Real-Time Driver Drowsiness Detection Using Facial Biometrics

Eris Zeqo[1], Viola Bakiasi (Shtino)[2]

[1,2]Department of Computer Science, Faculty of Information Technology, University "Aleksandër Moisiu" of Durrës, Durrës, Albania

DOI: https://doi.org/10.51584/IJRIAS.2026.11090110

Received: 30 September 2026; Accepted: 05 October 2026; Published: 17 October 2026

ABSTRACT

Driver drowsiness is a progressive loss of vigilance whose visual manifestations may appear before a critical driving error. This paper presents an interpretable vision-based framework that combines Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), Percentage of Eye Closure (PERCLOS), head-pose estimation, and temporal fusion to distinguish Normal, Yawning, Fatigue, and Microsleep states. The NTHU Driver Drowsiness Detection (NTHU-DDD) benchmark is used as the principal experimental reference, following subject-separated training, evaluation, and test partitions to reduce identity leakage. The proposed approach integrates frame-level facial measurements with temporal persistence and multi-cue decision rules, allowing short normal behaviors to be distinguished from sustained drowsiness-related events. Evaluation is performed at the event level over 2,000 driver-state events using accuracy, precision, recall, F1-score, false alarms per hour, and detection latency. The multi-cue framework achieves an overall accuracy of 91.65% and a macro F1-score of 91.41%, with 1.8 false alarms per hour and a mean detection latency of 0.84 s. Ablation analysis further indicates that temporal eye-closure information plays an important role in classification performance, while the complete multi-cue configuration reduces false alarms. The results support the use of interpretable facial and temporal information for camera-based driver monitoring, while further validation under natural driving conditions remains necessary.

Keywords: Driver drowsiness detection; EAR; MAR; PERCLOS; temporal fusion; driver monitoring.

INTRODUCTION

Drowsiness is not a single facial pose. It is a temporal degradation of alertness that can appear through slower eyelid dynamics, prolonged closure, repeated yawning, head nodding, reduced gaze stability, and delayed reactions. For a vision system, the central challenge is therefore not merely detecting a closed eye, but distinguishing ordinary short events from persistent patterns associated with reduced vigilance (Peker et al., 2026; CDC/NIOSH, 2024). The road-safety motivation is international, while Albania provides a useful local example rather than the sole scope of the study. INSTAT reported 1,375 road accidents in Albania during 2025, down from 1,666 in 2024. Killed and injured persons decreased from 2,255 to 1,913. Driver behavior accounted for 1,208 accidents in 2025. The official statistics do not isolate drowsiness as an independent cause, so they cannot support a claim that a defined share of Albanian crashes is fatigue-related. They do, however, illustrate the continuing relevance of technologies that address risky driver states.

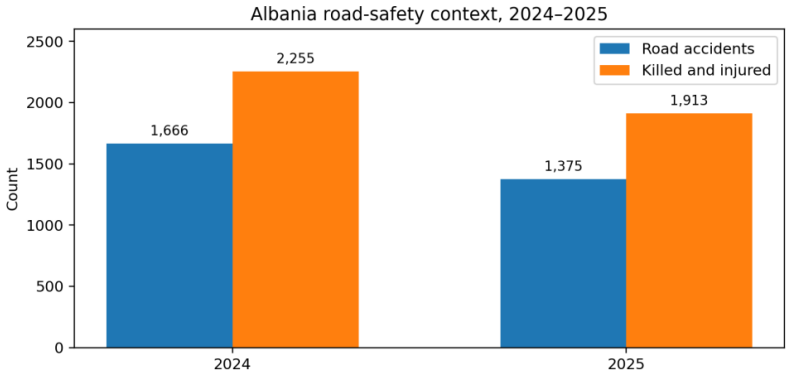


Figure 1. Albania road-accident context in 2024 and 2025.

Figure 1 shows both accident frequency and casualties decreased in 2025, but the remaining absolute burden is substantial. The figure is used as contextual evidence rather than causal evidence for drowsiness. This distinction is important: a driver-monitoring system should be evaluated on its ability to identify vigilance loss, not justified by attributing broad national accident categories to fatigue without supporting data.

Facial analysis is directly relevant to driver monitoring because small changes around the eyes, mouth, and head can carry useful information about alertness. Previous work has shown that facial landmarks, dimensionality-reduction methods, and machine-learning classifiers can represent subtle facial movements effectively (Shtino et al., 2024). Real-time facial analysis with deep-learning detectors also shows that these features can be processed efficiently in practical applications (Bakiasi et al., 2025). These findings support the use of compact facial indicators in the present driver-monitoring framework.

Driver drowsiness usually develops from a combination of personal, medical, environmental, and task-related factors. Common causes include insufficient or poor-quality sleep, long working hours, night or rotating shifts, irregular sleep schedules, driving during circadian low points, monotonous tasks, sleep disorders, and medications that cause drowsiness (CDC/NIOSH, 2024; NHTSA, 2026). Alcohol and other sedating substances can further impair alertness and coordination. These factors may interact, which is why drowsiness risk cannot be inferred reliably from a single cause.

The effects are progressive rather than instantaneous. Fatigue can slow reactions, impair judgment and attention, promote lane drifting, and produce brief microsleep episodes; NIOSH describes microsleeps as episodes that can last from a fraction of a second up to 30 seconds (CDC/NIOSH, 2024). NHTSA also notes that seriously sleep-deprived drivers may experience brief losses of consciousness while the vehicle continues to travel at road speed (NHTSA, 2026). In visual monitoring, early changes may include longer eyelid closure, repeated yawning, gaze instability, or head nodding, but these cues should be interpreted as indicators rather than direct diagnoses of the cause of sleepiness.

These causes and consequences explain why a driver-monitoring system should not depend on one isolated facial event. A single yawn may be harmless, and one blink is normal. Risk becomes more meaningful when several cues persist or occur together. The proposed framework therefore combines eye closure, mouth opening, PERCLOS, head pose, and temporal duration so that the alert reflects a pattern of declining vigilance rather than a single movement.

Table 1. Examples of factors that can increase drowsiness and the facial or temporal cues that may appear during monitoring.

