Traffic & Transportation

Modelling Driver Behaviour at Urban Signalised Intersections Using Logistic Regression and Machine Learning

Ahmad H. ALOMARI[1], Bara' W. AL-MISTAREHI[2], Areen A. AL-JAMMAL[3], Taqwa I. ALHADIDI[4], Motasem S. OBEIDAT[5]

Original Scientific Paper
Submitted: 25 Apr. 2023
Accepted: 9 Oct 2023

[1] Corresponding author, alomarish@yu.edu.jo, Department of Civil Engineering, Yarmouk University
[2] bwmistarehi@just.edu.jo, Department of Civil Engineering, Jordan University of Science & Technology
[3] aaaljammal@eng.just.edu.jo, Department of Civil Engineering, Jordan University of Science & Technology
[4] t.alhadidi@ammanu.edu.jo, Department of Civil Engineering, Al-Ahliyya Amman University
[5] 2019751019@alumni.yu.edu.jo, Department of Computer Science, Yarmouk University

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT

This study investigated several factors that may influence driver actions throughout the yellow interval at urban signalised intersections. The selected samples include 2,168 observations. Almost 33% of drivers stopped ahead of the stop line, 60% passed the intersection through the yellow interval, and 7% passed after the yellow interval was complete (red light running, RLR violations). Binary logistic regression models showed that the chance of passing went up as vehicle speed went up and down as the gap between the vehicle and the traffic light and green interval went up. The movement type and vehicle position influenced the passing probability, but the vehicle type did not. Moreover, multinomial logistic regression models showed that the legal passing probability declined with the growth in the green time and vehicle distance to the traffic signal. It also increased with the growth in the speed of approaching vehicles. Also, movement type directly affected the chance of legally passing, but vehicle position and type did not. Furthermore, the driver's performance during the yellow phase was studied using the k-nearest neighbours algorithm (KNN), support vector machines (SVM), random forest (RF) and AdaBoost machine learning techniques. The driver's action run prediction was the most accurate, and the run-on-red camera was the least accurate.

KEYWORDS

traffic signal; traffic safety; logistic regression; machine learning; yellow phase; red light running.

1. INTRODUCTION

At signalised intersections, the yellow phase plays a significant transitional role between the green and red intervals. When a signal light changes from green to yellow, drivers must decide whether to cross the intersection safely or stop before the stop line. Also, drivers at the onset of the yellow phase need to interact with other drivers in front and back to prevent unsafe decisions [1]. Wrong driver decisions during the yellow interval may lead to right-angle crashes, left-turn crashes, rear-end crashes, or red-light running (RLR) violations. RLR and inconsistent stopping behaviour are considered risky causes of traffic crashes at signalised intersections [1]. Several RLR violations happened due to drivers' presence in the dilemma zones during the yellow period. At the onset of the yellow phase, dilemma zones occur upstream of the intersection approach [2]. A dilemma zone is formed when the driver approaches the intersection at a speed greater than the speed limit; conversely, an "option zone" is formed when the driver traverses slower than the speed limit [3]. Li and Wei [4] showed that dynamic dilemma zone models could predict the dilemma zone more accurately than the traditional dilemma zone and the type II dilemma zone models, which is the area where more than 10% but less than 90% of drivers would choose to stop at the start of the yellow interval. Driver behaviour during the yellow interval can be categorised as aggressive, normal or conservative based on stop/go decisions and distance to the stop line at the beginning of the yellow interval [5].

Several studies developed linear, non-linear and logistic regression models to represent and predict driver behaviour in dilemma zones based on several influencing factors at urban, suburban and rural signalised intersections. Results found that RLR enforcement cameras increase driver-stopping decisions and decrease RLR violations [6–9]. Rakha et al. [10] found that the dilemma zone for the older drivers was larger and closer to the intersection than for the younger drivers. Also, results showed that the most significant factors affecting driver behaviour through the yellow phase, stop/go decisions and RLR violations were delay [8], approaching speed [11–15], acceleration and deceleration of vehicles [9, 13, 16], traffic conditions [8, 9], weather conditions [9], perception-reaction times [16], time to stop line at the beginning of yellow interval [10, 14, 17, 18], vehicle position and distance to stop line [9, 11–15, 17, 19], vehicle type [9, 11, 12, 14, 15, 17], intersection type [11, 12], roadway grade [18], duration of yellow interval [11, 12] presence of interruptions such as vehicles, bicycles or pedestrians on the side street [17], driver age and gender [14, 20].

Driving simulators were also used in several studies to investigate driver behaviour in dilemma zones. According to Swake et al. [21], driver decisions, deceleration rates and brake response duration all influence driver behaviour. Also, they state that driving simulators are a good way to predict how drivers will act in certain situations. Choudhary and Velaga [22] revealed that phones and music players' distractions decrease the probability of yellow signal crossing, where the crossing possibility was positively associated with driving speed and negatively associated with time to stop line, type of manoeuvre and the presence of the distractions. Bryant et al. [23] concluded that the clearance interval at signalised intersections should consider the truck's characteristics and how the driver acts. With the right design, trucks can get through the intersection before the green light changes to let other cars go. Hussain et al. [1] suggested that RLR violations were significantly decreased by installing red-LED earth lights combined with the regular traffic signal and RLR camera warning support. Also, it was seen that when the green signal was set to flash, people stopped in different ways. Banerjee et al. [24] investigated how red-light violation warning (RLVW) systems affect the way drivers act. Results showed that the tested system slowed down approaching vehicles by a lot, giving drivers more time to come to a safe stop at the red-light intersection.

Many countermeasures were tested to reduce RLR violations in the field and a simulated environment. Najmi et al. [25] showed that dilemma zones at signalised roundabouts were shorter and closer to the stop-line than regular signalised intersections because drivers move more conservatively at the roundabouts with a safer stopping ability. Also, Wang et al. [16] concluded that implementing five seconds of yellow interval proposed the best results for reducing risky behaviour at high-speed intersections. Moreover, Sun et al. [26] recommended an exclusive heavy vehicle lane as a future safety countermeasure to reduce vehicle conflict and intersection delay. Furthermore, Zhang et al. [9] focused on reducing RLR by installing red-light cameras and countdown timers to increase stopping behaviour at the onset of the yellow interval and reduce risky driving behaviour. Finally, Ni et al. [27] stated that a mandatory stop during a solid yellow light could control aggressive drivers efficiently, which reduces the approaching speeds significantly and enhances the acceptance of more significant headways between vehicles. However, it increased the rear-end collision probability, raising the demand for more drivers' educational programs for traffic safety and conservative behaviours.

Other useful modelling techniques were used to represent and predict motorist performance in dilemma zones at traffic signals, such as artificial neural networks [28], fuzzy logic and decision tree modelling [15, 28, 29, 30], hidden Markov modelling [31] and other machine learning (ML) algorithms [32–40]. Results showed that these techniques could produce a high accuracy level similar to the linear, non-linear and logistic regression models. Elhenawy et al. [32–34] specified that driver aggressiveness at signalised intersections significantly affected the driver's stop/go decisions and positively increased the models' accuracy. They also verified that all modelling approaches generated similar prediction accuracy. Khanfar et al. [35] studied the driving behaviour at signalised intersections using unsupervised ML and a driving simulator dataset. The approach confirmed that driving behaviour reflects drivers' habits and character rather than the signal condition; however, it still represents the nature of the intersection, which requires drivers to be more careful. Tawfeek [37] modelled the speed of unassisted drivers using ML as the yellow light turned on to improve connected and autonomous vehicle implementations at signalised intersections and enhance driver comfort. The findings suggest that the speed at the yellow light can be estimated using observations that account for the perceptual ability of drivers. Karri et al. [38, 39] examined driving behaviour (safe and unsafe stopping)at signalised intersections using ML

based on the driving features. Findings showed that the suggested method could assist in developing a system to alert drivers reaching signalised intersections, thereby reducing rear-end collisions and crashes.

