Accident Analysis and Prevention 83 (2015) 90–100

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Modeling driver stop/run behavior at the onset of a yellow indication considering driver run tendency and roadway surface conditions

Mohammed Elhenawy[a,1], Arash Jahangiri[b,1], Hesham A. Rakha[c,*], Ihab El-Shawarby[b,d,1]

[a] Bradley Department of Electrical and Computer Engineering, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States
[b] Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States
[c] Charles E. Via, Jr. Department of Civil and Environmental Engineering, Virginia Tech Transportation Institute, 3500 Transportation Research Plaza, Blacksburg, VA 24061, United States
[d] Ain-Shams University, Cairo, Egypt

ARTICLE INFO

Article history:
Received 1 November 2014
Received in revised form 19 April 2015
Accepted 18 June 2015
Available online 28 July 2015

Keywords:
Dilemma zone
Yellow indication
Run/stop behavior
Classification

ABSTRACT

The ability to model driver stop/run behavior at signalized intersections considering the roadway surface condition is critical in the design of advanced driver assistance systems. Such systems can reduce intersection crashes and fatalities by predicting driver stop/run behavior. The research presented in this paper uses data collected from two controlled field experiments on the Smart Road at the Virginia Tech Transportation Institute (VTTI) to model driver stop/run behavior at the onset of a yellow indication for different roadway surface conditions. The paper offers two contributions. First, it introduces a new predictor related to driver aggressiveness and demonstrates that this measure enhances the modeling of driver stop/run behavior. Second, it applies well-known artificial intelligence techniques including: adaptive boosting (AdaBoost), random forest, and support vector machine (SVM) algorithms as well as traditional logistic regression techniques on the data in order to develop a model that can be used by traffic signal controllers to predict driver stop/run decisions in a connected vehicle environment. The research demonstrates that by adding the proposed driver aggressiveness predictor to the model, there is a statistically significant increase in the model accuracy. Moreover the false alarm rate is significantly reduced but this reduction is not statistically significant. The study demonstrates that, for the subject data, the SVM machine learning algorithm performs the best in terms of optimum classification accuracy and false positive rates. However, the SVM model produces the best performance in terms of the classification accuracy only.

1. Introduction

With advances in sensing, communications, and computational technologies, research in the area of vehicle safety is increasing. Most new vehicles have active safety features including anti-lock braking and adaptive cruise control systems to reduce road accidents (Jones, 2001). In the US, the Department of Transportation (DOT) reported 32,367 fatalities caused by road accidents in 2011 (Tibshirani et al., 2009). A significant percentage of these road accidents occurred at signalized intersections as a result of driver behavior in the decision/dilemma zone while approaching signalized intersections (U.S. Department of Transportation and Federal Highway Administration, 2014).

Drivers approaching a traffic signal yellow indication have to decide whether to stop or proceed through the intersection. Typically, drivers far from the intersection, when a yellow indication is initiated, tend to stop while others near the intersection tend to proceed. A dilemma zone is a spatial stretch of roadway upstream of the intersection stop line that exists when the minimum stopping distance d_s is larger than the maximum clearing distance d_r, as illustrated in Fig. 1. In this case, drivers encountering the onset of yellow interval while traveling between d_s and d_r have no valid option (i.e., they cannot stop comfortably nor can they run before the traffic signal indication turns red). The minimum stopping distance is the distance required by a vehicle to safely come to a complete stop upstream of the intersection stop bar at a reasonable deceleration level (assumed to be 3 m/s^2). The maximum clearing distance is the distance within which the

* Corresponding author. Fax: +1 540 231 1555.
E-mail addresses: mohame1@vt.edu (M. Elhenawy), arashj@vt.edu (A. Jahangiri), hrakha@vt.edu (H.A. Rakha), shawarby@vt.edu (I. El-Shawarby).
[1] Fax: +1 540 231 1555.

http://dx.doi.org/10.1016/j.aap.2015.06.016
0001-4575/

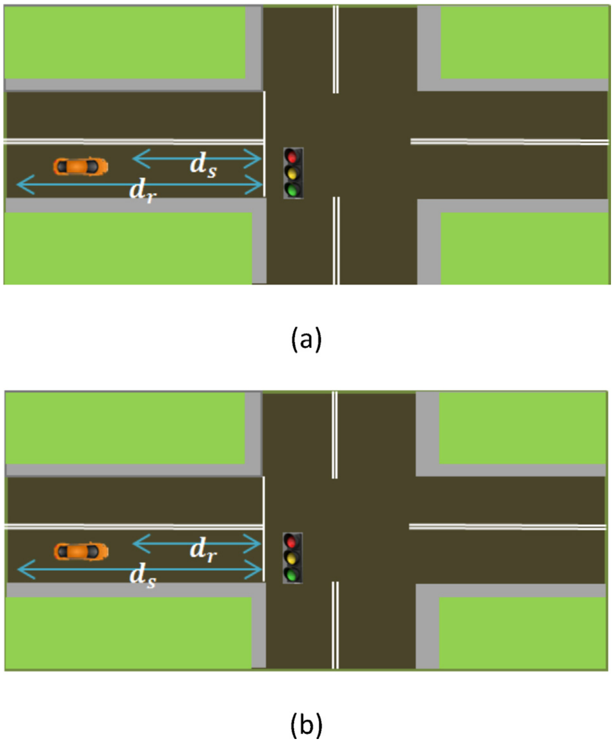

Fig. 1. Illustration of the option zone in panel (a) and DZ in panel (b).

vehicle can cross the intersection stop bar before the conclusion of the yellow interval. If the distance of an approaching vehicle to the intersection stop bar (DTI) at the onset of a yellow indication is between d_s and d_r (i.e., $d_r < \text{DTI} < d_s$) then the vehicle is in the DZ. Alternatively, if d_s is less than d_r, then vehicles at DTIs between d_s and d_r ($d_s < \text{DTI} < d_r$) are classified as being in the option zone and have two valid choices (stop safely or proceed safely). The DZ was first introduced by Gazis et al. (1960) and has been studied in many other studies (Rakha et al., 2007; Sheffi and Mahmassani, 1981; Bonneson et al., 2002; Gates et al., 2007; Pant et al., 2005; Chang et al., 1985; Zegeer, 1978; Liu et al., 2007; Wei et al., 2011; Ghanipoor Machiani and Abbas, 2014a; Abbas et al., 2014). The design of yellow timings is made in order to avoid the creation of the DZ.

Several factors influence driver behavior at the onset of yellow that result in the potential existence of a DZ. The factors can be divided into three categories; driver-related, intersection related and vehicle-related. These factors that have been studied throughout the literature (Rakha et al., 2007; Gates et al., 2007; Liu et al., 2007; Wei et al., 2011; Ghanipoor Machiani and Abbas, 2014a; Caird et al., 2007; El-Shawarby et al., 2007; Li et al., 2012; Jahangiri et al., 2015) include the driver perception–reaction time; the driver's acceptable deceleration level; the driver's age; the driver's gender; the time-to-intersection (TTI) at the onset of yellow; the distance-to-intersection (DTI) at the onset of yellow; the vehicle's approach speed; the vehicle type; presence of side-street vehicles, pedestrians, bicycles, or opposing vehicles waiting to turn left; the arrival rate; the length of the yellow interval; the cycle length; and presence of police vehicles in the vicinity of the intersection. Moreover, El-Shawarby et al. (2015) compared driver stopping/running probabilities in clear weather and in rainy weather and found a slight shift between the two probabilities. El-Shawarby et al. (2015) correlated this shift to the decrease in the probability of stopping in case of wet pavement surface and rainy weather conditions. Consequently, the roadway surface condition was added as an input variable to the proposed classifiers. The proposed driver aggressiveness factor has not been considered in past studies.

