

Real-Time Driver Drowsiness and Distraction Detection Using MediaPipe Face Mesh and YOLOv8 Classification on Consumer Hardware

Arik Dhaliwal
Western University
adhali69@uwo.ca

Abel Alexander
Western University
aalexa83@uwo.ca

Anthony Liao
Western University
aliao29@uwo.ca

Satvik Garimella
Western University
sgarimel@uwo.ca

Nathan Feng
Western University
nfeng8@uwo.ca

Abstract—Driver fatigue and distraction are two major contributors to road accidents. We present a real-time, open-source driver monitoring system that operates on a standard webcam and consumer laptop. The system combines geometric drowsiness signals derived from MediaPipe Face Mesh (eye aspect ratio, blink detection, and eye-closure duration) with a lightweight YOLOv8ncls classifier for distraction detection. A temporal smoothing module reduces per-frame noise, and a fusion module produces a single risk state (OK, MONITOR, WARNING, CRITICAL). The system runs at an average of 24 FPS using an iPhone X camera feed displayed on macOS and supports a plug-and-play model design for easy replacement of improved weights. Preliminary distraction results from a binary StateFarm-based dataset achieve 97% accuracy, 35% precision, and 30% recall, while drowsiness detection reliably identifies prolonged eye closure using configurable thresholds. These findings highlight both the practicality of our system architecture and the remaining challenges of reliable single-frame distraction classification in unconstrained settings.

I. INTRODUCTION

Drowsy and distracted driving reduce reaction time, impair situational awareness, and are leading causes of preventable collisions. While commercial driver-monitoring systems exist, most are proprietary and depend on specialized sensors or in-vehicle hardware. This project targets an accessible solution that can run using only an RGB webcam and open-source tools, enabling rapid prototyping, academic reproducibility, and low-cost deployment.

Our system addresses two complementary aspects of driver risk:

- Drowsiness detection using geometric facial cues, specifically eye aspect ratio (EAR), blink events, and closure duration, computed from MediaPipe Face Mesh landmarks.
- Distraction detection using a YOLOv8ncls model trained to classify distracted vs. non-distracted behavior.

The outputs are smoothed over time and fused into a single risk state, enabling clear alerts and minimizing flicker.

II. RELATED WORK

Classical drowsiness detection methods rely on facial landmark geometry, including PERCLOS and EAR-based blink detection. These approaches are computationally efficient and interpretable but depend on stable face tracking. For distraction detection, recent work uses deep learning classifiers trained on

datasets such as StateFarm Distracted Driver and NTHU-DDD, covering phone use, off-road gaze, and secondary tasks. Industry systems (e.g., in luxury vehicles) frequently use infrared sensors and proprietary pipelines, limiting reproducibility.

Our system combines these two approaches into a single, open, real-time pipeline that operates on consumer hardware.

III. SYSTEM OVERVIEW

The system processes webcam frames in two parallel branches:

Drowsiness pipeline (geometric):

- MediaPipe Face Mesh detects 468 landmarks.
- EAR is computed from eye landmarks, with blink and closure event detection.
- A rule-based alert logic triggers warnings based on closure duration.

Distraction pipeline (learned):

- YOLOv8ncls model outputs $P(\text{distracted})$ and label.
- Temporal smoothing computes a rolling window ratio and mean probability.
- A trigger is raised when distraction persists over the window.

A fusion module combines these signals into a single risk state with graded severity.

IV. METHODOLOGY

A. Drowsiness Signal Extraction

MediaPipe Face Mesh provides facial landmarks per frame. The system computes EAR for each eye and averages them:

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

where p_i are the eye landmarks. A hysteresis threshold is used:

- Enter closed state: $EAR < enter_th$
- Exit closed state: $EAR > exit_th$

The system tracks closure duration, blink counts, and uses an EMA-smoothed EAR value. A drowsiness alert is triggered when closure duration exceeds a threshold (default 1.5 s in `alerts.py`).

B. Distraction Classification

We use YOLOv8n-cls trained on a binary distracted vs. not distracted dataset derived from the StateFarm Distracted Driver dataset. Training settings in the notebook include:

- Model: yolov8n-cls
- Image size: 224
- Epochs: 10
- Batch size: 32

The classifier returns *prob_distracted* per frame.

C. Temporal Smoothing

Distraction outputs are smoothed using a sliding window (*window_size*=10 by default). The system computes:

- *ratio* = fraction of frames above *prob_threshold*
- *mean_prob* = average probability across window
- *triggered* = *ratio* $\geq$ *trigger_ratio*

This reduces flicker and false positives.

D. Fusion Policy

Fusion integrates both signals into one risk state:

- **CRITICAL**: drowsiness alert active (prolonged closure)
- **WARNING**: distraction window triggered
- **MONITOR**: eyes closed briefly or elevated distraction probability
- **OK**: otherwise

V. IMPLEMENTATION

The system is modular and implemented in Python:

- *camera.py* — webcam capture
- *landmarks.py* — MediaPipe face landmarks
- *signal.py* — EAR/blink/closure logic
- *distraction_model.py* — YOLOv8 inference
- *distraction_state.py* — temporal smoothing
- *fusion.py* — risk state combiner
- *alerts.py* — drowsiness alert logic
- *ui.py* — heads-up display
- *app.py* — system integration

