

Seeing the Room, Not the Student: A Privacy-Preserving Computer Vision Prototype for Anonymous Aggregate Classroom Visual-Engagement Trend Monitoring from Face-Visibility and Head-Orientation Cues

Hamzah Asyrani bin Sulaiman ^{1,*}, Nazreen bin Abdullasim ¹,
Mohamad Lutfi bin Dolhalit ¹, Iznora Aini Zolkifly ²

¹ Department of Interactive Media, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

² College of Computing and Informatics, Universiti Tenaga Nasional, Selangor, Malaysia

ABSTRACT

Camera-based classroom analytics can preserve a record of how a room behaved during a lesson, but conventionally built systems store face images that remain open to re-identification, name individual learners, and report weak visual cues as attention scores, the signal cannot support. The gap addressed here is an engineering one: neither the reported classroom-vision prototype nor the verification restricts storage to aggregate numbers, and both verify this restriction below the application layer. This paper presents the design, implementation and pilot evaluation of a prototype built to that constraint. Webcam frames are read in real time, passed through face detection and coarse head-orientation estimation, and released from memory; only numerical aggregate records reach disk, and a Streamlit dashboard renders the resulting trends. Three processing modules were compared under identical logging conditions: a Haar Cascade baseline, an OpenCV deep neural network (DNN) face detector, and a MediaPipe head-orientation module. A controlled pilot covering six predefined conditions retained 100 aggregate orientation records; the intended dominant category was recovered in 6 of 6 sessions (100%), with 29% front-facing, 21% looking left, 20% looking right, 17% looking down and 13% looking up. Of 23 DNN records, 22 carried a detection with a mean confidence of 0.890 to 1.000, all clearing the 0.60 acceptance threshold by at least 0.29, against a 4.3% zero-detection rate. A three-layer system-level audit recorded no write outside the aggregate CSV logs. The findings indicate that a research-safe aggregate visual analytics pipeline can be engineered, and its privacy boundary verified, before classroom-level deployment.

Keywords: Classroom analytics, Computer vision, Educational technology, Head-orientation estimation, Privacy-preserving AI, Data minimization.

*Corresponding author

Peer review under the responsibility of University of Baghdad.

<https://doi.org/10.31026/j.eng.2026.10.01>



This is an open access article under the CC BY 4 license (<http://creativecommons.org/licenses/by/4.0/>).

Article received: 21/07/2026

Article revised: 24/08/2026

Article accepted: 27/08/2026

Article published: 01/10/2026



1. INTRODUCTION

Student engagement is widely described as a multidimensional construct that spans behavioral, emotional and cognitive dimensions (**Fredricks et al., 2004**). Because these dimensions are not directly observable, teachers and researchers have long relied on indirect signals such as participation, questioning, body posture, or the general “feel” of a room, in order to judge whether a class is following a lesson. Learning analytics has shown that educational data, when interpreted with care, can help institutions and teachers reflect on the learning process and identify moments that deserve a second look (**Siemens and Baker, 2012; Ferguson, 2012**). In a physical classroom, however, many of the most useful signals remain difficult to capture at scale. A lecturer may sense that a room has become visually disengaged during a long explanation, yet that impression is informal, subjective and difficult to review once the session has ended.

Computer vision offers one way of turning such impressions into reviewable data. Modern detectors can locate faces and estimate head pose in real time on commodity hardware, which makes a webcam-based classroom sensor technically feasible. The same capability, however, raises sensitive engineering questions. A system that records faces, recognizes individuals, or labels a particular student as “not paying attention” may drift from a teaching-support tool into a surveillance instrument. Prior work on learning-analytics ethics has repeatedly stressed transparency, informed consent, data minimization and the careful interpretation of student data (**Slade and Prinsloo, 2013; Pardo and Siemens, 2014**). Engineers tend to underestimate a second point from the privacy literature. Stripping names out of a dataset does not make it safe, because behavioral and contextual attributes can still single a person out (**Sweeney, 2002; Dwork, 2006**). We treated both bodies of work as design input rather than as background reading.

Our prototype is therefore conservative by construction. Video is held in memory for as long as a detector needs it and no longer, and the only thing that outlives a frame is a row of aggregate numbers. No individual student record is created. We refer to the system as AI Vision Classroom Engagement, and we remain careful about what it claims to measure. The goal is not to establish that a particular student is attentive. Instead, the system produces a classroom-level visual proxy, namely how many faces are visible and how their coarse head orientation is distributed over time, which may later help researchers examine whether such broad patterns can support teaching reflection.

The contribution of this work is therefore primarily engineering and methodological rather than empirical, and it can be stated as five points.

- A privacy-first logging design in which no frame is written to disk at any stage and only aggregate numeric records survive a session.
- A comparison of a classical and a deep-learning face detector, run under identical logging conditions, so that the two can be read against one another rather than against published benchmarks collected elsewhere.
- A system-level verification procedure that tests the storage rule against file-system activity and a before-and-after directory manifest, instead of trusting the contents of the output folder.
- An interactive dashboard that keeps the analyst inside the claims the data can carry, together with an interpretation protocol and a set of evaluation metrics that a larger study can adopt without moving the privacy boundary.



- A controlled head-orientation pilot that exercises the whole pipeline end to end across six predefined conditions, including a mixed-orientation condition designed to break a naive categorizer.

Taken together, these five elements are what allow the privacy position of the system to be stated as a property of the code and of the file system, rather than as an assurance offered in a paper. It would be easy to read an aggregate-only design as a watered-down monitoring system. We think that reading is wrong, because the two answer different questions. A monitoring system answers questions about people: who was present, who looked away, and who seemed to switch off. The record it accumulates along the way is exactly what makes classroom cameras contentious. Ask instead how the room as a whole behaved, and no such record is needed. Lecturers reflecting on the rhythm of their own teaching usually want the room-level answer anyway, and it carries a far lighter ethical and legal load. We treated that as an opportunity rather than a compromise.

2. RELATED WORK

2.1 Engagement and Learning Analytics

There is no single event that counts as engagement. **(Fredricks et al., 2004)** treat it as a broad construct running across behavioral, emotional and cognitive dimensions, each surfacing differently and each calling for its own instrument. Later conceptual work has reinforced the point and added that what one ends up measuring depends heavily on context and on the grain of the analysis **(Appleton et al., 2008; Kahu, 2013; Sinatra et al., 2015)**. This distinction matters for any vision-based system, because a face that is visible to a camera is not the same thing as a mind that is cognitively engaged. Learning analytics research has consistently made a parallel argument at the level of data: educational data are valuable only when they are tied to sound pedagogical reasoning, and they can mislead when they are treated as ends in themselves **(Ferguson, 2012; Siemens and Baker, 2012; Gašević et al., 2015; Bond et al., 2020)**. **(Gašević et al., 2015)** put this bluntly in reminding the community that learning analytics are, ultimately, about learning.

Vision-based indicators therefore require careful wording. A body of work on automatic recognition of affect and engagement has shown that facial signals can carry useful information in controlled settings, for example when distinguishing boredom, confusion and frustration during computer-based learning **(Baker et al., 2010; Calvo and D'Mello, 2010; D'Mello and Graesser, 2012; Whitehill et al., 2014)**. These studies are valuable, but they also illustrate how quickly a weak signal can be over-interpreted once it is packaged as an automatic “engagement score.” In the present work we take the opposite starting point: face visibility and head orientation are treated as weak aggregate cues about a room, never as labels for the internal state of an individual learner.

More recent classroom work keeps that caution while improving the modelling. **(Goldberg et al., 2021)** predict visible engagement from classroom video and are explicit that the target is what an observer can see rather than what a learner feels. **(Sümer et al., 2023)** extend the approach to multimodal facial video and report the disagreement between human raters as part of the result rather than burying it in a preprocessing step. A systematic review of the area finds that engagement is still defined differently from one paper to the next, which is one reason why results rarely transfer between studies **(Karimah and Hasegawa, 2022)**. Both of those systems retain video and report per learner, and it is exactly that pair of properties which the present design removes.



