Journal of Electrical Engineering and Computer (JEECOM)
Vol. 8, No. 2 (2026), DOI: 10.33650/jeecom.v8i2.16982
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Performance Evaluation of YOLOv8-Pose for Vision-Based Fatigue Detection via Multi-Feature Behavioral Analysis

Dinda Ayu Permatasari [1*], Rifki Noviandra Lestari [2], Dimas Rossiawan Hendra Putra [3], Brahma Ratih Rahayu Fakhrunnia [4], Wahyu Tri Wahono [5], Mohammad Muallif [6]
[1-6] Teknik Elektronika, Jurusan Teknik Elektro, Politeknik Negeri Malang, Malang, Indonesia

Article Info

Article history:

Received August 15, 2026
Revised September 01, 2026
Accepted September 03, 2026

Keywords:

YOLOv8-Pose
fatigue detection
keypoint estimation
eye closure
head tilt

ABSTRACT

Abstract—Driver fatigue remains a major, largely preventable contributor to road traffic crashes, yet most existing YOLOv8-based fatigue-detection systems rely on a two-stage detect-then-landmark pipeline, adding complexity and latency. This study evaluates a single-stage alternative: a YOLOv8n-Pose model that jointly extracts eye closure, a hand-at-mouth gesture used as the visual cue for yawning, and head tilt from eight keypoints in a single detection pass, with eye state determined via class prediction rather than a computed Eye Aspect Ratio, since only one keypoint is annotated per eye. The model was trained on a 2,625-image dataset spanning varied lighting, subjects, accessories, and camera distances. The trained model achieved a bounding-box mAP50 of 0.967 and a keypoint mAP50 of 0.817, with per-indicator F1-scores of 0.90 for eyes open, 0.94 for eyes closed, and 0.97 for hand-at-mouth detection, with most misclassifications occurring between the two eye-state classes. Live testing showed a near-linear relationship between ear-to-shoulder keypoint distance and head-tilt angle, an effective operating range of 60–100 cm, and a processing rate of 3.3–3.5 FPS, indicating further optimization is needed for real-time deployment. The fatigue-decision logic evaluates eye closure through a one-minute PERCLOS window while treating a detected hand-at-mouth event as fatigue-relevant from a single instance. Error analysis identified two failure modes: confidently incorrect eye-state predictions under backlighting, and missed hand-at-mouth detections under partial hand occlusion. These findings demonstrate that a single-stage, pose-based pipeline can achieve competitive multi-indicator fatigue-detection performance without a separate landmark-extraction stage.

Corresponding Author:
Dinda Ayu Permatasari,
Politeknik Negeri Malang, Jalan Soekarno Hatta No 9, Malang, 65141, Indonesia
Email: dinda_ayu@polinema.ac.id Orchid ID : https://orcid.org/0000-0002-6695-4769

1. INTRODUCTION

Driver fatigue is a persistent and largely preventable contributor to road traffic crashes. Comprehensive crash investigations indicate that driver fatigue is implicated in approximately 8.8–9.5% of vehicle collisions, with this proportion rising to 10.6–10.8% in crashes involving substantial property damage [1]. In Indonesia, epidemiological reviews of crash causation similarly list driver condition including fatigue, drowsiness, illness, and psychological pressure among the recurring contributing factors, alongside speeding and traffic-rule violations, and note that such incidents tend to be underreported in official statistics [2]. Because fatigue impairs reaction time and judgment well before a driver perceives themselves as unsafe to drive, early and reliable detection is central to crash prevention.

Fatigue-assessment approaches fall into vehicle-based, physiological, and behavior-based categories [3]. Physiological methods (EEG, ECG, EMG) are accurate but require intrusive sensors [4], [1], although hybrid physiological–visual systems have shown improved responsiveness [5]. This has pushed research toward non-invasive, camera-based monitoring of cues such as eye closure, yawning, and head movement [6]. While eyelid-closure percentage (PERCLOS) is a common single predictor [7], it is fragile on its own, motivating multi-feature approaches that combine eye, mouth, and head-pose signals [8], [9].

This behavior-based line of research has been driven largely by computer vision: early EAR/MAR-based systems[10] gave way to CNNs[10], edge-deployed pipelines [11], embedded Indonesian YOLO-based systems [12], and Vision Transformers [13] typically chaining a detection stage to a separate classification or landmark stage. Within this space, YOLO detectors, especially YOLOv8, are favored for their speed–accuracy balance [14] with even lightweight variants performing well [15]. In addition to object detection, YOLOv8 also offers native pose estimation, regressing keypoints within the same forward pass as detection a capability that has been specifically validated for general human-pose tasks within this architecture [16], [17].