The intersection safety needs identification report published by the Federal Highway Administration (FHWA) in July 2009 showed that in 2007, 22% of the total fatal crashes were intersection-related with an estimated cost of 27.8 US billion dollars while 44.8% of the total injury crashes were also intersection-related with an estimated cost of 51.3 US billion dollars (Coakley and Stollof, 2009). Based on the National Highway Traffic Safety Administration (NHTSA), two-thirds of all fatal crashes are caused by aggressive driving (Wei, 2008). Consequently, aggressive driving is critical in modeling driver stop/run behavior at signalized intersections; however, measuring driver aggressiveness may not be plausible. In a previous research study, five driver actions were used to measure aggressive driving behavior. These five measures include: short or long honk of the horn, cutting in front of other vehicles in a passing lane maneuver, cutting in front of other vehicles in a multi-lane passing maneuver, and passing one or more vehicles by driving on the shoulder and then cutting in (Shinar and Compton, 2004). Other studies classified drivers into three categories aggressive, conservative, and normal drivers based on their decision (stop/run) and the distance to the stop line when the traffic signal turned yellow (Liu et al., 2012). In our previous work, we proposed the use of the frequency of running a yellow indication as a measure of driver aggressiveness (Elhenawy et al., 2014). In the current study a more formal definition and formulation of the driver aggressiveness parameter is proposed. The measure proposed here is a continuous measure of aggressiveness that varies from zero (not aggressive) to one (very aggressive) as opposed to a categorical variable as was done in previous studies.

Consider a vehicle approaching a signalized intersection, our goal is to build a model that predicts driver stop or run behavior at the onset of the yellow indication. This model uses many predictors such as the TTI and driver's age to predict the driver decision. Because in real-life, different drivers behave differently, we added the proposed predictor to explain some of the variation between drivers based on their history. Such a model should be one of the main building blocks in more advanced driver assistance systems. These systems should be able to predict the driver behavior and warn them if their decision is incorrect. Moreover, it would warn the driver if there is any potential violation from other drivers/vehicles approaching the intersection. The system should ensure the algorithm produces minimum false positives in order to encourage drivers trust their output.

The past two decades have seen numerous research efforts and advances in both machine learning techniques and computer computational power. Many machine learning techniques require a large number of computations and are infeasible without computers. The available machine learning algorithms and computational power have made such techniques feasible for real-time implementation. Transportation engineers are among people who are interested in applying these algorithms to address transportation problems. This interest increases with the availability of data sets from fixed detectors, data probes and intelligent transportation systems (ITSs). Recently, some machine learning algorithms were used in the transportation field, including: classifying and counting vehicles detected by multiple inductive loop detectors (Ali et al., 2012), identifying motorway rear-end crash risks using disaggregate data (Pham et al., 2010), automatic traffic incident detection (Liu et al., 2015), real-time detection of driver distraction (Yulan et al., 2007; Tango and Botta, 2013), transportation mode recognition using smartphone sensor data (Jahangiri and Rakha, 2014, 2015), and video-based highway asset segmentation and recognition (Balali and Golparvar-Fard, 2014). Modeling driver stop/run behavior at signalized intersections is very important and is ideal for applying machine learning techniques (Ghanipoor Machiani and Abbas, 2014b). At first glance, driver stop/run behavior modeling seems to be a good candidate for a straightforward application of machine learning algorithms. Observations of driver stop/run behavior from naturalistic driver datasets or from controlled field experiment datasets can be used to train machine learning algorithms. The trained models can then be used to predict future driver decisions for implementation in in-vehicle safety systems. However, machine learning modeling of driver stop/run behavior faces some challenges including the need for large labeled datasets, driver stop/run behavior drift, and computational complexity.

In this paper a new parameter related to the driver aggressiveness is proposed. This new predictor can be observed directly from historical driver stop/run behavior. Using this new predictor, we demonstrate that the modeling of driver stop/run behavior can be enhanced. The use of such models can then be integrated with in-vehicle safety systems to predict the action of a driver and thus warn other drivers or take action to ensure that no vehicle collisions occur.

2. Methods

We define our problem by defining the input variables (predictors) and the output (response). There are six predictors used as inputs to the model:

G is the gender (1 = female, 0 = male),

A is the age (years),

TTI is the time-to-intersection (s),

V is the approach speed (km/h),

S is the roadway surface condition (0 = rainy/wet surface, 1 = dry),

D is the new proposed driver aggressiveness predictor which will be discussed in the next section.

The output of the model is a label that indicates the driver decision to either stop or run. This type of modeling problem is a binary classification problem. Classification is one of the main areas in machine learning. In order to build such models, labeled data are needed. Once the model is built, it can be used to classify (predict) the behavior of the driver at onset of yellow light based on the values of the input predictors. There are numerous machine learning classification algorithms that are suitable for such type of modeling. A brief introduction to the logistic regression, random forest (RF), AdaBoost, and support vector machine (SVM) is presented to familiarize readers with these emerging techniques. The strengths of each modeling technique are presented given that the models will be compared later on the same dataset.

Logistic regression and the used machine learning algorithms require two efforts. The first effort is training and the second effort is testing. The training effort for the machine learning approach requires more computations compared to logistic regression, however this is not detrimental because training is usually done offline. Testing of both logistic regression and machine learning algorithms does not require much computation especially with the presence of existing powerful computers. Another issue that may emerge is re-training when new data become available. Regarding, the re-training issue, the data set used for training any model should constitute a representative sample of the entire population such that re training the model after adding new cases will not significantly improve the model. In the worst case if the training data were not a good representative of the population then additional retraining of the model may be required.

2.1. Generalized linear models

Generalized linear models (GLMs) are statistical models that model responses as linear combinations of regressors (predictors). It can be used as regression models for continuous dependent variables, models for rates and proportions, binary, ordinal and multinomial variables, and counts. The GLM approach has two significant advantages: (a) its theoretical framework is valid for many commonly encountered datasets; (b) the software implementation is simple because the same algorithm can be used for estimation, inference and assessing model adequacy for all GLMs. Logistic regression is a commonly used GLM when the response variable is a binomial response and is thus used to model driver stop/run behavior, as was demonstrated in an earlier study (Amer et al., 2010).

2.2. Random forests

The random forest (RF) is an ensemble approach that is an effective tool in prediction (Breiman, 2001). Breiman used the Strong Law of Large Numbers and proved that RFs do not suffer from over-fitting as additional trees are added (Breiman, 1999). The main idea behind ensemble methods is that building a large group of simple models will give an overall improved performance. The random forest is a large group of un-pruned decision trees with randomized selection of features at each split. The well-known machine learning technique, classification and regression tree (CART), is one of the common decision trees used in random forests (Breiman, 1984). Random forests start with the CART that, in ensemble terms, corresponds to the weak model. CART starts by splitting the feature space into two partitions (children) such that its objective function is locally optimized. For each child, CART repeats this splitting process until the stopping criteria is reached. The cases in each region have (almost) the same outcome. The random forest algorithm for classification can be simply described as, assuming the training dataset has *H* cases, *P* predictors, and *M* trees to build for each of the *M* iterations, in the following steps:

1. Sample *H* cases at random with replacement to create a bootstrap sample from the original dataset. The subset should be approximately 66% of the original training set and the other cases are repeated cases.
2. For some number, $\sqrt{p}$ at each node, $\sqrt{p}$ predictor variables are selected randomly from all the predictor variables.
3. The predictor variable out of the $\sqrt{p}$ predictors that provides the best split is used to produce a binary split on that node.
4. At the next node, choose randomly another $\sqrt{p}$ predictors from all regressor variables and do the same.
5. Do not perform cost complexity pruning and save the tree as is with the other built trees produced from earlier iterations.