Multimodal learning analytics broadens this perspective by combining sensor streams from physical learning settings, but its literature also distinguishes directly observable behavior from latent constructs that still require theory and human interpretation (**Blikstein, 2013; Ochoa and Worsley, 2016; Di Mitri et al., 2018**). Section 4 fixes the claim boundary of our prototype on the same basis.

2.2 Face Detection and Head-Orientation Cues

None of the detection components used here are new. Early neural-network face detectors, and the surveys that reviewed them, reported steady gains but also a set of failure modes that persisted throughout: pose, scale, illumination, and cluttered backgrounds (**Rowley et al., 1998; Yang et al., 2002**). Real-time detection became practical with a boosted cascade of simple features (**Viola and Jones, 2001**), which we retain as a baseline for the narrow reason that its firing behavior can be inspected. The associated Haar-like feature set was widened soon after (**Lienhart and Maydt, 2002**). Hand-designed descriptors, the histogram of oriented gradients in particular, stayed in use until deep learning displaced them (**Dalal and Triggs, 2005**). Faster R-CNN, SSD, RetinaNet and the YOLO family then reset accuracy and throughput expectations for object detection generally, and production detectors today mostly descend from that lineage (**Liu et al., 2016; Redmon et al., 2016; Lin et al., 2017; Ren et al., 2017**).

Work aimed specifically at faces has been more targeted. Cascaded convolutional pipelines such as MTCNN, and regression-tree landmark fitting, absorb pose, illumination and alignment variation at low cost (**Kazemi and Sullivan, 2014; Zhang et al., 2016**). Head-pose estimation carries an even longer record, normally posed as the recovery of yaw, pitch and roll; convolutional methods later pushed it into unconstrained imagery and, in some variants, dropped the key point stage altogether (**Murphy-Chutorian and Trivedi, 2009; Patacchiola and Cangelosi, 2017; Ruiz et al., 2018**). OpenFace shows what becomes possible once landmarks, pose and facial action units are analyzed together (**Baltrušaitis et al., 2018**); MediaPipe offers a modular cross-platform framework that suits rapid prototyping of real-time vision tasks (**Lugaresi et al., 2019**). We use these components but ask far less of them. No Euler angle is reported, and no behavioral inference is drawn from one. Pose cues collapsed into a handful of course, readable orientation labels, and those labels are what we aggregate.

Detection has moved on again since the components used here were released. Single-stage detectors with dense multi-level supervision now recover small and heavily occluded faces that the models in this prototype would miss (**Deng et al., 2020**), and head-pose estimation has been reformulated to avoid the discontinuities that Euler-angle regression introduces (**Hempel et al., 2022**). We stay with the older components deliberately, because the contribution of this work is the storage boundary rather than detection accuracy, and because small models run on the hardware a classroom already owns. One recent result does bear directly on the design. Face datasets, and the models trained on them, carry measurable demographic imbalance (**Kärkkäinen and Joo, 2021**). A system that attaches a label to an individual inherits that imbalance as unfairness towards a particular student; an aggregate count inherits it as a bias in a room-level number, which is still a problem, but a smaller and a more tractable one.

2.3 Privacy as a Design Constraint

One argument runs through most of the learning-analytics ethics literature: privacy has to be built in during design, not appended as a compliance statement once the system works



(Slade and Prinsloo, 2013; Pardo and Siemens, 2014; Rubel and Jones, 2016). For (Slade and Prinsloo, 2013), student data is an ethical and relational question before it is a technical one. (Pardo and Siemens, 2014) restate the same concerns as privacy principles specific to learning analytics, and (Rubel and Jones, 2016) approach them from information ethics. Work since then has been more operational: data-protection frameworks, trust checklists, survey evidence on what students say they want, and consent mechanisms (Cormack, 2016; Drachsler and Greller, 2016; Ifenthaler and Schumacher, 2016; Jones, 2019). The formal reason that removing identifiers is not, by itself, protective comes from k-anonymity and differential privacy (Sweeney, 2002; Dwork, 2006). Classroom cameras warrant more caution than most sources, since a single frame carries faces, seating positions and social behavior together. Our response is the design set out in Sections 3 and 4: no identity is learned, no image is kept, and the system's interpretation is constrained in advance.

The empirical side of that literature has grown since. A systematic review of ethics in learning analytics finds that most published work still argues from principle rather than from evidence about students themselves (Cerratto Pargman and McGrath, 2021), while survey work shows that willingness to consent is not uniform across a student body, so that an opt-in mechanism can introduce a selection bias of its own (Li et al., 2022). On the technical side, research on visual privacy in camera-based sensing has concentrated on obfuscating video after capture (Climent-Pérez and Flórez-Revuelta, 2021). Our position sits one step earlier in the chain, for a simple reason: a frame that is never written cannot be obfuscated incorrectly.

2.4 Positioning Against Recent Classroom-Vision Systems

Because most of the foundational work cited above predates the current generation of classroom-vision studies, it is worth setting the prototype directly against systems published between 2020 and 2026. **Table 1** does that on the dimensions that decide whether a system can be put into a teaching room.

Table 1. Comparison of recent classroom-vision and visual-sensing systems with the prototype reported here.

Study	Visual signal used	What is retained on disk	Reporting unit	Claim attached to the output
(Whitehill et al., 2014)	Facial expression	Labelled video and face crops	Individual learner	Engagement score
(Goldberg et al., 2021)	Face and head pose from classroom video	Annotated classroom video	Individual learner	Visible engagement rating
(Sümer et al., 2023)	Multimodal facial video	Video plus per-learner features	Individual learner	Engagement and attention estimate
(Karimah and Hasegawa, 2022)	Varied across the surveyed systems	Dataset-dependent, usually video	Mostly individual learner	Engagement estimate, inconsistently defined
(Climent-Pérez and Flórez-Revuelta, 2021)	RGB video of living spaces	Obfuscated or filtered video	Individual occupant	Activity and event detection
This work	Face visibility and coarse head orientation	Numeric aggregate CSV records only	Whole room, per session	Visual-trend proxy; no attention, emotion or comprehension claim

The visual signal it uses, what survives to storage, the unit at which it reports, and the strength of the claim attached to its output. Detection accuracy is deliberately not one of those dimensions.

The systems in the first rows of the table are more accurate than ours and will remain so, because they were built for accuracy and rest on larger models and labelled data that this study does not have.

The distinction the table is meant to expose lies elsewhere. Every one of those systems retains material that can be traced back to a person, whether as video held for annotation or as features computed per learner, and every one reports at a level that names an individual. The prototype described here does neither, and it cannot be made to do either without rewriting the logger. That is a narrower capability with a different risk profile, and this paper argues that the narrower capability is the one a lecturer reflecting on the rhythm of a lesson actually needs. Reviews from the infrastructure side reach a related conclusion about volume rather than content, namely that the sheer quantity of sensor data an installation generates is itself part of its risk surface (Wheeb and Khan, 2024; Barbhuyan et al., 2026).

3. SYSTEM DESIGN

3.1 Design Principle and Processing Pipeline

The prototype follows a single rule that can be checked by inspecting the output directory: frames may be processed, but only numbers may be saved. **Fig. 1** traces the resulting flow. Each webcam frame is read into a buffer and handed to a detector, which returns a per-frame face count and, in the case of the pose module, one coarse orientation category.

