Garefalakis *et al.*
European Transport Research Review (2024) 16:65
https://doi.org/10.1186/s12544-024-00691-9

European Transport Research Review

ORIGINAL PAPER

Open Access

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

Thodoris Garefalakis[1*], Eva Michelaraki[1], Stella Roussou[1], Christos Katrakazas[1], Tom Brijs[2] and George Yannis[1]

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. 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, data were collected through a driving simulator as well as a naturalistic driving study. To that end, four classification algorithms, namely support vector machines, random forest (RFs), AdaBoost, and multi-layer perceptron (MLP) neural networks were implemented and compared. In the simulator experiment, RFs and MLPs emerged as the top-performing models with an accuracy of 84% and 82%, respectively, demonstrating its ability to accurately classify driving behavior in a controlled environment. In the naturalistic driving study, RF and AdaBoost maintained robust performance, with high accuracy (i.e., 75% and 76.76% respectively) and balanced precision and recall. The outcomes of this study could provide essential guidance for practitioners and researchers on choosing models for driving behavior classification tasks.

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), 1.19 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 and 29 years old [1]. 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.

1.1 Main motivation

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

*Correspondence:
Thodoris Garefalakis
tgarefalakis@mail.ntua.gr
[1] Department of Transportation Planning and Engineering, National Technical University of Athens, 5 Heroon Polytechniou Str., 15773 Athens, Greece
[2] School for Transportation Sciences, Transportation Research Institute (IMOB), UHasselt, Agoralaan, 3590 Diepenbeek, Belgium

contributor to road crashes [2]. 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 [3].

The research motivation for this study is to explore ways to predict and analyze the dangerous driving behavior using simulator and naturalistic driving data. An analysis framework based on statistical and machine learning models, verified the significance of risky driving behavior classifications for a crash prediction model, was developed.

1.2 Innovative aspects

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) [4]. 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 level: Conditions suggest a low likelihood of a crash, with the driver operating safely within acceptable boundaries.
- Dangerous level: Conditions indicate an increased risk of a crash. Although a crash is not imminent, the likelihood has increased, requiring heightened awareness and potentially corrective actions from the driver.
- Avoidable accident level: Conditions are critical, with a high probability of collision unless immediate evasive action is taken by the driver. The need for action in this phase is urgent.

Nevertheless, it is important to note that headway levels, the Dangerous level, and the Avoidable accident level are also closely related to speeding. In our analysis, we found that many headway events occur due to speeding, where drivers maintain insufficient headway at higher speeds. This significantly increases the risk because the time available to react is reduced, and the stopping distance required in an emergency is greater.

Risky driving behavior is intrinsically connected to the STZ concept, which categorizes driving scenarios based on the probability of a collision. Within this framework:

- Normal driving behavior corresponds to the Normal Level of STZ, where the risk is low, and the driver maintains safe driving practices.
- Risky driving behavior encompasses the Dangerous Level and the Avoidable Accident Level of STZ. At the Dangerous Level, the driver's actions increase the likelihood of a crash, necessitating corrective measures. At the Avoidable Accident Level, immediate evasive actions are required to prevent an imminent collision.

The challenge lies in accurately defining and identifying risky driving behavior. The current literature on Collision Warning Systems includes variables of interest such as acceleration/deceleration, Time Headway (TH) and Time to Collision (TTC) along with respective thresholds. Based on the conceptualization of this work, a number of indicators were considered for the definition of the STZ levels.

1.3 Contributions/objectives

Based on the above framework, the aim of this paper is to develop and evaluate different classification models in order to predict risky driving behavior, 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.

To achieve this object, data were collected through a driving simulator and a naturalistic driving study and four classification algorithms, namely support vector machines (SVMs), random forest (RFs), AdaBoost and multilayer perceptron (MLP) neural networks were implemented.

The main contributions of this paper are as follows:

1. Based on the collected driving behavior data, the simultaneous application of both types of simulator and naturalistic driving data has been achieved. The introduction of new data sources enriches the model input features and compensates for the lack of dynamic data in traditional road safety research, thus effectively improving the accuracy of crash prediction.

2. Headway indicators were used to classify risky driving behavior. The contribution of the degree of different levels of headway to the risky driving behavior is shown and provides a reference for traffic management departments to develop effective traffic safety measures and reduce the likelihood of crashes.

1.4 Various parts of the manuscript

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 [5]. 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 [6] 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 instance, DSS presents a valuable opportunity for collecting a wide range of driving scenario data efficiently in well-controlled environments [7]. On the other hand, NDS has a higher degree of realism reflecting more accurately the natural driving situation [8].

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 [9–15] utilized models such as random forest (RFs), multilayer perceptron (MLP), support vector machines (SVMs), eXtreme gradient boosting (XGBoost), decision trees (DT), gradient boosting (GB) and logistic regression (LR).