Yet most YOLOv8-based fatigue systems still use YOLOv8 only for region detection, relying on a separate landmark model such as PFLD, Attention Mesh, or Dlib to obtain eye, mouth, and head-pose measurements [8], [9], [7], adding latency that works against embedded, real-time deployment [11], [12]. At the same time, YOLOv8's own pose-estimation task, which is capable of regressing keypoints directly in a single detection pass, has been extensively validated for general human pose estimation ever, none of the YOLOv8-based fatigue-detection studies reviewed above adopt this native pose head directly for eye closure, yawning, and head tilt each instead pairs YOLOv8 with a separate landmark-extraction model. This indicates that using YOLOv8's pose-estimation head as a single, unified mechanism for jointly capturing these three fatigue-relevant cues without a second, separate landmark-extraction stage has not yet been demonstrated in the literature surveyed here. Motivated by this gap, this study evaluates YOLOv8 pose estimation as a single-stage, multi-feature behavioral analysis approach to near real-time fatigue detection jointly capturing eye closure, yawning, and head tilt from a single keypoint set assessed using standard object-detection metrics and analyzed for performance and limitations under varied testing conditions, to clarify the effectiveness of this approach for near real-time, vision-based fatigue detection.

2. METHOD

This study applies a single-stage object-detection-and-pose-estimation approach based on YOLOv8 to detect a driver's fatigue level in real time, evaluated to date within a controlled indoor testing environment. The proposed method extracts three primary behavioral indicators eye closure, yawning, and head tilt directly through keypoint regression within a single computational pass, rather than through a separate face-detection stage followed by a second landmark-extraction model as in most prior work [7] [8], [9]. Figure 1 summarizes this overall pipeline, from data collection and annotation through model training to the final fatigue decision output, and the remainder of this section describes each stage in turn.

2.1 Dataset

The dataset used in this study consists of 2,625 images assembled from a combination of sources: primary data collected through personal recording, images drawn from the COCO dataset, and images sourced via Roboflow. The dataset covers a range of indoor conditions, including variation in lighting (bright and low-light conditions), variation in subjects (male, female wearing a hijab, and female not wearing a hijab), variation in accessories (regular eyeglasses and sunglasses), and variation in camera viewpoint (predominantly frontal angles, but with the head positioned at varying distances from the camera).

All images were standardized to a resolution of 1920×1080 pixels. The dataset was then split randomly into three subsets: 80% for the training set, 15% for the validation set, and 5% for the testing set. Model training was carried out using the Ultralytics library, with Google Colaboratory used as the runtime environment for the training process.

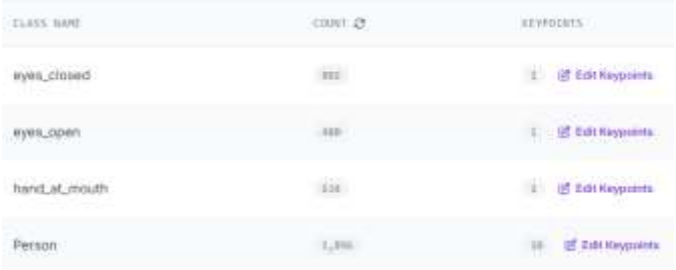

Figure 1. Overview of the detection feature pipeline, from the three data sources through Roboflow annotation and dataset splitting, YOLOv8n-Pose training, class and keypoint prediction, indicator computation, to the final fatigue decision logic.

ISSN: 2715-6427

2.2 Annotation Process

The model adopted in this study is a keypoint (pose) detection model. This model was selected because it can measure the head-to-shoulder distance used to estimate head tilt, while yawning is captured through a dedicated Hand-at-Mouth detection class rather than a geometric measurement between mouth keypoints. Eye state (open/closed) is likewise determined through the model's class prediction rather than a computed Eye Aspect Ratio (EAR): because each eye is annotated with a single keypoint rather than the multiple landmarks EAR conventionally requires, eye openness is classified directly as a detection class, with the eye keypoint serving only to localize the eye region rather than to compute a ratio. Annotation was performed manually using the Roboflow platform to produce labels in the YOLO-Pose annotation format. Each sample was annotated with a bounding box around the facial region together with eight specific keypoints located on the face and upper body, as defined in Figure 2.

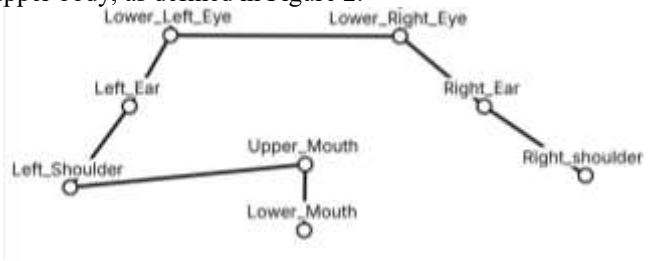

Figure 2. Person keypoint class definition.

The eight keypoints used in this study, along with their corresponding parameters and functions, are described in Table 1.

Table 1. List of keypoints

Index	Parameter	Function
1	Lower_Left_Eye	Detects the position of the left eye
2	Lower_Right_Eye	Detects the position of the right eye
3	Left_Ear	Detects the position of the left ear
4	Right_Ear	Detects the position of the right ear
5	Left_Shoulder	Detects the position of the left shoulder
6	Right_Shoulder	Detects the position of the right shoulder
7	Upper_Mouth	Detects the position of the upper lip
8	Lower_Mouth	Detects the position of the lower lip

The parameters described in Table 1 serve as the basis for detecting the physical features required to perform fatigue detection: the ear–shoulder keypoint pairs support estimation of head tilt, the upper- and lower-mouth keypoints mark the mouth's location so that the model can confirm whether a detected hand is positioned at the mouth, and the eye keypoints support localization of the eye region used in eye-state classification. Once labeling was complete, the next step was to train the model on the annotated dataset. The output of this training process is a PyTorch weight file together with a complete record of the model's detection capability. This record includes the confusion matrix, which summarizes how well the model distinguishes between the labeled classes and where confusion between classes occurs.