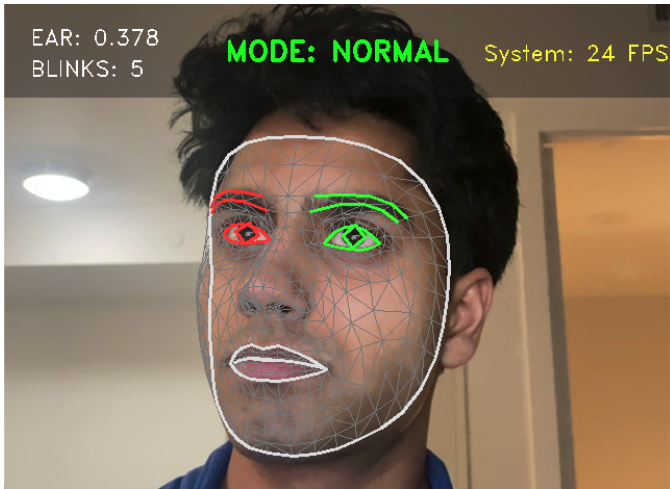


Fig. 1. System interface in a normal driving state. The overlay shows facial mesh landmarks, EAR, blink count, and fused risk mode.

VI. EXPERIMENTS AND RESULTS (DRAFT)

A. Drowsiness Detection

Drowsiness detection is evaluated using prolonged eye-closure scenarios. The system correctly identifies blinks and closure durations in real time using configurable thresholds (*enter_th*=0.23, *exit_th*=0.26 by default). Across repeated live runs, EAR tracking was stable and prolonged eye closure triggered reliable escalation behavior.

B. Distraction Classification

The YOLOv8n-cls model was trained on a binary dataset derived from StateFarm. Current results:

- Accuracy: 97%
- Precision: 35%
- Recall: 30%

Although overall accuracy appears high, precision and recall indicate weak distracted-class capture. This suggests class bias and confirms that frame-wise binary distraction classification is more challenging than expected in realistic conditions.

C. Temporal Aggregation Effects

To improve deployment stability, we aggregate frame outputs using a 10-frame sliding window and fuse this with drowsiness cues. This temporal aggregation reduced prediction flicker and produced clearer, more stable warning behavior during real-time testing.

D. Real-Time Performance

The system achieves real-time performance at an average of 24 FPS using an iPhone X camera feed displayed on macOS. YOLO inference is run every *N* frames to maintain responsiveness while preserving continuous landmark tracking.

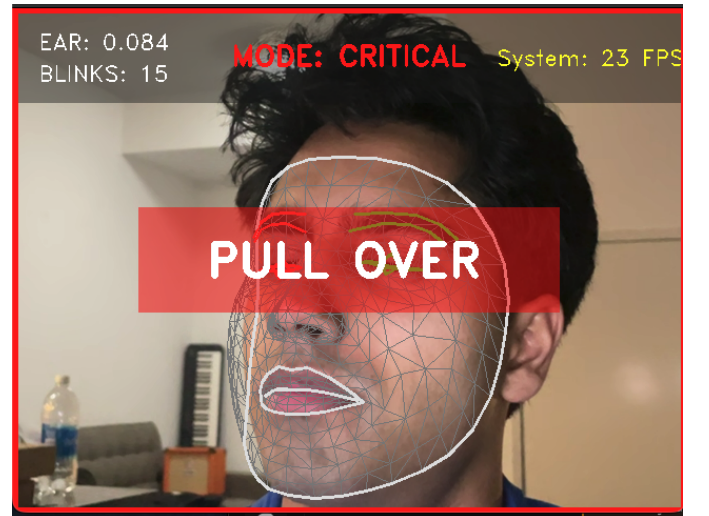


Fig. 2. Critical alert state during prolonged eye closure. The fusion policy escalates to CRITICAL when drowsiness persists.

E. Analysis of Underperformance

Our distraction branch underperformed relative to initial expectations, and we treat this as a key technical finding rather than a negative omission. The results suggest that single-frame distraction classification in unconstrained settings is inherently difficult due to class imbalance, subtle inter-class visual differences, and domain shift between curated datasets and live webcam conditions.

From a system perspective, this validates the need for temporal reasoning and multi-signal fusion. Even with limited distraction precision/recall, integrating geometric drowsiness cues with smoothed distraction estimates yields more useful runtime behavior than either branch alone.

VII. LIMITATIONS

- Distraction classifier performance depends on dataset quality and class balance.
- Face landmarks degrade under extreme lighting or occlusion.
- Binary classification simplifies complex distraction behaviors.
- Current YOLO wrapper assumes `.pt` format; ONNX export is not yet integrated.

VIII. THREATS TO VALIDITY

The reported metrics should be interpreted with care. First, distraction is framed as a binary label, which compresses diverse behaviors and may hide action-specific failure modes. Second, evaluation data and live deployment conditions are not identical, introducing domain shift. Third, camera placement, head pose, and lighting variability can materially affect both face-landmark reliability and classifier confidence.

To address these constraints, we report precision and recall alongside accuracy, include qualitative deployment observations, and avoid claims of production-level readiness.

IX. FUTURE WORK

Future work will prioritize improved distraction robustness through class rebalancing, hard-negative mining, richer labeling, and sequence-aware models that operate on short frame clips rather than independent images. We also plan to investigate confidence-aware fusion and adaptive thresholds to better personalize alert sensitivity across users and environments.

X. CONCLUSION

We present a modular, real-time driver monitoring system that fuses drowsiness signals from MediaPipe with distraction classification using YOLOv8. The system is open-source, runs on consumer hardware, and supports plug-and-play model replacement. Our experiments show that distraction detection in real-world settings is significantly harder than drowsiness detection via physiological cues, but temporal smoothing and multi-signal fusion provide a practical and stable monitoring foundation for continued research and development.