traffic flows in individual lanes might affect the lane changing behavior of drivers as well as the risk involved in lane changing maneuvers. Traffic flow parameters from all three lanes would be required to deduce the input variables representing the between-lane variations. It was noticed that out of 69 stations, data from all three lanes of the freeway were never available simultaneously from eight stations located on the two extremities (five on the west end and three on the east end) of the freeway corridor. Therefore, the corridor for this study, which only deals with lane-change related crashes, was limited to 32.37 miles instead of the 36.25 instrumented corridor of Interstate-4. A similar data availability problem was observed at stations 38, 40, and 41 and the crashes at those locations also could not be considered for analysis. In conclusion, lane-change related crashes in the vicinity of 58 loop detector stations were used in the analysis.

3.2. Loop data corresponding to crashes and non-crash cases

After assembling the crash data, the next step was to extract loop data corresponding to these crashes. First of all, the loop detector stations nearest to the location of each crash in upstream and downstream direction were determined. Station nearer to the crash location out of these two stations was named as "station of crash". Loop data were then extracted for every crash in a specific format. If a crash, for example, occurred on April 12, 1999 (Monday) 06:00 p.m., I-4 eastbound and the nearest loop detector in the upstream and downstream directions were at stations 30 and 31, respectively, then this crash case will have loop data table consisting of the 30 s averages of speed, volume, and occupancy for all three lanes at stations 30 and 31 (on eastbound direction) from 05:40 p.m. to 06:00 p.m. on April 12, 1999. Variable "*y*" was created with its value as 1 for all the crashes. It would later be used as the binary target variable.

It is worth mentioning that the reported time of crashes obtained from individual crash reports has been used for collecting the corresponding loop data. The accuracy of the reported time of crashes is a critical issue identified in some of the relevant literature (Lee et al., 2002, 2003). Fortunately, there is an automated system in place in Florida that records the exact time when a crash is reported to the Police. According to Florida Highway Patrol (FHP) officials, due to wide spread use of mobile phones, difference between time of crash occurrence and its reporting is minimal. It was also pointed out by local traffic management authorities that the reported time of the crash in accident reports is corroborated through the video surveillance system available on the freeway. To validate their claim, before proceeding with the collection of loop data according to the reported time of crashes, its concurrence with the actual time of crash was verified through a rule based shockwave methodology developed in one of our previous studies (Abdel-Aty et al., 2005a). It was found that most of the crashes where the methodology could be successfully applied, the estimated time of crash occurrence concurred with the reported time. These pieces of information indicated that the time obtained from the crash reports is in fact very close to the actual time of crash occurrence and can be used for collecting the loop data.

The aim of this research is to develop models with the ability to separate conditions prone to lane-change related crashes from 'normal' freeway traffic. Lane-change related crashes with corresponding loop data available constitute the sample that would 'teach' the neural network models about crash prone conditions. A random sample of non-crash loop detector data would be used to provide the models with a-priori information on what constitutes 'normal' traffic on the freeway. These non-crash data were selected from a sample of 150,000 random non-crash cases. To generate random non-crash cases, 5-year period may be divided into 2,629,440 1-min periods (60 min × 24 h × 1826 days over 5 years = 2,629,440 1-min periods), which would be the number of options available to choose the "time of non-crash". Similarly, we have 116 stations (58 stations in two directions: eastbound and westbound) to choose as "station of non-crash". In all, we can choose from 305,015,040 (2,629,440 1 min periods × 2 directions × 58 stations) options to draw a random combination of time, station, and direction to assign as random non-crash case. One lakh and fifty thousand such combinations were selected randomly as the non-crash cases. These cases were also assigned a random milepost location as per the corresponding "station of non-crash". Randomly selected combinations of time, station, and direction were used to extract sets of 20 min loop data prior to the assigned time of the non-crash from the station immediate upstream as well as immediate downstream of the random milepost assigned to it. It constituted a random non-crash sample. The variable "*y*" was given the value 0 for these cases. Out of these 150,000 random non-crash cases, a non-crash sample of appropriate size may be drawn depending on the sample size requirements of the methodology used for analysis. It was ensured that no crash cases were included in these random non-crash cases.

The milepost location of the ramps on the Interstate-4 corridor was known from the FDOT database. Using this information, along with the milepost location of each crash, the distances of nearest on and off-ramp from crash location, in both upstream and downstream direction, were determined. Essentially, we created four variables, namely "upstreamon", "upstreamoff", "downstreamon", and "downstreamoff" for each crash case; indicating the distance of nearest ramp of the respective type from crash location. These variables for non-crash cases were obtained based on the assigned milepost location.

