# Predicting risky driving behavior with classification algorithms: Results from a large-scale field-trial and simulator experiment

Thodoris Garefalakis[a]*, Eva Michelaraki[a], Stella Roussou[a], Christos Katrakazas[a], Tom Brijs[b] and George Yannis[a]

[a]National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Heroon Polytechniou str., Athens, GR-15773, Greece
[b]UHasselt, School for Transportation Sciences, Transportation Research Institute (IMOB), Agoralaan, 3590 - Diepenbeek, Belgium

*Corresponding author: tgarefalakis@mail.ntua.gr

**Abstract**
Road safety is a subject of significant concern and substantially affects individuals across the globe. Thus, real-time, and post-trip interventions have gained significant importance in the past few years. The European Union's Horizon 2020 project i-DREAMS has also directed its attention towards this aspect. In particular, i-DREAMS aimed to define, develop, test, and validate a 'Safety Tolerance Zone (STZ)' in order to prevent drivers from risky driving behaviors using interventions both in real-time and post-trip. This study aimed to analyze different classification techniques and examine their ability to identify dangerous driving behavior based on a dual-approach study. The analysis was based on the investigation of important risk factors such as average speed, harsh acceleration, harsh braking, headway, overtaking, distraction (i.e., mobile phone use), and fatigue. In order to achieve the objective of this study, significant data were collected through a driving simulator as well as a naturalistic driving experiment. Based on the data collected for each of the two approaches, several classification models were developed, analyzed, and compared, according to their performance. To that end, four classification algorithms, namely Support Vector Machines (SVMs), Random Forest (RFs), AdaBoost, and Multilayer Perceptron (MLP) Neural Networks were implemented. The proposed methods were compared based on different evaluation metrics and it emerged that RFs and MLPs outperformed the rest of the classifiers with 84% and 82% overall accuracy, respectively, while the maximum average speed of the vehicle was found to be the most crucial predictor for identifying the driving time at each safety level. Risky and aggressive driving behavior is a worldwide critical social and public health concern. The findings of this research could provide essential guidance for decision-makers to initiate concrete steps for engineering applications in road safety management.

**Keywords:** Driving behavior, Random Forests, Machine Learning models, Classification algorithms, Driving Simulator Study, Naturalistic Driving Study

## 1. Introduction

Despite global and extensive efforts to mitigate crashes, casualties have not disappeared - with significant social consequences constantly emerging. According to the World Health Organization (WHO), 1.3 lives are lost each year due to road crashes, becoming the 8th

cause of death for all ages and the 1st for people aged between 5-29 years old (WHO, 2018). Considering the evolution in transport and the complexity of modern transportation systems, an opportunity is offered for safer driving behavior, which of course poses certain challenges and risks. In line with this direction, the WHO and the European Union have set a 50% reduction goal in road crashes for the decade 2021-2030 focusing on using new technologies.

Driving behavior is a complex issue that is affected by a wide range of factors, including driver's characteristics as well as environmental and traffic variables. However, human error stands out as the most significant contributor to road crashes (Staubach, 2009). Cognitive processes such as attention, perception, and decision-making each play an essential role in how drivers adapt to changing road conditions and make split-second decisions. Understanding these factors and their interrelationship is essential for developing effective road safety interventions and integrating emerging technologies to mitigate human errors and reduce the number of road crashes. Emerging technology systems can significantly reduce the likelihood of such collisions by reducing cognitive overload and thus removing human involvement in driving tasks (Khoury & Hussein, 2023).

Based on the integration of emerging technologies in the European Union's commitment to improve road safety and minimize road fatalities, the European H2020 project i-DREAMS aims to define, develop, test, and validate a 'Safety Tolerance Zone' (STZ) (Michelaraki et al., 2021). Through a smart system, i-DREAMS aims to identify the level of 'STZ', by monitoring and evaluating risk indicators related to the complexity of the driving task as well as the ability to cope with the challenges posed by it, and thus support drivers to operate within safe boundaries. The STZ is classified into three risk levels: 'Normal', 'Dangerous', and 'Avoidable Accident'. The distinction between the three levels lies in whether the driver is operating with safety (i.e., 'Normal' level) or not (i.e., 'Dangerous', 'Avoidable Accident' level). Levels 'Dangerous' and 'Avoidable Accident refer to the high probability of collision, with the significant difference that in the case of 'Avoidable Accident' the need for action is more urgent.

Therefore, based on the above framework this paper aims to develop and evaluate different classification models, leveraging two distinct data sources: simulator data and naturalistic driving data. This dual-source methodology not only enhances the diversity and richness of the dataset but also allows for a comprehensive evaluation of machine learning models in both controlled and real-life driving conditions, thereby advancing our understanding of driver behavior across different contexts.

