Beyond Breathalyzers: Towards Pre-Driving Sobriety Testing with a Driver Monitoring Camera

Simon Stent*, John Gideon, Kimimasa Tamura, Avinash Balachandran, Guy Rosman

***Abstract*—Field sobriety tests and breathalyzers are commonly used to prevent alcohol-impaired driving, but are expensive and time-consuming to administer. We propose a set of sobriety tests which, in contrast, can feasibly be automated and deployed to modern vehicles equipped with a driver monitoring camera. Our tests are inspired by research on the physiological effects of alcohol, with particular focus on eye movements and gaze behavior. We run an exploratory in-lab study with N=50 subjects (20 alcohol-impaired, 30 control), and train a variety of models to detect alcohol impairment. We find that, using only 10 seconds of observations of the driver, one of the four proposed tests performs comparably to existing non-breathalyzer field sobriety tests. We make our code and data available to support further research efforts to combat alcohol-impaired driving: https://toyotaresearchinstitute.github.io/IV25-beyond-breathalysers/.**

I. INTRODUCTION

Alcohol intoxication is one of the leading causes of road traffic accidents. The World Health Organization estimated that in high-income countries in 2023, around one fifth of fatally injured drivers had a blood alcohol concentration (BAC) above the legal limit, while in middle- and low-income countries, 33-69% of fatally injured drivers had consumed alcohol prior to their crash [1]. The National Highway Traffic Safety Administration (NHTSA) estimated that alcohol-impaired driving costs the US economy around $280 billion/year in lost wages, lost quality of life, medical costs, and more [2].

Because of its significant societal cost, a variety of methods to detect alcohol impairment have become commonplace in driving: in particular, measurement of Breath Alcohol Concentration (BrAC) through a breathalyzer test, and the now-standardized field sobriety tests [3] which are used in roadside stops. The European New Car Assessment Programme (Euro NCAP) "Roadmap for 2030" highlighted the urgent need for automakers "to expand the scope of driver impairment adding specific detection of driving under the influence... with advanced vision and/or biometric sensors" [4], while in the US, the Bipartisan Infrastructure Law in 2021 created regulations requiring improved drunk and impaired driving prevention technology [5].

Prior work has proven the feasibility of detecting alcohol intoxication without breathalyzers by observing a driver's behavior over several minutes of driving (e.g. [6]). Here, we explore whether sobriety testing might be feasible within a short period (under 10 seconds) *prior to driving*, using only the sensors available in many modern cars: a driver monitoring camera to capture gaze and eye features, an instrument cluster or console to display messages, and a steering wheel for control feedback. If successful, such tests could help to identify or flag likely-impaired drivers before they begin to drive. They could provide objective feedback to drivers, who often underestimate their own alcohol levels [7], about their readiness to drive. They could even be deployed outside of a driving context, e.g. to monitor employee well-being in industries where sobriety is critical to safety.

Our paper makes the following contributions:

- We design and implement four candidate visuomotor sobriety tests suitable for in-vehicle deployment.
- We run a carefully designed human subjects study with 50 test subjects (20 alcohol-impaired, 30 control).
- We design and evaluate several machine learning models to assess the ability of the tests to discriminate between intoxicated and sober states.

II. RELATED WORK

The impairment of sensory, motor, and intellectual faculties following alcohol consumption has been extensively studied through laboratory, simulated, and on-road driving. For recent comprehensive reviews, see [8]–[10]. Below we review the most relevant work in (i) existing sobriety tests, (ii) gaze and eye behaviors under alcohol intoxication, and (iii) attempts to predict alcohol intoxication from sensory data.

A. Existing Sobriety Tests

The most reliable methods for assessing alcohol intoxication are chemical and invasive: they sample the blood, saliva, or breath of the individual. Although they are accurate and straightforward to administer, they require additional sensors not commonly available in vehicles. We use the most common of these methods, the breathalyzer test, to measure ground truth alcohol intoxication in our dataset.

Behavioral tests, which do not require specialist sensors, are also commonly used: the "Field Sobriety Tests" comprise several which examine for symptoms of intoxication through simple physical tasks (see Fig. 2). These include performing actions such as walking heel-to-toe in a straight line, standing on one leg, and the "horizontal gaze nystagmus" test [11], which involves following an object with the eyes (such as a pen) to examine for characteristic impaired eye movements. While useful in the field, these tests require physical activity

All authors are with the Toyota Research Institute, USA. This work reflects solely the opinions and conclusions of its authors, and not TRI or any other Toyota entity. *Corresponding author. Email: firstname.lastname@tri.global.

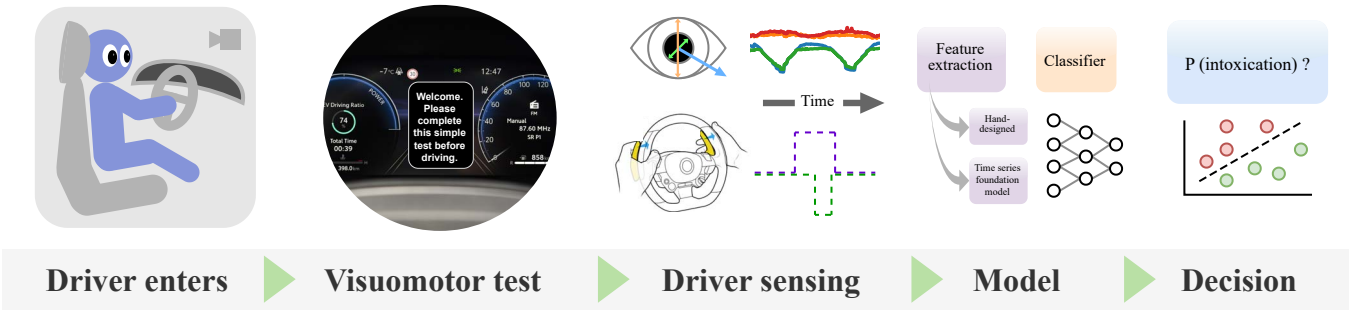


Fig. 1: **Paper summary.** We design a set of visuomotor tests which can be carried out in a vehicle in under 10 seconds, before driving, to assess driver sobriety using a driver monitoring camera. Driver responses to the test are assessed using machine learning models which analyze driver gaze tracking and response times against known sober behavior and predict between sober or impaired (approx. 0.10g/dl Blood Alcohol Concentration) states.