At the testing phase, the new arrived case is pushed down all the trees. Each tree votes for one class by providing a class label. The output of the random forest is the class which has most votes. This modeling technique could improve the modeling of driver stop/run behavior and thus will be tested as part of this research effort.

2.3. Adaptive boosting algorithm

The adaptive boosting (AdaBoost) is a machine learning algorithm that is based on the idea of incremental contribution (Freund and Schapire, 1999). AdaBoost was introduced as an answer to the question of whether a group of “weak” learner algorithms that each has low accuracy can be combined into a learning algorithm with high accuracy. Before introducing the idea of AdaBoost, the traditional approach in machine learning was based on choosing the most possible class of discriminating features. In other words algorithms are required to be as class discriminatory as possible. Subsequently, the algorithm uses these features to find the most discriminating learning algorithm to predict the class label for an unseen data instance, labeled data are collected and used as a training dataset. Afterwards, feature selection is done. If the number of features is larger than the number of instances, then the data suffer from the curse of dimensionality (Clarke et al., 2008). At this point, principal components analysis (Smith, 2002) or independent component analysis (Hyvärinen and Oja, 2000) can be used to reduce the space dimensions and address the dimensionality problem. The last step after defining the feature space is choosing an algorithm such as K-NN (Tibshirani et al., 2009), support vector machine (Hsu et al., 2003), or parametric models that give the highest accuracy.

AdaBoost does not use one classifier; instead it uses a set of weak classifiers, each is trained using the same training dataset but with a different weight distribution. Each of the weak learners focuses on the instances that are misclassified by the previous learner. The output of AdaBoost is the weighted average of all weak learner outputs. AdaBoost is likely to have smaller misclassification error compared to the summation of weak learners; also it has a bound on the generalization error (Freund and Schapire, 1999; Friedman et al., 2000) . To describe the AdaBoost algorithm, let us assume the training set consists of K instances $T = \{(x_1, y_1), \ldots, (x_K, y_K)\}$, here x_i is the vector of input predictors that can be represented by a point in the multidimensional feature (predictor) space and y_i is the corresponding label (response). Because the algorithm predicts whether the driver will run or stop, the focus will be on the binary classification problems with the label y_i is either +1 or −1. The pseudo-code of the classic AdaBoost is described in Table 1.

After training T weak learners the model is ready to predict the label for test instance (unseen) x_{test}. The label of the test instance is defined using Eq. (1).

Table 1
Proposed AdaBoost algorithm pseudo code.

Set a probability distribution $P_t(x_i)$ over all the training samples. Initially, $P_t(x_i)$ is set to be uniform then it is modified iteratively with each selection of a weak classifier.

for iteration t do

1. Train weak learner L_t with weighted sample.
2. Test weak learner L_t on all data and get the predicted label $L_t(x_i)$ for each x_i.
3. Compare the predicted labels $L_t(x_i)$ with y_i for i=1,…,n and calculate the classification error ϵ_t.
4. Calculate the trustiness level $\propto_t$ of the L_t the following equation

$$\propto_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t}$$

5. Update F_t such that misclassified instances weights are increased using,

$$P_{t+1}(x_i) = \frac{P_t(x_i) e^{-\propto_t y_{iL_t(x_i)}}}{Z_t}$$

where Z_t serves as a normalize such that $\sum_{i=1}^{n} P_{t+1}(x_i) = 1$

end for

$$\text{sign}\left(\sum_{t=1}^{T} \propto_t L_t(x_{\text{test}})\right) \tag{1}$$

The label is set equal to 1 if the output of Eq. (1) is positive and −1 if the output is negative. Again this algorithm will be tested on the dataset to develop driver stop/run models.

2.4. Support vector machine

Support vector machine (SVM) is a rather complex machine learning technique that can be employed in classification problems. SVM is known as a large margin classifier, which means that while this method attempts to find decision boundaries between different classes, it tries to maximize the gap or margin between classes.

The objective function of the SVM formulation and the associated constraints are presented below in Eq. (2) through Eq. (4) (Hsu and Lin, 2002). The sum of two terms are minimized in the objective function; minimizing the first term is basically equivalent to maximizing the margin between classes, and the second term consists of an error term multiplied by the regularization (penalty) parameter denoted by C. The regularization is designed to deal with the problem of over-fitting a model. The value of the C parameter should be adjusted to obtain the best possible performance.

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T\mathrm{w} + C\sum_{n=1}^{K}\xi_n\right) \tag{2}$$

Subject to:

$$y_n(w^T\phi(x_n)+b) \geq 1-\xi_n, n=1,\ldots,K \tag{3}$$

$$\xi_n \geq 0, n=1,\ldots,K \tag{4}$$

where$n$ index of the data observation$w$ parameters to define decision boundary between classes$C$ regularization (or penalty) parameterξ_n error parameter to denote margin violation$b$ intercept associated with decision boundaries$\phi(x_n)$function to transform data from X space into some Z space$K$ the number of observations in the dataset

When using SVM, the data are transformed from the X space to the Z space using some function $\phi(x_n)$. The reason for conducting the transformation is to obtain a space in which identifying decision boundaries between classes becomes easier. However, in solving the problem, there is no need to actually do the transformation. Instead, some other functions, known as kernels, are adopted. The kernels, which appear in the dual formulation of the problem, correspond to the vector inner product in the Z space. To construct the model, the kernel type should be selected (e.g., linear, polynomial, Gaussian). Depending on the problem, one kernel may perform better than the other. Some practical considerations suggest using certain kernels for specific problems based on the data size (Hsu et al., 2003).

3. Proposed driver aggressiveness predictor

Models used to classify the driver stop/run behavior, use driver- and intersection-related predictors. Driver's age and gender are used as the driver-related predictors. But these predictors are not sufficient to measure the level of aggressiveness of the driver. Thanks to advances in telecommunication and computation power, connected vehicles have become a reality in which vehicles can exchange information with each other (V2V) and with the traffic signal controllers (V2I). We can use this technology advantage to allow vehicles to learn from the behavior of its driver by maintaining a small record describing many behavioral-related parameters. These parameters can describe the level of aggressiveness of the driver. One intuitive parameter we propose is the probability of running at the onset of a traffic signal yellow indication when stopping is a better decision.