The paper is structured as follows. In the beginning, an overview of this paper's objective and the gaps it seeks to fill is provided. This is followed by the description of the research methodology, encompassing the theoretical foundations of the models utilized. Moreover, the collection process (i.e., simulator and field trials) and the processing of the dataset are described. Finally, the results of the analysis are presented accompanied by relevant conclusions on the different data collection approaches and road safety in general.

## 2. Literature Review

Driving Simulator Studies (DSS) and Naturalistic Driving Studies (NDS) are the two main approaches that have been extensively employed in driving behavior analysis research (Osman et al., 2019). These research methodologies have provided valuable insights into the multifaceted nature of risky driving behaviors and have become indispensable tools for understanding the factors that contribute to road safety challenges. A recent study (Wijayaratna et al., 2019) has examined the use of both methodologies to analyze the impact of mobile phone conversation on the task of driving. Results showed that DSS tend to reveal an increased risk of crash due to mobile phone use, while the NDS, suggested a reduction in crash risk. The benefit of each approach is different, and it would be helpful to compare them in order to draw comprehensive conclusions. For example, DSS offers a wide range of driving scenarios and requires less time to collect data in well-controlled environments compared to field studies (Nasr Azadani & Boukerche, 2022). On the other hand, NDS has a higher degree of realism reflecting more accurately the natural driving situation (Wang et al., 2022).

Due to their high accuracy, machine learning-based models are widely used in the field of road safety and are exploited to predict risky driving behavior. Given this context, recent studies utilized such models such as Random Forest (RFs; (Song et al., 2021)), Multilayer Perceptron (MLP; (Shangguan et al., 2021)), Support Vector Machines (SVMs; (K. Yang et al., 2021; Zhang et al., 2016)), eXtreme Gradient boosting (XGBoost; (Shi et al., 2019)), Decision Trees (DT; (K. Yang et al., 2021)), Gradient Boosting (GB; (Ghandour et al., 2021)) and Logistic Regression (LR; (Papadimitriou et al., 2019)).

Various methodologies have been proposed in recent studies to assess and predict risky driving behavior, each employing diverse approaches and algorithms. For instance, Shangguan et al. (2021) devised a framework encompassing feature extraction, clustering techniques, feature importance analysis, and the utilization of machine learning algorithms including RF, XGBoost, SVM, and MLP, demonstrating an accuracy exceeding 85%. Similarly, Yang et al. (2021) investigated a driving simulator dataset, developed clustering techniques to distinguish the different levels, and applied three classification algorithms (i.e., SVM, Decision Tree, and Nave Bayes classifier), with the highest accuracy being 95%, to classify and evaluate different risk levels of driving behavior. Additionally, Shi et al. (2019) introduced a risk prediction framework, incorporating feature selection, risk level labeling, addressing imbalanced datasets, and employing an XGBoost classification model with an overall accuracy of 89%.

Furthermore, Zhang et al. (2016) successfully classified driving behaviors by utilizing low-level sensors, combining smartphone and OBD data, and applying an SVM algorithm, resulting in an accuracy of 86.67%. Another study by Papadimitriou et al. (2019) quantified the correlation between dangerous driving and mobile phone usage through logistic regression, with a marked accuracy of 70%. Lastly, Ghandour et al. (2021) classified driving behavior based on psychological states, employing machine learning techniques, and identified Gradient Boosting as the optimal method for level prediction within this context.

Therefore, based on the gaps in the literature, this research aims to gain deeper knowledge and understanding regarding the development of driver behavior identification models and the factors that affect it. Through the dual approach (i.e., Driving Simulator Study and Naturalistic Driving Study), a holistic overview of the topic is pursued.

## 3. Data Collection

For the purpose of the study, a simulator experiment and a naturalistic driving experiment were carried out in order to collect and analyze data from Belgian car drivers. The value of the two data sources is that they address driving behavior in controlled conditions and a specific environment (i.e., Simulator experiment) as well as in a real-world context (i.e., Naturalistic Driving experiment). Both approaches have certain limitations. While in the first case simulator data are difficult to apply to real-world conditions, on the other hand, the absence of experimental control in the context of natural driving (ND) data collection inherently limits the possibility of establishing unambiguous causal relationships between specific variables and road user behavior (van Schagen & Sagberg, 2012).