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. [11] devised a framework encompassing feature extraction, clustering techniques, feature importance analysis, and the utilization of machine learning algorithms including RF, XGBoost, SVM, and MLP. This framework achieved an accuracy exceeding 85% in predicting driving risk statuses, which are defined as safe, low-risk, medium-risk, and high-risk based on factors like speed variations, headway distance, speed, and acceleration. Similarly, Yang et al. [14] exploited a driving simulator dataset, to develop a framework for classifying driving behaviors into different safety levels. They applied clustering techniques to identify three distinct levels of driving behavior: normal, low-risk, and high-risk. The study defined risky driving behaviors as actions like harsh acceleration and harsh braking. By using classification algorithms such as SVM, Decision Tree, and Naive Bayes, they achieved the highest accuracy of 95% with the Decision Tree model in evaluating these safety levels. Additionally, Shi et al. [12] developed a framework integrating feature selection, risk level labeling, and addressing imbalanced datasets. They defined risky driving behaviors using 1300 features related to speed, headway distance, and acceleration, achieving an overall accuracy of 89% with the XGBoost model.

Furthermore, Zhang et al. [15] 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. [10] quantified the correlation between dangerous driving and mobile phone usage through logistic regression, with a marked accuracy of 70%. Lastly, Ghandour et al. [9] 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.

Previous studies have often focused on either simulated or naturalistic environments, but a holistic understanding of driving safety lacks. Simulated scenarios provide controlled conditions that allow for targeted analysis, while naturalistic studies capture the complexity and variability of real-world driving. By combining these approaches, can overcome the limitations associated with singular methodologies, offering a more nuanced and valid assessment of driving behavior safety.

Based on the gaps in the literature, the adopted approach recognizes the variation in driving behavior across simulated and real-world conditions. By exploiting the advantages of both methodologies, a more robust and generalizable classification model is pursued. Through the dual approach (i.e., driving simulator study and naturalistic driving study), a holistic overview of the topic is pursued.

3 Methods and materials

3.1 Driving experiment and data collection

For the purpose of the study, a simulator experiment and a naturalistic driving study 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 study). 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 [16].

Within the framework of the simulator experiment, and to determine the three safety levels (i.e., the target variable of the classification process), various indicators such as speed, time to collision (TTC), and time headway (TH) were considered. Click or tap here to enter text. However, only the TH-based categorization of STZ consistently aligned with relevant studies, where dangerous behavior is a rarer phenomenon compared to safe driving behavior. The range of values for the headway corresponding to each safety level is:

- 'Normal' Level: Headway ≥ 2 s
- 'Dangerous' Level: Headway ≥ 1.4 s and Headway < 2 s
- 'Avoidable Accident' Level: Headway < 1.4 s

Headway is a vital parameter in analyzing traffic flow and safety, representing the time or distance between vehicles, which directly influences the likelihood of rear-end collisions [17].

Moreover, modern automotive industries are increasingly incorporating collision warning systems that rely heavily on Time Headway (TH) measurements. These systems utilize sensors to continuously monitor the distance to the vehicle ahead and calculate the TH. If the TH drops below a predefined threshold (typically two seconds), the system alerts the driver to increase the following distance. This threshold is based on extensive research indicating that shorter headways significantly increase the risk of rear-end collisions due to insufficient reaction time. Previous studies have shown that a time headway ranging from 1.1 to 1.7 s is considered a manageable margin [18]. However, numerous driver training programs advocate that maintaining at least a 2-s distance from the vehicle ahead is essential for safe following and preventing collisions, commonly referred to as the "2-s rule" [19].

Following the initial definition of the STZ, the analysis aimed to investigate key risk factors influencing driving behavior, with a focus on variables that play a crucial role in assessing road safety. The initial set of variables included average speed, harsh acceleration, harsh braking, headway, overtaking, distraction (i.e., mobile phone use), and fatigue. However, to improve the performance of the models, feature selection process was employed. Specifically, through the permutation feature importance technique the significance of each variable in predicting driving behavior was evaluated. 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 [20]. 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.

Following this approach, three variables emerged as particularly influential, demonstrating a substantial impact on the model's predictive capabilities. These selected variables were chosen for further analysis to ensure consistency in both approaches. Table 1 provides a detailed description of the chosen variables:

The chosen variables allow for a meaningful exploration of driving behavior, minimizing unnecessary complexity and ensuring a focused and effective investigation.

3.1.1 Simulator driving 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 [21, 22]. These principles, encompass defining outcomes, predictors, and hypotheses, determining sample size and statistical power, choosing the design type, allocating risk scenarios among participants, deciding on drive durations to prevent simulator sickness, preventing order and learning effects, and accounting for confounding factors.

