Identification of safety-critical events on rural roads using a driving simulator

Akrivi Varela[1], **Christos Katrakazas**[1], **George Yannis**[1]

[1]National Technical University of Athens, Department of Transportation Planning and Engineering,
5 Heroon Polytechniou str., GR-15773, Athens, Greece

Abstract

The aim of this paper is to identify events based on driving characteristics on rural roads and determine the main factors that can describe the situation before and during an event. The data used were collected from a driving simulator experiment in rural roads. Data analysis was performed using advanced statistical models and more specifically binomial logistic regression, random forests, as well as factor analysis. The models were applied on two different variable sets: i) the whole set of independent variables and ii) the four most important independent variables after the application of a feature selection algorithm. Results showed that the random forest model performed much better than the binomial logistic regression in identifying event occurrence with very few false alarms, in both variants. Speed and time to collision along with total distance driven from the beginning of the driving session, turned out to better describe the case of driving one minute prior to an event. Driving during an event can be sufficiently described through speed, the deviation of the vehicle from the middle of the road as well as time headway. In terms of data describing the situation one minute prior of each event, these are better expressed through speed, time headway the deviation from the median and the total distance driven. Nevertheless, a larger sample of drivers and a naturalistic driving dataset could offer better results in the future.

Keywords: event identification, event duration, binary logistic regression, random forests, classification, factor analysis.

[1] Corresponding author. Tel.: +30-210-772-1265;
E-mail address: ckatrakazas@mail.ntua.gr

1. Introduction

As road traffic injuries remain a major public health problem, ensuring road safety and traffic management is significantly correlated with safety-critical incidents on the road [1], [2]. Driving events refer to road accidents and other safety-critical events, such as the unexpected entrance of a parked vehicle or a pedestrian, which often end up in potentially dangerous situations.

The road users, the condition of the road and the vehicle play an important role in driving events [3]. Nevertheless, to-date, the most critical factor leading to an event or accident is driver error, as driving characteristics and reaction times can be affected by many factors in real-time [4]. Researchers have shown that distraction such as cell phone use, talking to passengers, eating and drinking, negatively contribute to driver behavior [5]. Distraction can affect reaction times, reduce safety distances and the ability to maintain a proper lane position as well as deteriorate the driver's perception of what is happening around [6], [7]. The driver's characteristics, such as gender, age and driving experience, also play a role in the driver's distraction [8].

The way in which an unexpected event affects speed-related driving characteristics has been the subject of further investigation. It was found that after such an event the driving changes from normal levels to more careful actions. This change is due to the multiple actions that the driver is required to perform, such as making decisions and adapting driving behavior to the circumstances. Drivers talking on a cell phone pay much more attention after an unexpected event, because the use of one hand on the phone serves as a reminder to the driver of the potential security threat posed by the use of the phone. On the other hand while talking to the passenger, the driver has a lower level of compensatory behavior, however his attention is more often diverted from the road [9].

In order to investigate the effect of driving behavior before accidents and safety-critical events, recent studies have been focusing on the entire collision sequence, i.e. from a normal driving situation to a collision event or near collision. The results haverevealed that longitudinal acceleration, lateral acceleration and deflection rate can be reliable indicators for detecting deviations from normal driving. Also the values of the time to collision are affected by the type of vehicle, the speed of the vehicle, the longitudinal acceleration and the time in the accident [10].

From the aforementioned studies on driving behavior, distraction and safety-critical events, it can be concluded that there is insufficient research on driving characteristics before and during a safety-critical event, as well as on the identification of such phases (i.e. before or during the incident) using classifiers. This forms the motivation of the current paper.

2. Methodology

In order to identify safety-critical events, two classification models were used, namely, binomial logistic regression and random forests, while factor analysis was used to search for common factors among of the independent variables.

The binomial logistic regression model [11] is used to search for the relationship between a binary dependent variable and one or more independent variables described by the model equation as shown in equation (1).