Within the framework of the simulation experiment, and to determine the three safety levels (i.e., the dependent variable of the classification process), specific headway thresholds were applied (Garefalakis et al., 2022) based on the literature. Conversely, in the second approach of the Naturalistic Driving experiment, these thresholds were integrated into mapping the different safety levels during the driving task. The range of values for the headway corresponding to each safety level is:

- ‘Normal’ Level: Headway > 2 sec
- ‘Dangerous’ Level: Headway > 1.4 sec and Headway < 2 sec
- ‘Avoidable Accident’ Level: Headway < 1.4 sec

The variables collected for the analysis were the same in both approaches to ensure consistency. In addition, the variables that were finally evaluated were three, as shown in Table 1, following the process of feature selection and permutation feature importance.

*Table 1: Description of variables collected*

| Variable | Description | Units | Type |
|---|---|---|---|
| Speed | Vehicle speed | Kilometers per hour | Numeric |
| Distance travelled | Distance driving | Meters | Numeric |
| Speed Limit | Current speed limit | Kilometers per hour | Numeric |

The permutation feature importance technique calculates the prediction error by permuting the feature value. This approach severs the connection between the feature and the objective, allowing one to discern the model's dependence on the feature by evaluating its prediction error after the feature's value has been permuted (Molnar et al., 2021). An added benefit of Permutation Feature Importance is its time-saving aspect, as it eliminates the need for model retraining, potentially saving a significant amount of time.

Moreover, this method offers another advantage by taking into account all interactions with other attributes.

### *3.1 Simulator Experiment*

The simulator experiment was carried out with the contribution of 36 drivers and was based on principles that have been comprehensively documented in the literature (Fisher et al., 2011; Tipton et al., 2014). The experiment was implemented based on three scenarios as shown in the Table 2.

*Table 2: Different scenarios applied during the driving simulator experiment.*

| Scenario | Road Section | Number of lanes | Speed Limits |
|---|---|---|---|
| A | 0-6300 m | 1x1 | 70 km/h |
| | 6300-11300 m | 2x2 | 90 km/h |
| | 11300-16500 m | 2x2 | 120 km/h |
| B | 0-6100 m | 2x2 | 90 km/h |
| | 6100-12000 m | 2x2 | 120 km/h |
| | 12000-18200 m | 1x1 | 70 km/h |
| C | 0-6000 m | 2x2 | 120 km/h |
| | 6000-11000 m | 2x2 | 90 km/h |
| | 11000-17200 m | 1x1 | 70 km/h |

Each participant performed three separate drives.

- Drive 1: No interventions
- Drive 2: Interventions
- Drive 3: Interventions with modifying condition

### *3.2 Naturalistic Driving Experiment*

The design and implementation of the on-road experiment was conducted following certain principles from the existing literature focusing on testing interventions to assist drivers in operating within safe boundaries. The ND experiment was divided into four phases and focused on monitoring driving behavior and the impact of real-time interventions (e.g., in-vehicle warnings) and post-trip interventions (e.g., post-trip feedback & gamification) on driving behavior. The description of the four phases as well as the drivers and trips that were collected are outlined in the following Table 3:

*Table 3: Description of each Phase.*

| Phases | Description | Drivers | Trips |
|---|---|---|---|
| Phase 1 | Monitoring (baseline measurement; no interventions) | 39 | 1,173 trips (23,725 minutes) |
| Phase 2 | In-vehicle intervention | 43 | 1,549 trips (31,414 minutes) |
| Phase 3 | Post-trip feedback on the smartphone | 51 | 1,973 trips (40,121 minutes) |
| Phase 4 | Post-trip feedback on smartphone + gamified web platform | 49 | 2,468 trips (52,077 minutes) |

## 4. Methods

### *4.1 Classification Algorithms*

According to the literature review, four classification models were applied to achieve the objective of this research, namely (i) Support Vector Machines, (ii) Random Forest, (iii) AdaBoost, and (iv) Multilayer Perceptron.

#### 4.1.1 Support Vector Machines (SVM)

SVMs are supervised machine-learning models used for data analysis, and pattern detection and apply to both classification and regression problems (Roy et al., 2015). The context of the SVM model is to develop a hyper-plane in a multidimensional space to separate different class boundaries (Ghosh et al., 2019). The key advantage of SVMs is that they can handle high-dimensional datasets (Xia, 2020). Utilizing the hyperparameter tuning technique called GridSearchCV from the scikit-learn Python library, the optimal values for SVMs hyperparameters were emerged as: (a) kernel type = 'rbf'; (b) regularization parameter C = 50; and (c) kernel coefficient gamma = 'scale'.