We propose a new predictor that can be used as a measure of the driver's aggressiveness. The new measure is based on a count of the number of runs the driver makes when the time-to-intersection at the onset of the yellow indication is greater than the yellow time and their speed is equal to or greater than the posted speed limit. The value of the new predictor θ_j for driver j can be estimated using the Bayesian approach by finding the posterior density distribution, as shown in Eq. (5).

$$f\left(\theta_j|y_{j1},y_{j2},\ldots,y_{jN_j}\right) \propto f\left(y_{j1},y_{j2},\ldots,y_{jN_j}|\theta_j\right)f(\theta_j) \tag{5}$$

where

$f\left(y_{j1},y_{j2},\ldots,y_{jN_j}|\theta_j\right)$ is the sampling distribution for driver j

$f(\theta_j)$ is the prior distribution for driver j

N_j the number of cases for driver j when the time-to-intersection at the onset of the yellow indication is greater than the yellow time and his/her speed is equal or greater than the posted speed limit

The distribution of the $f\left(y_{ji}|\theta_j\right)$ is Bernoulli because the random variable is either one or zero. The new predictor is unitless, and it takes any value between zero and one. This means the domain of the $f(\theta_j)$ is from zero to one [0,1]. Consequently, a reasonable choice of $f(\theta_j)$ is a Beta distribution and the problem can be viewed as a Beta-Bernoulli model, as shown in Eq. (6).

$$f(y_{ji}|\theta_j) \sim \text{Bernoulli}(\theta_j) \tag{6}$$

where y_{ji} is case i for driver j

$(\theta_j) \sim \text{Beta}(a,b)$

$a = 1$ and $b = 1000$

$$f\left(\theta_j|y_{j1},y_{j2},\ldots,y_{jN_j}\right) \propto \left\{\Pi_{i=1}^{N_j}\theta_j^{y_{ji}}(1-\theta_j)^{1-y_{ji}}\right\}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta_j^{a-1}(1-\theta_j)^{b-1}$$

The above equation can be simplified as shown in Eq. (7).

$$f\left(\theta_j|y_{j1},y_{j2},\ldots,y_{jN_j}\right) \propto \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left\{\theta_j^{\Sigma_{i=1}^{N_j}y_i+a-1}(1-\theta_j)^{N_j-\Sigma_{i=1}^{N_j}y_i+b-1}\right\} \tag{7}$$

By removing the constants $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ from the above equation, the kernel of the posterior is a Beta distribution with the parameters shown in Eq. (8).

$$f\left(\theta_j|y_{j1},y_{j2},\ldots,y_{jN_j}\right) \sim \text{Beta}\left(\sum_{i=1}^{N_j}y_i+a, N_j-\sum_{i=1}^{N_j}y_i+b\right) \tag{8}$$

From Eq. (8) we can estimate the expectation $E\left[f\left(\theta_j|y_{j1},y_{j2},\ldots,y_{jN_j}\right)\right]$ and use it as the aggressiveness measure for driver j, as shown in Eq. (9).

$$E\left[f\left(\theta_j|y_{j1},y_{j2},\ldots,y_{jN_j}\right)\right] = \frac{\sum_{i=1}^{N_j}y_i+a}{a+b+N_j} \tag{9}$$

The new predictor is a parameter that is estimated using Eq. (9). This estimate approaches the true value as N_j increases. In our dataset the minimum N_j is zero and the maximum N_j is seven. If we can choose the Beta distribution's parameters such that the prior is informative, then the required value N_j to approach the true value decreases. To calculate the new predictor for each driver, the number of cases N_j are computed for which both the driver's time-to-intersection at the onset of the yellow indication is greater than the yellow time and the speed of the vehicle is greater than or equal to the posted speed limit. Let $y_{ji} = 1$ if the driver's decision was to run and $y_{ji} = 0$ otherwise. We adopt Eq. (9) to compute the aggressiveness parameter for each driver. The proposed aggressiveness parameter is not simply the percentage of time a driver stops/runs when encountering a yellow interval; instead it is a measure of the proportion of time the driver proceeded when they should have stopped. If the value of the new predictor is close to one that means the driver rarely stops and thus is more aggressive than a driver who has a smaller value of the new predictor. This new predictor is important because it captures the stop/run tendencies of that specific driver.

It is envisioned that the computation of this new predictor can be done through some form of infrastructure-to-vehicle (I2V) communication in which the vehicle receives signal phasing and timing (SPaT) information to identify the indication of the traffic signal. Moreover, vehicle-to-vehicle (V2V) communication would be required to exchange information with surrounding vehicles to ensure that the driver was not forced to stop because the vehicle ahead of it stopped. Using the SPaT and surrounding vehicle information the vehicle would count the number of times the driver stopped and ran the yellow indication when they had the freedom to proceed.

It is important to show how this new driver aggressiveness predictor can deployed in different vehicles. We have two categories of vehicles. The first category of vehicles is that always driven by the same driver/drivers such as a private and/or family vehicle. The second category are vehicles driven by different drivers (e.g., rental cars). For private and family vehicles, the vehicle may recognize the driver from their key, their voice, image, or finger print and save data for each family member that drives the vehicle.

For the commercial vehicles or any vehicle that does not recognize its driver, the system can initially start with a conservative assumption by assuming the driver is very aggressive and set the new predictor close to one. Subsequently, the vehicle updates the value of the new predictor as the driver passes through signalized intersections and data are gathered.

Another solution that may become feasible with recent advances in electronics is to equip the driver license with a read/write chip that carries the driver's information including their aggressiveness parameter. The vehicle can read the information on the driver's license and use the aggressiveness parameter of the driver.

4. Data description

The data used in this paper were collected from two different field experiments that were conducted on two different roadway surface conditions. Both experiments were conducted on the Virginia Department of Transportation's (VDOT) Smart Road facility, located at the Virginia Tech Transportation Institute (VTTI). The length of the Smart Road is a 3.5 km (2.2 mile). It is a two-lane road with one four-way signalized intersection. The entrance to this test facility is through an electronic gate to make it safe location to conduct field tests. The horizontal layout of the test section is fairly straight, and the vertical layout has a substantial grade of 3% (Rakha et al., 2001).

4.1. Dry roadway surface field experiment

This experiment was only run in clear weather and dry pavement surface condition. In this experiment, three vehicles were used, one was driven by test participants and accompanied with in-vehicle experimenter (Amer et al., 2010). The other two vehicles were driven by trained experimenters who were involved in the study. One of them following the test vehicle, whereas the other vehicle was crossing the intersection from the conflicting approach when the traffic light was green. The test vehicle is equipped with a real-time data acquisition system (DAS), differential global positioning system (GPS) unit, a longitudinal accelerometer, sensors for accelerator position and brake application, and a computer to run the different experimental scenarios.

The vehicle data stream was synchronized with changes in the traffic signal controller by the communication channel that linked the data recording equipment to the intersection signal controller. The phase changes were triggered by the test vehicle using a GPS unit that determined the distance from the intersection. Twenty-four licensed drivers were recruited in three equal age groups (under 40-years-old, 40–59-years-old, and 60-years-old or older); each group was male–female balanced.

Participants were asked to follow all normal traffic rules and to obey all traffic laws while driving. They drove loops on the Smart Road at a 72.4 km/h (45 mile/h) instructed speed, crossing the four-way signalized intersection 24 times for a total of 48 trials, where a trial consisted of one approach to the intersection. Among the 48 trials, a 4-s yellow indication at the 72 km/h (45 mile/h) instructed speed were triggered for a total of 24 times (four repetitions at six distances). The yellow indications were triggered when the front of the test vehicle was 40.2, 54.3, 62.5, 70.4, 76.5, and 82.6 m (132, 178, 205, 231, 251, and 271 ft) from the intersection for the 72 km/h (45 mile/h) instructed speed to ensure that the entire dilemma zone was within the range. On the remaining 24 randomized trials the signal indication remained green. The dry dataset included a complete tracking data every deci-second of the subject vehicle within about 150 m (500 ft) before and after the intersection.

