Leveraging driver vehicle and environment interaction: Machine learning using driver monitoring cameras to detect drunk driving

Kevin Koch*
kevin.koch@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Martin Maritsch*
mmaritsch@ethz.ch
ETH Zürich
Zürich, Switzerland

Eva van Weenen
evanweenen@ethz.ch
ETH Zürich
Zürich, Switzerland

Stefan Feuerriegel
feuerriegel@lmu.de
LMU Munich
Munich, Germany

Matthias Pfäffli
matthias.pfaeffli@irm.unibe.ch
University of Bern
Bern, Switzerland

Elgar Fleisch
efleisch@ethz.ch
ETH Zürich and University of St. Gallen
Zürich and St. Gallen, Switzerland

Wolfgang Weinmann†
wolfgang.weinmann@irm.unibe.ch
University of Bern
Bern, Switzerland

Felix Wortmann†‡
felix.wortmann@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

ABSTRACT

Excessive alcohol consumption causes disability and death. Digital interventions are promising means to promote behavioral change and thus prevent alcohol-related harm, especially in critical moments such as driving. This requires real-time information on a person's blood alcohol concentration (BAC). Here, we develop an in-vehicle machine learning system to predict critical BAC levels. Our system leverages driver monitoring cameras mandated in numerous countries worldwide. We evaluate our system with $n = 30$ participants in an interventional simulator study. Our system reliably detects driving under any alcohol influence (area under the receiver operating characteristic curve [AUROC] 0.88) and driving above the WHO recommended limit of 0.05 g/dL BAC (AUROC 0.79). Model inspection reveals reliance on pathophysiological effects associated with alcohol consumption. To our knowledge, we are the first to rigorously evaluate the use of driver monitoring cameras for detecting drunk driving. Our results highlight the potential of driver monitoring cameras and enable next-generation drunk driver interaction preventing alcohol-related harm.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Ubiquitous and mobile computing systems and tools*; • **Applied computing** → *Consumer health.*

KEYWORDS

health; safety; driving; alcohol; eye movements; head movements; driver monitoring

ACM Reference Format:
Kevin Koch, Martin Maritsch, Eva van Weenen, Stefan Feuerriegel, Matthias Pfäffli, Elgar Fleisch, Wolfgang Weinmann, and Felix Wortmann. 2023. Leveraging driver vehicle and environment interaction: Machine learning using driver monitoring cameras to detect drunk driving. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/3544548.3580975

1 INTRODUCTION

Alcohol consumption is responsible for 5% of the global disease burden and is further the cause of 1 in 20 deaths worldwide [110]. To promote behavior change, digital interventions provide effective means to prevent harm in critical situations due to alcohol consumption and intoxication [3, 5, 64, 65]. In particular, digital interventions could promote behavior change by delivering real-time targeted feedback on alcohol consumption. However, to intervene early, real-time predictions of alcohol consumption are needed.

Alcohol consumption increases, among others, the risk of traffic crashes, making drunk driving one of the leading causes of severe crashes on public roads. For example, in the US, around 30 people die each day in traffic crashes in which one of the parties is under the influence of alcohol, and, together, alcohol-related crashes amount to 30% of all traffic fatalities [67]. To prevent alcohol-related crashes, in-vehicle systems are needed to detect drunk driving and enable targeted interventions. Examples of such interventions are, e.g., warnings of impairment and forced stops of the vehicle.

As of today, the only reliable measurement technology for identifying intoxicated driving are ignition interlock devices that analyze a driver's breath alcohol. However, ignition interlock devices

[*] Joint first author.
[†] Joint last author.
[‡] Corresponding author.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580975

are expensive and, further, require regular maintenance. State-of-the-art devices cost around USD 1000 and need yearly maintenance [77, 101]. Regulators are aware of these challenges: For example, the US Congress recently introduced a concrete timeline for drunk driving detection systems in vehicles [96], calling for scalable, low-cost, and easily accessible technologies.

A cost-effective and scalable approach could be to measure the driver performance on the basis of the existing sensor technology of today's vehicles. Even though progress has been made toward fully autonomous driving, experts agree that autonomous driving will not be widely available in the next two decades [6, 69]. Hence, in the coming years, driving will still require to interact with the vehicle as well as the environment, and, thus, detection systems are needed that build upon existing vehicle technology. Here, we develop and evaluate a machine learning system for detecting drunk driving based on driver monitoring cameras already built into modern vehicles. In fact, driver monitoring cameras will be introduced in the coming years in almost all new vehicles due to safety regulations, such as the European New Car Assessment Programme (Euro NCAP) or the EU General Safety Regulation (GSR), which make them mandatory from 2024 onwards [26, 27].

Contributions

In this paper, we develop and evaluate a novel machine learning system to detect drunk driving from driver monitoring cameras leveraging driver vehicle and environment interaction. Specifically, our system extracts information on gaze behavior and head movements from driver monitoring cameras and then predicts whether drivers exceed two critical thresholds for blood alcohol concentration (BAC): (1) BAC values above 0.00 g/dL yet below the World Health Organization (WHO) recommended [111] legal limit of 0.05 g/dL for early warnings and (2) BAC values above the WHO recommended limit. For this purpose, we conducted an interventional clinical trial with $n = 30$ healthy participants that completed driving tasks in a research-grade driving simulator with different levels of alcohol intoxication and in different driving environments.

The contribution, novelty, and significance of our work are as follows:

- **Contributions:** (1) Our system reliably detects driving under the influence of alcohol. (2) Our approach based on a driver monitoring camera outperforms previous approaches based on driving data (e.g., [47, 48, 52, 79, 92]) by a clear margin. (3) Our system is highly generalizable as the detection is robust to unseen individuals and driving scenarios. (4) Analysis of the learned patterns of our machine learning model shows that our system relies upon known pathophysiological mechanisms of alcohol intoxication.
- **Novelty:** To our knowledge, we are the first to rigorously evaluate the use of driver monitoring cameras for detecting drunk driving by conducting a clinical trial with participants driving in a research-grade simulator, both sober and drunk. Although existing work on driver state monitoring addresses safety-critical states such as drowsiness or distraction, previous attempts at detecting drunk driving with vehicle signals have not achieved sufficiently good results or were not rigorously evaluated (e.g., [17, 35, 47, 48, 52, 83, 98, 100]).
- **Significance:** Our system offers a viable approach based on existing technologies, allowing for a rapid implementation to prevent potential negative consequences of drunk driving after decades of stagnating high alcohol-related road crashes without any significant advancement by regulators and industry [67, 90]. To accelerate this development, we provide the source code of our evaluated machine learning system on GitHub: https://github.com/im-ethz/CHI-2023-paper-Leveraging-driver-vehicle-and-environment-interaction.

2 RELATED WORK

In recent years, the Human–Computer Interaction (HCI) community has already taken steps toward digital systems for monitoring alcohol consumption and intervening when needed. Previous work detects alcohol intoxication based on gait analysis [43], smartphone interactions and related contextual data [40, 56, 60], sometimes in conjunction with smart breathalyzer systems [39, 115], wrist-worn devices that measure physical or transdermal activity [45, 116], or social media data [74, 75]. Interventions targeting responsible alcohol consumption include self-management of use through diaries [118], informing peers and family about recent drinking behavior [114, 117], or virtual agents that intervene [53]. These systems demonstrate that it is feasible to detect alcohol consumption and that digital interventions are capable of successfully averting harmful behavior. However, the use scenarios of these systems are very broad; we believe that addressing harmful alcohol use in situations associated with truly harmful outcomes will significantly impact reducing alcohol-related harm.

2.1 Overview on driver state detection

In the realm of driving, research and industry have introduced increasingly advanced assistance and safety systems in recent decades that understand the interaction between drivers, vehicles, and environments. In the case of driver monitoring systems, there has been a strong focus on the condition of drowsiness. For several years, there are now systems commercially available that issue drowsiness warnings based on existing vehicle signals from the CAN bus (i.e., a central communication network with high-frequency messages between vehicle components, such as steering wheel or accelerator/brake pedal actuation, speed, or acceleration) [7, 23]. Due to increasing availability and higher detection performance, modern commercial drowsiness detection systems shifted to the use of cameras [23, 88]. Building on these widespread driver monitoring systems, research and industry have developed, validated, and commercialized various algorithms that cover additional use cases, often coming from non-vehicle domains, such as driver identification, distraction detection, and emotion detection based on eye, pupil or head movements, or facial expressions [2, 29, 44, 103]. Common across these systems is their warning approach when an impairment is detected. They intervene with a simple audiovisual warning, delivered via the vehicle's infotainment system on the center console, which emits a clearly audible warning tone and recommends taking a break. Figure 1 shows the implementation of such warnings by Volkswagen and Nissan [68, 102].

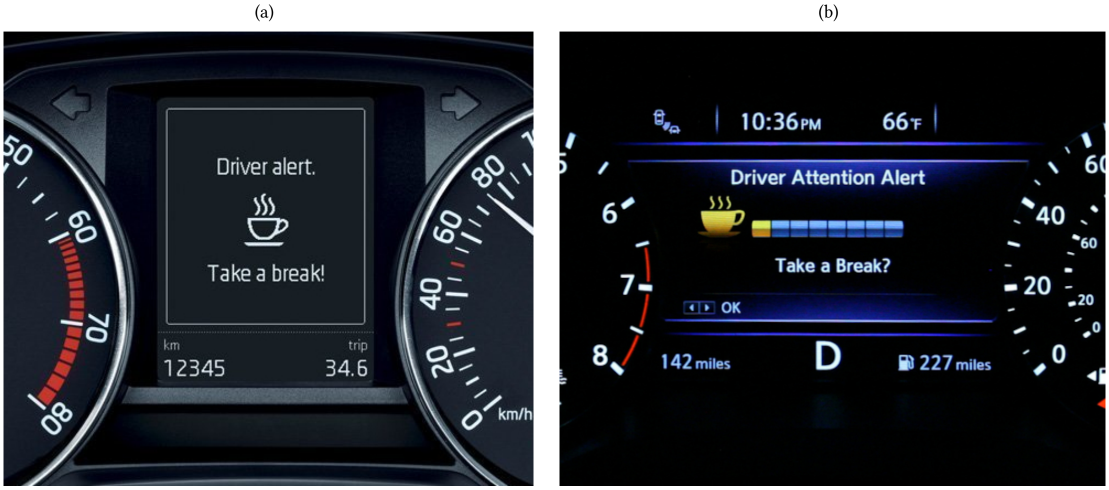


Figure 1: Examples of existing drowsiness and attention warning systems in today’s cars. (a) Drowsiness warning delivered in Volkswagen cars; (b) Attention warning delivered in Nissan cars.

2.2 Camera-based driver monitoring

Driver monitoring systems in industrial research are often proprietary approaches, and implementation details are not available to a broad audience. Fortunately, related academic work provides extensive research on driver monitoring (e.g., see the literature reviews in [23, 103]). Generally, camera-based driver monitoring focuses on a per-use-case solution of problems, and, if solved, a specific use case is combined with existing systems as part of an ensemble to monitor various driver states [8, 88]. In the following, we discuss the specific example of drowsiness detection to explain common approaches on the development and evaluation of driver monitoring systems. In general, the majority of detection systems rely on the detection of behavioral patterns that are associated with a specific driver state. In the case of drowsiness, these patterns are, for example, higher percentages of eye closure, yawning, or head nodding and scaling [103]. Hence, related studies recorded people acting and imitating typical drowsiness patterns using (driver monitoring) cameras (e.g., openly available datasets are [1, 106]). Subsequently, machine learning approaches are trained to detect these patterns. In general, two major approaches exist. First, traditional approaches rely upon basis signals describing low-level characteristics of visual behavior, such as gaze direction, eyelid closure, or facial landmarks, which are extracted with computer vision algorithms (e.g., OpenCV [9] and Viola Jones [99]) or neural networks (e.g., [66, 121]). Subsequently, those signals are used to either hand-craft features, such as Percentage Eye Closure (PERCLOS), or to train deep neural networks for the automated detection of drowsiness patterns (e.g., [15, 119]). Second, more recent work leverages end-to-end neural networks with the intention of directly detecting behavior patterns associated with drowsiness from raw images (e.g., [120]). If specific movements exceed predefined thresholds, the system raises an alarm.

While the prior described imitation of drowsy patterns by people acting has safety advantages and can be conducted with less effort (i.e., drivers do not drive in a critical drowsiness state), these imitations are only approximations. In comparison, statistical analyses on drowsiness detection commonly rely upon driving in actual drowsiness by either using long driving times or methods for sleep deprivation (e.g., see literature review [89]). To our knowledge, there are only very few studies on drowsiness detection based on cameras in which drivers were actually in a drowsy state [11, 93, 122]. Here, the entire visual, facial, and head movement behavior is used as a proxy for driver impairment. Awake and drowsy driving trips are partitioned into smaller sequences over which behavior signals are aggregated.

2.3 Drunk driving detection

Although there is considerable work on drunk driving, this work focuses largely on statistical analysis examining the influence of alcohol on driving performance (see, for example, the systematic review in [41]). These empirical findings provide a basis for regulators to decide on relevant legal thresholds for driving under the influence of alcohol [28]. However, these previous works did not focus on the real-time identification of drunk driving, which would allow for intervening when a driver is actually drunk. Prior research also covered the negative effect of alcohol on gaze behavior while driving (e.g., tunnel vision) [63, 87].

In contrast, related work on *detecting* drunk driving is still at comparatively early stages. Similar to the early work on detecting drowsiness, research focused on detecting drunk driving based on driving behavior such as steering, pedal usage, and vehicle speed. These past findings demonstrate the general feasibility of detecting drunk driving from in-vehicle signals [47]. Unfortunately, the results of this study and comparable work on the detection of drunk driving [19, 20, 35, 36, 48, 51, 52, 79, 92] show that, while machine learning models trained on driving behavior yield a good performance on previously seen drivers, they do not achieve the performance needed to generalize to unseen drivers. One main reason is that driving behavior is overlaid by additional influences from the

Table 1: Participant characteristics (n = 30). AUDIT: Alcohol Use Disorder Identification Test; PEth: phosphatidylethanol; SD: standard deviation.

	Mean ± standard deviation (SD)	Min	Max
Gender	15 female, 15 male	–	–
Age [years]	37.0 ± 9.2	21	59
Weight [kg]	75.0 ± 26.0	55.7	96.3
Height [cm]	173.5 ± 56.4	159	192
Driving experience [years]	13.3 ± 7.8	4	29
Driving distance [km/year]	8300 ± 6820	400	30000
Professional drivers	1 (truck driver)	–	–
AUDIT score	5.71 ± 2.57	1	11
PEth blood concentration [ng/mL]	4 below detection and 5 below quantification limit*	–	–
	21 above quantification limit* with 84.1 ± 48.7	28.6	199.0

*detection limit 10 ng/mL, quantification limit: 20 ng/mL

environment (highway, rural, and urban) and further varies across drivers (e.g., slow vs. fast drivers, defensive vs. aggressive drivers), which introduces additional noise and thus makes inferences of alcohol levels challenging. As a remedy, we propose to shift from driving behavior (i.e., how does the vehicle behave?) to driver behavior (i.e., how does the driver behave?). To this end, we leverage driver monitoring cameras to capture eye movements, gaze events, and head movements as fine-grained and frequent predictors of driving under the influence of alcohol.

Regarding camera-based driver monitoring to detect drunk driving, we found that related work either lacks rigor and clarity or only proposes a solution without evaluation. More specifically, studies miss key information about experimental designs, drinking procedures, targeted alcohol levels, or methods to measure alcohol levels. Furthermore, they apply unclear evaluation approaches, making it impossible to interpret the results, replicate these studies, or understand them at all (e.g., [17–19, 21, 80]). Although many studies elaborate on the possibility of detecting drunk driving and design systems, they often do not evaluate them (e.g., [83, 98, 100]).

Given the limitations of previous work on drunk driving detection, we conducted an interventional clinical trial following standardized principles in clinical research to rigorously evaluate driver monitoring cameras for drunk driving detection. In general, the standard regarding study quality in the field of driver monitoring cameras is insufficient as shown by both literature reviews on drowsiness and drunk driving detection. Ultimately, we believe it is necessary to analyze drivers in a reliable impairment state to thoroughly evaluate detection approaches.

3 DATA COLLECTION

In this paper, we aimed at developing and evaluating a novel machine learning (ML) system to detect critical BAC thresholds. For this purpose, we conducted a non-randomized, single-blinded, interventional, single-center study (ClinicalTrials.gov NCT04980846) called DRIVE (design and implementation of a drunk driving detection system). The DRIVE study took place between August 2021 and November 2021 in Bern, Switzerland. The study followed the Declaration of Helsinki, the guidelines of good clinical practice, the Swiss health laws, and the ordinance on clinical research. Each participant gave informed written consent prior to any study-related procedure. The study was further approved by the local ethics committee in Bern, Switzerland (ID 2021-00759). Throughout this paper, we report statistics as mean ± SD. In the following, we describe the overall study procedure and data collection in detail.

3.1 Sample size calculation

To determine the number of participants, we cannot rely on power calculation as traditional null hypothesis testing is not applicable to our study (i.e., there is no null hypothesis for the development of ML models). Therefore, we implemented an established methodology from a previous study [30] to extrapolate the discriminatory power of ML with increasing sample size. Due to the lack of pre-existing literature in the field, this method was applied to preliminary data that we retrieved in a pilot study (n = 10). Based on this approach, an area under the receiver operating characteristic curve (AUROC) of 0.85 to detect driving above the WHO recommended BAC of 0.05 g/dL was projected for a sample size of 30.