3.3. Loop data aggregation and preliminary analysis

The raw 30 s loop data have random noise and are difficult to work with in a modeling framework. Therefore, the 30 s raw data were combined as 5 min level averages and standard deviations of these traffic parameters. For 5 min aggregation, 20 min period was divided into four time-slices. The stations were named as "U" or "W", with "U" being station upstream of the crash loca-

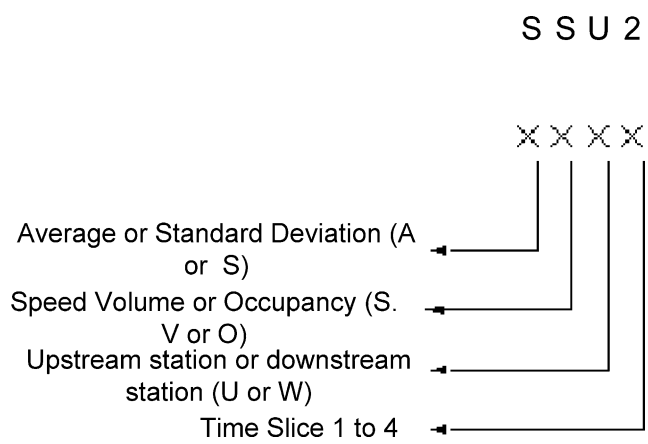

Fig. 2. Nomenclature for the factors used for analysis of lane-change related crashes.

tion and "W" being the downstream station. Similarly, the 5 min intervals were also given "IDs" from 1 to 4. The interval between time of the crash and 5 min prior to the crash was named as time-slice 1, interval between 5 and 10 min prior to the crash as time-slice 2, and so on. These parameters were further aggregated across the three lanes and the averages (and standard deviations) for speed, volume, and lane-occupancy at 5 min level were calculated based on 30 (10, 30 s observations in 5 min × 3 lanes) observations. The nomenclature for these independent variables is exemplified in Fig. 2. The variable "SSU2", for example, represents the standard deviation of 30 speed observations during the 5 min period of 5–10 min prior to a crash at station "U", which is the upstream station. According to the nomenclature shown in the figure, the same parameter measured at the station downstream of crash site would have been named "SSW2". The random non-crash data were also aggregated to 5 min level and traffic parameters similar to crash cases (refer to Fig. 2) were generated. These parameters explored as input variables were similar to the ones used for identification of conditions prone to rear-end crashes. The critical difference, of course, is that the rear-end crashes are related to formation/dissipation of queues as opposed to lane-changing maneuvers. Therefore, for rear-end crashes data from a series of stations upstream and downstream of crash location were analyzed. For lane-change crashes we are more interested in traffic conditions at or very close to the crash location. The traffic parameters from the stations located immediately upstream and downstream of the crash location are used as inputs to the models for lane-change related crashes.

Flow ratios representing a measure for the number of lane-changing maneuvers, identified by Chang and Kao (1991) and Lee et al. (2006), were also attempted as the input variables. The flow ratio devised by Chang and Kao (1991) was based on their field studies to identify macroscopic traffic factors related to lane changing behavior. Lee et al. (2006) proposed some modifications to the aforementioned flow ratio to overcome the limitations in applying this factor to investigate its effects on lane-change related crashes. It was noted that the work by Chang and Kao (1991) only relates the number of lane changes in specific lane to average flow ratios (AFR) in the corresponding lane but does not consider the total number of lane changes in all lanes in the form of overall AFR (OAFR). However, OAFR might be important in representing general traffic stability on freeways and its consequent impact on crash risk. Therefore, AFR calculated for each subject lane should be combined to reflect the total number of lane changes (Lee et al., 2006).

The objective of the study by Lee et al. (2006) was to be able to differentiate between rear-end and lane-change related crashes. First, the average flow ratios for the individual lanes were defined as follows:

$$\begin{aligned} \mathrm{AFR}_1(t) &= \frac{v_2(t)}{v_1(t)} \times \left(\frac{\mathrm{NL}_{2,1}(t)}{\mathrm{NL}_{2,1}(t) + \mathrm{NL}_{2,3}(t)} \right) \\ \mathrm{AFR}_2(t) &= \frac{v_1(t)}{v_2(t)} + \frac{v_3(t)}{v_2(t)} \\ \mathrm{AFR}_3(t) &= \frac{v_2(t)}{v_3(t)} \times \left(\frac{\mathrm{NL}_{2,3}(t)}{\mathrm{NL}_{2,1}(t) + \mathrm{NL}_{2,3}(t)} \right) \end{aligned} \tag{9}$$

where $\mathrm{AFR}_1(t)$ is the average flow ratio in lane 1 (left lane) during time interval t; $\mathrm{AFR}_2(t)$ is the average flow ratio in lane 2 (center lane) during time interval t; $\mathrm{AFR}_3(t)$ is the average flow ratio in lane 3 (right lane) during time interval t; $v_1(t)$, $v_2(t)$, and $v_3(t)$ are the average flow in lane 1, 2, and 3, respectively, during time interval t; $\mathrm{NL}_{2,1}(t)$ and $\mathrm{NL}_{2,3}(t)$ are the number of lane changes from lanes 2 to 1 and lanes 2 to 3, respectively, during time interval t.