4.2. Rainy/wet roadway surface field experiment

This experiment was only run in rainy weather and wet pavement surface conditions. in this experiment, two vehicles were used, because the platooning factor was not found to be a significant factor in the stop/go model (Amer et al., 2010). One vehicle was driven by test participants (accompanied by the in-vehicle experimenter) and the other vehicle was driven by a trained research assistant to simulate real-world conditions (Li et al., 2012). The confederate vehicle crossed the intersection from the side street when the signal was red for the test vehicle. The participant was asked to follow all normal traffic rules and to obey all traffic laws.

As was the case in the first study, the test vehicle was equipped with a differential global positioning system (GPS), a real-time data acquisition system (DAS), and a computer to run the different experimental scenarios. A communications link to the intersection signal control box was used by the data recording equipment to synchronize the vehicle data stream with changes in the traffic. The two vehicles were equipped with a communications system between vehicles, operated by the research assistants.

Twenty-six drivers were recruited in three age groups (under 40-years-old, 40–59-years-old, and 60 years of age or older), equal number of male and female participants were assigned to each group.

During the test runs the participants drove a distance of 1.6 km (1 mile) going downhill to approach the intersection followed by a 0.5 km (0.3 mile) leg to a high-speed turnaround, and another 0.5 km (0.3 mile) approach going back to the intersection. A 4-s yellow interval at the 72.4 km/h (45 mile/h) instructed speed was triggered for a total of 24 times (four repetitions at six distances).

The yellow indications were triggered when the front of the test vehicle was 54.3, 62.5, 70.4, 76.5, 82.6, and 92.7 m (178, 205, 231, 251, 271, and 304 ft) from the intersection to ensure that the entire dilemma zone was within the range. An additional 24 green trials were randomly introduced into the 24 yellow trials to introduce an element of surprise into the experiment. The run sequence was generated randomly and was different from one trial to another.

5. Results

This section presents the classification results of the logistic regression and the machine learning algorithms (i.e., random forest, AdaBoost, and SVM). Using the two datasets collected in the two previous studies, as described in the above section. The logistic model and the machine learning models are evaluated using both the classification accuracy and the false positive rate, computed using Eqs. (9) and (10), respectively. In our context stop is positive and run is negative. So, false positive (error type I) is predicting the driver stops while he actually runs which is very dangerous and may causes crashes.

Our goal is to identify the model that has;

1. The highest classification accuracy to increase the user's acceptance of the advanced driver assistance system.
2. The lowest false positive rate to provide the driver with the highest degree of safety.

$$\text{Classifiaction accuracy} = 100 \times \frac{\sum \text{True classified run} + \sum \text{True classified stop}}{\sum \text{Actual run} + \sum \text{Actual stop}} \qquad (9)$$

$$\text{False positive rate(FPR)} = \frac{\sum \text{misclassified run}}{\sum \text{Actual stop}} \qquad (10)$$

For each classifier, the average of classification accuracy and FPR of each set of the trials are calculated using the leave-one-out (LOO) cross-validation method (Evgeniou et al., 2004). In the LOO approach, the classifier is built using labeled data from all drivers except one single driver. This driver is used as the unseen test driver and the classification accuracy and FPR are calculated for that driver. The entire process is repeated where each driver is used once as a test driver and the average classification accuracy is calculated across all drivers.

5.1. Logistic regression

We built two logistic models, one without the new predictor and the other with the new predictor. The Bayesian information criterion (BIC) of the model without the new predictor is 1018.1 and after adding the new predictor becomes 877.5. The Akaike information criterion (AIC) of the model without the new predictor is 987.6 and after adding the new predictor is reduced to 841.9. Both the BIC and AIC parameters demonstrate that the model with the proposed aggressiveness predictor is better. The parameter estimates are shown in Tables 2 and 3.

Table 2
Parameter estimates of the logistic model without the new predictor.

Term	Estimate	Std error	ChiSquare	Prob > ChiSq
Intercept	4.05225895	2.192154	3.42	0.0645
Time to intersection (TTI)	−2.8928114	0.1708661	286.63	<0.0001
Age (A)	−0.0195942	0.0048147	16.56	<0.0001
Gender (G)	0.25835955	0.162566	2.53	0.1120
Speed (V)	0.07641985	0.0286772	7.10	0.0077
Road condition (S)	−0.2666143	0.1621986	2.70	0.1002

As shown in Table 2 all factors are significant except the roadway surface condition and the driver gender. After adding the new predictor, which is shown in Table 3 to be significant, the vehicle approach speed became insignificant in addition to the roadway surface condition and gender variables.

Eqs. (11) and (12) show the driver stop/run behavior model.

$$\ln\frac{P_s}{P_r} = \ln\frac{P_s}{1-P_s} = 4.05 - 2.89\text{TTI} - 0.02A + 0.26G + 0.08V - 0.27S \qquad (11)$$

$$\ln\frac{P_s}{P_r} = \ln\frac{P_s}{1-P_s} = 5.85 - 3.45\text{TTI} - 0.02A - 0.34G + 0.04V - 0.16S + 2696.5D \qquad (12)$$

where P_s is the probability of stopping and P_r is the probability of running. The classification accuracy and FPR of the logistic models without the new predictor are 80.76 and 0.2451, respectively. After adding the new predictor the classification accuracy is raised to 84.19 and the FPR is reduced slightly to 0.2417.

5.2. AdaBoost and random forest

The sequence in the above subsection was repeated using the AdaBoost machine and random forest learning algorithms. Figs. 2 and 3 show the classification accuracy and FPR of the AdaBoost models with and without the new predictor for different number of weak learners. As shown in the top panel of Fig. 2, the light blue bars show the change in the classification accuracy with the number of weak learners. Although there is a variation in the accuracy; the trend is increasing with an increase in the classification accuracy as the number of the weak learners increases. We can overcome this fluctuation by increasing the number of weak learners at the cost of increasing training time and testing time. The bottom panel in Fig. 2 shows some fluctuation in the false positive in both cases (with and without the new predictor). By increasing the number of weak learners we can reach a platoon where the false positive become almost unchanged as we increase the number of weak learners. These figures show that random forest and AdaBoost with the new predictor are always better in terms of classification accuracy and the false positive rate irrespective of the number of weak learners or trees.

5.3. Support vector machine

In implementing the SVM, the LibSVM library of SVMs was applied (Chang and Lin, 2011). With regard to the size of the data Gaussian kernel was selected to adopt for model development (Hsu et al., 2003). Furthermore, complete model selection was conducted by changing the regularization parameter and the Gaussian parameter to achieve the highest performance. Classification accuracy of about 90% and 86% were obtained with and without including the new predictor, respectively. Moreover, the SVM model resulted in FPR of about 24% and 17% with and without the new predictor, respectively. As mentioned earlier, LOO cross-validation technique was applied to assess the model. Fig. 4 presents the complete model selection for two of the SVM models, with and without the new predictor, while using LOO cross-validation technique. Fig. 4 illustrates how varying the model parameters (i.e., regularization and Gaussian parameters) affects the cross-validation accuracy. In fact the best performance (i.e.,

Table 3
Parameter estimates of the logistic model with the new predictor.