#### 4.1.2 Random Forest (RF)

The RF classifier is an ensemble approach that trains several decision trees in parallel employing bootstrapping and aggregation, often known as the bagging technique (Misra & Li, 2020). The bootstrapping technique concerns simultaneously training multiple decision trees using different subsets of the dataset. By aggregating the outcomes of these individual decision trees, the final decision is reached. Additionally, RF offers the advantage of overcoming the common overfitting problem associated with decision trees (Shangguan et al., 2021), making it a preferred choice for identifying risky driving behavior. In this case, Grid Search was also applied and the optimal hyperparameters were: (a) the number of estimators/trees of the forest = 200 and (b) the function to measure the quality of a split (criterion) = ‘entropy’.

#### 4.1.3 AdaBoost

The AdaBoost algorithm is extensively used due to its high speed, low complexity, and good compatibility (Liu, 2021). AdaBoost represents an ensemble technique that trains and deploys sequential trees using the boosting methodology, which involves linking a series of weak classifiers, each of which aims to improve the classification of samples previously misclassified by the previous weak classifier (Misra & Li, 2020). This approach effectively combines these weak classifiers into a series to produce a strong classifier. The ideal maximum number of estimators was determined to be 500 using GridSearchCV.

#### 4.1.4 Multilayer Perceptron (MLP)

The MLP is a feed-forward neural network complement and consists of three types of layers: (i) the input layer, (ii) the output layer, and (iii) the hidden layer (Abirami & Chitra, 2020). The main advantage of the MLP algorithm is its ability to handle non-linear problems with large datasets while providing quick predictions. Applying the Grid Search method for MLP resulted in the following six optimal hyperparameters: (a) number of hidden layers = (500, 500, 500,), (b) activation function = "relu" and (c) alpha parameter of the regularization term = 0.0001.

### *4.2 Multiclass Classification*

The three-level classification of driving behavior (i.e., "Normal", "Dangerous" and "Avoidable Accident") is a multi-classification problem. In order to assess the effectiveness of classification algorithms, the dataset is initially segmented into training

and testing datasets. The training dataset is structured as $X_{training} = \{(x_n, y_n), n = 1, N\}$, with $x_n$ representing predictor variables and $y_n$ taking values from the set {0, 1, 2} as the target variable. Through model training, it gains the capacity to accurately classify new data instances. The classification model's performance can easily be demonstrated with a confusion matrix, where one axis represents the actual class and the other denotes the predicted class. The results showcased in this paper were achieved by employing 10-fold cross-validation. The metrics utilized to evaluate the models are accuracy, precision, recall, f1-score, and false alarm rate defined by Equation (1) to Equation (5)

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{f1-score} = \frac{2x\,(Precision)x\,(Recall)}{(Precision)+(Recall)} \quad (4)$$

$$\text{False Alarm Rate} = \frac{FP}{FP+TN} \quad (5)$$

where True Positives (TP) denote instances from class i that were classified correctly within it. True Negatives (TN) represent instances not belonging to class i, correctly excluded from it. False Positives (FP) indicate instances not belonging to class i but incorrectly classified within it. False Negatives (FN) signify instances from class i that were erroneously not classified within it.

## 5. Results

This study aimed to comprehensively assess the performance of four machine learning classifiers (i.e., SVM, RF, AdaBoost, and MLP) across two distinct datasets (i.e., Simulator experiment dataset and Naturalistic Driving experiment dataset). Due to the phenomenon of "accuracy paradox" (Valverde-Albacete & Peláez-Moreno, 2014) the evaluation was conducted based on several metrics, such as accuracy, precision, recall, false alarm rate, and F1-score, as otherwise the evaluation of accuracy alone would be misleading.

Due to the fact that risky driving is less common than normal driving and since the classification algorithms operate on the assumption of equal distribution of samples, the Adaptive Synthetic (ADASYN) (He et al., 2008) technique was applied to address the imbalanced problem.

### *5.1 Classification Models on Simulator Experiment*

Considering Figure 1 and Table 4, overall, the four algorithms had insightful and satisfactory results in terms of accuracy and recall. Among the different algorithms, RF stands out with the highest accuracy of 84.00%, indicating its ability to accurately classify driving behaviors in a controlled environment. RF also achieves a well-balanced precision

(59.41%) and recall (70.27%), demonstrating its robustness and versatility. The MLP model also performs admirably with an accuracy of 81.28%, highlighting its capability in this simulation framework, balancing precision (57.51%) and recall (72.04%) effectively, achieving a competitive f1-score (61.79%).