In context of the simulator study, 36 drivers were participated with an average age of 42 years of whom 70%

Table 1 Description of the selected features emerged through permutation importance

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

were men and 30% were women. From the design perspective, the scenarios had minimal rapid changes in direction and acceleration. The total duration of the simulator did not exceed 2 h, and the duration of each drive did not exceed 1 h. Drives with more demanding scenarios were kept shorter than regular drives. Although there were no set rules for the duration of the drives, the general practice was to set them between 5 and 25 min with 10-min breaks in between. It had been shown that simulator sickness increased with the drive duration in one trial but decreased with successive trials in multiple sessions [23]. As such, a few practice drives prior to the main drive were designed to help reduce the effects of simulator sickness. However, these practice drives could have resulted in adaptation (or learning effects), which is a type of contamination that may have influenced the results. Overall, the simulator's high fidelity to real-world dynamics, encompassing precise auditory, visual, and motion cues, mitigated simulator sickness by aligning participants' sensory impressions with their expectations in a realistic driving context. Based on the above, it was possible that during and after the pilot driving, the driver experienced mild or intense discomfort, dizziness or nausea. In such cases, the experiment trials were stopped if symptoms of simulator sickness were apparent or if the participant reported feeling unwell.

Eligibility criteria included being between 20 and 65 years old, no history of motion sickness and no use of medication that could impair driving performance. Participants were given detailed instructions before the experiment, which included a demonstration of the driving simulator and a practice session. The instructions emphasized the importance of driving as naturally as possible and following the simulated traffic rules.

The experiment was based on the DriveSimSolutions (DSS) driving simulator which was developed for the purposes of the i-DREAMS project and conducted from December 2020 to January 2021. The simulator was design based on a Peugeot 206 based on the Peugeot 206 model, incorporating various genuine components such as the complete dashboard, operational instrument panel, and the driver's seat to accurately emulate the cockpit of this particular vehicle. The simulator operates on the STISIM Drive 3 software, showcased on three 49-inch screens with 4K resolution, delivering a 135° field of view. The experiment was implemented based on three scenarios as shown in the Table 2.

Each participant performed three separate drives.

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

Table 2 Characteristics of the scenarios implemented during the driving simulator experiment

Scenario	Road section (m)	Number of lanes	Speed limits (km/h)
A	0–6300	1×1	70
	6300–11,300	2×2	90
	11,300–16,500	2×2	120
B	0–6100	2×2	90
	6100–12,000	2×2	120
	12,000–18,200	1×1	70
C	0–6000	2×2	120
	6000–11,000	2×2	90
	11,000–17,200	1×1	70

3.1.2 Naturalistic driving study

The design and implementation of the on-road study was conducted following certain principles from the existing literature focusing on testing interventions to assist drivers in operating within safe boundaries. Data collection for the ND study was conducted 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. Although the NDS data collection was divided into four phases, the combined data (of all phases) were utilized for the classification process. The description of the four phases as well as the drivers and trips that were collected are outlined in the following Table 3:

In order to gather a range of vehicle and driver-related driving attributes, devices such as Mobileye system [24], a dash camera and the Cardio gateway (*CardioID* [25]) were used in order to record driving behavior (e.g., speed, acceleration, deceleration, steering). More specifically, the Mobileye system is as a network sensor and a camera-based system mounted on the windshield that measures parameters, like headway monitoring, traffic sign recognition, lane position monitoring and pedestrian recognition. The system can be connected to the CAN bus and enables the integration with several ADAS ecosystem products.

The Cardio gateway is a system based on sensors which is connected to the Mobileye equipment through the CAN bus of the vehicle and can transfer data through different communication technologies (BLE, CAN, I2C, SPI, WiFi). CardioID also provided a web API to support data access. The API completely following the REST architectural style, and the data were available in JSON format. In addition, Cardio Watch was also used to provide more reliable data about

Table 3 Description of each phase of the naturalistic driving study

Phases	Description	Drivers	Trips
Phase 1	Monitoring (baseline measurement; no interventions)	39	1173 trips (23,725 min)
Phase 2	In-vehicle intervention	43	1549 trips (31,414 min)
Phase 3	Post-trip feedback on the smartphone	51	1973 trips (40,121 min)
Phase 4	Post-trip feedback on smartphone + gamified web platform	49	2468 trips (52,077 min)

car drivers' fatigue and sleepiness compared to Cardio Wheel which required both hands on the steering wheel to provide fatigue index data.

Lastly, OSeven provided a state-of-the-art android-based smartphone application that also monitors and collect driving behavior of individuals using a variety of parameters. The app uses different smartphone sensors to collect such data. The app was used by drivers recruited for on-field trials. Drivers recruited for the field-trials were required to install this app on their smartphone. A standard procedure was followed every time a new trip is retrieved by the application: the application collects in real-time the data from the sensors of the mobile phone and then data processing takes place. All the variables in the analyzed data were derived from a combination of machine learning methods (data fusion, clustering and classification). Since OSeven has strict data sharing policies, further information cannot be provided at the moment. Nevertheless, additional details for data extraction regarding the OSeven application can be found in Papadimitriou et al. [10].