The form of the equation is as follows:

$$y_i = \text{logit}\,(P_i) = \ln\frac{Pi}{1-Pi} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_n x_{ni} \qquad (1)$$

where:
n: number of independent variables
$\beta_1, \beta_2, \ldots, \beta_n$: regression coefficients of independent variables $x_1, x_2, \ldots, x_n$
β_0: fixed term coefficient
P_i: predicted probability

Random Forests belong to the group of ensemble classifiers and more specifically to the group of bagging algorithms. Bagging algorithms make use of only one learning algorithm and modify the training set by using the bagging algorithm to create new training sets [12]. Random Forests are an evolution of bagged trees and uses the bagging algorithm along with the random subspace method proposed by Ho [13]. Each tree is built using the impurity Gini index.

In order to predict the occurrence of an event, two different databases are created, a training Set with 80% of the data and a testing set with 20% of the elements of the original base. In order to assess the results and acceptance of models some statistical checks must be carried out. These include for the binomial regression model, the logical explanation of the coefficients of the resulting model as well as the statistical significance of those coefficients. For both classification models the correlation of independent variables and the main metrics of the confusion matrix are considered. From the confusion matrix the following metrics as shown in equations (2) – (7):

$$\text{accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{2}$$

$$\text{recall} = \frac{\mathrm{TP}}{TP+FN} \tag{3}$$

$$\text{specificity} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \tag{4}$$

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{5}$$

$$\text{F-measure} = \frac{2*precision*recall}{\mathrm{precision+recall}} \tag{6}$$

$$\text{False alarm rate} = \frac{FP}{\mathrm{TN+FP}} \tag{7}$$

where:

TP: true positive. The number of instances of the positive class (i.e. unexpected event), categorized as unexpected events by the classifier.
TN: true negative. The number of instances belonging to the negative class (i.e. safe driving), classified as safe driving.
FP: false positive. The number of instances of safe driving incorrectly identified as unexpected events.
FN: false negative. The number of unexpected events instances, falsely identified as safe driving.

Finally, to perform the method of factor analysis other than the correlation between the variables, the sample adequacy was checked with the Kaiser-Meyer-Olkin (KMO>0.6) test and its sphericity was checked with the Bartlett's Test of Sphericity (p<0.05).

3. Analysis and Results

3.1 Data collection and preprocessing

The database used to gather the necessary data was the one that emerged from an experiment in a driving simulator for a Greek research project, which examined the behavior of healthy drivers and drivers with mild brain diseases while driving [14]. The driving simulator records measurements at intervals of 16-17ms, which means that the measurements are about 60 per second, and gives information about at least 32 variables, which relate to speed, acceleration, vehicle position on the road, and steering position. In addition to these variables, through questionnaires, the variables relating to the characteristics of each driver, as well as variables resulting from the execution of the experiment and the scenarios in which the drivers were invited to drive are collected. These scenarios concerned traffic volume (high, low), attention or distraction (e.g. using mobile phone, conversation with a passenger) as well as sudden events during driving. For the needs of this research, only the data related to healthy drivers on country roads were used, so that the brain disease is not taken into account in the results.

The independent variables that were used for the identification of unexpected events for each of the twenty seven drivers that made up the sample are presented in the table below (Table 1). The sample of 27 drivers tested was considered satisfactory as it included all age groups of both sexes with many different levels of training (Figure 1), with or without distraction (cell phone use, conversation with a passenger). During each driving scenario, some events (sudden entry of an animal on the road, diversion of the vehicle of the opposite current, etc.) occurred. These events were also the dependent variable of this research as the goal is to identify them with the variable Event = 0 to symbolize the non-existence of event and Event = 1, which shows the existence of event. Only critical events and normal conditions were included for classification purposes. The time lag between the visibility of an emerging event and driver reactions are not documented and therefore were not used in the analyses. The other variables used were independent and are presented in Table 1.

Table 1: Initial set of independent variables considered

Time	PersonID
rdist	Age
rspur	AgeGroup
Speed	Gender

HWay	Education
Dleft	Driving Experience
DRight	Disease
Wheel	Trial
THead	Traffic
TTC	Distractor
AccLat	Event
AccLon	State