In above equations, since the fractions of lane changes from lane 2 to lanes 1 and 3 were unknown, they were assumed to be equal (i.e., $\mathrm{NL}_{2,1}/(\mathrm{NL}_{2,1} + \mathrm{NL}_{2,3}) = \mathrm{NL}_{2,3}/(\mathrm{NL}_{2,1} + \mathrm{NL}_{2,3}) = 0.5$). In case of AFR in lane 2, since there is only one way of lane-change from lanes 1 and 3, there is no need to estimate the fractions of lane changes and OAFR (overall average flow ratio) can be calculated using the following expression:

$$\mathrm{OAFR}(t) = \sqrt[3]{0.5 \left(\frac{v_2(t)}{v_1(t)} \right) \times \left(\frac{v_1(t) + v_3(t)}{v_2(t)} \right) \times 0.5 \left(\frac{v_2(t)}{v_3(t)} \right)} \tag{10}$$

Eq. (10) in a more general form for an n-lane freeway may be represented as follows:

$$\begin{aligned} \mathrm{OAFR}(t) &= \sqrt[n]{\mathrm{AFR}_1(t) \times \mathrm{AFR}_2(t) \times \ldots \times \mathrm{AFR}_n(t)} \\ &= \left(\prod_{i=1}^{n} \mathrm{AFR}_i(t) \right)^{1/n} \end{aligned} \tag{11}$$

Note that Eqs. (10) and (11) represent geometric mean of the individual average flow ratios shown in Eq. (9) (Lee et al., 2006). This factor was found statistically significant in separating loop data prior to lane-change related crashes from loop data observed prior to rear-end crashes.

In the present study, these flow ratios along with the off-line factors (e.g., presence of ramps, milepost lactation) were first subjected to a preliminary analysis for variable selection. As mentioned earlier, some locations had to be excluded from analysis because the data from all three lanes at those locations were almost never available. Even among the remaining stations the loop failure pattern was not random. At some locations, detector at least one of the three lanes was more likely to fail. Hence, if among the randomly selected 150,000 cases one only considers 47,693 cases (which had data from all three lanes available) some locations were under-represented than others even though the original 150,000 cases were almost uniformly distributed over all stations.

It means that due to non-random failure patterns the sample with all three lane data available was not random. In other words, the underlying distribution of the sample changed due to data availability issues. The loop failure pattern also affected the underlying distribution of the crash cases. To overcome this, a weighted sampling procedure was adopted by over-sampling the random non-crash cases from under-represented locations and vice versa. The weights used for making the non-crash distribution uniform were then adopted for crashes. The underlying principle was that since the weighted sampling restored the underlying distribution (random with all locations uniformly represented) of non-crash cases; if applied it would do the same to crash cases.

Note that the sample of crashes resulting from the weighted sampling is not supposed to be uniform but should be comparable to what it was without taking data availability into consideration. The distribution of crash cases over the freeway locations was compared to their distribution in the original lane-change related crash sample (the later proportions were based on actual frequency of crashes without taking loop data availability into consideration). It was found that at 95% confidence level there was no difference between the two samples. The weighted sample of crash and non-crash cases was used for preliminary analysis. Note that without the weighted sampling we could not have simultaneously analyzed the effect of traffic parameters along with location specific characteristics on lane-change related crashes. The loop data availability would have affected the sample in such a way that location represented in the crash sample would not exactly be the locations with high frequency of crashes but would be locations with better functioning of the loop detectors. With such a sample the results of the analysis, about how the factors such as milepost location, presence of on and off-ramps affect the crash occurrence, would have been questionable.

The sample was subjected to the classification tree based variable selection procedure for the binary target "*y*". Variables included as potential inputs were the average and standard deviation of the speed, volume, and occupancy (SSU2, SSW2, etc.). In addition, the flow ratio (represented by Eq. (10)) from the station located upstream of the crash location was also subjected to the selection process. It turned out not to be significantly associated with the binary target, however. Also, none of the off-line factors, including the factors explicitly related to the presence of on and off-ramps, milepost location on the freeway had significant VIM. This conclusion was further confirmed by analyzing the data using the within stratum matched case–control sampling for crashes. Under this scheme all the crashes are sampled first and then non-crash cases are sampled corresponding to each crash. The correspondence means that, for example, if a crash occurred on April 12, 1999 (Monday) 06:00 p.m., I-4 eastbound and the nearest loop detector was at station 30, data were extracted from the same location for the 5 min period 5–10 min prior to the time of the crash for all Mondays of the same season for the year 1999 at the same time. This matched sample design controls for most of the critical off-line factors affecting crash occurrence such as time of day, day of week, location on the freeway, etc (thus implicitly accounting for these factors). A logistic regression model with step-wise variable selection procedure was estimated following the sampling procedure. It was found that the variables included in the logistic regression model were in fact a subset of the variables identified by the classification tree based selection procedure. It essentially means that the off-line factors used to create the strata (i.e., the control parameters) for matched case–control sampling, adopted in some of our earlier studies (Abdel-Aty et al., 2004, 2005b), are not critical for identifying conditions prone to lane-change related crashes. More details on this preliminary analysis may be found in Pande (2005).

In short, two critical conclusions were drawn from this preliminary analysis; one, geometric characteristics of the freeway segments are not as significantly associated with lane-change crashes as they were with the rear-end crashes (Pande and Abdel-Aty, 2006). Second, the flow ratios measured at 5 min level, although significant in separating lane-change related crashes from rear-end crashes (Lee et al., 2006), are not sufficient to separate crashes from random non-crash cases and therefore the between-lanes variation of traffic parameters must be examined in more detail.