2.3 YOLOv8 Pose Architecture

YOLOv8 extends the original YOLO single-stage detection paradigm with a dedicated pose-estimation task, regressing keypoint coordinates in the same forward pass that produces the bounding box and class prediction, instead of requiring a separate landmark-extraction network [16], [17]. This suits the present study, which must localize the face and estimate eight keypoints from Table 1 from a single frame in near real time. YOLOv8 is composed of three stages a backbone, a neck, and a decoupled detection/pose head — that together turn a raw input image into a class probability, a bounding box, and a keypoint set, as illustrated in Figure 3.

The backbone extracts features through a Modified CSPDarknet53 structure, where the Cross-Stage Partial (CSP) strategy splits each feature map into a processed branch and a skip branch that are later merged, keeping gradients flowing well into deeper layers [16]. Its C2f module further splits, processes, and concatenates feature branches so that information from every stage is retained rather than only the last, producing richer features than earlier YOLO versions' CSP modules [14], [16]. An SPPF block then applies

pooling at several effective scales (5×5, 9×9, 13×13) cheaply, giving the backbone a multi-scale view — important here since the eye and mouth keypoints are much smaller in the frame than the shoulder keypoints.

A PAN-FPN neck combines a top-down path (semantic information flowing from deep to shallow layers) with a bottom-up path (localization information flowing back up) [16]. This dual-flow design lets both small keypoints (eyes, mouth) and larger ones (ear, shoulder) be localized reliably by the same model.

The head is decoupled and anchor-free: classification and box regression run through separate parallel branches, improving convergence and accuracy over earlier coupled heads [6], [14]. A third branch runs in parallel for pose estimation, predicting the (x, y) coordinates and a confidence value for each of the eight keypoints from the same features. Because all three branches share one forward pass, the model localizes the face and regresses all eight keypoints in a single inference step, retaining near real-time speed while avoiding the need for a separate landmark-extraction stage.

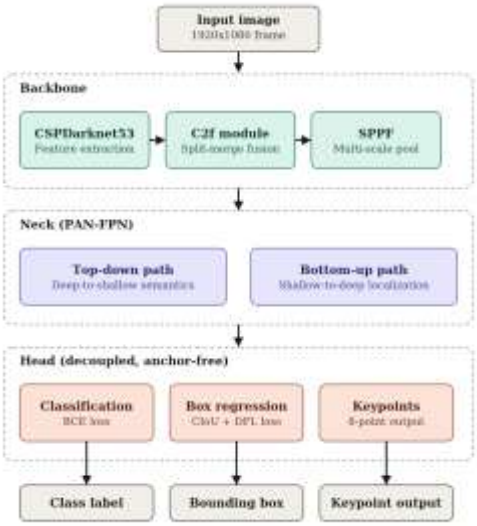

Figure 3. YOLOv8-pose Architecture

Three loss terms are used to train the head's three branches [18] Classification is trained with Binary Cross-Entropy (BCE) loss, which measures the difference between the predicted class probability p and the ground-truth label y ∈ {0, 1} for each candidate box:

$$L_{BCE} = -[y.\log(p) + (1-y).\log(1-p) \quad (1)$$

Bounding-box regression is trained using Complete Intersection over Union (CIoU) loss, which extends the standard IoU overlap metric by also penalizing the normalized distance between the predicted and ground-truth box centers and any inconsistency in their aspect ratio:

$$L_{CIoU} = 1 - IoU + \rho^2(b, b^{gt})/c^2 + \alpha v \quad (2)$$

where $\rho^2(b, b^{gt})/c^2$ is the Euclidean distance between the center points of the predicted box b and the ground-truth box b^{gt}, c is the diagonal length of the smallest box enclosing both, v measures aspect-ratio consistency, and α is a trade-off weight computed from the current IoU value. Because CIoU jointly considers overlap, center distance, and shape, it produces more stable gradients than plain IoU loss, particularly for small or elongated boxes such as those around the eye and mouth regions. In addition, box-coordinate regression uses Distribution Focal Loss (DFL), which models each box boundary not as a single regressed value but as a discrete probability distribution over a small range of candidate positions, and optimizes the two positions y_i and y_{i+1} nearest to the true value y:

$$L_{DFL} = -[(y_{i+1} - y).\log(S_i) + (y - y_i).\log(S_{i+1})] \quad (3)$$

where S_i and S_{i+1} are the predicted probabilities at the two neighboring candidate positions. DFL allows the model to express uncertainty about a box edge's exact location rather than committing to a single point estimate, which has been shown to improve localization precision for small or ambiguous objects [18].

In this study, the head-tilt indicator is derived from the relative positions of the ear and shoulder keypoints, the yawning indicator is derived from the distance between the upper- and lower-mouth keypoints, and the eye-closure indicator is derived from the detected state of the eye region localized by the corresponding eye keypoints. The architecture, dataset composition, and keypoint scheme follow the YOLOv8n-Pose configuration and 8-point annotation scheme.

