International Journal of Teaching, Learning and Education (IJTLE)
ISSN: 2583-4371
Vol-5, Issue-6, November-December 2026
Journal Home Page: https://ijtle.com/
Journal DOI: 10.22161/ijtle

# Student Engagement Monitoring System using Computer Vision

Amarpreet Kaur, Maged Nasser*

Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
amarpreet_22004311@utp.edu.com, maged.nasser@utp.edu.my

Received: 10 Oct 2026, Received in revised form: 05 Nov 2026, Accepted: 09 Nov 2026, Available online: 13 Nov 2026
©2026 The Author(s). Published by IJTLE. This is an open-access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

***Abstract***

*Lecturers in large classes cannot watch every student, so declining attention is usually noticed only after the lesson. This paper presents SEMS (Student Engagement Monitoring System), a computer vision system that classifies each student as Engaged, Partially Engaged or Disengaged from five visual cues (eye gaze, eye openness, yawning, head pose and body posture), alerts the lecturer in real time and suggests a teaching action that suits the lesson. The cues are measured with MediaPipe face and pose landmarks, converted into indicator levels and classified by explicit rules and by trained models. A dataset of 6,780 student images was built from the public SCB classroom datasets and split by classroom photographs, so that test classrooms were never seen in training. A rule baseline, a Decision Tree, a multilayer perception, an EfficientNetV2-B0 image model and a weighted fusion of cue and image predictions were compared. The fusion reached 79.8% accuracy and a macro F1 of 0.796 on 1,328 test students, doubling the 39.95% of the first model but remaining below the 90% target; sleeping students, whose faces are hidden, were the main source of error. In a simulated class of 24 students the live system flagged disengaged students 2.4 s after drowsiness began, raised no alerts while the class was attentive, named the affected rows and delivered alerts to a phone dashboard in about 3 ms.*

***Keywords*—student engagement, computer vision, facial landmarks, learning analytics, real-time alerts, teacher dashboard.**

## I. INTRODUCTION

Student engagement is closely linked to academic success, and its behavioral side is the one a lecturer can see: whether students look towards the front, keep their eyes open, yawn, hold their heads up and sit upright [1]. In Malaysian universities, however, active participation is limited and many students learn passively [2], while students value timely interaction with their instructor more than any other form of interaction [3]. In a hall of fifty or more students no lecturer can follow these signs for everyone at once, so disengagement is often noticed too late to change the lesson.

Three connected problems motivate this work. First, manual monitoring of engagement is impractical in large classes, and existing automatic systems usually rely on one or two cues and only two states. Second, lecturers lack real-time feedback: most systems report after the class or weekly. Third, even when lecturers know that students are disengaged, current tools do not tell them what to do about it.

This paper presents SEMS, a system that classifies each student into three engagement levels from five visual cues, shows the class on a mobile-friendly dashboard with alerts, and suggests teaching actions. The work had three objectives: (1) to build a computer vision system that classifies students as Engaged,

Partially Engaged or Disengaged from eye gaze, eye openness, yawning, head pose and body posture, (2) to deliver real-time notifications to lecturers through a mobile-friendly system, and (3) to give suggestions based on the lesson type and the engagement data.

The contributions are as follows. First, an explainable five-cue pipeline is built on lightweight on-device landmark models, with explicit indicator thresholds and classification rules. Second, a reproducible dataset is constructed from public classroom images with a split by photograph that prevents leakage, and rule-based, cue-based, image-based and fused classifiers are compared on it. Third, cues are judged over a five-second window and turned into Critical and Trend alerts and lesson-specific suggestions on a dashboard that runs on the lecturer's own laptop and phone. Fourth, the paper documents where the final system departs from the original proposal and why. Real classroom trials with lecturers, audio and physiological signals, and identity recognition are outside the scope of this study.

The rest of the paper is organized as follows. Section II reviews related work, Section III describes the method and the system, Section IV reports results, Section V discusses them against the objectives and prior work, and Section VI concludes with recommendations and future work.

## II. RELATED WORK

*A. Detecting and classifying engagement.* Observers agree well when they judge engagement from faces, and machines can learn these judgements [4], but intermediate levels of engagement are hard both to label and to recognize [5]. Stidi and Nasser [6] combined eye gaze and posture using YOLOv11, MoveNet and MediaPipe FaceMesh with MLP, SVM and Random Forest classifiers and reached 88% accuracy and an F1 of 0.86 on the SCB dataset, although the classification was binary and used static frames. Wang et al. [7] added attention to YOLOv5 to recognize seven classroom behaviors (88.13% mAP), and Jiang et al. [8] classified concentration from head pose with 94.6% accuracy, but only for two states. Across these studies, detection accuracy is reported, while the step from detection to support for the lecturer is rarely taken.

*B. Real-time feedback for teachers.* Anh et al. [9] tracked gaze and summarized attention in charts (93.33% gaze accuracy), leaving interpretation to the teacher. Sabuncuoğlu and Sezgin [10] built an engagement dashboard from face and pose features with a kNN classifier, but reached an F1 of only 0.684 and gave no recommendations. Dan et al. [11] delivered real-time alerts with 94 ms latency from visual, audio and physiological signals, at the cost of extra sensors that a normal classroom does not have.