3.2 Participants

Inclusion criteria for individuals eligible for participation were: (1) passing the driver examination at least 2 years before study inclusion; (2) possession of a driver's license that is valid in the European Union or Switzerland; and (3) reporting moderate alcohol consumption (i.e., neither total absence nor excess). The latter was validated based on two instruments. First, participants were asked with the Alcohol Use Disorder Identification Test (AUDIT) [84] multiple-choice questionnaire about their alcohol consumption (e.g., "How often do you have a drink containing alcohol?") and related experiences (e.g., "During the past year, how often have you been unable to remember what happened the night before because you had been drinking?"). Each answer is scored between 0 to 4 and their total sum is a proxy for drinking behavior. A sum of 0 indicates total absence. A score of 1 to 7 suggests low-risk consumption. Scores from 8 to 14 suggest hazardous or harmful alcohol consumption and a score of 15 or more indicates the likelihood of alcohol dependence (excess). Second, we collected the

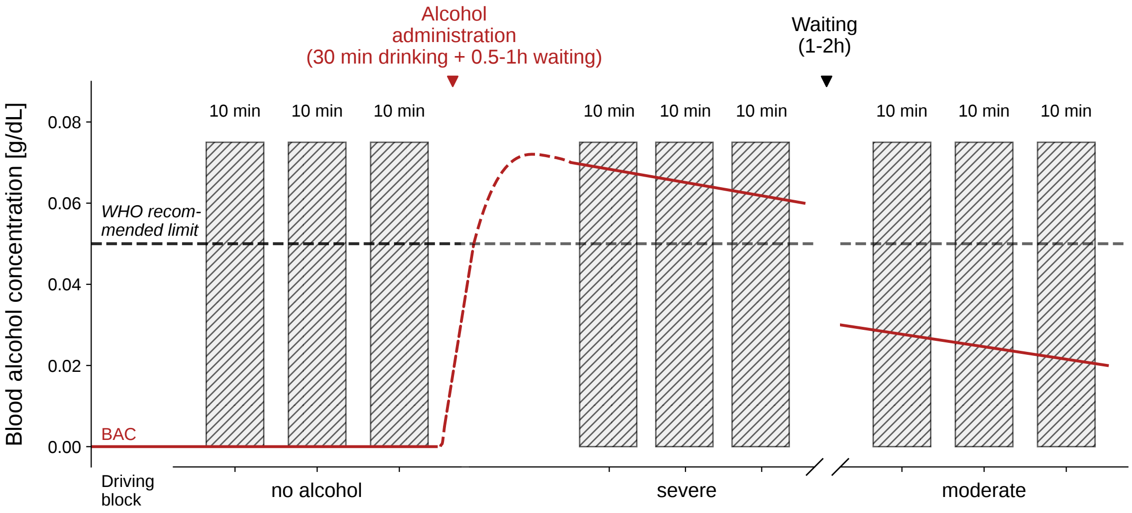


Figure 2: Study overview. Overview of study procedure where participants performed driving arranged in three blocks during which different alcohol levels were targeted (i.e., no alcohol, moderate, and severe). BAC: blood alcohol concentration; WHO: World Health Organization.

phosphatidylethanol (PEth) level of a capillary blood sample on the pre-screening day, which needed to be below 210 ng/mL [112]. Exclusion criteria for participation in the study were if participants met one or more of the following: pregnancy or an intention for pregnancy; health conditions incompatible with alcohol consumption; known or suspected non-compliance; participation in another study with investigational drug preceding and during the present study; personal dependencies with the study team (e.g., employees, family members, and other dependent persons); experience of motion sickness.

In total, we screened 39 participants for eligibility in our study. After checking for inclusion and exclusion criteria, we finally collected data from $n = 30$ participants (15 female, 15 male, age 37.03 ± 9.24 years). Detailed participant characteristics are reported in Table 1, summarizing the demographics, driving experience, and alcohol consumption of the participants. More details on the enrollment in the study are in Appendix A.

3.3 Study procedure

Following a telephone screening interview, participants were invited to the study location for on-site screening. After informed consent was obtained, a simulator training session followed. The training session was used to familiarize participants with the driving simulator and to test whether they experience motion sickness.

On the day of the study visit, participants arrived at the research facility in the morning after an overnight fast. After an initial, additional training session in the driving simulator, participants conducted driving tasks in a driving simulator during controlled alcohol administration. Here, the targeted alcohol levels were: (1) *no alcohol* (BAC of 0.00 g/dL) (no alcohol); (2) *severe*, that is, above the WHO recommended legal limit (i.e., 0.05 g/dL < BAC ≤ 0.07 g/dL) (severe); and (3) *moderate*, that is, below the WHO recommended legal limit (i.e., 0.00 g/dL < BAC ≤ 0.03 g/dL) (moderate). No driving took place between BAC levels of 0.03 g/dL and 0.05 g/dL. For the severe condition, we defined that participants should be above the WHO recommended BAC legal limit of 0.05 g/dL, since this (or a lower) limit is mandated by 97 countries worldwide [110]. While higher BAC levels would likely increase the expected effects on driving behavior, it would put participant's health at unnecessary risk [63]. For the moderate condition, we relied on past research indicating altered driving behavior already at a BAC of 0.02 g/dL [63].

For each alcohol level, the participants drove for 30 minutes in a research-grade driving simulator. The driving was split across three scenarios (highway, rural, and urban). Driving in each scenario lasted for 10 minutes, separated by breaks of 1–2 minutes for intermediate breath alcohol measurements. During driving, participants were captured by a driver monitoring camera. Participants further had breaks of 1–2 hours between each driving block to reduce the potential effects of drowsiness. Participants received food and non-caffeinated drinks during the study day. The procedure for alcohol administration and driving sessions is shown in Figure 2.

3.4 Alcohol administration and measurement

To obtain specific target levels for BAC, we used established procedures for alcohol administration [47]. Depending on the gender, weight, and age of a participant, we calculated the amount of alcohol to be administered with an updated version of the established Widmark formula [10, 104, 107]. Afterward, participants received the calculated amount of alcohol divided over three mixed drinks (vodka and a non-alcoholic beverage in equal parts) that all had to be consumed within 30 minutes at a steady pace. To achieve comparable conditions for all participants, that is, the same target BAC of 0.07 g/dL when the driving starts, we administered the amount of alcohol for a BAC of 0.08 g/dL and then waited until the BAC dropped to our target of 0.07 g/dL. The additional amount further ensured that physiological differences (e.g., resorption deficits [86]) were

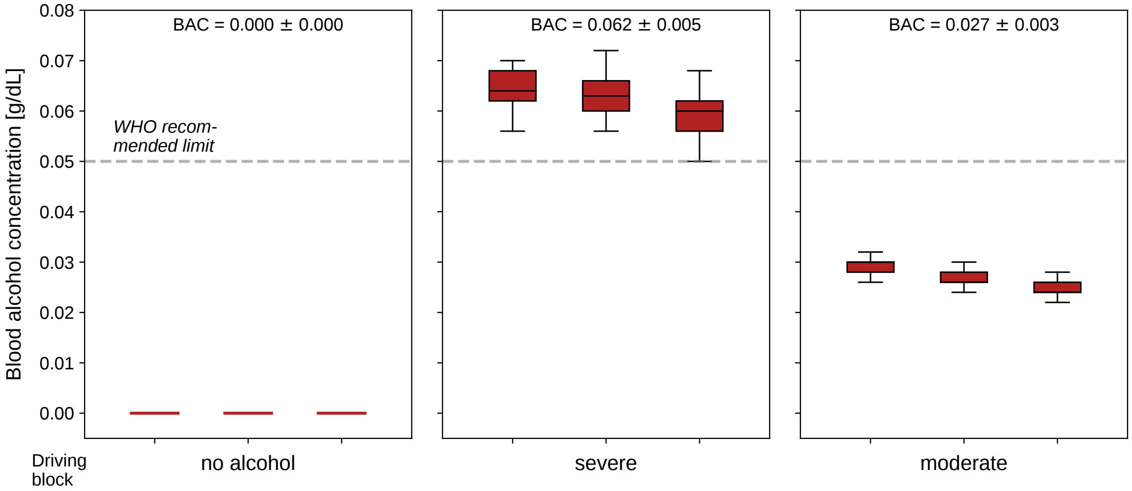


Figure 3: Blood alcohol levels. Observed BAC levels across participants during driving. BAC: blood alcohol concentration; WHO: World Health Organization.

mitigated. Participants were informed before the study that they would consume alcohol during the study, but they were blinded to the total amount and their current BAC values during the procedure.

Alcohol levels of the participants were measured throughout the study with a certified and calibrated breath alcohol measurement device (Dräger Alcotest 6820, Drägerwerk AG & Co KGaA, Lübeck, Germany), that is authorized to be used by law enforcement in Switzerland [25]. We measured the BAC of participants prior to the first driving session (*no alcohol*) and started to measure again 20 minutes after the intake of the last alcoholic beverage (i.e., before the second driving session) to avoid that mouth alcohol can influence the breath alcohol concentration (BrAC) measurement [14]. Measurements were conducted repeatedly (every 2–5 minutes) until two consecutive measurements were at the target BAC of 0.07 g/dL or below. To guarantee that the BAC is in the target range for each driving block, participants conducted BAC measurements prior to each driving scenario. After the second driving block, participants had their first food consumption during a long break in which the study team occasionally measured the BAC (every 2–30 minutes) depending on how close participants were to the target level for the third and last driving session (BAC of 0.03 g/dL). As soon as two consecutive measurements were at or below a BAC of 0.03 g/dL, participants commenced the last driving session.

In each of the three driving blocks (i.e., no alcohol, severe, moderate), participants drove for a total of 30 minutes. The time was split across three scenarios (highway, rural, and urban), each with a duration of 10 minutes and in random order. Before each scenario, the BrAC was measured, which thus resulted in three measurements per driving block. To convert between BrAC and BAC, we used a factor of 0.2 (e.g., a BrAC of 0.35 mg/L corresponds to a BAC of 0.07 g/dL) as defined by national law in Switzerland [12]. The corresponding observations of BAC are reported in Figure 3. In our clinical study, we recorded a BAC of 0.000 ± 0.000 g/dL during the first driving block (no alcohol); 0.062 ± 0.005 g/dL during the second driving block (severe); and 0.027 ± 0.003 g/dL during the third driving block (moderate). Hence, the observed BAC levels are within the intended target ranges.

3.5 Driving simulator

The driving tasks were conducted in a driving simulator (Carnetsoft BV, Groningen, The Netherlands). The simulator offers a realistic setting and is widely used in medical research [49, 97]. The simulator consists of three screens with a field-of-view of 270 degrees, a steering wheel, and pedals for an immersive driver experience. All participants used automatic transmission. The driver's field-of-view mimics that of a real vehicle, including a dashboard with all standard instruments as well as rear-view mirrors. The driving simulator setup is shown in Figure 4.

To cover a variety of driving situations, we implemented three different driving scenarios with distinct characteristics: highway, rural, and urban. These are as follows: (1) The highway scenario comprised of a two-lane highway with one-way traffic. Here, the route was mostly straight with a few wide curves. The speed limit varied between 80 and 120 km/h. Drivers experience varying traffic densities, ranging from free flow to slow-moving traffic. (2) The rural scenario consisted of two-lane rural roads with traffic in both directions and several intersections with and without yield signs. The speed limit was between 60 and 100 km/h. Drivers experienced other traffic participants and had to react to occasional events, such as a stopping bus or slower speeds in front of a school. (3) The urban scenario was used to reflect driving in a city. The route consisted of shorter and narrower roads compared to the two other scenarios. In addition, there were a large number of warning signs, intersections with and without yield or stop signs, and special events, such as pedestrians crossing streets. The speed limit was between 30 and 50 km/h. In all scenarios, variations of the following parameters were randomly assigned within limits: traffic density, behavior of other traffic participants, traffic light circuits, maneuvers of other road users, vehicle types, and traffic at intersections. The order

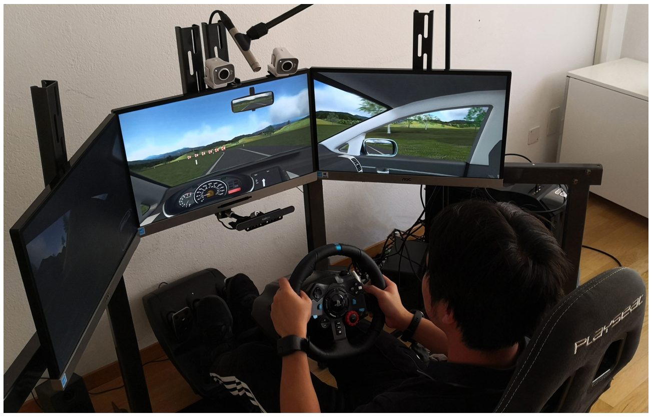


Figure 4: Driving simulator setup. Driving simulator with the driver monitoring camera mounted below the center screen.

of the driving scenarios was randomized for each driver in every driving block. The routes in each driving scenario were intentionally kept the same across all three blocks. To avoid learning effects over time during the experiment itself, people had an extensive training session before the experimental procedure to make them familiar with the routes. In line with existing research, we kept the routes in each driving scenario the same because of two fundamental reasons: (1) familiarity with routes is often an excuse for drunk driving [95], and (2) known routes are often associated with an increased chance of alcohol-related road crashes [13].

Participants were instructed to adhere to local traffic laws (in Switzerland), act as they would in normal road traffic, and make use of all provided vehicle facilities, e.g., turn signal lights. While driving, participants had to follow the guidance of an on-board navigation system.

3.6 Collecting gaze behavior and head movements

During driving, gaze behavior and head movements were recorded using a driver monitoring camera. Here, we used an infrared camera system (Tobii Pro Nano, Tobii AB, Stockholm, Sweden), which was directly mounted below the center screen of the driving simulator. The camera comes with a pre-validated eye tracking algorithm that calculates the gaze positions of drivers as Cartesian coordinates on the center screen with a frequency of 60 Hz. Moreover, this algorithm for eye tracking calculates the current position of the eyes, which we used to infer head movements. At the beginning of each driving day, the eye tracker was calibrated for the participant with respect to the center screen of the driving simulator.

4 MACHINE LEARNING SYSTEM

We developed a novel machine learning system to detect drunk driving using driver monitoring cameras (see Figure 5). Our system proceeds as follows: In the first step, the driver monitoring camera is used to capture information on gaze behavior (velocity, acceleration, fixations, saccades) and head movements. Second, feature engineering is applied using a sliding window approach to train a machine learning model. For training, we varied the underlying classification task (i.e, the label of the prediction), so that two different BAC thresholds are classified as "drunk driving". In the following, we explain the single steps of this system in detail.

4.1 Feature generation

We intentionally chose intelligible features that allow for post hoc explainability in order to interpret the ML model against previous knowledge about pathophysiological mechanisms. In our system, we use three feature groups, which are able to reflect well-known pathophysiological changes due to an alcohol intoxication [58, 63]. The feature groups are: (1) eye movements, (2) gaze events, and (3) head movements. To compute the final features, we first preprocessed the gaze behavior and head movement data into high-frequent signals, which we then summarized into feature vectors for our ML model using sliding windows and statistical aggregation functions. This procedure is explained in the following.

4.1.1 Data pre-processing. We first calculated the following signals from the gaze behavior and the head movements: (1) Eye movements. We calculated the velocity and acceleration of vertical, horizontal, and the combination of both gaze directions. In addition, we used the absolute vertical and horizontal coordinate positions on the screen. (2) Gaze events. We distinguish fixations and saccades. Fixations are time periods in which drivers concentrate their gaze on a certain point or region, whereas saccades are rapid movements of a driver's gaze after and before fixations. To identify both, we applied the REMoDNaV algorithm [22]. The REMoDNaV algorithm calculates the duration for each fixation and saccade. Further, it identifies the amplitude (i.e., distance traveled) as well as the peak and average velocity for each of these gaze events. (3) Head movements. We computed the velocity and acceleration of the head across three dimensions: vertical, horizontal, and depth (i.e., distance to the eye tracker). We further aggregated them into a combined vector.

4.1.2 Sliding window and feature calculation. We used a sliding window approach to split the time-series data into time windows [73]. The window length is subject to an inherent trade-off. On the one hand, long time windows capture more variance and should thus be more informative. On the other hand, short time windows are necessary for near real-time predictions and thus early warnings. Informed by prior literature on detecting driver drowsiness [122], we set the window length to 60 seconds with a shift of 1 second. We perform a sensitivity analysis with other window lengths in Figure 9d) and Appendix D, providing empirical evidence supporting our choice.

Each window was then processed by different aggregation functions in order to map the time-series data onto single features. Here, we used the following mathematical functions (see Appendix B): mean, standard deviation, 0.05 and 0.95 quantiles, skewness, kurtosis, and power (i.e., sum of squares divided by the amount of each signal or event). We intentionally preferred the 0.05 and 0.95 quantiles over minimum and maximum, respectively, as we found the former to be more robust to outliers. For gaze event features, we additionally counted the overall number of fixations and saccades. As a result, we have 56 features for eye movements, 58 features

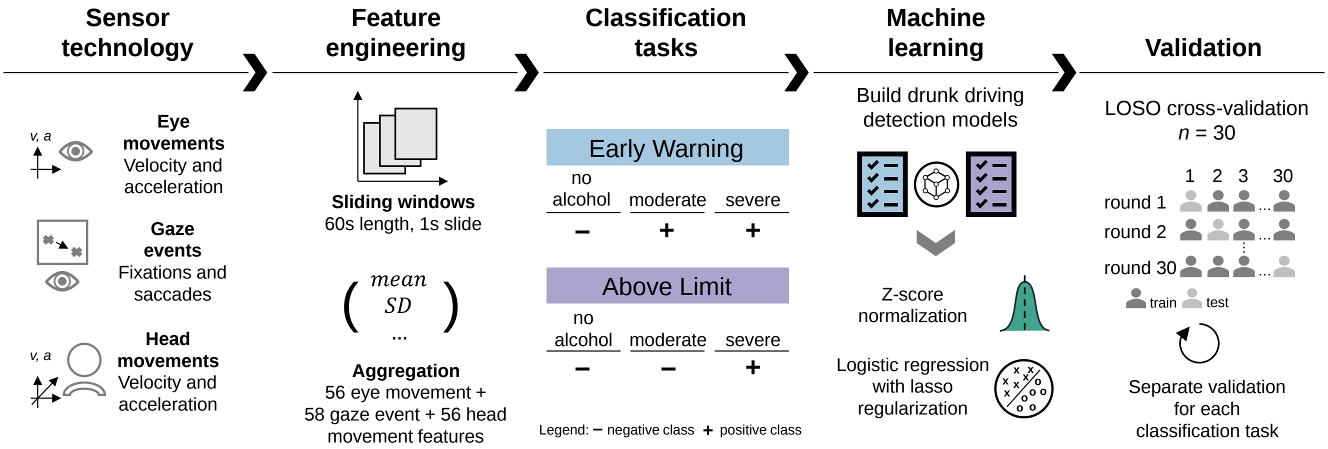


Figure 5: Machine learning system and evaluation. Overview of our machine learning system based on driver monitoring cameras to detect drunk driving at different thresholds. LOSO: leave-one-subject-out; SD: standard deviation.

for gaze events, and 56 features for head movements. In total, this approach led to 151,200 samples (30 subjects × 81 minutes driving [i.e., 90 minutes of driving, but in the first 60 seconds of each driving phase no window was created] × every second a window aggregating 60 seconds).

