

Personalized Temporal Fatigue Analysis and Adaptive Risk Scoring for Camera-Based Driver Monitoring: Prototype and Evaluation Framework

Anshuman Gaikwad

`anshuman.gaikwad_24uds@sanjivani.edu.in`

other <https://orcid.org/0009-0000-7932-3010>

Research Article

Keywords: driver drowsiness, fatigue risk score, personalised calibration, temporal analysis, PERCLOS, eye aspect ratio, MediaPipe, false-alarm rate.

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11416808/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.

Personalized Temporal Fatigue Analysis and Adaptive Risk Scoring for Camera-Based Driver Monitoring: Prototype and Evaluation Framework

Anshuman Sachin Gaikwad

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

gaikwad.anshuman0728@gmail.com

Abstract—Camera-based driver fatigue monitors often apply one fixed threshold to every driver and judge individual frames, which can raise false alarms from normal blinks, talking, or brief head movements. This paper presents a prototype that combines a per-driver calibration baseline, sliding-window temporal indicators (PERCLOS, blink-rate deviation, yawn frequency, head droop), and a bounded risk score with persistence-based alert logic and a live ALERT/DROWSY dashboard. The scoring method is specified as a mathematical formulation and a reference implementation, which is exercised on nine controlled synthetic scenarios covering normal blinking, talking, brief head movement, repeated yawning, camera occlusion, a narrow-eyed driver, and sustained and gradual eye closure. In these scenarios the scoring module ran at 2.4–10.1 microseconds per frame on a single CPU thread, withheld its score during occlusion, and avoided DROWSY states for non-fatigue behaviours. A four-condition ablation (B0 fixed threshold, B1 personalisation only, B2 temporal only, B3 full method) repeated over 30 random seeds of the same scenarios gave precision 1.00, recall 0.76 and about 0.1 false alarms per hour for the full method, against 3.4 per hour for temporal analysis alone and several hundred per hour for per-frame rules; the scenarios are authored together with the method, so this checks the logic of the design and is not evidence of accuracy on real drivers. A subject-independent evaluation framework is specified for NTHU-DDD, YawDD, DROZY, and UTA-RLDD, comparing a fixed threshold, personalisation only, temporal analysis only, and the full method on identical subject-disjoint splits, with false alarms per hour, detection lead time, and subject-level confidence intervals as primary outcomes.

Index Terms—driver drowsiness, fatigue risk score, personalised calibration, temporal analysis, PERCLOS, eye aspect ratio, MediaPipe, false-alarm rate

I. INTRODUCTION

Drowsy driving degrades attention and reaction time, and non-intrusive camera monitoring is an attractive way to detect its visible signs. Recent surveys describe both physiological and vision-based approaches and note recurring difficulties with lighting, occlusion, hardware limits, and bias [3]. Many deep-learning systems report high accuracy on image-level eye or yawn classification [1], [2], [4]–[6], [10], but a single global decision rule can behave differently across drivers whose eye shape, blink habits, and posture differ, and frame-wise decisions do not by themselves separate a blink from sustained eye closure.

This work asks a narrow question: does adding (i) a per-driver calibration baseline and (ii) temporal aggregation with a persistence rule to a transparent, bounded risk score reduce false alarms and/or improve detection compared with fixed-threshold and single-ingredient baselines, under subject-independent evaluation? The contributions are (i) a documented prototype with a live dashboard, (ii) an explicit, bounded and configurable risk-score specification with a reference implementation verified on controlled synthetic scenarios, and (iii) a subject-independent evaluation framework for public datasets with a pre-defined ablation.

Revised research questions and hypotheses. RQ1: Does personalisation improve subject-independent detection over a fixed threshold? RQ2: Does temporal aggregation reduce false alarms per hour at matched sensitivity? RQ3: Does the full method (personalised + temporal) outperform each single-ingredient baseline? RQ4: Which indicators contribute most (ablation and permutation analysis)? Hypotheses H1–H3 correspond to RQ1–RQ3 and are tested as paired, subject-level

comparisons. Earlier drafts also posed “early prediction”; this is retained only as a secondary outcome measured as detection lead time against an annotated reference onset.

II. RELATED WORK AND RESEARCH GAP

Table I summarises the ten studies discussed in the project documents. Their characterisations are taken from the author’s review notes; metadata for [2] was checked against its arXiv record, and the remaining entries are summarised from the author’s review of the original papers.