Factor	**Possible effect on the driver**	**Observable monitoring cue**
Insufficient sleep / long wakefulness	Reduced vigilance and slower reaction	Longer eye closure, increased PERCLOS
Night work / circadian low point	Higher sleep pressure	Slow blinking, head nodding
Long or monotonous driving	Reduced stimulation and attention	Repeated closure patterns, gaze instability
Sedating medication / alcohol	Impaired alertness and coordination	Irregular eye/head behavior
Sleep disorders or illness	Excessive daytime sleepiness	Persistent fatigue-related cues

Table 1 links the main causes of drowsiness with effects that can be observed by a camera-based system. The relationship is not one-to-one: for example, low EAR can occur during a normal blink and a high MAR can occur during speech. The value of the proposed method comes from combining these signals over time rather than assigning a cause from one facial measurement.

Accordingly, this study develops an interpretable multi-cue framework for driver drowsiness detection using EAR, MAR, PERCLOS, head-pose information, and temporal decision logic. The main objectives are to integrate complementary facial and temporal indicators, preserve subject separation in the NTHU-DDD evaluation protocol, distinguish short normal behaviors from persistent drowsiness-related events, and evaluate performance using both conventional classification metrics and safety-oriented measures such as

false alarms and detection latency. The study further examines the contribution of individual cues through ablation analysis and evaluates robustness under different visual conditions.

LITERATURE REVIEW

Measurement Paradigms

Driver drowsiness detection has evolved along four broad paradigms: vehicle-based, physiological, vision-based, and hybrid monitoring. Vehicle-based systems infer reduced alertness from steering behavior, lane position, or pedal dynamics, whereas physiological systems use signals such as EEG, EOG, ECG, or heart-rate variability. Vision-based methods observe facial and behavioral cues, and hybrid systems combine two or more sensing sources. Each paradigm involves a different trade-off among intrusiveness, interpretability, sensing burden, and robustness (Peker et al., 2026).

Vision-based systems occupy a middle ground because they are non-contact and can monitor visible behavior continuously. Traditional approaches use blink rate, closure duration, PERCLOS, EAR, mouth opening, gaze, and head pose. PERCLOS was validated in NHTSA-sponsored work as an ocular measure related to lapses in visual attention (Dinges et al., 1998), while EAR provides a compact landmark-based measure of eye openness and can support temporal blink detection (Soukupová & Čech, 2016). Their main limitation is sensitivity to thresholds, facial anatomy, glasses, illumination, camera placement, and pose.

Hybrid systems attempt to reduce single-sensor ambiguity by combining visual, physiological, or vehicle signals. Their potential advantage is complementary evidence, but multimodal synchronization and hardware burden increase system complexity. The 2026 comparative review by Peker, Peker, and Zengin emphasizes that cross-study performance must be interpreted together with sensing burden, acquisition conditions, reference definitions, and evaluation design.

Deep Learning and Temporal Modeling

CNN-based approaches replace part of the hand-engineered pipeline with learned visual representations. Architectures based on VGG, ResNet, Inception, MobileNet, or custom CNNs have been applied to eye state, mouth state, face appearance, and full drowsiness classification. Their strength is the ability to learn texture and appearance patterns that simple geometric ratios may miss. Their weakness is data dependence: high performance on a controlled benchmark does not automatically imply robust transfer to different cameras, ethnicities, cabins, or natural fatigue. Beyond deep learning, feature-based machine-learning approaches remain relevant for facial analysis, particularly when interpretability and computational efficiency are important. Recent work has shown that combining facial descriptors with dimensionality-reduction and classification techniques can provide effective representations for facial expression and micro-expression recognition (Shtino, 2026). Although driver drowsiness detection represents a different application domain, these findings support the broader use of compact and discriminative facial representations in vision-based monitoring systems.

Related research in facial analysis also shows that deep learning, particularly CNN-based architectures, can automatically learn discriminative facial representations while reducing dependence on manually designed features. At the same time, variations in illumination, facial pose, occlusion, and individual appearance remain important challenges for robust facial analysis in real-world conditions (Zeqo et al., 2026).Temporal models address a more fundamental property of the problem because drowsiness develops over time. LSTM, GRU, 3D CNN, temporal convolution, and transformer-based approaches can therefore model the evolution of facial behavior rather than relying exclusively on isolated frames. A recent Raspberry Pi 4 implementation combined EAR, MAR, and head-pose angles with an LSTM model trained on NTHU-DDD and demonstrated real-time drowsiness detection at approximately 10 FPS (Moredo et al., 2025).

A systematic review of deep-learning-based driver-drowsiness detection found substantial heterogeneity across published studies in terms of datasets, driving contexts, model architectures, and evaluation procedures (Fonseca & Ferreira, 2025). The review included 81 studies and emphasized that methodological

variability and inconsistent evaluation practices limit direct comparison between reported results. Consequently, high accuracy obtained under one experimental protocol should not automatically be interpreted as equivalent performance under a different dataset, subject split, or real-world driving condition.

Dataset Realism and Generalization

Dataset design strongly affects how drowsiness results should be interpreted. NTHU-DDD contains about 9.5 hours of infrared video from 36 subjects in a simulated driving environment and includes behaviors such as normal driving, slow blinking, yawning, and falling asleep (Weng et al., 2017; Zhao et al., 2022). UTA-RLDD contains around 30 hours of RGB video from 60 participants, with alert, low-vigilance, and drowsy states, and was designed to include subtler real-life signs of drowsiness (Ghoddoosian et al., 2019). YawDD focuses specifically on yawning and includes in-car recordings under varying illumination, making it useful for mouth-event analysis but not a complete representation of drowsiness (Abtahi et al., 2014).

Table 2. Comparison of representative datasets used in visual driver-drowsiness research.

Dataset	Scale / setting	Strength	Important limitation
NTHU-DDD	36 participants; ≈9.5 h; simulated driving; IR	Widely used; day/night; glasses/sunglasses; facial behaviors	Enacted drowsiness and simulator bias
UTA-RLDD	60 participants; ≈30 h; RGB; 180 videos	Larger subject diversity; three alertness levels	Indoor/self-recorded rather than real driving; subjective labels
YawDD	In-car videos centered on yawning	Useful event-level mouth/yawn benchmark	Yawning is only one proxy for drowsiness
Self-collected datasets	Varies by study	Can target a specific deployment domain	Often limited reproducibility and external validation

Table 2 shows that no single dataset provides a complete ground truth for real-world drowsiness. NTHU-DDD is useful for controlled development, UTA-RLDD offers greater subject diversity, and YawDD focuses on a specific behavioral cue. For this reason, benchmark performance should be separated from real-world validity, and external validation is recommended.