The primary goal of this study is to investigate the main factors that may influence driver actions throughout the yellow interval of the traffic signal at urban intersections in Jordan. This paper classified driver actions during the yellow interval at traffic signals into: "stopping before the stop line," "crossing the intersection before the end of the yellow phase" and "crossing the intersection after the end of the yellow phase." This paper developed logistic regression and ML models to investigate the relationships between motorist actions through the yellow interval and influencing factors. The rest of this article is organised as follows: Section 2 introduces the methodology used for modelling driver actions using logistic regression and ML techniques, including binary logistic regression, multinomial logistic regression, k-nearest neighbours algorithm (KNN), support vector machines (SVM), random forest (RF) and AdaBoost. It also defines the study area and the data collected in this work. Section 3 shows the modelling results by analysing the methods employed. Finally, Section 4 introduces the essential conclusions of this paper.

2. METHODOLOGY

Eight intersections controlled by traffic signals with channelised right-turn movements (*Figure 1*) were chosen in Irbid City, Jordan [41]. Four of them were fully actuated with RLR cameras, and RLR cameras did not control the rest. Intersection characteristics were also gathered, including the speed limit (60 km/h), lanes on the studied approach, lanes crossed, approach width (meters), width of lane (meters), flow (vehicles/hour/lane), number of approaches on the intersection and pavement marking conditions. An approach operating vehicle speed was measured using a laser radar gun. Three-legged and four-legged intersections were considered. Data were gathered during peak hours in fine weather and dry road conditions. *Table 1* presents the summary of data collection.

Table 1 – Intersection characteristic and traffic signal timing data

Intersection	Studied approach	No. of lanes crossed	No. of lanes	Width of lane [m]	Width of intersection [m]	Traffic flow [veh/h]	No. of phases	No. of legs	Operating speed [km/h]	Grade
T1	NB	8	3	3	23.2	476	4	4	34	Level
T2	SB	9	2	3.5	37	382	4	4	39	Upgrade
T3	WB	9	4	3	31	394	4	4	47	Level
T4	NB	8	3	3.57	39	502	4	4	43	Level
T5	EB	5	4	2.925	32	342	3	3	37	Level
T6	EB	5	4	3.12	33.5	359	3	3	41	Level
T7	WB	6	3	2.933	27.5	221	4	4	29	Level
T8	EB	10	3	3	43.7	272	4	4	35	Level
Intersection	Cycle range [s]	Red period [s]	Yellow period [s]	Green period [s]	Green split	All red period [s]	RLR	Green flash	Pavement markings	Pedestrians
T1	131	95	4	30	0.229	2	Yes	Yes	Yes	Low
T2	139	102	3	32	0.23	2	Yes	Yes	Yes	Medium
T3	109	75	2	30	0.275	2	Yes	Yes	Yes	Low
T4	146	104	3	37	0.253	2	Yes	Yes	Yes	Low
T5	64	44	5	15	0.234	0	No	Yes	No	Low
T6	82	56	3	21	0.256	2	No	No	No	Low
T7	112	82	2	26	0.232	2	No	Yes	Yes	Heavy
T8	126	95	3	26	0.206	2	No	Yes	No	Low

** NB: Northbound, WB: Westbound, SB: Southbound, EB: Eastbound.*

T1 (32°32’36.6”N 35°52’50.4”E)

T2 (32°31’54.0”N 35°51’08.9”E)

T3 (32°32’05.3”N 35°51’36.3”E)

T4 (32°32’33.3”N 35°51’31.8”E)

T5 (32°33’00.2”N 35°51’44.2”E)

T6 (32°32’41.3”N 35°52’29.2”E)

T7 (32°33’24.3”N 35°50’58.2”E)

T8 (32°33’26.7”N 35°51’47.8”E)

Figure 1 – Bird’s-eye view of the studied intersections

Binary and multinomial logistic regression models were developed to predict motorist actions throughout the yellow interval of the traffic signals at urban intersections, whether or not they have RLR cameras. Logistic regression can be represented as the following formula (*Equation 1*) [42]:

$$Logit(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (1)$$

where P is the probability of a decision to pass (Y=1), β_0 is the model constant, β_0 is the coefficient of variable and X_i represents the predictor variable.

In this paper, the dependent variable was categorical. So, binary and multinomial logistic regression models are the best choices for predicting the probability of a categorical dependent variable [42]. Also, they were selected to overcome the problem of violating the linearity assumption. The proposed models involved two types of variables: categorical and continuous. *Table 2* describes all the variables involved in the proposed models.

Previous research has widely used several ML classification algorithms. To predict driver behaviour, the commonly used KNN, SVMs, RF and AdaBoost are used in this paper. The same training dataset was employed to train the different ML techniques, and the models' performance was reported using the same testing dataset.

The KNN algorithm is a straightforward non-parametric modelling technique [43]. It is based on the probability that similar data points belong to the same cluster. KNN begins by locating the K nearest neighbourhoods of the training dataset and then predicts the major class within the K nearest neighbours. Due to its simplicity and ability to predict in less time, it has been chosen as one of the best data mining algorithms [44].

Table 2 – Description of variables in binary and multinomial regression models

Variable	Variable Type	Unit
Speed	Continuous	km/h
Vehicle distance to stop line	Continuous	Meter
Volume in selected approach	Continuous	veh/h/lane
Yellow period	Continuous	Second
Green period	Continuous	Second
Red period	Continuous	Second
Cycle period	Continuous	Second
Green split	Continuous	Unitless
Vehicle type	Categorical	0 pc (regular cars),1 taxi, 2 pickup, 3 truck, 4 van, 5 bus
Intersection type	Categorical	0 3 legs, 1 4 legs
Type of movement	Categorical	0 left, 1 U-turn, 2 through
Vehicle position	Categorical	0 platoon, 1 not platoon
Lane width	Continuous	Meter
Intersection width	Continuous	Meter
No. of lanes in	Categorical	2 two-lanes, 3 three-lanes, 4 four-lanes
No. of lanes crossed	Categorical	0 5-lanes, 1 6-lanes, 2 8-lanes, 3 9-lanes, 4 10-lanes
Grade	Categorical	0 level, 1 upgrade, 2 downgrade
RLR cameras	Categorical	0 yes, 1 no
Pavement marking	Categorical	0 yes, 1 no
Green flash	Categorical	0 yes, 1 no
Pedestrians	Categorical	0 low, 1 medium, 2 heavy

The KNN accuracy depends on choosing the best cluster size; the optimum K was selected based on prediction accuracy. Afterwards, the response (i.e. driver action in our problem) is classified by considering the majority vote of the K closest points within the class as shown in *Equation 2*; where *R* is the number of assigned classes based on checking the model accuracy for each value, y_j^{test} is the test observation which is assigned to class R based on the majority of class R voting after training the model using x_j^{train} as input variables and y_j^{train} as response variable [45].

$$y_j^{test} = \frac{1}{R} \sum_{X_j^{train} \in R_k} y_j^{train} \tag{2}$$

The SVM algorithm is a supervised learning method that sorts data into groups based on how different the groups are. *Equation 3* shows that the algorithm looks for the hyperplane (also called a "splitter") that is the closest to the training data. The SVM looks for the weight (w) with the most significant margin near the hyperplane and meets the two constraints (see *Equations 4 and 5* [46]).