- Phases: while Phase 1 established a baseline by monitoring driving behavior post-i-DREAMS system installation, Phase 2 introduced in-vehicle real-time warnings, adapting to drivers' behaviors identified during Phase 1. Consequently, Phase 3 integrated post-trip feedback via the i-DREAMS smartphone app alongside in-vehicle warnings, creating a more comprehensive intervention framework. Finally, Phase 4 expanded upon the feedback mechanism by incorporating gamification features. Additionally, providing web-dashboard support further enhanced participant engagement and intervention effectiveness throughout the trial.
- Drivers: refers to the total number of individuals participating as drivers in each phase of the study. Each driver contributes to the dataset through their trips and interactions with the interventions.
- Trips: indicates the total count of trips recorded during each phase of the study. A trip is defined as a single journey made by a driver from one location to another. Also, the cumulative duration of all trips combined during each phase, measured in minutes

It should be noted that in this study, a simulated environment was utilized to closely approximate the conditions and driving routes observed in the ND study. In terms of road scene design, the simulator environment was carefully constructed to replicate key features of the ND study routes, including similar roadway types such as highways and urban streets, along with comparable environmental elements like road signage, lane markings, and intersections. For the on-road testing, a selection of real-world routes was chosen to mirror those observed in the ND study, based on factors such as roadway type, typical traffic density, and the complexity of the driving environment.

In addition, traffic configuration was designed to align with the conditions observed in the ND study. In both simulated and on-road settings, traffic scenarios were created with similar volume and speed, as well as the presence of common driving challenges, such as merging, lane changes, and intersections, to reflect the real-world conditions experienced by drivers. Traffic in the simulator was programmed dynamically, with variability in the speed and behavior of surrounding vehicles to mimic the unpredictable nature of real-world traffic.

To ensure alignment with the ND study routes, specific on-road testing routes were chosen for their resemblance to those used in the ND study. This selection involved analyzing key route metrics from the ND study, such as average speed, traffic density, and road type distribution, to identify routes that best captured similar driving experiences. Lastly, scenarios in the simulator were carefully calibrated to match the spatial and temporal aspects observed in the ND study, ensuring that the experiment remained valid and comparable with real-world findings.

3.2 Classification algorithms

According to the literature review, four classification models were applied to achieve the objective of this research, namely (1) support vector machines, (2) random forest, (3) AdaBoost, and (4) multilayer perceptron. All models were implemented using the scikit-learn library, with hyperparameters optimized through GridSearchCV to ensure each model's configuration was tailored to the dataset.

The use of machine learning (ML) models in experimental studies, such as driving simulator studies (DSS), offers significant advantages despite smaller sample sizes compared to naturalistic driving studies (NDS). ML models enhance predictive accuracy by identifying complex, non-linear relationships and key features influencing driving behavior that traditional methods might overlook [26]. The dual-source methodology, combining DSS and NDS data, enriches the dataset and enables a comprehensive evaluation of ML models, thereby advancing the understanding of driving behavior in both controlled and real-world conditions. This approach addresses the limitations inherent in each data source, leading to more robust and generalizable findings.

3.2.1 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 [27]. The context of the SVM model is to develop a hyper-plane in a multi-dimensional space to separate different class boundaries [28]. The key advantage of SVMs is that they can handle high-dimensional datasets [29]. The GridSearchCV tool from the scikit-learn library was employed to test various combinations of kernel types (linear, poly, rbf, sigmoid, regularization parameters C (1, 10, 50, 100, and kernel coefficients gamma (scale, auto. The configuration that provided the optimal balance of accuracy and generalization was identified with (a) kernel type = 'rbf', (b) regularization parameter C = 50; and (c) kernel coefficient gamma = 'scale'.

3.2.2 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 [30]. 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 [11], making it a preferred choice for identifying risky driving behavior. During the Grid Search, the number of trees (100, 200, 300, 400) and the criteria for splits (gini, entropy) were varied. 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', verified through cross-validation to ensure effectiveness across diverse data subsets.

3.2.3 AdaBoost

The AdaBoost algorithm is extensively used due to its high speed, low complexity, and good compatibility [31]. 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 [30]. This approach effectively combines these weak classifiers into a series to produce a strong classifier. Through the GridSearchCV method, various numbers of estimators (100, 200, 300, 300, 400, 500) were tested, with the ideal maximum number of estimators determined to be 500.

3.2.4 Multilayer perceptron (MLP)

The MLP is a feed-forward neural network complement and consists of three types of layers: (1) the input layer, (2) the output layer, and (3) the hidden layer [32]. The main advantage of the MLP algorithm is its ability to handle non-linear problems with large datasets while providing quick predictions. GridSearchCV facilitated the exploration of different network architectures, testing numbers of hidden layers (1–3), neurons per layer (100, 200, 300, 500), activation functions (relu, tanh), and regularization parameters (alpha values: 0.0001, 0.001, 0.01, 0.1). The most effective configuration featured (a) number of hidden layers = (500, 500, 500,), (b) activation function = "relu" and (c) alpha parameter of the regularization term = 0.0001.