The former conclusion leads to the inference that as far as identification of conditions prone to lane-change related crashes based on real-time traffic data is concerned, there is no significant difference among different sections of the I-4 corridor. It means that classification model(s) developed using data from certain segments of the freeway corridor may be applied to other segments of Interstate-4, loop data belonging to which were not used at the modeling stage. Based on this inference, the under-represented locations in the dataset may be excluded altogether from the sample at the modeling stage. Even without including these locations in the modeling sample the estimated model(s) would be able to assess crash risk at those locations in real-time provided requisite data are available. After this exclusion there were 162 crashes, loop data for which were used for further analysis and neural network model calibration.

Following the preliminary analysis, variables more precisely representing between lane variations in traffic flow parameters were calculated to examine their effect on lane-change related crashes. Two sets of such parameters were calculated. The first set of parameters measuring 5 min average of between lanes variations of speed/volume/occupancy are defined in the

following equation:

$$\begin{aligned}
\mathrm{ABLVSU2} &= \frac{1}{10}\sum_{i=1}^{10}|\mathrm{LS}-(\mathrm{LS}+\mathrm{CS}+\mathrm{RS})/3|+|\mathrm{CS}-(\mathrm{LS}+\mathrm{CS}+\mathrm{RS})/3|+|\mathrm{RS}-(\mathrm{LS}+\mathrm{CS}+\mathrm{RS})/3| \\
\mathrm{ABLVVU2} &= \frac{1}{10}\sum_{i=1}^{10}|\mathrm{LV}-(\mathrm{LV}+\mathrm{CV}+\mathrm{RV})/3|+|\mathrm{CV}-(\mathrm{LV}+\mathrm{CV}+\mathrm{RV})/3|+|\mathrm{RV}-(\mathrm{LV}+\mathrm{CV}+\mathrm{RV})/3| \\
\mathrm{ABLVOU2} &= \frac{1}{10}\sum_{i=1}^{10}|\mathrm{LO}-(\mathrm{LO}+\mathrm{CO}+\mathrm{RO})/3|+|\mathrm{CO}-(\mathrm{LO}+\mathrm{CO}+\mathrm{RO})/3|+|\mathrm{RO}-(\mathrm{LO}+\mathrm{CO}+\mathrm{RO})/3|
\end{aligned} \tag{12}$$

LS, CS, and RS, respectively, represent left, center, and right lane speed values observed every 30 s. First, the average of 30 s speeds over the three lanes is calculated as (LS + CS + RS)/3. The absolute value of the difference between individual lane speeds and this average is then added together, which is the term inside the summation in Eq. (12). The parameter is then averaged over ten 30 s observations that are recorded during the 5 min slice. The parameters shown are calculated for station located upstream of the crash location for time-slice 2 (5–10 min period before the crash) as indicated by the term "U2" at the end of each parameter. The term "ABLV" represents "average between lane variations". ABLV for volume and occupancy are calculated in an identical manner. Note that this is just one way to represent the between lane variation of traffic parameters. Another set of parameters calculated to represent them is provided below in Eq. (13):

$$\begin{aligned}
\mathrm{ADALSU2} &= \frac{1}{10}\sum_{i=1}^{10}|\mathrm{LS}-\mathrm{CS}|+|\mathrm{CS}-\mathrm{RS}| \\
\mathrm{ADALVU2} &= \frac{1}{10}\sum_{i=1}^{10}|\mathrm{LV}-\mathrm{CV}|+|\mathrm{CV}-\mathrm{RV}| \\
\mathrm{ADALOU2} &= \frac{1}{10}\sum_{i=1}^{10}|\mathrm{LO}-\mathrm{CO}|+|\mathrm{CO}-\mathrm{RO}|
\end{aligned} \tag{13}$$

In Eq. (13), the absolute difference between speeds in adjacent lanes is added together and averaged over the 5 min slice. The term "ADAL" represents "average difference between adjacent lanes".

The two sets of parameters are two different measures of representing the same traffic characteristics (i.e., variation of speed/volume/occupancy between the three lanes) and as expected, the correlation coefficients between the parameters shown in Eqs. (12) and (13) were in the vicinity of 0.95. Therefore, these parameters were not attempted together in the variable selection/modeling procedure and were tried one set at a time. Note that data from all three lanes of the freeway would be required to compute the variables shown in Eqs. (12) and (13).

3.4. Final variable selection process and results

The dataset with 162 crashes and 3650 non-crash cases (after removing crash and non-crash observations belonging to under-represented locations) was then partitioned into training (70%) and validation (30%) datasets. The datasets were subjected to classification tree based variable selection process. While the parameters from multiple time-slices were available, parameters from only one of the four slices (20 min period was divided into four 5 min time-slices) at a time were attempted in the variable selection process. At this stage the variables included as potential inputs from the downstream station were the average and standard deviation of the speed, volume, and occupancy (AS/SS/AV/SV/AO/SOW2). From the upstream station average of speed, volume, and occupancy (ASU2, AVU2, and AOU2) were included. In addition, three sets of between-lane variation (speed/volume/occupancy) measures at the upstream station were also subjected to the selection process one at a time. The first set included SSU2, SVU2, and SOU2 and the other two were the ones represented by Eqs. (12) and (13), respectively. The list of significant variables identified by classification tree models employing entropy maximization criterion for optimal split is provided in Table 1.