*C. Prescriptive support.* Firat [12] predicted disengagement early (73.3% accuracy) but for intelligent tutoring systems rather than lecturers, and Pillai [13] computed weekly engagement scores, too late for intervention during a lesson. A review of 80 studies concluded that few systems turn recognition results into actionable support for teachers [14], although evidence-based strategies such as quizzes, think-pair-share and short breaks are well documented [15]. Table I summarizes these studies.

*Table I: Summary of related studies*

| Study | Method | Key finding | Limitation |
|---|---|---|---|
| Stidi and Nasser [6] | Gaze and posture; MLP, SVM, RF | 88% accuracy, F1 0.86 | Binary classes; static frames |
| Wang et al. [7] | YOLOv5 with attention, 7 behaviours | 88.13% mAP | False detections at side angles |
| Jiang et al. [8] | Head detection and head pose | 94.6% accuracy | One cue; two states |
| Anh et al. [9] | Face detection and gaze | 93.33% gaze accuracy | Charts only; teacher interprets |
| Sabuncuoğlu and Sezgin [10] | Face and pose; kNN dashboard | F1 0.684 | No suggestions |
| Dan et al. [11] | Visual, audio, physiological; BiLSTM | Alerts in 94 ms | Needs extra sensors |
| Firat [12] | LSTM on visual cues | 73.3% accuracy | Tutoring systems; no teacher alerts |
| Pillai [13] | YOLOv4, two classes | 77.08% mAP50 | Weekly, not real time |
| Liu et al. [14] | Review of 80 studies | Four recognition purposes | Few systems support teachers |

*D. Research gap and SEMS's response.* Three gaps follow from the literature. Few works combine several visual cues into three engagement levels; few deliver feedback


Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.2

to the lecturer during the lesson on ordinary hardware; and none recommends an action that depends on both the lesson type and the kind of disengagement observed. SEMS addresses the first with five cues, explicit thresholds and a fusion with an image model, the second with a five-second judging window, Critical and Trend alerts and a phone dashboard that needs only a webcam, and the third with a suggestion module and chatbot that combine the dominant cue with the lesson type. Unlike most prior work, it also benchmarks the trained models against a rule baseline that needs no training.

## III. METHODOLOGY

### *A. System Architecture*

SEMS has three phases that match the three objectives (Fig. 1). In Phase 1 the system detects every student in the webcam frame and extracts the five cues. In Phase 2 it classifies each student and computes class-level engagement every second. In Phase 3 it presents the results on a dashboard with alerts and suggestions. The classifiers used in Phase 2 are trained beforehand in Google Colab on the same cue extractor. The model was developed following the CRISP-DM process [16]: a first model trained on an initial behavior mapping and a random split performed poorly, and its error analysis led back to data preparation, as reported in Section IV.

### *B. Phase 1: Data Acquisition and Preprocessing*

Two public datasets were downloaded from Kaggle: the SCB Classroom Dataset with about 13,300 classroom images [17] and the SCB-05 subset used in a public notebook with about 3,185 images. In both, every student is marked with a bounding box and a behavior label. Duplicate images were found with MD5 hashes and kept once. Each labelled student was cropped with a 25% margin, made square and resized to 256×256 pixels, and students smaller than 48 pixels were removed because their faces cannot be measured.

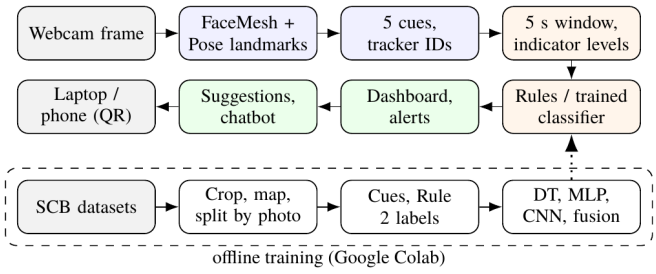



*Fig. 1. SEMS architecture as implemented: live pipeline on the lecturer's laptop (top) and offline training in Google Colab (bottom).*

Four preparation steps followed. (i) Seven behaviors were mapped to engagement categories: reading, writing, listening and hand raising to Engaged; sleeping (including "leaning on the desk"), using phones and looking away (including "turn head") to Disengaged. Ambiguous labels such as talking or standing were not used. (ii) The five cues were extracted from every crop (Section III-D). (iii) Because no behavior corresponds to partial engagement, an Engaged behavior student whose cues met classification Rule 2 (two or three negative signs) was relabeled Partially Engaged, provided that at least three cues could be measured. (iv) The data were split 70/10/20 into training, validation and test sets by classroom photograph, so that all students from one photograph fall in the same set, and each set was sampled with a fixed seed.

### *C. Dataset*

The SCB images show real classrooms photographed from the front, with dozens of students per image [17]. Table II gives the final dataset. Engaged and Disengaged reached their planned sizes, while fewer students met the Partially Engaged condition than planned, so all of them were used and the total is 6,780 rather than the planned 7,145. The data contain no "listening" label, so the Engaged class consists of reading, writing and hand raising.