TABLE I. Studies reviewed (as summarised by the author)

Ref.	Approach	Relevance to this work
[1]	CNN/SVM, transfer learning, EAR/MAR, webcam	Feature-level detection; no per-driver baseline described
[2]	Deep CNN + OpenCV, eye/mouth cues; NTHU-DDD	Real-time warning; image-level
[3]	Survey: EEG/ECG and vision methods	Challenges: light, occlusion, bias
[4]	Lightweight CNN, limited data	Eye-state focus; embedded
[5]	CNN + Vision Transformer	Higher compute; image-level
[6]	GA-optimised CNN, transfer learning	Accuracy focus
[7]	IoT, edge/cloud, mobile app	Remote monitoring
[8]	Personalised eye open/close threshold (FaceMesh)	Closest prior art for personalisation
[9]	Federated transfer learning	Privacy; not behaviour baseline
[10]	HOG/PCA, ensemble, custom CNN	Hybrid classifier

Personalised eye thresholds [8] and temporal drowsiness models on realistic data [18] already exist, so neither personalisation nor temporal analysis is claimed as new here. The remaining, **unproven** candidate contribution is the combination of (a) a multi-indicator per-driver z-score baseline, (b) a reliability gate that withholds the score when tracking is poor, (c) a persistence

rule against short non-fatigue movements, and (d) a controlled four-way ablation with subject-level statistics. If the ablation shows no improvement over the best single-ingredient baseline, the work should be reported as a negative or null result.

III. SYSTEM ARCHITECTURE

Fig. 1 shows the pipeline. Blue blocks correspond to behaviour visible in the prototype (landmark overlay, four indicator bars, risk gauge, trend plot, a Recalibrate control, and Calibrating/No Face states in the interface code); dashed amber blocks are specified but not yet validated. The prototype is written in Python with OpenCV and MediaPipe [13]; the interface code shows separate modules for the engine, configuration, alert manager, and a worker thread that reads the camera and queues features and risk values. The scoring method specified here uses landmark geometry and per-driver statistics rather than a trained fatigue classifier.

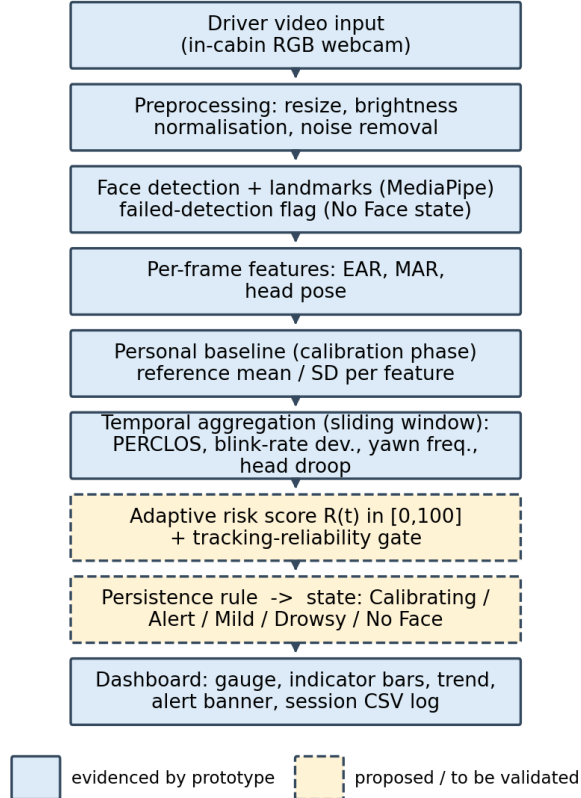


Fig. 1. System architecture. Solid blue: evidenced by prototype; dashed amber: proposed or to be validated.

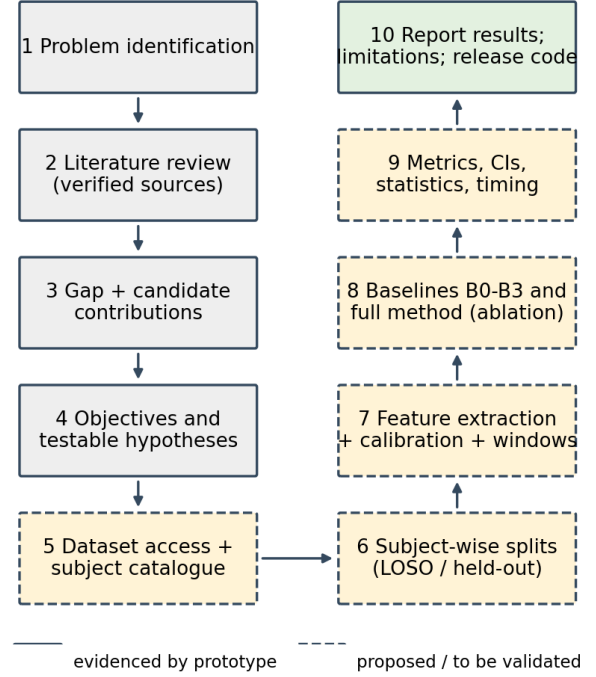


Fig. 2. Research methodology and experimental workflow (redrawn from the project flowchart, with dataset-access and subject-split steps made explicit).



Fig. 3. Prototype project structure (left) and excerpt of the interface/worker code (right), showing the engine, configuration, alert-manager modules and the Alert / Mild Fatigue / Drowsy / Calibrating / No Face states.