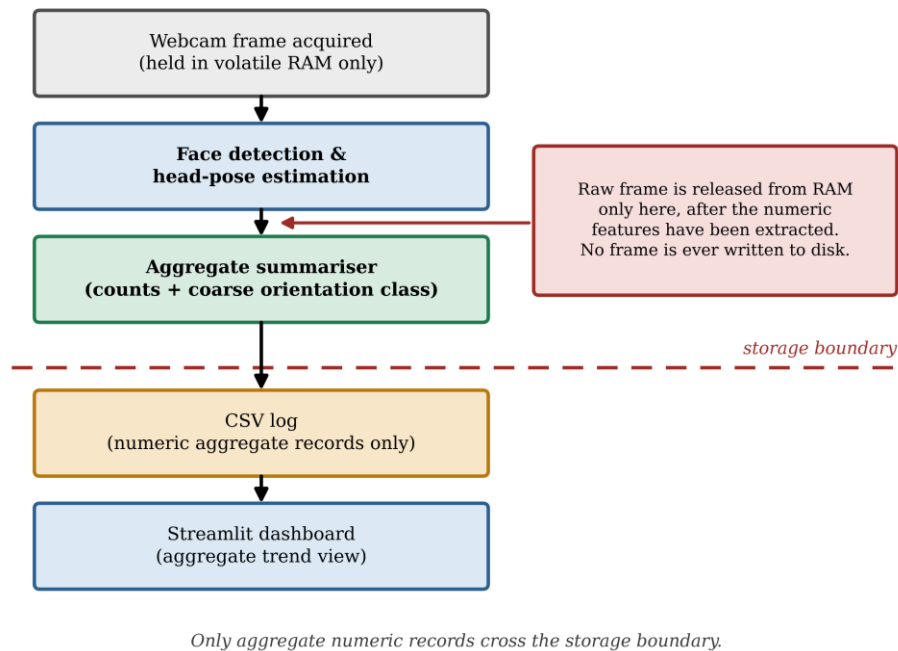


Figure 1. Privacy-preserving visual-engagement trend-monitoring pipeline.

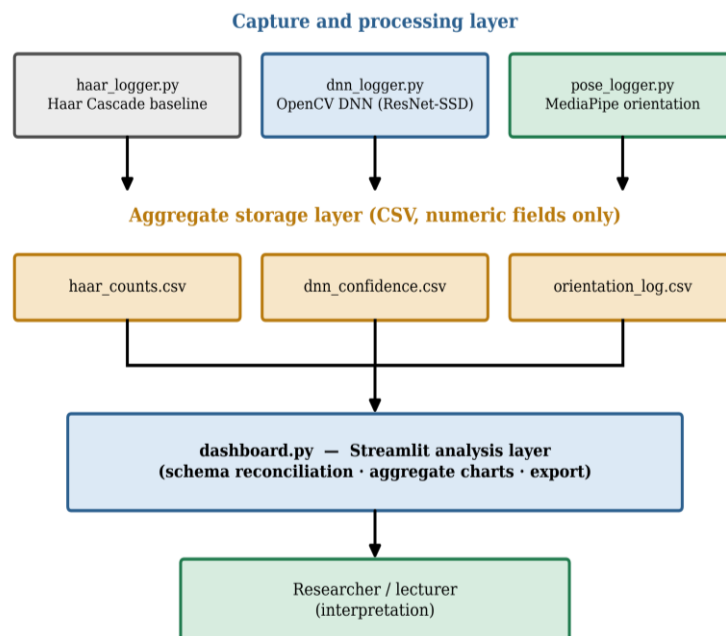
A frame is held in volatile memory for the duration of detection and is released only after the numeric features have been extracted, which is the point marked on the right of the diagram. The dashed line marks the storage boundary, and the only artefacts that cross it are aggregate numbers.

The logger summarizes that output immediately and appends it to a CSV file, which the dashboard reads afterwards. The raw video frame is never written to disk; it is released as soon as the numeric summary has been extracted. We regard this as the load-bearing decision in the design, since re-identification requires a stored artefact and the pipeline produces none. A rule of this kind, however, is only as strict as the evidence that it holds. Inspecting the output directory shows what remains at the end of a session, not what the process wrote and then removed, and not what it may have written somewhere else; on its own it is weak evidence for either privacy or correctness. Section 3.6 therefore describes the system-level checks we added so that the rule can be tested rather than asserted.

Three modules sit behind the pipeline. A Haar Cascade baseline supplies a fast reference detector whose decisions can be traced. An OpenCV DNN detector, built on a ResNet-SSD Caffe face model, records a face count together with a detector-confidence value for every frame. A MediaPipe Face Landmarker module estimates head pose and maps it onto five coarse categories: front-facing, looking left, looking right, looking down and looking up. We kept all three rather than settling on one for two reasons. The classical and the deep detector can then be compared under identical logging conditions, and orientation estimation stays separate from raw face counting, so a fault in one does not contaminate the other.

3.2 File-level Architecture and Module Separation

Fig. 2 gives the file-level view. Each capture module runs as an independent logger writing to its own CSV output, so that `haar_counts.csv`, `dnn_confidence.csv` and `orientation_log.csv` are never shared between modules. The separation costs little and returns three things. The classical baseline survives untouched as a reference, no later run can silently overwrite an earlier one, and comparing detectors reduces to opening two files side by side. Records are appended with timestamps, which means a session can be replayed as an ordered sequence without going anywhere near the original video.



No module reads another module's file, and no module writes an image.

Figure 2. File-level prototype architecture.

Each capture module owns one aggregate CSV file; the Streamlit dashboard reads those logs, reconciles their schemas and renders them for a human analyst.

Above those files sits the dashboard, which is where a person actually looks at the data. It loads each CSV, reconciles column names when the logger schemas disagree, and shows aggregate summaries and nothing else. We arranged the layers this way so that a stricter evaluation could be bolted on later without touching the privacy position. A future pilot might, for instance, tag each session with its lighting, camera distance, participant count and seating layout. Labels of that kind describe a test environment rather than a student, so the evaluation gets stronger while the identity-free policy stays exactly as it is.

3.3 The Dashboard as a Human-Facing Analysis Layer

Nothing in the dashboard decides anything. It is an aid to analysis, meant to let a researcher or lecturer follow a trend, spot the moments where detection dropped to zero, set one detector against another, and export summaries to work on elsewhere. **Fig. 3** sketches the layout. Controls live in a sidebar: active detector, session filter, export. A row of summary cards carries the headline figures, namely how many frames were logged, what fraction of them saw nothing, and the mean detector confidence. Two chart panels then plot face visibility across the session and the orientation mix over it. At the bottom is a table of the aggregate records themselves, so any number on the screen can be walked back to the row it came from. Since numbers are all it ever holds, the dashboard can be projected, screenshotted or emailed without exposing a single student. Monitoring dashboards generally cannot say that, because live video tiles and per-student rows are the norm there. Here, the most detailed object the analyst can reach is a timestamped count, not a face.

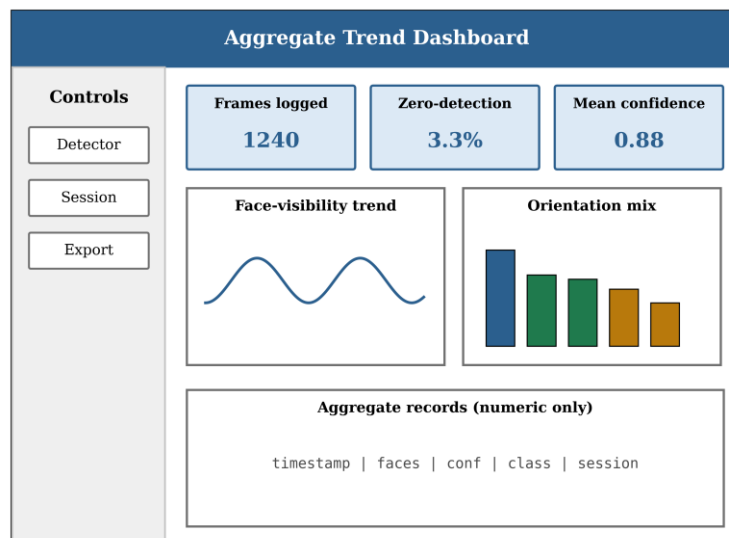


Figure 3. Conceptual layout of the Streamlit analysis dashboard. Every widget draws on numeric aggregate records; no video frame or face crop is shown or kept.

3.4 Aggregate Data Schema

Table 2 summarizes the aggregate CSV fields retained by the current prototype. The schema is intentionally minimal, because the first research question is not whether the system can classify individuals; it is whether a classroom camera pipeline can be engineered so that only low-risk aggregate indicators survive to storage. Every field in the schema is either a



timestamp, a count, a detector metadata value, a confidence value or a coarse category. None of them, alone or in combination, is intended to identify a person.