Term	Estimate	Std error	ChiSquare	Prob > ChiSq
Intercept	5.85016118	2.3571039	6.16	0.0131
Time to intersection (TTI)	−3.4510272	0.2081167	274.97	<0.0001
Age (*A*)	−0.0174145	0.0052297	11.09	0.0009
Gender (*G*)	−0.3409965	0.1854972	3.38	0.0660
Speed (*V*)	0.03591772	0.030821	1.36	0.2439
Roadway condition (*S*)	−0.1600429	0.1768756	0.82	0.3656
New predictor	2696.46047	255.80208	111.12	<0.0001

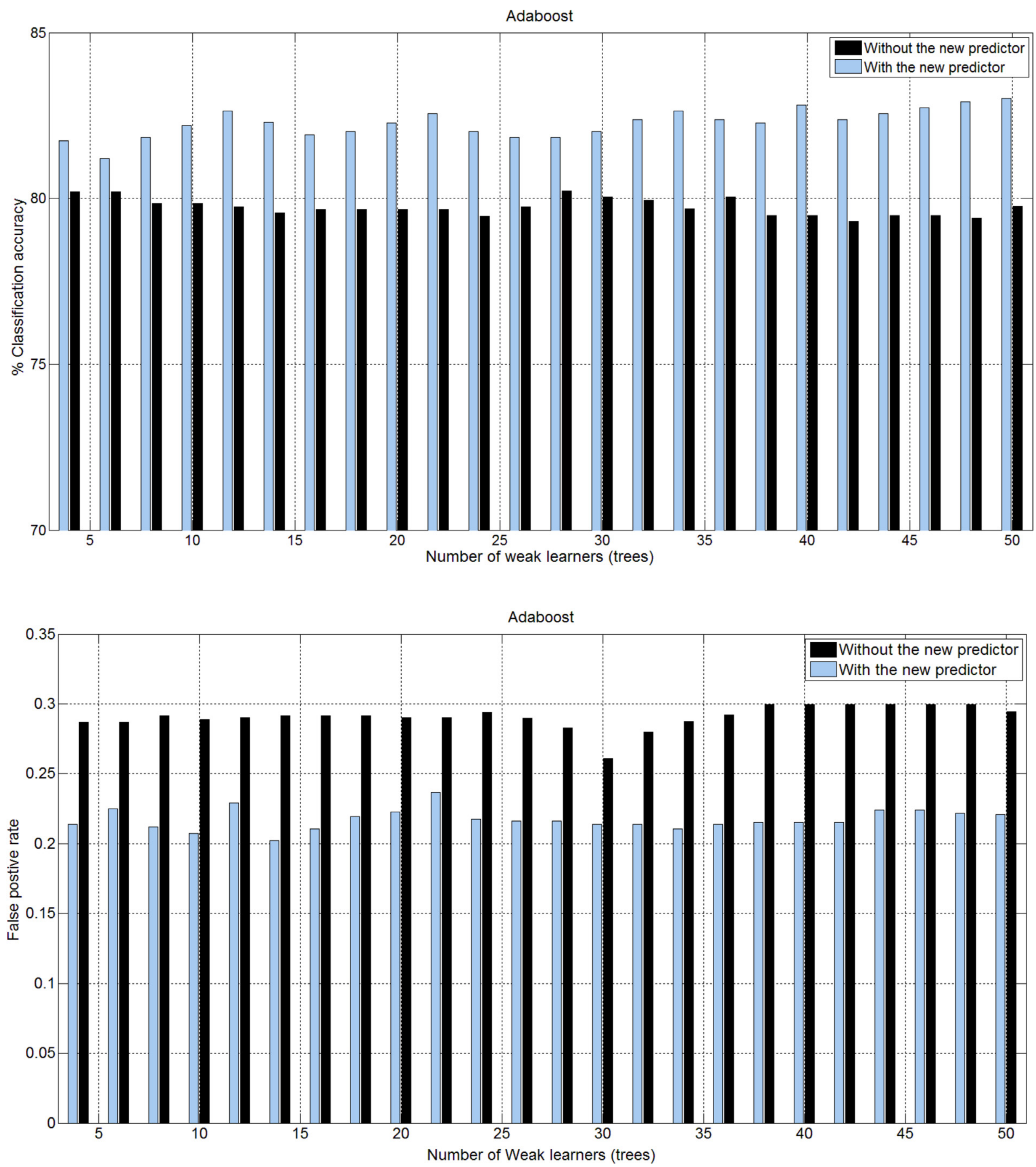

Fig. 2. The classification accuracy (above panel) and the FPR (bottom panel) using a different number of weak learners. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