IV. METHODS

A. Input, preprocessing, and failed detections

Input is RGB video from an in-cabin webcam (live) or recorded clips (datasets). Frames are resized, brightness-normalised, and denoised with OpenCV, and face landmarks are obtained with MediaPipe FaceMesh [13]. Frames without a valid face are flagged and excluded from feature statistics; the interface exposes a No Face state. Sampling rate, resolution, and normalisation parameters are listed in the released configuration; the reference implementation assumes 15 frames per second.

B. Frame-level features

With six eye landmarks $p_1 \dots p_6$ (corner, upper lids, corner, lower lids) the eye aspect ratio is [11]

$$\text{EAR} = (\|p_2 - p_6\| + \|p_3 - p_5\|) / (2\|p_1 - p_4\|) \quad (1)$$

and the mouth aspect ratio uses analogous vertical/horizontal mouth distances:

$$\text{MAR} = (\|m_2 - m_8\| + \|m_3 - m_7\| + \|m_4 - m_6\|) / (2\|m_1 - m_5\|) \quad (2)$$

Head pitch θ_i is estimated from landmark geometry. The landmark index sets actually used must be listed from the code. A yawn is a sustained high-MAR event; it is evidence for, not proof of, fatigue, and eye closure alone likewise does not define a fatigued state.

C. Personal baseline (calibration)

During a calibration phase the system records valid frames and computes, for each indicator k , a baseline mean μ_k and standard deviation σ_k . The prototype provides a Recalibrate control and a Calibrating state. In the reference implementation calibration lasts $T_{cal}=60$ s, uses valid frames only, and expects the driver to be alert and facing forward; the baseline is fixed after calibration and is refreshed only through Recalibrate. Proposed handling: if the valid-frame count is below a minimum, remain in Calibrating and output no score; do not update the baseline while the score is elevated, to avoid absorbing fatigue into the “normal” reference.

D. Temporal indicators

Indicators are computed over a sliding window of W seconds (30 s in the reference implementation, Table II; to be tuned on validation subjects only). PERCLOS is the fraction of valid frames in the window with the eyes counted as closed [12]:

$$\text{PERCLOS}_w(t) = (1/N_w) \sum_{i \in w} 1[\text{EAR}_i < \alpha \cdot E_{\text{open}}] \quad (3)$$

where E_{open} is the driver’s calibrated open-eye EAR and α is the closure fraction (the 80%-closure convention corresponds to $\alpha \approx 0.2$; the value used in the prototype is unspecified). Blink-rate deviation compares the windowed blink rate with the baseline, $\Delta b = (b_w - \mu_b)/\sigma_b$; yawn frequency counts MAR events above a duration-qualified threshold; head droop is the fraction of the window with pitch deviation beyond a threshold. Isolated blinks are short and barely move the windowed PERCLOS, whereas sustained closure raises it (Fig. 4).

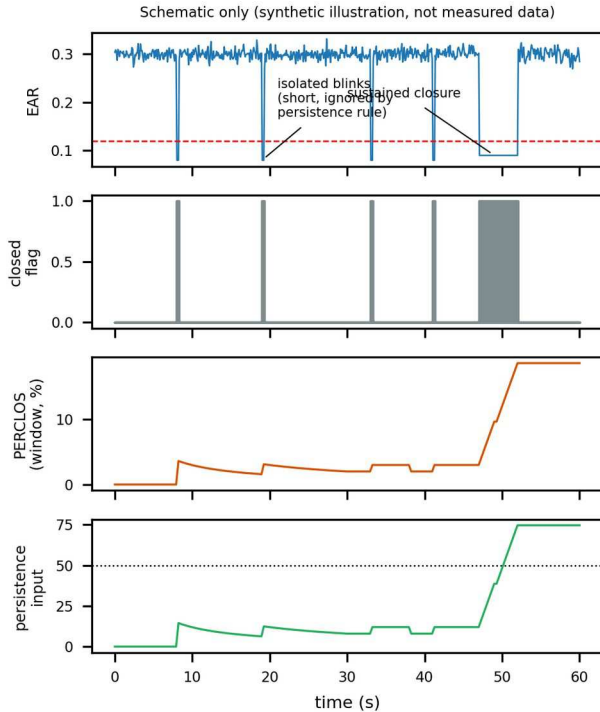


Fig. 4. Schematic of windowed temporal analysis. The traces are synthetic illustrations of the concept, not measured data.