and/or interaction with an examiner, and cannot obviously be transferred to a vehicle environment.

In contrast to these existing tests, we focus on the in-vehicle setting. This affords access to gaze tracking through a driver monitoring camera, with stricter requirements in terms of observation time since the window for collecting observations prior to driving is short. While the tests may require compliance from the driver, the actions required might feasibly be incorporated into the routine of vehicle startup to 'unlock' it for driving.

B. Gaze and Eye Behavior under Alcohol Intoxication

Alcohol intoxication is known to affect a vast range of cognitive, perceptual, attentional, and motor functions, even at low doses [9]. For a summary of its effects on driving behavior, see [5]. Focusing on the visual system, some examples of negatively impacted processes include motion parallax [12], visual scanning [13], vestibulo-ocular reflex [14], [15], and the ability to perform spatial vigilance tasks [16].

In driving, alcohol has been found to generally impair performance on all driving sub-tasks (e.g. lane keeping, brake reactions, situational awareness), but the level of impairment on visual tasks has been related to the information processing demand [9]. A study of the eye movements of alcohol-intoxicated subjects observing realistic traffic scenes showed that alcohol has a significant effect on the latency, velocity and accuracy of saccades, even at low BACs (0.04%-0.06%), reducing the inflow of visual information [17]. Further studies have found significant alcohol-induced differences in gaze behavior, e.g. as measured by gaze transition entropy and stationary gaze entropy [18], although these were in simulated driving within specific driving scenarios.

Given the many ways in which gaze and eye behaviors are affected by alcohol intoxication, developing automatic ways to use those cues for earlier detection of intoxication is useful [19], and made possible by the growing presence of driver monitoring cameras in modern vehicles.

C. Automatic Alcohol Impairment Prediction

Alcohol impairment has been predicted in various contexts, including from gait using smartphone sensors [20] or combinations of smartphone and wrist-worn sensors [21], from facial and gaze features in RGB video [22], and from speech [23]. While using other modalities such as speech or RGB can certainly help in impairment prediction, in this work, we focus on using gaze tracking data, as might be available from a driver monitoring camera. Gaze tracking has the significant benefit that the domain gap between in-lab and in-vehicle environments is relatively narrow (compared to, for example, RGB which can be impacted by external lighting conditions, or speech which can be affected by background noise or music), making results more likely to generalize to a real deployment. The gaze modality also better protects user privacy.

In closely related work to ours, Makowski et al. [24] gathered eye tracking data from 44 participants completing a set of vigilance tasks under sober and intoxicated conditions. Their results demonstrate that contactless inebriation detection based on eye gaze is possible. However, they gather eye tracking data in conditions that are not feasible in a vehicle cabin (2 kHz with participants' heads stabilized using a chin and forehead rest). They conclude that future research is needed to explore the use of lower frequency devices such as the one we use in our study. Here we aim to provide a similar proof of concept within the constraints of the driving application.

The work of [25] explored the use of smartphones to measure how alcohol affects a person's motor coordination and cognition. They designed various "drunk user interfaces" and trained models on human performance metrics and sensor data to estimate a person's BAC, finding that results with high correlation to breathalyzer measurements can be achieved after user-specific learning. Our work is similar, in that we design interactions with the goal of eliciting signs of alcohol intoxication. However, we constrain our application to short interactions in a driving environment, assuming the presence of a gaze tracker, rather than a prolonged interaction with a smartphone. Interactions with smartphones and wearable devices, when available, can be used to gather further evidence of driver state and are complementary to our own approach.

Attempts to identify alcohol-intoxicated drivers in real-time using driving performance data have been explored. For example, performance comparable to the standardized field sobriety tests has been achieved using a variety of machine learning

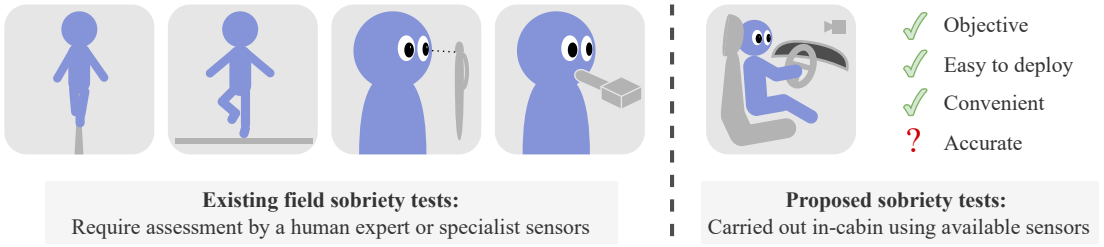


Fig. 2: **Our approach compared to existing field sobriety tests (FSTs).** *Left*: Existing FSTs (walk-and-turn, one-leg stand, horizontal gaze nystagmus, breathalyzer test) are carried out by law enforcement officers. They are time-consuming and subjective. *Right*: We investigate whether tests might feasibly be run without drivers exiting their vehicles, in vehicles fitted with a driver monitoring camera. We design four candidate tests which are objective and relatively quick to administer. We evaluate their accuracy in this paper and encourage further exploration of this design space of in-cabin sobriety testing.

models operating on eight minutes of driving observations [6], although differences between drivers and roadway situations had a large influence on algorithm performance. Other recent work used a logistic regression model over eye, gaze and head movement features extracted from a driver monitoring camera [26]. Our work is again complementary, since we aim to produce the strongest possible prediction for driver intoxication, *before* driving begins. Earlier predictions could help inform downstream driver models and interactions.