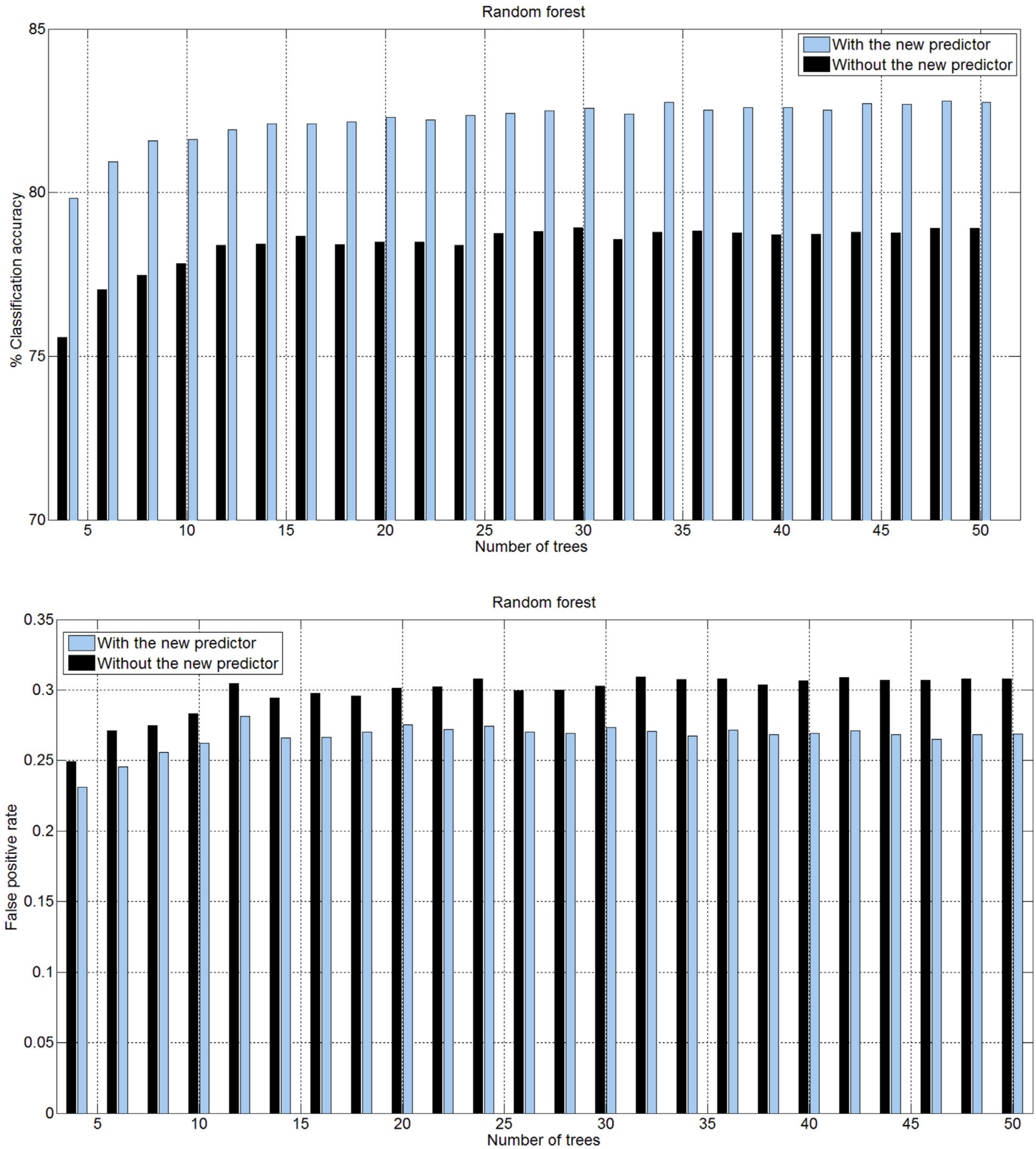

Fig. 3. The classification accuracy (above panel) and the FPR (bottom panel) using a different number of CARTs.

accuracy of 90% and 86%) for the two SVM models were obtained by adjusting the model parameters as presented in the figure.

5.4. Model comparison

The four developed classifiers were compared in terms of the false positive rate (false alarms) and the classification accuracy, as shown in Table 2. All four classifiers were comparable with and without the new predictor in terms of classification accuracy. The AdaBoost was the best classifier in terms of low false alarms when compared with the other three classifiers. However, if only the overall classification accuracy is considered, the SVM model with the new predictor achieved the highest accuracy despite the slightly higher false positive rate. The paired *t*-test was used to test the significance of the improvement in both FPR and classification accuracy between the classifier without the new predictor and after adding the new predictor. The sample test statistics is $T = \overline{d}/\mathrm{SE}(\overline{d})$ here di is the difference between the two observations for each pair, $\overline{d}$ is the mean of the difference and $\mathrm{SE}(\overline{d})$ is the standard error of the mean difference. In both tests (FPR and classification accuracy tests), our null hypothesis is that the distribution of differences come from a distribution with a mean of zero, assuming both classifiers (without and with the new predictor) are equivalent and the new predictor does produce any significant effect. The alternative hypothesis is that the differences tend to be different from zero and the new predictor significantly improves the classifier.

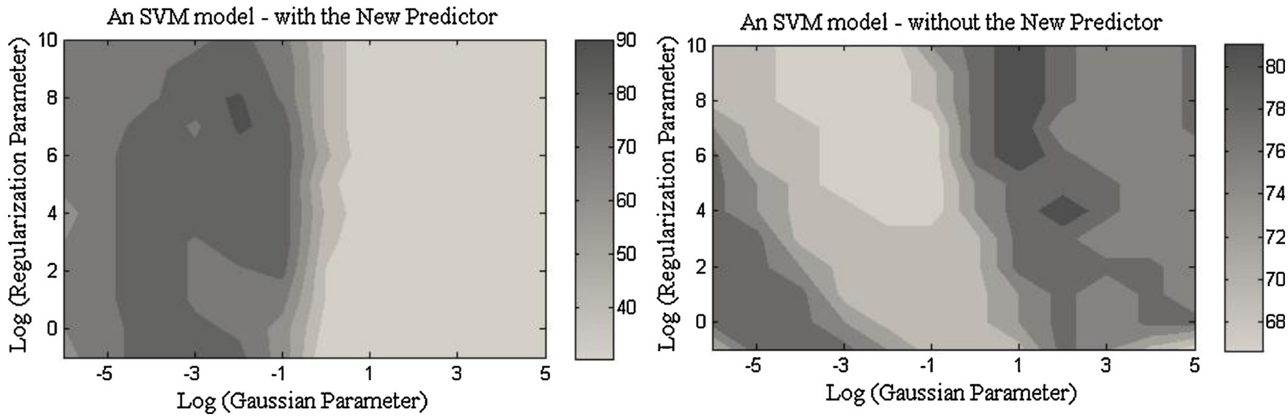

Fig. 4. Complete model selection for SVM models.

Table 4
Comparison between different classifiers.

Method	Evaluation measure	Without new predictor	With new predictor	*p*-value
Logistic regression	Classification accuracy (%)	80.76	84.19	0.0120
	False positive (%)	24.51	24.17	0.9290
AdaBoost	Classification accuracy (%)	79.77	83.00	0.0499
	False positive (%)	29.42	22.06	0.0441
Random forest	Classification accuracy (%)	78.90	82.77	0.0356
	False positive (%)	30.80	26.88	0.1974
SVM	Classification accuracy (%)	86.46	90.02	0.0187
	False positive (%)	22.34	20.73	0.4219

Table 4 shows the *p*-values that are extracted from the t-distribution with degree of freedom equal to the number of drivers less one. The experimental results show that the AdaBoost model with the added new predictor has a statistically significant lower FPR and statistically significant higher classification accuracy at a 0.05 significance level. The other three classifiers have only one significant improvement in terms of classification accuracy. In other words, including the new predictors for these three models resulted in a statistically significant improvement in the classification accuracy, but the reduction in the FPR was not statistically significant at the 0.05 significance level.

6. Study conclusions and future work

In this paper, we introduced a measure of driver aggressiveness into the modeling of driver stop/run behavior at the onset of a yellow indication. The driver aggressiveness parameter can be estimated by monitoring the driver historical response to yellow indications. The new aggressiveness parameter is based on the count of the number of runs the driver makes when the time to intersection, at the onset of the yellow indicator, is greater than the yellow time and their speed is equal to or greater than the posted speed limit. The parameter can then be added to the model after some period of monitoring. The experimental results demonstrate the ability of the new predictor to explain part of the variability in the driver stop/run decision. Specifically, there is a statistically significant increase in the classification accuracy. The FPR is significantly reduced but this reduction is not statistically significant. The study also demonstrates that the AdaBoost machine learning algorithm is the best algorithm in terms of its statistically significant improvement in FPR. However, the SVM model had the best performance in terms of the classification accuracy.

Further enhancements to the model are required to model driver stop/run behavior under more severe inclement weather (such as snow, freezing rain, ice and fog) and road surface conditions, considering the impact of the vehicle type (bus or truck) on the driver behavior, and developing real-time machine learning techniques that can adapt to changes in driver behavior.

Acknowledgements

The authors acknowledge the support of the Center for Technology Development and Smart Road Operations Group at the Virginia Tech Transportation Institute (VTTI), Ahmed Amer, Huan Li for running the tests. This work was supported in part by grants from the Virginia Center for Transportation Innovation and Research (VCTIR), the Mid-Atlantic Universities Transportation Center (MAUTC), and SAFETEA-LU funding.

References

Abbas, M, Machiani, S.G., Garvey, P.M., Farkas, A., Lord-Attivor, R., 2014. Modeling the Dynamics of Driver's Dilemma Zone Perception Using Machine Learning Methods for Safer Intersection Control.

Ali, S.S.M., Joshi, N., George, B., Vanajakshi, L., 2012. Application of random forest algorithm to classify vehicles detected by a multiple inductive loop system. 15th International IEEE Conference on Intelligent Transportation Systems (ITSC) 491–495.