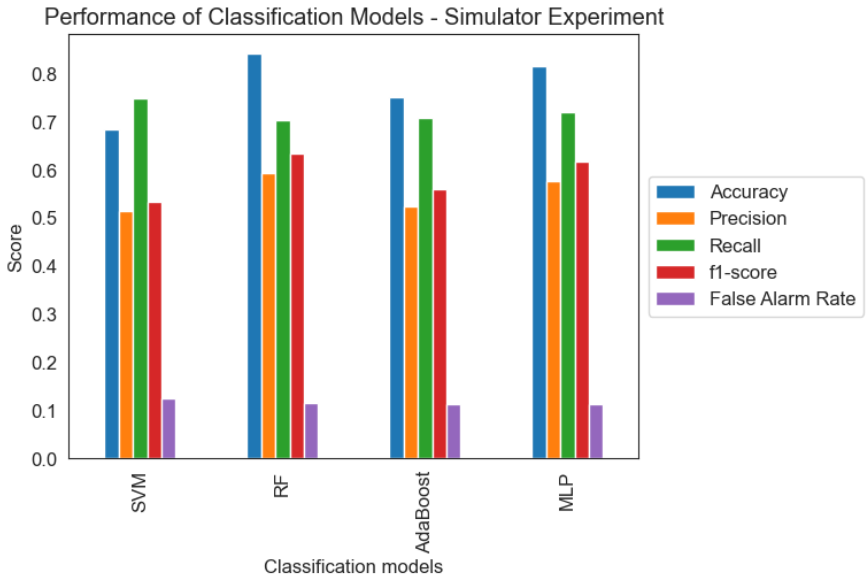


*Figure 1: Classification metrics of the four machine learning models*

*Table 4: Classification metrics for the Simulator Experiment dataset*

| Classifier | Accuracy | Precision | Recall | False Alarm Rate | f1-score |
|---|---|---|---|---|---|
| SVM | 68.67 % | 51.35 % | 74.72 % | 12.47 % | 53.22 % |
| RF | 84.00 % | 59.41 % | 70.27 % | 11.47 % | 63.42 % |
| AdaBoost | 75.08 % | 52.31 % | 70.71 % | 11.30 % | 55.87 % |
| MLP | 81.28 % | 57.51 % | 72.04 % | 11.37 % | 61.79 % |

Furthermore, the AdaBoost model achieves reasonable accuracy (75.08%) but has lower precision (52.31%) and recall (70.71%) compared to RF and MLP. While the SVM shows a strong recall of 74.72%, indicating its ability to effectively capture true positive instances, it comes at the cost of lower precision (51.35%), resulting in a trade-off between recall and precision.

### *5.2 Classification Models on Naturalistic Driving Dataset*

The results of the naturalistic driving experiment were similar to those of the simulation experiment. RF achieved an adequate accuracy of 75.00% and a balanced precision (56.77%) and recall (66.28%) demonstrating a robust performance in classifying real-world driving behavior. In the Naturalistic Driving Dataset, MLP maintains its strong performance with an accuracy of 73.26% but faces challenges with lower accuracy (52.14%) and recall (56.57%) which is reflected in the f1-score (52.65%).

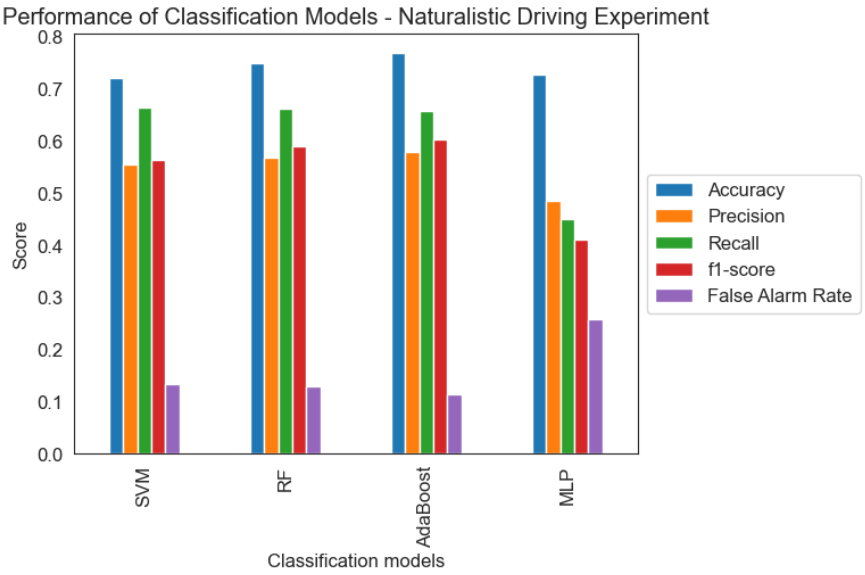


*Figure 2: Classification metrics of the four machine learning models*

*Table 5: Classification metrics for the Naturalistic Driving Experiment dataset*

