Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Investigating Driver-Specific Variability in Vision-Based Driver Fatigue Monitoring: A Subgroup Evaluation Design Tested in a Controlled Simulation

**Prasad Tribhuvan**

prasadtribhuvan47@gmail.com

Sanjivani University

**Research Article**

**Keywords:** Driver Fatigue Monitoring, Computer Vision, Driver-Specific Variability, Eye Aspect Ratio (EAR), PERCLOS, Personalized Calibration, Subgroup Evaluation, Vision-Based Driver Monitoring, Simulation-Based Analysis, Intelligent Transportation Systems

**Posted Date:** November 4th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11578167/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Investigating Driver-Specific Variability in Vision-Based Driver Fatigue Monitoring: A Subgroup Evaluation Design Tested in a Controlled Simulation

**Prasad Rajesh Tribhuvan**

Department of Artificial Intelligence and Data Science, Sanjivani University, Ahilyanagar, Maharashtra, India

Corresponding author: tribhuvanprasad47@gmail.com

## Abstract

Vision-based driver fatigue monitors are usually validated on pooled data in which the same or similar drivers appear in both training and test sets, so a single headline accuracy hides how performance differs from one driver to another. This paper treats driver-specific variability (eye geometry, eyewear, head pose and illumination) as an explicit experimental factor and contributes a reusable evaluation design: four annotated subgroup factors, subject-disjoint splits, per-subgroup metrics, a null-model check, an effect-size sweep and sample-size planning. The quantitative part is a controlled simulation of 240 drivers replicated 30 times, in which a fixed-threshold rule and a per-driver calibrated rule are compared per subgroup, confirmed by an independent re-implementation that agrees within about 0.05. In the simulation, eye geometry mainly affected the fixed rule, whereas sunglasses and illumination mainly affected the calibrated rule, and head pose had little effect at the assumed magnitudes. The weakness of the calibrated rule with sunglasses is largely a design-parameter effect, and about 8, 16–20 and 30–40 test drivers per subgroup are needed to detect large, moderate and small effects. The findings are conditional on the model's assumptions and are not evidence about real drivers; they define hypotheses, a reporting checklist and sample requirements for a planned real-data study.

**Keywords:** driver fatigue detection; driver-specific variability; eye aspect ratio; PERCLOS; per-driver calibration; subject-independent evaluation; simulation

## 1. Introduction

Drowsy driving is a recognised contributor to road crashes, which has motivated extensive work on non-intrusive monitoring from in-cabin cameras. Eye closure, blink behaviour, yawning and head pose can be obtained from a standard camera with facial landmark detectors, and recent deep-learning systems report high accuracy [3]–[5]. Most results are summarised as one number over a pooled test set. A driving population is not homogeneous: drivers differ in eye opening, wear spectacles or sunglasses, sit and turn differently, and drive under very different lighting. A rule that performs well on average can behave differently for particular groups of drivers, and a pooled figure cannot reveal it.

This paper makes driver-specific variability the object of study. Its objectives are to (O1) investigate how this variability affects the generalisability of vision-based fatigue detection, (O2) identify and compare the driver-related factors involved, and (O3) review how public datasets and protocols represent them. The scope of the evidence is stated plainly: the quantitative experiments are simulations designed to expose mechanisms and to specify the analysis that real data will need; they do not estimate the accuracy of any system on real drivers. The contributions are a subject-disjoint, per-subgroup evaluation design with four annotated factors; a documented generative simulation with a null-model check and effect-size sweep comparing a fixed-threshold and a per-driver calibrated rule; a review of five public datasets; extended analyses from an independent re-implementation (closure-fraction sensitivity, sample-size curves and a subject-level bootstrap demonstration); and a subgroup-reporting checklist.

## 2. Related Work

Kolus [1] reviewed 41 studies of drowsiness detection from eye-activity measures and called for standardised evaluation and more generalised systems; Fu et al. [2] named bias, real-time performance and hardware limits as open problems and argued for personalised detection; Albadawi et al. [7] reviewed a decade of systems by signal type. Learning-based detectors such as DrowsyDetectNet [3], transfer-learning approaches

[4] and real-time deep-learning systems [5] centre on accuracy and efficiency, leaving the effect of eye shape, head pose or eyewear across drivers open as a separate factor.

Ersoy et al. [6] compare generalised EAR/MAR thresholds with per-driver calibrated thresholds and report that personalisation improves accuracy, which supports the premise that driver characteristics matter. Herath et al. [8] benchmark YOLOv5–v11 on UTA-RLDD, Sawant et al. [9] implement MediaPipe-based EAR monitoring with one fixed threshold, and Essahraui et al. [10] compare KNN, SVM, CNN, YOLO and Faster R-CNN across NTHU-DDD, YawDD and UTA-RLDD; the last is organised by algorithm, and attributing differences to driver characteristics is the natural next step. Across this literature, evaluation is aggregate, subject-dependent, or organised by algorithm. The present work adds explicit driver subgroups on subject-disjoint splits with per-subgroup, subject-level reporting; a null-model check and effect-size sweep; a quantified view of how fast subgroup samples thin out; and a reporting checklist.