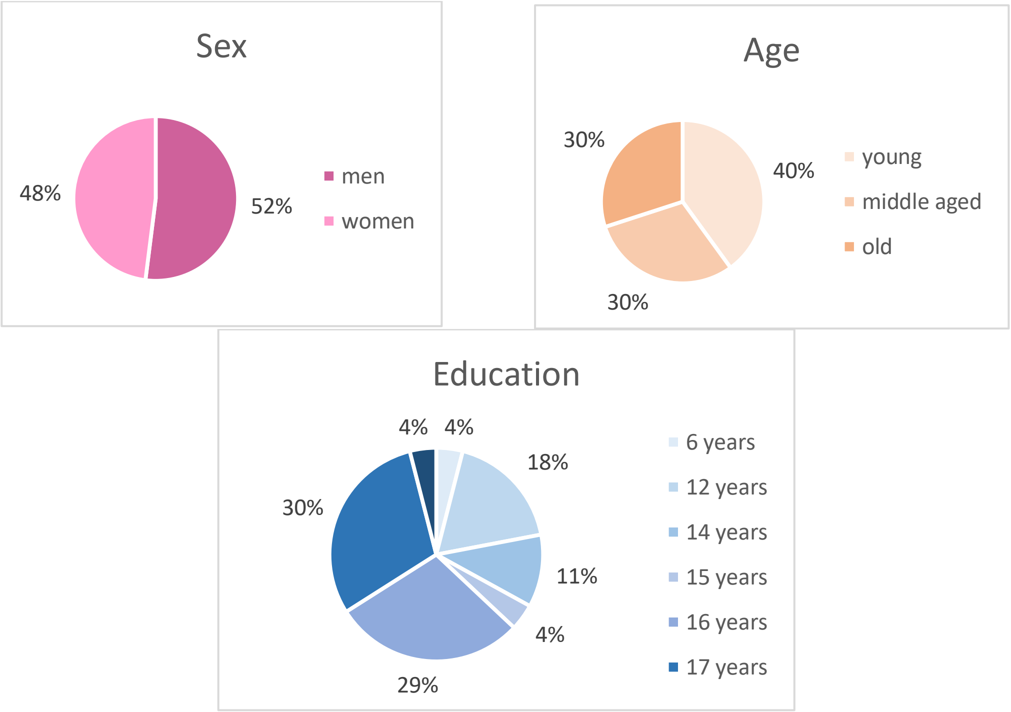

Figure 1: Sample characteristics with regards to sex, age and education years

The data was processed in the RStudio environment. From the initial database (RuralControl) that emerged from the experiment in the simulator, the data related to the duration of one minute before each event until its outcome were isolated. From these data, the starting time and the end time for each event, for each driver, were determined (index) and three separate databases were created, one consisting of the data one minute before each event (PreEvent), one with the data during each event (DurEvent) and a database that was the sum of the two previous databases (Events).

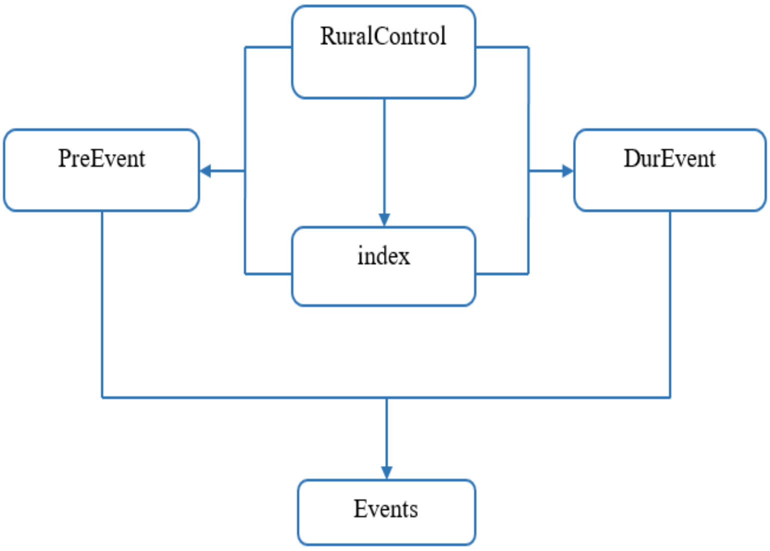

Figure 2: Flow chart for the extraction of the final database (Events) used in the classification analysis

For the first two tables (PreEvent, DurEvent) descriptive statistics were performed for the continuous variables as well as for the driving experience and age. These variables are:

Table 2: Variable description

Variable	Description
Speed	actual speed in km/h.
Acclat	lateral acceleration, in m/s^2.
Acclon	longitudinal acceleration, in m/s^2.
HWay	headway, distance to the ahead driving vehicle in m.
THead	time headway, i.e. to collision with the ahead driving vehicle, in seconds.
TTC	time to collision (all obstacles), in seconds.
DLeft	distance to the left road border in meters.
DRight	distance to the right road border in meters.
rdist	distance travelled in m.
rspur	distance of the vehicle from the median in m.
Wheel	Steering wheel position in degrees.
Age	age in years.
Driving Experience	driving experience in years.