$$\min_{w,b,\xi} \left(\frac{1}{2} w^T w + c \sum_{n=}^{N} \xi_n \right) \tag{3}$$

subjected to:

$$y_n \left(w^T \phi\left(X_n \right) + b \right) \geq 1 - \xi_n \quad , n = 1, \ldots\ldots, N \tag{4}$$

$$\xi_n \geq 0 \, , n = 1, \ldots N \tag{5}$$

w – set of parameters used to define class boundaries
c – penalty parameter
ξ_n – parameter to express the margin error
b – intercept is linked with the hyperplane functions to change data from X space
$\phi(X_n)$ – transform data from X space to Z space
y_n – target value

The objective function is simplified by adding the two terms in *Equation 2*. Primarily, the first term aims to clarify the difference between classes. Reducing its length is identical to enlarging the gap between classes. The other term aims to reduce the penalty (regularisation) parameter times the error term. The penalty term is intended to address overfitting, whereas the term c is intended to optimise the performance of the model. Therefore, n represents the index of the data observation, w denotes the decision border between classes, c denotes the regularisation (or penalty) parameter and j_n represents the margin violation error parameter. K is the number of observations in the X space that the $\phi(X_n)$ function moved to another space. The transformation is done to make a Z space that can be used to make class boundaries easier to define. On the other hand, certain functions (i.e. kernels) can be used directly to create transformations more easily, as demonstrated in this paper. Meanwhile, *Equation 2* can be solved using kernels or $\phi(X_n)$ to transform data to the Z plane. Before model construction, the kernel type should be determined (i.e. linear). One kernel could work better than another. Some realistic recommendations propose using different kernels at various data sizes and problems [46].

Random forest (RF) is a successful ensemble prediction technique. Breiman [47] used the strong law of large numbers (SLLN) to demonstrate that there is no overfitting of RFs as more trees are established. The fundamental concept underlying ensemble approaches is that creating many simple models will improve overall performance. An RF is a collection of unpruned decision trees with random feature selection at each split. Classification and regression tree (CART), a well-known ML technique, is a frequently used decision tree in RFs [48]. In ensemble terms, RFs start with the CART, which refers to the weak model.

CART partitions the feature space into two regions to optimise its objective function locally (children). This procedure is repeated for every child until the termination criteria are met. Cases from each region have (nearly) identical outcomes. Using the assumption that the training dataset consists of H cases, P predictors and M trees to generate for each of the M iterations, the RF classification algorithm is as follows:

- Build a bootstrap trial from the first dataset by randomly selecting H cases and replacing them. The subset must comprise around 66 percent of the initial training set; the left cases should be duplicated.
- For certain numbers, $\sqrt{p}$ predictor variables are chosen randomly from all predictor variables at each node.
- From the $\sqrt{p}$ predictor variables, the best predictor variable is employed to generate a binary split on that node.

– Avoid value-complexity pruning and keep the tree in its current state, along with other constructed trees from prior iterations. During the testing phase, the recently delivered case is moved down each tree. By supplying a class label, each tree votes for one class. RF determines the class with the most votes. This method will be evaluated as part of this research effort because it can improve driver stop/run behaviour modelling.

The adaptive boosting (AdaBoost) algorithm is an incremental contribution-based ML algorithm [48]. AdaBoost was developed in response to whether it was possible to combine a cluster of "weak" learner algorithms with low accuracy to generate a learning algorithm with a high one. Prior to the running of AdaBoost, the conventional ML technique consisted of selecting the highest-discriminating class of features. In other terms, algorithms should be classified. AdaBoost employs a collection of weak classifiers, each of which is trained on the same training dataset but has a different weight distribution. Every learner concentrates on the instances where the previous learner failed. AdaBoost's output is the weighted average of all weak learners' outputs. It has a minor misclassification than the sum of weak learners and a generalisation error limit [48, 49].

In a classification problem, the output could be a true positive prediction (TP), a true negative prediction (TN), a false positive prediction (FP) or a false negative prediction (FN). These distinct outcomes were utilised to compute the various evaluation metrics. Precision, recall, F1-score and support are the evaluative metrics (*Equations 6–8*). Precision and recall are two methods for evaluating the performance of a classifier in binary and multiclass classification problems. Precision is determined by dividing the number of accurate positives by the accurate and false positives summation. Recall is the proportion of correctly classified instances (true positives) to the total instances that should have been classified as positive (true positives plus false negatives). The F1-score is utilised to evaluate the accuracy of a model on a dataset. It assesses classification systems that categorise instances as "positive" or "negative". The F-score combines the model's precision and recall [40]. Support refers to the number of actual class occurrences in the dataset. It is the count of true instances for each class. These indices are calculated as follows:

$$Precision = \frac{TP}{TP + NP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

These metrics are essential for evaluating the performance of classification models, and they help in understanding how well a model is doing in correctly identifying positive and negative cases.

3. RESULTS AND DISCUSSION

The data extraction process yielded a total of 2,168 samples, including stop, pass and RLR violations. Only 721 (33%) drivers stopped ahead of the stop line, 1,296 (60%) passed the intersection through the yellow interval and 151 (7%) passed after the yellow interval was complete (RLR violations). According to the findings, drivers had more potential to stop ahead of the stop line during the yellow interval on intersections with RLR cameras, green light flashing, heavy pedestrian activity, pavement markings and intersections with four approaches. In addition, platoon-positioned vehicles had more pass actions (69.8%) than non-platoon-positioned vehicles (46.6%). Moreover, the van carried the most significant percentage of pass action among all vehicle types (68.1%), while the taxi was the lowest (54.5%). In contrast, the pass rates for trucks and pickups were 64% and 65.9%, respectively. The percentages for straight, left and U-turn manoeuvres were 58.4%, 61.3% and 57.8%, respectively. However, straight movement had the highest RLR violation rate (8.6%). *Table 3* displays the frequencies and percentages of driver actions for the studied intersections.

3.1 Binary logistic regression models

In the Statistical Package for the Social Sciences (SPSS), sequential logistic regression models were made to predict how drivers would do during the yellow phase. Two models were considered. Model one top-level logit, including a stop or go action, and model two bottom-level logit models, including only legal passes through the yellow phase or RLR violations. *Figure 2* shows a two-step decision process for binary logistic regression.

Table 3 – Descriptive statistics for the major categorical variables

Influencing parameters		Action			Total
		Stop	Pass	RLR	
Presence of RLR cameras	Yes (33.3%)	406 (56.3%)	305 (42.2%)	11 (1.5%)	722
	No (66.7%)	315 (21.8%)	991 (68.5%)	140 (9.7%)	1446
Presence of flash green device	Yes (68.1%)	622 (42.1%)	772 (52.3%)	83 (5.6%)	1477
	No (31.9%)	99 (14.3%)	524 (75.8%)	68 (9.8%)	691
Presence of pedestrian	Low (85.82%)	576 (31.0%)	1156 (62.1%)	129 (6.9%)	1861
	Medium (6.5%)	87 (61.3%)	53 (37.3%)	2 (1.4%)	142
	High (7.6%)	58 (35.2%)	87 (52.7%)	20 (12.1%)	165
Presence of pavement marking	With (40.9%)	464 (52.3%)	392 (44.2%)	31 (3.5%)	887
	Without (59.1%)	257 (20.1%)	904 (70.6%)	120 (9.4%)	1281
Intersection type	3-leg (59.2%)	256 (20%)	894 (69.7%)	133 (10.4%)	1283
	4-leg (40.8%)	465 (52.5%)	402 (45.4%)	18 (2.0%)	885
Vehicle type	PC (70.8%)	513 (33.4%)	911 (59.3%)	111 (7.2%)	1535
	Taxi (11.4%)	102 (41.5%)	134 (54.5%)	10 (4.1%)	246
	Pickup (3.8%)	21 (30.9%)	54 (65.9%)	7 (5.1%)	82
	Van (5.2%)	26 (23.0%)	77 (68.1%)	10 (8.8%)	113
	Truck (6.3%)	42 (30.9%)	87 (64.0%)	7 (5.1%)	136
	Bus (2.5%)	17 (30.9%)	32 (58.2%)	6 (10.9%)	55
Turning movement type	Left (47.7%)	345 (33.4%)	633 (61.3%)	55 (5.3%)	1033
	U-Turn (4.2%)	32 (35.6%)	52 (57.8%)	6 (6.7%)	90
	Through (48.2%)	344 (33.0%)	610 (58.4%)	90 (8.6%)	1044
Vehicle position	Platoon (56.8%)	306 (24.9%)	859 (69.8%)	66 (5.4%)	1231
	Not platoon (43.2%)	415 (44.3%)	437 (46.6%)	85 (9.1%)	937
Grade	Level (93.5%)	634 (31.3%)	1243 (61.4%)	149 (7.4%)	2026
	Upgrade (6.5%)	87 (61.3%)	53 (37.3%)	2 (1.4%)	142
	Downgrade (0%)	0	0	0	0