# 3. Methods

### *3.1 Research questions, hypotheses and dataset review*

**Research questions.** RQ1: How does variability across drivers change performance under subject-independent evaluation? RQ2: Which of eye geometry, eyewear, head pose and illumination has the largest effect? RQ3: How do existing datasets represent these factors? **Hypotheses.** For each factor F, $H_0$: subject-level accuracy does not differ across the subgroups of F; $H_1$: it differs. A second pair compares the two classifiers on paired subjects. The hypotheses are evaluated inside the simulation, where effects are assumed by construction; a rejection shows that the pipeline can detect an effect of the assumed size, not that the effect exists in real drivers.

Table 1 summarises five public resources that are candidates for the real-data study. Entries are taken from the dataset papers and will be re-verified against official documentation and licences at access time.

**Table 1.** Candidate datasets and the driver factors they expose.

| Dataset | Size / setting | Eyewear and light | Label source |
|---|---|---|---|
| NTHU-DDD [14] | 36 subjects; simulated driving | glasses, sunglasses; day and night (IR) | annotated behaviours (normal, slow blink, yawn, nod, look aside) |
| YawDD [15] | drivers in cars; dash and mirror cameras | with / without glasses; in-car light | yawning vs talking vs silent |
| UTA-RLDD [16] | 60 participants, ≈30 h; own cameras | uncontrolled home / office | self-reported alert, low vigilance, drowsy |
| DROZY [17] | 14 subjects; laboratory sleep deprivation | controlled NIR video | KSS with EEG / EOG / ECG |
| MRL Eye [18] | ≈85 k eye images, 37 subjects | glasses flag, reflection and light level annotated | eye open / closed only |

*Sizes as reported in the dataset papers. Standardised skin-tone or head-pose-angle annotation is not known to the author to be provided by any listed resource.*

Four observations follow. (i) Eyewear and illumination are explicit in some resources (NTHU-DDD, MRL Eye) but only implicit in others, so they would be annotated per subject. (ii) Head pose and appearance are not annotated in a standard way; skin tone would require a documented, ethically reviewed procedure. (iii) "Drowsy" means different things (annotated behaviour, self-report, physiological sleepiness), so labels from different sources should be reported separately. (iv) The video datasets with a stated size contain 14–60 subjects, and the subject, not the frame or clip, is the independent unit, so subgroup inference is limited by the number of subjects per subgroup. The recommended protocol is therefore: subject-disjoint splits; calibration data separated from evaluated windows; metrics aggregated per subject and reported per subgroup; and uncertainty from a subject-level (cluster) bootstrap.

### *3.2 Simulation design and generative driver model*

Four factors define the subgroups: eye geometry (open-eye EAR $E_0$ below 0.27, 0.27–0.33, above 0.33; a proxy for facial appearance), eyewear (none, spectacles, sunglasses), head pose (frontal, moderate, large) and illumination (normal, low, night). Skin tone is not simulated because it acts through image contrast and cannot be represented credibly without real imagery. The eye aspect ratio is $EAR = (\|p_2 - p_6\| + \|p_3 - p_5\|) / (2\|p_1 - p_4\|)$ [11], computed from six eye landmarks (for example from a MediaPipe face mesh [13]). Each simulated driver has an open-eye level $E_0$, a blink rate and fixed factor levels. At 15 frames per second the

true eye state $e_t$ is open ($E_0$) or closed (0.25 $E_0$); alert windows have brief blinks (0.10–0.30 s), drowsy windows have slow blinks (0.35–1.2 s at 60–100% of the driver's rate) and a partial lid droop to 72–92% of $E_0$. The measured value is

$$m_t = P + (1 - c)[\, e_t \cdot g(\varphi_t, \psi_t) - P \,] + \varepsilon_t \quad (1)$$

where $g = \cos\varphi \cdot [1 + 0.5(1 - \cos\psi)]$ models pitch $\varphi$ and yaw $\psi$ distortion, c is an eyewear "compression" toward a generic face prior P = 0.26 (the detector partly guesses a hidden eye), and $\varepsilon_t \sim N(0, \sigma Z)$ with

$$\sigma = \sigma_0 \sqrt{(\, a^2 + b^2 + q^2 - 2 \,)} \quad (2)$$

combining eyewear (a), illumination (b) and pose (q) noise factors, $\sigma_0 = 0.012$. Frames are dropped with a probability that depends on illumination and pose. All parameters (Table 2) are plausible assumptions, not measurements. A scale s multiplies every departure from the neutral level: s = 0 removes all factor effects, s = 1 is the reference model and s = 2 doubles them. The pose noise factor q was not listed for the main run; the independent re-implementation (Section 4.4) fixes q = 1/1.15/1.4 for frontal/moderate/large pose and agrees with the main run within about 0.05, so the conclusions do not depend on this choice.