A consistent methodological framework was applied in the implementation of GridSearchCV across all classification models to ensure comparability and methodological rigor. Each model utilized a uniform cross-validation strategy, ensuring that the evaluation of model performance was consistent. While the specific hyperparameters tested were customized to each model's unique characteristics and theoretical requirements, the approach to selecting the optimal parameters (based on achieving the best cross-validation performance) remained consistent across models. This structured yet flexible approach allowed for the rigorous tuning of each model while maintaining a standardized evaluative methodology [33]. To facilitate understanding of the hyperparameter tuning process, Table 4 summarizes each model, the hyperparameters explored, the defined search spaces, and the optimal hyperparameters identified through the GridSearchCV method. This table provides a concise overview of the systematic tuning approach applied across all models, enhancing comparability and interpretability.

Table 4 GridSearch Hyperparameter Tuning Summary

Model	Hyperparameters	Search space	Best Hyperparameters
SVM	Kernel type, Regularization parameter (C), Kernel coefficient (gamma)	Kernel: ['linear', 'poly', 'rbf', 'sigmoid'], C: [1, 10, 50, 100], gamma: ['scale', 'auto']	Kernel: 'rbf', C: 50, gamma: 'scale'
RF	Number of trees (Estimators), Split criterion	Estimators: [100, 200, 300, 400], Criterion: ['gini', 'entropy']	Estimators: 200, Criterion: 'entropy'
AdaBoost	Number of estimators	Estimators: [100, 200, 300, 400, 500]	Estimators: 500
MLP	Number of hidden layers, Neurons per layer, Activation function, Regularization parameter (alpha)	Hidden layers: [1, 2, 3], Neurons: [100, 200, 300, 500], Activation: ['relu', 'tanh'], Alpha: [0.0001, 0.001, 0.01, 0.1]	Hidden layers: (500, 500, 500), Activation: 'relu', Alpha: 0.0001

3.3 Evaluation metrics

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. A tenfold cross-validation method was employed, dividing the dataset into 10 equal parts. Each fold was used once as a validation set while the remaining nine folds served as the training set. This random division did not account for stratification by driver or trip groups. The rationale was to maintain statistical integrity and ensure randomness, allowing each data point an equal probability of inclusion in training or validation sets. This approach aimed to capture the overall diversity and variability of driving behavior, providing a comprehensive evaluation of the model's performance. The metrics utilized to evaluate the models are accuracy, precision, recall, f1-score, and false alarm rate defined by Eqs. (1) to (5):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

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

$$\text{f1} - \text{score} = \frac{2 \times (Precision) \times (Recall)}{(Precision) + (Recall)} \tag{4}$$

$$\text{False alarm rate} = \frac{FP}{FP + TN} \tag{5}$$

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.

4 Results

As mentioned above, the categorization of the target variable (e.g., output variable in the modelling process) of STZ level was based on certain headway thresholds. Furthermore, based on the analysis of the importance of the features in the dataset, the most important ones were selected as input variables (e.g., Speed, Distance travelled, Speed Limit) in the modelling process. Building upon these foundational steps, the dataset underwent preprocessing to enhance its suitability for modeling. This included performing tenfold cross-validation, where in each fold the dataset was divided into 90% for training and 10% for testing. Additionally, the adaptive synthetic sampling method (ADASYN) was applied to address class imbalance. Furthermore, the labels were scaled using the MinMaxScaler to standardize their ranges and improve model performance. With these preprocessing techniques applied, the models were trained and evaluated on the respective datasets. The subsequent subsections present the results of this analysis, highlighting the impact of these preprocessing steps on the performance and effectiveness of the machine learning algorithms deployed.

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 study dataset). Due to the phenomenon of "accuracy paradox" [34] 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) [35] technique was applied to address the imbalanced problem.

4.1 Classification models on simulator experiment dataset

Considering Fig. 1 and Table 5, 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 f1-score 63.42%, demonstrating its robustness and versatility. The MLP model also performs admirably with an accuracy of 81.28%, highlighting its capability in the simulator framework, achieving a competitive f1-score (61.79%).

Furthermore, the AdaBoost model achieves reasonable accuracy (75.08%) but has lower f1-score (55.87%) compared to RF and MLP. Although SVM shows a high recall, its lower accuracy (68.67%) and f1-score (53.22%) suggest a trade-off with precision. This implies that while SVM is effective at identifying true positive instances, it tends to include more false positives.

4.2 Classification models on naturalistic driving dataset

The results of the naturalistic driving study were similar to those of the simulator experiment. As illustrated in Fig. 2 and summarized in Table 6, RF achieved an adequate accuracy of 75.00%, 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 precision and recall, which is reflected in the f1-score (52.65%).

AdaBoost, scored the highest accuracy (76.76%) maintaining a competitive performance consistent with the simulator data, achieving the highest f1-score (60.19%). Finally, SVM maintains its proficiency in recall, showing consistency in capturing true positives. Lastly, SVM achieves an accuracy of 72.05% and an f1-score of 56.37%, which is relatively competitive but falls behind compared to the other models.