The descriptive statistics for conditions before and during unexpected events are summarized in Tables 3 and 4.

Table 3: Descriptive statistics for conditions before unexpected events

Variable	Min	Median	Max	Variance	Standard Deviation
Speed (km/h)	0.00	45.68	103.60	351.013	18.735
AccLat (m/s^2)	-0.724	0.00	0.377	0.00	0.017
AccLon (m/s^2)	-9523.00	13.433	6399.00	37251.45	193.006
HWay (m)	1.20	956.10	1241.90	6368759.00	2523.64
THead (s)	0.20	1304.36	9993.70	11093839.00	3330.742
TTC (s)	0.40	6436.66	29.90	22907929.00	4786.223
DLeft (m)	-1.28	0.713	2.17	0.077	0.277
DRight (m)	-0.66	0.80	2.76	0.078	0.279
Rdist (m)	5.00	1142.392	2510.90	262085.20	511.943
Rspur (m)	-0.44	1.532	2.98	0.077	0.278
Wheel (deg)	-156.00	-4.126	109.00	232.217	15.239
Age (years)	22	41.322	78	270.506	15.239
DrExp (years)	3	19.898	46	182.114	13.495

Table 4: Descriptive statistics for conditions during unexpected events

Variable	Min	Med	Max	Var	Sd
Speed (km/h)	0.00	29.28	104.80	620.617	24.912
AccLat (m/s2)	-11.328	0.108	7048.00	619.669	24.893
AccLon (m/s2)	-487799.00	-49.50	3561.00	3597746.00	1896.773
HWay (m)	0.00	2018.80	1135.60	14169430.00	3764.230
THead (s)	0.00	3389.50	9969.10	21670860.00	4655.197
TTC (s)	0.00	5783.60	29.90	24347950.00	4934.364
DLeft (m)	-1.37	0.672	2.07	0.094	0.307
DRight (m)	-0.53	0.838	2.79	0.093	0.305
Rdist (m)	850.20	1551.80	2047.70	138608.90	372.302
Rspur (m)	-0.49	1.492	2.87	0.093	0.305
Wheel (deg)	-125.00	-4.094	137.00	222.189	14.906
Age (years)	22	41.41	78	272.435	16.506
DrExp (years)	3	20.01	46	184.354	13.578

From Tables 3 and 4, it is observed that if the values of the variables are compared, the following basic observations emerge:

- The average speed (Speed) was reduced during the events.

- The variation and standard deviation increased for lateral acceleration (AccLat) during the events.
- During the events the average value of the longitudinal accelerator (AccLon) decreased significantly, while variation and standard deviation increased.
- Also the average distance to the ahead driving vehicle (HWay) has increased, as well as time to collision with the ahead driving vehicle (THead).

In the overall table (Events) which contains both the precursors and the conditions during the events, the correlation of the independent variables was checked, so that the final databases consist only of the variables that do not have large correlation with each other. The results are shown in the following heat map of Figure 1. From this map it appears that the lateral acceleration (AccLat) with the longitudinal acceleration have a large correlation between them, as has the distance to the ahead driving vehicle (HWay) with the time to collision with the ahead driving vehicle (THead), the distance of the vehicle from the beginning of the drive (rdist) with the distance from the right (DRight) and from the left (DLeft) road border and driving experience with the age of the driver (Age).