Table 2. Aggregate CSV schema retained by the prototype (numeric and categorical fields only).

Field	Type	Example	Purpose
timestamp	ISO datetime	2026-03-11T09:14:02	Time-orders records within a session
total_faces	integer	3	Number of visible face regions in the frame
avg_confidence	float	0.89	Mean detector confidence (DNN module)
max_confidence	float	0.97	Highest single-face confidence in the frame
detector	string	opencv_dnn	Identifies the logging module
threshold	float	0.60	Confidence acceptance threshold used
dominant_class	category	front_facing	Coarse head-orientation category (pose module)
session	string	S3_looking_right	Non-identifying session label / test condition

3.5 Real-time Implementation and Reproducibility

The prototype is implemented in Python and is deliberately built from widely available, well-documented components so that the whole pipeline can be reproduced on commodity hardware without specialized accelerators. Frame capture and the Haar and DNN detectors use OpenCV; the orientation module uses the MediaPipe Face Landmarker; and the dashboard is a Streamlit application that reads the CSV logs produced by the loggers. Because each stage communicates only through plain CSV files, the capture layer and the analysis layer are fully decoupled: a logger can be replaced, re-run or moved to a different machine without touching the dashboard, and an archived CSV can be re-analyzed long after the corresponding session has ended. This file-based contract is what makes the system both reproducible and auditable, because the exact data a result depends on is always a small, human-readable file rather than an opaque binary stream.

Several constraints arise from running the pipeline in real time. On a commodity webcam the capture loop only stays responsive if frames are taken in the camera's own default mode and never buffered, and if the only thing appended to the log is the numeric summary. Sampling is consequently event-driven rather than exhaustive. Not every frame has to be logged, and the cadence can be traded against file size depending on how much temporal resolution a session need. We can afford this because the system reports trends, not per-frame ground truth. A useful side effect is that less data reaches disk at all, which quietly strengthens the boundary described in the next section.

3.6 Verifying the Privacy Boundary at System Level

The design rule stated in Section 3.1 makes a claim about process behavior, and process behavior cannot be confirmed by looking at a folder. We verified it in three layers, each sitting below the one above it, and we report what each layer returned.

The first layer is the source. Each of the four capture modules was searched for every entry point through which an image could leave the process: the OpenCV writers `cv2.imwrite`, `cv2.VideoWriter` and `cv2.imencode`, the Pillow imaging library, and the NumPy and pickle serializes. None of them occurs anywhere in the capture modules. Every file that handles the loggers open is a CSV path opened in text mode, with the modes restricted to read, write and append; no logger opens a binary handle at all. The code path that would be needed to serialize a frame is therefore absent from the running process, rather than merely unused.



The second layer is the running process. We instrumented a full capture-and-log cycle with an interpreter-level audit hook that records every file-system event the process raises, covering opens, renames, deletions and directory creation, and drove the detector with a synthetic frame so that the trace would contain nothing but the pipeline's own behavior. Over five consecutive cycles the trace held exactly five events: one append-mode open of the single aggregate CSV file per cycle, and nothing else. No other path was touched and no binary handle was opened.

The third layer is the file system. The output directory is enumerated after every session and each entry is checked by type. Across the pilot the directory held CSV files only; no image, no video and no serialized array was produced at any point. Repository policy reinforces the same boundary, since the generated logs are excluded from version control and the only binary files the project keeps are the three read-only detector model files listed in **Table 3**. We would put the emphasis on the procedure rather than on the result. A clean trace on one machine shows that this build did not write an image; it does not show that no build ever could. Any deployment should repeat all three checks on its own hardware, and a classroom deployment should add an operating-system level trace filtered on the capture process identifier, so that writes issued below the Python layer are covered as well. **Table 3** in Section 5.1 lists the audit steps alongside the hardware the sessions ran on.

4. PRIVACY BOUNDARY

4.1 What is and is not Stored

The boundary belongs to the technical design, and it can be written down exactly. Not stored: names, face images, face crops, facial embeddings, identity labels, attendance records. Nor is there any individual-level dashboard. Stored: timestamped aggregate counts, detector metadata, confidence values, non-identifying session labels, coarse orientation categories. The list matters because classroom camera data is easy to misuse once it exists at all. A face crop that exists on disk can be matched against other images and re-identified, and whether a name was ever attached to it makes little difference. Not creating the file is therefore preferable to controlling access to it.

The same boundary constrains the vocabulary available to the interface. A face detector reports the presence of a face-like region and nothing about learning; an orientation category reports an approximate direction of the head and nothing about intention or cognitive state. Dashboard labels were written to match, and terms implying attendance verification, emotion detection or comprehension measurement were excluded during implementation. We note that the shortcuts here are cheap, since mapping a down-facing head onto “distracted” is a single conditional, and that is precisely why the constraint had to be applied in the code rather than left to the analyst.

4.2 Framing the System as an Environmental Sensor

Functionally, the prototype sits closer to an environmental sensor than to a student-monitoring system. A light meter registers that illumination in a room has changed without identifying who opened the blind. Ours does the same for visible faces and coarse orientations: it registers the change and cannot attribute it. For some applications that is a real limitation. It is also the main privacy protection on offer. Describing the tool as an environmental sensor keeps lecturer and reviewer expectations where they belong, and it puts the engineering in line with the data-minimization principle that the learning-analytics



literature has argued for (Slade and Prinsloo, 2013; Pardo and Siemens, 2014; Rubel and Jones, 2016).

5. METHODOLOGY

5.1 Experimental Setup

Reproducing this pilot requires the hardware and the software stack to be stated, and, because part of the claim made in this paper concerns what the software does not do, the audit steps have to be stated as well. **Table 3** gives all three. Every session ran on a single desktop workstation with a consumer USB webcam. The machine does contain a discrete graphics card, but no stage of the pipeline uses it: the stock OpenCV build carries no CUDA backend, and the MediaPipe task runs on its CPU delegate, so every detection reported in this paper was computed on the CPU. That is the configuration we want, because the paper argues that the pipeline should run on hardware a teaching room already has rather than on hardware it has to buy.

Table 3. Hardware, software and audit configuration used for the controlled pilot.

Component	Specification used in the pilot
Processor	AMD Ryzen 5 7500F, 6 cores / 12 threads, 3.70 GHz base clock
Graphics	NVIDIA GeForce RTX 5070 Ti present in the machine but unused; all detection ran on the CPU
Memory	32 GB DDR5
Storage	1 TB NVMe solid-state drive
Camera	Logitech C922 Pro Stream USB webcam, operated in its default capture mode
Operating system	Windows 11 Pro, 64-bit, build 26200
Runtime	Python 3.14.4 (64-bit), isolated virtual environment
Face detection	OpenCV 4.13.0: Haar Cascade frontal-face baseline, and ResNet-10 SSD Caffe DNN with a 300×300 input blob, mean subtraction (104, 177, 123) and an acceptance threshold of 0.60
Head orientation	MediaPipe 0.10.35 Face Landmarker; yaw threshold $\pm 20^\circ$, pitch threshold $\pm 15^\circ$
Analysis layer	Streamlit 1.57.0 with pandas 3.0.3 and NumPy 2.4.6
Logging cadence	One aggregate record every 5 s; frames processed continuously between records
Source-level audit	Static search of the four capture modules for image-writing and serialization calls
Runtime audit	Interpreter-level file-system audit hook over a full capture-and-log cycle
Output audit	File-type enumeration of the output directory after every session

The logger does not override the capture resolution. It takes whatever mode the camera driver supplies by default, holds no frame buffer, and resizes internally for the detector. One aggregate record is written every five seconds while frames are processed continuously in between, so a log is a periodic sample of a continuous process rather than a per-frame dump, for the reasons set out in Section 3.5. The detector model files, the 0.60 confidence acceptance threshold and the $\pm 20^\circ$ yaw and $\pm 15^\circ$ pitch orientation thresholds were held constant across every session, so that differences between logs cannot be attributed to configuration drift. The audit steps in the lower part of the table are the ones described in Section 3.6.