By examining the classification tree model closely it was noticed that high average speed downstream of crash site (ASW2) along with low average speeds upstream (ASU2) increases the likelihood of lane-change related crashes. It indicates that when drivers perceive a chance to increase speed, while traveling from low average speed regime (measured at the station upstream) to high average speeds (measured at the station located downstream of the crash site) they might make lane-changing maneuvers, thereby, increasing chances of conflicts. It was also noticed that if both upstream and downstream are operating at high speeds (around or greater than 50 mph) small average differences between adjacent lane occupancies

Table 1
Results of variable selection procedure for lane-change related crashes

Name	Variable importance measure (VIM)	Variable description
ASW2	1.0000	Average speed at station downstream of crash location
ASU2	0.6179	Average speed at station upstream of crash location
AOW2	0.5142	Average occupancy at station downstream of crash location
ADALOU2	0.2692	Average of absolute difference between 30 s occupancy observations on adjacent lanes
SVW2	0.2591	Standard deviation of volume at station downstream of crash location
SSW2	0.2006	Standard deviation of speed at station downstream of crash location

upstream of the crash site involve more risk than the sites with this parameter (ADALOU2) being high. Hence, if the difference in occupancy between adjacent lanes is small then caution should be exercised while changing lanes. Standard deviation of volume and speed (SVW2 and SSW2) downstream of crash site were found to be positively associated with lane-change related crashes.

4. Neural network based classification models

Following variable selection neural network based modeling procedure was initiated with variables shown in Table 1 as inputs. The best models were identified through the lift plot having cumulative percentage of captured response for the validation dataset on the vertical axis. The output of the neural network based classification models for any observation is termed as the posterior probability of the event (i.e., a lane-change crash in this case). Posterior probability is a number between 0 and 1. The closer it is to unity the more likely, according to the model, it is for that observation to be a crash. In a lift chart, the observations in the validation dataset are sorted from left to right by the output posterior probability obtained from each model. The sorted group is lumped into 10 deciles[1] (one decile represents 10 percentiles) along the horizontal axis. The left-most decile is the 10% of observations with highest posterior probability, i.e., most likely to be a lane-change related crash. The performance of each model may be measured by determining how well the models capture the target event across various deciles. From a practical application point of view it must be understood that crashes are rare events and one would need to be parsimonious in issuing warnings for crashes. Therefore, it might not be reasonable to assign more than 20–30% of observations as crashes and it was decided to evaluate the individual neural network models at the validation stage based on the percentage of crashes identified within first three deciles of posterior probability. The threshold may be altered at the application stage based on desired number of warnings. A more elaborate discussion on this issue is provided in the next section. It should also be noted the posterior probability is not the probability of crash occurrence at a given point in time but is a measure providing the relative likelihood of crash occurrence given the composition of the sample.

The first neural network architecture explored for classification is the multi-layer perceptron with Levenberg–Marquardt training algorithm. The training procedure starts with an arbitrary randomly chosen set of interconnection weights and then it tries to minimize the difference between network output and the desired outputs for the training dataset. All runs have been carried out with a maximum number of epochs (a complete list presentation) of 1500 and error goal of 0.01. It has been proven in the literature that an MLP network with one hidden layer and non-linear activation functions for the hidden nodes can learn to approximate virtually any continuous function to any degree of accuracy (Cybenko, 1989). Therefore, the most critical issue was to estimate the number of neurons in the hidden layer. The underestimation of hidden neurons leads to a network having an incomplete representation of inputs and by contrast, the over representation reduces the network to a simple look-up table. The methodology adopted for selecting appropriate number of nodes in the hidden layer was to evaluate the performance of the models having hidden nodes varying from 2 to 8. Unconstrained normalized radial basis function neural network (NRBFUN) were also used for classification of lane-change related crashes. To select appropriate number of nodes in the hidden layer, performance of seven different NRBF networks (with hidden nodes varying from 2 to 8) was examined.

Table 2
Structure and percentage of captured response within the first three deciles for two classes of neural network models along with the validation root mean square error (RMSE)

Neural network architecture	Number of hidden neurons	Crash identification in first three deciles (%)	Validation error (RMSE)
NRBF	2	31.42	0.2060
NRBF	3	48.00	0.2039
NRBF	4	32.87	0.2030
NRBF	5	44.00	0.2033
NRBF	6	44.29	0.2034
NRBF	7	32.00	0.2042
NRBF	8	37.26	0.2038
MLP	2	38.73	0.2030
MLP	3	44.44	0.2035
MLP	4	50.00	0.2037
MLP	5	40.44	0.2041
MLP	6	33.26	0.2039
MLP	7	34.26	0.2039
MLP	8	45.90	0.2039

Table 2 depicts the performance of various NRBF and MLP neural networks having varied number of hidden neurons. The performance is shown in terms of validation root mean square error (RMSE) as well as percentage of crashes identified within 30% observations with highest posterior probability output. Note that being a binary classification problem there is not much difference between RMSE values for various models. It may be seen that NRBF network with three hidden neurons and MLP network with four hidden neurons provide the best crash identification within the first three deciles of posterior probability. The row corresponding to the two models are highlighted in the table.