4.2 Predictive modeling

For ML, we make use of two different classification tasks with different prediction labels. Specifically, both vary in the BAC thresholds that are classified as "drunk driving". Since there is no universal threshold across countries for when driving is forbidden, we base our ML on the recommend legal limit by the WHO, that is, a BAC limit of 0.05 g/dL [111]. Accordingly, the following definitions of labels were used for training and testing: (1) early warning for predicting when the BAC has already reached a moderate BAC level and (2) above the limit for predicting when the BAC is above the WHO recommended legal limit [111] of 0.05 g/dL (severe). We refer to the two classification tasks as Early Warning and Above Limit.

The predictive models for ML were set to logistic regression with lasso regularization (i.e., L1 penalty). This choice was due to two reasons. First, logistic regression with lasso regularization results in parsimonious models, and, therefore, the risk of overfitting is comparatively low. Second, the model allows for straightforward interpretability, which allows us later to compare the model coefficients against prior knowledge on pathophysiological mechanisms.

The ML models were trained using log loss. We used the default implementation in Python 3.8 from the package scikit-learn (version 1.0.2) [72]. We centered and standardized each feature wrt. to the training data using z-score normalization (centering with mean and dividing by SD). Further, we set the class weights to be balanced and left the inverse regularization strength at the default value of 1.0. We tested the robustness of our hyperparameter choice in Appendix D. Overall, our system remained robust to varying choices of lambda and regularization methods (i.e., L1 or L2 penalty).

4.3 Model evaluation

We evaluate the performance of our ML system primarily based on the AUROC [42]. The AUROC has several benefits: it is widely used for classification tasks, considers the complete spectrum of decision thresholds, and accounts for class imbalances [105]. In addition, we report the area under the precision-recall curve (AUPRC). The AUPRC is useful for settings where the positive class is particularly important (here: correct detection of drunk driving). Results for the AUPRC are in Appendix C. We also report further measures, namely balanced accuracy and F1 score (weighted by class). Here, we used a default 0.5 probability for the decision threshold.

To evaluate our ML system, we make use of leave-one-subject-out (LOSO) cross-validation [82]. Accordingly, a model is trained using the data from $n - 1$ participants (i.e., all subsets except one) and then tested on the remaining n-th participant. This procedure is repeated for all n participants. LOSO cross-validation implies one important benefit: In contrast to standard cross-validation (i.e, k-fold within-participant cross-validation), LOSO cross-validation evaluates the generalization capabilities of the ML model to unseen participants [82].

Results are reported as the out-of-sample prediction performance averaged across participants (i.e., macro-average). Further, the standard deviation is reported. This allows us to compare the variability in the performance across participants. Reassuringly, we remind that all hyperparameters are fixed and thus the same across models and participants to ensure generalizability.

4.4 ML baseline based on driving behavior only

We further introduce a baseline based on Controller Area Network (CAN) data (i.e., all vehicle signals but without eye tracking) to evaluate the performance of our ML system. In this evaluation, we apply the same feature processing pipeline as for the camera data. This pipeline also performed well in the past in detecting other driving states using CAN data, for example, detecting distraction [59], emotions [54], low blood glucose levels [50], and even intoxication [47] of drivers. The simulator in our study captured CAN data regarding driver and vehicle behavior with a frequency of 30 Hz. Signals directly reflecting the driver behavior are the steering wheel angle as well as gas and brake pedal positions. For each of these signals, we further calculated the first and second derivative (i.e., velocity and acceleration). The vehicle behavior signals are the latitudinal as well as the longitudinal velocities and accelerations. Finally, we included the lateral position of the vehicle within the lane since related statistical analyses have described it as the most consistent factor in behavior change due to an alcohol intoxication [41]. We applied the same feature generation methods to the signals as for the camera data, that is, the same statistical aggregations and sliding window parameters of 60 seconds every 1 second.

4.5 Post hoc interpretability

To interpret how the ML systems arrive at predictions, we proceeded as follows. In line with existing research [71, 124], we assess the coefficients in our trained models to understand their underlying patterns. Features with a positive coefficient lead to a positive classification (i.e., a larger propensity to classify as drunk), whereas features with a negative coefficient lead to a negative classification (i.e., a lower propensity to classify as drunk). Moreover, as we normalize all features, the absolute size of each feature describes its importance on the output of the model. Hence, the coefficients explain the contribution of each feature to the model output.

5 RESULTS

In this section, we present the results of the evaluation of our machine learning system to detect drunk driving. First, we report the performance metrics of our machine learning system for the two classification tasks Early Warning and Above Limit. Then, we compare our camera-based approach with a CAN baseline, before examining the interpretability of our ML system. Finally, we provide further insights into applicability by analyzing the decision time of our system.

5.1 Performance evaluation

The performance of the ML system for detecting drunk driving is shown in Figure 6. The overall AUROCs for the two classification tasks with different BAC thresholds are 0.88 ± 0.09 (Early Warning) and 0.79 ± 0.10 (Above Limit). The respective SDs show that the performance across drivers is fairly stable. The ML system further achieves a similar performance across different driving scenarios (highway, rural, and urban). For example, the mean AUROC in the Early Warning task varies only between 0.87 (urban) and 0.90 (highway). We observe here and in the following analyses that the prediction performance is better for Early Warning than for Above Limit but only to a small extent.

Confusion matrices comparing the relative frequency of actual alcohol levels against predictions are shown in Figure 7. For Early Warning, both true positives and false negatives are comparatively infrequent, that is, the rate of false alarms and misses is low. A similar pattern is observed for Above Limit prediction. Here, the rate of misses is again comparatively low (18%). The rate of false alarms is also low for when drivers have no alcohol (14%), while false positive tend to be more frequent when drivers have a moderate alcohol yet below the WHO limit (41%), implying that our ML system is sensitive even to driving under little alcohol influence. This is further confirmed when our ML model predicts each driving state separately. Here, we see that our model achieves a 70% true positive (TP) rate for detecting the no alcohol driving state. Moreover, the two alcohol driving states have also a high TP rate with 45% (moderate) and 55% (severe), respectively, but with 30% confusion between both intoxication states.

5.2 Comparison of CAN vs. camera approaches

Here, we compare our proposed camera-only approach against a CAN baseline (i.e., vehicle signals only without camera data) and an approach combining both data sources (i.e., vehicle signals and camera data). The results are shown in Figure 8. In direct comparison to a baseline using CAN-only, our camera-only approach performs consistently better. For example, the CAN-only baseline achieves only an AUROC of 0.74 ± 0.10 (for Early Warning) and 0.66 ± 0.12 (for Above Limit), and thus has an AUROC that is lower than that of the camera-only approach by around 0.10. In line with this, the other performance metrics are also inferior for the CAN-only approach. For example, our camera-based ML system records an AUPRC of 0.93 ± 0.05 (for Early Warning) and 0.65 ± 0.16 (for Above Limit). In contrast, the CAN-only baseline reaches an AUPRC of only 0.84 ± 0.07 (for Early Warning) and 0.50 ± 0.15 (for Above Limit). Furthermore, the camera-only ML system achieves a balanced accuracy of 0.76 ± 0.10 (for Early Warning) and 0.68 ± 0.10 (for Above Limit), whereas the CAN-only baseline results in a balanced accuracy of only 0.65 ± 0.08 (for Early Warning) and 0.60 ± 0.09 (for Above Limit). Comparable results are achieved for the F1 score, where the camera-only system achieves 0.75 ± 0.14 (for Early Warning) and 0.67 ± 0.12 (for Above Limit). Again, the CAN-only baseline is substantially outperformed. The latter registers an F1 score of 0.64 ± 0.11 (for Early Warning) and 0.60 ± 0.08 (for Above Limit).

The camera-only approach achieves a performance similar to that of an approach combining both CAN and camera data. As an example, the ML system combining both data sources achieves an AUROC of 0.91 ± 0.07 (for Early Warning) and 0.81 ± 0.11 (for Above Limit). In comparison, the camera-only ML system records 0.88 ± 0.09 (for Early Warning) and 0.79 ± 0.10 (for Above Limit). In sum, using CAN instead of camera signals is inferior and combining both CAN and camera signals does not lead to a statistically significant improvement. As such, the results corroborate the relevance of driver monitoring cameras to detect drunk driving.

5.3 Robustness checks

Additional sensitivity analyses were performed to evaluate the robustness of the ML system (see Figure 9 and Appendix D). First, we evaluated how our ML system performs on unseen driving scenarios. Therefore, we introduced a leave-one-driving-scenario-out cross-validation on top of our LOSO cross-validation. More specifically, we performed three evaluations for each participant: each driving scenario was omitted once in training and evaluation was based on the left-out driving scenario. Our ML system achieves similar results as with the LOSO cross-validation alone, showing

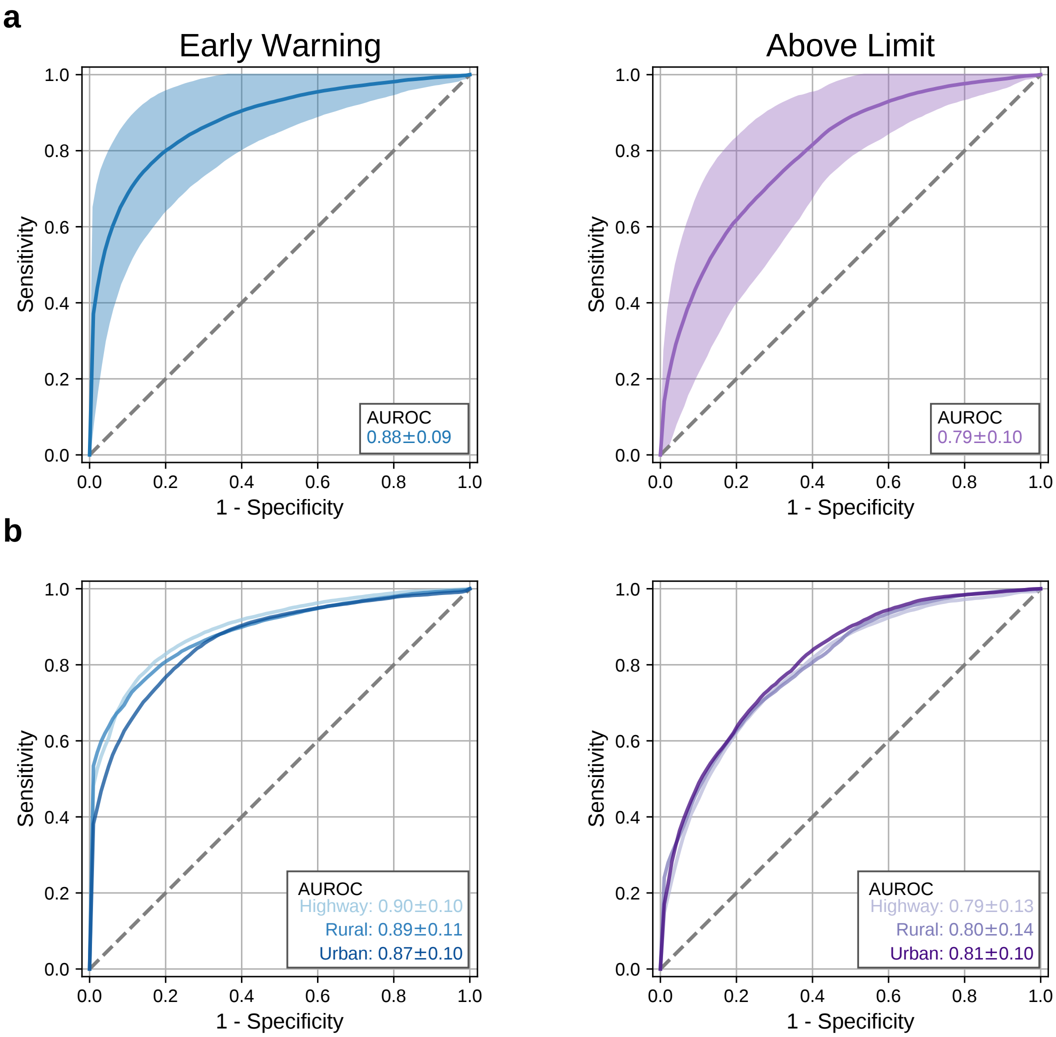


Figure 6: Performance of drunk driving detection. The machine learning system for detecting drunk driving is evaluated based on the area under the receiver operating characteristic curve (AUROC). (a) Performance across participants for different BAC thresholds. (b) Performance by driving scenario (i.e., highway, rural, and urban). The dashed, gray line shows an AUROC of 0.50 as a naïve baseline (i.e., a random guess). BAC: blood alcohol concentration.

that our system works independent of driver and scenario. Second, we varied the length of the sliding window for feature engineering. Overall, the results of our ML system remain robust. We observe a tendency that longer sliding windows are associated with a larger AUROC. However, informed by literature on detecting other critical driving events [122], we set the window length to 60 seconds in the above analyses, as it allows for timely feedback. Second, we compared the predictive power of the different features from eye movements, gaze events, and head movements. Here, we observe a larger AUROC for gaze events (i.e., fixations and saccades), followed by eye movements (i.e., velocity and acceleration) and head movements (i.e., velocity and acceleration). Across all, the mean AUROC remains consistently above 0.70. Third, we repeated the analysis with alternative ML models. We used both linear models (logistic regression with lasso, ridge, and elastic net regularization) and non-linear models (support vector machine, random forest, gradient boosting model, and multi-layer perceptron). Overall, the logistic regression with lasso regularization performs best and thus justifies our model choice. Nevertheless, all models achieved a mean AUROC of 0.73 or better, which corroborates the robustness of our prediction.

5.4 Interpretability

To explain the decision logic in the ML model, we interpret the coefficients and thereby assess how features are associated with the predictions. First, we focus on the absolute magnitude of the

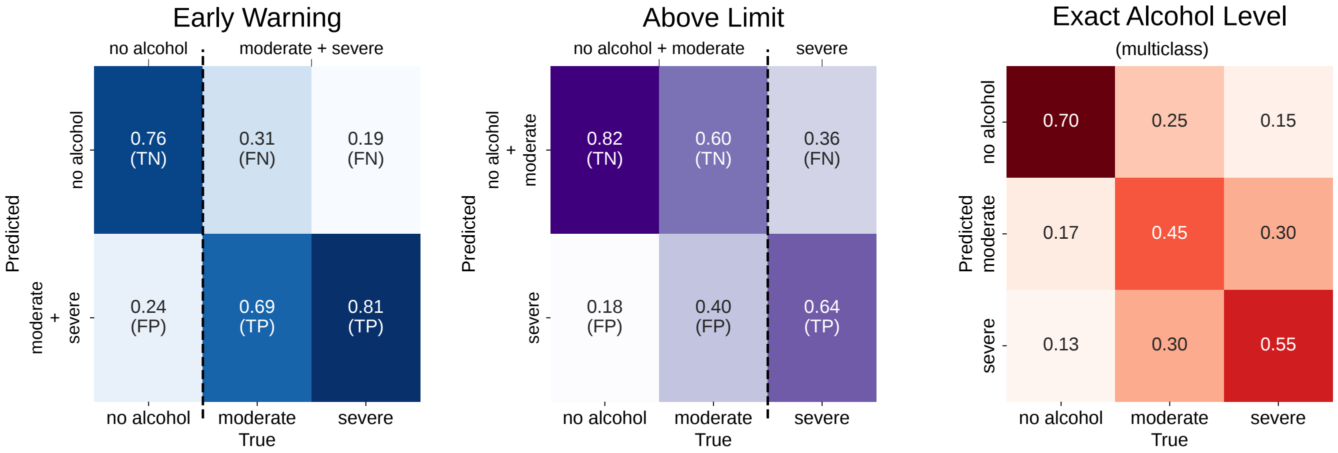


Figure 7: Confusion matrices for drunk driver detection. The confusion matrices compare the relative frequency of the actual alcohol level (horizontal) vs. the predicted alcohol level (vertical). For comprehensiveness, we further report a granular breakdown for the actual alcohol level (horizontal) comparing no alcohol, moderate, and severe intoxication separately. The cells correspond to FN: false negative; FP: false positive; TN: true negative; TP: true positive. In addition, we provide a confusion matrix when our machine learning system is trained and evaluated on predicting each driving block/alcohol level separately.