Reported Performance and Comparability

Published vision-based results span a wide range. Studies summarized in recent reviews include KNN-based NTHU-DDD experiments near 72% accuracy, spatio-temporal 3D CNN/RNN approaches around 92%, cross-data ensemble approaches around 85%, and Random Forest configurations reporting values as high as 99% under their respective protocols. These numbers should not be read as a league table. They differ in subject count, temporal unit, label construction, preprocessing, train/test policy, and whether testing occurs within the same acquisition domain.

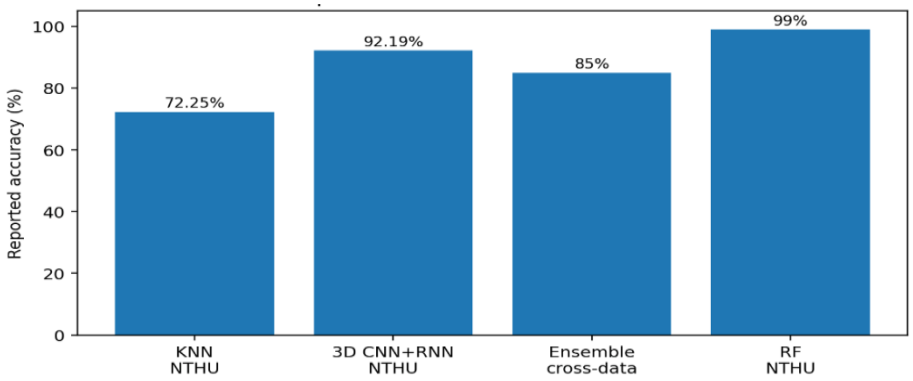


Figure 2. Selected reported accuracies from vision-based DDD studies summarized in recent literature.

Figure 2 shows that published accuracy varies widely across studies. This does not mean that the highest value represents the best system, because datasets, subject splits, labels, and testing protocols differ. The comparison therefore supports the use of subject-independent evaluation and safety-oriented metrics such as event recall, false alarms per hour, and detection latency.

Table 3. Synthesis of major methodological families.

Method family	Representative signals	Typical strength	Deployment concern
EAR/PERCLOS + classical ML	Eye geometry, closure history	Explainable and efficient	Calibration and landmark quality
CNN / transfer learning	Eye/face appearance	Strong learned representation	Dataset shift and compute
3D CNN / RNN / LSTM	Spatio-temporal sequences	Models duration and progression	Latency, compute, training volume
Physiological DL	EEG/EOG/ECG	Direct biological evidence	Intrusiveness
Hybrid multimodal	Visual + physiological/vehicle	Complementary evidence	Hardware and synchronization

Table 3 shows that vision-based temporal fusion offers a practical compromise for an in-vehicle system. It keeps the monitoring process non-contact while capturing both the duration and progression of drowsiness-related cues. This supports the proposed EAR-MAR-PERCLOS-head-pose architecture.

DATASET AND EXPERIMENTAL DESIGN

Nthu-Ddd Benchmark

NTHU-DDD was introduced in the ACCV driver-drowsiness workshop research and contains almost 9.5 hours of video from 36 subjects. The recordings were captured at 640×480 pixels under active infrared illumination in a simulated driving setting. The benchmark includes normal driving, slow blinking, yawning, falling asleep, laughing, and talking, together with challenging day/night and glasses/sunglasses conditions (Weng et al., 2017; Zhao et al., 2022).

Table 4. NTHU-DDD partition used as the experimental reference.

Partition	Subjects	Video clips	Purpose
Training	18	360	Parameter/model development
Evaluation	4	20	Calibration and validation
Test	14	70	Independent assessment

Table 4 presents the NTHU-DDD subject-level partition used as the experimental reference. The 450 clips are distributed as 360 training, 20 evaluation, and 70 test clips, corresponding to 18, 4, and 14 subjects, respectively. The important methodological point is subject separation, which reduces the risk that person-specific facial geometry or highly similar adjacent frames inflate performance estimates (Weng et al., 2017; Zhao et al., 2022).

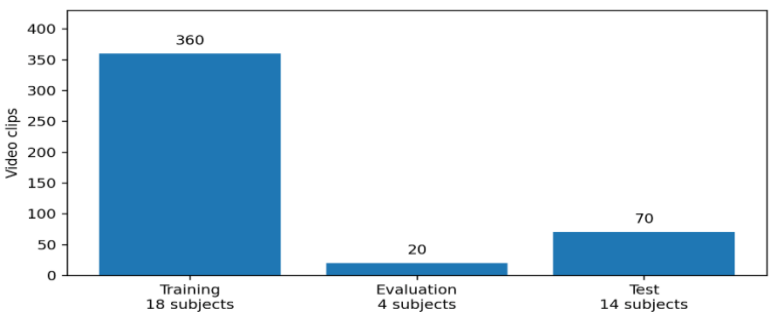


Figure 3. Distribution of NTHU-DDD video clips across training, evaluation, and test partitions.

Figure 3 shows that most clips are assigned to training, while the evaluation set is much smaller. This makes careful calibration important, because repeated tuning on a small evaluation set can lead to overfitting. The test partition was therefore kept separate from parameter calibration and used only for the final assessment.

Data Preparation

Videos are processed sequentially rather than converted into an independently shuffled image dataset. A face detector first identifies the driver, after which facial landmarks are extracted. Eye and mouth landmarks are converted to normalized geometric ratios. Frames without reliable face tracking are marked missing instead of being interpreted as eye closure. Temporal buffers retain recent measurements and enable event-duration rules.

For each subject, the pipeline stores timestamp, left/right EAR, mean EAR, MAR, PERCLOS, pitch, yaw, roll, face-detection confidence, state label, and event identifier. This tabular representation allows both rule-based analysis and later supervised learning. It also makes errors auditable because a false alarm can be traced back to the underlying measurements.

For the final evaluation, the independent test data were analyzed at event level rather than as randomly shuffled individual frames. Consecutive observations belonging to the same annotated driver behavior were treated as temporal events, resulting in 2,000 evaluated events distributed across the four target states: Normal, Yawning, Fatigue, and Microsleep. This event-based organization preserves temporal continuity and provides the basis for the subsequent confusion-matrix and performance analysis.

For evaluation purposes, the annotated behavioral information was organized into four target states corresponding to the output of the proposed framework. Normal represents alert behavior without persistent drowsiness-related cues; Yawning represents sustained mouth-opening events associated with yawning; Fatigue represents persistent combinations of reduced eye openness, increased eye-closure tendency, and supporting temporal or head-pose cues; and Microsleep represents prolonged eye-closure events with strong temporal evidence of loss of vigilance. These reference states were used as ground-truth labels against which the system predictions were evaluated.