5 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.

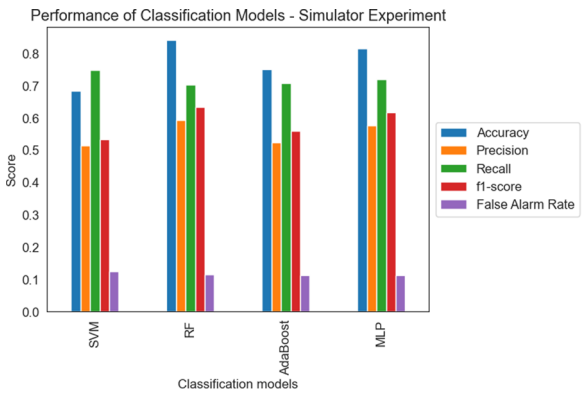

Fig. 1 Classification metrics of the four machine learning models for the simulator experiment dataset

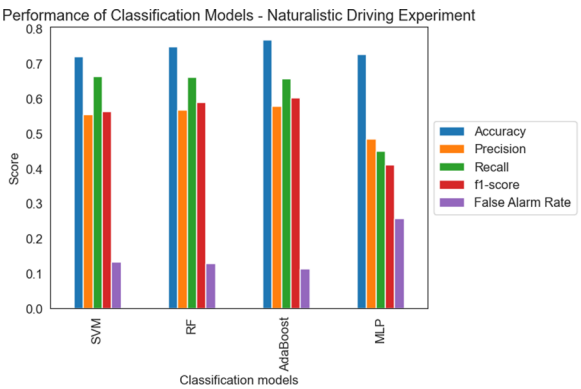

Fig. 2 Classification metrics of the four machine learning models for the naturalistic driving study dataset

Table 5 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

Table 6 Classification metrics for the naturalistic driving study 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

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, 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 study), the RF model emerged as the most efficient one.

The f1-scores, found to be moderate in both the simulator and naturalistic datasets, are indicative of the unique challenges that each data type presents to classification models. Although the simulator dataset offers a controlled environment that assists in model consistency, the nuanced nature of hazardous driving behaviors makes it difficult to achieve high f1-scores. In the naturalistic dataset, real-world factors such as environmental unpredictability and varying traffic conditions introduce additional noise, resulting in slightly lower f1-scores in comparison. Furthermore, the precision and recall of both datasets are influenced by the class imbalance of the dataset, which has fewer instances of hazardous behavior than normal driving. In order to capture a broader spectrum of driving behaviors, future studies could improve f1-scores by incorporating more diverse input features, such as driver biometrics or environmental factors, and investigating deep learning approaches, such as long short-term memory (LSTM) networks.

The observed variations in classification model performance between datasets derived from (A) simulator experiment and (B) naturalistic driving study, may be attributed to inherent dissimilarities in the data acquisition environments. Naturalistic driving study data depicts real-world driving scenarios with dynamic and unexpected aspects, adding a higher level of complexity than Simulator Experiment data, which is generated within a controlled virtual environment. The nuanced characteristics of real-world driving, such as diverse traffic conditions, weather variations, and unanticipated events, may challenge the models' ability to generalize effectively from the simulated environment. Differences in data distribution, noise levels, and the authenticity of driving behavior across the two sources may all contribute to observed model performance variances.

In the Simulator Experiment, Random Forest (RF) exhibited the highest accuracy at 84.00%, surpassing other classifiers, while in the Naturalistic Driving Study, AdaBoost achieved the highest accuracy at 76.76%. Notably, the precision of classifiers in the Simulator Experiment generally tended to be lower compared to the Naturalistic Driving Study. The Simulator Experiment RF model demonstrated precision, recall, and f1-score of 59.41%, 70.27%, and 63.42%, respectively, while in the Naturalistic Driving Study, AdaBoost achieved precision, recall, and f1-score of 57.91%, 65.81%, and 60.19%. The controlled nature of the simulator might influence RF to generalize effectively, as it excels at capturing and leveraging the inherent structure within the data. On the other hand, the dynamic and unpredictable aspects of real-world driving could have influenced the performance of AdaBoost in the Naturalistic Driving Study. These variations highlight the necessity of taking into account the contextual aspects of datasets when evaluating model performance, as well as the need for adaptable models that can efficiently address the complexities presented by many different types of experimental conditions.

Based on comparable driving behavior studies, the findings of this study were very similar to those described in the literature. For instance, Yang et al. [36] 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 study). In another study by Song et al. [13], 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. [13] considered variables such as gender, age, and driver perception.

Compared with the predicted results, Gan et al. [37] predicted that the accuracy rate was 75% when the data was 10,000. The model of this study achieved 80% prediction result, indicating that if the sample size is increased, the prediction accuracy will be also improved. This also reflects the superiority of the prediction model in this paper.

In contrast to the outcomes of this research, findings from the literature regarding the SVM classifier showed higher performance, especially with Yang et al. [14] having an outstanding accuracy rate of 95%. Additionally, in contrast to the research of Shangguan et al. [11], 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.