| Classifier | Accuracy | Precision | Recall | False Alarm Rate | f1-score |
|---|---|---|---|---|---|
| SVM | 72.05 % | 55.51 % | 66.31 % | 13.39 % | 56.37 % |
| RF | 75.00 % | 56.77 % | 66.28 % | 12.97 % | 59.03 % |
| AdaBoost | 76.76 % | 57.91 % | 65.81 % | 11.47 % | 60.19 % |
| MLP | 73.26 % | 52.14 % | 56.57 % | 16.66 % | 52.65 % |

AdaBoost, scored the highest accuracy (76.76%) maintaining a competitive performance consistent with the simulator data, with a balanced precision (57.91%), and recall (65.81%) achieving the highest f1-score (60.19%). Finally, SVM maintains its proficiency in recall (66.31%), showing consistency in capturing true positives. However, similar to the Simulator Experiment, this is accompanied by lower precision (55.51%). Also, SVM achieves an accuracy of 72.05%, which is relatively competitive but falls behind compared to the other models.

## 6. Discussion

Overall, the findings of this study provided valuable insights while supporting its objective, which was the investigation of various classification models utilizing two distinct data sources. These findings are essential for advancing the understanding of driving behavior across various contexts, ultimately contributing to the development of safer and more efficient transportation systems.

The evaluation of the four machine learning classifiers (SVM, RF, AdaBoost, and MLP) revealed varying performance across the two datasets. In the simulator experiment, RF emerged as the top-performing model with an accuracy of 84%, demonstrating its ability to accurately classify driving behavior in a controlled environment. Following the MLP

model which also performed well scoring a notable 81.28% accuracy. Regarding, AdaBoost and SVM models, they underperformed compared to the other two underperformed compared to the other two, displaying a lower weighted accuracy and recall. In the naturalistic driving dataset, RF and AdaBoost maintained robust performance, with high accuracy (i.e., 75% and 76.76% respectively) and balanced precision and recall. Furthermore, MLP while still effective, faced challenges with lower accuracy (73.26%) and recall (56.57%) compared to the simulator experiment. Finally, SVM, although competitive, lagged behind other models. These performance variations underscore the importance of selecting the right model based on data characteristics and precision-recall trade-offs, essential for real-world applications. Since, in the context of the current study, it is more dangerous to misidentify driving behavior as less dangerous, the recall metric is the most significant metric to consider. Thus, evaluating the results of both approaches (i.e., the Driving Simulator experiment and the Naturalistic Driving experiment), the RF model emerged as the most efficient one.

Based on comparable driving behavior studies, the findings of this study were very similar to those described in the literature. For instance, J. Yang et al. (2023) achieved an 80% accuracy, which is relatively close to the accuracy of the two approaches (84% and 75%), as well as better performance in terms of False Alarm Rate. However, in terms of recall the RF model of this research underperforms by 13% (for the simulator experiment) and 17% (for the naturalistic driving experiment). In another study by Song et al. (2021), the RF classifier exhibited a remarkable 90% accuracy, surpassing the performance in this study. This discrepancy may be attributed to differences in input variables, as this study focused on driving behavior characteristics while Song et al. (2021) considered variables such as gender, age, and driver perception. In contrast to the outcomes of this research, findings from the literature regarding the SVM classifier showed higher performance, especially with K. Yang et al. (2021) having an outstanding accuracy rate of 95%. Additionally, in contrast to the research of (Shangguan et al., 2021), this study's accuracy metric findings for the MLP classifier were identical. Nonetheless, the MLP classifier that was developed in previous literature exhibited better performance than the one employed in this study, with a notable 20% difference in the f1-score between them. Finally, regarding the AdaBoost model, it showed promising findings for real-world data. Since its application is limited in the literature, to the author’s knowledge, in the field of road safety it offers a robust approach.

In conclusion, the findings of this study not only contribute to a better understanding of driving behavior in various circumstances, but they also show the crucial importance of model selection and data features in establishing accurate classifications. The findings highlight the RF model's effectiveness, particularly in controlled environments, while also shining light on AdaBoost's potential for real-world driving data analysis.