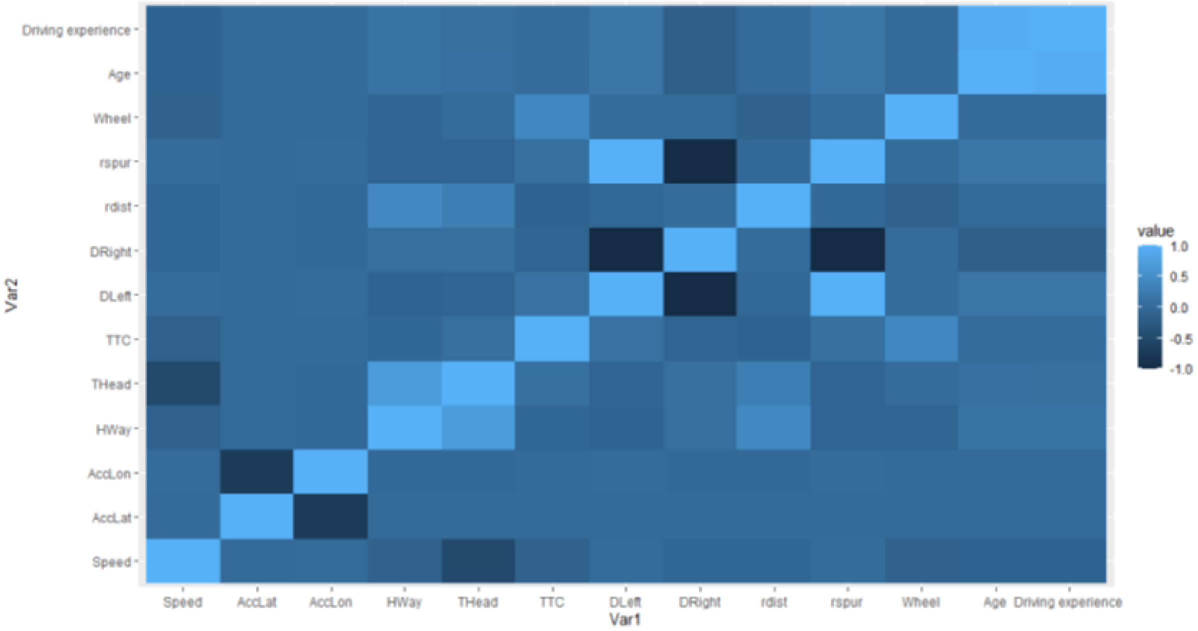

Figure 3: Correlation heat map

According to the above, the final databases used for classification models and factors analysis were formed. For the classification models, the data concerning the time duration one minute before until the outcome of each event were used. Two variants were performed, with the variable Event as the dependent variable with values 0 and 1 and the independent variables:

- Speed
- Longitudinal Acceleration
- Time Headway
- Distance travelled
- Distance from the road median
- Steering Wheel position
- Driving Experience

The aforementioned variables apart from driving experience were also used for the factor analysis, so as to identify characteristics among kinematic variables only with regards to the conditions before and during unexpected events.

3.2. Analysis results

For both the classification models, i.e. the binomial logistic regression and the random forest model, 2 different variants were performed in terms of the independent variables used to detect the events. The number and type of variables used in each variant were derived from the Boruta feature selection algorithm [15]which examines the mean significance of each independent variable in determining the dependent one.

Variant A included all statistically significant variables (table DATA1) while variant B included the 4 most important independent variables (table DATA2). So:

- Variant A: rdist, Speed, AccLon, DrExp, rspur, THead
- Variant B: rdist, Speed, AccLon, DrExp

The variable Wheel emerged as non-statistically significant for the event identification.

In each database used for the analysis, 80% was utilized for the training set and the remaining 20% was included in the test set (Figure 3).

Variant B for binomial logistic regression did not give satisfactory results so it is not presented in this paper. The results obtained from the confusion matrix for the developed models are presented in the following tables (i.e. Tables 5 and 6).

Table 5: Binomial logistic regression model results

	Variant A
Metrics	**Value**
accuracy	83.80%
Recall	27.90%
Specificity	96.60%
Precision	65.70%
F-measure	39.20%
False alarm rate	3.30%
AUC	80.00%

Table 6: Random Forests model results

	Variant A	**Variant B**
Metrics	**Value**	**Value**
accuracy	99.80%	99.20%
Recall	99.50%	96.60%
Specificity	99.90%	99.80%
Precision	99.60%	98.90%
F-measure	99.50%	97.80%
False alarm rate	0.09%	0.23%
AUC	99.99%	99.94%

The factor analysis for the expression of the independent variables through certain factors was applied to each data table separately, to the one that consisted of the data for the period of one minute before each event (PreEvent), the one with data during each event (DurEvent) and the third which is the sum of the two previous (Events) and the results were obtained and presented in Table 7. Scree plot was usef for factor extaction.