Two features of the data affected the design. Among the 15,932 candidate crops, the eye-based cues (gaze, openness and yawning) could be measured for only 24% of students, while head pose was measurable for 74% and posture for 68%, because many faces are small, bowed or turned. In addition, a student asleep on the desk hides the face, so the cues of a sleeping student resemble those of a student writing with the head down. The main limitation of the data is that they are still photographs, so the time-based indicators could only be approximated in training.

*Table II Final dataset by split and category*

| Category | Train | Val. | Test | Total |
|---|---|---|---|---|
| Engaged | 2,000 | 285 | 572 | 2,857 |
| Partially Engaged | 1,282 | 170 | 327 | 1,779 |
| Disengaged | 1,500 | 215 | 429 | 2,144 |
| **Total** | **4,782** | **670** | **1,328** | **6,780** |

### *D. Phase 1: Detection and Feature Extraction*

For each frame, MediaPipe FaceLandmarker detects up to 30 faces with 478 landmarks each, including the


Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.2

irises [18], [19], and MediaPipe Pose locates 33 body landmarks around each student [20]. A tracker assigns anonymous identifiers and keeps a student for 8 s after the face is lost, so that a student who puts the head down is still measured from the body. The five cues are then computed (Fig. 2). Eye gaze is the offset of each iris from the centre of the eye, scaled from 0 to 1. Eye openness uses the eye aspect ratio [21]:

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

where $p_1$ to $p_6$ are six eye landmarks; an open eye has an EAR of about 0.27–0.30 and a closed eye about 0.10. Yawning uses the analogous mouth aspect ratio (MAR), and a yawn is counted when MAR ≥ 0.6 for at least 0.8 s. Head pose is expressed as yaw, pitch and roll from the nose, ears and eyes, and posture as a slouch score from head drop, shoulder tilt and spine lean. A three-second calibration records the neutral pose of the class, all values are scaled to 0–1, and missing cues are filled from recent frames (live) or the training median (offline), giving a five-value feature vector per student.

### *E. Phase 2: Classification and Engagement Detection*

Each cue is mapped to a level using the thresholds in Table III; a cue in the Engaged column is a positive sign and the other two columns are negative signs. The rules in Table IV are applied in the order Rule 1, Rule 3, Rule 2, so that a clearly disengaged student is never reported as only partially engaged; the overlap of Rules 2 and 3 at exactly three negative signs is resolved by requiring three cues at the Disengaged level for Rule 3.

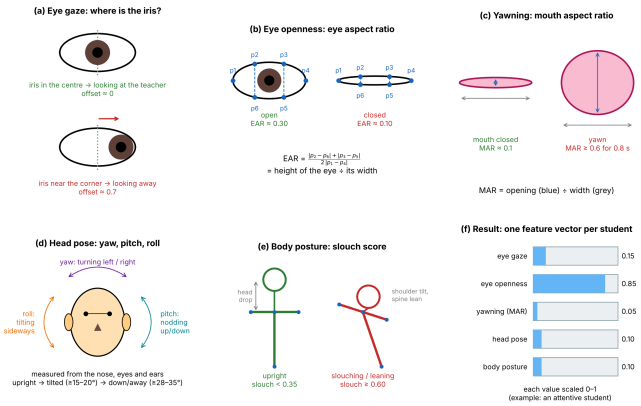



*Fig. 2. Measurement of the five visual cues from the MediaPipe landmarks.*

*Table III Engagement indicators and thresholds*

| Cue | Engaged | Partially Engaged | Disengaged |
|---|---|---|---|
| Eye gaze | Away < 30% of time | Away 30–65% | Away ≥ 65% |
| Eye openness | Score ≥ 0.55 | 0.30–0.55 | < 0.30 |
| Yawning | None in 60 s | 1–2 times | ≥ 3 times |
| Head pose | Upright | Roll ≥ 15°, yaw ≥ 20° or pitch ≥ 15° | Yaw ≥ 35° or pitch ≥ 28° |
| Body posture | Slouch < 0.35 | 0.35–0.60 | ≥ 0.60 |

*Table IV Classification rules*

| Rule | Condition | Class |
|---|---|---|
| Rule 1 | ≥ 4 cues show positive signs | Engaged |
| Rule 2 | 2–3 cues show negative signs | Partially Engaged |
| Rule 3 | ≥ 3 cues at the Disengaged level, or ≥ 4 negative signs | Disengaged |

Besides the rules, four models were developed (Table V). A Decision Tree (DT) and a multilayer perceptron (MLP) learn from the five-cue vector, with missing values imputed by the training median and features scaled to 0–1; their settings were chosen by grid search on validation macro F1. Because many faces are too small or turned for the cues to be measured, an EfficientNetV2-B0 network [22-25] pretrained on ImageNet was fine-tuned on the student images, which also shows objects such as pens, books and phones. Finally, the better cue model and the image model were fused:

$$p = wp_{\mathrm{image}} + (1 - w)p_{\mathrm{cues}} \quad (2)$$

where p is the class probability and w the image weight, chosen on validation from 0 to 1 in steps of 0.1, following the multimodal approach of [6].