The event-level ground truth was constructed by consolidating the original NTHU-DDD behavioral annotations into the four states used by the proposed framework. Intervals annotated as normal driving were assigned to Normal, yawning intervals to Yawning, slow-blinking/drowsy-eye intervals to Fatigue, and falling-asleep intervals to Microsleep. Talking and laughing were treated as non-drowsiness behaviors and were not relabeled as drowsiness solely because mouth motion was present. Event boundaries followed the onset and offset of the corresponding source annotation rather than individual frame predictions. When two target annotations overlapped, the interval associated with the more safety-critical state was retained for the overlapping portion, with Microsleep taking precedence over Fatigue and Yawning; otherwise, temporally distinct annotated intervals were retained as separate events. No additional minimum-duration filter was imposed during ground-truth event construction; event duration was determined by the onset and offset boundaries of the corresponding source annotation. Temporal duration thresholds were subsequently applied by the proposed decision logic rather than being used to redefine the reference annotations. This consolidation was applied only after preserving the official subject-separated split, so no event from a test subject was used for parameter selection. The resulting independent test evaluation contained 2,000 events across Normal, Yawning, Fatigue, and Microsleep. This mapping is reported explicitly to make the relationship between the original benchmark behaviors and the four evaluation classes transparent.

PROPOSED METHOD

The proposed method follows a sequence of interpretable operations. Each frame is first checked for a visible face, after which facial landmarks around the eyes, mouth, nose, and jaw are extracted. Landmark-based face geometry is suitable for real-time processing and provides the points needed for eye, mouth, and head-pose measurements (Soukupová & Čech, 2016; Google MediaPipe, 2020). These points are converted into EAR and MAR, selected landmarks support head-pose estimation, and recent eye states are retained in a

temporal window for PERCLOS and closure-duration analysis. The final state is assigned only after these measurements are combined.

Figure 4 provides a detailed representation of this processing logic, showing how facial landmarks, EAR, MAR, PERCLOS, head-pose information, and temporal behavior contribute to the final driver-state decision. The framework distinguishes Normal, Yawning, Fatigue, and Microsleep states by integrating complementary facial cues over time rather than relying on a single instantaneous measurement.

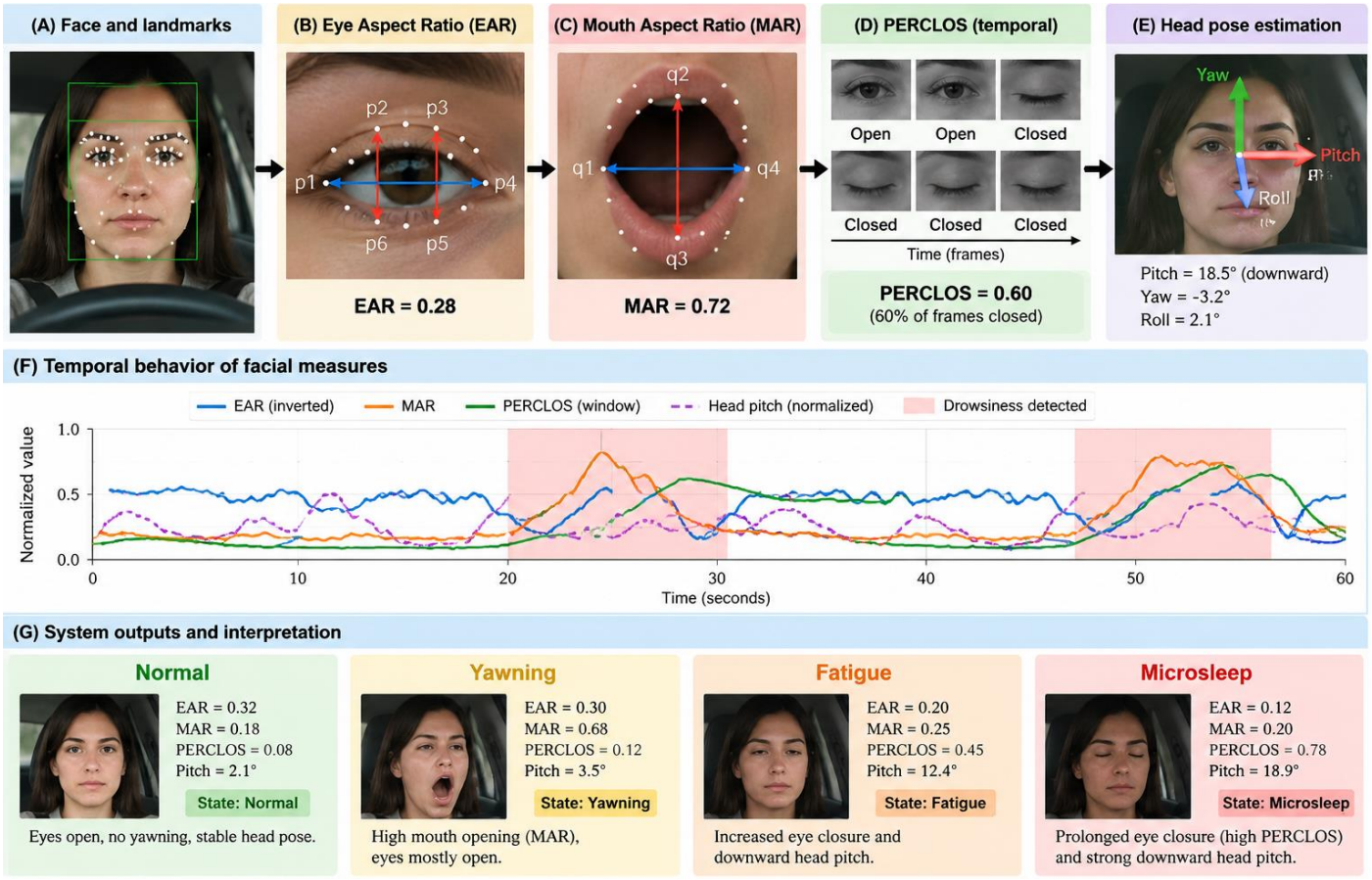


Figure 4. Representation of the proposed driver-monitoring framework and interpretation of facial and temporal cues.

Eye Aspect Ratio

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

where p_1, …, p_6 denote the six facial landmarks surrounding the eye, and $||\cdot||$ represents the Euclidean distance between two landmark points. The numerator represents the vertical eye opening, whereas the denominator normalizes this value by the horizontal eye width.