ISSN: 2715-6427

2.4 Training Configuration

The YOLOv8-pose model was trained using the Ultralytics implementation on the 80/15/5 train–validation–test split described earlier, executed on Google Colaboratory, following common practice in comparable YOLOv8-pose fatigue-detection studies and mirroring the related thesis built on the same dataset and keypoint scheme [18].

A YOLOv8n-pose (nano) checkpoint was used, consistent with [19], chosen for its lowest parameter count and computational cost within the YOLOv8-pose family important given the goal of near real-time, resource-constrained inference rather than server-side deployment. Training was initialized from COCO-pretrained weights rather than from scratch, following standard practice, since COCO's existing human-pose priors reduce the epochs needed to converge on a smaller, task-specific dataset. Input images were resized to 640×640 pixels via letterbox resizing from the original 1920×1080 resolution, the Ultralytics default.

Reported epoch counts in comparable YOLOv8-based fatigue-detection literature range from roughly 100 to several hundred depending on dataset size; given the moderate dataset size here (2,625 images), training ran for 283 epochs with early stopping enabled to reduce overfitting risk. Batch size was set according to available GPU memory on the Colaboratory runtime (typically 16 for a T4-class GPU at 640×640), following a hyperparameter study showing YOLOv8's mAP is sensitive to batch size and should be tuned against hardware rather than fixed a priori [20].

Standard Ultralytics augmentations were applied mosaic (disabled in the final epochs), horizontal flipping, HSV jittering, and random translation/scaling consistent with augmentation strategies reported in related fatigue-detection studies for improving robustness to lighting and pose variation [7] [8].

2.5 Evaluation Metrics

Model performance was evaluated using standard object-detection and pose-estimation metrics, consistent with the evaluation approach adopted in related fatigue-detection studies [7] [8]. Four outcome categories were used to construct the confusion matrix: True Positive (TP), the number of instances correctly predicted as positive; False Positive (FP), the number of instances incorrectly predicted as positive; True Negative (TN), the number of instances correctly predicted as negative; and False Negative (FN), the number of instances incorrectly predicted as negative.

From these four quantities, Precision and Recall were computed as:

$$Precision = TP/(TP + FP) \tag{4}$$

$$Recall = TP/(TP + FN) \tag{5}$$

Precision reflects the proportion of the model's positive predictions that were actually correct, while Recall reflects the proportion of actual positive instances that the model successfully identified. The F1-score, computed as the harmonic mean of Precision and Recall, was used to summarize the balance between the two:

$$F1 = 2 \times (Precision \times Recall) \,/\, (Precision + Recall) \tag{6}$$

Mean Average Precision (mAP) was used to summarize detection performance across confidence thresholds, reported separately for the bounding-box (detection) task and the keypoint (pose) task at an IoU threshold of 0.5 (mAP50), following common practice in YOLO-based fatigue-detection evaluation. This box/pose separation matters in practice: in the related thesis using the same architecture and keypoint scheme, bounding-box detection reached a substantially higher mAP50 (0.967) than keypoint estimation (0.808), indicating that keypoint localization is a harder sub-task than simply detecting that a class is present, and should be evaluated and reported separately rather than assumed to track the detection score.

In addition, the confidence score, defined as the probability that a predicted bounding box and its associated keypoints correctly correspond to a given class was analyzed per indicator (eye closure, yawning, and head tilt) to assess how reliably the model distinguishes each fatigue-relevant condition. Precision-Confidence, Recall-Confidence, and F1-Confidence curves were used to identify the confidence threshold at which the model achieves the best balance between Precision and Recall, following the same analytical approach used in prior YOLO-based drowsiness-detection. As a reference point from the related thesis using the same keypoint scheme, box-level detection reached its optimal F1-score (0.96) around confidence 0.465, while pose-level keypoint estimation required a notably higher confidence threshold (0.99) to reach maximum precision indicating that, in general, keypoint-based indicators may need a stricter confidence threshold than simple object-presence detection to avoid false positives.

2.6 Fatigue Decision Logic

Because a single video frame is not, on its own, a reliable basis for judging fatigue a closed-eye frame may simply be an ordinary blink the three detected indicators are converted into a final Normal/Fatigue decision

using different evaluation windows, following the decision logic established in the related thesis this study builds on. Eye closure is evaluated through PERCLOS (the percentage of eyelid closure over time), computed as the proportion of frames within a rolling one-minute window in which the eyes are detected as closed. A Fatigue condition is flagged for this indicator only once PERCLOS exceeds a 15% threshold within that window, rather than from any single closed-eye frame. This accumulation-based rule is what allows the system to distinguish a normal blink, which contributes only briefly to the one-minute window, from sustained or repeated eye closure consistent with drowsiness.

Yawning is evaluated differently, by occurrence rather than accumulated duration: because the hand-at-mouth gesture used as its visual cue is already a distinctive, self-contained behavioral event, a single detected instance is sufficient to register as fatigue-relevant, without requiring it to persist across a time window. Head tilt is evaluated from the ear-to-shoulder keypoint distance described in Section 2.3. Since this distance decreases monotonically as the head tilts further from upright, a drop in the measured distance below a calibrated threshold is interpreted as a tilt consistent with drowsiness-related head nodding, in the same way that eye closure and yawning are interpreted as fatigue-relevant states.