**Table 2.** Simulation parameters (all assumed).

| Parameter | Value at s = 1 |
|---|---|
| Population | 240 drivers per replicate; 24 windows (12 alert, 12 drowsy) of 10 s; 60 s alert calibration |
| Open-eye EAR $E_0$ | N(0.30, 0.04Z), clipped to [0.20, 0.42]; blink rate N(16, 4Z)/min |
| Eyewear | none/spec./sun. prob. 0.6/0.3/0.1; noise factor 1/1.4/2.5; compression c = 0/0.10/0.55 |
| Pose | frontal/mod./large prob. 0.5/0.35/0.15; yaw SD 5/15/30°; pitch SD 3/8/15°; frame drop 0/1/5%; q 1/1.15/1.4 (re-implementation) |
| Light | normal/low/night prob. 0.5/0.3/0.2; noise factor 1/2/3.5; frame drop 0/3/12% |

### *3.3 Classifiers, statistics and DriverVision demonstration*

**C0** (fixed threshold) counts a frame as closed when $m < 0.20$ for every driver. **C1** (per-driver calibrated) first estimates the driver's open-eye level $\hat{E}_{open}$ as the median measured EAR over 60 s of alert driving, then counts a frame as closed when $m < \alpha \hat{E}_{open}$; the closure fraction $\alpha = 0.6$ was fixed a priori and not tuned. In both cases a 10 s window is flagged drowsy when PERCLOS [12] reaches a cut-off $\tau$ chosen on the training drivers only by maximising balanced accuracy. Drivers are split 50/50 by subject, so no test driver appears in training. Per-driver accuracy, recall and false-positive rate (FPR) are averaged within each subgroup; windows with fewer than 30% valid frames are discarded.

In each of 30 replicates a Kruskal–Wallis test compares per-driver accuracy across the three subgroups of each factor, with $\varepsilon Z = H(n+1)/(nZ-1)$ as effect size, and C1 and C0 are compared on paired drivers with a Wilcoxon signed-rank test. Results report the share of replicates with $p < 0.05$ and the median $\varepsilon Z$; intervals (2.5–97.5 percentiles over replicates) reflect simulation noise only. Two checks guard against over-reading: a null model (s = 0) and a sweep over s = 0, 0.5, 1, 1.5 and 2. Random seeds are fixed (7 for the re-implementation, 11 for resampling).

DriverVision is the author's browser-based EAR/MAR prototype (468-point face mesh [13], about 29–30 frames per second) with fixed thresholds EAR > 0.25 (alert) and EAR < 0.20 (drowsy). In one live session by the author (one driver, one camera position, one lighting condition) it displayed ALERT at EAR 0.304, DROWSY at 0.188 and MICROSLEEP at 0.170. This is a qualitative functional check only: it does not establish accuracy, reliability or validity, and a single driver is exactly the case in which per-driver differences remain invisible, which motivates the subgroup analysis. No webcam images are included.

## 4. Results

### *4.1 Mechanisms*

Figure 1 shows simulated EAR traces for four drivers. The narrow-eyed driver (B) has an alert open level of about 0.22, so the fixed 0.20 line sits almost on the signal, while the calibrated line adapts to the driver. With sunglasses at night and a large pose (D) the signal is noisy and compressed, and neither threshold separates closure cleanly.

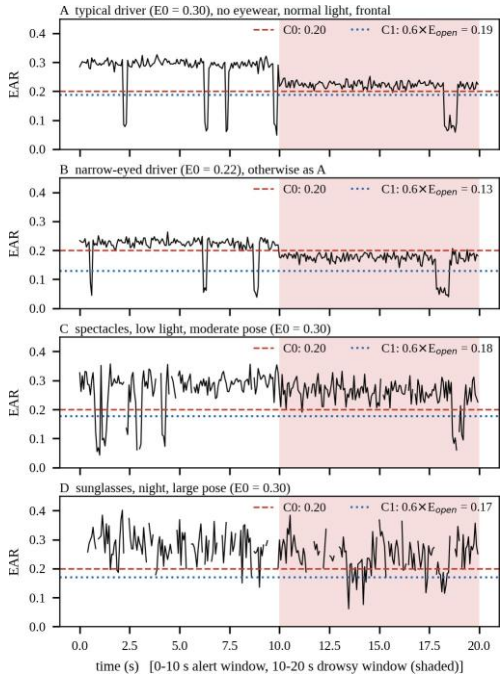



**Figure 1.** Simulated EAR traces (one alert and one drowsy 10 s window) for four driver conditions. Red dashed: C0 threshold; blue dotted: C1 threshold. Synthetic data.