III. TEST DESIGN

A. Test Desiderata

An ideal in-vehicle sobriety test should meet four key characteristics:

- *Quick to run:* so it can be carried out prior to driving with minimal disruption to journey time.
- *Intuitive:* easy for a user to understand/perform, with little or no learning curve.
- *Easy to deploy:* not relying on additional sensors (such as breathalyzers), making it cheap to implement.
- *Accurate:* should make impairment at low BAC apparent with as few false positives as possible.

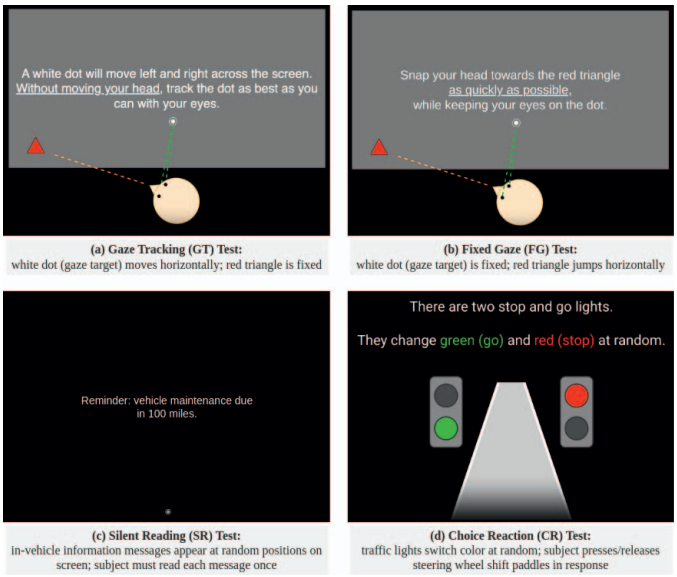


Fig. 3: **Screenshots from instructional videos for our four proposed tests.** The tests were designed to assess various aspects of behavior and decision-making that are impacted by alcohol impairment. Videos of the instructions shown to experiment participants are included in the Supplementary Video.

B. Candidate Tests

Drawing from our understanding of the behavioral impact of alcohol intoxication from Sec. II and the desired characteristics of an in-vehicle sobriety test, we generated a set of four candidate tests, illustrated in Fig. 3. Each test uses visual stimuli such as moving dots and text on an LCD screen, which could reasonably be displayed on an instrument cluster, console or heads-up display in a vehicle.

1) Gaze Tracking (GT) Test: We simplify the horizontal gaze nystagmus (HGN) test [11], which is commonly used in field sobriety testing, to make it more compatible for an in-vehicle deployment. The HGN test requires a subject to track a horizontally moving object at varying speeds to the maximum angle of lateral deviation either side of center. Alcohol intoxication often results in a lack of smooth pursuit and the onset of nystagmus (rhythmical, repetitive and involuntary eye movements) at and beyond 45 degrees of lateral deviation. The test evaluates the ability to smoothly estimate gaze over a wide lateral range, but this is challenging if using centrally-mounted eye trackers (as are common in driver monitoring systems), because gaze estimation accuracy worsens at high lateral angles due to perspective distortion of the pupils. We adapt this test for in-vehicle deployment by testing only one side of vision: subjects tilt their head such that the gaze tracking camera is mounted at about 30 degrees offset from their forward direction. Keeping their head fixed, they track a dot as it moves back and forth across the screen in one hemisphere of their vision. This adjustment means that gaze tracker estimates are improved for higher lateral gaze angles, halves the time taken to administer the test (at the cost of half of the behavioral observations), and halves the screen size required to display the moving target.

2) Fixed Gaze (FG) Test: The "head impulse test" is a commonly used clinical test in which a subject, seated in front of an examiner, is asked to fixate on the examiner's nose while their head is guided in a series of unpredictable, limited amplitude but high-velocity movements [27]. This tests for vestibulo-ocular reflex (VOR), the mechanism by which vision is stabilized during head motion by coordinating eye

movements in the opposite direction. Since VOR is known to be impacted at low levels of BAC [14], [28], it is an effective method for determining alcohol intoxication. It is not possible, or safe, to physically guide a driver's head through high-velocity movements, but we explore whether VOR impairment can be determined by asking subjects to quickly move their heads in response to random left and right stimuli. With a normally functioning VOR, the driver should be able to maintain their gaze on a fixed target, in a central position near the gaze tracking camera.

3) Silent Reading (SR) Test: A study into the effect of alcohol on reading found that the number and duration of eye fixations increased significantly with increased breath alcohol concentration [29]. Since reading is a natural task to perform in a vehicle, we explore whether this effect is significant enough to be detected through short observations of the driver reading messages on an instrument cluster. We create a set of realistic messages related to vehicle use.