5.2 Detector Comparison Workflow

The baseline logger uses a frontal Haar Cascade detector and writes timestamped face counts to a CSV file. To reduce the false positives that classical cascades are prone to, a stricter configuration was used, combining a larger minimum face size with an aspect-ratio filter that rejects implausibly wide or narrow candidate regions. The improved logger uses the OpenCV DNN detector with a confidence acceptance threshold of 0.60; its CSV output includes the timestamp, total face count, average and maximum confidence, the detector name, the threshold and the session metadata. Running both loggers over the same conditions makes their behavior directly comparable.

The comparison dashboard reads the Haar and DNN logs and displays detector status, face-count metrics, DNN confidence trends, zero-detection moments and the raw records. **Fig. 4** shows the confidence log the DNN module produced over a short session, together with the 0.60 acceptance threshold and the zero-detection sample that the dashboard flagged automatically. Those markers turned out to be among the more useful things in the dashboard. A run of consecutive zero-detection frames almost never means the room emptied. Far more often the camera is badly placed, the lighting has shifted, or the participant has drifted out of frame.

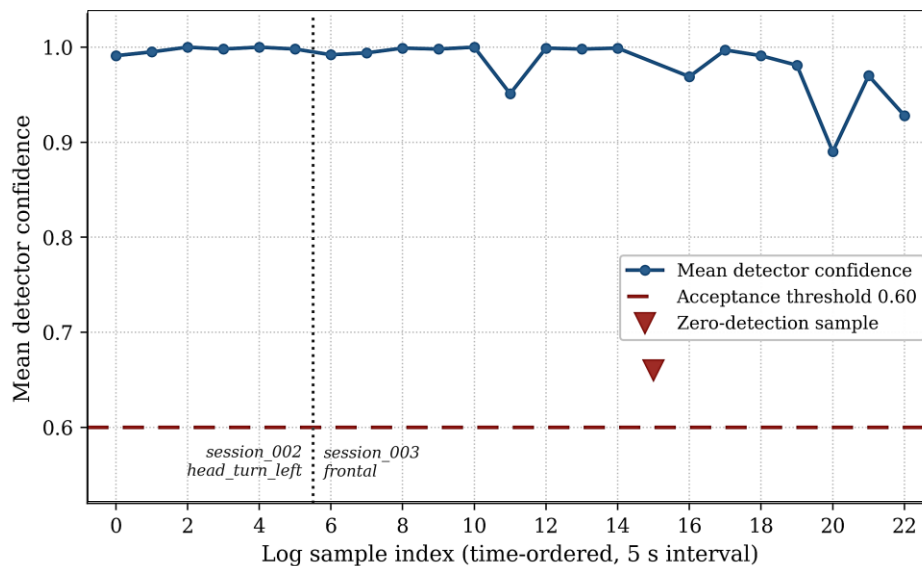


Figure 4. Mean detector confidence per aggregate record in the OpenCV DNN log, plotted directly from the prototype's own CSV output (23 time-ordered samples at five-second intervals). The dashed line is the 0.60 acceptance threshold, the marker at index 15 is the single zero-detection sample, and the dotted divider separates the two logged test conditions.

5.3 Controlled Head-Orientation Pilot

The orientation pilot used a sequence of controlled sessions, each of which asked a single participant to hold or perform a predefined head direction. The sessions covered front-facing, looking left, looking right, looking down, looking up and a mixed-orientation condition. **Fig. 5** shows the workflow; the mixed session was included deliberately to test whether the summarization step could avoid forcing a single label when no category clearly dominated. This is an important robustness check, because a real classroom rarely presents a single clean orientation.

After each session, the aggregate output was reviewed manually. This review never inspected identity or assigned a learner label; it only checked whether the dominant logged category matched the intended session and whether the session was usable for pilot interpretation. Sessions with obvious logging faults, such as long zero-detection gaps caused by the participant leaving the frame, were discarded at this stage. The review therefore functions as a lightweight sanity check on the pipeline rather than as a source of ground-truth labels.

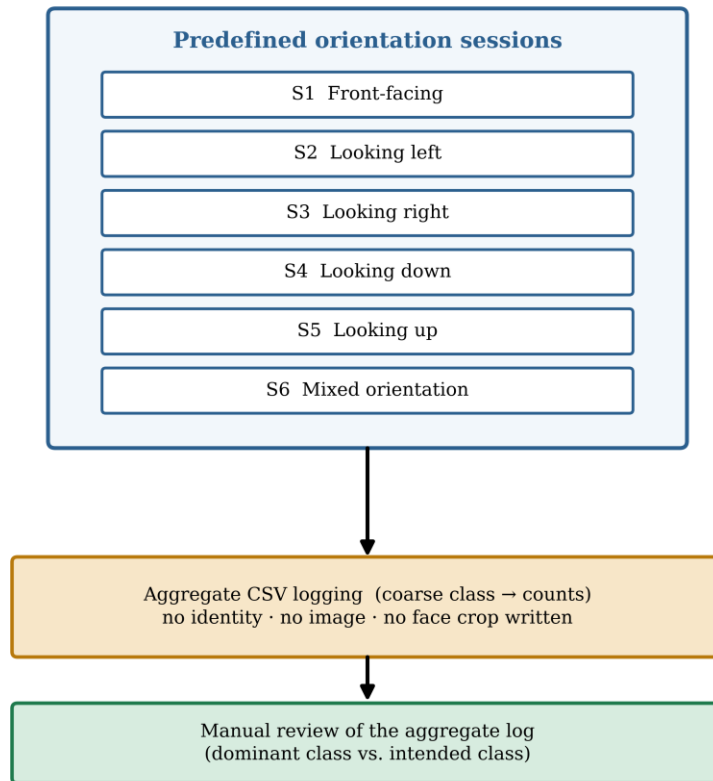


Figure 5. Controlled pilot workflow for coarse head-orientation categories. Each predefined session is logged as aggregate counts and checked in a manual review that never inspects identity or stores images.

5.4 Interpretation Protocol

The analysis follows three strict rules. First, every result must be described at the aggregate level; no statement is permitted about an individual participant. Second, detector confidence is read as detector confidence and nothing else, never as emotional certainty or as engagement strength. Third, should a visual-engagement score be introduced later, it must be labelled a proxy wherever it appears, and it must be checked against independent learning or classroom evidence before any teaching decision rests on it. These rules are intentionally conservative, because early prototypes fail most often when their claims outrun their evidence.

For a larger future pilot, the single-reviewer sanity check described above should be replaced by two independent reviewers labelling sampled frames against a shared rubric, with inter-rater agreement reported before any detector metric is computed. The current prototype also separates live monitoring from offline analysis: during live operation, the system shows detection overlays for debugging, but for research reporting, the paper uses



recreated diagrams and aggregate charts rather than screenshots containing faces. This keeps the manuscript itself consistent with the privacy boundary of the system it describes. All figures in this paper are original work by the authors. **Figs. 1 to 5** are schematic diagrams drawn for this manuscript, and **Figs. 4** and **6** are plotted directly from the prototype's own aggregate logs. No figure is reproduced or adapted from another source.

5.5 Evaluation Metrics for a Labelled Study

Although the present pilot is not a labelled accuracy study, the prototype is instrumented so that standard detection metrics can be computed once a labelled set exists. For a future evaluation, we define the face-visibility ratio (FVR) for a session as the mean number of detected faces relative to the number expected under the test condition, given in Eq. (1).

$$FVR = (1/N) \times \sum_i (f_i / f_{\text{expected}}) \quad (1)$$

where N is the number of sessions in the session set under analysis; f_i is the number of faces detected in session i , averaged over the frames logged for that session; and f_{expected} is the number of visible faces expected under the test condition, which for the single-participant pilot reported here is one. All three quantities are dimensionless counts, so FVR is itself dimensionless. A value of 1.0 means the detector recovered exactly the number of faces the condition placed in front of the camera; values below 1.0 indicate missed detections and values above 1.0 indicate spurious ones.