### *4.2 Overall and per-subgroup performance*

Over 30 replicates, C0 reached mean accuracy 0.81 (range 0.79–0.83), recall 0.75 and FPR 0.13; C1 reached accuracy 0.78 (0.77–0.79), recall 0.68 and FPR 0.12. C1 was lower than C0 in accuracy (Wilcoxon $p < 0.05$ in 97% of replicates), and excluding sunglasses drivers the gap largely disappears (0.81 vs 0.80). Calibration is therefore not uniformly beneficial in this model; its effect depends on the subgroup (Table 3, Figure 3).

**Table 3.** Per-subgroup results (mean over 30 simulated replicates).

| Subgroup | n | Acc C0 | Acc C1 | Rec C0 | Rec C1 | FPR C0 | FPR C1 |
|---|---|---|---|---|---|---|---|
| **Eye geometry** | | | | | | | |
| narrow | 27 | .79 | .78 | .95 | .69 | .36 | .13 |
| medium | 65 | .84 | .78 | .74 | .67 | .07 | .11 |
| wide | 28 | .77 | .78 | .59 | .67 | .04 | .11 |
| **Eyewear** | | | | | | | |
| none | 71 | .81 | .80 | .75 | .73 | .13 | .13 |
| spectacles | 37 | .81 | .80 | .76 | .72 | .14 | .12 |
| sunglasses | 12 | .80 | .62 | .73 | .24 | .14 | .00 |
| **Head pose** | | | | | | | |
| frontal | 59 | .81 | .77 | .74 | .66 | .12 | .11 |
| moderate | 43 | .81 | .78 | .75 | .68 | .13 | .12 |
| large | 18 | .82 | .80 | .79 | .74 | .16 | .13 |
| **Illumination** | | | | | | | |
| normal | 59 | .81 | .76 | .70 | .63 | .08 | .11 |
| low | 37 | .82 | .78 | .76 | .67 | .12 | .11 |
| night | 23 | .80 | .84 | .87 | .81 | .26 | .14 |
| **All drivers** | **120** | **.81** | **.78** | **.75** | **.68** | **.13** | **.12** |

*n = mean number of test drivers per replicate. Acc = accuracy, Rec = recall. C0 fixed threshold; C1 per-driver calibrated. Simulated data.*

**Eye geometry.** Under C0, narrow-eyed drivers have FPR 0.36 against 0.07 and 0.04 for medium and wide eyes, while recall falls from 0.95 to 0.59: a single threshold trades false alarms for narrow eyes against misses for wide eyes. C1 largely removes this imbalance (FPR 0.13/0.11/0.11). **Eyewear.** C0 is almost unaffected in accuracy (none 0.81, spectacles 0.81, sunglasses 0.80). C1 is most sensitive to sunglasses (accuracy 0.62, recall 0.24): compression lifts the closed-eye signal, so the closure line scaled from the compressed open level falls below it and closures are missed. Calibration cannot restore information the landmark detector has lost, which points to the value of a reliability gate. **Illumination.** C0 shows more false

alarms at night (FPR 0.26 against 0.08 in normal light); C1 keeps FPR at 0.14. The higher C1 accuracy at night (0.84 against 0.76) should not be read as an advantage of darkness: a plausible, untested explanation is that extra noise pushes partially drooped lids across the calibrated line more often than fully open eyes. **Head pose.** Accuracy differences are small at the assumed distortion magnitudes; real landmark detectors degrade at extreme poses in ways the simple model does not capture, so this result is specific to the model.

**Table 4.** Kruskal–Wallis tests of per-driver accuracy across subgroups.

| Factor | C0 reject | C0 $\varepsilon^2$ | C1 reject | C1 $\varepsilon^2$ | Null: C0 / C1 |
|---|---|---|---|---|---|
| Eye geometry | 73% | 0.078 | 7% | 0.015 | 97% / 13% |
| Eyewear | 7% | 0.016 | 97% | 0.183 | 3% / 3% |
| Head pose | 7% | 0.010 | 17% | 0.019 | 0% / 3% |
| Illumination | 10% | 0.018 | 80% | 0.088 | 13% / 13% |

*"reject" = share of 30 replicates with $p < 0.05$; $\varepsilon^2$ = median effect size. Last column: same test under the null model (s = 0). Rejections demonstrate sensitivity to the assumed effects, not real-world effects.*