From a broader perspective, the implications of this study extend to traffic safety management and policy-making. The effective classification of driving behavior can inform various interventions aimed at reducing road crashes. For instance, the findings can aid in refining vehicle design, improving road infrastructure, and implementing targeted traffic regulations. Additionally, the development of advanced driver assistance systems (ADAS) can be guided by these insights, enhancing their ability to provide real-time warnings and post-trip feedback to drivers. The establishment of a Safety Tolerance Zone related to Time Headway (TH) is supported by research indicating that shorter headways are significantly associated with a higher likelihood of rear-end collisions due to the reduced reaction time available for drivers to respond to sudden changes in traffic conditions [17, 38].

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.

In conclusion, the study's outcomes highlight the necessity of considering the contextual aspects of datasets when evaluating model performance, as well as the need for adaptable models that can efficiently address the complexities presented by many different types of experimental conditions. These insights may inform various interventions such as refining vehicle body structures, enhancing road surface conditions, revising speed limits, and identifying hazardous road sections. This study lays a foundation for future research exploring the integration of deep learning techniques and expanding the diversity of datasets to enhance the generalizability and applicability of driving behavior models.

6 Conclusions

The research aimed to develop and evaluate four classification models on two distinct data sources (i.e., simulator experiment and naturalistic driving study) in order to predict risky driving behavior. 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 studies. 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 data, highlighting key road safety factors.

The observed performance variations among classification models have implications for real-world applications, especially regarding the potential misidentification of dangerous driving behavior. Discrepancies in accuracy, precision, and recall may compromise the reliable detection and classification of critical driving events, jeopardizing the effectiveness of automated systems designed to enhance road safety. Addressing these variations is

crucial for developing robust models with enhanced generalization capabilities.

In this context, future research could explore the utility of deep learning techniques, such as long short-term memory (LSTM) [39, 40]. While this paper focuses on conventional machine learning models, it is important to note that deep learning (DL) models have shown significant promise in surpassing traditional methods in similar applications. For instance, Saleh et al. [41] demonstrated that an LSTM-based model significantly outperformed traditional machine learning models in the classification of driving behavior using sensor data fusion. Their study found that the proposed Stacked-LSTM model achieved an F1-measure score of 91%, which was more than a 10% improvement over the closest compared approaches using conventional ML models. This highlights that LSTM networks are especially efficient at tasks requiring temporal relationships and intricate sequential patterns. Additionally, Naji et al. [42] found that deep learning models, particularly LSTM networks, outperformed traditional ML models (i.e., SVM, RF and MLP) in classifying the risk levels of near crashes, achieving a classification accuracy of 96%. To effectively apply DL models such as LSTM to datasets such as the one used in this study, an alternative dataset configuration approach would be required to maintain the temporal structure of the driving data recorded at 30-s intervals in order. The data should be arranged by driver and trip, and sequential segments of consecutive intervals (e.g., sets of five 30-s intervals) should be applied as each input to the model. This is due to the efficiency of DL models in identifying sequential patterns. Consequently, each trip should be regarded as a chronologically ordered sequence of intervals. DL models can leverage patterns over time to enhance prediction accuracy and reveal more complex driving behavior patterns. Furthermore, diversifying datasets, particularly by incorporating Naturalistic Driving study datasets involving drivers from different countries or transport modes, is crucial for a more holistic understanding of driver behavior. This diversity contributes to the development of models with robust generalization capabilities, ultimately enhancing their reliability and applicability in real-world scenarios. Addressing these research directions advances the field and contributes significantly to the development of automated systems that can effectively improve road safety measures.

Abbreviations

WHO	World Health Organization
DSS	Driving simulator studies
NDS	Naturalistic driving studies
SVM	Support vector machines
RF	Random forest
MLP	Multilayer perceptron
TP	True positives
TN	True negatives
FP	False positives
FN	False negatives

Acknowledgements
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).

Author contributions
TG: Conceptualization, Methodology, Data curation, Data analysis, Software, Writing, Revision; EM: Conceptualization, Methodology, Data curation, Data analysis, Writing, Revision; SR: Data curation, Data analysis, Writing, Revision; CK: Conceptualization, Methodology, Revision; TB: Supervision, Resources, Revision; GY: Supervision, Resources, Revision. All authors have read and agreed to the published version of the manuscript.

Funding
This article has been published open access with support of the TRA2024 project funded by the European Union. 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).

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors state that they do not have any clear financial conflicts or personal interests that might have seemed to affect the work described in this paper.