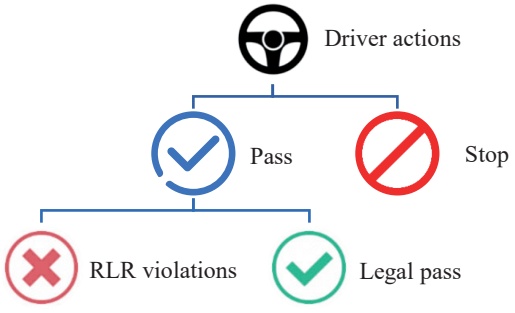

Figure 2 – Step decision process for binary logistic regression

Model-I (stop and pass action) looked at 2,168 observations, including people stopping before the stop line and passing through the intersection. Also, Model-II (legal pass and RLR violations), which looked at 1,450 observations, only looked at legal passes through the intersection during the same phase and RLR violations. Model-I and Model-II have different sample sizes because the driver takes different actions during the same phase in each model. *Table 4* presents the binary logistic regression analysis for Model-I and Model-II.

Table 4 – Estimated parameter of binary regression for Model-I and Model-II

Model	Changing	B	S.E	Wald	DF	Sig.	Exp (B)
Model-I	Operating speed (S)	0.078	0.024	10.364	1	0.001	1.081
	Vehicle distance (D)	-0.091	0.007	189.127	1	0.0	0.913
	Green interval (G.I)	-0.084	0.039	4.545	1	0.003	0.920
	Movement type			37.939	2	0.0	
	Movement type 1, M1, left (0)	-0.698	0.117	35.313	1	0.0	0.498
	Movement type 2, M2, U-turn (1)	-0.832	0.295	7.990	1	0.005	0.435
	Vehicle position 1, platoon, V.p1 (0)	0.694	0.116	35.731	1	0.0	2.002
	Presence of RLR cameras, P. of RLR (1)	-1.284	0.529	5.889	1	0.015	0.277
	Constant	1.175	1.421	0.683	1	0.409	3.237
Model-II	Operating speed (S)	0.200	0.058	11.852	1	0.001	1.222
	Vehicle distance (D)	-0.291	0.022	176.281	1	0.0	0.747
	Yellow interval (Y.I)	0.746	0.193	15.004	1	0.0	2.109
	Movement type			13.056	2	0.001	
	Movement type 1, M1, left (0)	-0.798	0.289	7.629	1	0.006	0.450
	Movement type 2, M2, U-turn (1)	-2.001	0.680	8.671	1	0.003	0.135

For Model-I, the negative sign of the variable's vehicle distance to the stop line and the green interval indicates that the probability of passing action increases with the raising of these variables. The positive sign of the variable operating vehicle speed suggests that the likelihood of passing action increases with the raising of this variable. Moreover, the passing probability is found to be safely affected by the presence or absence of RLR cameras, movement type and vehicle position. The reference movement type and vehicle position were taken through movement, not platoon position. The platoon vehicle was more likely to pass than not, given the platoon vehicle's position. Also, the left movement was more likely to pass than the U-turn movement. Finally, drivers at locations with no RLR cameras had a greater chance of passing than locations with RLR cameras.

The odds ratio of the operating vehicle speed means that for each unit raised in the variable operating vehicle speed, the odds of passing probability increase by 1.081 times. Also, the chance of passing decreases by 0.913 times for every unit where the distance between the vehicle and the stop line increases. For Model-II, the negative sign of the variable "vehicle distance" to the stop line at the beginning of the yellow interval indicates that the probability of legally passing an action decrease with the increase in this variable. The positive sign of the variables operating vehicle speed and yellow interval suggests that the likelihood of passing legally increased with the raising of these variables. Moreover, the passing likelihood was safely influenced by movement type. The reference movement type was taken as a through movement; the left movement was more likely to pass than the U-tern movement.

The odds ratio of the operating vehicle speed means that for each unit raised in the variable operating vehicle speed, the odds of passing probability increase by 1.222 times. Also, for each unit raised in the variable "vehicle distance", the odds of passing probability decrease by 0.747 times.

Table 5 shows the classification predicted for Models I and II. The overall prediction accuracy for Model-I and Model-II was 76.7% and 94.4%, respectively, indicating that the prediction results are close to reality. The Negelkerke R-squared for Model-I and Model-II was found to be 0.364 and 0.645.

3.2 Multinomial logistic regression models

MLR (multinomial logistic regression) models were made to predict how drivers would do during the yellow phase. In the proposed model, driver actions, including stopping before the stop line, passing through the intersection, and breaking RLR rules, were considered. *Figure 3* shows the step-decision process for multinomial logistic regression.

Table 5 – Classification of predicted for Model-I and Model-II

Model	Action	Percentage correct	
Model-I	Stop	51.6%	-2 log likelihood (2097.703) Cox and snell R^2 (0.262) Nagelkerke R^2 (0.364)
	Pass	88.8%	
	Overall percentage	76.7%	
Model-II	RLR violation	60.9%	-2 log likelihood (421.008) Cox and snell R^2 (0.315) Nagelkerke R^2 (0.645)
	Legal pass	98.3%	
	Overall percentage	94.4%	

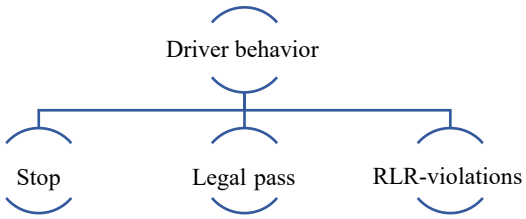

Figure 3 – Step decision process for multinomial logistic regression

The proposed model describes the driver actions as "stop" (Y=0), "pass through the yellow phase" (Y=1) and "RLR violations" (Y=2). Also, two types of variables were included in the proposed models: categorical and continuous variables. Multinomial logistic regression analysis for a stop to RLR violations and a legal pass to RLR violation models are presented in *Table 6.*

Table 6 – Estimated parameters of multinomial logistic regression for "Stop Action" and "Legal-Pass Action" models

Model	Variable	B	S.E	Wald	D.F	Sig.	Exp(β)	95% C.I for Exp(β)	
								Low	Up
Model-I	Vehicle distance (D)	-0.192	0.021	87.363	1	0.00	0.825	0.792	0.859
	Vehicle position, platoon, V.P1 (0)	-0.794	0.245	10.520	1	0.001	.452	0.280	0.730
Model-II	Approach speed (S)	0.405	0.136	8.860	1	0.003	1.499	1.148	1.957
	Vehicle distance (D)	-0.380	0.022	288.288	1	0.0	0.684	0.655	0.715
	Green interval (G.I)	-3.684	1.525	5.834	1	0.016	0.025	0.001	0.499
	Movement type1, M1, left (0)	-0.705	0.256	7.561	1	0.006	0.494	0.299	0.817
	Movement type 2, M2, U-turn (1)	-1.410	0.617	5.221	1	0.22	0.244	0.073	0.818

For the action model, the negative sign of the "vehicle distance" indicates that the probability of a stop action for RLR violations decreases with the increase in this variable. Moreover, for the legal-pass action model, the stopping possibility was discovered to be safely affected by vehicle position. Also, for a legal-pass action, the negative sign of the variable's "vehicle distance" and green interval indicates that the probability of a legally passing action decreases with the increase in these variables. The positive sign of the variable operating vehicle speed suggests that the likelihood of legally passing action increased with this variable's increase. In addition, movement type had an impact on the passing likelihood.