### *4.3 Null check and sensitivity*

With all factor effects removed (s = 0), the eyewear and head-pose tests rejected in at most 3% of replicates, close to the nominal 5%. The illumination test rejected in 13% (4 of 30), slightly above nominal but not unusual for 30 replicates (probability of at least four rejections at a true 5% rate ≈ 0.06). Eye geometry still showed an effect on C0 (97%) because variation in $E_0$ is intrinsic to drivers and is not scaled by s. Figure 2 shows the accuracy range across subgroups as s grows: the eyewear range for C1 rises from 0.03 at s = 0 to 0.32 at s = 2, C0 is affected as well at s ≥ 1.5 (0.16, 0.37), the illumination range grows more modestly (C1: 0.03 to 0.09), and head pose stays below 0.06. The ranking of factors is thus conditional on effect size and should be re-estimated with real data. Crossing eyewear with illumination, C1 is lowest for sunglasses in normal and low light (0.59 and 0.61) with only 63 pooled test drivers in the sunglasses-at-night cell, which illustrates how fast subgroup samples thin out when factors are combined.

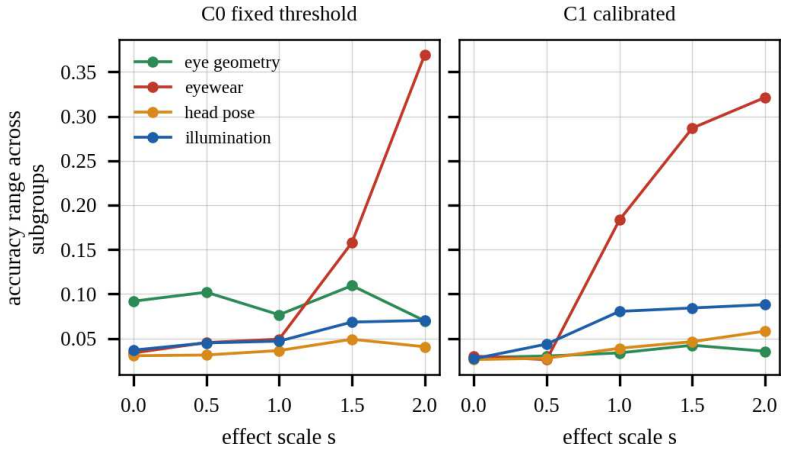



**Figure 2.** Accuracy range across subgroups of each factor versus effect scale s for C0 (left) and C1 (right). Eye geometry remains non-zero at s = 0 because it is intrinsic.

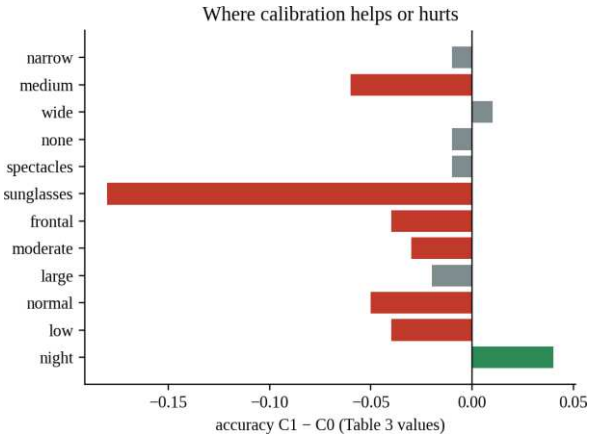



**Figure 3.** Accuracy difference C1 − C0 by subgroup (Table 3). Red: below −0.02; green: above +0.02; grey: within ±0.02. Calibration helps only at night and costs most for sunglasses.

### *4.4 Independent re-implementation and extended analyses*

The analyses in this section come from an independent re-implementation of the model of Section 3.2 (ext/ext_sim.py, seed 7, 30 replicates, 240 drivers each, 50/50 subject-disjoint split) with the pose noise factor set to 1/1.15/1.4 (an added assumption) and Poisson blink onsets. Numbers therefore differ slightly from Section 4.2 and are reported as a robustness check; all results remain simulated. Table 5 shows that the qualitative pattern is reproduced: C1 is lower than C0 overall (0.75 vs 0.80; Wilcoxon $p < 0.05$ in 100% of replicates, C1 below C0 in every replicate), the narrow-eye false-alarm excess of C0 is removed by calibration, sunglasses are the weak point of C1, and night-time false alarms are higher for C0. Quantities agree within about 0.05, which supports the conclusion that the findings follow from the model structure and not from one implementation.