E. Adaptive risk score (candidate specification)

The following specification is the reference implementation of the scoring module and the version to be evaluated; it is a heuristic score, not a statistically validated fatigue predictor. For indicators $k \in \{\text{PERCLOS}, \text{blink}, \text{yawn}, \text{droop}\}$:

$$d_k(t) = \max(0, (x_k(t) - \mu_k) / (\sigma_k + \epsilon)) \quad (4)$$

$$s_k(t) = \min(1, d_k(t) / d_{\max}) \quad (5)$$

$$R(t) = 100 \cdot \sum_k w_k v_k s_k / \sum_k w_k v_k \quad (6)$$

Here $v_k \in \{0,1\}$ marks a valid indicator and $w_k \geq 0$ are weights, set by the author for the reference implementation (Table II) and, if tuned, fitted on validation subjects only. If the valid-frame ratio $q(t)$ falls below $q_{\min}$, the system reports UNRELIABLE rather than a score, so occlusion cannot masquerade as high risk. An exponential moving average smooths $R(t)$, and an alert requires

$$\tilde{R}(t) \geq \theta \text{ for at least } T_p \text{ seconds.} \quad (7)$$

The remaining parameters are listed in Table II; for evaluation they are selected on validation subjects and then frozen. Fig. 5 summarises the computation and Algorithm 1 gives the procedure. Because only the first three indicators are available when the face is tracked but the mouth is closed, eye-closure evidence alone can reach $R=70$, so the DROWSY band starts at 55.

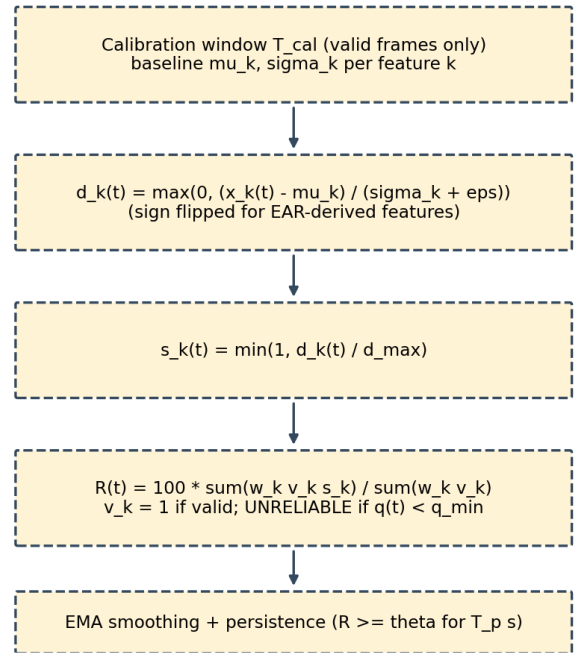


Fig. 5. Personalised baseline and risk-score computation (heuristic specification).

TABLE II. Parameters of the reference implementation

Parameter	Value	Meaning
T_{cal}	60 s	calibration, valid frames only
W	30 s	sliding window for indicators
α	0.2	closure if $\text{EAR} < \alpha \cdot E_{\text{open}}$
blink duration	0.05–0.5 s	closure counted as a blink
yawn	$\text{MAR} > 0.6$ for ≥ 1.5 s	one yawn event
droop	$ \text{pitch dev.} > 15^\circ$	fraction of window
σ floors	0.03; 4/min; 1; 0.05	PERCLOS; blink; yawn; droop
$d_{\max}, w_k$	5; 0.5/0.2/0.15/0.15	PERCLOS/blink/yawn/droop

Parameter	Value	Meaning
q_{min} ; EMA λ	0.7 over 10 s; 0.3	reliability gate; smoothing
bands; T_p	30 / 55; 3 s	MILD / DROWSY, persistence

Algorithm 1: personalised temporal risk scoring (executed once per second after calibration).

```

1 Calibrate: over  $T_{cal}$  s of valid frames set  $E_{open} =$ 
  median(EAR),  $pitch_0 = \text{median}(\text{pitch})$ , blink rate  $\mu_b$ .
2 For each frame: append (EAR, MAR, pitch) to the W-second
  buffer, or NaN if no valid face.
3  $q = \text{valid-frame ratio over the last 10 s}$ . If  $q < q_{min}$ :
  state = UNRELIABLE; reset persistence counters; stop.
4 Compute PERCLOS,  $|\text{blink rate} - \mu_b|$ , yawn count, droop
  fraction on valid frames of the window.
5 For each k:  $d_k = \max(0, x_k)/\sigma_k$ ;  $s_k = \min(1,$ 
   $d_k/d_{max})$ .
6  $R = 100 * \sum(w_k s_k)/\sum(w_k)$ ;  $R_{smooth} = \lambda R +$ 
   $(1-\lambda)R_{smooth}$ .
7 Count consecutive seconds with  $R_{smooth} \geq 30$  and  $\geq 55$ .
8 state = DROWSY if the 55-counter  $\geq T_p$ ; MILD if the 30-
  counter  $\geq T_p$ ; else ALERT.
9 Log (t, features, R, state) to the session CSV; update
  the dashboard.
```

F. Alerts and dashboard

The prototype displays a risk gauge, per-indicator bars, a rolling trend plot, a state banner (e.g., “Alert — no signs of fatigue”, “Drowsiness detected — pull over safely”), and writes a session CSV log. Occlusion is handled by the reliability gate, and repeated yawns contribute through the windowed yawn count and are never sufficient for DROWSY on their own (Section VII-B).