*Table V Models and settings*

| Model | Settings searched |
|---|---|
| Rule baseline | Tables III and IV; no training |
| Decision Tree | Max depth 4, 6, 8, 12; min leaf 2, 5, 10; balanced weights |


Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.2

| Model | Settings searched |
|---|---|
| MLP | Hidden (32,16), (64,32), (128,64), (64,32,16); L2 $10^{-4}$, $10^{-3}$, $10^{-2}$; 500 iterations |
| EfficientNetV2-B0 | 224×224; rotation, flip, zoom; head 4 epochs, fine-tune ≤ 20 epochs; label smoothing 0.1; early stopping |
| Fusion | w = 0, 0.1, …, 1 on validation |

In the live system a single frame can mislead, since a student who blinks or glances away briefly is not disengaged. Cues are therefore aggregated over the last 5 s (yawns over 60 s), so that "sometimes away" is measured as a share of time, and a class score is computed every second:

$$S = \frac{100n_E + 50n_P}{n} \quad (3)$$

*F. Phase 3: Dashboard and Prescriptive Suggestions*

The results are served by a Flask web server on the lecturer's laptop and pushed every second to a mobile-friendly dashboard through Server-Sent Events (Fig. 3). It shows class counts, the score, a timeline, the disengagement indicators, a classroom heatmap by row, the alerts and the current suggestion; a phone on the same network opens it by scanning a QR code, and a one-click launcher starts the whole system. A Critical alert is raised when at least five students, or 30% of a small class, are Disengaged, and a Trend alert when the score falls by 15 points within 60 s, naming the row with the largest drop; each alert type has a 45 s cool-down.

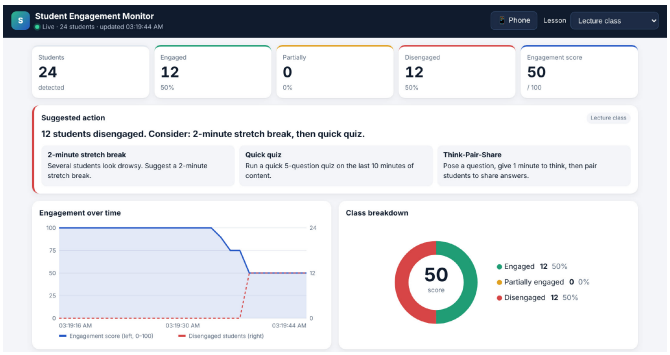



*Fig. 3. Teacher dashboard during a simulated drowsy period: class counts, score, Critical alert, suggestion and timeline.*

The suggestion module combines the dominant problem with the lesson type selected by the lecturer. If most disengaged students are drowsy, it suggests a two-minute stretch break, if they are looking away it advises walking closer to the named row, and otherwise it suggests a group discussion; two strategies for the lesson type are then added, such as a quick quiz in a lecture, a case study in a hands-on session or peer coaching in an interactive class [15]. A rule-based chatbot answers questions such as "who is disengaged?" and "why is S8 disengaged?" from the live data, without a generative model.

*G. Evaluation Protocol*

Models were selected on the validation set and scored once on the test set of 1,328 students from unseen classrooms. Table VI lists the metrics; all are macro averages over the three classes, with precision P = TP/(TP + FP), recall R = TP/(TP + FN) and F1 = 2PR/(P + R). The targets were an accuracy of at least 90% and precision, recall and F1 of at least 85%, set from the 88% reported in [6]. The live system was tested with eight automated tests, a simulated 24-student class with a drowsy period and a period in which the back rows look away, and 100 alert-delivery trials.

*Table VI Evaluation metrics and their purpose*

| Metric | Purpose | Target |
|---|---|---|
| Accuracy | Overall correctness | ≥ 90% |
| Precision | Few false alarms | ≥ 85% |
| Recall | Catch disengaged students | ≥ 85% |
| F1-score | Balance of precision and recall | ≥ 85% |
| Detection delay | Time to flag after onset | Seconds |
| Alert latency | Decision to dashboard | Milliseconds |

*H. Tools and Reproducibility*

The system was written in Python with MediaPipe, OpenCV, scikit-learn, TensorFlow/Keras and Flask. Models were trained on Google Colab with an NVIDIA T4 GPU from three notebooks (data preprocessing; feature extraction and training; real-time alerts and dashboard), and the live system runs on a laptop without a GPU. Random seeds were fixed and the split is stored in a manifest, so the dataset and results can be rebuilt. No images or video are stored at run time, and students are identified only by anonymous IDs.

*I. Changes from the Proposed Methodology*

Table VII lists where the final implementation differs from the methodology proposed at the start of the project, and why.


Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.2

*Table VII Changes from the proposed methodology*

| Proposed | Implemented | Reason |
|---|---|---|
| MediaPipe or YOLO for detection | MediaPipe FaceMesh and Pose | Gives landmarks for all five cues on a laptop |
| Decision Tree or MLP | Rules, DT, MLP, image model and fusion | Cues often unmeasurable on small faces |
| Three rules as written | Rule 3 checked first; overlap resolved by levels | Rules 2 and 3 overlap at three negatives |
| 7,145 images | 6,780 images | Fewer students met the Rule 2 condition |
| Listening as an Engaged behaviour | Not present | No listening label in the SCB data |
| Single-frame judgement | 5 s window; yawns over 60 s | Needed to measure "sometimes away" and repeated yawns |
| HTML/CSS/JS or Flutter app | Flask web dashboard with QR code | Runs on laptop and phone without installation |

## IV. RESULTS

### *A. Preliminary Findings*

The first model, a Decision Tree trained before the mapping and the split were corrected, reached only 39.95% accuracy and a macro F1 of 0.377 on 11,952 test crops, and recognized only 14.3% of Engaged students; most were predicted Partially Engaged (2,215) or Disengaged (2,207). Three causes were found: (i) labels outside the engagement mapping, such as bowing the head or talking, had been assigned to categories and overlapped in the cues; (ii) writing students look down like sleeping students; and (iii) a random split put students from the same photograph in both training and testing, while many crops were too small to measure. Restricting the data to the seven mapped behaviors, splitting by photograph, removing tiny faces and requiring three measurable cues produced the final design.

### *B. Model Results*

Table VIII compares the final approaches on the 1,328 test students (Fig. 4). The cue models were close to each other: the grid search chose a depth of 6 and a minimum leaf of 10 for the Decision Tree (validation F1 0.623) and hidden layers (64, 32, 16) with L2 $10^{-3}$ for the MLP (validation F1 0.620), and the Decision Tree was kept as the cue model. The image model alone reached 65.9%. The fusion gave the best result, 79.8% accuracy and a macro F1 of 0.796, about 19 points above the Decision Tree and 14 points above the image model. No model reached the 90% target.

*Table VIII Test results (macro average, 1,328 students)*

| Approach | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| Preliminary model | 0.400 | 0.431 | 0.434 | 0.377 |
| Rule baseline | 0.453 | 0.506 | 0.522 | 0.433 |
| Decision Tree | 0.605 | 0.586 | 0.631 | 0.600 |
| MLP | 0.607 | 0.590 | 0.627 | 0.603 |
| EfficientNetV2-B0 | 0.659 | 0.665 | 0.663 | 0.658 |
| **Fusion (DT + image, w = 0.4)** | **0.798** | **0.797** | **0.810** | **0.796** |
| Stidi and Nasser [6] (binary) | 0.880 | 0.870 | 0.860 | 0.860 |

Preliminary model scored on 11,952 crops before the design was corrected. Bold: main model.

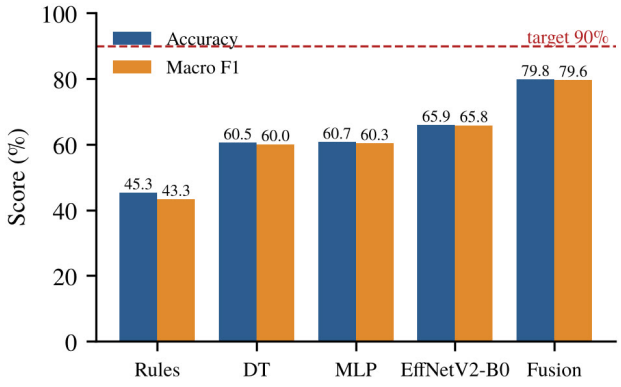



*Fig. 4. Test accuracy and macro F1 of the five final approaches; the dashed line marks the 90% target.*

The fusion weight was chosen on the validation set (Fig. 5). Cues alone (w = 0) gave 62.8% validation accuracy and images alone (w = 1) 70.4%, while w = 0.4 gave 81.9%, showing that the two sources complement each other rather than one simply dominating.

### *C. Per-Class and Per-Behavior Results*

Table IX and Fig. 6 give the fusion results by class. Partially Engaged had the highest recall (0.963) and

Disengaged the lowest (0.657): of 429 Disengaged students, 91 were predicted Engaged and 56 Partially Engaged. By behavior (Fig. 7), hand raising (86.9%) and looking away (82.6%) were recognized best, while writing (73.0%) and especially sleeping (44.8%) were hardest; 42.8% of sleeping students were predicted Engaged. Partially Engaged students of every behavior were recognized at 95.2–98.3%.

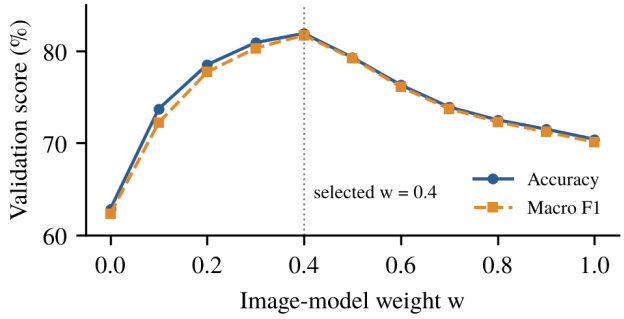



*Fig. 5. Validation accuracy and macro F1 of the fusion for different image-model weights.*

*Table IX Fusion results by class (test set)*

| Class | Prec. | Rec. | F1 | Support |
|---|---|---|---|---|
| Engaged | 0.825 | 0.809 | 0.817 | 572 |
| Partially Engaged | 0.743 | 0.963 | 0.839 | 327 |
| Disengaged | 0.822 | 0.657 | 0.731 | 429 |

