World Conference on Transport Research (WCTR)
6-10 July 2026, Toulouse, France

Identifying Dangerous Driving Behaviour through Big Data and Machine Learning Techniques

Hector Kamvoussioras, Thodoris Garefalakis, Eva Michelaraki, George Yannis

National Technical University of Athens, Department of Transportation Planning and Engineering, Athens, Greece

Introduction

Road safety is an important area that affects society, economy and many other areas. Europe has made significant efforts to reduce the number of traffic crashes. The **significance of modern automobile technology** and transportation automation to improving road safety has received special attention in recent decades.

Human mistakes are **responsible for 94% of major crashes**. Therefore, it is crucial to investigate new technology in order to improve road safety by reducing human error. The goal of using Advanced Driver-Assistance Systems (ADAS) is to either prevent or lessen the severity of unavoidable collisions. These systems use a range of sensors, including cameras, radar and ultrasonic technologies, to collect vital information about the vehicle's environment in order to proactively prevent collisions or lessen their effects. Even though ADAS is a huge advancement in improving road safety, there are still issues with using these technologies to accurately forecast and reduce crash risks in real time.

Objectives

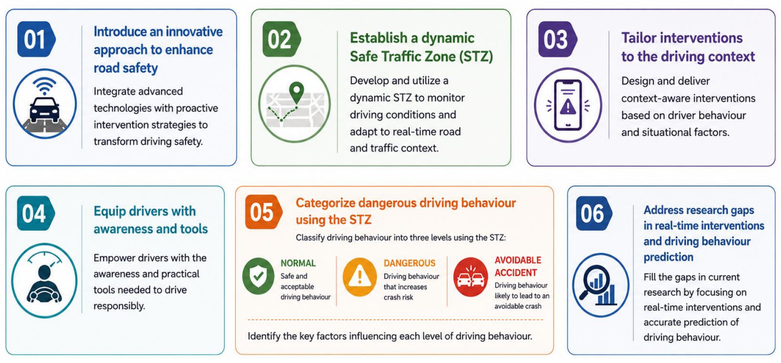

***Figure 1**: Objectives of the present study*

Data Overview

To achieve this objective, a **driving simulator experiment** was conducted, involving 36 drivers aged 29-60, with an average age of 42 years. Although the sample size was relatively modest, it is consistent with the requirements and common practice of driving simulator studies, where data collection is resource-intensive and conducted under highly controlled experimental conditions. Figure 2 illustrates the system and its integration within the simulator environment.

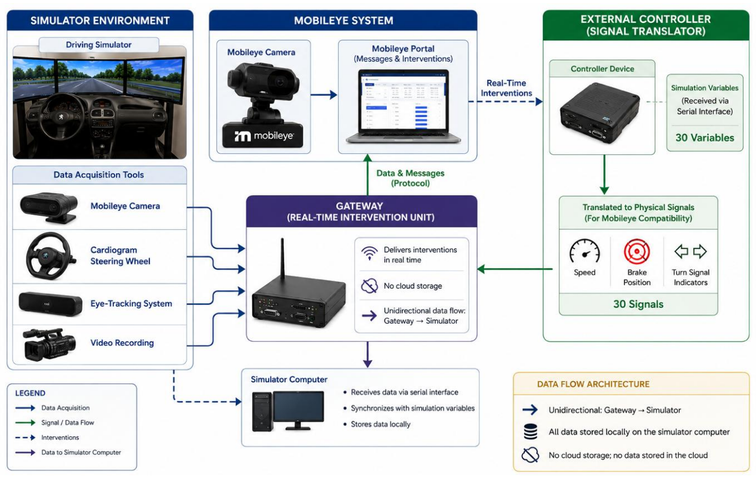

***Figure 2:** Mobileye system and its integration within the simulator environment*

Table 1 presents the variables used in this study, including their descriptions, units of measurement and data types.