V. DATASETS AND LABEL HANDLING

Datasets differ in labels, sensors, and purpose and **are not merged into one classification task**. Table III lists each source and its intended role. Items marked “confirm” come from the project notes or the original papers and must be checked against official documentation and licence terms before use.

TABLE III. Candidate datasets and intended roles

Dataset	Content / labels	Limits	Role
NTHU-DDD [15]	36 subjects; simulated driving, day/night, glasses; frame-level behaviours (normal, yawn, slow blink, nodding, looking aside)	Behaviour labels, not physiological sleepiness; simulated	Eye, head, blink features; baseline comparison
YawDD [16]	In-vehicle yawning videos, dashboard/mirror views; yawn vs talking vs silent (participant count: confirm)	Yawn label $\neq$ fatigue	Validate yawn/MAR module only
DROZY [17]	14 subjects; sleep-deprivation sessions; KSS [14] and EEG/EOG/ECG; Kinect-based video	Small, lab; camera differs; KSS per session	Construct validity: indicators vs KSS
UTA-RLDD [18]	60 participants; alert / low-vigilance / drowsy videos, self-reported (confirm)	Self-report labels; consumer cameras	Per-subject baseline + temporal test
SUST-DDD [19], UL-DD [20]	Real-drive / multimodal datasets (confirm access, labels)	Not yet reviewed	Optional external check

Label reconciliation. The primary binary task (alert vs. drowsy) is defined separately per dataset using only labels that exist there: clip-level drowsy/non-drowsy annotations where provided (NTHU-DDD), self-reported alert/drowsy classes (UTA-RLDD), and KSS thresholds for DROZY (a pre-declared cut-off such as $KSS \geq 7$ with a sensitivity analysis over thresholds). YawDD supports only yawn detection. Results are

reported per dataset; pooled numbers, if any, are secondary. Annotation sources (annotator, self-report, physiology) are reported as such and not mapped onto each other.

Sampling. The earlier survey-style formula ($n \approx 384$, finite-population corrected to ≈ 350 clips) is not a valid power argument for machine-learning evaluation, because the independent unit is the **subject**, not the clip or frame. All usable clips are used where licences allow, after exclusions are counted; precision is reported through subject-level confidence intervals. Actual per-dataset subject, clip, and class counts will be reported after cataloguing, with exclusions, missing values, and class imbalance documented.

VI. EXPERIMENTAL PROTOCOL

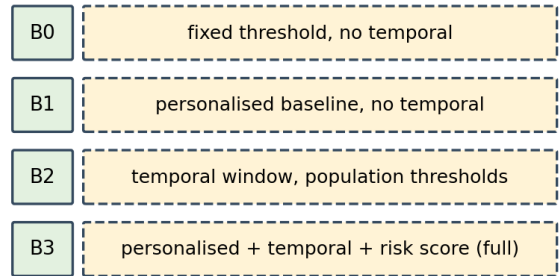
Subjects are partitioned into disjoint training, validation, and test sets, or evaluated by leave-one-subject-out or grouped K-fold (Fig. 6); no subject’s clips appear in more than one role in any fold. Calibration uses an initial segment of each test subject’s own recording judged alert by annotation, so that calibration data and evaluated windows do not overlap. Four conditions are run on identical splits: B0 fixed threshold; B1 personalised, no temporal window; B2 temporal window with population thresholds; B3 full method.

Subject-wise partitioning (no subject in two sets)



Repeat as leave-one-subject-out (LOSO) or grouped K-fold; metrics aggregated per subject, then cluster bootstrap CI

Compared methods (identical splits)



Metrics: acc., precision, recall, specificity, F1, FPR, false alarms/hour, lead time, FPS/latency

Fig. 6. Evaluation protocol: subject-wise partitioning and compared conditions.

Metrics. The positive class is “drowsy/sleepy”. For each method, report accuracy, precision, recall, specificity, F1, and the confusion matrix with explicit denominators, plus false alarms per hour of alert driving and detection lead time relative to annotated onset. Metrics are computed at the window/event level and aggregated per subject; confidence intervals use a cluster (subject-level) bootstrap. Paired comparisons use subject-level paired tests (Wilcoxon signed-rank unless normality holds) with effect sizes; frames are never treated as independent samples. A PERCLOS–risk-score correlation is not evidence of validity because the score includes PERCLOS; the

score is instead compared with independent labels (annotations, KSS). Real-time claims require reporting hardware, resolution, pipeline stages, and repeated timing runs (mean, SD, percentile latency, FPS).

VII. RESULTS

A. Prototype dashboard outputs

Fig. 7 shows the dashboard in two states (webcam preview omitted to protect privacy). In the ALERT example the indicators read PERCLOS 38%, yawn frequency 0%, blink-rate deviation 12%, head droop 3% and a risk score of 35; the DROWSY example reads PERCLOS 74%, yawn 0%, blink deviation 71%, droop 0% and a score of 100. These are single live demonstrations of the interface by the author; they show that the display responds to eye-closure behaviour and are not used as performance evidence.