4) Choice Reaction (CR) Test: Recent work on designing mobile user interfaces to detect alcohol intoxication [25] found that the best performing discriminatory task involved reacting to a divided attention stimulus. A task involving attention, peripheral vision and rapid motor control can place sufficient information processing load on subjects to make alcohol impairment evident at lower BAC. We adapt this type of task to a driving setting: we design a choice reaction test in which two traffic light stimuli are presented at random to the driver, and the driver responds by pressing or releasing two steering wheel buttons in response.

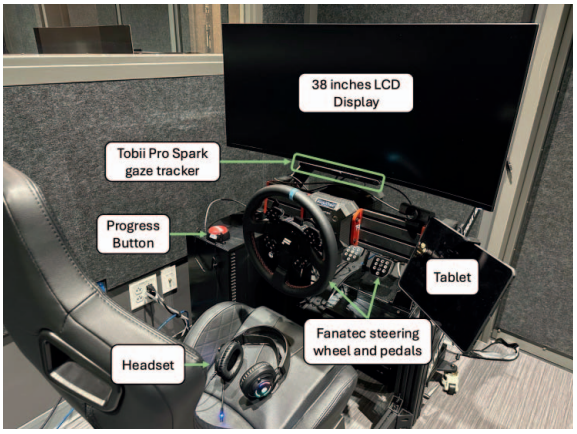
C. Implementation

Our tests were implemented using PyQt, a Python binding of the GUI toolkit Qt. Each test was preceded with an instructional video, explaining the goal of the test to participants (see Supplementary Video). To improve ease of understanding and execution, the videos and the tests were tested and improved with five sober pilot study participants, whose data was discarded. Sensor recordings from the tests were captured and synchronized using ROS2 [30]. All of the tests involved multiple repeats: the gaze tracking test involved multiple horizontal scans, while the other three tests each contained 20 iterations of the action-provoking stimulus during a single run, yielding larger sample sizes.

IV. DATA COLLECTION

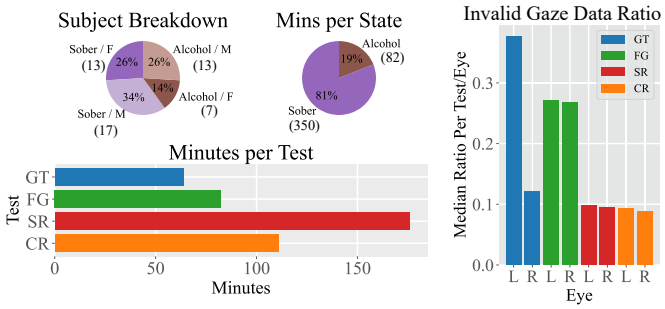
A. Collection Setup and Protocol

Each subject participated in two sessions in a driving simulator, shown in Fig. 4a. For tracking eye and gaze behaviors, we used a Tobii Pro Spark desktop gaze tracker, chosen for its ease of use and availability while having a comparable spec to modern driver monitoring systems. During each session, subjects first completed an 8-point gaze calibration procedure, then watched instructional videos for and executed each of the four tests once, taking approximately 10 minutes in total.



(a) Driving simulator setup

Measurement	Sensor	Details
3D gaze measurements	Tobii Pro Spark	60 Hz
Pupil diameter	Tobii Pro Spark	60 Hz
Vehicle control inputs	Fanatec Gran Turismo DD Pro	async.



(b) Dataset summary statistics

Fig. 4: **Dataset overview**. Using (a) a driving simulator with a Tobii Pro Spark gaze tracker, we captured data of sober and alcohol-impaired subjects completing a set of visuomotor tests while seated in a vehicle. We captured test data from (b) 50 subjects (30 sober, 20 alcohol-impaired), comprising 7.2 hours of gaze tracking data and vehicle control inputs.

Between sessions, subjects participated in other driving experiments (described in [31]) and rested to minimize the effects of fatigue.

Session 1 was completed in a sober state by all 50 subjects. Session 2 was completed after alcohol consumption for 20 of the subjects, while the remaining 30 completed it sober. Alcohol was administered in shots of 80 proof (40% Alcohol By Volume) spirits selected by the subject, to a target BAC of 0.10% using the Widmark formula. BAC was monitored using a Alco-Sensor FST breathalyzer breathalyzer until it had reached the target concentration or peaked, at which point the test was administered. BACs were recorded for alcohol-impaired subjects just prior to the second session and again, 15 minutes after its completion, yielding an average BAC across participants of $0.096 \pm 0.021\%$.

B. Safety and Ethical Considerations

Subjects who were intoxicated as part of the study were monitored after study completion, and released only once their BAC had fallen to safe levels (where "safe" was determined

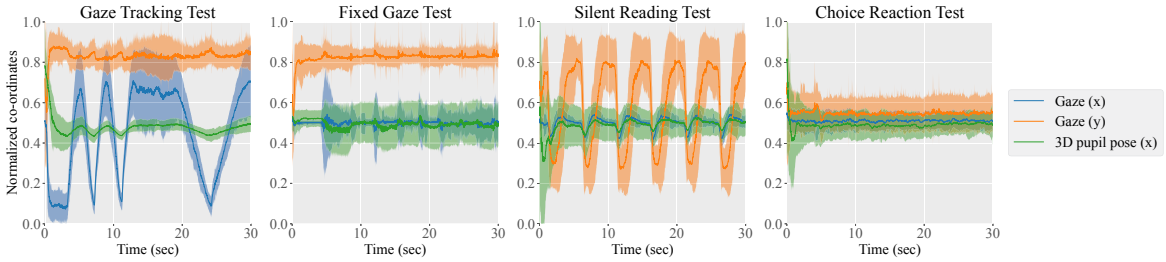


Fig. 5: **Mean and standard deviation gaze location** (normalized x and y position on screen) across all subjects for first 30s of each test.

based on the elected mode of transport for each participant). During the study, any subjects who felt uncomfortable were allowed to cease the study immediately, and subjects who were not comfortable with the target amount of alcohol were given the chance to opt out of consuming alcohol. From an initial pool of 62 subjects (including pilot test subjects), we obtained a final dataset of 50 subjects who completed both Sessions correctly. The test protocol was approved through an Institutional Review Board with WCG (IRB Protocol #20241945). Participants were aged 23-65 (average 35.2) and provided written informed consent. While we made best efforts to obtain a diverse subject pool, a larger dataset would be necessary to properly assess for certain biases such as age and ethnicity.