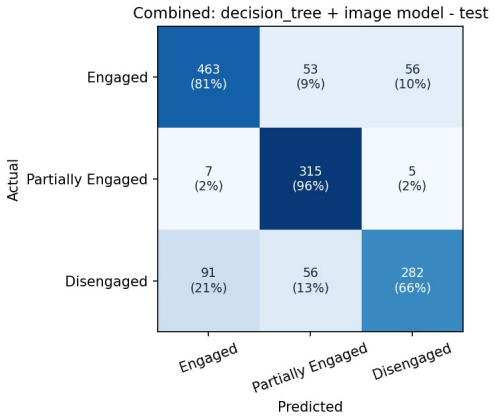



*Fig. 6. Confusion matrix of the fusion on the test set (1,328 students).*

The image model overfitted: its training accuracy rose above 95% while validation accuracy stayed at 67–69%, and early stopping ended training after 15 epochs.

*D. Real-Time System*

All eight automated tests passed. Table X and Fig. 8 summaries the simulated class. In calm periods all 24 students were Engaged and no alert was raised. In the drowsy period the first students were flagged Disengaged 2.4 s after onset, 12 students were Disengaged at the steady state, the score fell to 50, and Critical and Trend alerts named the front row (−35 points, 42.4 s) and the middle rows (−48 points, 87.4 s). In the distracted period ten students who looked away with turned heads were classified Partially Engaged, as Rule 2 intends, and a Trend alert named the back row. The class returned to all Engaged 3.4 s after recovery.

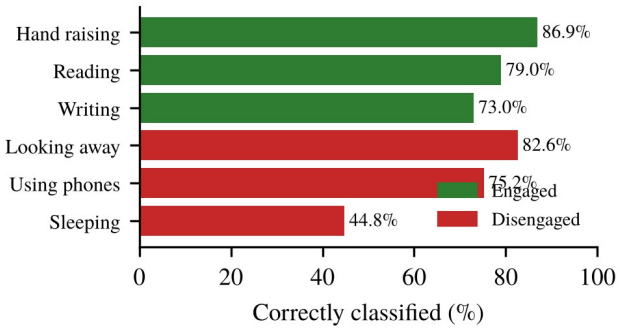



*Fig. 7. Share of test students classified correctly by the fusion, by behavior.*

*Table X Simulated classroom (24 students)*

| Period | E/P/D | Score | System response |
|---|---|---|---|
| Calm | 24/0/0 | 100 | No alerts |
| Drowsy | 12/0/12 | 50 | Flagged after 2.4 s; Critical and Trend alerts (front, middle rows) |
| Recovery | 24/0/0 | 100 | All Engaged after 3.4 s |
| Distracted | 14/10/0 | 79.2 | Trend alert (back row) |

E/P/D: Engaged / Partially Engaged / Disengaged.

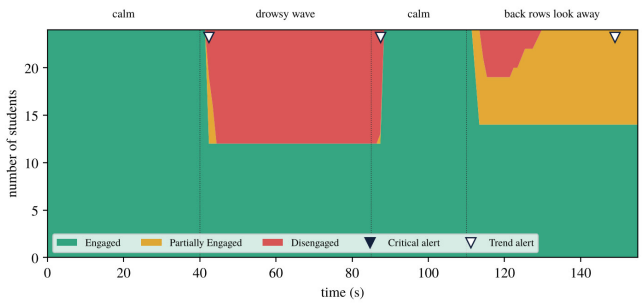



*Fig. 8. Class breakdown over time in the simulated classroom.*


Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.2

Over 100 trials the alert decision took 3.4 ms on average (95th percentile 4.1 ms) and delivery to the dashboard 3.3 ms (95th percentile 4.2 ms, maximum 6.0 ms). One Critical alert arrived 2.4 s after the drowsy period had ended because the window still held drowsy frames. With a webcam in Google Colab the pipeline processed 2.4 frames per second, a figure that includes sending each frame between the browser and the cloud server.

*E. Suggestions and Chatbot*

Table XI shows the suggestions produced by the system during the simulation. In every case the first suggestion matched the dominant problem and the other two matched the lesson type. The lesson strategies changed from minute to minute, so the lecturer receives new ideas rather than the same advice. During the drowsy period the chatbot listed the 12 disengaged students (S8–S15, S19, S20, S22, S23), located them mainly in the middle rows and explained each decision, for example "S8 is disengaged: drowsy, yawning, head down, slouching".

