Factors contributing to safety-critical events in urban areas: A driving simulator study

Fotini Bardi[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

Background: The emergence of automated vehicles and new Advanced Driver Assistance Systems (ADAS), has brought about the need of identifying safety-critical events in real-time in urban environments. Nevertheless, the identification of critical factors contributing to such critical events is still ongoing.
Objective: The aim of this paper is to determine critical factors for the identification of safety critical driving events in urban areas using highly disaggregated data from a driving simulator experiment.
Method: Two statistical models were developed, namely a binomial logistic regression and a random forest one, in order to compare statistical and machine learning approaches for the identification of critical events. Furthermore, factor analysis was performed in order to investigate the existence of common factors in the group of independent variables.
Results: The random forest model provided more reliable results in predicting events when compared to binomial logistic regression. More specifically, including more independent variables in the model was found to be more effective and with a lower false alarm rate. Moreover, factor analysis demonstrated that speed, the deviation of the vehicle from the middle of the road and the distance from the right boundary line can be event precursors, while speed and acceleration are representative variable during the event timeline.
Conclusions: Speed as well as lateral and longitudinal acceleration along with lateral distances were found to be the most critical factors for identifying events in urban roads. Nevertheless, a larger sample of drivers and driving conditions (e.g. weather or distraction) and a naturalistic driving dataset could offer better results in future analyses.

Keywords: safety-critical events, urban areas, driving simulator, random forests, logistic regression, factor analysis

[1] * Corresponding author. Tel.: +302107721265;
*E-mail address:*ckatrakazas@mail.ntua.gr

1 Introduction

Road safety is an issue of great social importance as the number of road traffic deaths and injuries is too high worldwide (WHO, 2018). The analysis of factors that lead to wrong and unsafe driving behavior, the effect of distraction on driving performance as well as how the driving characteristics change in the occurrence of an event have been heavily researched in the literature (e.g. [1]–[3]).

Factors that affect the probabiblity of an accident include driving behavior (e.g. aggressive driving, drowsiness, frequency of sudden accelerations or decelerations in [4], distraction [5], [6] as well as the influence of the driving environment [7].

The effects of distraction have been thoroughly investigated in many studies. Surprisingly, according to several researches, the involvement of drivers in minor actions (such as texting) draws their attention to the main task of driving and prevent accidents [6], especially when dangerous changes occur [5]. After an unexpected event, likewise, drivers talking on a mobile phone are more careful due to the sense of security threat they feel. On the other hand, when they talk to a fellow passenger, their attention is more easily distracted and they do not react in the same careful way [8].

In addition to distraction, research has shown that the influence of driver characteristics as well as the type of driving area [7] are factors that lead to incorrect driving behavior. Extensive research has been done on how driving behavior and the possibility of an accident are affected by the driver's distraction and participation in minor actions. The influence of unexpected events on driving characteristics as well as the change of these characteristics before an accident has been analyzed using various statistical analysis models [1], [9].

With regards to unexpected events and traffic conflicts, the majority of studies are focused on the identification of such events for real-time safety evaluation (e.g. [3], [10]), but there is yet no focus on factors that play an important role both before and during such an event. This forms the motivation for the current paper, which aims at determining critical factors for the identification of safety- critical events in urban areas. For that purpose, this research uses data collected from a driving simulator experiment investigating the behavior of 41 healthy drivers. For the data analysis, two statistical models are developed, namely a binomial logistic regression and a random forest one, both of which considered the occurrence of an event as the dependent variable. Furthermore, factor analysis was performed with regards to data concerning one minute before the event, the duration of the event as well as the combination of these cases, in order to investigate the existence of common factors in the group of independent variables.

2 Methodology

The problem of identifying an unexpected events is a binary classification problem, and therefore binary classification models were developed in this paper. Binomial logistic regression was chosen as a statistical model in this study due to the binary form of the dependent variable (Event), which takes the value 1 for the existence of an event and the value 0 for non – existence respectively. Similarly, the random forest method [11] was chosen to be used in the present analysis as it can manage large data, it provides the highest accuracy among many classification methods used for similar purposes[2] and it offers more capabilities in working with missing data.

Binomial logistic regression is a statistical model that looks for the relationship between a distinct dependent variable and one or more independent variables. Dependent variable is the variable whose value is predicted and independent is the variable which is given and used to predict the dependent one. In the present study the dependent variable (Event) takes the value 1 for the existence of an event and the value 0 for non-existence respectively. The form of the model equation is:

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

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 [11]. Random Forests are an evolution of bagged trees and uses the bagging algorithm along with the random subspace method proposed by Ho [12]. Each tree is built using the impurity Gini index.