Table 7: Factor Analysis results

Data Table	**Factor 1**	**Factor 2**	**Factor 3**
PreEvent	Speed THead	rdist	
DurEvent	Speed	rspur	THead
Events	Speed Thead	rspur	rdist

From Table 8, it is evident that speed as well as total distance travelled and distance of the vehicle from the median play a significant role prior and during events. The conditions before unexpected events can be better described through values from speed and time headway, while during the event speed and distance from the median are extremely important.

4. Discussion

For the logistic regression and random forests models, two variants, namely A and B, were created in order to examine the possibility of predicting an unexpected event. The results showed that the random forests model works best for this taask. More specifically, the variant A revealed that the logistic regression model did not have the ability to predict well the existence of an incident, although it was quite effective to predict safe driving. This is a known problem in imbalanced datasets[16], [17]. On the contrary, the random forests model gave very satisfactory results in both variants, both for the prediction of incidents and safe driving, with a very small probability of error. The reason behind this finding is their know ability to work well with imbalanced sets [1].Comparing the two variants of the random forest model, variant A seems to give better results.

Factor analysis was carried out separately for the data relating to the time before the event, during the event and the sum of them. It has been observed that the results in these 3 cases differed both in the number of factors and variables expressed by them. The data relating to the period one minute before the event revealed that they can be expressed through speed, acceleration and the total distance that the driver has traveled. This could be reasonably explained as, as it has been found in previous researches that speed and acceleration vigorously alter before an event [18]. Furthermore, the feeling of fatigue that may have been created in the driver due to the long distance is likely to increase the likelihood of involvement in an event [14].

It has been found that speed is an important factor also in the case of data during events while acceleration and distance cease to play a dominant role, with their places overtaken by the vehicle deviation from the median of the road and time to collision.

The models developed for the detection of incidents in this research can be used to further improve road safety. For example, the event detection algorithm based on the driving characteristics of one minute before incidents, can be exploited by a Traffic Management Center for a proactive traffic management system initiation[19]. Evaluating driving characteristics continuously, a warning message could also be displayed to the driver through the driver-vehicle interfaces (e.g. HMIs), in case of driving behavior that may lead to an event, so that the driver can take the necessary corrective actions to avoid it. Furthermore, smartphone application could be developed, which will warn the driver about the need to perform corrective actions to avoid the predicted event [20].

5. Conclusions and future research

The purpose of this paper was to develop classification models in order to detect rural road events based on driving characteristics. According to the mean significance of the independent variables estimated by the Boruta algorithm, it turned out that the variables that describe the vehicle speed, longitudinal acceleration, total distance traveled as well as the driver's driving experience, are the variables with the greatest importance for identifying an unexpected event.

From the classification models developed, it was found that the random forest model was much more efficient than the binomial logistic regression in identifying safety-critical events. Unfortunately, the use of the binomial logistic regression model was observed to be ineffective in locating an event with a small number of independent variables, demonstrating low recall, relatively high probability of error of classification, as well as low accuracy and F-measure values. Comparing the two variants utilized with the random forest model for statistical analysis, it was found that although both variants produce useful and reliable results and are acceptable, the one containing the largest number of variables gave better classification results.

With regards to the factor analysis, it was shown that different factors can be found in i) the data that describe the duration of one minute before each event, ii) those that describe the duration of each event, and iii) the sum of them. The data describing the situation one minute before each event was found to be better described with two factors, one describing the influence of speed and time until the collision and one describing the influence of the total distance traveled by the vehicle. These two factors are likely to arise as, as it has been observed in the international literature, that speed and time to the collision play a decisive role in the likelihood of getting involved in an unexpected incident on the road. In contrast to the duration of one minute before, during an event, the variables that have the greatest influence along with speed are the deviation of the vehicle from the middle of the road and the time to collision. Finally, the data describing the situation one minute before and during each event, turned out to be expressed by three factors: one for the influence of speed and time until the collision from the vehicle in front, one that describes the effect of the deviation of the vehicle from the middle of the road and the factor of the influence of the total distance traveled of the vehicle on the existence of an event.