In the next step, these two models were hybridized by averaging posterior probabilities from the individual models. For a binary target, a hybrid model may alternatively be achieved by classifying the cases into the classes assigned to them by majority of the individual models. This method is called voting and is not equivalent to averaging posteriors. While voting could provide a predicted target value, it would not produce posterior probability estimates consistent with the individual posteriors. When an individual classifier assigns an output class label, the decision is based on a pre-determined threshold. If the estimated posterior probability is less than this threshold then the classifier would produce 0 indicating non-crash; otherwise, it would return a value of unity to indicate a crash. The output of a hybrid

[1] Decile is defined as any of nine points that divided a distribution of ranked scores into equal intervals where each interval contains one-tenth of the scores.

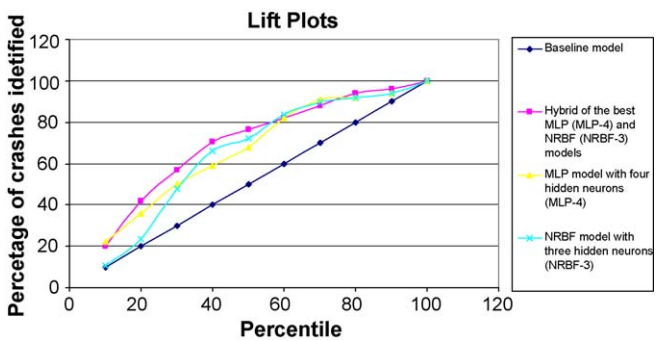

Fig. 3. Percentage of captured response lift plot for the best models belonging to different modeling techniques along with the hybrid model.

Table 3
Performance of the classification models over the validation dataset

Percentiles of posterior probability	Percentage of crashes identified in the validation dataset			
	Baseline model	Hybrid model	NRBF-3	MLP-4
10	10	20 (+10)	11 (+01)	22 (+12)
20	20	42 (+22)	24 (+04)	36 (+16)
30	30	57 (+27)	48 (+18)	50 (+20)
40	40	70 (+30)	66 (+26)	59 (+19)
50	50	77 (+27)	72 (+22)	68 (+18)

The margin in the parentheses shows the differential between crashes identified by the corresponding model and the baseline model.

classifier, according to the voting method, would be based on the majority of class labels from multiple classifiers. In that case, observations assigned as crash according to the "majority-rule" hybrid classifier cannot be compared amongst each other. In other words, there would be no way to judge which pattern is *more* crash prone among all the patterns that are identified as potential crashes. However, if the hybrid model is estimated by averaging the posterior probabilities, it is still possible to rank the observations in the validation dataset to create lift plots. It will in turn help in evaluating the performance of the hybrid model vis-à-vis the individual models.

In fact, for the present research problem, a significant improvement in crash identification was achieved through the hybrid model created by averaging the outputs from the best MLP and NRBF models. Fig. 3 shows the lift plot for the two individual models (NRBF-3 and MLP-4) highlighted in Table 2 along with the hybrid model. The curve shows the percentage of the lane-change crashes in the validation dataset captured within various deciles of posterior probability by each model on y-axis. On the x-axis the percentiles are shown at equal intervals of 10. Fig. 3 also demonstrates '*performance*' of a random baseline model which represents the expected percentage of crashes identified in the validation sample if one randomly assigns validation dataset observations as crash and non-crash. A model can be assessed by examining the separation of its corresponding lift curve from the random baseline curve. It may be seen that in the first half (up to 50 percentiles) the lift curve for the hybrid model remains consistently above the curves for the best individual models.

The performance of the best individual models, the hybrid model, and the baseline model is summarized in Table 3. The performance is measured in terms of the percentages of crashes identified at various deciles (1–5). The percentage of crashes 'identified' by the baseline model is equal to the corresponding percentile values. For the other three models (MLP-4, NRBF-3, and the hybrid model) the table shows the percentage of crashes identified at the five deciles along with the differential of these percentages vis-à-vis the baseline model in the parentheses.

It may be seen from Table 3 that the hybrid model identifies 57, 70, and 77% crashes in the validation dataset, respectively, at 30, 40, and 50 percentiles. As we increase the percentage of observations declared as crash, the crash identification will obviously improve but the percentage of non-crash cases correctly identified would decrease. Hence, there is a trade-off involved since as we declare more patterns as crashes we also increase the 'false alarms'. Also, note that the performance of the hybrid model created by combining the outputs of the best individual models is much better than that of the best individual models. The comparison of the performance of the hybrid model with that of the baseline model suggests that the hybrid model is in fact capable of identifying conditions prone to lane-change related crashes.