Detection quality against a labelled set can then be summarised with the usual precision, recall and F1 score, given in Eqs. (2) to (4), where TP, FP and FN denote true positives, false positives and false negatives at the level of detected face regions.

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

Eq. (5) defines the zero-detection rate (ZDR), which counts how often the pipeline saw no face at all. Read it as a diagnostic on camera placement and lighting, not on engagement. We report it alongside precision and recall, since a detector that fires rarely can otherwise appear accurate.

$$ZDR = (\text{frames with zero detections}) / (\text{total frames}) \quad (5)$$

6. RESULTS AND DISCUSSION

6.1 Detector Behavior

The Haar Cascade was trivial to run and served its purpose as a reference point, but it leans hard on the assumption of a frontal, well-lit face; moderate head tilt or a side profile was enough to lose detections. The OpenCV DNN detector held up better across the same variations, including side profile, tilt and partial occlusion, and its mean confidence over the logged samples sat in a 0.890 to 1.000 band. We note that these are preliminary observations rather than formal accuracy results, because a fully labelled classroom ground-truth set has not yet been collected. **Table 4** summarizes the qualitative comparison of the three modules across the properties that matter for a privacy-preserving deployment.

**Table 4.** Qualitative comparison of the three processing modules under identical logging conditions

Property	Haar Cascade	OpenCV DNN	MediaPipe pose
Primary output	Face count	Face count + confidence	Coarse orientation class
Robustness to tilt / profile	Low	Moderate-high	Moderate
Explain ability	High	Moderate	Moderate
Confidence logging	No	Yes (0.890–1.000 observed)	Indirect (class stability)
Stores identifiable data	No	No	No
Role in prototype	Baseline reference	Confidence-aware detector	Orientation categorizer

Fig. 4 repays a closer reading than a single summary statistic allows, and it is worth taking from several directions at once. The log holds 23 time-ordered samples drawn at five-second intervals across two conditions, six from a head-turn-left session and seventeen from a frontal session. Read as a distribution, the 22 samples that carried a detection occupy a narrow band from 0.890 to 1.000, with a mean of 0.984 and a median of 0.996. Every one of them clears the 0.60 acceptance threshold, and the smallest margin above it is 0.29, so in these conditions the threshold is nowhere near binding; it could be raised substantially before it began to discard data. Read as a time series, the trace is close to flat for the first fifteen samples and becomes uneven afterwards: the three lowest values in the whole log, 0.951, 0.969 and 0.890, all fall in the final third, which is also the part of the session in which a second face is present in the frame. Read as a diagnostic, the single zero-detection sample at index 15 gives a zero-detection rate of 4.3% for this log. It does not sit beside the lowest confidence values, and that is the useful observation: a zero detection is the absence of a detection rather than a weak one, and the two have different causes and different fixes. Read against the face count, finally, every sample reporting two faces carries a lower mean confidence than the single-face samples around it, which is what an average over two detections should do when the second face is smaller or further off-axis. The Haar baseline, run over comparable conditions, lost detections where the DNN module did not, which is consistent with the qualitative summary in **Table 4** and with what the detection literature reports about frontal cascades under pose variation (**Viola and Jones, 2001; Yang et al., 2002**).

Two cautions follow, and we would rather state them than let the analysis run ahead of the evidence. This is one short log from one detector on one machine, so the confidence band is a property of that configuration and not of a room; a classroom with different lighting and camera geometry would have to be measured again. And 23 samples cannot carry a claim about a rate: the 4.3% zero-detection figure rests on a single event, and it would fall to zero or double on the next session. Establishing either as a relationship rather than an observation needs the labelled multi-participant study described in Section 9.

6.2 Orientation Pilot Outcome

The controlled head-orientation pilot produced 100 aggregate records after filtering. The overall category distribution was 29 front-facing, 21 looking left, 20 looking right, 17 looking down and 13 looking up, as shown in **Fig. 6** and tabulated in **Table 5**. This distribution is not intended to represent how a real classroom behaves. It serves as a workflow-verification artefact, confirming that the logging, categorization, export and review steps behave coherently across a set of predefined conditions. That each intended orientation produced a plausible dominant category, and that the mixed session did not collapse to a single spurious label, is the outcome of interest.

Fig. 6 also carries more than its headline distribution. Taken as proportions, front-facing accounts for 29% of the 100 retained records and the four off-axis categories for the remaining 71%, split into 41% horizontal (21% left and 20% right) and 30% vertical (17% down and 13% up). The horizontal split is very nearly symmetric, one record apart, which is the result a workflow check is supposed to produce: a categorizer leaning to one side would appear here as an imbalance between left and right, and it does not. The vertical pair is less even, with looking up returning four fewer records than looking down. That is consistent with how the sessions were run, since an upward head takes the eyes off the screen and the participant tended to hold the pose for shorter spells, but it is also where a coarse categorizer is most fragile, because pitch and the vertical placement of the camera are confounded whenever the camera sits below eye level.

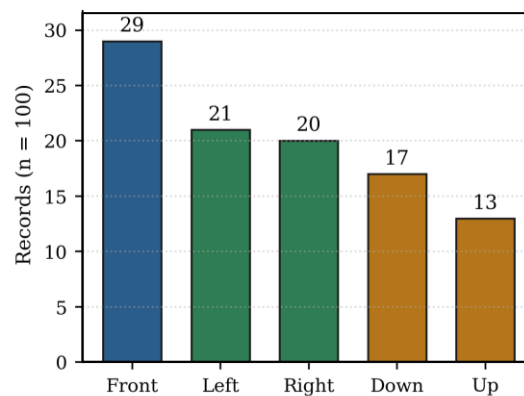


Figure 6. Head-orientation categories aggregated over the controlled pilot (n = 100 records). The distribution verifies workflow coherence and is not a sample of real classroom behavior.

The same plot answers three different questions depending on what is asked of it. Read against the intended conditions, all six sessions recovered their intended dominant category, and the mixed-orientation session spread its records across categories instead of collapsing onto one. Read as a claim about classrooms, it says nothing whatsoever, and we want to be explicit about that: the shares reflect how long a single participant was asked to hold each pose, not how a room behaves during a lesson. Read as an instrument check, though, it is informative, because the three failure modes it was designed to expose, a categorizer stuck on the dominant class, a systematic left-right bias, and a mixed session forced into a single label, would each have been visible in this plot, and none of them appeared.

Table 5. Counts and percentage shares per orientation category, controlled pilot, n = 100.

Orientation category	Records	Share (%)
Front-facing	29	29
Looking left	21	21
Looking right	20	20
Looking down	17	17
Looking up	13	13
Total	100	100



6.3 Interpretation and Methodological Value

The contribution of the pilot is methodological rather than behavioral. It establishes that the chain holds end to end, from live camera input through structured CSV logs and dashboard summaries to a manual review sheet, without an identifiable image being stored at any point along it. A secondary observation is that reflection on visual-trend data remains possible using vocabulary that is safer for educational research than the attention and emotion labels comparable systems adopt.

The results still need a conservative reading, for three specific reasons. A visible face is not an attentive one, since students look straight at a camera while thinking about something unrelated. A downward head is not a failure to understand, and in a lecture, it usually indicates note-taking. Detector confidence is confidence that a face region exists, nothing more. Holding those three distinctions open is what separates responsible classroom analytics from automated judgement of students. If we had to name the defensible outcomes of this study, they would be the working privacy-preserving pipeline and the controlled pilot workflow. The detector observations are context; they stay observations until a labelled, multi-participant study says otherwise.

There is a wider lesson here for anyone engineering classroom analytics. Institutions want a confident automated label, and the reason is not mysterious: one number fits on a slide and can be acted on. When the signal underneath is weak, a responsible system has to push back. Ours does that by reporting only what it saw, namely face-visibility counts, detector confidence and coarse orientation categories, and by insisting that any claim about engagement be carried by evidence from outside the camera pipeline. The same design remains useful to lecturers, because it can offer an aggregate reflective summary after class (for example, that face visibility dropped during a long explanation segment, or that many faces were not visible because of poor camera placement) without creating a record of an individual student.