Further research may be conducted to better address the aim of this research. For example, driving characteristics 5 minutes before the event could be exploited or naturalistic driving data could be used instead of data from a driving simulator. Another future investigation could investigated data from other road environments such as rural roads or highways. Finally, exploring a larger sample of participants and different methods of statistical analysis and machine or deep learning could further enhance the results presented in this paper.

References

[1] C. Katrakazas, M. Quddus, and W. H. Chen, "A simulation study of predicting real-time conflict-prone traffic conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 10, pp. 3196–3207, 2018, doi: 10.1109/TITS.2017.2769158.

[2] A. P. Tarko, "Traffic conflicts as crash surrogates," *Meas. Road Saf. Using Surrog. Events*, pp. 31–45, 2020, doi: 10.1016/b978-0-12-810504-7.00003-3.

[3] E. Michelaraki, C. Katrakazas, T. Brijs, and G. Yannis, "Modelling the Safety Tolerance Zone : Recommendations from the i-DREAMS project," in *10th International Congress on Transportation Research*, 2021, pp. 1–18.

[4] H. J. Foy and P. Chapman, "Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation," *Appl. Ergon.*, vol. 73, no. June 2017, pp. 90–99, 2018, doi: 10.1016/j.apergo.2018.06.006.

[5] M. Q. Khan and S. Lee, "A comprehensive survey of driving monitoring and assistance systems," *Sensors (Switzerland)*, vol. 19, no. 11, 2019, doi: 10.3390/s19112574.

[6] P. Choudhary and N. R. Velaga, "Effects of phone use on driving performance: A comparative analysis of young and professional drivers," *Saf. Sci.*, vol. 111, no. July 2018, pp. 179–187, 2019, doi: 10.1016/j.ssci.2018.07.009.

[7] G. Yannis, E. Papadimitriou, and P. Papantoniou, "Distracted driving and mobile phone use: Overview of impacts and countermeasures," *Commun. Technol. Road Saf.*, pp. 1–23, 2014.

[8] P. Papantoniou, G. Yannis, and E. Christofa, "Which factors lead to driving errors? A structural equation model analysis through a driving simulator experiment," *IATSS Res.*, vol. 43, no. 1, pp. 44–50, 2019, doi: 10.1016/j.iatssr.2018.09.003.

[9] P. Papantoniou, C. Antoniou, G. Yannis, and D. Pavlou, "Which factors affect accident probability at unexpected incidents ? A structural equation model approach," *J. Transp. Saf. Secur.*, vol. 0, no. 0, pp. 1–18, 2018, doi: 10.1080/19439962.2018.1447523.

[10] Papazikou et al., "What came before the crash? An investigation through SHRP2 NDS data," *Saf. Sci.*, vol. 119, no. March, pp. 150–161, 2019, doi: 10.1016/j.ssci.2019.03.010.

[11] S. P. Washington, M. G. Karlaftis, and F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press, 2010.

[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45.1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[13] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998, doi: 10.1109/34.709601.

[14] D. Pavlou, "Traffic and safety behaviour of drivers with neurological diseases affecting cognitive functions," 2016.

[15] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[17] C. Katrakazas, C. Antoniou, and G. Yannis, "Identification of driving simulator sessions of depressed drivers: A comparison between aggregated and time-series classification," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 75, pp. 16–25, 2020, doi: 10.1016/j.trf.2020.09.015.

[18] D. Pavlou, I. Beratis, E. Papadimitriou, C. Antoniou, G. Yannis, and S. Papageorgiou, "Which Are the Critical Measures to Assess the Driving Performance of Drivers with Brain Pathologies?," *Transp. Res. Procedia*, vol. 14, pp. 4393–4402, 2016, doi: 10.1016/j.trpro.2016.05.361.

[19] M. Hossain, M. Abdel-Aty, M. A. Quddus, Y. Muromachi, and S. N. Sadeek, "Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements," *Accid. Anal. Prev.*, vol. 124, no. July 2018, pp. 66–84, 2019, doi: 10.1016/j.aap.2018.12.022.

[20] A. M. Stavrakaki, D. I. Tselentis, E. Barmpounakis, E. I. Vlahogianni, and G. Yannis, "Estimating the necessary amount of driving data for assessing driving behavior," *Sensors (Switzerland)*, vol. 20, no. 9, 2020, doi: 10.3390/s20092600.