The odds ratio of the operating vehicle speed means that for each unit increase in the variable operating vehicle speed, the odds of a legal passing probability increase by 1.957 times. Also, for each unit raised in the

variable "vehicle distance", the odds of a legal passing probability decrease by 0.715 times. *Table 7* shows the classification of predicted stop-action and legal-pass action models.

Table 7 – Classification results, R-squared and models summary

Action	Percentage correct
Stop	60.2%
Legal-pass	88.6%
RLR violations	49%
Overall percentage	76.4%

Model	-2 log likelihood	Cox and Snell R-squared	Nagelkerke R-squared	Mc Fadden
1	1745.320	0.483	0.589	0.384

Model		Equation	
Binary logistic regression	Model-I stop- and pass-action	Logit (P) = Ln [Pi/(1-Pi)] = 1.175 + 0.078 S − 0.091 D − 0.084 G.I − 0.698 M1 − 0.832 M2 + 0.694V.P1 1.284 P.of RLR	76.7%
	Model-II legal-pass and RLR violations	Logit (P) =Ln [Pi/(1-Pi)] = 0.20 S – 0.291 D + 0.746 Y.I – 0.798 M1 – 2.001 M2	94.4%
Multinomial logistic regression	Stop action model	P (Stop/RLR) = – 0.192 D – 0.794 V.P1	76.4%
	Legal-pass action model	P (Legal pass / RLR) = 0.405 S – 0.380 D – 3.684 G.I – 0.705 M1 – 1.410 M2	

The total prediction accuracy was 76.4%, indicating that the prediction results were close to reality. *Table 7* also shows R-squared results for stop action and legal-pass action models. The Mc Fadden R-squared was found at 0.384, which indicates that it is effective enough to forecast driver performance through the yellow phase.

3.3 Machine learning (ML) models

This section discusses the outcomes of the Python-based ML methods applied in this paper. The first step in modelling data was feature engineering. It began with a data type check, followed by a report of the original data correlation matrix and a review of the problem's most relevant variables. *Figure 4* shows the correlation matrix of the original dataset for the different collected variables.

All the various variables have different correlation values. Red interval, no. of lanes crossed, cycle length, green interval, pavement markings, intersection type and presence of RLR cameras are highly inversely correlated variables. The number of lanes is related to driver behaviour. Nonetheless, the most relevant variables were chosen using the P-value and F-score. The selected characteristics were determined based on P-values and F-scores exceeding 0.05 and 5, respectively.

Table 8 presents the original variables along with their P-values and F-scores. There were 2,168 total instances in this dataset, out of which 1,734 (80%) random data were used for training, and the remaining 434 (20%) were used for testing and validating the model. According to *Table 8*, the chosen variables are the intersection type, green interval, cycle length, number of lanes crossed, number of lanes in, location, volume in selected approaches, lane width, grade, vehicle position, and the existence of RLR cameras, pavement marking, green flash, yellow interval and pedestrians. These variables were selected as the X matrix, while driver actions were chosen as their output. This study used four methods: KNN, SVM, RF and AdaBoost.

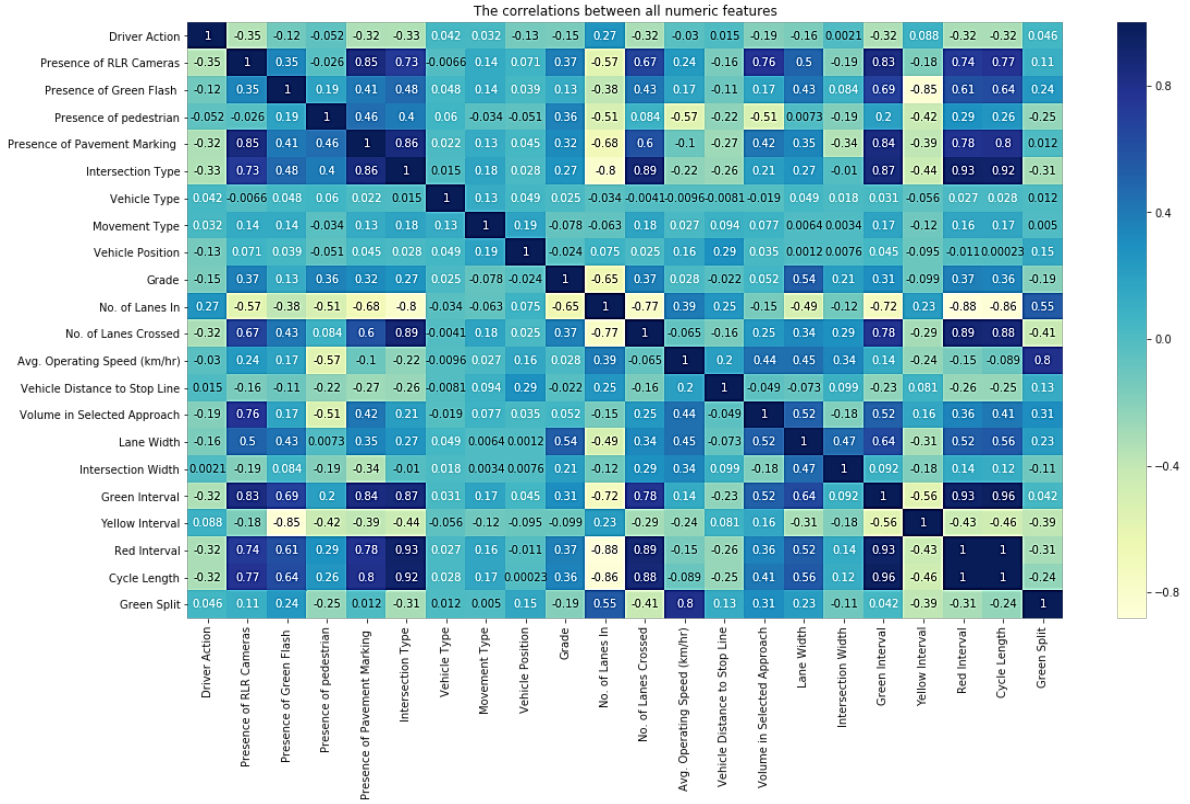

Figure 4 – Correlation matrix

Table 8 – Variables with their P-value and F-score

Input variable	P-value	F-score	Input variable	P-value	F-score
Presence of RLR cameras	1.00E-62	298.3186	Grade	5.66E-13	52.60281
Intersection type	1.06E-57	272.2201	Vehicle position	2.81E-10	40.17894
Presence of pavement marking	1.98E-54	255.3854	Presence of green flash	4.56E-08	30.10771
Green interval	8.67E-54	252.1052	Yellow interval	4.04E-05	16.92263
Cycle length	2.37E-53	249.8698	Presence of pedestrian	0.01559	5.85778
No. of lanes crossed	2.14E-52	244.9839	Green split	0.033383	4.531813
Red interval	9.26E-52	241.7396	Vehicle type	0.047264	3.940452
No. of lanes ln	4.35E-37	168.1459	Movement type	0.142152	2.156078
Location	2.45E-34	154.6135	Avg. operating speed [km/h]	0.1598	1.977477
Volume in selected approach	6.43E-20	85.13198	Vehicle distance to stop line	0.497443	0.460544
Lane width	1.54E-14	59.86941	Intersection width	0.922979	0.00935

Table 9 presents the overall classification report and confusion matrix of the test data. This table demonstrates four metrics: precision, recall, F1-score and support.