C. Dataset

Summary statistics from the final collected dataset are shown in Fig. 4b. In total, we collected over seven hours of data from participants (20 Female/30 Male, 19% alcohol-impaired/81% sober) and 100 completions of each test. The ratio of invalid to valid gaze data was found to be significantly higher on average during the Fixed Gaze test and for subjects' left (L) eyes in the Gaze Tracking test. This was due to the severe head motions required to complete the tests, which caused failures within the gaze tracker. Fig. 5 shows the mean and standard deviation of gaze location for a 30 second portion of each of the four tests, averaged over all participants, highlighting the variety of behaviors tested and the uniformity of eye motions across subjects within each test.

V. MODEL

The goal of our model is to best infer an individual's impairment state from a short window of observational test data. The high noise levels in our raw data (as shown by higher invalid gaze data ratios in Fig. 4b) pose a challenge to typical gaze feature extraction pipelines. We opt to use a deep learning approach, which is known to reliably yield strong performance at noisy pattern classification tasks given sufficient training [32]. Our model, shown in Fig. 6, takes two samples of data, one from a known sober (baseline) state, and one from an unknown (test) state, extracts features and compares the features to determine the state prediction. Code for the models tested will be made available for reproducibility; here we summarize the key components of the model at a high level:

Random sampler, R. The random sampler extracts a windowed segment of W time steps from a given subject's Session 1 and Session 2 data for a single test type. During training, the samples are extracted at the same randomly chosen point, with some time jitter, and sample pairs from sober subjects have their session order flipped at random 50% of the time. Left or right eye data is also selected at random during training time, to create a model with lower dimensional input and double the training data. Since the dataset is relatively small, we limit the dimensionality of the data, D, to 4: left or right eye 2D gaze position on screen, gaze tracking validity, and an event message, which provides the timing of test stimuli and button responses.

Feature embedding model, F. The $D \times W$ samples are fed to a feature embedding model, F. We explore two types of F: (1) a deep convolutional neural network, trained from a random initialization; and (2) a state-of-the-art foundation model for general-purpose time-series analysis [33], which is frozen during training. The latter has the significant advantage of being pre-trained through masked time-series prediction on large and diverse time series data from many domains, meaning it is likely to be able to capture useful temporal features in noisy data.

Comparison network, C. The Comparison network takes paired input features from F and combines them to output a predicted Session 2 state (where the correct state should be $y = 1$ for alcohol-impaired, $y = 0$ otherwise). We find that direct differencing (in which Session 2 features are subtracted from Session 1 features at the input of the network) empirically works better than feature concatenation, so we adopt this approach for all experiments.

A. Training

From a paired sober (x_1) and unknown (x_2) input sample, our model computes the predicted unknown state $\hat{y}$:

$$\hat{y} = C\Big(F\big(R(x_1)\big), F\big(R(x_2)\big)\Big). \tag{1}$$

We minimize the weighted binary cross-entropy loss against ground truth y:

$$\mathcal{L}(y, \hat{y}) = -\big(w_1 y \log(\hat{y}) + w_0 (1-y) \log(1-\hat{y})\big), \tag{2}$$

where weights w are set according to normalized inverse class frequency, accounting for the class imbalance between sober