In order to evaluate the two classification models (i.e. binary logistic regression and random forests), the confusion matrix was utilized, as well as the logical explanation of the coefficient signs and the z-test for statistical significance with regards to the binary logistic regression model. In order to exploit the confusion matrix, in the present study, the existence of an event was defined as positive and the non-existence as negative class, respectively. The performance metrics that were used are:

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision= $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$
- F-measure = $\frac{2*Precision*Recall}{Precision+Recall}$
- False alarm rate = $\frac{FP}{TN+FP}$

where TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive.

In order to identify relationships between variable before and during the occurrence of unexpected events, factor analysis [13] is a linear statistical model used to explain the variation between observed variables and to summarize a set of them into unobserved variables called factors. Factor analysis is an accurate method and was chosen to be used in order to reduce the number of variables by combining two or more of them into a single factor and to identify groups of inter-related variables and examine how they relate to each other.

Two indicators are provided to control the quality of the data:
• The Keiser-Meyer-Olkin (KMO) Index, which assesses the adequacy of the sample (KMO> 0.50)
• The Bartlett's Test of Sphericity Index (p), which assesses whether the correlations between the variables allow the application of factor analysis ($p < 0.05$)

3 Analysis and Results

3.1 Data collection

Data are obtained from a large dataset created through a driving simulator experiment for the preparation of a doctoral thesis [14] and they relate to driving characteristics. It was decided that useful variables that could be used in this research are time, total distance traveled, vehicle deviation from the middle of the road, speed, distance from the vehicle in front, distance from the right and left lane, steering wheel position, time to collision with the vehicle in front, time to collision with all the obstacles, lateral and longitudinal acceleration. All of them are crucial factors in the way the vehicle moves and is positioned on the road and therefore in the probability of its involvement in an unexpected incident. Each driver's age, gender, education, driving experience were also collected from questionnaires they were requested to fill in. Furthermore, traffic conditions as well as the existence or not of an unexpected event emerged from the different scenarios performed by the drivers in the simulator. From the above

data, the UrbanControl database was created and used in the present research, which contains data on 41 drivers, men and women, from all ages and levels of education with different driving experience and only applies to urban areas. Their social characteristics as well as their driving experience are illustrated below in Figure 1.

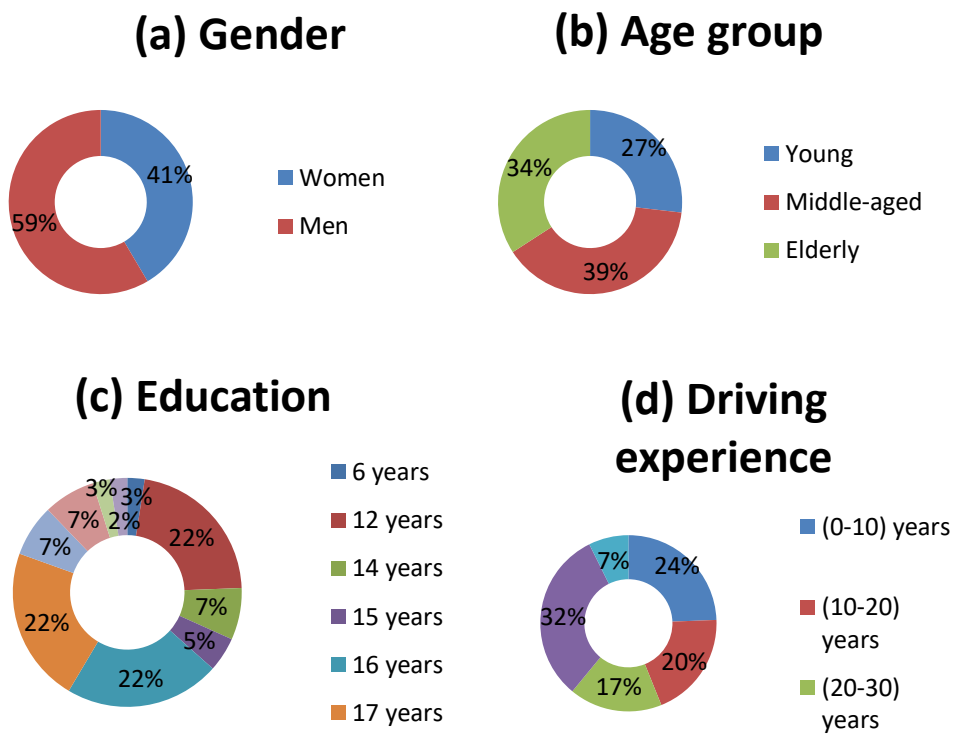

Figure 1: Participants allocation based on a) Gender, b) Age group, c) Education and d) Driving experience

3.2 Data Processing

The above database, was divided into three individual data tables, using the R programming language, so that the detection of incidents based on driving characteristics is possible. A table containing the data during events was created (DurEventU), a table containing data one minute before the event (PreEventU) and a total table containing the sum of the aforementioned data (EventU). The tables were separated based on the Event variable. Also, in order to help the separation of tables, the Index table was created which records the start and end times of each event.