**Table 5.** Reproduction check: main run vs re-implementation (mean over 30 replicates).

| Quantity | Main run | Re-impl. |
|---|---|---|
| All drivers: accuracy C0 / C1 | 0.81 / 0.78 | 0.80 / 0.75 |
| All drivers: FPR C0 / C1 | 0.13 / 0.12 | 0.12 / 0.13 |
| Narrow eyes: FPR C0 / C1 | 0.36 / 0.13 | 0.31 / 0.13 |
| Sunglasses: accuracy C0 / C1 | 0.80 / 0.62 | 0.79 / 0.60 |
| Night: FPR C0 / C1 | 0.26 / 0.14 | 0.21 / 0.13 |
| Large pose: accuracy C0 / C1 | 0.82 / 0.80 | 0.80 / 0.77 |

**Threshold-free view and closure fraction.** Sweeping the PERCLOS cut-off gave an area under the ROC curve of 0.84 for C0 and 0.79 for C1 over all test drivers (0.85 and 0.79 for sunglasses), so the C1 deficit reflects reduced separability of the closure signal at the chosen closure line, not only the operating point of $\tau$. Figure 4 varies the closure fraction $\alpha$ of C1 from 0.5 to 0.8, with $\tau$ re-selected on training drivers at each value. Mean accuracy rises monotonically from 0.73 to 0.92, and for sunglasses from 0.50 to 0.88, because a closure line closer to the open-eye level keeps the compressed closed-eye signal below it. This identifies $\alpha$ as the key parameter of the calibrated rule and shows that the sunglasses weakness is largely a consequence of the a-priori $\alpha$ rather than of calibration itself. Because $\alpha$ was varied on the same simulated test drivers, this is a sensitivity analysis; a real study must select $\alpha$ on training drivers only. A simple blink-depth gate accepted 99.3% of drivers and left accuracy unchanged (0.755 vs 0.754), so an effective gate needs a quality indicator that captures eyewear compression directly, such as landmark confidence or glare features.

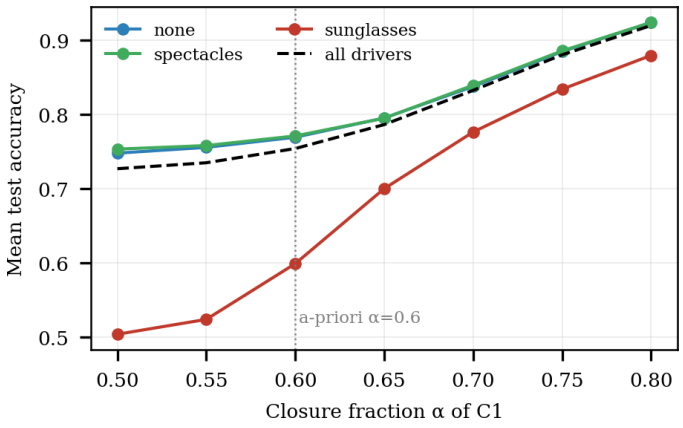



**Figure 4.** Mean test accuracy of C1 versus closure fraction α for eyewear subgroups (re-implementation). Dotted line: the a-priori value used in the main analysis.

**Sample size and subject-level reporting.** Figure 5 translates the simulation into sample-size guidance (Mann–Whitney tests on per-driver accuracy, 1,500 draws per cell): a large effect (eyewear, C1) is detectable with about 8 test drivers per subgroup, a moderate effect (illumination, C1) needs about 16–20, a small effect (narrow vs wide eyes, C0) needs roughly 30–40, and the head-pose contrast never reaches useful power because the underlying effect is near zero. Because the public video datasets with stated sizes contain 14–60 subjects (Table 1), only the larger effects could be tested within a single dataset, which quantifies the value of pooling subjects across datasets with harmonised annotation. Figure 6 shows the recommended reporting format on a single replicate with a realistic sample size: the sunglasses subgroup (n = 16) shows a clearly negative difference (−0.253, 95% subject-level bootstrap CI −0.299 to −0.208), whereas several other subgroups, including night-time driving (n = 19), have intervals that include zero. A pooled accuracy figure would have hidden both facts.

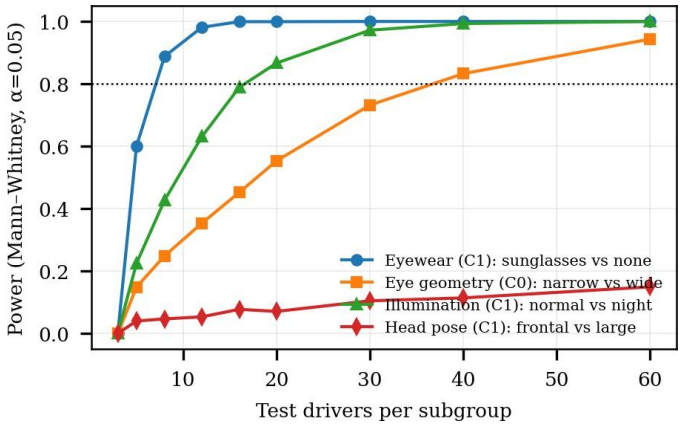



**Figure 5.** Power versus test drivers per subgroup for four factor contrasts; dotted line marks 80% power (re-implementation, simulated data).