In this study, the overall system classification is assigned as Fatigue when at least one predefined fatigue-related indicator exceeds its corresponding threshold. Eye closure is evaluated using the PERCLOS criterion, while a detected Hand-at-Mouth event is treated as a fatigue-relevant visual cue according to the predefined decision rule. Head tilt is determined based on the calibrated ear-to-shoulder distance threshold.

3. RESULTS AND DISCUSSION

3.1 Dataset Evaluation

Before training began, the fully annotated dataset was exported from Roboflow and its class composition was reviewed to confirm that every category needed for fatigue detection was adequately represented. Each image had already been labeled with a Person bounding box, a set of keypoints, and an eye-state or mouth-state class where applicable, and the counts below reflect the totals recorded at export time.

Table 2. Dataset composition by class.

Class	Number of Images	Number of Labels
Eyes Open	972	972
Eyes Closed	1,635	1,635
Yawning and Hand at Mouth	1,227	1,227
Person	2,625	2,625
Total	**2,625**	—

The dataset consists of 2,625 total images, each annotated with a Person bounding box and keypoint set, alongside class labels indicating eye state (open/closed) and mouth/hand state (yawning or hand-at-mouth) where present. Because a single image can contain more than one labeled indicator for example, a frame showing closed eyes together with a yawning mouth the sum of the individual class counts (972 + 1,635 + 1,227 = 3,834) exceeds the total image count, reflecting overlapping annotations rather than a labeling error. The class distribution is imbalanced, with Eyes Closed labels occurring approximately 1.7 times more frequently than Eyes Open labels. This imbalance should be considered when interpreting the per-class performance.

3.2 Training Performance

Once training reached its final epoch, the model's performance was evaluated on the held-out validation data using standard object-detection metrics, computed separately for the bounding-box (detection) output and the keypoint (pose) output, since the two prediction heads are optimized independently and can diverge in performance.

Table 2a. Training performance — bounding-box (detection) task.

Epoch	Precision	Recall	mAP50	mAP50-95	F1
283	0.91	0.97	0.967	0.614	0.96

Table 3b. Training performance — keypoint (pose) task.

Epoch	Precision	Recall	mAP50	mAP50-95	F1
283	0.98	0.89	0.817	0.62	0.87

The detection (box) task on Table 3a reached a high mAP50 of 0.967, indicating that the model reliably localizes the person, eye-state, and hand/mouth regions when only a loose overlap with the ground truth is required. The much lower mAP50-95 of 0.614 shows that this reliability drops substantially once a

ISSN: 2715-6427

stricter, multi-threshold IoU criterion is applied i.e., the model finds the correct region confidently but does not always draw a tightly fitting box, which is a common and expected pattern for small facial regions such as the eyes. The keypoint (pose) task shows the same qualitative pattern but at a lower level overall (mAP50 = 0.817, mAP50-95 = 0.62), confirming that regressing exact keypoint coordinates is a harder sub-task than simply detecting that a region is present.

It is also worth noting that the reported F1-scores (0.96 for box, 0.87 for pose) do not exactly equal the harmonic mean of the Precision and Recall values shown in the same row (which would give 0.94 and 0.93, respectively). This is expected: training logs of this kind typically report Precision and Recall at a representative operating point, while the F1-score is taken from the confidence threshold that maximizes the F1-confidence curve, which is not necessarily the same threshold.

3.3 Precision and Recall

To understand where the aggregate numbers above originate, the same validation results were broken down by individual class as shown I Table 4. A separate row is reported for keypoint performance on the Person class, since keypoint estimation is applied exclusively to that class the eye-state and hand-at-mouth classes are evaluated purely as detection targets, without an associated keypoint set.

Table 3. Precision, Recall, and F1-score per indicator.

Indicator	Precision	Recall	F1
Eyes Open	0.89	0.91	0.90
Eyes Closed	0.97	0.91	0.94
Hand at Mouth (Yawning)	0.98	0.97	0.97
Person (box)	1.00	1.00	1.00
Person (keypoints)	0.98	0.89	0.93

The Person class achieves perfect detection performance (Precision = Recall = 1.00), which is expected since the presence of a human face/upper body in the frame is the easiest and most consistent condition to detect across the dataset. The Hand-at-Mouth indicator is used in this study as the visual cue for yawning: rather than detecting the mouth's shape directly, which can be ambiguous (a wide-open mouth can resemble talking, laughing, or yawning), the system relies on the accompanying hand-to-mouth gesture that commonly occurs during a yawn as a more distinctive, easier-to-localize signal. This indicator performs strongly (F1 = 0.97), likely because the hand-to-mouth gesture produces a visually distinctive silhouette regardless of lighting, supporting its use as a practical stand-in for direct mouth-state detection. The lower precision for Eyes Open (0.89) indicates that some predictions classified as Eyes Open correspond to other conditions, making this class more susceptible to false-positive predictions than Eyes Closed. The keypoints (pose) task for the Person class shows the same precision-over-recall pattern seen at the aggregate level, reinforcing that the model is conservative about which keypoints it commits to rather than missing detections outright it tends to omit uncertain keypoints rather than mislocate them.