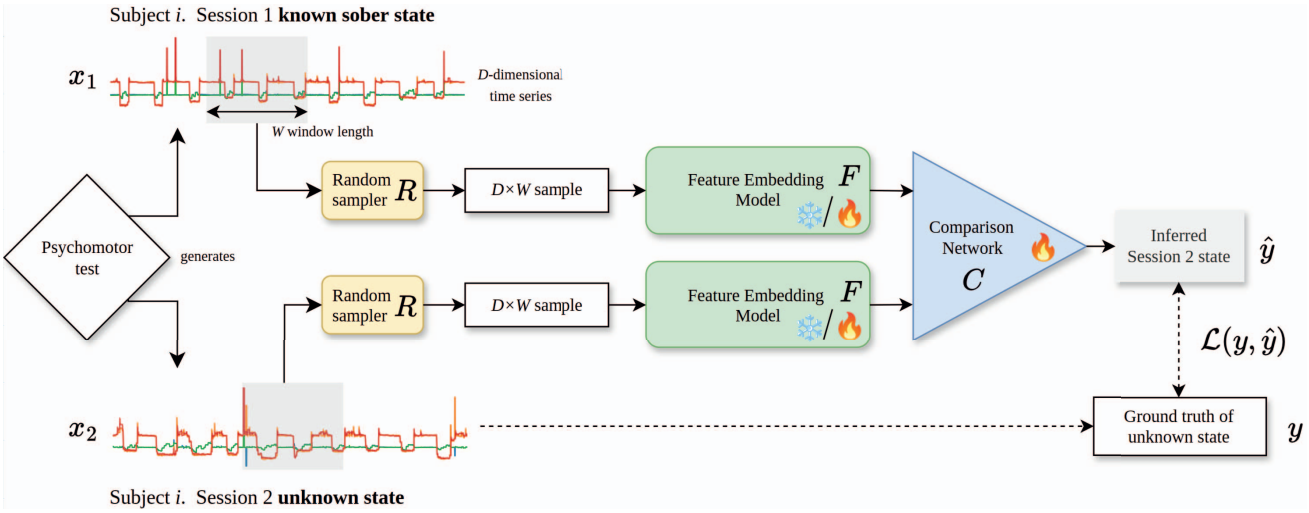


Fig. 6: **Model overview**. Each Subject i sits a test in a known sober state (during Session 1) and an unknown second state (during Session 2). A Random sampler, R, extracts random, augmented samples from the data from each test, and passes them to a Feature Embedding Model, F, which captures different temporal characteristics in the signal. A Comparison Network, C, has access to information from both embeddings to produce an estimate of the inferred state of Session 2. During training, the Session 2 state is known and used to backpropagate prediction errors through C and optionally F.

and impaired training data. All models are implemented in PyTorch [34] and trained using AdamW [35] on a single NVIDIA Quadro RTX 6000 GPU. For the pre-trained model, we use the 'MOMENT-1-small' version [33]. We set window length $W = 512$. This corresponds to 8.5 seconds of observational data, which we assume is a reasonable duration for a pre-driving test to minimize inconvenience to a driver.

B. Evaluation

We are interested in the generalization performance of our model to new test subjects, so we adopt a five-fold cross-validation approach with held-out subjects. We measure performance using Balanced Accuracy (BAcc - which takes class imbalance into account) and F1 scores, averaged across held-out folds. Test samples are taken from the first 30s of each test, and left and right eye predictions, $\hat{y}_l$ and $\hat{y}_r$, are combined using a max function $\hat{y} = \max(\hat{y}_l, \hat{y}_r)$ for simplicity.

VI. EXPERIMENTS

A. CNN vs. Foundation Model

For the feature embedding model, F, we compared the average performance across all data of a high-capacity transformer model pre-trained on diverse time-series data [33], with the performance of a CNN trained from scratch. We found that, despite exploring different CNN architectures (varying depth, width, normalization), we were unable to make the CNN approach perform on average any better than chance (F1 = 0.42 ± 0.06, BAcc = 0.49 ± 0.03. However, the transformer model succeeded in finding a signal in at least some of the tests (F1 = 0.54 ± 0.06, BAcc = 0.57 ± 0.06). It is possible that the training set size is insufficient to support learning features from scratch, although further exploration of the model space is possible in future work. Given the success of the transformer model, we adopted it for subsequent experiments.

B. Individual Test Performance

We next explored the breakdown of individual test performance for the transformer model. We trained on data exclusively from each visuomotor test, as well as from all tests combined. Fig. 7a and Fig. 7b show the training set vs. test performance, averaged over 5 training runs with different random seeds. Best performance was observed on the Choice Reaction test, while other tests were at or slightly above chance. This does not imply that the other sobriety tests did not elicit observable changes in behavior, but that such changes if any were too infrequent or too subtle to be reliably detected using our model.

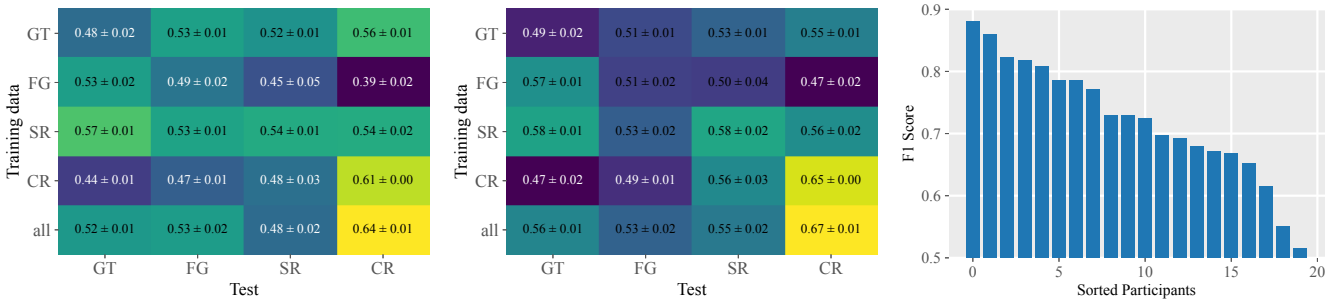
Training on all data did not appear to yield a significant benefit compared to training on same-domain data. However, training on data from the Silent Reading test (where gaze is both diverse and consistently tracked) appeared to be most beneficial to learn a general-purpose signal, sometimes even out-performing same-domain training. This suggests that diverse high-quality data is most helpful to learn useful features.

While the detection of alcohol impairment is far from solved by even the best-performing of our models, we note the wide distribution of F1 score across subjects (Fig. 7c). This implies that certain subjects do not exhibit easily detectable effects of alcohol impairment as readily as others. The personalization of models and prediction of model efficacy based on individuals may be an interesting direction for future research.

C. Sensitivity Analyses

We next evaluated the sensitivity to various practical capture and input variables of the test with most promising prediction performance (Choice Reaction, CR).

Effect of sampling rate. While we collected gaze tracking data at 60 Hz, a common sampling rate for current generation driver monitoring cameras is 30 Hz or below. We



(a) F1 score (mean ± std)

(b) Balanced accuracy (mean ± std)

(c) F1 Breakdown by participant

Fig. 7: **Results**. *Left and middle*: Test performance breakdown using models trained on different data splits showing out-performance of the Choice Reaction test with modest results for other tests. *Right*: Sorted F1 scores by participant showing diversity of performance.

found in Fig. 8a that the Choice Reaction test maintains reasonable performance at sampling rates down to 20 Hz, suggesting feasible use with a lower frame-rate tracker.