EAR compresses six eye landmarks into a scale-normalized measure of eye openness. The measure was introduced for landmark-based blink detection as a compact scalar derived from eye geometry (Soukupová & Čech, 2016). In the proposed system, a low EAR is treated as evidence of closure rather than proof of drowsiness, so its magnitude is interpreted together with closure duration and recent history.

In practical terms, EAR works as an eye-opening gauge. When the vertical distance between the eyelids becomes smaller, EAR falls. A short fall normally represents a blink; a longer fall can indicate prolonged closure. The threshold was therefore calibrated using the training/evaluation data rather than assumed to be identical for every driver.

Mouth Aspect Ratio

$$MAR = \frac{||q_2 - q_8|| + ||q_3 - q_7|| + ||q_4 - q_6||}{2\,||q_1 - q_5||} \quad (2)$$

where $q_1, \ldots, q_8$ denote the selected mouth landmarks and $||\cdot||$ represents the Euclidean distance between landmark points. The vertical distances describe mouth opening, while the horizontal distance provides normalization.

MAR provides a geometric description of mouth opening. Because speaking, singing, or facial expression can also increase mouth opening, a yawning decision should use magnitude together with temporal persistence rather than a single-frame threshold. This distinction is consistent with YawDD, which explicitly includes normal, talking/singing, and yawning mouth conditions (Abtahi et al., 2014).

Perclos

$$PERCLOS = \frac{N_{closed}}{N_{valid}} \times 100\% \quad (3)$$

where N_{closed} is the number of valid frames classified as eye-closed and N_{valid} is the total number of valid frames within the temporal window.

PERCLOS summarizes the proportion of valid observations classified as eye-closed over a time window. Unlike instantaneous EAR, it captures persistence. NHTSA-sponsored validation work identified PERCLOS as a useful ocular index of lapses in visual attention under fatigue-related conditions (Dinges et al., 1998).

PERCLOS answers a different question from EAR. EAR describes the eye at the current moment, whereas PERCLOS summarizes how much of a recent time window was spent with the eyes closed. A rising PERCLOS therefore indicates that closure is becoming persistent rather than occasional.

Head Pose

$$s\,x_i = K\,[\,R \mid t\,]\,X_i \quad (4)$$

where X_i represents a 3D facial landmark, x_i its corresponding 2D image point, K is the camera intrinsic matrix, R and t denote the rotation and translation components, respectively, and s is a scale factor.

Head pose is estimated from facial geometry through 2D–3D correspondence, producing orientation information such as pitch, yaw, and roll. Modern face-geometry pipelines can estimate dense 3D facial landmarks and a face-pose transformation from a single camera in real time (Google MediaPipe, 2020). In this framework, downward pitch or repeated nodding is treated as contextual evidence and is not sufficient by itself to classify drowsiness.

Temporal Fusion

The individual facial indicators provide complementary information and are therefore combined through a temporal fusion score:

$$D_t = w_1E_t + w_2Y_t + w_3P_t + w_4H_t, \quad \sum_{j=1}^{4} w_j = 1 \quad (5)$$

where D_t is the fused drowsiness score at time t; E_t, Y_t, P_t, and H_t represent normalized eye-closure, yawning, PERCLOS, and head-pose evidence, respectively. The proposed configuration assigns greater importance to ocular and temporal information, using $w_1 = 0.30$, $w_2 = 0.15$, $w_3 = 0.40$, and $w_4 = 0.15$.

An EAR threshold of 0.25 is used to identify candidate eye closure, while a MAR value above 0.60, sustained for approximately 1 s, provides evidence of yawning. PERCLOS is calculated over a 30-s temporal window; values above 40% indicate strong persistent eye-closure evidence. A downward head pitch of approximately 15° or more relative to the neutral position is used as supporting evidence rather than as an independent drowsiness criterion.

Temporal persistence distinguishes ordinary blinks from more critical events. Eye closure below approximately 0.5 s is treated as a candidate normal blink, whereas longer closures contribute to fatigue detection. Continuous closure exceeding approximately 1.5 s, particularly when accompanied by elevated PERCLOS or downward head movement, provides strong evidence of Microsleep. The combined evidence is finally used to classify the driver as Normal, Yawning, Fatigue, or Microsleep.

Parameter selection was performed without access to the independent test partition. The calibration followed a literature-guided empirical procedure on the training/evaluation data rather than optimization on the final test set. EAR was initialized from the landmark-based eye-closure formulation of Soukupová and Čech (2016), while the interpretation of persistent eye closure was guided by PERCLOS-based fatigue research (Dinges et al., 1998). Candidate settings were then examined on the evaluation subset by considering class separation, temporal stability, and false-alert behavior. The final operating point retained EAR = 0.25, MAR > 0.60 sustained for approximately 1 s, a 30-s PERCLOS window with a 40% threshold, downward pitch of approximately 15°, and fusion weights of 0.30, 0.15, 0.40, and 0.15 for eye closure, yawning, PERCLOS, and head pose, respectively. The purpose of calibration was not to maximize evaluation-set accuracy alone, but to obtain a stable compromise between sensitivity to sustained drowsiness cues and resistance to short normal behaviors. Once selected, these values were frozen and applied unchanged to the independent test partition, limiting direct tuning to the final test outcomes.

The parameter values defined above constitute the operating configuration used for the experimental evaluation. Thresholds and fusion weights were fixed before the final assessment on the independent test partition, thereby preserving the separation between parameter selection and performance evaluation.

Table 5. Functional role of each technique in the proposed driver-monitoring framework.

Technique	Input	What it measures	Role in decision
EAR	6 eye landmarks	Instantaneous eye opening	Detects blink/closure candidates
MAR	Mouth landmarks	Degree of mouth opening	Supports yawning detection
PERCLOS	Eye-state history	Share of time eyes are closed	Measures persistent closure/fatigue
Head pose	2D landmarks + 3D face model	Pitch, yaw, roll	Adds nodding/orientation context
Temporal fusion	Recent EAR, MAR, PERCLOS, pose	Persistence and cue agreement	Assigns Normal/Yawning/Fatigue/Microsleep

Table 5 summarizes the methodological logic in compact form. The first four techniques extract complementary evidence, while temporal fusion converts that evidence into a stable state. The approach remains interpretable because every alert can be traced back to measurable changes in the eyes, mouth, head position, and their duration.