Fig. 7. Prototype dashboard output (cropped; webcam view omitted). (a) ALERT state, risk 35. (b) DROWSY state, risk 100. Qualitative demonstration.

B. Scoring-logic verification on synthetic scenarios

The reference implementation (supplementary file `ref_impl.py`) was run on nine synthetic 210 s scenarios at 15 frames per second with a fixed random seed. Each has 60 s of calibration followed by 150 s of evaluation. The signals are generated by simple models (Gaussian noise on EAR, MAR and pitch; random blinks at 15 per minute; scripted closure, yawn, head-movement and occlusion episodes); they test that the logic behaves as specified and are **not** real driver data and not a measure of detection accuracy. Table IV and Fig. 8 summarise the outputs.

TABLE IV. Synthetic-scenario outputs (150 s evaluation after calibration)

Scenario	Max R	DROWSY (%)	MILD (%)	Fixed per-frame EAR<0.2 alarm frames (%)
Normal blinking	38.2	0	16.0	4.2
Talking (MAR varies)	38.2	0	16.0	4.2
Brief head movements	41.7	0	18.0	4.2
Single yawn	38.2	0	16.0	4.2
Repeated yawns	40.2	0	17.3	4.2
Occlusion from 90 s	10.6	0	0 (63.4% UNRELIABLE)	n/a

Scenario	Max R	DROWSY (%)	MILD (%)	Fixed per-frame EAR<0.2 alarm frames (%)
Narrow-eyed driver	0.0	0	0	91.1
Sustained closure	60.0	71.7	4.0	53.5
Gradual fatigue	59.7	47.7	16.0	35.4

Four observations follow. (1) No non-fatigue scenario produced a DROWSY state. (2) The narrow-eyed driver, whose normal open-eye EAR of 0.19 is below the global value of 0.2, is flagged on 91.1% of frames by the fixed per-frame rule but never by the calibrated score, which illustrates the mechanism of personalisation. (3) In the occlusion scenario the score was withheld for 63.4% of the evaluation period (the remainder is the 10 s gate delay before invalid frames dominate and the interval before the face was lost), so lost tracking did not produce a high score. (4) In the sustained-closure scenario the DROWSY state appeared 12.9 s after onset, and 48.9 s after onset for gradual fatigue; these are synthetic delays, not detection lead times. Blink-rate deviation from a noisy 60 s baseline places the score in MILD for 16% of the normal-blinking scenario, which motivates longer calibration or larger σ floors and is evaluated in the ablation. The scoring module cost 2.4–10.1 μ s per frame (mean over the 210 s run, one CPU thread, Python, sandbox hardware); this excludes camera capture and landmark detection, whose latency must be measured on the target device.

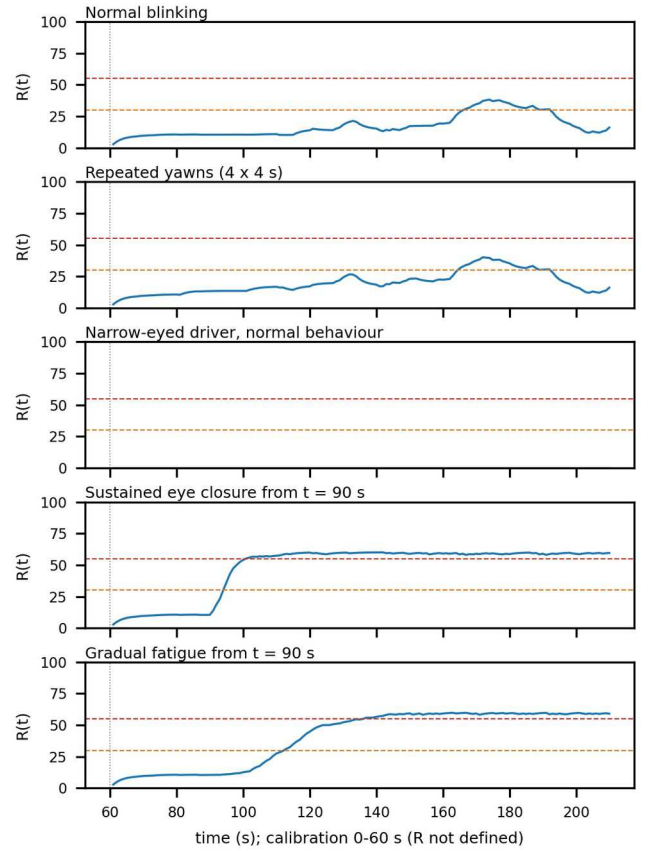


Fig. 8. Smoothed risk score $R(t)$ for five synthetic scenarios. Dashed lines: MILD (30) and DROWSY (55) bands. Synthetic data, for logic verification only.