Received: 10 October 2023 Accepted: 5 November 2024
Published online: 21 November 2024

References

1. World Health Organization. (2023). *Global status report on road safety 2023*. World Health Organization. https://www.who.int/publications/i/item/9789240086517
2. Staubach, M. (2009). Factors correlated with traffic accidents as a basis for evaluating advanced driver assistance systems. *Accident Analysis and Prevention, 41*(5), 1025–1033. https://doi.org/10.1016/j.aap.2009.06.014
3. 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.
4. Michelaraki, E., Katrakazas, C., Brijs, T., & Yannis, G. (2021). Modelling the safety tolerance zone: Recommendations from the i-DREAMS project. In *10th International congress on transportation research*.
5. 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 and Prevention, 123*, 274–281. https://doi.org/10.1016/j.aap.2018.12.005
6. 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 and Prevention, 129*, 108–118. https://doi.org/10.1016/j.aap.2019.04.017
7. 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
8. 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

9. 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), 10562. https://doi.org/10.3390/app112210562
10. 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
11. 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 and Prevention, 156*, 106122. https://doi.org/10.1016/j.aap.2021.106122
12. 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 and Prevention, 129*, 170–179. https://doi.org/10.1016/j.aap.2019.05.005
13. 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 and Prevention, 153*, 106038. https://doi.org/10.1016/j.aap.2021.106038
14. Yang, K., Haddad, C. Al, Yannis, G., & Antoniou, C. (2021). Driving behavior safety levels: Classification and evaluation. In *2021 7th International conference on models and technologies for intelligent transportation systems (MT-ITS)* (pp. 1–6). https://doi.org/10.1109/MT-ITS49943.2021.9529309
15. Zhang, C., Patel, M., Buthpitiya, S., Lyons, K., Harrison, B., & Abowd, G. D. (2016). Driver classification based on driving behaviors. In *Proceedings of the 21st international conference on intelligent user interfaces* (pp. 80–84). https://doi.org/10.1145/2856767.2856806
16. 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
17. Knipling, R. R., Mironer, M., Hendricks, D. L., Tijeripa, L., Everson, J., Allen, J. C., Wilson, C., et al. (1993). *Assessment of IVHS countermeasures for collision avoidance: Rear-end crashes*.
18. Ohta, H. (1993). Individual differences in driving distance headway. *Vision in Vehicles, 4*, 91–100.
19. Michael, P. G., Leeming, F. C., & Dwyer, W. O. (2000). Headway on urban streets: Observational data and an intervention to decrease tailgating. *Transportation Research Part F: Traffic Psychology and Behaviour, 3*(2), 55–64. https://doi.org/10.1016/S1369-8478(00)00015-2
20. 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
21. 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
22. 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
23. Kennedy, R. S., Stanney, K. M., & Dunlap, W. P. (2000). Duration and exposure to virtual environments: sickness curves during and across sessions. *Presence: Teleoperators & Virtual Environments*, *9*(5), 463–472
24. Mobileye. (2023). https://www.mobileye.com/
25. CardioID Technologies. (2023). https://www.cardio-id.com/automotive/
26. Wang, C., Liu, L., Xu, C., & Lv, W. (2019). Predicting future driving risk of crash-involved drivers based on a systematic machine learning framework. *International Journal of Environmental Research and Public Health, 16*(3), 334. https://doi.org/10.3390/ijerph16030334
27. 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
28. Ghosh, S., Dasgupta, A., & Swetapadma, A. (2019). A study on support vector machine based linear and non-linear pattern classification. In *2019 International conference on intelligent sustainable systems (ICISS)* (pp. 24–28). https://doi.org/10.1109/ISS1.2019.8908018
29. 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.
30. 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
31. 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
32. Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. *Advances in Computers, 117*(1), 339–368. https://doi.org/10.1016/bs.adcom.2019.09.007
33. Ahmad, G. N., Fatima, H., Ullah, S., Salah Saidi, A., & Imdadullah. (2022). Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. *IEEE Access, 10*, 80151–80173. https://doi.org/10.1109/ACCESS.2022.3165792
34. 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
35. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322–1328). https://doi.org/10.1109/IJCNN.2008.4633969
36. 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
37. Gan, J., Li, L., Zhang, D., Yi, Z., & Xiang, Q. (2020). An alternative method for traffic accident severity prediction: Using deep forests algorithm. *Journal of Advanced Transportation, 2020*, 1–13. https://doi.org/10.1155/2020/1257627
38. Ding, N., Zhu, S., Jiao, N., & Liu, B. (2020). Effects of peripheral transverse line markings on drivers' speed and headway choice and crash risk in car-following: A naturalistic observation study. *Accident Analysis and Prevention, 146*, 105701. https://doi.org/10.1016/j.aap.2020.105701
39. 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/10.1016/j.aquaeng.2020.102053
40. 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
41. Saleh, K., Hossny, M., & Nahavandi, S. (2017). Driving behavior classification based on sensor data fusion using LSTM recurrent neural networks. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)* (pp. 1–6). https://doi.org/10.1109/ITSC.2017.8317835
42. Naji, H. A. H., Xue, Q., Lyu, N., Duan, X., & Li, T. (2022). Risk levels classification of near-crashes in naturalistic driving data. *Sustainability, 14*(10), 6032. https://doi.org/10.3390/su14106032

Publisher's Note