## 7. Conclusions

The research aimed to develop and evaluate four classification models on two distinct data sources (i.e., Simulator Experiment and Naturalistic Driving Experiment). This methodological approach has facilitated a comprehensive evaluation of machine learning

models within controlled and real-life driving contexts. Consequently, this study has significantly contributed to advancing the understanding of driver behavior across diverse scenarios (i.e., controlled, and real-world) as well as the ability of machine learning models to effectively capture driving behavior, as well as the performance of various models in the two distinct experiments. RF model emerged as a strong performer, offering a balanced approach between precision and recall in both simulated and real-world driving scenarios. Given that misidentifying dangerous driving behavior as less dangerous would have serious implications for road safety, recall is a key metric with SVMs outperforming in capturing true positive instances in both datasets.

The findings of this study offer valuable guidance to researchers and practitioners in model selection for driving behavior classification tasks. Considering the dual-source methodology, drivers' risky behavior can be assessed by comparing both simulator and field-trials experiment data, highlighting key road safety factors.

Future research could examine the usefulness of deep learning (DL) techniques on this matter, such as Long Short-Term Memory (LSTM) (Banan et al., 2020; Chen et al., 2022). DL models are increasingly utilized due to their ability to capture complex temporal dependencies of features, thus potentially improving the accuracy and predictive capabilities of driver behavior classification models. Furthermore, the examination of additional data sources such as Naturalistic Driving experiment dataset involving drivers from different countries or transport modes, would assist in the comprehensive understanding and evaluation of the models utilized.

**Abbreviations**

WHO: World Health Organization
DSS: Driving Simulator Studies
NDS: Naturalistic Driving Simulator
SVM: Support Vector Machines
RF: Random Forest
MLP: Multilayer Perceptron
TP: True Positives
TN: True Negatives
FP: False Positives
FN: False Negatives

**Acknowledgments**

The research was funded by the European Union's Horizon 2020 i-DREAMS project (Project Number: 814761) funded by European Commission under the MG-2-1-2018 Research and Innovation Action (RIA).

The IVORY Industrial Doctoral Network (101119590 IVORY HORIZON-MSCA-2022-DN-01) is funded by the European Commission for the Doctoral Networks programme under the Horizon Europe (HORIZON) Marie Skłodowska-Curie Actions. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those

of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

**Availability of data and materials**
Not applicable.

**References**


Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. In *Advances in Computers* (Vol. 117, Issue 1, pp. 339–368). Academic Press Inc. https://doi.org/10.1016/bs.adcom.2019.09.007

Banan, A., Nasiri, A., & Taheri-Garavand, A. (2020). Deep learning-based appearance features extraction for automated carp species identification. *Aquacultural Engineering*, *89*, 102053. https://doi.org/https://doi.org/10.1016/j.aquaeng.2020.102053

Chen, W., Sharifrazi, D., Liang, G., Band, S. S., Chau, K. W., & Mosavi, A. (2022). Accurate discharge coefficient prediction of streamlined weirs by coupling linear regression and deep convolutional gated recurrent unit. *Engineering Applications of Computational Fluid Mechanics*, *16*(1), 965–976. https://doi.org/10.1080/19942060.2022.2053786

Fisher, D., Caird, J., & Rizzo, M. (2011). Handbook of Driving Simulation for Engineering, Medicine and Psychology. In *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. https://doi.org/10.1201/b10836-2

Garefalakis, T., Katrakazas, C., & Yannis, G. (2022). Data-Driven Estimation of a Driving Safety Tolerance Zone Using Imbalanced Machine Learning. *Sensors*, *22*(14). https://doi.org/10.3390/s22145309

Ghandour, R., Potams, A. J., Boulkaibet, I., Neji, B., & Al Barakeh, Z. (2021). Driver Behavior Classification System Analysis Using Machine Learning Methods. *Applied Sciences*, *11*(22). https://doi.org/10.3390/app112210562

Ghosh, S., Dasgupta, A., & Swetapadma, A. (2019). A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification. *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, 24–28. https://doi.org/10.1109/ISS1.2019.8908018

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Khoury, M. A., & Hussein, F. A. (2023). Efficiency and Safety: The Impact of Autonomous Controls on Transportation. *International Journal of Information and Cybersecurity*, *7*(1), 13–39. https://publications.dlpress.org/index.php/ijic/article/view/11

Liu, H. (2021). Data mining and processing for train unmanned driving systems. In *Unmanned Driving Systems for Smart Trains* (pp. 211–252). Elsevier. https://doi.org/10.1016/B978-0-12-822830-2.00005-2

Michelaraki, E., Katrakazas, C., Brijs, T., & Yannis, G. (2021, September). Modelling the Safety Tolerance Zone: Recommendations from the i-DREAMS project. *10th International Congress on Transportation Research*.