The KNN was used to evaluate which of three possible driver actions occurred. A 10-fold cross-validation was used to select the best model for each value of K, and the 10-fold with the highest average accuracy was selected. The optimal value of K was determined by comparing different values of K to overall classification accuracy. *Figure 5* illustrates the classification accuracy of KNN with varying K neighbours.

As illustrated in *Figure 5*, using pooled features in the proposed hierarchical framework yielded higher classification accuracy than just time-domain features. The optimal K was determined to be nine, with an accuracy of 67.5%.

Table 9 – Classification report and confusion matrix

	Precision	Recall	F1-score	Support	0	1	2
KNN							
0	0.59	0.51	0.55	138	70	68	0
1	0.71	0.84	0.77	264	41	223	0
2	0	0	0	32	7	25	0
SVM							
0	0.73	0.36	0.48	138	49	89	0
1	0.67	0.94	0.78	264	17	247	0
2	0	0	0	32	1	31	0
RF							
0	0.66	0.53	0.59	138	73	65	0
1	0.71	0.87	0.78	264	35	229	0
2	0	0	0	32	3	29	0
AdaBoost							
0	0.66	0.53	0.59	138	73	65	0
1	0.71	0.87	0.78	264	35	229	0
2	0	0	0	32	3	29	0

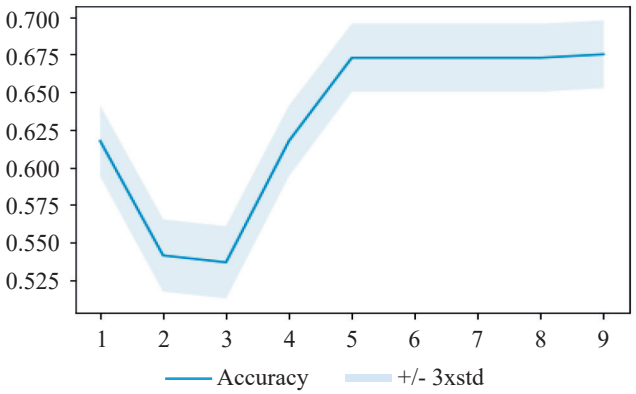

Figure 5 – Classification accuracy of KNN

Regarding SVM, a 10-fold cross-validation was used to get the optimal model for each value of K. The model with the highest average accuracy across all 10 folds was chosen. In addition, different kernels were used to train the model, and the best kernel was RBF with a gamma of 0.001, with an accuracy of 68.2%. In RF, many trees were trained, with the best accuracy of 69.58% coming from 400 trees. Finally, the AdaBoost method achieved an accuracy of 69.58%. The driver action run prediction had the highest accuracy for all methods (KNN, SVM, RF and Adaboost). In contrast, the run-on-red camera had the lowest precision because the number of its samples in the train and test data is low, and that is normal since not many people run on a red camera. The overall accuracy of all models was 68.7%.

The correct configuration of optimal ML models is crucial for practitioners implementing them. For example, practitioners can replicate these results with SVM by employing a 10-fold cross-validation approach and choosing the best kernel (RBF with a gamma of 0.001). This method suits scenarios where a balance between precision and recall is essential. On the other hand, with the knowledge that 400 trees yield the best accuracy, practitioners can set up their RF classifier accordingly. RF is known for its robustness and is suitable for handling large datasets. Also, the AdaBoost ensemble method achieved a competitive accuracy score. It can be applied when emphasis needs to be placed on the classification of harder-to-detect instances. Moreover,

the selection of K is vital for KNN and impacts the model's performance. The experiments found K=9 to be optimal for this specific dataset. Practitioners should consider a similar tuning process when applying KNN to their data.

Overall, the investigated models, binary logistic regression, multinomial logistic regression and the ML models, can be applied in practice in three major approaches.

Classification of driver actions. The primary application of the investigated models is in classifying driver actions based on data obtained from various sensors or sources. For instance, these models can be deployed in a real-time setting within a vehicle to predict and classify driver actions such as "run" actions. This prediction can be utilised for several practical purposes, including:

- driver assistance systems - these models can be integrated into driver assistance systems, providing real-time feedback to the driver. For example, if the model predicts a "run" action, it can trigger warnings or corrective actions, such as automatic braking or steering assistance, to prevent accidents.
- traffic safety - the models' ability to predict driver actions can contribute to improving overall traffic safety. Law enforcement or traffic management authorities can use this information to identify and address risky behaviour patterns, making roads safer for all users.
- insurance industry - insurance companies could leverage this model to assess driver behaviour and risk. It could be used to offer more accurate and personalised insurance premiums based on individual driving habits, ultimately promoting safer driving practices.
- fleet management - companies with large vehicle fleets can use these models to monitor driver behaviour and enhance the efficiency and safety of their operations. It can help identify drivers who consistently exhibit risky behaviour and may require additional training or supervision.

Optimal model selection. Detailed information about the optimal models and their configurations is crucial for practitioners who want to implement these techniques.

Data considerations. It is essential to mention that the models' performance may vary depending on the quantity and quality of training data. In cases like "run-on-red camera", where the sample size is limited, practitioners should be cautious about the model's reliability for such specific scenarios.

In summary, the investigated models' practical applications extend to driver assistance systems, traffic safety, insurance, fleet management and more. The detailed model configurations and optimal parameters provided in the paper can serve as a valuable starting point for practitioners looking to implement similar systems in their respective domains.

4. CONCLUSIONS

The objective of this study was to construct statistical models representing the relationships between various parameters and driver actions during the yellow interval at urban intersections controlled by traffic signals, whether or not they have red-light running (RLR) cameras. A video camera was utilised and positioned at an appropriate height ahead of the intersection to observe traffic signals, driver actions and parameters that may influence driver behaviour. Around 2,168 observations of motorist behaviour have been gathered from the data. Results showed that only 33% of drivers stopped ahead of the line, 60% passed the intersection in the yellow interval, and 7% passed after the yellow interval was complete (RLR violations). The following are the main findings:

- The likelihood of vehicles stopping before the line through the yellow interval with RLR cameras, the green flash tool, multiple pedestrians, pavement markings and intersections with four legs.
- At 68.1%, vans had the most significant proportion of pass actions among all vehicle types. In comparison, the taxis experienced the lowest pass rate, at 54.5%, although trucks and pickups had comparable pass rates, at 64% and 65.9%, respectively.
- The pass rates for through, left and U-turn manoeuvres were 58.4%, 61.3% and 57.8%, respectively. Nevertheless, the through direction had the most significant percentage of RLR violations. In addition, platoon-positioned vehicles had more pass actions (69.8%) than non-platoon-positioned vehicles (46.6%).
- The prediction accuracy of binary logistic regression Model-I was 76.7%, and Model-II's was 94.4%. Model-I (stop and pass action) indicated that the probability of a pass action increased with the rise in speed and dropped with the growth in the green interval and the length to the stop line. Also, the presence of RLR

cameras, movement type, and vehicle position significantly influenced the passing probability, but vehicle type did not.