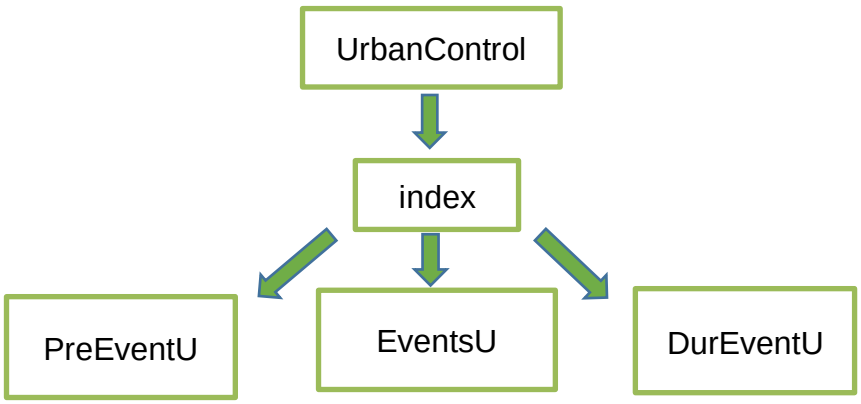

Figure 2: Creation of PreEventU, DurEventU and EventsU tables

Subsequently, descriptive statistics were performed in DurEventU and PreEventU tables, and the correlation between the independent variables was determined. The variables which were highly correlated are:

- Distance from the vehicle in front with time to collision with the vehicle in front
- Distance from the left lane with distance from the right lane
- Age with Driving experience.

Thereafter, the variables decided to be removed are distance from the vehicle in front, distance from the left lane and age.

Table 1: Correlation matrix

	Speed	AccLat	AccLon	HWay	THead	DLeft	DRight	rdist	rspur	Wheel	Age	Driving experience
Speed	1.00											
AccLat	0.00	1.00										
AccLon	-0.01	0.12	1.00									
HWay	0.07	-0.03	0.00	1.00								
THead	-0.23	-0.03	0.00	0.83	1.00							
DLeft	0.01	-0.08	0.02	0.31	0.25	1.00						
DRight	0.01	-0.08	0.02	0.31	0.25	1.00	1.00					
rdist	0.26	0.01	-0.03	-0.11	-0.20	-0.27	-0.27	1.00				
rspur	0.01	0.08	-0.03	0.00	0.00	-0.05	-0.05	0.11	1.00			
Wheel	0.06	0.04	0.05	-0.01	-0.03	0.03	0.03	-0.01	-0.03	1.00		
Age	-0.10	0.00	-0.02	0.12	0.09	0.01	0.01	0.04	0.07	0.00	1.00	
Driving experience	-0.06	0.00	-0.02	0.11	0.07	0.01	0.01	0.04	0.04	-0.01	0.87	1.00

Table 2: Variable description

Variables	Explanation
Speed	Speed (km/h)
AccLat	Lateral acceleration (m/s^2)
AccLon	Longitudinal acceleration (m/s^2)
HWay	Distance from the vehicle in front (m)
THead	Time to collision with the vehicle in front (s)
DLeft	Distance from the left lane (m)
DRight	Distance from the right lane (m)
rdist	Total distance traveled (m)
rspur	Vehicle deviation from the middle of the road (m)
Wheel	Steering wheel position (degrees)
Age	Age (years)
Driving experience	Driving experience (years)

According to the above, the final data table (Model) was created by extracting from the EventU table the columns: event, speed, lateral acceleration, longitudinal acceleration, time headway with the vehicle in front, distance from the right lane, total distance travelled, vehicle deviation from the middle of the road, steering wheel position and driving experience. Both statistical analysis models were developed using the Model table.

Finally, the Factor Analysis method was performed in the EventsU2, PreEventU2 and DurEventU2 tables. The EventsU2 table was created by isolating speed, lateral acceleration, longitudinal acceleration, time to collision with the vehicle in front, distance from the right lane, total distance traveled, vehicle deviation from the middle of the road and steering wheel position columns from the Model table. The PreEventU2 table was created by isolating the same columns from the Model table only when Event = 0, i.e. the data related to the non-existence of an event. The DurEventU2 table was created by isolating the same columns from the Model table only when Event = 1, i.e. the data related to the existence of the event. The development of the aforementioned tables is depicted in Figure 3.