3.4 Confusion Matrix

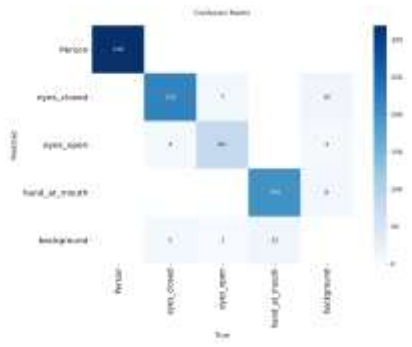

Figure 4. Confusion matrix for the trained YOLOv8n-Pose model

The confusion matrix provides a detailed view of the classification errors across the evaluated classes. The Person class was correctly detected in 319 instances without visible misclassification. For Eyes Closed, 216 instances were correctly classified, while 4 instances were predicted as Eyes Open and 5 as background. For Eyes Open, 81 instances were correctly classified, with 7 instances misclassified as Eyes Closed and 1 as background. The Hand-at-Mouth class achieved 199 correct predictions, while 11 instances were classified as

background. These results indicate that the dominant classification errors occurred between the two eye-state classes, suggesting that distinguishing Eyes Open from Eyes Closed remains more challenging than detecting Person or Hand-at-Mouth.

3.5 Confidence Analysis

Beyond whether a prediction is correct, it is useful to know how confident the model is when it gets things right. For each class, the confidence score assigned to every correctly classified instance in the validation set was averaged, giving a sense of how strongly the model commits to a correct answer once it makes one.

Table 4. Average confidence score per indicator.

Indicator	Average Confidence
Eyes Open	0.84
Eyes Closed	0.88
Hand at Mouth	0.82
Person	0.93

The Person class again shows the highest average confidence (0.93), consistent with its perfect precision and recall. Among the fatigue-relevant indicators, Eyes Closed is detected with the highest average confidence (0.88), slightly ahead of Eyes Open (0.84) and Hand-at-Mouth (0.82). This indicates that Eyes Closed produced the highest average confidence among the fatigue-related indicators. However, confidence scores alone do not determine detection reliability, which should be interpreted together with Precision, Recall, and F1-score. Because the confidence level required to achieve optimal performance may differ across indicators and between detection and keypoint tasks, a single fixed confidence threshold may not be optimal for all indicators. Per-indicator threshold tuning could therefore be considered in future work.

3.6 Effect of Head-Angle Variation

A live test examined sensitivity to head orientation. The test subject tilted their head to four target angles relative to an upright, forward-facing position 90° (upright), 70°, 50°, and 30° (increasingly tilted) and at each angle, the model's live output was used to measure the pixel distance between the ear and shoulder keypoints, the geometric quantity the head-tilt indicator is built on.

Table 5. Ear-to-shoulder keypoint distance across head angles.

Head Angle	Ear–Shoulder Distance (px)
90° (upright)	132.1
70°	103.3
50°	80.0
30°	61.8

The measured ear-to-shoulder distance decreased monotonically as the head angle changed from upright (90°) to a more tilted position (30°), decreasing from 132.1 px to 61.8 px. This represents approximately a 53% reduction in pixel distance over the tested 60° angular range. The observed monotonic relationship indicates that ear-to-shoulder distance can potentially be used as a simple geometric indicator of head tilt within the tested range. However, additional measurements across more subjects and head orientations are required to establish a robust threshold and assess its generalizability.

3.7 Error Analysis

To characterize how and why the system fails, recordings from the live testing sessions were reviewed manually, and frames where the model's near real-time prediction disagreed with the condition actually presented by the test subject were isolated. For each such frame, the model's output, its confidence score, and a probable cause were recorded based on visual inspection.

3.7.1 False Positive

Table 6. Sample on False Positive Condition

Sample	Actual Condition	YOLO Result	Detection Image
1	Eyes closed	Eyes open	

ISSN: 2715-6427

2	Eyes closed	Eyes open	

In two sampled failure cases in Table 7, the actual condition was closed eyes, but the model predicted eyes open, with confidence scores between 0.84 and 0.86, these confidence values are comparable to or higher than the average confidence for correct eyes-open predictions, meaning these errors would not have been caught by a stricter confidence threshold alone. In both cases, the identified cause was backlighting: strong light behind the subject appears to wash out the contrast between the eyelid and the eye region, making closed eyes visually resemble the open-eye appearance the model was trained on. This indicates a more concerning failure mode because the model produces a confident but incorrect prediction, which cannot be easily resolved through confidence-based filtering alone. This points toward lighting-normalization preprocessing (automatic exposure/backlight compensation) or targeted data augmentation with backlit training examples as a more effective mitigation than adjusting the confidence threshold.

3.7.2 False Negative

Table 7. Sample on False Negative Condition

Sample	Actual Condition	YOLO Result	Detection Image
1	Yawn	Normal	
2	Yawn	Normal	

In two sampled failure cases in Table 8, the actual condition was yawning, but the model predicted a normal (non-yawning) state. The identified cause was hand positioning: when a hand covers the mouth during a yawn, the model appears to key on the visible finger/hand silhouette near the mouth, but when a larger portion of the hand (beyond just the fingers) is visible, the resulting confidence score fell below the 0.6 operating threshold and the detection was suppressed. This suggests the Hand-at-Mouth class may have been trained on a narrower range of hand poses (e.g., fingers lightly covering the mouth) than actually occurs during natural yawning, where a full palm or fist is common. Expanding the training data to include a wider variety of natural hand-to-mouth poses during yawning, or lowering the confidence threshold specifically for this indicator, are both reasonable directions to reduce this false-negative rate.