- The binary logistic regression Model-II (legal pass and RLR violations) showed that the likelihood of legally passing rose with the increase in vehicle speed and yellow interval and dropped as the distance from the stop line increased. Also, movement type had a meaningful impact on the passing probability, but vehicle type and vehicle position did not.
- The prediction accuracy of the proposed multinomial logistic regression model was 76.4%, and McFadden's R-square was 0.384. The proposed models showed that the likelihood of stopping before the stop line declined with the increase in vehicle distance to the stop line. Also, vehicle position had an essential effect on the stopping probability, but movement and vehicle types did not. The likelihood of passing in the yellow interval decreased with the increase in vehicle distance to the stop line and green interval and increased with the increase in speed. Moreover, movement type had a meaningful impact on the passing probability, but vehicle position and vehicle type did not.
- This paper also used the commonly utilised KNN, SVM, RF and AdaBoost ML techniques to predict driver behaviour. The same training dataset was employed to train the different ML methods, and the models' performance was reported using the same testing dataset. As a result, the driver action run prediction had the highest accuracy, while the run-on-red camera had the lowest precision. The overall accuracy of all models was 68.7%.

Additional research is suggested to explore the influence of geometric design features, asphalt conditions, the characteristics of drivers, whether there are any passengers in the vehicle and the usage of mobile phones throughout the day.

REFERENCES

[1] Hussain Q, et al. Innovative countermeasures for red light running prevention at signalized intersections: A driving simulator study. *Accident Analysis & Prevention*. 2020;134:105349. DOI: 10.1016/j.aap.2019.105349.

[2] Elmitiny N, et al. Classification analysis of driver's stop/go decision and red-light running violation. *Accident Analysis & Prevention*. 2010;42(1):101-111. DOI: 10.1016/j.aap.2009.07.007.

[3] Papaioannou P. Driver behaviour, dilemma zone and safety effects at urban signalised intersections in Greece. *Accident Analysis & Prevention*. 2007;39(1):147-158. DOI: 10.1016/j.aap.2006.06.014.

[4] Li Z, Wei H. Modeling dynamics of dilemma zones by formulating dynamical contributing factors with video-observed trajectory data. *Procedia-Social and Behavioral Sciences*. 2013;80:880-900. DOI: 10.1016/j.sbspro.2013.05.048.

[5] Liu Y, et al. Empirical observations of dynamic dilemma zones at signalized intersections. *Transportation Research Record: Journal of the Transportation Research Board*. 2007;2035(1):122-133. DOI: 10.3141/2035-14.

[6] Gates T, Savolainen P, Maria HU. Impacts of automated red-light running enforcement cameras on driver behavior. *Transportation Research Board (TRB) Annual Meeting, Washington D.C., United States, 2014*. Paper No. 14-0943. 2014.

[7] Savolainen PT, Sharma A, Gates TJ. Driver decision-making in the dilemma zone – Examining the influences of clearance intervals, enforcement cameras and the provision of advance warning through a panel data random parameters probit model. *Accident Analysis & Prevention*. 2016;96:351-360. DOI: 10.1016/j.aap.2015.08.020.

[8] Awad W, et al. Drivers' behavior at signalized intersections. *Proceedings of the Seventh Traffic Safety Conference, 12-13 May, 2015, Amman, Jordan*. 2015.

[9] Zhang Y, Fu C, Hu L. Yellow light dilemma zone research: A review. *Journal of Traffic and Transportation Engineering.* 2014;1(5):338-352. DOI: 10.1016/S2095-7564(15)30280-4.

[10] Rakha H, Amer A, El-Shawarby I. Modeling driver behavior within a signalized intersection approach decision-dilemma zone. *Transportation Research Record: Journal of the Transportation Research Board*. 2008;(2069):16-25. DOI: 10.3141/2069-03.

[11] Pathivada BK, Perumal V. Modeling driver behavior in dilemma zone under mixed traffic conditions. *Transportation Research Procedia*. 2017;27:961-968. DOI: 10.1016/j.trpro.2017.12.120.

[12] Pathivada BK, Perumal V. Analyzing dilemma driver behavior at signalized intersection under mixed traffic conditions. *Transportation Research Part F: Traffic Psychology and Behaviour*. 2019;60:111-120. DOI: 10.1016/j.trf.2018.10.010.

[13] Li J, Jia X, Shao C. Predicting driver behavior during the yellow interval using video surveillance. *International JOUrnal of Environmental Research and Public Health*. 2016;13(12):1213. DOI: 10.3390/ijerph13121213.

[14] Alex S, Isaac KP, Varghese V. Modelling driver behavior at signalized intersection in Indian roads. *Transportation Research Board (TRB) Annual Meeting, Washington D.C., United States, 2013*. Paper No. 13-0257. 2013.

[15] Dong S, Zhou J. A comparative study on drivers' stop/go behavior at signalized intersections based on decision tree classification model. *Journal of Advanced Transportation*. 2020;2020(2):1-13. DOI: 10.1155/2020/1250827.

[16] Wang F, et al. Modeling risky driver behavior under the influence of flashing green signal with vehicle trajectory data. *Transportation Research Record*. 2016;2562(1):53-62. DOI: 10.3141/2562-07.

[17] Gates TJ, Noyce DA, Laracuente L, Nordheim EV. Analysis of driver behavior in dilemma zones at signalized intersections. *Transportation Research Record*. 2007;2030(1):29-39. DOI: 10.3141/2030-05.

[18] El-Shawarby I, Abdel-Salam ASG, Rakha H. Evaluation of driver perception–reaction time under rainy or wet roadway conditions at onset of yellow indication. *Transportation Research Record: Journal of the Transportation Research Board*. 2013;2384(1):18-24. DOI: 10.3141/2384-03.

[19] Campisi T, et al. Comparison of red-light running (RLR) and yellow light running (YLR) traffic violations in the cities of Enna and Thessaloniki. *Transportation Research Procedia*. 2020;45:947-954. DOI: 10.1016/j.trpro.2020.02.072.

[20] Ingale A, et al. Understanding driver behavior at intersection for mixed traffic conditions using questionnaire survey. *Transportation Research.* 2020:647-661. DOI: 10.1007/978-981-32-9042-6_51.

[21] Swake J, Jannat M, Islam M, Hurwitz D. Driver response to phase termination at signalized intersections: Are driving simulator results valid. *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: Driving Assessment 2013, 17-20 June 2013, Bolton Landing, New York, USA*. 2013. p. 278-284. DOI: 10.17077/drivingassessment.1501.

[22] Choudhary P, Velaga NR. Driver behaviour at the onset of yellow signal: A comparative study of distraction caused by use of a phone and a music player. *Transportation Research Part F: Traffic Psychology & Behaviour*. 2019;62:135-148. DOI: 10.1016/j.trf.2018.12.022.

[23] Bryant CW, Rakha HA, El-Shawarby I. Study of truck driver behavior for design of traffic signal yellow and clearance timings. *Transportation Research Record*. 2015;2488(1):62-70. DOI: 10.3141/2488-07.

[24] Banerjee S, Jeihani M, Khadem NK, Kabir MM. Influence of red-light violation warning systems on driver behavior – A driving simulator study. *Traffic Injury Prevention*. 2020;21(4):265-271. DOI: 10.1080/15389588.2020.1744135.

[25] Najmi A, Choupani AA, Aghayan I. Characterizing driver behavior in dilemma zones at signalized roundabouts. *Transportation research part F: Traffic Psychology and Behaviour*. 2019;63:204-215. DOI: 10.1016/j.trf.2019.04.007.