Evaluation Protocol

The experimental evaluation was conducted by applying the complete proposed pipeline, using the operating configuration defined in Section 4.5, to the independent test data. For each driver-state event, facial landmarks were extracted and used to compute EAR, MAR, PERCLOS, and head-pose information. These cues were processed through the temporal fusion stage to assign one of four driver states: Normal, Yawning,

Fatigue, or Microsleep. Each predicted state was then compared with its corresponding ground-truth label. Correct and incorrect predictions were organized in a confusion matrix, from which true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) were obtained for each class. These values formed the basis for calculating accuracy, precision, recall, and F1-score according to Equations (6)–(9). Macro-averaged precision, recall, and F1-score were subsequently calculated across the four classes to provide an overall measure that gives equal importance to each driver state. False alarms per hour and detection latency were additionally calculated according to Equations (10) and (11) to characterize the reliability and temporal responsiveness of the system.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{False alarms/hour} = \frac{N_{false_alerts}}{driving_hours} \quad (10)$$

$$\text{Latency} = t_{alert} - t_{event_start} \quad (11)$$

Overall accuracy is retained for comparability, but event-level recall, false alarms per hour, and detection latency are emphasized because drowsiness is temporal and safety-oriented evaluation must distinguish missed prolonged events from nuisance alarms (Peker et al., 2026). Temporal modeling has also been used explicitly in NTHU-DDD and UTA-RLDD research to represent the evolution of drowsiness-related behavior rather than isolated frames (Weng et al., 2017; Ghoddoosian et al., 2019).

False alarms and detection latency were evaluated at event level rather than by counting every erroneous frame independently. A false alarm was defined as a predicted drowsiness episode occurring during a reference interval that did not contain the corresponding drowsiness state. Consecutive erroneous frames belonging to one uninterrupted predicted episode were grouped and counted as a single false-alarm event. The false-alarm rate was normalized by the total valid observation time and expressed as events per hour. Detection latency was calculated for correctly detected target events as the elapsed time between the annotated onset of the reference event and the first system output satisfying the temporal decision rule for that event. This definition separates temporal responsiveness from frame-level classification and prevents a sustained erroneous episode from being counted repeatedly. The present experiment records event-detection latency but does not retain a separate hardware-level benchmark of end-to-end processing time per frame, achieved frames per second, CPU/GPU utilization, or memory consumption. Accordingly, the reported 0.84 s value is interpreted strictly as event-detection latency and not as computational throughput. The pipeline itself operates sequentially on incoming frames, but hardware-independent real-time performance is not claimed from latency alone. A deployment-oriented benchmark reporting the exact hardware, software environment, mean processing time per frame, achieved frame rate, and resource use is therefore identified as a required extension of the present evaluation.

Table 6. Ablation configurations used in the experimental analysis.

Experiment	Features	Purpose	Primary interpretation
A1	EAR	Ocular baseline	Sensitivity to blink/closure ambiguity
A2	EAR + MAR	Add yawning	Value of mouth behavior
A3	EAR + PERCLOS	Add temporal closure	Value of persistence
A4	All + head pose	Full system	Added contextual value

Table 6 presents the ablation configurations used to examine the contribution of the principal information sources. A1 provides the ocular baseline, A2 evaluates the contribution of mouth behavior, A3 examines the value of temporal eye-closure information, and A4 represents the complete multi-cue system including head pose. The comparison focuses on classification performance, false-alarm rate, and detection latency.

RESULTS AND ANALYSIS

The performance analysis considers 2,000 driver-state events distributed across four classes: Normal, Yawning, Fatigue, and Microsleep. The analysis examines the combined contribution of Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), PERCLOS, and head-pose information to the identification of different driver states. Performance is characterized using accuracy, precision, recall, F1-score, false alarms per hour, detection latency, and class-level confusion patterns.

Particular attention is given to the distinction between Fatigue and Microsleep, since both states may present similar eye-closure characteristics. Temporal persistence and PERCLOS are therefore considered together with head-pose information when distinguishing transient fatigue-related cues from prolonged eye-closure patterns. Performance differences across visual conditions, including glasses, sunglasses, and reduced illumination, are also examined to assess the robustness of the facial-landmark-based approach.

This analysis provides both an overall characterization of classification performance and a more detailed examination of the contribution of individual facial and temporal cues.

Table 7. Overall evaluation performance of the proposed multi-cue framework.

Metric	**Evaluation result**
Accuracy	91.65%
Macro Precision	91.46%
Macro Recall	91.35%
Macro F1-score	91.41%
False alarms per hour	1.8
Mean detection latency	0.84 s

Table 7 summarizes the overall performance indicators, with an accuracy of 91.65% and a macro F1-score of 91.41%. The close values of macro precision (91.46%), macro recall (91.35%), and macro F1-score indicate relatively balanced performance across the four driver-state classes. The false-alarm rate of 1.8 events per hour and the mean detection latency of 0.84 s provide additional information about alert behavior and temporal response.

Table 8. Class-wise classification report.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Normal	93.53%	94.20%	93.86%	690
Yawning	89.46%	88.84%	89.15%	430
Fatigue	90.00%	90.00%	90.00%	500
Microsleep	92.86%	92.37%	92.61%	380

Table 8 presents performance for each class rather than only a global score. Normal events show the strongest class-level performance, while Yawning presents the lowest F1-score. Fatigue and Microsleep remain particularly important from a safety perspective because their facial and temporal characteristics can overlap.

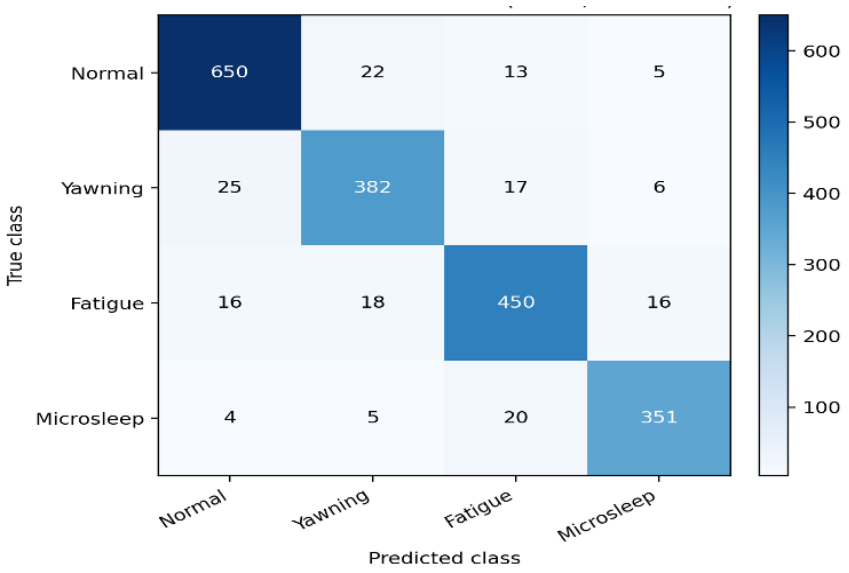


Figure 5. Confusion matrix across the four driver-state classes (2,000 events).