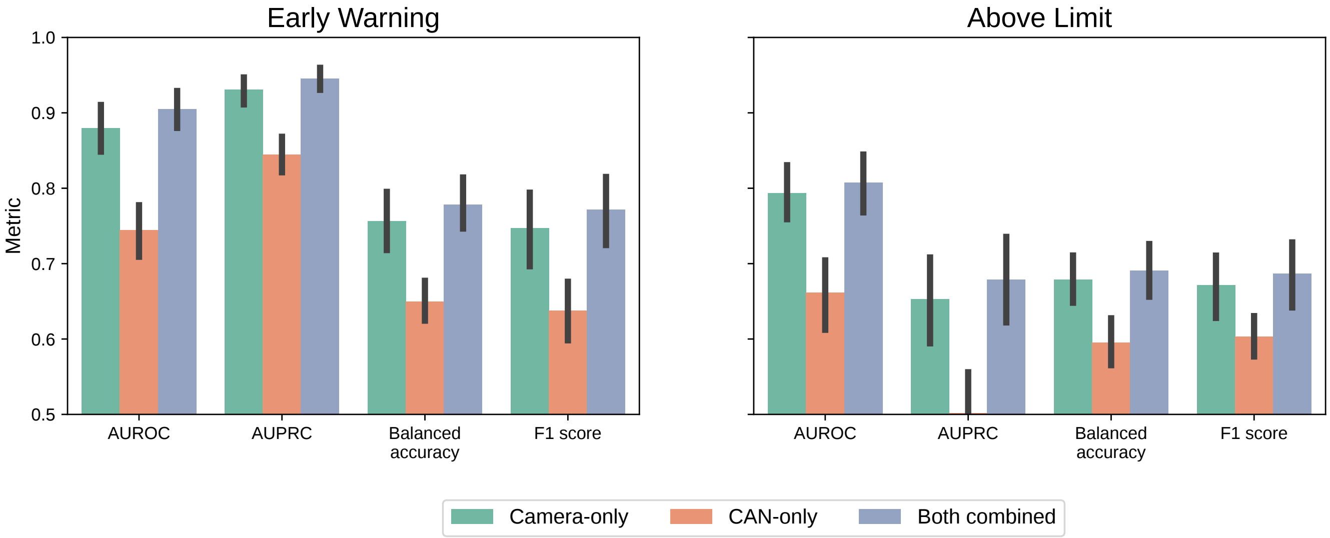


Figure 8: Comparison of CAN vs. camera signals for detecting drunk driving. We report the performance of our machine learning system while using different data sources: camera-only (our main system), CAN-only, and both combined. Overall, the results demonstrate the relevance of driver monitoring cameras. The following performance metrics are computed: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), balanced accuracy, and F1 score (weighted by classes). Reported: mean ± standard deviation. CAN: Controller Area Network (i.e., vehicle signals).

coefficients (Figure 10a), which allows us to identify features that are important for the ML model. Here, we group features by eye movements (velocity, acceleration), gaze events (fixations, saccades), and head movements (velocity, acceleration). Overall, features from gaze events receive coefficients with larger absolute value than the other feature groups and are thus highly important for the ML. Moreover, for detecting an alcohol intoxication, eye movements are more important than head movements. We also evaluated each feature group separately for their predictive performance (see Appendix C).

The model coefficients are shown in Figure 10b. For example, the ML model for Above Limit is ceteris paribus more likely to

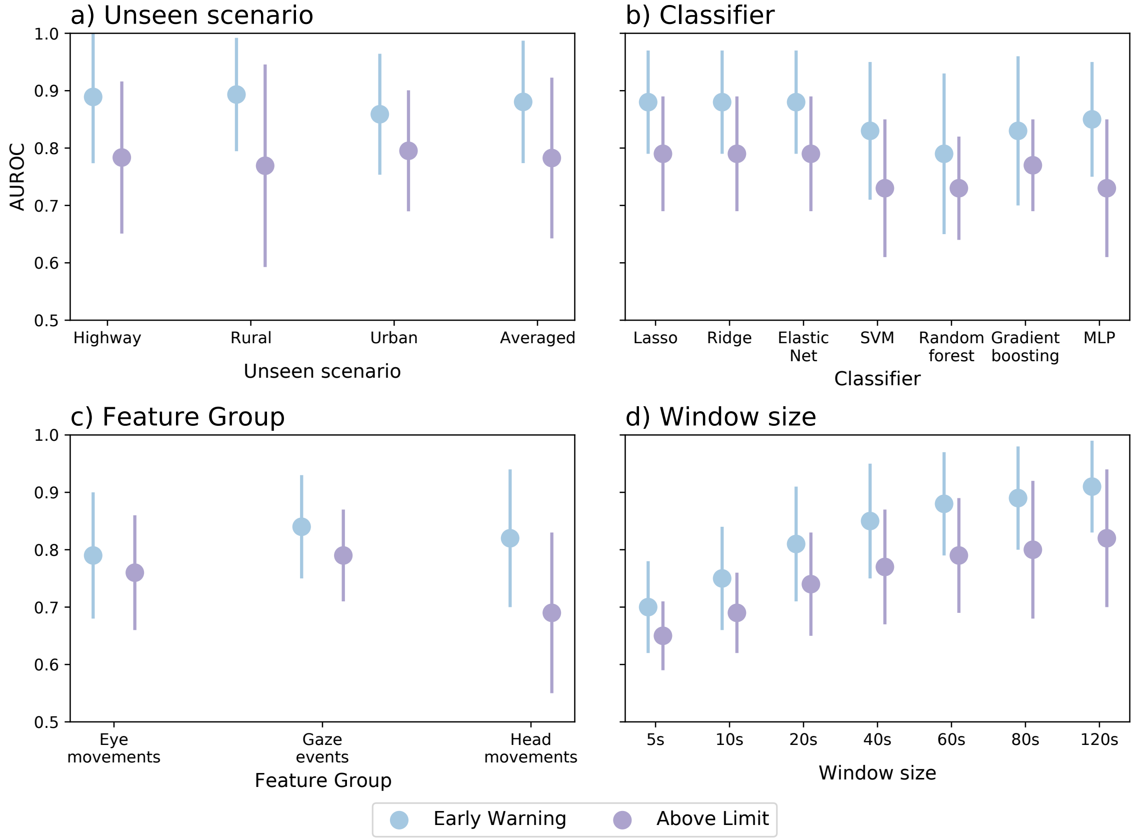


Figure 9: Sensitivity analysis for drunk driver detection. The machine learning system for detecting drunk driving is evaluated based on the area under the receiver operating characteristic curve (AUROC) while varying different components of our ML system: (a) leaving one scenario in training out and testing on it (i.e., leave-one-subject and leave-one-scenario out cross validation), (b) the classifier, (c) the features, and (d) the size of the sliding window. Across all configurations, the machine learning system has a robust, high prediction performance. MLP: multi-layer perceptron; SVM: support vector machine.

predict drunk driving when drivers have a longer mean duration of fixations and saccades and a shorter mean amplitude (i.e., distance traveled) of saccades.

5.5 Decision time

Decision time plays a critical role in achieving timely as well as reliable decisions for our system before interventions are delivered. To examine the decision time, we computed the balanced accuracy for each second of a trip. To do this, we applied a rolling cumulative moving average to the predicted probability of each window in each separate trip across all drivers. The predictions reach a high plateau early on. After 90 seconds of driving, this approach already achieves a balanced accuracy with 95% confidence interval (CI) of 0.77 [0.72, 0.82] (Early Warning) and 0.69 [0.64, 0.75] (Above Limit). The decision frequency of our system is every second, and, hence, a filter is needed to prevent a potentially volatile decision behavior and assure stable and reliable intervention delivery. Therefore, we evaluated a non-overlapping majority vote for the predicted windows (such as in [62]). The prediction performance of our ML system reaches a peak at the aggregation of 150 windows with an AUROC of 0.91 ± 0.08 (Early Warning) and 0.85 ± 0.11 (Above Limit). Both analyses show that our ML system provides timely and reliable decision after only a few minutes.

6 DISCUSSION

6.1 Contributions

Alcohol consumption is responsible for a major share of the global disease burden and overall mortality [109, 110]. There is thus a need for scalable and cost-effective HCI technology towards behavioral change with the aim of reducing alcohol-related harms. Here, we developed and evaluated a novel machine learning system for drunk driver detection based on driver monitoring cameras. To the best of our knowledge, our system is the first to detect drunk driving from camera-based sensor technology.

Our system achieves a high detection performance with an AUROC of 0.88 (for detecting driving with a BAC > 0.00 g/dL and ≤ 0.03 g/dL) and 0.79 (for detecting driving above the WHO

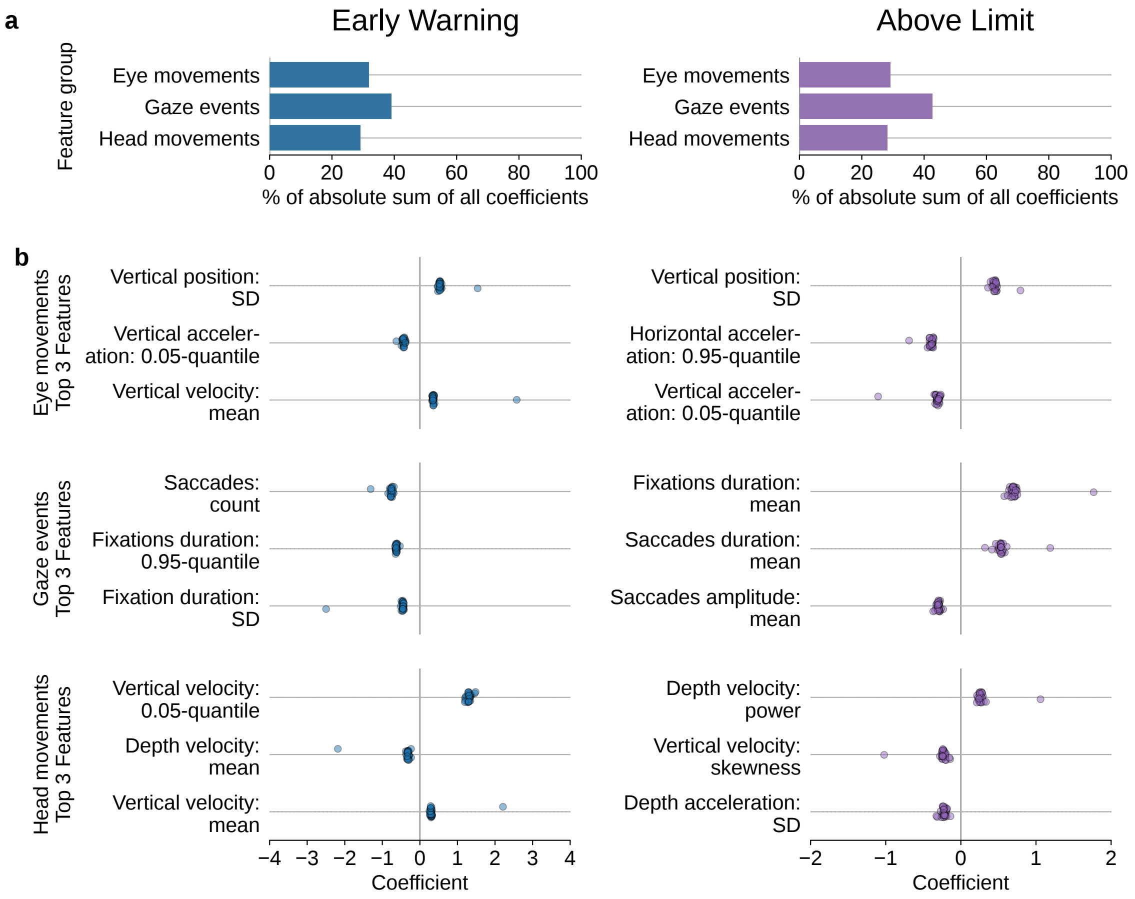


Figure 10: Interpretability for machine learning. The coefficients are analyzed in order to examine how the logistic regression arrives at predictions. (a) The importance of different feature groups (eye movements, gaze events, and head movements) is compared. Formally, we sum over the absolute values of the coefficients in each group and then normalize them to one. This yields the % of the sum of the coefficients (absolute value) per feature group and thus quantifies the relative importance. (b) The regression coefficients of the top-3 features in each feature group are reported. Separate dots ($n = 30$) are shown for the coefficients from the different splits during cross-validation. SD: standard deviation.

recommended legal limit of 0.05 g/dL) while testing with a leave-one-subject-out cross-validation. Moreover, we achieve similar performances when applying a leave-one-subject-out and leave-one-driving-scenario-out cross-validation with an AUROC of 0.82 (BAC > 0.00 g/dL and ≤ 0.03 g/dL) and 0.78 (above the WHO recommended legal limit of 0.05 g/dL). Hence, the results demonstrate that our system can even provide early warnings when there is only a moderate alcohol intoxication.

Our ML system relies upon driver monitoring cameras as input. Driver monitoring cameras are nowadays common in modern vehicles as part of driver assistance systems. Moreover, driver monitoring cameras are to become mandatory for all new vehicles due to safety regulations. Examples of such regulations are the Euro NCAP and the EU GSR, which make driver monitoring cameras mandatory from 2024 onwards [26, 27]. Notwithstanding, other technologies such as breath-based sensors [123] allow for the detection of drunk driving. However, the current state of such technologies is expensive and requires regular maintenance [77, 101]. In contrast to that, the growing availability of driver monitoring cameras makes them a scalable, low-cost, and easily accessible technology.

6.2 Comparison with previous work

Previous works on detecting drunk driving are based on driving behavior (e.g., steering, pedal usage, vehicle speed) [19, 20, 35, 36, 47, 48, 51, 52, 79, 92]. However, to the best of our knowledge, no work has so far developed or evaluated a system based on driver cameras. This is our novelty.

Here, we propose to shift from driving behavior to driver behavior and, specifically, to leverage eye gaze and head movements. Our choice has important benefits. (1) Visual and perceptual impairments due to alcohol already occur at a BAC of 0.005 g/dL. In comparison, changes in vehicle control occur only for a much

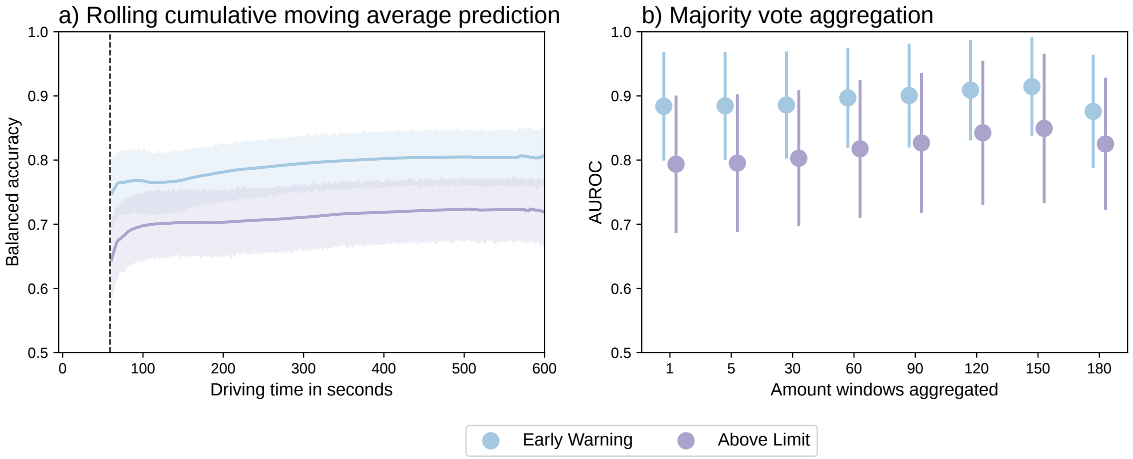


Figure 11: Insights into the decision-making time for the machine learning system for drunk driving detection. Drunk driving detection needs fast and stable decision-making, thus the decision time and frequency is evaluated. (a) Decision time and the performance when applying a cumulative moving average to the predicted probability of each window along the driving time. (b) Decision time and the performance when applying a non-overlapping majority vote aggregation of the predicted windows. AUROC: area under the receiver operating characteristic curve.

larger BACs [41, 63]. This thus makes predictions from eye gaze particularly effective for early warnings of drunk driving. (2) Driver monitoring cameras have recently been applied to a related task, namely detecting driver drowsiness [16, 23]. Here, predictive gains have been achieved by moving from driving behavior to driver behavior, analogous to our work. (3) Driving behavior (e.g., steering, pedal usage, vehicle speed) is subject to large variability across driving scenarios (highway, rural, and urban) and across drivers (e.g., slow vs. fast drivers, defensive vs. aggressive drivers) and thus overlays the raw signals with noise [47].

Our work shows that the predictive power of eye gaze and head movements is robust and can reliably generalizes across participants and to unseen driving scenarios. We found that the camera-based approach clearly outperforms a CAN-only baseline as used by related work (e.g., [47]). For example, the CAN-only baseline achieves only an AUROC of 0.74 ± 0.10 (for Early Warning) and 0.66 ± 0.12 (for Above Limit), and thus has an AUROC that is lower than that of the camera-only approach by around 0.10. Combining both camera and CAN, we achieve similar performances. This underlines that the camera information are the major factor in detecting drunk driving. Moreover, this hypothesis is confirmed when we compare our prediction performance with different previous studies. For example, in one study, the US National Highway Traffic Safety Administration (NHTSA) examined the extent to which driving below and above a BAC of 0.08 g/dL (the legal limit in the US) can be detected [47]. Despite the higher alcohol threshold, we can put into context the results with our study because the study design, driving task, and analysis are comparable to ours. For the same driving scenarios, the work based on driving behavior (e.g., steering, pedal usage) achieves an AUROC of 0.77 ± 0.08. Thereby, data from all drivers in sober and intoxicated states is used. For comparison, our system achieves a better prediction performance for a lower BAC threshold of 0.05 g/dL (the WHO recommended limit) with an AUROC of 0.79 ± 0.10 validated out-of-subject. In essence, this demonstrates the effectiveness of using driver monitoring cameras for detecting drunk driving.

6.3 Interpretations of machine learning model in relation to pathophysiology

The post hoc interpretation of our models is in line with known pathophysiological effects that describe impairments induced by alcohol. Examples of such pathophysiological effects are divided attention, lower reaction times, changes in vigilance, tracking, perception, and psychomotor functions [63]. As we see in our analysis, these pathophysiological effects are especially pronounced for gaze events. Our system learns to recognize that, with increased intoxication, the mean duration of fixations increases and fixation frequencies decrease. This is consistent with pathophysiological effects according to which people under the influence of alcohol take longer time to process visual information [58, 63]. Moreover, we observe changes in saccadic eye movements (e.g., fewer and longer saccades). This can be explained by prior literature on visual scanning [31, 85, 87]. Specifically, drunk drivers tend to follow scattered and irregular patterns of visual sampling, which is reflected by changes in saccade velocities and amplitudes [31, 85, 87]. In sum, the observed changes in fixations and saccades are consistent with existing research and well-described phenomena such as tunnel vision [61, 63].