3.8 Live Testing Overview

Beyond the specific failure modes above, it is useful to characterize the conditions under which live testing was carried out overall, and to show representative examples of the system working as intended.

3.8.1 Testing Conditions

Overall, direct testing involving variations in head angle demonstrated a clear effective operating range for the system show in Table 9. Live testing was conducted at distances between 60 cm and 100 cm; within this range, the model successfully localized the keypoints ears, shoulders, and mouth required for accurate indicator collection, whereas closer distances of 20–40 cm fell outside the limits of keypoint detection accuracy.

Table 8. Live Testing Conditions

Sample	Distance	Result
1	80 cm	
2	60 cm	
3	40	
4	20 cm	

3.8.2 Sample Detection Results

To illustrate correct detection across different combined conditions rather than summarize aggregate performance, several representative frames from the live testing sessions were selected show in Table 10. Each entry below is a single captured frame, with the actual condition, the model's predicted output, and the associated confidence score recorded at the moment of capture.

Table 9. Sample detection results across combined conditions

Detection Figure	Actual Condition	Predicted Condition	Confidence	FPS	Results
	Normal, Eyes Open	Eyes Open	0.85	3.5	Normal
	Normal, Eyes Closed	Eyes Closed	0.82	3.3	Normal

ISSN: 2715-6427

	Eyes Open, Hand at Mouth	Eyes Open, Hand at Mouth	0.75	3.4	Fatigue
	Eyes Closed	Eyes Closed	0.82	3.4	Normal

In all four samples, the model's predicted condition matches the actual condition exactly, with confidence scores ranging from 0.75 to 0.85. The comparatively lower confidence for the combined eyes-open-and-yawning case (0.75) versus the single-indicator cases (0.82–0.85) is consistent with the pattern observed throughout this section: the model is most confident when a single, clearly trained indicator is present, and somewhat less confident while still correct when multiple indicators must be resolved simultaneously in the same frame.

The Result column illustrates the behavior of the implemented fatigue-decision logic. The eyes-closed sample is classified as Normal because a single closed-eye frame does not exceed the one-minute PERCLOS threshold, whereas the Hand-at-Mouth event is classified as Fatigue according to the predefined decision rule. This demonstrates the distinction between instantaneous visual cues and time-window-based fatigue assessment.

CONCLUSION

This study evaluated a single-stage YOLOv8n-Pose approach for vision-based driver fatigue detection using multiple behavioral indicators, including eye closure, a hand-at-mouth gesture as a yawning-related visual cue, and head tilt. The proposed approach achieved a bounding-box mAP50 of 0.967 and a keypoint mAP50 of 0.817, with F1-scores of 0.90 for Eyes Open, 0.94 for Eyes Closed, and 0.97 for Hand-at-Mouth detection. These results demonstrate that YOLOv8n-Pose can simultaneously provide object detection and task-specific keypoint estimation for extracting fatigue-related behavioral cues within a single inference framework. The head-angle evaluation showed a monotonic relationship between ear-to-shoulder keypoint distance and head orientation within the tested conditions, indicating its potential as a simple head-tilt indicator. The implemented fatigue-decision logic further distinguishes transient eye closure from fatigue by evaluating eye closure using a one-minute PERCLOS window, while Hand-at-Mouth events are treated as fatigue-relevant cues. Live testing identified an effective camera distance of 60–100 cm. Although live inference was demonstrated, the observed processing rate of 3.3–3.5 FPS indicates that further optimization is required for practical real-time deployment. Error analysis showed that backlighting affected eye-state classification, while hand occlusion contributed to missed Hand-at-Mouth detections. Overall, the findings indicate that a single-stage pose-based approach has potential for multi-indicator vision-based fatigue detection. Future work should focus on improving inference efficiency, expanding subject and environmental diversity, addressing lighting and occlusion-related errors, and validating the fatigue-decision logic using larger and subject-independent datasets.

ACKNOWLEDGEMENTS

The authors would like to thank the DIPA of Politeknik Negeri Malang 2026 for funding this research.

REFERENCES

[1] G. Liu, K. Wu, W. Lan, and Y. Wu, “YOLO-FDCL: Improved YOLOv8 for Driver Fatigue Detection in Complex Lighting Conditions,” *Sensors*, vol. 25, no. 15, Aug. 2025, doi: 10.3390/s25154832.

[2] I. Zainafree, N. Syukria, S. Addina, and M. Z. Saefurrohim, “EPIDEMIOLOGI KECELAKAAN LALU LINTAS: TANTANGAN DAN SOLUSI,” *Bookchapter Kesehatan Masyarakat Universitas Negeri Semarang*, no. 1, pp. 92–127, May 2022, doi: 10.15294/km.v1i1.70.

[3] M. I. B. Ahmed *et al.*, "A Deep-Learning Approach to Driver Drowsiness Detection," *Safety*, vol. 9, no. 3, Sep. 2023, doi: 10.3390/safety9030065.

[4] Q. Abbas and A. Alsheddy, "Driver fatigue detection systems using multi-sensors, smartphone, and cloud-based computing platforms: A comparative analysis," Jan. 01, 2021, *MDPI AG*. doi: 10.3390/s21010056.