***Table 1:** Variables used in the driving behaviour classification model*

Variable	Description	Units	Type
Calculated in OM			
TTC	Time to collision with the vehicle ahead	Seconds	Numeric
Headway	Time headway to the vehicle ahead in the same lane	Seconds	Numeric
Speed	Vehicle speed	Km / h	Numeric
Distance_travelled	Distance driving	Meters	Numeric
Data From Mobileye			
ME_ForwardCollisionWarning	Whether a forward collision warning is active	bool	Logical (False or True)
ME_LaneDepartureWarningActive	Whether Lane Departure Warning is active	bool	Logical (False or True)
Data From Cardiogram Steering Wheel			
HandsOnEvent	Whether hands are on the steering wheel	None / both	Discrete
FatigueEvent	KSS score	1 - 9	Discrete

Methodology

Four classification algorithms: Ridge Classifier (RC), Support Vector Machines (SVM), Random Forests (RF) and eXtreme Gradient Boosting (XGBoost) were developed to predict driving behaviour within 30-second intervals. To better understand the contribution of individual features, SHAP analysis was conducted.

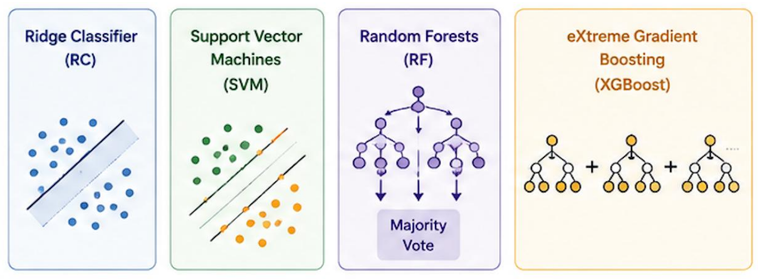

***Figure 3:** Four classification algorithms applied*

Results

Through the systematic **feature selection process**, the most relevant variables were identified and organized into two groups: Group A, consisting of TTC, headway, speed and forward collision warning, while Group B, included TTC, headway, Hands on wheel, fatigue event, lane departure warning, speed and distance travelled. Figure 4 illustrates these relationships, providing a visual representation of both the magnitude and direction of each feature's effect. Features with higher absolute coefficient values are considered more influential, offering valuable insights for feature selection and supporting the interpretability of the classification model.

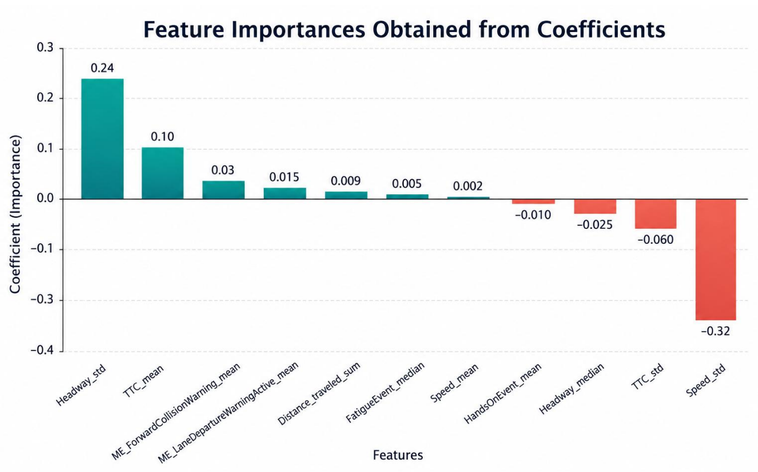

***Figure 4:** Feature importance obtained from coefficients*

The RC achieved moderate accuracy in predicting the normal level but performed poorly in identifying the avoidable accident and dangerous levels, indicating limited capability in recognizing risky behaviour. The SVM demonstrated reasonable accuracy across all three levels, with an overall accuracy of 73.5%. However, it also struggled with accurate predictions of the avoidable accident level. In contrast, the **RF model showed strong performance across all categories**, achieving 90% accuracy with minimal error rates. Its superiority was further confirmed in Group B, where accuracy increased to 95%, outperforming results from Group A.

***Table 2:** Performance metrics for Group A and B*

	Precision	Recall	F-means	False Alarm Rate	G-means
Group A					
Ridge Classifier	43%	35%	27%	95%	19%
Support Vector Machine	59%	71%	62%	29%	77%
Random Forest	80%	88%	83%	12%	90%
XGBoost	80%	88%	83%	12%	90%
Group B					
Ridge Classifier	50%	33%	23%	67%	5%
Support Vector Machine	92%	83%	86%	17%	83%
Random Forest	92%	90%	91%	7%	95%
XGBoost	91%	93%	92%	7%	96%

Similarly, the **XGBoost model delivered results comparable to RF**, with 90% accuracy in Group A and improved performance in Group B. The comparison between the two groups highlights that incorporating additional variables enhances predictive accuracy, as Group B consistently outperformed Group A. The outcomes of these evaluations for both Groups are presented in Table 4.