Amer, A.M.M., Rakha, H.A., El-Shawarby, I., 2010. A behavioral modeling framework of driver behavior at onset of yellow a indication at signalized intersections. Presented at the Transportation Research Board 89th Annual Meeting, Washington, D.C.

Balali, V., Golparvar-Fard, M., 2014. Scalable nonparametric parsing for segmentation and recognition of high-quantity, low-cost highway assets from car-mounted video streams. Construction Research Congress 2014@ sConstruction in a Global Network 120–129.

Bonneson, J.A., Middleton, D., Zimmerman, K., Charara, H., Abbas, M., 2002. Intelligent Detection-control System for Rural Signalized Intersections. Texas Transportation Institute, Texas A&M University System.

Breiman, L., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, California.
Breiman, L., 1999. Random forests—random features.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Caird, J.K., Chisholm, S., Edwards, C.J., Creaser, J.I., 2007. The effect of yellow light onset time on older and younger drivers' perception response time (PRT) and intersection behavior. Transp. Res. Part F: Traffic Psychol. Behav. 10, 383–396.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2, 27.
Chang, M.-S., Messer, C.J., Santiago, A.J., 1985. Timing traffic signal change intervals based on driver behavior.
Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A., et al., 2008. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat. Rev. Cancer 8, 37–49.
Coakley, Richard C., Stollof, E.R., 2009. Intersection Safety Needs Identification Report.
El-Shawarby, I., Rakha, H.A., Inman, V.W., Davis, G.W., 2007. Age and gender impact on driver behavior at the onset of a yellow phase on high-speed signalized intersection approaches. Transportation Research Board 86th Annual Meeting .
El-Shawarby, I., Abdel-Salam, A.-S.G., Li, H., Rakha, H., 2015. Driver Behavior at the Onset of Yellow Indication for Rainy/Wet Roadway Surface Conditions.
Elhenawy, M., Rakha, H., El-Shawarby, I., 2014. Enhanced modeling of driver stop-or-run actions at a yellow indication. Transp. Res. Rec.: J. Transp. Res. Board 2423, 24–34.
Evgeniou, T., Pontil, M., Elisseeff, A., 2004. Leave one out error, stability, and generalization of voting combinations of classifiers. Mach. Learn. 55, 71–97.
Freund, Y., Schapire, R., 1999. A short introduction to boosting. Jpn. Soc. Artif. Intell. 14, 771–780.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28, 337–407.
Gates, T.J., Noyce, D.A., Laracuente, L., Nordheim, E.V., 2007. Analysis of driver behavior in dilemma zones at signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 2030, 29–39.
Gazis, D., Herman, R., Maradudin, A., 1960. The problem of the amber signal light in traffic flow. Oper. Res. 8, 112–132.
Ghanipoor Machiani, S., Abbas, M., 2014a. Dynamic driver's perception of dilemma zone: experimental design and analysis of driver's learning in a simulator study. The 93rd Annual Meeting of the Transportation Research Board, Washington, D. C.
Ghanipoor Machiani, S., Abbas, M., 2014b. Predicting drivers decision in dilemma zone in a driving simulator environment using canonical discriminant analysis. The 93rd Annual Meeting of the Transportation Research Board, Washington, D. C.
Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification.
Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 13, 411–430.
Jahangiri, A., Rakha, H., 2014. Developing a support vector machine (SVM) classifier for transportation mode identification by using mobile phone sensor data. Transportation Research Board 93rd Annual Meeting .
Jahangiri, A., Rakha, H.A., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans. Intell. Transp. Syst. PP, 1–12.
Jahangiri, A., Rakha, H., Dingus, T.A., 2015. Predicting Red-light Running Violations at Signalized Intersections Using Machine Learning Techniques.
Jones, W.D., 2001. Keeping cars from crashing. IEEE Spectr. 38, 40–45.
Li, H., Rakha, H., El-Shawarby, I., 2012. Designing yellow intervals for rainy and wet roadway conditions. Int. J. Transp. Sci. Technol. 1, 171–190.
Liu, Y., Chang, G.-L., Tao, R., Hicks, T., Tabacek, E., 2007. Empirical observations of dynamic dilemma zones at signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 2035, 122–133.
Liu, Y., Chang, G., Yu, J., 2012. Empirical study of driver responses during the yellow signal phase at six Maryland intersections. J. Transp. Eng. 138, 31–42.
Liu, Qingchao, Lu, Jian, Chen, S., 2015. Traffic incident detection using random forest. Presented at the Transportation Research Board 92nd Annual Meeting, Washington, D.C.
Machiani, S.G., Abbas, M., 2014. Predicting drivers decision in dilemma zone in a driving simulator environment using canonical discriminant analysis. Transportation Research Board 93rd Annual Meeting .
Pant, P.D., Cheng, Y., Rajagopal, A., Kashayi, N., 2005. Field Testing and Implementation of Dilemma Zone Protection and Signal Coordination at Closely-spaced High-speed Intersections. University of Cincinnati.
Pham, M.-H., Bhaskar, A., Chung, E., Dumont, A.-G., 2010. Random forest models for identifying motorway rear-end crash risks using disaggregate data. Presented at the 13th International IEEE Annual Conference on Intelligent Transportation Systems, Madeira Island, Portugal.
Rakha, H., Lucic, I., Demarchi, S.H., Setti, J.R., Van Aerde, M., 2001. Vehicle dynamics model for predicting maximum truck acceleration levels. J. Transp. Eng. 127, 418–425.
Rakha, H., El-Shawarby, I., Setti, J.R., 2007. Characterizing driver behavior on signalized intersection approaches at the onset of a yellow-phase trigger. IEEE Trans. Intell. Transp. Syst. 8, 630–640.
Tibshirani, R., Hastie, T., Friedman, J., 2009. The Elements of Statistical Learning Data Mining, Inference, and Prediction.
Sheffi, Y., Mahmassani, H., 1981. A model of driver behavior at high speed signalized intersections. Transp. Sci. 15, 50–61.
Shinar, D., Compton, R., 2004. Aggressive driving: an observational study of driver, vehicle, and situational variables. Accid. Anal. Prev. 36, 429–437.
Smith, L., 2002. A Tutorial on Principal Components Analysis.
Tango, F., Botta, M., 2013. Real-time detection system of driver distraction using machine learning. IEEE Trans. Intell. Transp. Syst. 14, 894–905.
U.S. Department of Transportation and Federal Highway Administration, 2014. Safety at Signalized Intersections. U.S. Department of Transportation and Federal Highway Administration. http://safety.fhwa.dot.gov/Intersection/signalized/presentations/sign_int_pps051508/long/index.cfm.
Wei, H., Li, Z., Yi, P., Duemmel, K.R., 2011. Quantifying dynamic factors contributing to dilemma zone at high-speed signalized intersections. Transp. Res. Rec.: J. Transp. Res. Board 2259, 202–212.
Wei, H., 2008. Characterize dynamic dilemma zone and minimize its effect at signalized intersections.
Yulan, L., Reyes, M.L., Lee, J.D., 2007. Real-time detection of driver cognitive distraction using support vector machines. IEEE Trans. Intell. Transp. Syst. 8, 340–350.
Zegeer, C.V., 1978. Green-extension systems at high-speed intersections.