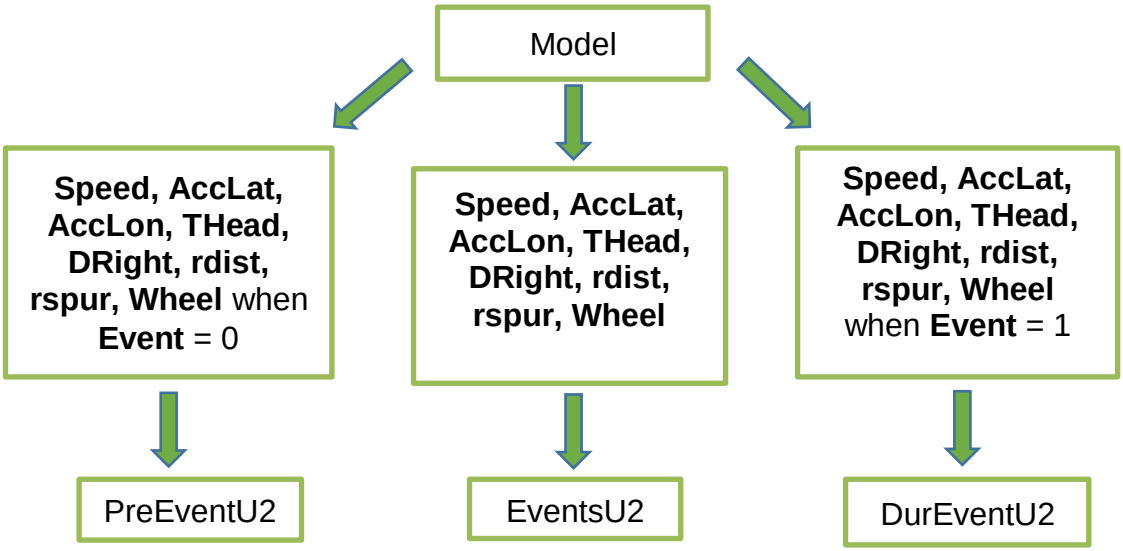

Figure 3: Creation of PreEventU2, DurEventU2 and EventsU2 tables

3.3 Results

After defining the methodology, the analysis method was applied. Two different independent variable settings were explored; Version A including all the independent variables in the EventsU2 dataset, and Version B, which included statistically significant variables using the Boruta variable importance algorithm [15]. Version B included speed, time headway, distance from the right side of the road, distance driven, steering wheel position and driving experience.

Table 3: Variable importance

Variable	Boruta Importance
Wheel	228.78
rdist	177.71
speed	173.15
Dright	160.84
Driving Experience	143.86
Thead	129.7

Table 4: Models evaluation results

Performance metrics	Binomial logistic regression (Version A)(%)	Random Forest (Version A) (%)	Random Forest (Version B) (%)
Accuracy	79.26	87.17	81.19
Recall	13.44	65.56	53.51
Specificity	99.04	93.67	89.51
Precision	80.8	75.68	60.53
F-measure	23.05	70.26	56.81
False alarm rate	0.96	6.33	10.49

According to the results, the **binomial logistic regression model** (Version A) presents a very low recall index (recall), so it does not predict at all the positive moments, i.e. the existence of an event, while it has the ability to predict satisfactorily the non-existence of an event (high specificity). The overall accuracy is satisfactory, ie the model is reliable for the correct predictions as well as for the degree of accuracy of the process (precision). The possibility of false alarm rate (false alarm rate) is also quite satisfactory but the F-measure which expresses the harmonic means of accuracy and recall is very low.

The binomial logistic regression model for Version B was not considered worthy of analysis as it did not predict the existence of an event satisfactorily.
It follows from the above that the binomial logistic regression model A does not work satisfactorily in event detection. For this reason, the random forest model is also investigated.

The **random forest model** (Version A) predicts marginally satisfactorily the positive moments (recall), ie the existence of an event, and has the ability to predict very well the non-existence of an event (high specificity). The model is also reliable for the correct predictions as it presents a high accuracy index (accuracy), but also for the degree of accuracy of the classification process (high precision). The possibility of incorrect classification of positive snapshots (false alarm rate) is satisfactory as well as the F-measure which expresses the harmonic means of accuracy and recall.
The random forest model (Version B) presents a marginally satisfactory recall index, ie prediction of the existence of an event, while it has the ability to predict well the non-existence of an event (high specificity). The overall accuracy is satisfactory, i.e. the model is reliable for the correct predictions. The probability of incorrect classification of positive snapshots (false alarm rate) is marginally satisfactory as well as the measure F (F-measure) which expresses the harmonic mean of accuracy. The degree of accuracy of the classification process (precision) is also marginally satisfactory.