Misra, S., & Li, H. (2020). Chapter 9 - Noninvasive fracture characterization based on the classification of sonic wave travel times. In S. Misra, H. Li, & J. He (Eds.), *Machine Learning for Subsurface Characterization* (pp. 243–287). Gulf Professional Publishing. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., & Bischl, B. (2021). *Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process*. arXiv. https://doi.org/10.48550/ARXIV.2109.01433

Nasr Azadani, M., & Boukerche, A. (2022). Driving Behavior Analysis Guidelines for Intelligent Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 6027–6045. https://doi.org/10.1109/TITS.2021.3076140

Osman, O. A., Hajij, M., Karbalaieali, S., & Ishak, S. (2019). A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accident Analysis & Prevention*, *123*, 274–281. https://doi.org/10.1016/j.aap.2018.12.005

Papadimitriou, E., Argyropoulou, A., Tselentis, D. I., & Yannis, G. (2019). Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving. *Safety Science*, *119*, 91–97. https://doi.org/10.1016/j.ssci.2019.05.059

Roy, K., Kar, S., & Das, R. N. (2015). Selected Statistical Methods in QSAR. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (pp. 191–229). Elsevier. https://doi.org/10.1016/B978-0-12-801505-6.00006-5

Shangguan, Q., Fu, T., Wang, J., Luo, T., & Fang, S. (2021). An integrated methodology for real-time driving risk status prediction using naturalistic driving data. *Accident Analysis & Prevention*, *156*, 106122. https://doi.org/10.1016/j.aap.2021.106122

Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis & Prevention*, *129*, 170–179. https://doi.org/10.1016/j.aap.2019.05.005

Song, X., Yin, Y., Cao, H., Zhao, S., Li, M., & Yi, B. (2021). The mediating effect of driver characteristics on risky driving behaviors moderated by gender, and the classification model of driver's driving risk. *Accident Analysis & Prevention*, *153*, 106038. https://doi.org/10.1016/j.aap.2021.106038

Staubach, M. (2009). Factors correlated with traffic accidents as a basis for evaluating Advanced Driver Assistance Systems. *Accident Analysis & Prevention*, *41*(5), 1025–1033. https://doi.org/10.1016/j.aap.2009.06.014

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample Selection in Randomized Experiments: A New Method Using Propensity Score Stratified Sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114–135. https://doi.org/10.1080/19345747.2013.831154

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PloS One*, *9*(1), e84217–e84217. https://doi.org/10.1371/journal.pone.0084217

van Schagen, I., & Sagberg, F. (2012). The Potential Benefits of Naturalistic Driving for Road Safety Research: Theoretical and Empirical Considerations and Challenges for the Future. *Procedia - Social and Behavioral Sciences*, *48*, 692–701. https://doi.org/10.1016/j.sbspro.2012.06.1047

Wang, X., Xu, R., Zhang, S., Zhuang, Y., & Wang, Y. (2022). Driver distraction detection based on vehicle dynamics using naturalistic driving data. *Transportation Research Part C: Emerging Technologies*, *136*, 103561. https://doi.org/10.1016/j.trc.2022.103561

Wijayaratna, K. P., Cunningham, M. L., Regan, M. A., Jian, S., Chand, S., & Dixit, V. V. (2019). Mobile phone conversation distraction: Understanding differences in impact between simulator and naturalistic driving studies. *Accident Analysis & Prevention*, *129*, 108–118. https://doi.org/10.1016/j.aap.2019.04.017

World Health Organization. (2018). *Global Status Report On Road Safety 2018*. World Health Organization. https://www.who.int/publications/i/item/9789241565684

Xia, Y. (2020). Chapter Eleven - Correlation and association analyses in microbiome study integrating multiomics in health and disease. In J. Sun (Ed.), *Progress in Molecular Biology and Translational Science* (Vol. 171, pp. 309–491). Academic Press. https://doi.org/https://doi.org/10.1016/bs.pmbts.2020.04.003

Yang, J., Han, S., & Chen, Y. (2023). Prediction of Traffic Accident Severity Based on Random Forest. *Journal of Advanced Transportation*, *2023*, 1–8. https://doi.org/10.1155/2023/7641472

Yang, K., Haddad, C. Al, Yannis, G., & Antoniou, C. (2021). Driving Behavior Safety Levels: Classification and Evaluation. *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 1–6. https://doi.org/10.1109/MT-ITS49943.2021.9529309

Zhang, C., Patel, M., Buthpitiya, S., Lyons, K., Harrison, B., & Abowd, G. D. (2016). Driver Classification Based on Driving Behaviors. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 80–84. https://doi.org/10.1145/2856767.2856806