7. LIMITATIONS AND THREATS TO VALIDITY

This study has four principal limitations. First, the pilot is small and controlled, so it cannot support any claim about real classroom accuracy. Second, the prototype has not yet been tested with a large group of students under varied lighting, seating and occlusion conditions. Third, no formal precision, recall, false-positive or false-negative rates are reported, because a labelled ground-truth dataset has not yet been prepared; the metrics defined in Section 5.5 are provided precisely so that such an evaluation can be run consistently later. Fourth, the system currently uses coarse head-orientation categories and does not model temporal behavior across a full teaching session.

Table 6 organizes the threats to validity by type. Threats to internal validity include camera placement, background variation, lighting, distance and apparent face size, all of which can change detector counts independently of any change in the room. Threats to construct validity are the most consequential, because face visibility and head orientation are only indirect cues; they cannot replace student feedback, assessment evidence, teacher observation or learning-outcome measures. Threats to external validity arise because the pilot was not conducted in a real classroom with diverse students and natural movement. Recognizing these threats does not weaken the aim of the prototype; it defines the boundary of the present contribution, which is best viewed as a working, privacy-preserving pipeline and a preparation step toward a stronger evaluation study.

**Table 6.** Threats to validity and the corresponding mitigation planned for future work.

Validity type	Main threat	Planned mitigation
Internal	Lighting, distance, camera placement bias counts	Log environment condition labels per session
Construct	Visibility/orientation are not engagement	Validate proxy against independent evidence
External	Single-participant, non-classroom pilot	Multi-participant classroom study with consent
Conclusion	No labelled ground truth yet	Two-rater labelling; report Eq. (2)–Eq. (5)

8. CONCLUSIONS

We have described an early-stage computer vision prototype that monitors classroom face-visibility and head-orientation trends anonymously and in aggregate. The system combines a Haar Cascade baseline, an OpenCV DNN detector, a MediaPipe head-orientation categorizer, CSV logging, an interactive dashboard and a manual pilot review; identity recognition, attendance tracking, face-image storage and claims about comprehension are absent by construction rather than by policy. Across the controlled pilot, 100 aggregate orientation records survived filtering, the intended dominant category was recovered in all six sessions, and 22 of 23 DNN records carried a detection with a mean confidence of 0.984 against a single zero-detection sample (4.3%); the Haar baseline lost detections under pose variation where the DNN module did not. Throughput averaged about 31 frames per second on the CPU, peak resident memory stayed below 175 MB per module, and what reaches storage is a plain CSV file of a few kilobytes per session, with the storage audit recording no write outside those logs. Two bottlenecks bound these results: the fixed 300×300 detector input, which sets the practical ceiling on how far from the camera a seat can contribute regardless of capture resolution, and single-machine CPU inference, which does not extend to multi-camera deployments without edge devices or a deliberately enabled GPU backend. This is not a finished engagement detector and should not be read as one. It is a reproducible, privacy-aware route from live vision input to aggregate trend interpretation, and an engineering template on which a larger, labelled and ethically governed study can be built without anyone having to renegotiate the privacy boundary first.

NOMENCLATURE

Symbol	Description	Symbol	Description
AI	Artificial Intelligence	FN	False Negatives (count)
CSV	Comma-Separated Values	FP	False Positives (count)
DNN	Deep Neural Network	FVR	Face-Visibility Ratio (dimensionless)
f_{expected}	Expected number of visible faces under a test condition (dimensionless)	N	Number of sessions in a session set (dimensionless)
f_i	Number of detected faces in session i (dimensionless)	TP	True Positives (count)
F1	F1 score, harmonic mean of precision and recall (dimensionless)	ZDR	Zero-Detection Rate (dimensionless)
		Σ	Summation operator

Acknowledgements

The authors are grateful to colleagues in the Faculty of Information and Communication Technology, whose comments on the privacy framing shaped several parts of Section 4, and to the participant who sat through the controlled orientation sessions.



Credit Authorship Contribution Statement

Hamzah Asyrani bin Sulaiman: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review and editing. Nazreen bin Abdullasim: Methodology, Validation, Writing – review and editing. Mohamad Lutfi bin Dolhalit: Validation, Investigation, Writing – review and editing. Iznora Aini Zolkifly: Validation, Supervision, Writing – review and editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

- Appleton, J.J., Christenson, S.L., and Furlong, M.J., 2008. Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, 45(5), pp. 369–386. <https://doi.org/10.1002/pits.20303>
- Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., and Graesser, A.C., 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), pp. 223–241. <https://doi.org/10.1016/j.ijhcs.2009.12.003>
- Baltrušaitis, T., Zadeh, A., Lim, Y.C., and Morency, L.P., 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pp. 59–66. <https://doi.org/10.1109/FG.2018.00019>
- Barbhuyan, M.F.K., Wheeb, A.H., and Morshedur, M., 2026. Exploring the future of connectivity: A comprehensive review of wireless and optical communications. In *Data-Driven AI: A Multidisciplinary Approach – Techniques, Applications, and Insights from Multiple Domains*, 1st ed. Boca Raton: CRC Press, pp. 75–79. <https://doi.org/10.1201/9781003753377-14>
- Blikstein, P., 2013. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)*, pp. 102–106. <https://doi.org/10.1145/2460296.2460316>
- Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., and Kerres, M., 2020. Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International Journal of Educational Technology in Higher Education*, 17(1), Article 2. <https://doi.org/10.1186/s41239-019-0176-8>
- Calvo, R.A., and D'Mello, S., 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), pp. 18–37. <https://doi.org/10.1109/T-AFFC.2010.1>
- Cerratto Pargman, T., and McGrath, C., 2021. Mapping the ethics of learning analytics in higher education: A systematic literature review of empirical research. *Journal of Learning Analytics*, 8(2), pp. 123–139. <https://doi.org/10.18608/jla.2021.1>
- Climent-Pérez, P., and Flórez-Revuelta, F., 2021. Protection of visual privacy in videos acquired with RGB cameras for active and assisted living applications. *Multimedia Tools and Applications*, 80(15), pp. 23649–23664. <https://doi.org/10.1007/s11042-020-10249-1>



- Cormack, A.N., 2016. A data protection framework for learning analytics. *Journal of Learning Analytics*, 3(1), pp. 91–106. <https://doi.org/10.18608/jla.2016.31.6>
- Dalal, N., and Triggs, B., 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–893. <https://doi.org/10.1109/CVPR.2005.177>
- Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S., 2020. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5203–5212. <https://doi.org/10.1109/CVPR42600.2020.00525>
- Di Mitri, D., Schneider, J., Specht, M., and Drachsler, H., 2018. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4), pp. 338–349. <https://doi.org/10.1111/jcal.12288>
- D'Mello, S., and Graesser, A., 2012. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), pp. 145–157. <https://doi.org/10.1016/j.learninstruc.2011.10.001>
- Drachsler, H., and Greller, W., 2016. Privacy and analytics - it's a DELICATE issue. A checklist to establish trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge (LAK '16)*, pp. 89–98. <https://doi.org/10.1145/2883851.2883893>
- Dwork, C., 2006. Differential privacy. In *Automata, Languages and Programming (ICALP 2006)*, pp. 1–12. https://doi.org/10.1007/11787006_1
- Ferguson, R., 2012. Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), pp. 304–317. <https://doi.org/10.1504/IJTEL.2012.051816>
- Fredricks, J.A., Blumenfeld, P.C., and Paris, A.H., 2004. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), pp. 59–109. <https://doi.org/10.3102/00346543074001059>
- Gašević, D., Dawson, S., and Siemens, G., 2015. Let's not forget: Learning analytics are about learning. *TechTrends*, 59, pp. 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
- Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., Kasneci, E., and Trautwein, U., 2021. Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33(1), pp. 27–49. <https://doi.org/10.1007/s10648-019-09514-z>
- Hempel, T., Abdelrahman, A.A., and Al-Hamadi, A., 2022. 6D rotation representation for unconstrained head pose estimation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2496–2500. <https://doi.org/10.1109/ICIP46576.2022.9897219>
- Ifenthaler, D., and Schumacher, C., 2016. Student perceptions of privacy principles for learning analytics. *Educational Technology Research and Development*, 64(5), pp. 923–938. <https://doi.org/10.1007/s11423-016-9477-y>
- Jones, K.M.L., 2019. Learning analytics and higher education: A proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education*, 16(1), Article 24. <https://doi.org/10.1186/s41239-019-0155-0>