The performance of the hybrid model in terms of a traditional classification table is depicted in Table 4. It shows that if the 30 percentile posterior probability value is used as the threshold to separate crashes from non-crash cases, 30% of 1145 (=1096 + 49) validation dataset observations, i.e., 344 observations, will be classified as crashes. Hence, according to the hybrid model, more than 57% of the crashes (i.e., 28 of 49) will be identified by declaring 344 patterns as crash. Among the rest 801 (=1145 − 344) observations, there will be 21 missed crashes and 780 non-crash cases which are correctly identified. It translates into about 71.17% (780 of 1096) non-crashes correctly identified. Therefore, the model achieves more than 71% classification accuracy over non-crash cases and 57% accuracy over crash cases. The cells of Table 4 depicting these percentages are highlighted in Table 4.

Table 4
Classification performance of the hybrid mode over the validation dataset if 30 percentile posterior probability output is used as the threshold

	Predicted		Total
	0	1	
Actual			
0	Frequency = 780	316	1096
	Percent = 68.12	27.60	
	Row Pct = 71.17	28.83	
	Col Pct = 97.38	91.86	
1	21	28	49
	1.83	2.45	
	42.86	57.14	
	2.62	8.14	
Total	801	344	1145
			100.00

5. Discussion of results

The hybrid model utilized traffic parameters from the stations located immediately upstream and downstream of the historical crash locations as inputs. Therefore, it may be used to assess the crash risk between the sections of the freeway located between a pair of loop detector stations.

The formulation of the problem along with the solution approach adopted here is somewhat similar to incident detection. However, the objective of the analysis is to identify crash prone conditions, i.e., the conditions in which drivers are more likely to make errors resulting in lane-change related crashes, rather than pin point the occurrence of a crash. It allows for more flexibility since conditions prior to crashes (present research problem) are not as readily identifiable (possibly due to significant human factor involvement) as the conditions following the crashes (approach for incident detection). Crashes being such rare events, it is not possible to fully avoid the false alarms. As depicted in Table 4, even the modest 30% positive decisions (resulting from using 30 percentile values as the threshold) would result in a significant number of 'false alarms'. One may bring it down to an extent by using a higher threshold (e.g., 20 percentile value for the posterior probability), it would still remain significant. Traffic parameters from time-slice 1, if used as inputs instead of the parameters from time-slice 2, are also expected to provide slight improvement. However, time-slice 1 being too close to time of the crash would leave absolutely no leverage in terms of time available to process, analyze and disseminate the information that may in turn be used to avoid crashes.

It should be noted that 'false alarms' are not as detrimental in the present application as they are in case of incident detection algorithms. In fact, the ultimate goal of this research would, or at least should be, to 'achieve' a 'false alarm' every time a crash warning is issued. The goal would be based on the expectation that with some form of proactive countermeasure or warnings to the motorists, potential crashes following the crash prone conditions may be avoided. Such countermeasures are obviously a matter of detailed investigation but even without the countermeasures it is neither improbable nor unacceptable to have these 'false alarms'. Crash prone traffic conditions, which could be identified by the hybrid model developed in this paper, would not always result in a crash occurrence even though a significant proportion of historical crashes did occur under those conditions. These conditions are worth warning the drivers and drivers need to be more attentive under such traffic conditions even if they may not culminate in a lane-change related crash every time.

The justification or inevitability of false alarm does not mean that an unlimited number of warnings could be issued; especially if the information based on the model output is being transferred to the drivers on the freeway. The reason for being judicious about the number of warnings would be to ensure that the drivers do not perceive the number of warnings to be "too many" and become immune to them. The whole notion of warnings and drivers' reaction to them are beyond the scope of the present work and require further investigation.

6. Concluding remarks

A data mining based approach to identify potential lane-change related freeway crashes was presented in this paper. Based on the findings from Lee et al. (2006) and an extensive review of crash reports, it was concluded that all sideswipe crashes and angle crashes on inner lanes of the freeway may be attributed to lane changing maneuvers. These crashes were referred to as lane-change related crashes and are analyzed in this study. Based on variable selection procedures based on random as well as within stratum matched data it was concluded that the location specific characteristics do not have a significant effect on occurrence of a lane-change related crash. Note that it does not imply that all locations on the freeway are expected to have similar frequency of these crashes. It means that if locations with certain geometric characteristics experience more lane-change related crashes, these occurrences are better correlated with the traffic conditions existing before these crashes than the geometric characteristics that might be causing the crash prone traffic conditions. It was also noticed that the intensity of lane changes, measured in terms of overall average flow ratio (OAFR), was not significant to separate crashes from non-crash cases. It is interesting because the OAFR was successfully used to classify lane-change related crashes from rear-end crashes by Lee et al. (2006).

The variables found significant in the final analysis were average speeds upstream and downstream of the crash site. Average differences between adjacent lane occupancies upstream of the crash site (ADALOU2) along with standard deviation of volume and speed (SVW2 and SSW2) downstream were also found to be associated with lane-change related crashes. These variables (shown in Table 1) were used as inputs to classification models based on two neural network architectures (MLP and NRBF). It was found that the MLP model with four and NRBF model with three hidden neurons were the single best models in their respective classes. The hybrid model created by combining these two models bettered the performance of individual models in terms of crash identification over the validation dataset. This model is recommended to assess the risk of a lane-change crash between two loop detector stations on the freeway. It should be mentioned that even though only the models using data from time-slice 2 (5–10 min before the crash) are described here, models using data from time-slice 3 to 4 were also attempted but as expected they did not achieve the performance comparable to the models described. Also, time-slice 1 traffic parameters might have produced slightly better results, but being too close to actual time of crash they cannot be used in a real-time application due to practical considerations.