The **SHAP value analysis further highlighted distinct variable importance patter**ns across the two groups. In Group A, time-to-collision emerged as the most critical predictor, with lower values strongly associated with higher-risk classifications (dangerous and avoidable accident).

Discussion

Headway also played a central role, particularly in differentiating between normal and risky behaviour, while speed was more strongly linked to normal driving. Results indicated that time-to-collision and headway variability are central indicators of driver safety when fewer variables are considered. In contrast, Group B emphasized speed as the dominant predictor, with higher values correlating with increased risk. Furthermore, hands on wheel, time-to-collision and headway contributed significantly, suggesting that average speed, distance to other vehicles and sustained driver engagement with the steering wheel are strong determinants of risk.

Exposure variables such as distance travelled also played a role, reinforcing that driving context and duration can influence the likelihood of risky behaviour. Although **fatigue event** and lane departure warning exhibited relatively limited effects, the results suggest that well-rested drivers tend to engage in less hazardous behaviour, consistent with previous findings on fatigue-related risk. Figure 5 illustrates that a decrease in TTC_mean acts as a critical trigger for class transitions, particularly influencing the categories labeled as dangerous (1) and avoidable accident (2).

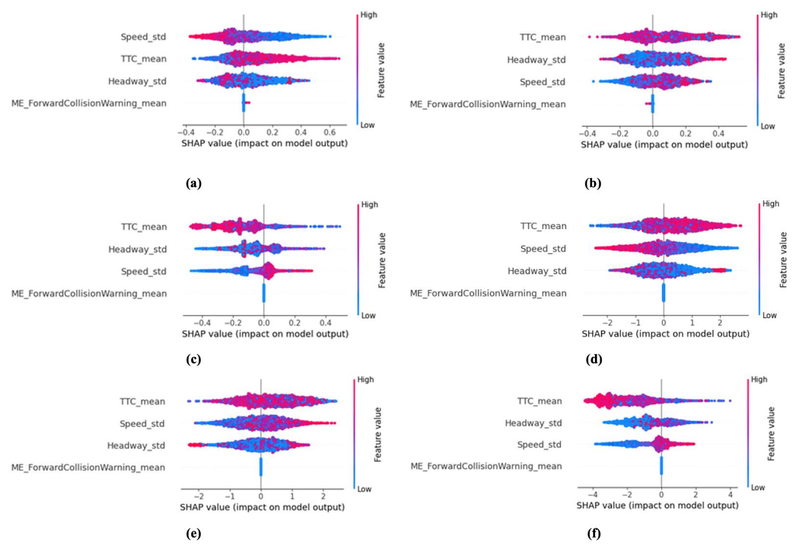

***Figure 5**: SHAP values (a) for RF - Class 0 - Group A; (b) for RF - Class 1 - Group A; (c) for RF - Class 2 - Group A; (d) for XGBoost - Class 0 - Group A; (e) for XGBoost - Class 1 - Group A; (f) for XGBoost - Class 2 - Group A*

The comparison of Groups A and B revealed that expanding the feature set allows models to capture a **more holistic view of driver risk**, shifting emphasis from immediate situational factors (TTC, headway) toward broader behavioural indicators (average speed, hands-on control, exposure). This reflects the complexity of driving behaviour, which is shaped not only by instantaneous performance but also by accumulated behavioural patterns over time.

Conclusions

- The findings of this study provide strong evidence of the value of **predicting risky driving behaviour** for road safety policy and practice
- The **identification of the specific factors** that contribute to risky driving behaviour, such as time to collision, headway variability, speed and driver attentiveness, transportation Authorities can design more targeted and effective interventions.
- The **integration of predictive models** into ADAS represents a significant opportunity for policymakers and industry stakeholders
- **Real-time interventions** based on behavioural predictions can provide immediate feedback to drivers, fostering safer practices and reducing crash likelihood
- Beyond in-vehicle applications, insurance providers and regulatory bodies could utilize these **predictive tools to inform risk-based policies**, promote responsible driving through incentives and ensure fairer premium structures

Contact Information: Eva Michelaraki, PhD, Research Associate NTUA
Email: evamich@mail.ntua.gr
Website: https://www.nrso.ntua.gr/p/evamich/