From the above it can be concluded that the random forest model for Version A satisfactorily predicts the existence of an event and is considered reliable. As for Version B it is considered marginally reliable.

Table 5: Factor analysis results

Table	Factor 1	Factor 2	Factor 3
PreEventU2 (1 min before the event)	DRight	Speed	rspur
DurEventU2 (during the event)	Speed	AccLat AccLon	
EventsU2 (before and during events)	AccLat	Speed	DRight

Initially, the number of factors was determined by the method of the main components and then, for this number, the method of factor analysis was performed. From the results emerged, it turned out that the situation one minute before the event can be described through the influence of the distance from the right borderline, the speed and the deviation of the vehicle from the middle of the road. The situation during the event can be expressed through the influence of speed and longitudinal and lateral acceleration while the whole is expressed through lateral acceleration, speed and distance from the right boundary line.

4 Conclusions

The present research aimed to determine critical factors for the identification of safety - critical events in urban areas. For this reason, data created through a driving simulator experiment were utilized, investigating the behavior of drivers.

After data collection and processing, two statistical analysis models were developed, the binomial logistic regression model and the random forest model, in order to investigate their ability to predict the existence of an event. Furthermore, factor analysis was performed with regards to data concerning one minute before the event, the duration of the event as well as the combination of the aforementioned cases.

Results of the developed models lead to very useful and important conclusions. Firstly, the variables with great significance to event identification in urban roads are speed, total distance traveled, distance from the right lane,

steering wheel position and time to collision with the vehicle in front. With regards to statistical analysis models, the results reveal that the random forest model provided more reliable results in event detection than the binomial logistic regression model. More specifically, version A of the random forest model is considered more effective as it presents an increased probability of event prediction with a lower false alarm rate. In both models of statistical analysis, the version included the highest number of variables showed more reliable results. Moreover, factor analysis demonstrated that speed as well as lateral and longitudinal acceleration along with lateral distances were found to be the most critical factors for identifying events in urban roads.

Further research should focus on different methodological approaches as well as in examination of more road types, in addition to urban environment, and alternative driving scenarios, for example high or low traffic conditions with rain or fog. Naturalistic driving data could also be utilized in the future, instead of driving simulator data, as they present a better view of the real conditions prevailing on the roads. Finally, the investigation of the change of driving characteristics in a longer period of one minute before the event would be interesting in future research.

References

[1] 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.

[2] 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.

[3] C. Katrakazas, A. Theofilatos, M. A. Islam, E. Papadimitriou, L. Dimitriou, and C. Antoniou, "Prediction of rear-end conflict frequency using multiple-location traffic parameters," *Accid. Anal. Prev.*, vol. 152, no. December 2019, 2021, doi: 10.1016/j.aap.2021.106007.

[4] M. Monselise, O. S. Liang, and C. C. Yang, "Identifying Important Risk Factors Associated with Vehicle Injuries Using Driving Behavior Data and Predictive Analytics," *2019 IEEE Int. Conf. Healthc. Informatics, ICHI 2019*, 2019, doi: 10.1109/ICHI.2019.8904860.

[5] O. A. Osman, M. Hajij, S. Karbalaieali, and S. Ishak, "A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data," *Accid. Anal. Prev.*, vol. 123, no. May 2018, pp. 274–281, 2019, doi: 10.1016/j.aap.2018.12.005.

[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] 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.

[8] P. Papantoniou, D. Pavlou, G. Yannis, and E. Vlahogianni, "How an unexpected incident affects speed related driving performance measures," *Transp. Res. Procedia*, vol. 41, no. 2016, pp. 529–531, 2019, doi: 10.1016/j.trpro.2019.09.087.

[9] X. Shi, Y. D. Wong, M. Z. F. Li, and C. Chai, "Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory," *Accid. Anal. Prev.*, vol. 117, no. May, pp. 346–356, 2018, doi: 10.1016/j.aap.2018.05.007.

[10] L. Zheng, T. Sayed, and F. Mannering, "Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions," *Anal. Methods Accid. Res.*, vol. 29, p. 100142, 2021, doi: 10.1016/j.amar.2020.100142.

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

[12] 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.

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

[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.