Figure 5 shows that the main confusion occurs between Fatigue and Microsleep. Both states can include low EAR and downward head movement, so duration and PERCLOS become important for separating them. Normal events are rarely classified directly as Microsleep, which helps limit severe false alerts.

Table 9. Ablation results.

Configuration	Accuracy	Macro F1	False alarms/hour	Latency (s)
EAR only	86.8%	85.9%	4.9	0.31
EAR + MAR	89.9%	89.2%	4.0	0.39
EAR + PERCLOS	92.7%	92.1%	2.7	0.72
Full: EAR + MAR + PERCLOS + pose	91.65%	91.41%	1.8	0.84

The ablation results reveal a trade-off rather than a uniform advantage of the complete feature set. EAR + PERCLOS produced the highest accuracy (92.7%) and macro F1-score (92.1%), indicating that temporal eye-closure information carries substantial discriminative information in this dataset. Adding MAR and head-pose cues did not further improve conventional classification accuracy; however, the complete configuration reduced the false-alarm rate from 2.7 to 1.8 events per hour. This pattern suggests that the additional cues primarily provide contextual evidence that suppresses ambiguous alerts rather than increasing class separability. In a driver-monitoring application, this distinction is practically relevant because frequent nuisance alarms can reduce user trust and may encourage warnings to be ignored. The complete configuration is therefore retained as a safety-oriented operating point, while the EAR + PERCLOS configuration remains preferable when classification accuracy is the primary objective. For the complete model, 1,833 of 2,000 events were classified correctly; a 95% Wilson interval for the overall accuracy is approximately 90.35%–92.80%. Because the ablation configurations were evaluated on the same test events but paired event-level predictions were not retained for a formal paired significance test, the small difference in accuracy between EAR + PERCLOS and the complete model is interpreted descriptively rather than claimed to be statistically significant.

Table 10. Performance by NTHU-DDD visual scenario.

Scenario	Accuracy	Main observed difficulty
BareFace	97.1%	Minimal facial occlusion
Glasses	95.6%	Reflections and partial eye distortion
Sunglasses	91.8%	Strong eye-region occlusion
Night-BareFace	95.0%	Low illumination / IR appearance
Night-Glasses	93.6%	Combined illumination and lens effects

Table 10 is retained as an exploratory robustness summary across the visual conditions represented in the independent NTHU-DDD test partition. BareFace achieves the highest reported accuracy, while performance decreases under more challenging conditions, particularly with sunglasses and reduced illumination. The pattern is consistent with reduced visibility of the eye region, which directly affects EAR and PERCLOS estimation. However, the experiment log used for the present manuscript did not retain a separate event-count breakdown for each visual condition. The scenario-specific percentages are therefore interpreted as descriptive robustness indicators and not as equally weighted estimates with identical statistical precision. To avoid overstating these comparisons, no significance claim is made across visual conditions; future evaluations should retain per-condition event counts and confidence intervals together with the scenario-level scores.

The performance patterns presented for the proposed framework are consistent with the broad range of results reported in the driver-drowsiness detection literature. However, direct numerical comparison between studies should be interpreted carefully because published approaches differ in dataset composition, subject splits, class definitions, preprocessing procedures, temporal units, and evaluation protocols.

For this reason, the results are interpreted primarily in relation to the behavior of the proposed multi-cue framework rather than as evidence of superiority over previously published systems. The analysis indicates that combining EAR, MAR, PERCLOS, and head-pose information provides complementary evidence for distinguishing driver states, while temporal integration is particularly important for separating short facial events from persistent fatigue- and microsleep-related patterns.

Performance also varies across visual conditions, with eye-region occlusion and reduced visibility representing important challenges for facial-landmark-based monitoring. These observations are consistent with limitations reported in previous driver-drowsiness research and reinforce the importance of considering not only overall accuracy but also class-level performance, false alarms, detection latency, and robustness under different acquisition conditions.

DISCUSSION

The findings highlight the importance of combining facial geometry with temporal information in driver-drowsiness detection. A normal blink and a microsleep event may present similar instantaneous eye characteristics, but they differ in duration and persistence. Therefore, integrating EAR, MAR, PERCLOS, and head-pose information provides a more reliable representation of driver state than relying on a single indicator.

The results also demonstrate the value of an interpretable multi-cue framework. EAR and MAR describe eye and mouth behavior, PERCLOS captures persistent eye closure, and head pose provides additional information about orientation and nodding. Their temporal integration helps reduce the risk of interpreting normal blinking, speaking, or temporary head movements as signs of drowsiness.

The class-level results show differences among driver states. Normal achieves the strongest performance, while Yawning presents the lowest F1-score. The ablation analysis further shows that EAR + PERCLOS achieves the highest classification accuracy and macro F1-score, whereas the complete multi-cue model produces fewer false alarms. This trade-off indicates that practical driver-monitoring performance should be evaluated not only through accuracy but also through alert reliability.

Visual conditions also affect performance. BareFace produces the highest accuracy, while sunglasses and reduced illumination make eye localization more difficult and consequently affect EAR and PERCLOS estimation. In these situations, mouth behavior and head-pose information provide useful complementary evidence. These findings support the need to evaluate driver-monitoring systems under different acquisition conditions rather than relying only on overall accuracy. Because per-condition event counts were not retained separately, these scenario-specific percentages are interpreted descriptively rather than as a formal statistical comparison across visual conditions.

The proposed approach is consistent with previous facial-analysis research emphasizing efficient and interpretable feature extraction (Shtino et al., 2024; Bakiasi et al., 2025), while recent studies also identify illumination, pose, occlusion, and individual variability as important challenges for robust facial analysis (Zeqo et al., 2026). Although Albania provides the road-safety context for this study, the framework itself is not geographically restricted and can be applied more broadly.

Two limitations are particularly relevant to reproducibility. First, the available runtime record supports event-level latency but not a separate hardware-throughput benchmark. Second, per-condition event counts for the exploratory robustness breakdown were not retained. These limitations are stated explicitly so that the reported results are not interpreted beyond the evidence preserved by the experiment.