6.4 Time until a decision

Most alcohol-related road crashes take place on free, straight roads (i.e., out of city, higher speed, and no curves) [90]; therefore, our system needs to make a reliable decision before drivers reaches such roads. In our evaluation (see Figure 11), we find that our system achieves a confident decision already after around 90 seconds (balanced accuracy with 95% CI of 0.77 [0.72, 0.82] (Early Warning) and 0.69 [0.64, 0.75] (Above Limit)). Moreover, a majority vote aggregation would further increase the reliability. The prediction performance of our ML system reaches a peak at the aggregation of 150 windows with an AUROC of 0.91 ± 0.08 (Early Warning) and 0.85 ± 0.11 (Above Limit). Detecting the alcohol level before driving would be the safest and most reliable way in preventing drunk driving. However, current technology such as alcohol ignition locks are highly expensive and require regular maintenance so that a widespread application of this preventive technology is highly unlikely [77, 101]. Given the number of alcohol-related driving incidents, waiting for the highly effective and affordable "silver bullet" to prevent drunk driving seems rather unethical. Instead of insisting on a perfect solution and waiting for years until manufacturers and/or regulators decide on a sensor technology that has yet to be developed and evaluated, smaller steps can be taken as of today. Our approach relies on existing technologies of modern vehicles, which allows a swift introduction of measures against the problem of drunk driving.

6.5 Limitations

The strength of this study is the rigorous development and evaluation of a ML system for the detection of drunk driving using driver monitoring cameras. For this, we performed an interventional study based on a standardized procedure for alcohol administration [24, 55, 113] and, further, followed best practice in ML [37]. Nevertheless, this study has limitations.

Environment and driving-related factors. First, as with all simulator studies, there may be risk of latent effects, such as learning or drowsiness. However, we follow best practice for simulator studies to prevent such latent effects to the extent possible [24, 55, 113]. Specifically, we mitigated learning effects by giving participants sufficient time for practicing before the start of the driving tasks and reduced drowsiness by introducing comparatively long breaks between each driving block. Second, our study examined driving in a simulator rather than in a real vehicle. In real-world driving, drivers may adapt their visual activity to the environment (e.g., at night or in rain) or perform secondary tasks while driving (e.g., talking to passengers or making phone calls). Real-world driving studies are recommended to address these issues, and we discuss how these issues should be addressed in future work (see Section 6.7). However, previous work suggests that simulators reliably reproduce changes in driver behavior under the influence of alcohol [38, 41]. By covering a wide range of scenarios and traffic situations, we demonstrated the generalizability of our approach to different driving situations. This is underlined by the stable results of the leave-one-driving-scenario-out cross-validation. Finally, we also want to highlight that our results rely on camera technology that produces reliable results even in demanding situations. Our system is based on an infrared camera (as also used in industry [8]). These cameras work in various light conditions and still provide reliable results even in situations with strong light changes, night driving or even when drivers wear (sun)glasses.

Individual driver factors. In our study, we were able to show that our system is able to have a high detection accuracy for a demanding leave-one-subject-out cross-validation across different age groups and genders. Therefore, our system is capable of handling individual tolerance, at least for our study population, which consisted of healthy individuals from Switzerland with regular alcohol consumption. We see two potential barriers why our system performance could potentially deteriorate when applying it to other populations. First, due to individual alcohol tolerances, the effect of alcohol levels can potentially vary from person to person [57]. For example, different ethnicities have a different sensitivity to alcohol [108]. However, the WHO recommendation of a BAC of 0.05 g/dL as the legal limit for driving under the influence is chosen to reflect in which the majority of people show signs of impaired driving [28]. Even experienced drinkers show impairments in their driving behavior as early as a BAC of 0.02 g/dL [63]. Therefore, our current system should already cover the problem of individual tolerances even for non-included healthy populations (e.g., a different ethnicity). Second, our system may make incorrect classifications for populations with different visual scanning behaviors. Most likely, the underlying gaze detection algorithm of our driver monitoring system will fail for individuals with health problems that affect their eye and pupil movements, such as strabismus or nystagmus. Very young or very old drivers might also challenge our system. Young drivers have less situation-aware visual activity due to their limited driving experience [76], whereas older people have slower visual processing times than other age groups [70]. Additional experiments are needed for these two populations. However, specifically in the case of elderly people, our system could create value by identifying fundamental and safety-critical changes in behavior. In the following, we will discuss possible interventions to support drivers.

6.6 Implications

The accurate detection of moderate alcohol levels by our system is an important prerequisite for providing effective digital interventions to prevent alcohol-related harm [33, 64, 65]. This is especially relevant as people regularly fail to correctly self-assess their alcohol levels [3, 81]. Drivers consistently underestimate their alcohol intoxication and therefore overestimate their ability to drive. Here, our system could be used to trigger behavioral interventions. One example is drunk driving warnings which promote transparency similar to self-tracking and thus train people in BAC discrimination [4] and drinking control strategies [91].

Figure 12 illustrates a potential future warning intervention. We envision a comprehensive driver warning system that conducts a fit to drive assessment. Such a system could address drunk driving but also other impairments, for example, drowsiness or lack of attention. In addition, driving under the influence of cannabis, ecstasy, or other illicit drugs could be also included in the future as they are known to impair the driving behavior as well [23, 46, 78]. Improved transparency has been shown to be a key driver of behavioral change in the context of self-tracking [4, 33, 118]. Accordingly, the system could inform the driver of the observed behavior that led to the warning. In the example of Figure 12, the machine learning system identified longer fixation times and thus slower information processing of the driver as the reason for the warning.

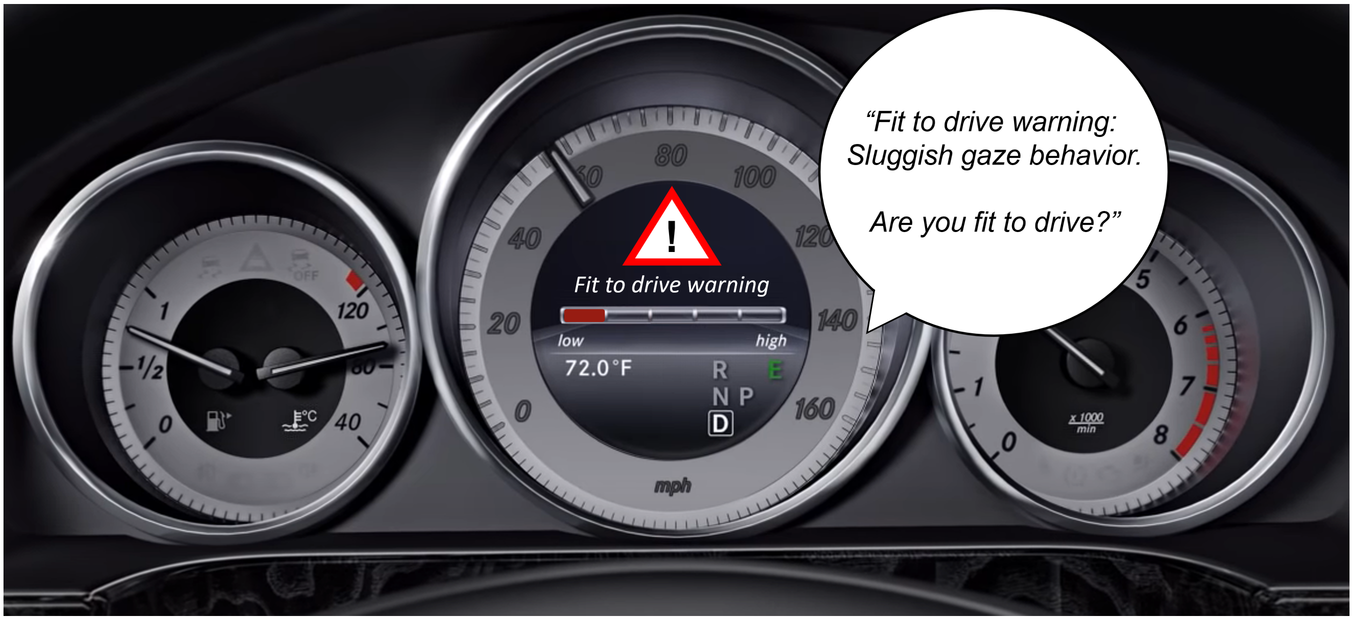


Figure 12: Example for a fit to drive intervention system. The system warns the driver with a visual- and audio-based intervention including a reason for the warning.

Beyond warnings, there are also more restrictive interventions for escalation. For example, modern vehicles may limit the maximum allowed speed, increase sensitivity to emergency braking, activate dedicated safety systems, increase the assistance provided by (partially) autonomous systems, or even force a full standstill for safety [27, 32]. A drunk driving detection could also be useful for existing digital intervention solutions outside of the vehicle, such as *Drink Less*, *Daybreak*, or *SoberDiary* [34, 94, 118]. Once informed by our system, such digital interventions could react on the detected drunk driving events and foster alcohol behavior change beyond the driving context.

6.7 Roadmap to implementation

Our drunk driving detection system can be easily integrated into existing camera-based systems for monitoring driving states such as drowsiness and distraction. These camera systems have the same input data, such as gaze positions, that our system requires, and therefore our system could be directly added as a simple software component (e.g., in [8, 88]).

To bring our system to market, we recommend a few calibration and evaluation steps before our drunk driving detection system is used in real vehicles. First, the current simulator study should be replicated with sober and drunk drivers in a real vehicle. Since driving under the influence of alcohol is a criminal offense in almost all countries (even in the context of a study), the most likeliest way would be to conduct such a study on a test track with a driving instructor next to the driver. Second, our system should prove its ability in everyday traffic. Therefore, we recommend collecting everyday driving data of non-impaired drivers on open roads in a vehicle equipped with a driver monitoring system. This represents a comparably small effort to a drunk driving study as less ethical concerns exist and no clinical trial is needed. The drunk driving study would allow to validate the detection performance of our system in real vehicles, while the open-road study can be used to calibrate against overly sensitive warnings (i.e., false alarms) that are a key threat for long-term adoption and might occur due to the increased diversity of driving conditions in a natural environment.

We also see clear advantages in embedding our drunk driving detection into an existing driver monitoring system on top of existing algorithms for detecting drowsiness or distraction. For example, our drunk driving detection algorithm is validated on the basis of rather undisturbed and focused driving. A distraction detection algorithm can filter out distracted driving to prevent potential erroneous classifications of our detection algorithm. Moreover, having an ensemble of driver monitoring systems for various impairments may further add to the reliability of the overall system.

7 CONCLUSION

To the best of our knowledge, our system is the first to detect drunk driving from camera-based sensor technology. Thereby, we directly address needs in practice for the HCI community: Policy initiatives and regulations around the world (such as in the US [96]) call for new drunk driving detection technologies and interventions and thereby reduce alcohol-related harm. Even though progress has been made toward fully autonomous driving, experts agree that autonomous driving will not be widely available in the next two decades [6, 69]. Hence, for the coming years, detection systems are needed that build upon existing HCI technologies in vehicles that leverage driver vehicle and environment interaction. Here, a cost-effective and scalable approach is offered by our novel machine learning system that uses existing driver monitoring cameras.

To this end, our system provides new opportunities for digital interventions to reduce alcohol-related harms, particularly traffic fatalities.

Data availability

The following procedure is required by our local ethics committee. Any requests for raw data (i.e., blood alcohol concentrations, camera data, driving data, de-identified patient characteristics) will be reviewed by the scientific study board leading the involved research group. Only applications for non-commercial use will be considered and should be sent to the corresponding author. Applications should outline the purpose for the data transfer. Any data that can be shared will need approval from the scientific study board and a Material Transfer Agreement in place. All data shared will be de-identified.

ACKNOWLEDGMENTS

This research was funded and supported by the Bosch IoT Lab at the University of St. Gallen and ETH Zurich. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors. M. Maritsch acknowledges funding from the Swiss National Science Foundation (183569). E.v. Weenen acknowledges funding from the Swiss National Science Foundation (186932). The funding bodies had no control over design, conduct, data, analysis, review, reporting, or interpretation of the research conducted.

REFERENCES

[1] Shabnam Abtahi, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariri. 2014. YawDD: A yawning detection dataset. In *Proceedings of the Multimedia Systems Conference*. 24–28.

[2] Affectiva. 2022. Affectiva Automotive AI. http://go.affectiva.com/auto Accessed: 08/11/2022.

[3] Kirstin Aschbacher, Christian S. Hendershot, Geoffrey Tison, Judith A. Hahn, Robert Avram, Jeffrey E. Olgin, and Gregory M. Marcus. 2021. Machine learning prediction of blood alcohol concentration: a digital signature of smart-breathalyzer behavior. *npj Digital Medicine* 4, 1 (2021), 74. https://doi.org/10.1038/s41746-021-00441-4

[4] Elizabeth R. Aston and Anthony Liguori. 2013. Self-estimation of blood alcohol concentration: A review. *Addictive Behaviors* 38, 4 (2013), 1944–1951. https://doi.org/10.1016/j.addbeh.2012.12.017

[5] Sangwon Bae, Tammy Chung, Denzil Ferreira, Anind K. Dey, and Brian Suffoletto. 2018. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive Behaviors* 83 (2018), 42–47. https://doi.org/10.1016/j.addbeh.2017.11.039

[6] Prateek Bansal and Kara M. Kockelman. 2017. Forecasting Americans' long-term adoption of connected and autonomous vehicle technologies. *Transportation Research Part A: Policy and Practice* 95 (2017), 49–63.

[7] Bosch Mobility Solutions. 2022. Driver Drowsiness Detection. https://www.bosch-mobility-solutions.com/en/solutions/interior/driver-drowsiness-detection/ Accessed: 08/11/2022.

[8] Bosch Mobility Solutions. 2022. Interior Monitoring Systems. https://www.bosch-mobility-solutions.com/en/solutions/interior/interior-monitoring-systems/ Accessed: 11/12/2022.

[9] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* 25, 11 (2000), 120–123.

[10] John Brick. 2006. Standardization of alcohol calculations in research. *Alcoholism: Clinical and Experimental Research* 30, 8 (2006), 1276–1287.

[11] Timothy Brown, John D. Lee, Chris Schwarz, Dary Fiorentino, and Anthony McDonald. 2014. *Assessing the feasibility of vehicle-based sensors to detect drowsy driving.* Report DOT HS 811 886.

[12] Bundesversammlung. 2012. Verordnung der Bundesversammlung über Alkoholgrenzwerte im Strassenverkehr. *Systematische Rechtssammlung* 741.13 (2012).

[13] Bridget R. D. Burdett, Nicola J. Starkey, and Samuel G. Charlton. 2017. The Close to Home Effect in Road Crashes. *Safety Science* 98 (2017), 1–8. https://doi.org/10.1016/j.ssci.2017.04.009

[14] Glenn R. Caddy, Mark B. Sobell, and Linda C. Sobell. 1978. Alcohol breath tests: Criterion times for avoiding contamination by "mouth alcohol". *Behavior Research Methods & Instrumentation* 10, 6 (1978), 814–818.

[15] Alimed Celecia, Karla Figueiredo, Marley Vellasco, and René González. 2020. A portable fuzzy driver drowsiness estimation system. *Sensors* 20, 15 (2020), 4093.

[16] Mario I. Chacon-Murguia and Claudia Prieto-Resendiz. 2015. Detecting Driver Drowsiness: A survey of system designs and technology. *IEEE Consumer Electronics Magazine* 4, 4 (2015), 107–119.

[17] Robert Chen-Hao Chang, Chia-Yu Wang, Hsin-Han Li, and Cheng-Di Chiu. 2021. Drunk Driving Detection Using Two-Stage Deep Neural Network. *IEEE Access* 9 (2021), 116564–116571.

[18] Ipshita Chatterjee and Apoorva Sharma. 2018. Driving fitness detection: A holistic approach for prevention of drowsy and drunk driving using computer vision techniques. In *IEEE South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA CECNSM)*. IEEE, 1–6.

[19] Huiqin Chen and Lei Chen. 2017. Support vector machine classification of drunk driving behaviour. *International Journal of Environmental Research and Public Health* 14, 1 (2017), 108.

[20] Huiqin Chen, Xiuhong Yuan, Hongxian Ye, Lei Chen, and Guanjun Zhang. 2019. The effect of alcohol on the physiological performance of the driver. *International Journal of Crashworthiness* 24, 6 (2019), 656–663.

[21] Abdelkader Dairi, Fouzi Harrou, and Ying Sun. 2022. Efficient Driver Drunk Detection by Sensors: A Manifold Learning-Based Anomaly Detector. *IEEE Access* (2022).

[22] Asim H. Dar, Adina S. Wagner, and Michael Hanke. 2021. REMoDNaV: robust eye-movement classification for dynamic stimulation. *Behavior Research Methods* 53, 1 (2021), 399–414.

[23] M. Doudou, A. Bouabdallah, and V. Berge-Cherfaoui. 2020. Driver drowsiness measurement technologies: Current research, market Solutions, and challenges. *International Journal of Intelligent Transportation Systems Research* 18, 2 (2020), 297–319. https://doi.org/10.1007/s13177-019-00199-w

[24] Luke A. Downey, Rebecca King, Katherine Papafotiou, Phillip Swann, Edward Ogden, Martin Boorman, and Con Stough. 2013. The effects of cannabis and alcohol on simulated driving: Influences of dose and experience. *Accident Analysis & Prevention* 50 (2013), 879–886. https://doi.org/10.1016/j.aap.2012.07.016

[25] Eidgenössisches Institut für Metrologie. 2016. Verordnung des EJPD über Atemalkoholmessmittel (AAMV). *Amtliche Sammlung* AS 2016 2841 (2016).

[26] Euro NCAP. 2017. *Euro NCAP 2025 Roadmap.* Report.

[27] European Parliament and Council of the European Union. 2019. Regulation (EU) 2019/2144. *Official Journal of the European Union* L 325 (2019). http://data.europa.eu/eli/reg/2019/2144/oj

[28] James C. Fell and Robert B. Voas. 2014. The Effectiveness of a 0.05 Blood Alcohol Concentration Limit for Driving in the United States. *Addiction* 109, 6 (2014), 869–874.

[29] Alberto Fernández, Rubén Usamentiaga, Juan Luis Carús, and Rubén Casado. 2016. Driver distraction using visual-based sensors and algorithms. *Sensors* 16, 11 (2016), 1805.

[30] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. Predicting Sample Size Required for Classification Performance. *BMC Medical Informatics and Decision Making* 12, 1 (2012), 1–10.