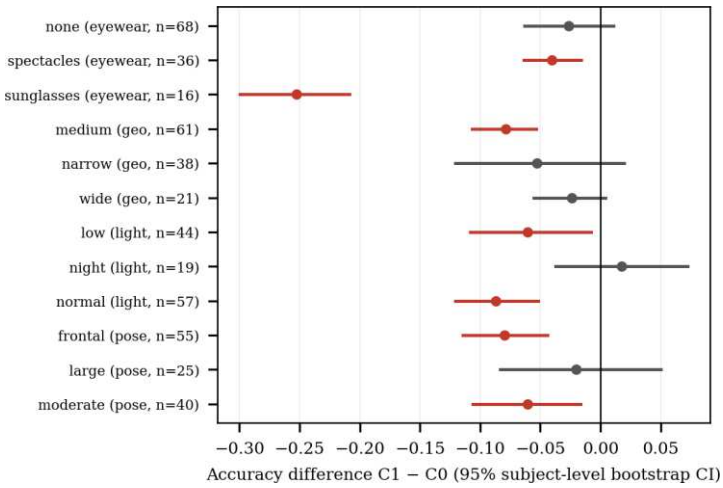



**Figure 6.** Forest plot of the accuracy difference C1 − C0 per subgroup with subject-level bootstrap intervals (single simulated replicate, 3,000 resamples). Red: CI below zero; grey: includes zero.

### *4.5 Dashboard and reporting checklist*

To make the analysis reusable, the study ships with an interactive dashboard (Streamlit and Plotly; dashboard/app.py) with views for overview, per-subgroup performance, effect sizes, the DriverVision rule demonstration, the dataset review and the extended analyses. In its current version it displays the values reported here and demonstrates the reporting design; pointing it at the outputs of a real-data run requires only replacing the data tables. Table 6 lists the subgroup-reporting checklist that the design implements.

**Table 6.** Subgroup-reporting checklist derived from this study.

| Checklist item | Purpose |
|---|---|
| Subject-disjoint train/test split | Measures generalisation to unseen drivers |
| Calibration data separate from evaluated windows | Prevents optimistic calibrated results |
| Per-subgroup metrics with number of subjects | Reveals disparities hidden by pooled accuracy |
| Subject-level (cluster) bootstrap intervals | Treats the driver as the independent unit |
| Null-model check and effect-size sweep | Shows what the analysis can detect |
| Labels from different sources reported separately | Behaviour, self-report and physiological labels differ |
| Reliability gate for degraded signal (e.g. sunglasses) | Withholds decisions when tracking is poor |

## 5. Discussion

**RQ1:** under subject-independent evaluation, performance differs markedly across subgroups even when pooled accuracy looks acceptable (narrow-eyed FPR 0.36 for C0 against a pooled 0.13). **RQ2:** in this model, eyewear has the largest effect on the calibrated rule and eye geometry on the fixed rule, with illumination intermediate and head pose smallest; the ranking depends on effect size (Figure 2). **RQ3:** public datasets expose eyewear and illumination only partly and head pose and appearance rarely, and their small subject counts limit subgroup inference (Table 1, Figure 5). Calibration moves error rather than removing it: it addressed the eye-geometry and night-time false-alarm problems of the fixed rule, and its benefit is clearest where the signal is intact. A practical system would pair calibration with a reliability gate and report performance per eyewear condition, consistent with reports that personalised thresholds help [6].

For researchers, the checklist and power curves allow a dataset to be assessed for subgroup analysis before experiments are run. For system developers, the results indicate three actions: calibrate thresholds per driver, select and validate the closure fraction on training drivers, and add a signal-quality gate that targets eyewear compression. All of these statements concern the simulated model and are intended to guide the real-data study, not to claim performance on real drivers.

### *5.1 Limitations and planned real-data study*

The study is synthetic by construction: factor effects are assumed, so outcomes largely follow from the assumptions, and no result supports or refutes $H_1$ for real drivers. Landmark noise is Gaussian and independent across frames, whereas real detectors show correlated errors, pose-dependent failures and

eyewear glare; skin tone is not modelled. Only a window-level PERCLOS rule with one a-priori α is studied, labels are noise-free, and some patterns, such as higher C1 accuracy at night, arise from how noise interacts with the assumed lid droop and should not be generalised. The DriverVision check involves one subject, one camera and one lighting condition, Table 1 is a literature-based summary to be verified against official documentation, and the main-run pose noise factor is not listed (Section 3.2). The next stage will obtain access to NTHU-DDD, UTA-RLDD and DROZY (and YawDD and MRL Eye for yawn and eye-state modules), annotate eyewear, illumination and head pose per subject (skin tone only under a documented, ethically reviewed procedure), run C0, C1 and a gated variant on subject-disjoint folds, and test subgroup differences with cluster bootstrap intervals and paired subject-level tests, reporting negative results as such. The number of subjects will be set by a power analysis based on pilot subject-level effect sizes.