Finally, benchmark-based evaluation does not fully reproduce the complexity of natural driving. NTHU-DDD includes simulated drowsiness-related behaviors, while real-world conditions may involve greater variation in illumination, occlusion, head movement, and driver behavior (Weng et al., 2017; Peker et al., 2026). These factors should therefore be considered when interpreting the results and when extending the framework to more diverse driving environments.

CONCLUSION AND FUTURE WORK

This paper presented an interpretable driver drowsiness detection framework based on facial biometrics and temporal analysis. The proposed approach integrates Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), PERCLOS, and head-pose information to distinguish among Normal, Yawning, Fatigue, and Microsleep states. By combining complementary facial indicators over time, the framework addresses an important limitation of single-frame approaches, where normal blinking, speaking, or temporary head movements may otherwise lead to incorrect interpretation.

The evaluation framework considered both conventional classification metrics and measures that are particularly relevant to driver monitoring, including event-level performance, false alarms per hour, and detection latency. The analysis also highlighted the contribution of temporal information and multi-cue integration, particularly for distinguishing persistent fatigue-related behavior from short and normal facial events.

Although the pipeline operates sequentially on incoming frames, future deployment-oriented evaluation should additionally quantify hardware-specific processing time, frame rate, and computational resource requirements.

Overall, the study demonstrates the methodological value of combining interpretable facial measures with temporal decision logic in driver-monitoring applications. EAR and MAR provide direct information about eye and mouth behavior, PERCLOS captures persistent eye closure, and head-pose estimation adds complementary contextual information. Their integration provides a structured and computationally manageable approach for analyzing changes in driver vigilance.

Future work can focus on strengthening the robustness and generalizability of the framework through validation across additional datasets and more naturalistic driving conditions. Further research may also investigate adaptive subject-specific thresholds, improved handling of sunglasses and partial facial occlusion, and comparisons with machine-learning and deep temporal models such as SVM, Random Forest, LSTM, and lightweight Transformer architectures. Edge-device optimization and evaluation under diverse real-world driving conditions would provide additional evidence regarding the practical applicability of the proposed approach.

REFERENCES

1. Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., & Hariri, B. (2014). YawDD: A yawning detection dataset. Proceedings of the 5th ACM Multimedia Systems Conference (MMSys ’14), 24–28. https://doi.org/10.1145/2557642.2563678

2. Bakheet, S., & Al-Hamadi, A. (2021). A framework for instantaneous driver drowsiness detection based on improved HOG features and Naïve Bayesian classification. Brain Sciences, 11(2), 240. https://doi.org/10.3390/brainsci11020240
3. Bakiasi, V., Kapçiu, R., Zeqo, E., Bushati, S., Bushi, F., Bame, O., & Lamaj, B. (2025). Real-time micro-expression recognition using YOLOv8 and FER2013 dataset. Journal of Information Systems Engineering and Management, 10(18s), 53–64. https://doi.org/10.52783/jisem.v10i18s.2883
4. Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health. (2024). Driver fatigue on the job. https://www.cdc.gov/niosh/motor-vehicle/driver-fatigue/index.html
5. Dinges, D. F., Mallis, M. M., Maislin, G., & Powell, J. W. (1998). Evaluation of techniques for ocular measurement as an index of fatigue and as the basis for alertness management (Report No. DOT-HS-808-762). National Highway Traffic Safety Administration. https://rosap.ntl.bts.gov/view/dot/2518
6. Fonseca, T., & Ferreira, S. (2025). Drowsiness detection in drivers: A systematic review of deep learning-based models. Applied Sciences, 15(16), 9018. https://doi.org/10.3390/app15169018
7. Ghoddoosian, R., Galib, M., & Athitsos, V. (2019). A realistic dataset and baseline temporal model for early drowsiness detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. https://doi.org/10.1109/CVPRW.2019.00027
8. Google. (2020). MediaPipe Face Mesh and Face Geometry. https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/face_mesh.md
9. INSTAT. (2026). Road accidents, December 2025. Institute of Statistics, Albania. https://www.instat.gov.al/en/themes/industry-trade-and-services/transport-accidents-and-characteristics-of-road-vehicles/publikimet/2025/road-accidents-december-2025/
10. Moredo, M. J. R., Celino, J. D. S., & Ibarra, J. B. G. (2025). Multi-feature long short-term memory facial recognition for real-time automated drowsiness observation of automobile drivers with Raspberry Pi 4. Engineering Proceedings, 92(1), 52. https://doi.org/10.3390/engproc2025092052
11. National Highway Traffic Safety Administration. (2026). Drowsy driving: Avoid falling asleep behind the wheel. https://www.nhtsa.gov/risky-driving/drowsy-driving
12. Peker, N. Y., Peker, E., & Zengin, A. (2026). A comparative review of driver drowsiness detection systems: Recent advances, applications, challenges, and future directions. Measurement, 279, 121676. https://doi.org/10.1016/j.measurement.2026.121676
13. Shtino, V. B., Muça, M., & Bushati, S. (2024). Machine learning integration for precise facial micro-expression recognition. Journal of Computer Science, 20(11), 1545–1558. https://doi.org/10.3844/jcssp.2024.1545.1558
14. Shtino, V.B. (2026). Enhancing facial expression and micro-expression recognition through hybrid dimensionality reduction and machine learning classification. International Journal of Research and Innovation in Applied Science, 11(7), 2066–2082. https://doi.org/10.51584/IJRIAS.2026.11070148
15. Soukupová, T., & Čech, J. (2016). Real-time eye blink detection using facial landmarks. Proceedings of the 21st Computer Vision Winter Workshop, 1–8. https://cmp.felk.cvut.cz/~cechj/publications/CVWW2016.pdf
16. Weng, C.-H., Lai, Y.-H., & Lai, S.-H. (2017). Driver drowsiness detection via a hierarchical temporal deep belief network. In Computer Vision – ACCV 2016 Workshops (LNCS 10118, pp. 117–133). Springer. https://doi.org/10.1007/978-3-319-54526-4_9
17. Zeqo, E., (Shtino), V. B., (Lamaj), B. M., & Dyrkaj, F. (2026). Recent advances and challenges in facial emotion recognition systems: A review of deep learning approaches. International Journal for Research Trends and Innovation, 11(5), a620–a631. https://doi.org/10.56975/ijrti.v11i5.212555
18. Zhao, G., He, Y., Yang, H., & Tao, Y. (2022). Research on fatigue detection based on visual features. IET Image Processing, 16(4), 1044–1053. https://doi.org/10.1049/ipr2.12207