[31] Per-Anders Fransson, Fredrik Modig, Mitesh Patel, Stephen Gomez, and Måns Magnusson. 2010. Oculomotor deficits caused by 0.06% and 0.10% blood alcohol concentrations and relationship to subjective perception of drunkenness. *Clinical Neurophysiology* 121, 12 (2010), 2134–2142. https://doi.org/10.1016/j.clinph.2010.05.003

[32] Rikard Fredriksson, Michael G. Lenné, Sjef van Montfort, and Colin Grover. 2021. European NCAP Program Developments to Address Driver Distraction, Drowsiness and Sudden Sickness. *Frontiers in Neuroergonomics* 2 (2021). https://doi.org/10.3389/fnrgo.2021.786674

[33] Claire Garnett, David Crane, Robert West, Jamie Brown, and Susan Michie. 2015. Identification of behavior change techniques and engagement strategies to design a smartphone app to reduce alcohol consumption using a formal consensus method. *JMIR mHealth and uHealth* 3, 2 (2015), e73. https://doi.org/10.2196/mhealth.3895

[34] Claire Garnett, David Crane, Robert West, Jamie Brown, and Susan Michie. 2019. The development of Drink Less: An alcohol reduction smartphone app for excessive drinkers. *Translational Behavioral Medicine* 9, 2 (2019), 296–307.

[35] Hasanin Harkous and Hassan Artail. 2019. A two-stage machine learning method for highly-accurate drunk driving detection. In *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*.

[36] Hasanin Harkous, Carine Bardawil, Hassan Artail, and Naseem Daher. 2018. Application of hidden Markov model on car sensors for detecting drunk drivers.

In *IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*.

[37] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[38] Arne Helland, Gunnar D. Jenssen, Lone-Eirin Lervåg, Andreas Austgulen Westin, Terje Moen, Kristian Sakshaug, Stian Lydersen, Jørg Mørland, and Lars Slørdal. 2013. Comparison of driving simulator performance with real driving after alcohol intake: A randomised, single blind, placebo-controlled, cross-over trial. *Accident Analysis & Prevention* 53 (2013), 9–16.

[39] Pei-Yi Hsu, Ya-Fang Lin, Jian-Lun Huang, Chih-Chun Chang, Shih-Yao Lin, Ya-Han Lee, Chuang-Wen You, Yaliang Chuang, Ming-Chyi Huang, Hsin-Tung Tseng, and Hao-Chuan Wang. 2017. A mobile support system to assist DUI offenders on probation in reducing DUI relapse. In *UbiComp '17: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 77–80. https://doi.org/10.1145/3123024.3123154

[40] Jittrapol Intarasirisawat, Chee Siang Ang, Christos Efstratiou, Luke Dickens, Naranchaya Sriburapar, Dinkar Sharma, and Burachai Asawathaweeboon. 2020. An Automated Mobile Game-based Screening Tool for Patients with Alcohol Dependence. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), Article 110. https://doi.org/10.1145/3411837

[41] S. Jongen, E. F. P. M. Vuurman, J. G. Ramaekers, and A. Vermeeren. 2016. The sensitivity of laboratory tests assessing driving related skills to dose-related impairment of alcohol: A literature review. *Accident Analysis & Prevention* 89 (2016), 31–48. https://doi.org/10.1016/j.aap.2016.01.001

[42] Sujay Kakarmath, Andre Esteva, Rima Arnaout, Hugh Harvey, Santosh Kumar, Evan Muse, Feng Dong, Leia Wedlund, and Joseph Kvedar. 2020. Best practices for authors of healthcare-related artificial intelligence manuscripts. *npj Digital Medicine* 3 (2020).

[43] Hsin-Liu Kao, Bo-Jhang Ho, Allan C. Lin, and Hao-Hua Chu. 2012. Phone-based gait analysis to detect alcohol usage. In *Proceedings of the ACM Conference on Ubiquitous Computing*. 661–662. https://doi.org/10.1145/2370216.2370354

[44] Minsong Ki, Bore Cho, Taejun Jeon, Yeongwoo Choi, and Hyeran Byun. 2018. Face Identification for an in-vehicle Surveillance System Using Near Infrared Camera. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6. https://doi.org/10.1109/AVSS.2018.8639472

[45] Sahiti Kunchay and Saeed Abdullah. 2020. WatchOver: using Apple watches to assess and predict substance co-use in young adults. In *UbiComp '20: Proceedings of the 2020 ACM International Joint Conference and 2020 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 488–493. https://doi.org/10.1145/3410530.3414373

[46] Kim P.C. Kuypers, N. Samyn, and Johannes Gerardus Ramaekers. 2006. MDMA and alcohol effects, combined and alone, on objective and subjective measures of actual driving performance and psychomotor function. *Psychopharmacology* 187, 4 (2006), 467–475.

[47] John D. Lee, Dary Fiorentino, Michelle L Reyes, Timothy L. Brown, Omar Ahmad, James Fell, Nic Ward, and Robert Dufour. 2010. *Assessing the feasibility of vehicle-based sensors to detect alcohol impairment*. Report DOT HS 811 358.

[48] Kang Hee Lee, Keon Hee Baek, Su Bin Choi, Nak Tak Jeong, Hyung Uk Moon, Eun Seong Lee, Hyung Min Kim, and Myung Won Suh. 2019. Development of three driver state detection models from driving information using vehicle simulator; normal, drowsy and drunk driving. *International Journal of Automotive Technology* 20, 6 (2019), 1205–1219.

[49] Vera Lehmann, Afroditi Tripyla, David Herzig, Jasmin Meier, Nicolas Banholzer, Martin Maritsch, Jörg Zehetner, Daniel Giachino, Philipp Nett, Stefan Feuerriegel, Felix Wortmann, and Lia Bally. 2021. The impact of postbariatric hypoglycaemia on driving performance: A randomized, single-blind, two-period, crossover study in a driving simulator. *Diabetes, Obesity and Metabolism* 23, 9 (2021), 2189–2193. https://doi.org/10.1111/dom.14456

[50] Vera Lehmann, Thomas Zueger, Martin Maritsch, Mathias Kraus, Caroline Albrecht, Caterina Bérubé, Stefan Feuerriegel, Felix Wortmann, Tobias Kowatsch, Naïma Styger, Sophie Lagger, Markus Laimer, Elgar Fleisch, and Christoph Stettler. 2023. Machine learning for non-invasive sensing of hypoglycemia while driving in people with diabetes. *Diabetes, Obesity and Metabolism* (2023). https://doi.org/10.1111/dom.15021

[51] Zhenlong Li, Xue Jin, and Xiaohua Zhao. 2015. Drunk driving detection based on classification of multivariate time series. *Journal of Safety Research* 54 (2015), 61. e29–64.

[52] ZhenLong Li, HaoXin Wang, YaoWei Zhang, and XiaoHua Zhao. 2020. Random forest–based feature selection and detection method for drunk driving recognition. *International Journal of Distributed Sensor Networks* 16, 2 (2020), 1550147720905234.

[53] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Transactions on Management Information Systems* 4, 4 (2013), Article 19. https://doi.org/10.1145/2544103

[54] Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. 2021. The Empathetic Car: Exploring Emotion Inference via Driver Behaviour and Traffic Context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), Article 117.

[55] Cecile A. Marczinski, Emily L.R. Harrison, and Mark T. Fillmore. 2008. Effects of alcohol on simulated driving and perceived driving impairment in binge drinkers. *Alcoholism: Clinical and Experimental Research* 32, 7 (2008), 1329–1337. https://doi.org/10.1111/j.1530-0277.2008.00701.x

[56] Alex Mariakakis, Sayna Parsi, Shwetak N. Patel, and Jacob O. Wobbrock. 2018. Drunk user interfaces: Determining blood alcohol level through everyday smartphone tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Paper 234.

[57] Teri L. Martin, Patricia A. M. Solbeck, Daryl J. Mayers, Robert M. Langille, Yvona Buczek, and Marc R. Pelletier. 2013. A review of alcohol-impaired driving: The role of blood alcohol concentration and complexity of the driving task. *Journal of Forensic Sciences* 58, 5 (2013), 1238–1250.

[58] Pierre Maurage, Nicolas Masson, Zoé Bollen, and Fabien D'Hondt. 2020. Eye tracking correlates of acute alcohol consumption: A systematic and critical review. *Neuroscience & Biobehavioral Reviews* 108 (2020), 400–422. https://doi.org/10.1016/j.neubiorev.2019.10.001

[59] Anthony D. McDonald, Thomas K. Ferris, and Tyler A. Wiener. 2020. Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors* 62, 6 (2020), 1019–1035.

[60] Lakmal Meegahapola, Florian Labhart, Thanh-Trung Phan, and Daniel Gatica-Perez. 2021. Examining the Social Context of Alcohol Drinking in Young Adults with Smartphone Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), Article 121. https://doi.org/10.1145/3478126

[61] Kenneth C. Mills, Susan E. Spruill, Roy W. Kanne, Katherine M. Parkman, and Ying Zhang. 2001. The influence of stimulants, sedatives, and fatigue on tunnel vision: risk factors for driving and piloting. *Human Factors* 43, 2 (2001), 310–327.

[62] Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. 2014. RecoFit: Using a Wearable Sensor to Find, Recognize, and Count Repetitive Exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3225–3234.

[63] Herbert Moskowitz and D. Florentino. 2000. *A review of the literature on the effects of low doses of alcohol on driving-related skills*. Report DOT HS 809 028. National Highway Traffic Safety Administration.

[64] Inbal Nahum-Shani, Eric B. Hekler, and Donna Spruijt-Metz. 2015. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology* 34 (2015), 1209–1219. https://pubmed.ncbi.nlm.nih.gov/26651462https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4732268/

[65] Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. 2017. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2017), 446–462. https://doi.org/10.1007/s12160-016-9830-8

[66] Rizwan Ali Naqvi, Muhammad Arsalan, Ganbayar Batchuluun, Hyo Sik Yoon, and Kang Ryoung Park. 2018. Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. *Sensors* 18, 2 (2018), 456.

[67] National Highway Traffic Safety Administration. 2022. *Traffic safety facts 2020*. Report No. DOT HS 813 294.

[68] Nissan. 2015. Driver Attention Alert System. https://www.nissanusa.com/experience-nissan/news-and-events/drowsy-driver-attention-alert-car-feature.html Accessed: 09/01/2022.

[69] Ashley Nunes, Bryan Reimer, and Joseph F. Coughlin. 2018. People must retain control of autonomous vehicles. *Nature* 556, 7700 (2018), 169–171.

[70] Cynthia Owsley, Gerald McGwin Jr., and Karen Searcey. 2013. A Population-based Examination of the Visual and Ophthalmological Characteristics of Licensed Drivers Aged 70 and Older. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 68, 5 (2013), 567–573.

[71] Ji Hwan Park, Han Eol Cho, Jong Hun Kim, Melanie M. Wall, Yaakov Stern, Hyunsun Lim, Shinjae Yoo, Hyoung Seop Kim, and Jiook Cha. 2020. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *npj Digital Medicine* 3, 1 (2020), 46. https://doi.org/10.1038/s41746-020-0256-0

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[73] Ignacio Perez-Pozuelo, Dimitris Spathis, Emma A. D. Clifton, and Cecilia Mascolo. 2021. *Digital Health*. Elsevier, Book section Wearables, smartphones, and artificial intelligence for digital phenotyping and health, 33–54.

[74] Thanh-Trung Phan, Skanda Muralidhar, and Daniel Gatica-Perez. 2019. #Drink Or #Drunk: Multimodal Signals and Drinking Practices on Instagram. In *PervasiveHealth'19: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare.* 71–80. https://doi.org/10.1145/3329189.3329193

[75] Thanh-Trung Phan, Skanda Muralidhar, and Daniel Gatica-Perez. 2019. Drinks & Crowds: Characterizing Alcohol Consumption through Crowdsensing and Social Media. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), Article 59. https://doi.org/10.1145/3328930

[76] Anuj Kumar Pradhan, Kim R. Hammel, Rosa DeRamus, Alexander Pollatsek, David A. Noyce, and Donald L. Fisher. 2005. Using eye movements to evaluate effects of driver age on risk perception in a driving simulator. *Human Factors* 47, 4 (2005), 840–852.

[77] Igor Radun, Jussi Ohisalo, Sirpa Rajalin, Jenni E. Radun, Mattias Wahde, and Timo Lajunen. 2014. Alcohol ignition interlocks in all new vehicles: A broader perspective. *Traffic Injury Prevention* 15, 4 (2014), 335–342.

[78] Johannes G. Ramaekers, Günter Berghaus, Margriet van Laar, and Olaf H. Drummer. 2004. Dose related risk of motor vehicle crashes after cannabis use. *Drug and Alcohol Dependence* 73, 2 (2004), 109–119.

[79] Audrey Robinel and Didier Puzenat. 2014. Alcohol consumption detection through behavioural analysis using intelligent systems. *Expert Systems with Applications* 41, 5 (2014), 2574–2581.

[80] Paul D Rosero-Montalvo, Vivian Félix López-Batista, and Diego Hernán Peluffo-Ordóñez. 2020. Hybrid embedded-systems-based approach to in-driver drunk status detection using image processing and sensor networks. *IEEE Sensors Journal* 21, 14 (2020), 15729–15740.

[81] Matthew E. Rossheim, Dennis L. Thombs, Kwynn M. Gonzalez-Pons, Jordan A. Killion, John D. Clapp, Mark B. Reed, Julie M. Croff, Danielle E. Ruderman, and Robert M. Weiler. 2016. Feeling no buzz or a slight buzz is common when legally drunk. *American Journal of Public Health* 106, 10 (2016), 1761–1762. https://doi.org/10.2105/AJPH.2016.303321

[82] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. 2017. The need to approximate the use-case in clinical machine learning. *GigaScience* 6, 5 (2017). https://doi.org/10.1093/gigascience/gix019

[83] Suparna Sahabiswas, Sourav Saha, Prachatos Mitra, Retabrata Chatterjee, Ronit Ray, Paramartha Saha, Rajarshi Basu, Saurav Patra, Pritam Paul, and Bidrohi Ananya Biswas. 2016. Drunken driving detection and prevention models using Internet of Things. In *IEEE Information Technology, Electronics and Mobile Communication Conference (IEMCON).* IEEE, 1–4.

[84] John B. Saunders, Olaf G. Aasland, Thomas F. Babor, Juan R. De la Fuente, and Marcus Grant. 1993. Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* 88, 6 (1993), 791–804.

[85] Kai-Uwe Schmitt, Christian Lanz, Markus H. Muser, Felix Walz, and Urs Schwarz. 2013. Saccadic eye movements after low-dose oral alcohol exposure. *Journal of Forensic and Legal Medicine* 20, 7 (2013), 870–874. https://doi.org/10.1016/j.jflm.2013.06.023

[86] Stephan Seidl, Uwe Jensen, and Andreas Alt. 2000. The calculation of blood ethanol concentrations in males and females. *International Journal of Legal Medicine* 114, 1 (2000), 71–77.

[87] Brook A. Shiferaw, David P. Crewther, and Luke A. Downey. 2019. Gaze entropy measures detect alcohol-induced driver impairment. *Drug and Alcohol Dependence* 204 (2019), 107519. https://doi.org/10.1016/j.drugalcdep.2019.06.021

[88] Smart Eye AB. 2022. Driver monitoring system. https://smarteye.se/automotive/driver-monitoring-systems/ Accessed: 14/02/2022.

[89] Sónia Soares, Sara Ferreira, and António Couto. 2020. Driving simulator experiments to study drowsiness: A systematic review. *Traffic Injury Prevention* 21, 1 (2020), 29–37.

[90] Statistisches Bundesamt. 2020. *Verkehrsunfälle - Unfälle unter dem Einfluss von Alkohol oder anderen berauschenden Mitteln im Straßenverkehr.* Report 5462404-19700-4. Statistisches Bundesamt.

[91] Dawn E. Sugarman and Kate B. Carey. 2009. Drink less or drink slower: the effects of instruction on alcohol consumption and drinking control strategy use. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors* 23, 4 (2009), 577–585. https://doi.org/10.1037/a0016580

[92] Yifan Sun, Jinglei Zhang, Xiaoyuan Wang, Zhangu Wang, and Jie Yu. 2018. Recognition method of drinking-driving behaviors based on PCA and RBF neural network. *Traffic & Transportation* 30, 4 (2018), 407–417.

[93] Mika Sunagawa, Shin-ichi Shikii, Wataru Nakai, Makoto Mochizuki, Koichi Kusukame, and Hiroki Kitajima. 2019. Comprehensive drowsiness level detection model combining multimodal information. *IEEE Sensors Journal* 20, 7 (2019), 3709–3717.

[94] Robert J. Tait, Raquel Paz Castro, Jessica Jane Louise Kirkman, Jamie Christopher Moore, and Michael P. Schaub. 2019. A digital intervention addressing alcohol use problems (the "Daybreak" program): Quasi-experimental randomized controlled trial. *Journal of Medical Internet Research* 21, 9 (2019), e14967. https://doi.org/10.2196/14967

[95] Q. Thurman, S. Jackson, and J. Zhao. 1993. Drunk-Driving Research and Innovation: A Factorial Survey Study of Decisions To Drink and Drive. *Social Science Research* 22, 3 (1993), 245–264. https://doi.org/10.1006/ssre.1993.1012

[96] United States Congress. 2021. Public Law 117-58 - Infrastructure Investment and Jobs Act. https://www.congress.gov/bill/117th-congress/house-bill/3684/text Accessed: 01/09/2022.

[97] Wim van Winsum. 2019. Optic flow and tunnel vision in the detection response task. *Human Factors* 61, 6 (2019), 992–1003.

[98] Renju Rachel Varghese, Pramod Mathew Jacob, Joanna Jacob, Merlin Nissi Babu, Rupali Ravikanth, and Stephy Mariyam George. 2021. An Integrated Framework for Driver Drowsiness Detection and Alcohol Intoxication using Machine Learning. In *IEEE International Conference on Data Analytics for Business and Industry (ICDABI).* IEEE, 531–536.

[99] Paul Viola and Michael Jones. 2001. Robust real-time object detection. *International journal of computer vision* 4, 34-47 (2001), 4.