- Kahu, E.R., 2013. Framing student engagement in higher education. *Studies in Higher Education*, 38(5), pp. 758–773. <https://doi.org/10.1080/03075079.2011.598505>
- Karimah, S.N., and Hasegawa, S., 2022. Automatic engagement estimation in smart education/learning settings: A systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 9(1), Article 31. <https://doi.org/10.1186/s40561-022-00212-y>
- Kärkkäinen, K., and Joo, J., 2021. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1548–1558. <https://doi.org/10.1109/WACV48630.2021.00159>
- Kazemi, V., and Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1867–1874. <https://doi.org/10.1109/CVPR.2014.241>
- Li, W., Sun, K., Schaub, F., and Brooks, C., 2022. Disparities in students' propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education*, 32(3), pp. 564–608. <https://doi.org/10.1007/s40593-021-00254-2>
- Lienhart, R., and Maydt, J., 2002. An extended set of Haar-like features for rapid object detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 1-900–1-903. <https://doi.org/10.1109/ICIP.2002.1038171>
- Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollár, P., 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. <https://doi.org/10.1109/ICCV.2017.324>
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A.C., 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision - ECCV 2016*, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
- Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M., 2019. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*. <https://doi.org/10.48550/arXiv.1906.08172>
- Murphy-Chutorian, E., and Trivedi, M.M., 2009. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4), pp. 607–626. <https://doi.org/10.1109/TPAMI.2008.106>
- Ochoa, X., and Worsley, M., 2016. Editorial: Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, 3(2), pp. 213–219. <https://doi.org/10.18608/jla.2016.32.10>
- Pardo, A., and Siemens, G., 2014. Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), pp. 438–450. <https://doi.org/10.1111/bjet.12152>
- Patacchiola, M., and Cangelosi, A., 2017. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71, pp. 132–143. <https://doi.org/10.1016/j.patcog.2017.06.009>
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>



- Ren, S., He, K., Girshick, R., and Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), pp. 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- Rowley, H.A., Baluja, S., and Kanade, T., 1998. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1), pp. 23–38. <https://doi.org/10.1109/34.655647>
- Rubel, A., and Jones, K.M.L., 2016. Student privacy in learning analytics: An information ethics perspective. *The Information Society*, 32(2), pp. 143–159. <https://doi.org/10.1080/01972243.2016.1130502>
- Ruiz, N., Chong, E., and Rehg, J.M., 2018. Fine-grained head pose estimation without keypoints. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2155–2164. <https://doi.org/10.1109/CVPRW.2018.00281>
- Siemens, G., and Baker, R.S.J.d., 2012. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)*, pp. 252–254. <https://doi.org/10.1145/2330601.2330661>
- Sinatra, G.M., Heddy, B.C., and Lombardi, D., 2015. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1), pp. 1–13. <https://doi.org/10.1080/00461520.2014.1002924>
- Slade, S., and Prinsloo, P., 2013. Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), pp. 1510–1529. <https://doi.org/10.1177/0002764213479366>
- Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., and Kasneci, E., 2023. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2), pp. 1012–1027. <https://doi.org/10.1109/TAFFC.2021.3127692>
- Sweeney, L., 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), pp. 557–570. <https://doi.org/10.1142/S0218488502001648>
- Viola, P., and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, pp. I-511–I-518. <https://doi.org/10.1109/CVPR.2001.990517>
- Wheeb, A.H., and Khan, M.F., 2024. A survey of several machine learning (ML) algorithms for security solution in Internet of Things (IoT) networks. *Journal of Artificial Intelligence Research & Advances*, 12(1), pp. 1–11.
- Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., and Movellan, J.R., 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1), pp. 86–98. <https://doi.org/10.1109/TAFFC.2014.2316163>
- Yang, M.-H., Kriegman, D.J., and Ahuja, N., 2002. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), pp. 34–58. <https://doi.org/10.1109/34.982883>
- Zhang, K., Zhang, Z., Li, Z., and Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), pp. 1499–1503. <https://doi.org/10.1109/LSP.2016.2603342>

رؤية القاعة لا الطالب: نموذج أولي للرؤية الحاسوبية يحافظ على الخصوصية لمراقبة اتجاهات المشاركة البصرية الإجمالية المجهولة في الفصل الدراسي اعتماداً على إشارات ظهور الوجه واتجاه الرأس

حمزة عسراني بن سليمان^{1*}، نازرين بن عبد العليم¹، محمد لطفي بن دولحالت¹، وإزنورا عيني ذوالكفلي²

¹ قسم الوسائط التفاعلية، كلية تكنولوجيا المعلومات والاتصالات، الجامعة التقنية الماليزية ملاكا، ملاكا، ماليزيا

² كلية الحوسبة والمعلوماتية، جامعة تنغا الوطنية، سيلانغور، ماليزيا

الخلاصة

يمكن لتحليلات الفصول الدراسية القائمة على الكاميرا أن تحفظ سجلاً لسلوك القاعة أثناء الدرس، غير أن الأنظمة المبنية بالطريقة المعتادة تخزن صور الوجوه فتبقى قابلة لإعادة تحديد الهوية، وتسمى المتعلمين فردياً، وتعرض إشارات بصرية ضعيفة بوصفها درجات انتباه لا تسند لها البيانات. والفجوة التي تعالجها هذه الورقة هندسية: لا يوجد نموذج أولي منشور يقصر التخزين على الأرقام الإجمالية ويتحقق من ذلك القيد على مستوى النظام. تعرض الورقة تصميم هذا النموذج وتنفيذه وتقييمه التجريبي: تُقرأ إطارات الكاميرا في الوقت الفعلي وتُمرر بكشف الوجوه وتقدير اتجاه الرأس التقريبي ثم تُحرر من الذاكرة، ولا يصل إلى القرص سوى سجلات رقمية إجمالية تعرضها لوحة معلومات Streamlit. قورنت ثلاث وحدات معالجة تحت ظروف تسجيل متطابقة: أساس Haar Cascade، وكاشف وجوه بالشبكات العصبية العميقة (DNN)، ووحد MediaPipe لاتجاه الرأس. غطت التجربة المضبوطة ست حالات محددة مسبقاً واحتفظت بـ 100 سجل اتجاه إجمالي؛ استُرجعت الفئة المهيمنة المقصودة في 6 من 6 جلسات (100%)، وتوزعت السجلات بنسبة 29% مواجهة للأمام و 21% نظراً إلى اليسار و 20% نظراً إلى اليمين و 17% نظراً إلى الأسفل و 13% نظراً إلى الأعلى. ومن بين 23 سجلاً للكاشف العميق حمل 22 سجلاً كشفاً بثقة بين 0.890 و 1.000، متجاوزة جميعها عتبة القبول 0.60 بمقدار 0.29 على الأقل، في مقابل معدل انعدام كشف بلغ 4.3%. ولم يسجل تدقيق ثلاثي الطبقات أي عملية كتابة خارج سجلات CSV الإجمالية. تشير النتائج إلى إمكانية هندسة خط أنابيب تحليل بصري إجمالي آمن للبحث والتحقق من حدود خصوصيته قبل أي نشر على مستوى الفصول الدراسية.

الكلمات المفتاحية: التحليلات الصفية، الرؤية الحاسوبية، الذكاء الاصطناعي الحافظ للخصوصية، تقدير اتجاه الرأس، تكنولوجيا التعليم.