## 6. Conclusions

This paper set out a design for studying driver-specific variability in vision-based fatigue monitoring, applied it in a controlled simulation, checked it with an independent re-implementation and packaged it as a reusable dashboard and reporting checklist. In the simulation, eye geometry mainly affected a fixed-threshold rule, sunglasses and illumination mainly affected a per-driver calibrated rule, and head pose had little effect at the assumed magnitudes; calibration reduced some subgroup disparities and its value depended on signal quality and on the closure fraction. These findings are conditional on the model and are intended to shape the real-data study, in which subject-disjoint, per-subgroup evaluation will determine whether the effects seen here exist in practice.

## References


[1] A. Kolus, “A systematic review on driver drowsiness detection using eye activity measures,” *IEEE Access*, vol. 12, pp. 97969–97993, 2024, doi: 10.1109/ACCESS.2024.3424654.

[2] B. Fu, F. Boutros, C.-T. Lin, and N. Damer, “A survey on drowsiness detection: Modern applications and methods,” *IEEE Trans. Intell. Veh.*, vol. 9, no. 11, pp. 7279–7300, 2024, doi: 10.1109/TIV.2024.3395889.

[3] M. Venkateswarlu and V. R. R. Ch, “DrowsyDetectNet: Driver drowsiness detection using lightweight CNN with limited training data,” *IEEE Access*, vol. 12, pp. 110476–110491, 2024, doi: 10.1109/ACCESS.2024.3440585.

[4] H. A. Madni, A. Raza, R. Sehar, N. Thalji, and L. Abualigah, “Novel transfer learning approach for driver drowsiness detection using eye movement behavior,” *IEEE Access*, vol. 12, pp. 64765–64778, 2024, doi: 10.1109/ACCESS.2024.3392640.

[5] S. Vidyanathan Dixith, S. Jadhav, Y. Kim, and N. Jayakumar, “Deep learning-based drowsiness detection system for driver's safety,” *IEEE Access*, pp. 154080–154102, 2025, doi: 10.1109/ACCESS.2025.3604728.

[6] G. Ersoy, M. A. Tatar, E. Tonbul, and S. Kırbız, “Improving driver drowsiness detection via personalized EAR/MAR thresholds and CNN-based classification,” arXiv:2604.22479, 2026.

[7] Y. Albadawi, M. Takruri, and M. Awad, “A review of recent developments in driver drowsiness detection systems,” *Sensors*, vol. 22, no. 5, p. 2069, 2022, doi: 10.3390/s22052069.

[8] D. Herath et al., “Vision-based driver drowsiness monitoring: Comparative analysis of YOLOv5–v11 models,” arXiv:2509.17498, 2025.

[9] A. G. Sawant, S. S. Kamble, R. S. Kanade, R. N. Kanugo, T. A. Kapse, and K. A. Bhapse, “A real-time driver drowsiness detection system using MediaPipe and eye aspect ratio,” arXiv:2511.13618, 2025.

[10] S. Essahraui et al., “Real-time driver drowsiness detection using facial analysis and machine learning techniques,” *Sensors*, vol. 25, no. 3, p. 812, 2025, doi: 10.3390/s25030812.

[11] T. Soukupová and J. Čech, “Real-time eye blink detection using facial landmarks,” in *Proc. 21st Comput. Vis. Winter Workshop (CVWW)*, Rimske Toplice, Slovenia, 2016.

[12] D. F. Dinges and R. Grace, “PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance,” Federal Highway Administration, Rep. FHWA-MCRT-98-006, 1998.

[13] C. Lugaresi et al., “MediaPipe: A framework for building perception pipelines,” arXiv:1906.08172, 2019.

[14] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, “Driver drowsiness detection via a hierarchical temporal deep belief network,” in *Computer Vision – ACCV 2016 Workshops*, Taipei, Taiwan, 2016.

[15] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, “YawDD: A yawning detection dataset,” in *Proc. ACM Multimedia Syst. Conf. (MMSys)*, 2014.

[16] R. Ghoddoosian, M. Galib, and V. Athitsos, “A realistic dataset and baseline temporal model for early drowsiness detection,” in *Proc. IEEE/CVF CVPR Workshops*, 2019.

[17] Q. Massoz, T. Langohr, C. François, and J. G. Verly, “The ULg multimodality drowsiness database (called DROZY) and examples of use,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–7.

[18] Media Research Lab, “MRL Eye Dataset,” Brno University of Technology / VSB. [Online]. Available: http://mrl.cs.vsb.cz/eyedataset