C. Provenance of numerical values

TABLE V. Provenance of numerical statements in the project documents

Statement	Status	Evidence to be added
≈350 clips; class table	Planned allocation	Real catalogue and counts
PERCLOS means 14.2% vs 61.8%	Illustrative	Per-subject measured PERCLOS
$r \approx 0.85, 0.6, 0.3; p < 0.001$	Expected	Independent-label analysis
Confusion matrix and derived metrics	Illustrative	Subject-wise test results
15–20 FPS	Author-reported	Hardware and timing log
Scores 35 and 100	Genuine screenshot	Annotated sessions

D. B0–B3 ablation on synthetic scenarios; dataset status

Results on the public datasets (NTHU-DDD, YawDD, DROZY, UTA-RLDD) are not yet available: those runs have not been performed, and no dataset numbers are reported or implied. Table VI instead reports the same four conditions on the nine synthetic scenarios of Section VII-B, repeated over 30 random seeds (script `ablation.py`, supplementary). B0 flags a frame when $EAR < 0.2$; B1 flags a frame when $EAR < 0.2$ ·Eopen using the driver’s own calibration; B2 applies the windowed scoring and persistence rule with population constants in place of the personal baseline (closure at $EAR < 0.2$, 15 blinks/min, zero pitch offset); B3 is the full method. A second counts as an alarm when a per-frame flag (B0, B1) or the DROWSY state (B2, B3) is active in it. The positive class is the period after scripted fatigue onset (sustained and gradual closure); all other evaluation seconds are negative. Intervals are 2.5–97.5 percentiles across seeds and reflect simulation noise only; they are not subject-level intervals.

TABLE VI. B0–B3 ablation on the synthetic scenarios (30 seeds; mean [2.5–97.5 percentile across seeds])

Condition	Precision / recall	F1	False alarms per hour	Lead time (s)
B0 fixed threshold	0.37 / 0.88	0.52 [0.48, 0.56]	450 [322, 593]	0 (per-frame)
B1 personalised only	0.51 / 0.88	0.65 [0.58, 0.70]	447 [319, 590]	0 (per-frame)
B2 temporal only	0.55 / 0.78	0.65 [0.63, 0.66]	3.4 [3.2, 4.1]	26.8 [24.6, 30.4]
B3 full method	1.00 / 0.76	0.86 [0.82, 0.89]	0.11 [0.0, 0.9]	28.6 [23.0, 35.7]

Within this synthetic setting, personalisation alone (B1 vs. B0) removes the narrow-eyed driver’s false alarms but leaves per-frame rules alarming on every blink, so both stay near 450 false alarms per hour. The temporal window (B2) removes the blink alarms (about 3.4 per hour, all from the narrow-eyed scenario under population thresholds), and adding calibration to the window (B3) removes that remaining case, giving precision 1.00 at recall 0.76 and 0.11 false alarms per hour. The 0 s lead time of B0 and B1 only reflects that a per-frame rule fires on the first closed frame, which in these scenarios is also an ordinary blink, so it is not comparable with B2 and B3. The lower recall of B3 relative to B0 and B1 comes from the window and persistence rule, whose detection lag was about 29 s after scripted onset. Because the scenarios, onsets and parameters were written by the author together with the method, this ordering is expected by construction: it checks the logic of the design and does not

estimate performance on real drivers. Dataset-level results in the same layout will be added when the subject-disjoint runs are completed.

VIII. DISCUSSION

The prototype provides a transparent, calibratable, windowed indicator display. The synthetic scenarios show the intended behaviour of each design element: calibration removes the narrow-eyed false alarms, the persistence rule and windowing prevent blinks and brief head movements from reaching DROWSY, and the reliability gate withholds the score under occlusion. Whether these properties carry over to real drivers is the question the four-way ablation on public datasets addresses, and results will be limited to the datasets used.

IX. LIMITATIONS AND THREATS TO VALIDITY

- Datasets are simulated or laboratory recordings with small, geographically narrow subject pools [15], [17]; on-road validity is untested.
- Labels come from different sources (annotators, self-report, physiology), so “drowsy” is not uniform across datasets.
- IR and RGB cameras alter landmark behaviour and EAR/MAR scales; thresholds must be calibrated per source.
- Class imbalance and subject leakage can inflate metrics; subject-wise splits and per-subject reporting mitigate but do not remove them.
- Several implementation details (calibration length, weights, thresholds, frame rate) are undocumented and may change conclusions.
- This is a research prototype and not a certified safety device.