[5] D. Ayu Permatasari, E. Rizqi Kusuma Pradani, O. Desta Triswidrananta, F. Budi Irawan, and A. Adibah, "Sistem Deteksi Kantuk Real-Time Berbasis YOLOv5 dan Detak Jantung," *Indonesian Journal of Electronics and Instrumentation Systems (IJEIS*, vol. 16, No.1, pp. 1–9, Apr. 2026, doi: 10.22146/ijeis.111510.

[6] A. Jones, "Cutting-Edge Detection of Fatigue in Drivers: A Comparative Study of Object Detection Models," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.15030

[7] X. Ran, S. He, and R. Li, "Research on Fatigued-Driving Detection Method by Integrating Lightweight YOLOv5s and Facial 3D Keypoints," *Sensors*, vol. 23, no. 19, Oct. 2023, doi: 10.3390/s23198267.

[8] C. Chen, X. Liu, M. Zhou, Z. Li, Z. Du, and Y. Lin, "Lightweight and Real-Time Driver Fatigue Detection Based on MG-YOLOv8 with Facial Multi-Feature Fusion," *J. Imaging*, vol. 11, no. 11, Nov. 2025, doi: 10.3390/jimaging11110385.

[9] M. Arava and D. M. Sundaram, "Integrating lightweight YOLOv5s and facial 3D keypoints for enhanced fatigued-driving detection," *PeerJ Comput. Sci.*, vol. 10, pp. 1–27, 2024, doi: 10.7717/peerj-cs.2447.

[10] F. Majeed, U. Shafique, M. Safran, S. Alfarhood, and I. Ashraf, "Detection of Drowsiness among Drivers Using Novel Deep Convolutional Neural Network Model," *Sensors (Basel)*, vol. 23, no. 21, Oct. 2023, doi: 10.3390/s23218741.

[11] R. Florez, F. Palomino-Quispe, A. B. Alvarez, R. J. Coaquira-Castillo, and J. C. Herrera-Levano, "A Real-Time Embedded System for Driver Drowsiness Detection Based on Visual Analysis of the Eyes and Mouth Using Convolutional Neural Network and Mouth Aspect Ratio," *Sensors*, vol. 24, no. 19, Oct. 2024, doi: 10.3390/s24196261.

[12] D. A. Permatasari, G. Al Azhar, M. R. Zharfan, A. D. Risdhayanti, A. R. Hidayat, and D. Dewatama, "Design of Emergency Alarm System for Drowsiness Detection Using YOLO Method Based on Raspberry Pi," *Journal of Electrical, Electronic, Information, and Communication Technology*, vol. 6, no. 2, p. 78, Nov. 2024, doi: 10.20961/jeeict.6.2.93201.

[13] O. F. Hassan, A. F. Ibrahim, A. Gomaa, M. A. Makhlouf, and B. Hafiz, "Real-time driver drowsiness detection using transformer architectures: a novel deep learning approach," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-02111-x.

[14] C. Zhou, S. Liu, Y. Zhao, Y. Zhao, X. Li, and C. Cheng, "Research on Driver Facial Fatigue Detection Based on Yolov8 Model."

[15] G. C. Chang, B. H. Zeng, and S. C. Lin, "Efficient Eye State Detection for Driver Fatigue Monitoring Using Optimized YOLOv7-Tiny," *Applied Sciences (Switzerland)*, vol. 14, no. 8, Apr. 2024, doi: 10.3390/app14083497.

[16] T. Lu, K. Cheng, X. Hua, and S. Qin, "KSL-POSE: A Real-Time 2D Human Pose Estimation Method Based on Modified YOLOv8-Pose Framework," *Sensors*, vol. 24, no. 19, Oct. 2024, doi: 10.3390/s24196249.

[17] C. Aulton, C.-Y. Chiu, and S.-Y. Chiou, "Advancing functional reach assessments: a comparison of YOLOv8 human pose estimation and 3D motion capture in a young healthy cohort," *Journal of Biomechanics Open*, vol. 1, no. 2, p. 100007, 2026, doi: https://doi.org/10.1016/j.jbmo.2026.100007.

[18] R. N. Lestari, D. A. Permatasari, and G. Al Azhar, "Deteksi Kelelahan Real-Time Berbasis Pengolahan Citra Digital Sebagai Media Deteksi Perilaku Pengguna," *Techno.Com*, vol. 25, no. 3, Aug. 2026, doi: 10.62411/tc.v25i3.16841.

[19] D. Lasmana, N. L. Husni, and R. Kusumanto, "Deteksi Objek Menggunakan YOLOv5 dan YOLOv8 pada Perangkat Pemantauan Lingkungan," *Jurnal Elektronika dan Otomasi Industri*, vol. 12 No. 2, pp. 276–282, Jul. 2025, doi: http://dx.doi.org/10.33795/elkolind.v12i2/7615.

[20] Z. S. Hidayat, Y. A. Wijaya, and R. Kurniawan, "OPTIMIZING YOLOV8 FOR AUTONOMOUS DRIVING: BATCH SIZE FOR BEST MEAN AVERAGE PRECISION (MAP)," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 1147–1153, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.1626.