[100] V. Viswanatha, Ashwini Kumari, and Pradeep Kumar. 2021. Internet of things (IoT) based multilevel drunken driving detection and prevention system using Raspberry Pi 3. *International Journal of Internet of Things and Web Services* 6 (2021).

[101] Willem Vlakveld, Paul Wesemann, Eline Devillers, R. Elvik, and K. Veisten. 2005. *Detailed cost-benefit analysis of potential impairment countermeasures.* Report R-2005-10.

[102] Volkswagen. 2022. Driver Alert System. https://www.volkswagen-newsroom.com/en/driver-alert-system-3932 Accessed: 09/01/2022.

[103] J. Wang, W. Chai, A. Venkatachalapathy, K. L. Tan, A. Haghighat, S. Velipasalar, Y. Adu-Gyamfi, and A. Sharma. 2022. A Survey on Driver Behavior Analysis From In-Vehicle Cameras. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2022), 10186–10209. https://doi.org/10.1109/TITS.2021.3126231

[104] Patricia E. Watson, Ian D. Watson, and Richard D. Batt. 1981. Prediction of blood alcohol concentrations in human subjects. Updating the Widmark Equation. *Journal of Studies on Alcohol* 42, 7 (1981), 547–556.

[105] Gary M. Weiss. 2013. *Imbalanced learning: foundations, algorithms, and applications.* IEEE Press, New York, USA, Chapter Methods for addressing imbalanced data.

[106] Ching-Hua Weng, Ying-Hsiu Lai, and Shang-Hong Lai. 2016. Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision.* Springer, 117–133.

[107] Erik Matteo Prochet Widmark. 1932. *Die theoretischen Grundlagen und die praktische Verwendbarkeit der gerichtlich-medizinischen Alkoholbestimmung.* Urban & Schwarzenberg.

[108] Peter H. Wolff. 1972. Ethnic Differences in Alcohol Sensitivity. *Science* 175, 4020 (1972), 449–450.

[109] World Health Organization. 2018. *Global status report on road safety 2018: Summary.* Report WHO/NMH/NVI/18.20. World Health Organization.

[110] World Health Organization. 2019. *Global status report on alcohol and health 2018.* World Health Organization.

[111] World Health Organization. 2019. *The SAFER technical package: five areas of intervention at national and subnational levels.* Report 9241516410.

[112] Friedrich M. Wurst, Natasha Thon, Michel Yegles, Alexandra Schrück, Ulrich W. Preuss, and Wolfgang Weinmann. 2015. Ethanol metabolites: their role in the assessment of alcohol intake. *Alcoholism: Clinical and Experimental Research* 39, 11 (2015), 2060–2072.

[113] Ankit Kumar Yadav and Nagendra R. Velaga. 2019. Modelling the relationship between different blood alcohol concentrations and reaction time of young and mature drivers. *Transportation Research Part F: Traffic Psychology and Behaviour* 64 (2019), 227–245. https://doi.org/10.1016/j.trf.2019.05.011

[114] Chuang-Wen You, Yaliang Chuang, Hung-Yeh Lin, Jui-Ting Tsai, Yi-Ching Huang, Chia-Hua Kuo, Ming-Chyi Huang, Shan Jean Wu, Frank Wencheng Liu, Jane Yung-Jen Hsu, and Hui-Ching Wu. 2019. SoberComm: Using Mobile Phones to Facilitate Inter-family Communication with Alcohol-dependent Patients. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), Article 119. https://doi.org/10.1145/3351277

[115] Chuang-Wen You, Ya-Fang Lin, Yaliang Chuang, Ya-Han Lee, Pei-Yi Hsu, Shih-Yao Lin, Chih-Chun Chang, Yi-Ju Chung, Yi-Ling Chen, Ming-Chyi Huang, Ping-Hsuan Shen, Hsin-Tung Tseng, and Hao-Chuan Wang. 2018. SoberMotion: Leveraging the Force of Probation Officers to Reduce the Risk of DUI Recidivism. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), Article 146. https://doi.org/10.1145/3264956

[116] Chuang-Wen You, Lu-Hua Shih, Hung-Yeh Lin, Yaliang Chuang, Yi-Chao Chen, Yi-Ling Chen, and Ming-Chyi Huang. 2019. Enabling Personal Alcohol Tracking using Transdermal Sensing Wristbands: Benefits and Challenges. In *MobileHCI '19: Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services.* Article 37. https://doi.org/10.1145/3338286.3344384

[117] Chuang-Wen You, Jui-Ting Tsai, Hung-Yeh Lin, Yu-Ting Wang, Yi-Ching Huang, Diane Lu, Yi-Ju Chung, Yaliang Chuang, Chia-Hua Kuo, Ming-Chyi Huang, Jane Yung-Jen Hsu, and Hui-Ching Wu. 2018. Using Mobile Phones to Facilitate Alcohol Dependent Patients to Improve Family Communication. In *UbiComp '18:*

Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 311–314. https://doi.org/10.1145/3267305.3267566

[118] Chuang-Wen You, Kuo-Cheng Wang, Ming-Chyi Huang, Yen-Chang Chen, Cheng-Lin Lin, Po-Shiun Ho, Hao-Chuan Wang, Polly Huang, and Hao-Hua Chu. 2015. SoberDiary: A Phone-based Support System for Assisting Recovery from Alcohol Dependence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 3839–3848. https://doi.org/10.1145/2702123.2702289

[119] Feng You, Xiaolong Li, Yunbo Gong, Hailwei Wang, and Hongyi Li. 2019. A real-time driving drowsiness detection algorithm with individual differences consideration. *IEEE Access* 7 (2019), 179396–179408.

[120] Jongmin Yu, Sangwoo Park, Sangwook Lee, and Moongu Jeon. 2018. Driver drowsiness detection using condition-adaptive representation learning framework. *IEEE Transactions on Intelligent Transportation Systems* 20, 11 (2018), 4206–4218.

[121] Kevan Yuen, Sujitha Martin, and Mohan M Trivedi. 2016. On looking at faces in an automobile: Issues, algorithms and evaluation on naturalistic driving dataset. In *IEEE International Conference on Pattern Recognition (ICPR)*. IEEE, 2777–2782.

[122] Ali Shahidi Zandi, Azhar Quddus, Laura Prest, and Felix J. E. Comeau. 2019. Non-intrusive detection of drowsy driving based on eye tracking data. *Transportation research record* 2673, 6 (2019), 247–257.

[123] Abdullatif K. Zaouk, Michael Willis, Eric Traube, and Robert Strassburger. 2019. Driver alcohol detection system for safety (DADSS)–a Non-regulatory approach in the research and development of vehicle safety technology to reduce alcohol-impaired driving – A status update. In *26th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*.

[124] Yijun Zhao, Tong Wang, Riley Bove, Bruce Cree, Roland Henry, Hrishikesh Lokhande, Mariann Polgar-Turcsanyi, Mark Anderson, Rohit Bakshi, Howard L. Weiner, Tanuja Chitnis, and Summit Investigators. 2020. Ensemble learning predicts multiple sclerosis disease course in the SUMMIT study. *npj Digital Medicine* 3, 1 (2020), 135. https://doi.org/10.1038/s41746-020-00338-8

A DETAILS ON PARTICIPANTS

The study flow diagram is shown in Supplementary Figure 1. Inclusion and exclusion criteria for participation are explained in the following.

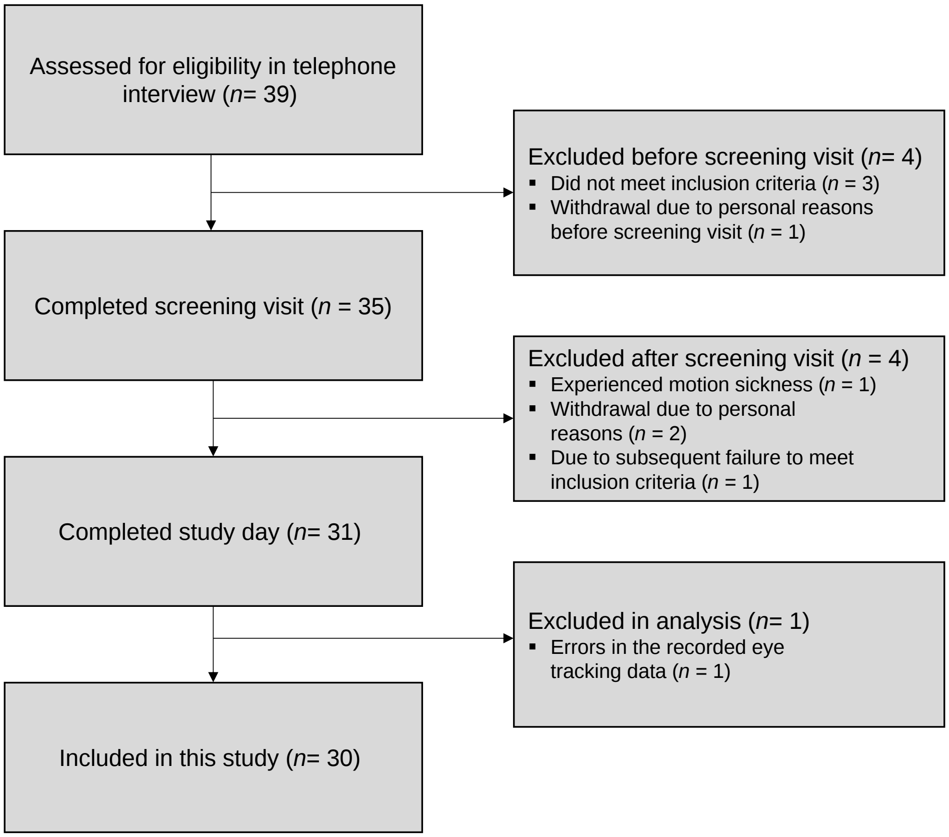
A.1 Inclusion and exclusion criteria

Inclusion criteria for individuals eligible for participation in the DRIVE study are: (1) passing the driver examination at least 2 years before study inclusion; (2) possession of a driver's license that is valid in the European Union or Switzerland; and (3) reporting moderate alcohol consumption (i.e., neither total absence nor excess). The latter was examined via the AUDIT [84] and the PEth level of a capillary blood sample (below 210 ng/mL) [112].

Exclusion criteria for participation in the DRIVE study were if participants met one or more of the following: women who are pregnant, breast feeding, or intend to become pregnant during the course of the study; other clinically significant concomitant disease states as judged by the investigator (e.g., renal failure, hepatic dysfunction, cardiovascular disease etc.); known or suspected non-compliance or drug abuse; inability to follow the procedures of the study, e.g., due to language problems, psychological disorders, dementia, etc.; participation in another study with investigational drug within the 30 days preceding and during the present study; specific concomitant therapy washout requirements prior to and/or during study participation; previous enrollment into the current study; personal dependences with the study team (e.g., employees, family members, and other dependent persons); physical or psychological disease likely to interfere with the normal conduct of the study and interpretation of the study results as judged by the investigator (especially coronary heart disease or epilepsy); current treatment with drugs known to interfere with metabolism (e.g., systemic corticosteroids, statins etc.) or driving performance (e.g., opioids, benzodiazepines); patients not capable of driving with a driving simulator or patients experiencing motion sickness during the simulator test driving session at the introductory study visit.

A.2 Study flow

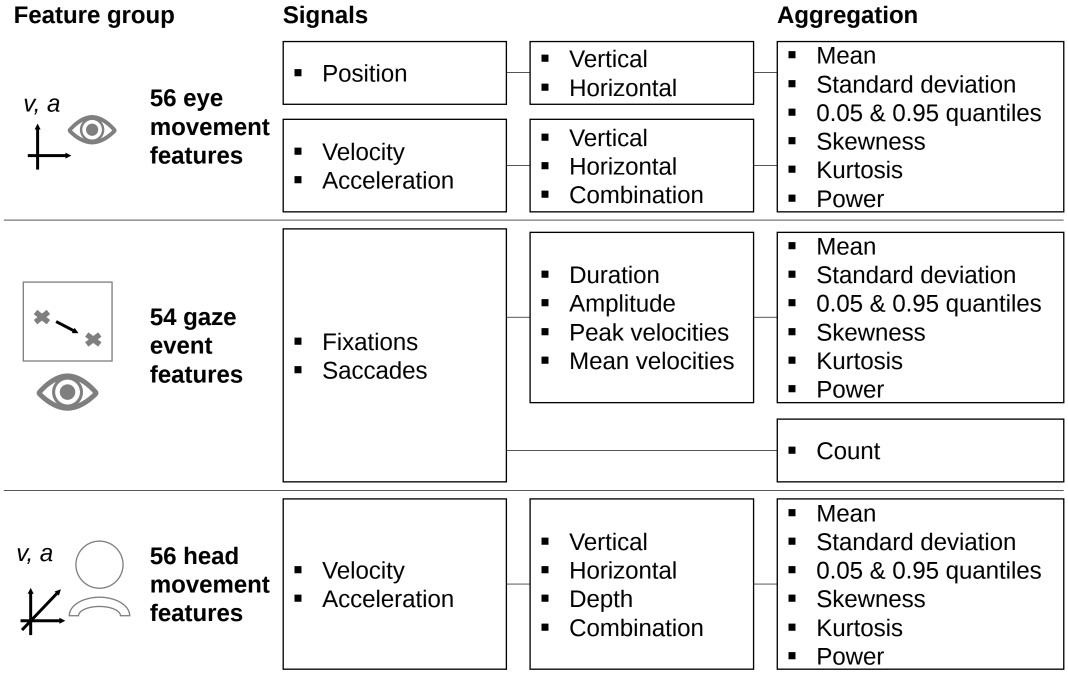
Of the screened 39 participants, we had to exclude 9 participants as they did not meet the inclusion/exclusion criteria. Three participants were excluded because they did not meet the inclusion criteria from above and reported – at least occasional – consumption of illegal drugs. One participant withdrew from the study due to time conflicts. After screening, one individual was excluded due to experiencing motion sickness while driving in the simulator, two participants withdrew for personal reasons, and one because of subsequent failure to comply with the participation requirements (i.e., drug prescription after initial screening). Finally, one individual was excluded due to errors in the eye-tracking recording. A diagram showing the study flow is in Supplementary Figure 1.


Assessed for eligibility in telephone interview (*n*= 39)
Excluded before screening visit (*n*= 4)
▪ Did not meet inclusion criteria (*n* = 3)
▪ Withdrawal due to personal reasons before screening visit (*n* = 1)
Completed screening visit (*n* = 35)
Excluded after screening visit (*n* = 4)
▪ Experienced motion sickness (*n* = 1)
▪ Withdrawal due to personal reasons (*n* = 2)
▪ Due to subsequent failure to meet inclusion criteria (*n* = 1)
Completed study day (*n*= 31)
Excluded in analysis (*n*= 1)
▪ Errors in the recorded eye tracking data (*n* = 1)
Included in this study (*n*= 30)

Supplementary Figure 1: Study flow diagram. Participants subject to screening and inclusion/exclusion in the clinical study.

B FEATURE GENERATION

The feature generation is shown in Supplementary Figure 2.



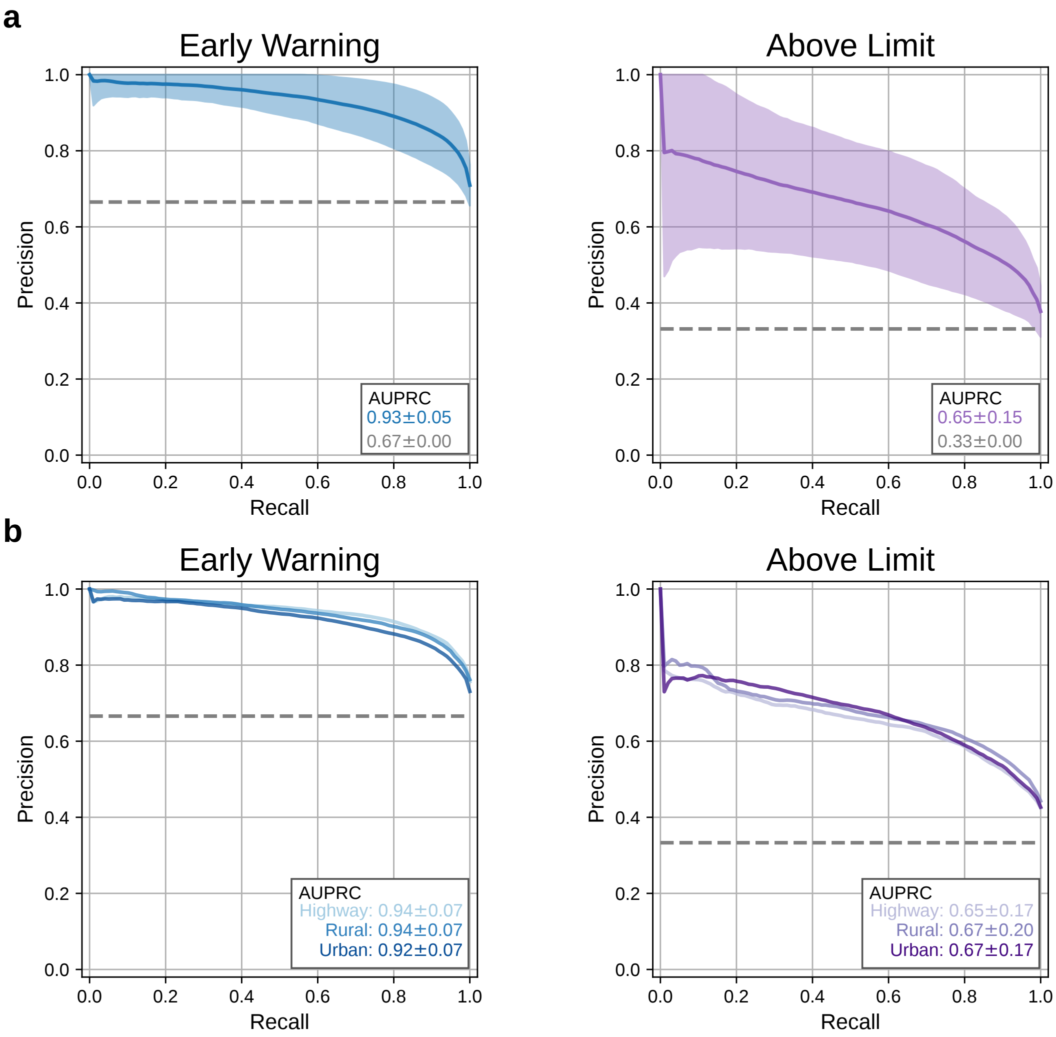
Supplementary Figure 2: Overview of feature generation. Features are arranged in three groups: eye movement features, gaze event features, and head movements, which are all derived from the driver monitoring camera. Each feature group is further processed by aggregation functions (i.e., to map a time series onto a single value).