TABLE VII. Failure cases and mitigation

Case	Effect on score	Mitigation
Occlusion, face lost	Features undefined	Reliability gate; UNRELIABLE state; baseline frozen
Poor or IR lighting	Landmark jitter; EAR shift	Brightness normalisation; per-source calibration
Glasses, sunglasses	Eye landmarks unreliable	Report per-condition performance
Talking or laughing	MAR rises briefly	Yawn needs duration; no DROWSY from MAR alone
Head turn, looking aside	EAR foreshortened	Pose-aware validity check (proposed)
Noisy calibration	Biased baseline	Longer T_{cal} ; σ floors; Recalibrate
Microsleep	Short closures missed	Window plus event-length features (proposed)

X. CONCLUSION AND FUTURE WORK

A personalised, temporal, risk-score driver-monitoring prototype and a subject-independent protocol to evaluate it have been specified. No performance claim is made. Next steps are to document the implemented formula, obtain dataset access, run the B0–B3 ablation on the public datasets, and report results with intervals; later work may add edge deployment, explainability, and physiological or vehicle signals.

DECLARATIONS

Competing interests. The author declares no competing interests. **Funding.** No external funding was received. **Ethics.** The study uses existing public datasets under their terms and one live demonstration by the author; no new participant data were collected, and the webcam image is omitted from the figures.

Author contributions. A.S.G. designed and implemented the prototype and wrote the manuscript. **AI assistance.** Generative AI (Claude, Anthropic) assisted with drafting, figure generation and the synthetic-scenario script; the author reviewed all content and is responsible for it.

DATA AND CODE AVAILABILITY

Datasets are available from their owners under their own terms and are not redistributed. The reference implementation of the scoring module (ref_impl.py), which generates Table IV and Fig. 8, and the script ablation.py, which generates Table VI, accompany this preprint. The full prototype code, configuration and dataset split manifests will be deposited in a public repository with a DOI.

ACKNOWLEDGMENT

The author thanks Sanjivani University, the Department of Artificial Intelligence and Data Science, and the supervisor, Ashish Katyal, for guidance and support.

REFERENCES

- [1] S. Vidyanathan Dixith, S. Jadhav, Y. Kim, and N. Jayakumar, “Deep learning-based drowsiness detection system for driver’s safety,” *IEEE Access*, vol. 13, pp. 154080–154102, 2025.
- [2] A. N. K. Zaman, P. Chatterjee, and R. Sharma, “Real-time drivers’ drowsiness detection and analysis through deep learning,” in *Proc. 2nd Int. Conf. Artif. Intell., Blockchain, Internet of Things (AIBThings)*, 2024, doi: 10.1109/AIBThings63359.2024.10861885; also arXiv:2511.12438.
- [3] 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, Nov. 2024.
- [4] M. Venkateswarlu and V. Rami Reddy Ch, “DrowsyDetectNet: Driver drowsiness detection using lightweight CNN with limited training data,” *IEEE Access*, vol. 12, pp. 110476–110491, 2024.
- [5] M. Venkateswarlu and V. R. R. Chirra, “CNN-ViT: A multi-feature learning based approach for driver drowsiness detection,” *Array*, vol. 27, Art. no. 100425, Sep. 2025.
- [6] Y. Jebrailey, Y. Sharafi, and M. Teshnehlal, “Driver drowsiness detection based on convolutional neural network architecture optimization using genetic algorithm,” *IEEE Access*, vol. 12, pp. 45709–45726, 2024.
- [7] M. A. Khan, T. Nawaz, U. S. Khan, A. Hamza, and N. Rashid, “IoT-based non-intrusive automated driver drowsiness monitoring framework for logistics and public transport applications to enhance road safety,” *IEEE Access*, vol. 11, pp. 14385–14397, 2023.
- [8] C. Ming and Y. Yunbing, “Perception-free calibration of eye opening and closing threshold for driver fatigue monitoring,” *IEEE Access*, vol. 10, pp. 125469–125476, 2022.
- [9] L. Zhang, H. Saito, L. Yang, and J. Wu, “Privacy-preserving federated transfer learning for driver drowsiness detection,” *IEEE Access*, vol. 10, pp. 80565–80574, 2022.
- [10] M. Ramzan, A. Abid, M. Fayyaz, T. J. Alahmadi, H. Nobanee, and A. Rehman, “A novel hybrid approach for driver drowsiness detection using a custom deep learning model,” *IEEE Access*, vol. 12, pp. 126866–126884, 2024.
- [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] T. Åkerstedt and M. Gillberg, “Subjective and objective sleepiness in the active individual,” *Int. J. Neurosci.*, vol. 52, no. 1–2, pp. 29–37, 1990.
- [15] 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.
- [16] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, “YawDD: A yawning detection dataset,” in *Proc. ACM Multimedia Syst. Conf. (MMSys)*, 2014.
- [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, doi: 10.1109/WACV.2016.7477715.
- [18] 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.
- [19] E. K. Yılmaz and M. A. Akcayol, “SUST-DDD: A real-drive dataset for driver drowsiness detection,” [venue and year to be verified].
- [20] “UL-DD: A multimodal drowsiness dataset using video, biometric signals, and behavioral data,” arXiv:2507.13403, 2025 [author list to be completed].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [refimpl.py](#)
- [ablation.py](#)
- [simresults.json](#)
- [ablationresults.json](#)