[26] Sun J, Wang Z, Yang J, Ouyang J. Comparison of dilemma zone and driver behavior of trucks and passenger cars at high-speed signalized intersections. *Transportation Research Board 94th Annual Meeting, 2015, Washington DC, United States*. 2015.

[27] Ni Y, Wang M, Li K, Xue N. Impacts of Chinese's new regulation of yellow signal on driving behavior and rear-end collision potential. *Transportation Research Board 93rd Annual Meeting, 2014, Washington DC, United States*. 2014.

[28] Biswas S, Ghosh I. Modeling of the drivers' decision-making behavior during yellow phase. *KSCE Journal of Civil Engineering*. 2018;22:4602-4614. DOI: 10.1007/s12205-018-0666-6.

[29] Hurwitz DS, et al. Fuzzy sets to describe driver behavior in the dilemma zone of high-speed signalized intersections. *Transportation Research Part F: Traffic Psych. & Behavior*. 2012;15(2):132-143. DOI: 10.1016/j.trf.2011.11.003.

[30] Yang Z, et al. Research on driver behavior in yellow interval at signalized intersections. *Mathematical Problems in Engineering*. 2014;2014:518782. DOI: 10.1155/2014/518782.

[31] Tang K, Zhu S, Xu Y, Wang F. Modeling drivers' dynamic decision-making behavior during the phase transition period: An analytical approach based on hidden markov model theory. *IEEE Transactions on Intelligent Transportation Systems*. 2016;17(1):206-214. DOI: 10.1109/TITS.2015.2462738.

[32] Elhenawy M, Rakha HA, El-Shawarby I. Enhanced modeling of driver stop-or-run actions at a yellow indication: Use of historical behavior and machine learning methods. *Transportation Research Record*. 2014;2423(1):24-34. DOI: 10.3141/2423-04.

[33] Elhenawy M, Jahangiri A, Rakha HA, El-Shawarby I. Classification of driver stop/run behavior at the onset of a yellow indication for different vehicles and roadway surface conditions using historical behavior. *Procedia Manufacturing*. 2015;3:858-865. DOI: 10.1016/j.promfg.2015.07.342.

[34] Elhenawy M, Jahangiri A, Rakha HA, El-Shawarby I. Modeling driver stop/run behavior at the onset of a yellow indication considering driver run tendency and roadway surface conditions. *Accident Analysis & Prevention*. 2015;83:90-100. DOI: 10.1016/j.aap.2015.06.016.

[35] Khanfar NO, et al. Driving behavior classification at signalized intersections using vehicle kinematics: Application of unsupervised machine learning. *International Journal of Injury Control and Safety Promotion*. 2022:30(3):1-11. DOI: 10.1080/17457300.2022.2103573.

[36] Jahangiri A, Rakha H, Dingus TA. Predicting red-light running violations at signalized intersections using machine learning techniques. *Transportation Research Board 94th Annual Meeting, 2015, Washington DC, United States*. 2015.

[37] Tawfeek MH. Perceptual-based driver behaviour modelling at the yellow onset of signalised intersections. *Journal of Transportation Safety & Security*. 2022;14(3):404-429. DOI: 10.1080/19439962.2020.1783414.

[38] Karri SL, et al. Identification and classification of driving behaviour at signalized intersections using support vector machine. *International Journal of Automation and Computing*. 2021;18:480-491. DOI: 10.1007/s11633-021-1295-y.

[39] Karri SL, et al. Classification and prediction of driving behaviour at a traffic intersection using SVM and KNN. *SN Computer Science*. 2021;2:1-11. DOI: 10.1007/s42979-021-00588-7.

[40] Alomari AH, Al-Mistarehi BW, Alnaasan TK, Obeidat MS. Utilizing different machine learning techniques to examine speeding violations. *Appl. Sci*. 2023;13:5113. DOI: 10.3390/app13085113.

[41] Al-Mistarehi BW, Alomari AH, Obaidat MT, Al-Jammal AA. Driver performance through the yellow phase using video cameras at urban signalized intersections. *Transport Problems*. 2021:16(1):51-64. DOI: 10.21307/tp-2021-005.

[42] Long K, Liu Y, Han LD. Impact of countdown timer on driving maneuvers after the yellow onset at signalized intersections: An empirical study in Changsha, China. *Safety Science*. 2013;54:8-16. DOI: 10.1016/j.ssci.2012.10.007.

[43] Zhang S, et al. Learning k for KNN classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2017;8(3):1-19. DOI: 10.1145/2990508.

[44] Wu X, et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*. 2008;14:1-37. DOI: 10.1007/s10115-007-0114-2.

[45] Friedman JH, Baskett F, Shustek LJ. An algorithm for finding nearest neighbors. *IEEE Transactions on Computers*. 1975;100(10):1000-6.

[46] Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*. 2002;13(2):415-25. DOI: 10.1109/72.991427.

[47] Breiman, L. Random forests. *Machine Learning*. 2001;45:5-32. DOI: 10.1023/A:1010933404324.

[48] Freund Y, Shapire R. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*. 1999;14(5):771-780.

[49] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*. 2000;28(2):337-407. DOI: 10.1214/aos/1016218223.

نمذجة سلوك السائق على التقاطعات الحضرية المحكومة بإشارات ضوئية باستخدام الانحدار اللوجستي والتعلم الآلي.

أحمد هاني العمري*، براء المستريحي، عرين الجمال، تقوى الحديدي، معتصم عبيدات.

الخلاصة

بحثت هذه الدراسة في العديد من العوامل التي قد تؤثر على تصرفات السائق خلال مرحلة اللون الأصفر عند التقاطعات الحضرية المحكومة بإشارات ضوئية. شملت العينات المختارة 2168 عينة وأظهرت النتائج توقف ما يقرب من 33% من السائقين قبل خط التوقف، في حين 60% عبروا التقاطع خلال مرحلة اللون الأصفر على الاشارة الضوئية، و7% عبروا بعد اكتمال اللون الأصفر (بداية الضوء الأحمر، بما يسمى قطع الاشارة الحمراء). أظهرت نماذج الانحدار اللوجستي الثنائي أن فرصة المرور ارتفعت مع ارتفاع سرعة السيارة وانخفاضها مع زيادة الفجوة بين السيارة وإشارة المرور والفاصل الأخضر. كذلك أظهرت النتائج أن نوع الحركة وموقع المركبة يؤثر على احتمالية نجاح العبور بسلام، لكن نوع المركبة لا يؤثر. علاوة على ذلك، أظهرت نماذج الانحدار اللوجستي متعدد الحدود أن احتمال المرور القانوني انخفض مع زيادة الوقت الأخضر ومسافة السيارة إلى الإشارة الضوئية. كما أنها تزداد مع نمو سرعة المركبات المقتربة. كما أن نوع الحركة أثر بشكل مباشر على فرصة المرور بشكل قانوني، لكن موقع السيارة ونوعها لم يؤثرا. علاوة على ذلك، تمت دراسة أداء السائق خلال مرحلة اللون الاصفر باستخدام خوارزمية (KNN)، و(SVM)، و(RF)، وتقنيات التعلم الآلي (AdaBoost). كان توقع حركة السائق هو الأكثر دقة، وكانت المواقع التي تعمل عليها الكاميرات التي تراقب قاطعي الاشارة الحمراء هي الأقل دقة.

الكلمات الدالة

الإشارة الضوئية، السلامة المرورية، الانحدار اللوجستي، التعلم الآلي، المرحلة الصفراء في الاشارة الضوئية، قطع الاشارة الحمراء.