*Table XI Suggestions produced during the simulation*

| Moment | Lesson | Suggestions shown |
|---|---|---|
| Drowsy | Lecture | Stretch break; purposeful pause; quick write |
| Drowsy | Hands-on | Stretch break; smaller groups; concept mapping |
| Drowsy | Interactive | Stretch break; students as partners; student-led discussion |
| Distracted | Lecture | Walk closer to back row; think-pair-share; purposeful pause |
| Distracted | Hands-on | Walk closer to back row; think-aloud; smaller groups |
| Distracted | Interactive | Walk closer to back row; student-led discussion; peer coaching |
| Calm | Any | "Class is engaged – keep going!" |

## V. DISCUSSION

*Objective 1: classification.* The system classifies students into three engagement levels from five cues, but the best accuracy of 79.8% is below the 90% target, so this objective was partly achieved. The most important design decision was the return to data preparation, which doubled accuracy and raised Engaged recall from 0.143 to 0.809. Four findings explain the remaining gap. First, the photographs limited the cues: with eye cues measurable for only 24% of students, the cue models were often decided from head and body alone and stayed near 60%. Second, the cues and the image model are complementary, since the cues work when the face is visible and the image model recognizes phones, books and pens when it is not. Third, sleeping on the desk looks like writing with a bowed head in a single photograph, which explains both the low recall for sleeping and for the Disengaged class. Fourth, the high recall of Partially Engaged must be read with care, because that class was labelled with the same cue rules the models learn; it measures agreement with the rules rather than with human judgement.

*Objective 2: real-time notification.* This objective was achieved. Disengagement was flagged within 2.4 s, alerts reached the dashboard in milliseconds, no alerts were raised while the class was attentive, and the Trend alerts named the correct rows, telling the lecturer where to act. The delay the lecturer sees comes from the five-second window, which is also what prevents alerts from blinks and short glances.

*Objective 3: suggestions.* This objective was achieved: the suggestions responded to both the dominant cue and the lesson type in every case, and the chatbot answered from the same data as the dashboard. Their effect on learning was not measured.

*Relation to prior work.* The 88% reported in [6] is not directly comparable, because that study used two classes, whereas SEMS separates three levels and is tested only on unseen classrooms. Compared with single-cue systems [8], [9], SEMS uses five cues and explains each decision; compared with weekly reports [13] and charts [9], it informs the lecturer during the lesson; and unlike [11], it needs only a webcam.

*A. Limitations*

Training and testing used still photographs, so the time-based indicators could only be approximated offline; the Partially Engaged class is defined by rules rather than by human raters; and the image model overfitted the small training set. The live system was tested with a simulated class and a webcam rather than with lecturers in real classes, and the phone must be on the same network as the laptop. One camera position and one laptop were used, so frame rate and accuracy in large halls remain to be measured.


Crossref DOI: https://dx.doi.org/10.22161/ijtle.5.6.2

## VI. CONCLUSION

SEMS classifies students as Engaged, Partially Engaged or Disengaged from five visual cues measured with lightweight landmark models, alerts the lecturer within seconds on a laptop or phone, and suggests actions that fit the lesson and the observed problem. On unseen classrooms the fusion of a Decision Tree on the cues with an EfficientNetV2-B0 image model reached 79.8% accuracy and a macro F1 of 0.796, doubling the first model; the live system flagged disengagement within 2.4 s, named the affected rows and raised no false alerts in the simulation.

### *A. Recommendations*

Three recommendations follow. First, treat an alert as a prompt to look at the named row rather than as a judgement about a student. Second, place the camera so that faces in the back rows are large enough to measure, since cue coverage limited accuracy more than the choice of classifier. Third, inform students and obtain consent in line with university policy before any classroom use.

### *B. Future Work*

Four directions would extend this work. The first is a dataset of labelled classroom videos with human-rated partial engagement, so that the time-based cues can be trained and tested directly. The second is object detection for pens, books and phones to separate writing and reading from sleeping. The third is higher-resolution or multiple classroom cameras and server-hosted push notifications. The fourth is a semester-long trial with lecturers to measure the effect of the suggestions on engagement.

## REFERENCES

[1] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Rev. Educ. Res.*, vol. 74, no. 1, pp. 59–109, 2004, doi: 10.3102/00346543074001059.

[2] M. Mat Halif *et al.*, "Moderating effects of student motivation on the relationship between learning styles and student engagement," *Asian J. Univ. Educ.*, vol. 16, no. 2, pp. 94–103, 2020, doi: 10.24191/ajue.v16i2.10301.

[3] K. S. Gurman, Y. F. Chan, M. W. Khor, A. Samsul Kamil, A. A. Norazah, and P. Krishnan, "Learner engagement in 21st century technology integrated classrooms: A Malaysian perspective," *Asian J. Univ. Educ.*, vol. 21, no. 3, 2025, doi: 10.24191/ajue.v21i3.76.

[4] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, 2014, doi: 10.1109/TAFFC.2014.2316163.

[5] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," arXiv:1609.01885, 2016.

[6] M. Stidi and M. Nasser, "Multi-modal student engagement detection using eye-gaze and posture analysis under obstructed conditions," *Int. J. Teach. Learn. Educ.*, vol. 4, no. 6, pp. 1–7, 2025.

[7] Z. Wang, J. Yao, C. Zeng, W. Wu, H. Xu, and Y. Yang, "YOLOv5 enhanced learning behavior recognition and analysis in smart classroom with multiple students," in *Proc. Int. Conf. Intell. Educ. Intell. Res. (IEIR)*, 2022, pp. 23–29, doi: 10.1109/IEIR56323.2022.10069015.

[8] B. Jiang, W. Xu, C. Guo, W. Liu, and W. Cheng, "A classroom concentration model based on computer vision," in *Proc. ACM Turing Celebration Conf.–China*, 2019, pp. 1–6, doi: 10.1145/3321408.3322856.