C ADDITIONAL INSIGHTS ON MACHINE LEARNING PERFORMANCE

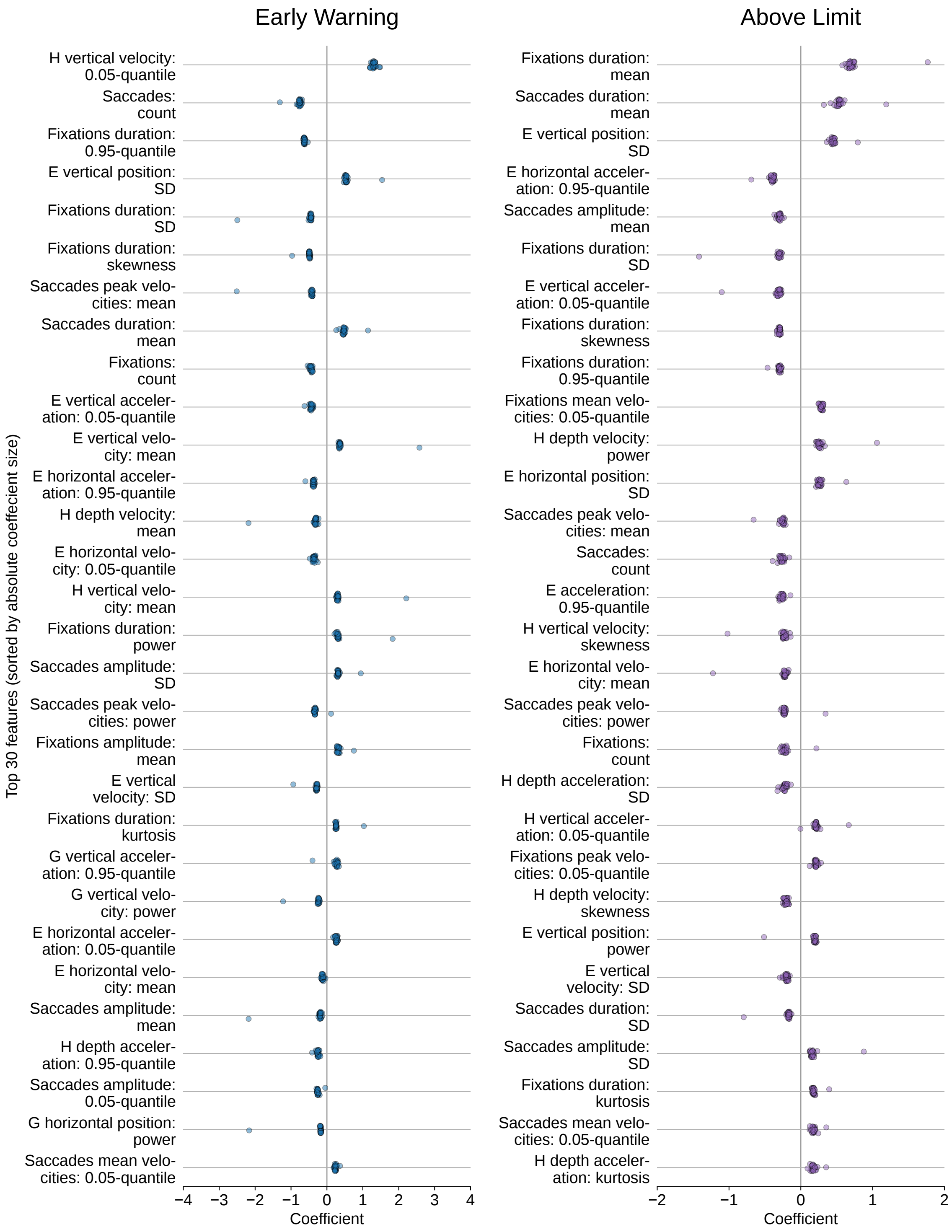
Here, we report additional performance metrics (Supplementary Table 1, Supplementary Figure 3, and Supplementary Table 2) and results for machine learning interpretability (Supplementary Figure 4).

Supplementary Table 1: Comparison of performance metrics with additional data sources as a baseline. We report the performance of our machine learning system while using different data sources: camera-only, CAN-only, and both combined. The following performance metrics are computed: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), balanced accuracy, and F1 score (weighted by classes). Reported: mean ± standard deviation. CAN: Controller Area Network (i.e., vehicle signals).

Data source		AUROC	AUPRC	Balanced accuracy	F1 score
Camera-only	Early Warning	0.88±0.09	0.93±0.05	0.76±0.10	0.75±0.14
	Above Limit	0.79±0.10	0.65±0.16	0.68±0.10	0.67±0.12
CAN-only	Early Warning	0.74±0.10	0.84±0.07	0.65±0.08	0.64±0.11
	Above Limit	0.66±0.12	0.50±0.15	0.60±0.09	0.60±0.08
Both combined	Early Warning	0.91±0.07	0.95±0.04	0.78±0.10	0.77±0.13
	Above Limit	0.81±0.11	0.68±0.16	0.69±0.10	0.69±0.12



Supplementary Figure 3: Performance of drunk driving detection. The machine learning system for detecting drunk driving is evaluated based on the area under the precision-recall curve (AUPRC). (a) Performance across participants for different BAC thresholds. (b) Performance by driving scenario (i.e., highway, rural, and urban). The dashed, gray line shows an AUPRC of a naïve classifier predicting the majority class, which is consistently outperformed by our ML system. BAC: blood alcohol concentration.



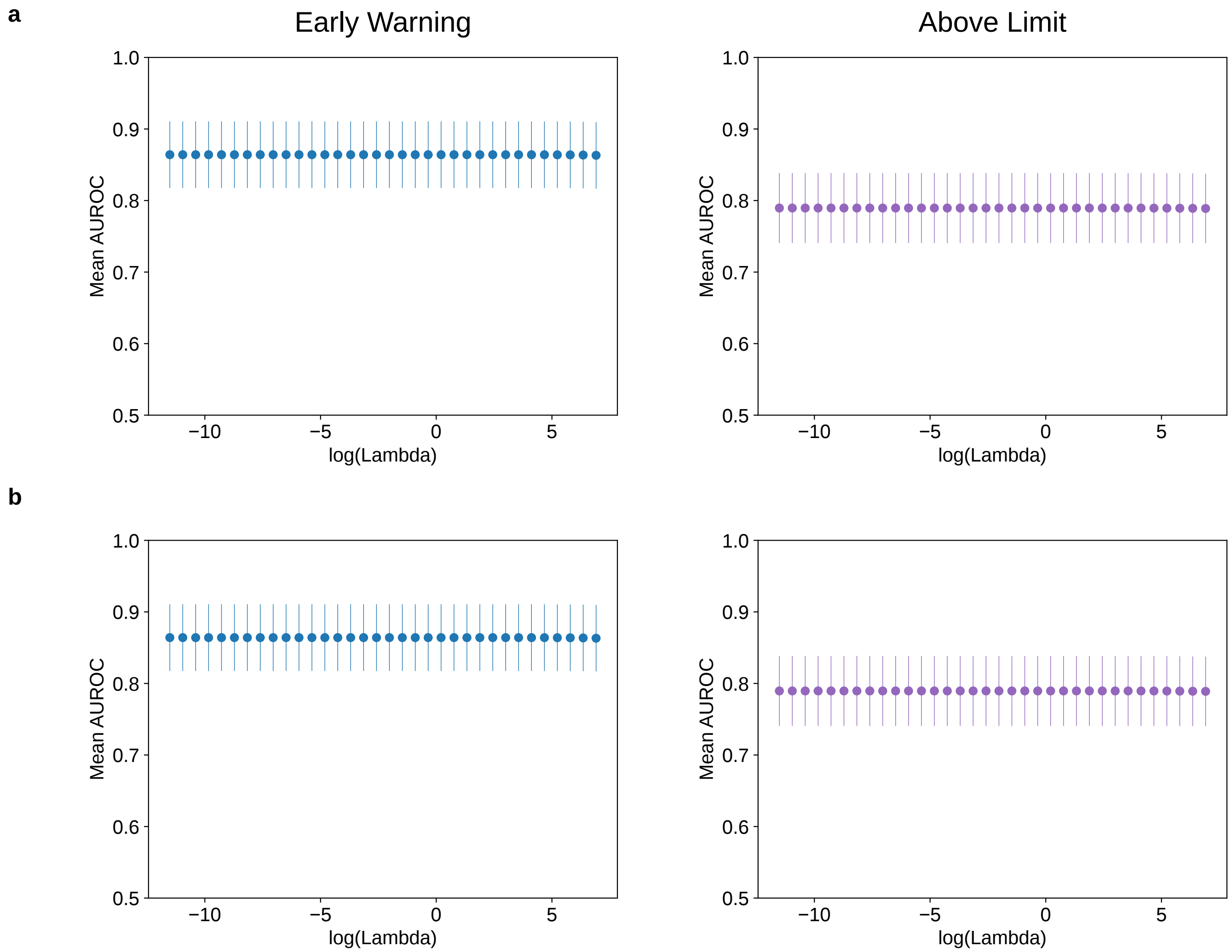
Supplementary Figure 4: Selected features in the machine learning system. Shown are the top 30 features. Features are ranked descending by their absolute coefficient size and, therefore, reflect their importance for the detection task. E: Eye movement; H: Head movement; SD: standard deviation.

Supplementary Table 2: Detailed performance metrics for a majority vote aggregation. We report the performance of our machine learning system when aggregating non-overlapping single windows over different amount of windows. With an increasing number of of aggregated windows, the performance metrics improve. The following performance metrics are computed: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), balanced accuracy, and F1 score (weighted by classes). Reported: mean ± standard deviation.

Window amount		AUROC	AUPRC	Balanced accuracy	F1 score
None	Early Warning	0.88±0.09	0.93±0.05	0.76±0.10	0.75±0.14
	Above Limit	0.79±0.10	0.65±0.16	0.68±0.10	0.67±0.12
5	Early Warning	0.88±0.08	0.93±0.05	0.76±0.10	0.75±0.13
	Above Limit	0.80±0.11	0.66±0.16	0.68±0.09	0.67±0.12
30	Early Warning	0.89±0.08	0.93±0.05	0.75±0.11	0.75±0.13
	Above Limit	0.80±0.10	0.67±0.15	0.67±0.10	0.67±0.12
60	Early Warning	0.90±0.08	0.94±0.04	0.77±0.11	0.76±0.13
	Above Limit	0.82±0.11	0.70±0.15	0.68±0.11	0.67±0.14
90	Early Warning	0.90±0.08	0.93±0.05	0.76±0.11	0.75±0.13
	Above Limit	0.83±0.11	0.71±0.15	0.68±0.11	0.67±0.13
120	Early Warning	0.91±0.08	0.94±0.06	0.78±0.12	0.77±0.14
	Above Limit	0.84±0.11	0.75±0.16	0.68±0.12	0.67±0.15
150	Early Warning	0.91±0.08	0.94±0.06	0.79±0.13	0.77±0.15
	Above Limit	0.85±0.11	0.76±0.17	0.68±0.13	0.67±0.16
180	Early Warning	0.88±0.09	0.89±0.09	0.73±0.12	0.72±0.15
	Above Limit	0.83±0.10	0.71±0.14	0.66±0.12	0.65±0.15

D ROBUSTNESS CHECKS

Here, we report the following robustness checks for the regularization (Supplementary Figure 5), an evaluation based on leaving one scenario in training out and evaluating on it (Supplementary Table 3), the length of the sliding window (Supplementary Table 4), the predictive power of different feature groups (Supplementary Table 5), and alternative ML models (Supplementary Table 6).



Supplementary Figure 5: Sensitivity analysis for regularization. The following evaluations compare the performance when varying the hyperparameter lambda inside the regularization. Across all lambdas, the machine learning system yields a robust, high prediction performance. (a) The prediction performance for different hyperparameters (lambda) that weight the L1 penalty in a logistic regression with lasso regularization. (b) The prediction for different hyperparameters (lambda) that weight the L2 penalty in a logistic regression with ridge regularization. Reported: mean ± standard deviation. AUROC: area under the receiver operating characteristic curve.

Supplementary Table 3: Sensitivity analysis for leaving-one-driving-scenario-out. The following evaluations compare the performance when our machine learning system is trained on two driving scenarios and is evaluated on on the third driving scenario. This analysis is done on top of our leave-one-subject out validation. The machine learning system has a robust, high prediction performance even for previously unseen scenarios. Reported: mean ± standard deviation. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

Unseen scenario	Task	AUROC	AUPRC	Balanced accuracy	F1 score
Highway	Early Warning	0.89±0.12	0.93±0.08	0.75±0.13	0.77±0.13
	Above Limit	0.78±0.13	0.64±0.17	0.66±0.12	0.60±0.17
Rural	Early Warning	0.94±0.05	0.88±0.06	0.76±0.14	0.75±0.15
	Above Limit	0.77±0.18	0.64±0.20	0.66±0.13	0.64±0.15
Urban	Early Warning	0.92±0.07	0.91±0.05	0.73±0.10	0.70±0.17
	Above Limit	0.80±0.11	0.65±0.16	0.66±0.11	0.67±0.11
Averaged	Early Warning	0.82±0.11	0.93±0.07	0.75±0.12	0.74±0.15
	Above Limit	0.78±0.14	0.64±0.18	0.66±0.12	0.64±0.15

Supplementary Table 4: Sensitivity analysis for window size. The following evaluations compare the performance when varying the size of the sliding window. Across all sizes, the machine learning system has a robust, high prediction performance. Reported: mean ± standard deviation. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

Window size	Task	AUROC	AUPRC	Balanced accuracy	F1 score
5s	Early Warning	0.70±0.08	0.81±0.06	0.63±0.07	0.64±0.09
	Above Limit	0.65±0.06	0.45±0.07	0.60±0.05	0.60±0.06
10s	Early Warning	0.75±0.09	0.84±0.06	0.67±0.08	0.67±0.10
	Above Limit	0.69±0.07	0.50±0.09	0.63±0.06	0.63±0.08
20s	Early Warning	0.81±0.10	0.88±0.06	0.71±0.09	0.71±0.11
	Above Limit	0.74±0.09	0.56±0.11	0.65±0.08	0.65±0.10
40s	Early Warning	0.85±0.10	0.91±0.05	0.74±0.10	0.73±0.13
	Above Limit	0.77±0.10	0.62±0.14	0.67±0.09	0.66±0.12
60s	Early Warning	0.88±0.09	0.93±0.05	0.76±0.10	0.75±0.14
	Above Limit	0.79±0.10	0.65±0.16	0.68±0.10	0.67±0.12
80s	Early Warning	0.89±0.09	0.94±0.04	0.77±0.10	0.76±0.13
	Above Limit	0.80±0.12	0.67±0.17	0.68±0.11	0.67±0.13
120s	Early Warning	0.91±0.08	0.95±0.04	0.79±0.11	0.78±0.13
	Above Limit	0.82±0.12	0.69±0.19	0.68±0.12	0.67±0.14

Supplementary Table 5: Sensitivity analysis for feature group. The following evaluations compare the performance when varying the selected features. Across all the features groups, gaze events perform best. However, the prediction performance drops compared to combining all features. Reported: mean ± standard deviation. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

Feature group	Task	AUROC	AUPRC	Balanced accuracy	F1 score
Eye movements	Early Warning	0.79±0.11	0.87±0.07	0.69±0.10	0.69±0.12
	Above Limit	0.76±0.10	0.60±0.14	0.67±0.09	0.65±0.11
Gaze events	Early Warning	0.84±0.09	0.91±0.06	0.71±0.11	0.70±0.15
	Above Limit	0.79±0.08	0.63±0.13	0.68±0.08	0.67±0.11
Head movements	Early Warning	0.82±0.12	0.89±0.09	0.67±0.11	0.65±0.17
	Above Limit	0.69±0.14	0.54±0.16	0.62±0.11	0.59±0.16

Supplementary Table 6: Sensitivity analysis for machine learning model. The following evaluations compare the performance when varying the machine learning model. Across all models, the machine learning system has a robust, high prediction performance. Reported: mean ± standard deviation. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; MLP: multi-layer perceptron; SVM: support vector machine.

Model	**Task**	**AUROC**	**AUPRC**	**Balanced accuracy**	**F1 score**
Lasso	Early Warning	0.88±0.09	0.93±0.05	0.76±0.10	0.75±0.14
	Above Limit	0.79±0.10	0.65±0.16	0.68±0.10	0.67±0.12
Ridge	Early Warning	0.88±0.09	0.93±0.05	0.76±0.11	0.75±0.14
	Above Limit	0.79±0.10	0.65±0.16	0.68±0.10	0.67±0.12
Elastic Net	Early Warning	0.88±0.09	0.93±0.05	0.76±0.10	0.75±0.14
	Above Limit	0.79±0.10	0.65±0.16	0.68±0.10	0.67±0.12
SVM	Early Warning	0.83±0.12	0.90±0.08	0.71±0.09	0.73±0.10
	Above Limit	0.73±0.12	0.57±0.15	0.61±0.09	0.64±0.08
Random forest	Early Warning	0.79±0.14	0.87±0.09	0.61±0.12	0.65±0.12
	Above Limit	0.73±0.09	0.55±0.13	0.57±0.08	0.61±0.08
Gradient boosting	Early Warning	0.83±0.13	0.90±0.08	0.68±0.12	0.71±0.12
	Above Limit	0.77±0.08	0.60±0.11	0.61±0.08	0.65±0.08
MLP	Early Warning	0.85±0.10	0.92±0.06	0.73±0.09	0.75±0.09
	Above Limit	0.73±0.12	0.56±0.15	0.61±0.09	0.64±0.08