Effect of varying observation window W. We next evaluated the effect of varying input window W on performance, finding that longer observation windows are critical to gather evidence for impairment detection, as might be expected (Fig. 8b). Expanding the input window would likely improve performance further, up to a saturation point.

Effect of varying input dimensionality D. Finally, we examined the effect of adding and removing input information. Our base model uses gaze and test event information, which contains the timings of test stimuli and, for the Choice Reaction test, the physical input response. Fig. 8c shows the effect of removing gaze, or adding pupil diameter data to the input. We found that gaze features extracted by the model appear to be critical to performance, while a smaller input dimensionality may help with training in a small data regime.

D. Summary

Quantitative assessment. Of the four tests we have proposed and evaluated, the Choice Reaction yielded the strongest detectable signal for alcohol impairment detection, with a best balanced accuracy of 0.67. This is comparable to some existing non-breathalyzer field sobriety tests (one-leg stand = 0.65, walk-and-turn = 0.68, horizontal gaze nystagmus = 0.77, for BAC $>$ 0.10% [36]), while only requiring 10 seconds of driver time inside the vehicle. We found that the tests could be used on a lower frame-rate gaze tracker without substantial performance degradation, but that larger observation windows (i.e. longer tests) were preferable to maximize performance.

The other three tests did not yield as promising results, perhaps due to insurmountable gaze tracking noise (in the case of the gaze tracking and fixed gaze tests) or the short observation window for passive tasks such as reading. Further work is needed to investigate whether more reliable signals may be derived from these tests, or whether other modalities of data such as video would be beneficial.

Qualitative assessment. In Sec. III we listed the design criteria for an ideal in-vehicle sobriety test. After implementing each test and running them twice over 50 subjects, we found that the Choice Reaction test would be best overall. The Choice Reaction test could be run using just stimuli displayed on an instrument cluster, making it easy to deploy. Although less intuitive than other tasks such as reading (requiring some initial explanation about how to respond to the stimuli), it showed the most promising discriminative performance, perhaps because it requires rapid motor responses from the user and as a result can elicit detectable effects of alcohol intoxication more readily.

VII. CONCLUSIONS

We have proposed and evaluated a set of automated sobriety tests that can feasibly be deployed in modern passenger vehicles equipped with gaze-tracking technology. Of the four candidate tests, one in particular showed promising results, comparable to existing non-breathalyzer field sobriety tests. To further build on our approach, dataset and findings, future research should employ larger, more diverse datasets gathered under a range of field conditions (e.g. varied lighting, subject appearance, in real vehicles using production driver monitoring systems), while also incorporating longitudinal data to account for factors such as fatigue, mood, compliance over time, and other individual differences. Ultimately, trying to leverage existing in-vehicle sensor technologies to assess driver impairment may help to bring about faster and more effective measures against alcohol-impaired driving.

Acknowledgments. We thank Laporsha Dees and Emily Sumner for their assistance running human subjects trials, Todd Rowell and Thomas Balch for supporting the simulator setup, and Jean Costa and Hiro Yasuda for helpful discussions.

REFERENCES

[1] World Health Org., "Global status report on road safety," 2023.
[2] National Highway Traffic Safety Administration (NHTSA), "Advanced notice of proposed rule making: Advanced impaired driving prevention technology," *Docket No. NHTSA-2022-0079*, 2023.
[3] ——, "Instructor guide: DWI detection and standardized field sobriety testing (SFST) refresher," 2015.
[4] EuroNCAP, "Vision 2030: a safer future for mobility," 2022.
[5] National Highway Traffic Safety Administration (NHTSA), "Advanced impaired driving prevention technology," Jan. 2024.
[6] J. D. Lee, D. Fiorentino, M. L. Reyes, T. L. Brown, O. Ahmad, J. Fell, N. Ward, and R. Dufour, "Assessing the feasibility of vehicle-based sensors to detect alcohol impairment," *Washington, DC: National Highway Traffic Safety Administration*, vol. 1, no. 2, p. 7, 2010.

(a) Effect of sampling rate

(b) Effect of observation window W

(c) Effect of input dimensionality D

Fig. 8: **Choice Reaction (CR) test sensitivity analyses**. *Left*: Prediction performance is reasonably robust to a reduction in eye tracking sampling rate. *Middle*: A longer observation window W is essential for performance, at the cost of increased inconvenience to the subject. *Right*: Adding pupil diameter data does not help performance; the model does not rely solely on reaction times (the "event" channel) to predict test outcomes.

[7] M. E. Rossheim, D. L. Thombs, K. M. Gonzalez-Pons, J. A. Killion, J. D. Clapp, M. B. Reed, J. M. Croff, D. E. Ruderman, and R. M. Weiler, "Feeling no buzz or a slight buzz is common when legally drunk," *Amer. j. of public health*, vol. 106, no. 10, p. 1761, 2016.

[8] T. L. Martin, P. A. M. Solbeck, D. J. Mayers, R. M. Langille, Y. Buczek, and M. R. Pelletier, "A review of alcohol-impaired driving: The role of blood alcohol concentration and complexity of the driving task," *Journal of Forensic Sciences*, vol. 58, no. 5, pp. 1238–1250, 2013.

[9] H. Moskowitz and D. Florentino, "A review of the literature on the effects of low doses of alcohol on driving-related skills, Tech. Rep. DOT-HS-809-028, Apr. 2000.

[10] M. Dong, Y. Y. Lee, J. S. Cha, and G. Huang, "Drinking and driving: A systematic review of the impacts of alcohol consumption on manual and automated driving performance," *Journal of Safety Research*, vol. 89, pp. 1–12, June 2024.