[9] B. N. Anh *et al.*, "A computer-vision based application for student behavior monitoring in classroom," *Appl. Sci.*, vol. 9, no. 22, Art. no. 4729, 2019, doi: 10.3390/app9224729.

[10] A. Sabuncuoğlu and T. M. Sezgin, "Developing a multimodal classroom engagement analysis dashboard for higher-education students," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. EICS, Art. no. 188, 2023, doi: 10.1145/3593240.

[11] Z. Dan, Z. Yali, and Z. Tong, "A multimodal AI framework for real-time student engagement detection and adaptive feedback in higher education," *J. King Saud Univ.–Comput. Inf. Sci.*, vol. 38, Art. no. 243, 2026, doi: 10.1016/j.jksuci.2026.243.

[12] M. Firat, "Predicting student disengagement: Harnessing visual cues for intelligent tutoring systems," *London J. Soc. Sci.*, vol. 6, p. 107, 2023, doi: 10.31039/ljss.2023.6.107.

[13] A. S. Pillai, "Student engagement detection in classrooms through computer vision and deep learning: A novel approach using YOLOv4," SSRN preprint, 2021, doi: 10.2139/ssrn.5187374.

[14] Q. Liu, X. Jiang, and R. Jiang, "Classroom behavior recognition using computer vision: A systematic review," *Sensors*, vol. 25, no. 2, Art. no. 373, 2025, doi: 10.3390/s25020373.

[15] T. Lowe and Y. El Hakim, Eds., *A Handbook for Student Engagement in Higher Education: Theory into Practice*. London, U.K.: Routledge, 2020, doi: 10.4324/9780429023033.

[16] P. Chapman *et al.*, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. SPSS Inc., 2000.

[17] F. Yang, "SCB-dataset: A dataset for detecting student classroom behavior," arXiv:2304.02488, 2023.

[18] C. Lugaresi *et al.*, "MediaPipe: A framework for building perception pipelines," arXiv:1906.08172, 2019.

[19] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs," arXiv:1907.06724, 2019.

[20] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," arXiv:2006.10204, 2020.

[21] T. Soukupová and J. Čech, "Real-time eye blink detection using facial landmarks," in *Proc. 21st Comput. Vis. Winter Workshop*, 2016.

[22] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 139, 2021, pp. 10096–10106.

[23] Danyaro, K. U., Nasser, M., Zakari, A., Abdullahi, S., Khanzada, A., Yakubu, M. M., & Shoaib, S. (2025). LLM-based code generation: a systematic literature review with technical and demographic insights. IEEE Access, 13, 194915-194939.

[24] Nasser, M. (2024). Personalized learning through AI: enhancing student engagement and teacher effectiveness. International Journal of Teaching, Learning and Education, 3(6), 23-26.

[25] Nasser, M., Yusof, U. K., & Salim, N. (2023). Deep learning based methods for molecular similarity searching: a systematic review. Processes, 11(5), 1340.

## Appendix A: Supplementary Tables

*Table A1 Per-class F1 of every approach on the test set*

| Approach | Engaged | Partially | Disengaged | Macro |
|---|---|---|---|---|
| Rule baseline | .34 | .59 | .38 | .43 |
| Decision Tree | .63 | .82 | .36 | .60 |
| MLP | .64 | .80 | .37 | .60 |
| EfficientNetV2-B0 | .65 | .55 | .77 | .66 |
| **Fusion** | **.817** | **.839** | **.731** | **.796** |

Partially: Partially Engaged. Bold: main model.

*Table A2 Confusion matrices of the cue models (rows: actual E/P/D)*

| Approach | Pred. E | Pred. P | Pred. D |
|---|---|---|---|
| Rules, E | 127 | 237 | 208 |
| Rules, P | 0 | 327 | 0 |
| Rules, D | 58 | 223 | 148 |
| DT, E | 354 | 35 | 183 |
| DT, P | 3 | 314 | 10 |
| DT, D | 203 | 91 | 135 |
| MLP, E | 368 | 32 | 172 |
| MLP, P | 3 | 297 | 27 |
| MLP, D | 199 | 89 | 141 |

*Table A3 Fusion weight on the validation set (670 students)*

| w | Acc. | F1 | w | Acc. | F1 |
|---|---|---|---|---|---|
| 0.0 | .628 | .623 | 0.6 | .763 | .761 |
| 0.1 | .737 | .722 | 0.7 | .739 | .737 |
| 0.2 | .785 | .777 | 0.8 | .725 | .723 |
| 0.3 | .809 | .803 | 0.9 | .715 | .712 |
| **0.4** | **.819** | **.817** | **1.0** | **.704** | **.701** |
| 0.5 | .793 | .792 | | | |

Bold: selected weight.

*Table A4 Cue measurability and alert timing*

| Item | Value |
|---|---|
| Candidate student crops | 15,932 |
| Eye gaze / openness / yawning measurable | 24% each |
| Head pose measurable | 74% |
| Body posture measurable | 68% |
| Alert decision, mean (p95, max) | 3.4 ms (4.1, 5.8) |
| Dashboard delivery, mean (p95, max) | 3.3 ms (4.2, 6.0) |