Through an online application of the final hybrid model the risk of a lane-change related crash may be continuously estimated between any two loop detector stations provided the data from all three lanes are available at those stations. Based on the measure of risk, i.e., the posterior probability output from the model, decision can be made about warning the motorists on the freeway. A reasonable number of warnings based on the hybrid model output can potentially play a critical role in proactive traffic management. These warnings may be issued to the

motorists driving on the freeway locations through variable message signs (VMS). Messages discouraging the drivers to change lanes could also be an alternative for reducing the risk of lane-change related crashes. However, the frequency and impacts of such warnings/messages on driver behavior call for further research and should therefore be pursued in the future.

References

Abdel-Aty, M.A., Pande, A., 2005. Identifying crash propensity using specific traffic speed conditions. J. Saf. Res. 36, 97–108.

Abdel-Aty, M.A., Pande, A., Hsia, L., Abdalla, F., 2005a. The potential of loop detector data for improving safety. ITE J. Web, 69–75.

Abdel-Aty, M.A., Uddin, N., Pande, A., 2005b. Split models for predicting multi-vehicle crashes under high-speed and low-speed operation conditions on freeways. Transport. Res. Rec. 1908, 51–58.

Abdel-Aty, M.A., Uddin, N., Abdalla, F., Pande, A., Hsia, L., 2004. Predicting freeway crashes based on loop detector data using matched case–control logistic regression. Transport. Res. Rec. 1897, 88–95.

Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Chang, G., Kao, Y., 1991. An empirical investigation of macroscopic lane-changing characteristics on uncongested multilane freeways. Transport. Res. Part A 25 (6), 375–389.

Christodoulou, C., Georgiopoulos, M., 2001. Applications of Neural Networks in Electromagnetics. Artech House, Boston, MA.

Cybenko, C., 1989. Approximations by superposition of sigmoid functions. Math. Control Signals Syst. 2, 303–314.

Golob, T.F., Recker, W.W., 2001. Relationships among urban freeway accidents, traffic flow, weather and lighting Conditions. California PATH Working Paper UCB-ITS-PWP-2001-19. Institute of Transportation Studies. University of California, Berkeley, CA.

Golob, T.F., Recker, W.W., 2004. A method for relating type of crash to traffic flow characteristics on urban freeways. Transport. Res. Part A 38 (1), 53–80.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. M.I.T Press, Cambridge, MA.

Lee, C., Abdel-Aty, M.A., Hsia, L., 2006. Potential real-time indicators of sideswipe crashes on freeways. In: Proceedings of the 85th Annual Meeting of Transportation Research Board (Paper #06-0017), Washington, DC.

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transport. Res. Rec. 1784, 1–8.

Lee, C., Saccomanno, F., Hellinga, B., 2003. Real-time crash prediction model for the application to crash prevention in freeway traffic. Transport. Res. Rec. 1840, 68–77.

Oh, C., Oh, J., Ritchie, S., Chang, M., 2001. Real time estimation of freeway accident likelihood. In: Proceedings of the 80th Annual Meeting of Transportation Research Board, Washington, DC.

Pande, A., 2005. Estimation of hybrid models for real-time crash risk assessment on freeways. Ph.D. Dissertation, University of Central Florida, Orlando, FL.

Pande, A., Abdel-Aty, M.A., 2006. A comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. In: Proceedings of the 85th Annual Meeting of Transportation Research Board (Paper #06-0016), Washington, DC.

Powell, M.J.D., 1987. Radial basis function approximations to polynomials. In: Proceedings of 12th Biennial Numerical Analysis Conference, Dundee.

Rumelhart, D.E., Hinton, G., Williams, R., 1986. Learning internal representation by error propagation, parallel distributed processing. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Explorations in the Microstructure of Cognition, vol. 1: Foundations. MIT Press, Cambridge, MA, pp. 318–362.

SAS Institute, 2001. Getting Started with Enterprise Miner Software. Release 4.1. SAS Institute, Cary, NC.

Tao, K.M., 1993. A closer look at the radial basis function (RBF) networks. In: Singh, A. (Ed.), Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, vol. 1. IEEE Comput. Soc. Press, Los Alamitos, CA.

Tarassenko, L., Roberts, S., 1994. Supervised and unsupervised learning in radial basis function classifiers. IEEE Proc. Vis., Image Signal Process. 141, 210–216.

Wang, J., Knipling, R., 1994. Lane change/merge crashes problem size assessment and statistical description. National Highway Traffic Safety Administration, Report No. DOT HS 808 075. US Department of Transportation, Washington, DC.

Wilamowski, B.M., Iplikci, S., Kaynak, O., Efe, M.O., 2001. An algorithm for fast convergence in training neural networks. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN'01), Washington, DC.