[11] S. J. Rubenzer and S. B. Stevenson, "Horizontal gaze nystagmus: A review of vision science and application issues," *Journal of Forensic Sciences*, vol. 55, no. 2, pp. 394–409, Mar. 2010.

[12] M. Nawrot, "Depth perception in driving: Alcohol intoxication, eye movement changes, and the disruption of motion parallax," *Driving Assessment Conference*, vol. 1, no. 2001, Aug. 2001, number: 2001 Publisher: University of Iowa.

[13] B. Shiferaw, C. Stough, and L. Downey, "Drivers' visual scanning impairment under the influences of alcohol and distraction: A literature review," *Current Drug Abuse Reviews*, vol. 7, no. 3, pp. 174–182, Dec. 2014.

[14] H. G. Choi, S. K. Hong, S. K. Park, H.-J. Lee, and J. Chang, "Acute alcohol intake impairs the velocity storage mechanism and affects both high-frequency vestibular-ocular reflex and postural control," *International Journal of Environmental Research and Public Health*, vol. 19, no. 7, p. 3911, Jan. 2022.

[15] R. B. Post, L. A. Lott, J. I. Beede, and R. J. Maddock, "The effect of alcohol on the vestibulo-ocular reflex and apparent concomitant motion," *Journal of Vestibular Research: Equilibrium & Orientation*, vol. 4, no. 3, pp. 181–187, 1994.

[16] H. S. Koelega, "Alcohol and vigilance performance: a review," *Psychopharmacology*, vol. 118, no. 3, pp. 233–249, Apr. 1995.

[17] A. Buser, B. Lachenmayr, F. Priemer, A. Langnau, and T. Gilg, "Effect of low alcohol concentrations on visual attention in street traffic," *Der Ophthalmologe: Zeitschrift Der Deutschen Ophthalmologischen Gesellschaft*, vol. 93, no. 4, pp. 371–376, Aug. 1996.

[18] B. A. Shiferaw, D. P. Crewther, and L. A. Downey, "Gaze entropy measures detect alcohol-induced driver impairment," *Drug and Alcohol Dependence*, vol. 204, p. 107519, Nov. 2019.

[19] E. Tivesten, V. Broo, and M. L. Aust, "The influence of alcohol and automation on drivers' visual behavior during test track driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 95, pp. 215–227, May 2023.

[20] Z. Arnold, D. Larose, and E. Agu, "Smartphone inference of alcohol consumption levels from gait," in *International Conference on Healthcare Informatics*, Oct. 2015.

[21] B. Nassi, J. Shams, L. Rokach, and Y. Elovici, "Virtual breathalyzer: Towards the detection of intoxication using motion sensors of commercial wearable devices," *Sensors*, vol. 22, no. 9, p. 3580, 2022.

[22] E. Keshtkaran, B. von Berg, G. Regan, D. Suter, and S. Z. Gilani, "Estimating blood alcohol level through facial features for driver impairment assessment," in *WACV*, 2024, pp. 4539–4548.

[23] A. A. Bonela, Z. He, A. Nibali, T. Norman, P. G. Miller, and E. Kuntsche, "Audio-based deep learning algorithm to identify alcohol inebriation (ADLAIA)," *Alcohol*, vol. 109, pp. 49–54, 2023.

[24] S. Makowski, A. Bätz, P. Prasse, L. A. Jäger, and T. Scheffer, "Detection of alcohol inebriation from eye movements," *Procedia Computer Science*, vol. 225, pp. 2086–2095, 2023, 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023).

[25] A. Mariakakis, S. Parsi, S. N. Patel, and J. O. Wobbrock, "Drunk User Interfaces: Determining blood alcohol level through everyday smartphone tasks," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[26] K. Koch, M. Maritsch, E. Van Weenen, S. Feuerriegel, M. Pfäffli, E. Fleisch, W. Weinmann, and F. Wortmann, "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*. Hamburg Germany: ACM, Apr. 2023, pp. 1–32.

[27] E. Armato, "What is the head impulse test?" [Online]. Available: https://www.audiologyonline.com/ask-the-experts/inventis-what-head-impulse-test-28458

[28] T. N. Roth, K. P. Weber, V. G. Wettstein, G. B. Marks, S. M. Rosengren, and S. C. A. Hegemann, "Ethanol consumption impairs vestibulo-ocular reflex function measured by the video head impulse test and dynamic visual acuity," *Journal of Vestibular Research: Equilibrium & Orientation*, vol. 24, no. 4, pp. 289–295, 2014.

[29] R. Watten and I. Lie, "Effects of alcohol on eye movements during reading," *Alcohol and alcoholism (Oxford, Oxfordshire)*, vol. 32, pp. 275–80, May 1997.

[30] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot Operating System 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, May 2022.

[31] J. Gideon, K. Tamura, E. Sumner, L. Dees, P. Reyes Gomez, B. Haq, T. Rowell, A. Balachandran, S. Stent, and G. Rosman, "A simulator dataset to support the study of impaired driving," 2025. [Online]. Available: https://toyotaresearchinstitute.github.io/IDD/

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[33] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A family of open time-series foundation models," in *ICML*, 2024.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *NeurIPS*, vol. 32, 2019.

[35] I. Loshchilov, F. Hutter, *et al.*, "Fixing weight decay regularization in Adam," *arXiv preprint arXiv:1711.05101*, vol. 5, 2017.

[36] International Assoc. of Chiefs of Police, "Standardized field sobriety testing," *NCJ Number 106113*, 1987.