

Paper Type: Original Article

BrightPath AI: A Privacy-Preserving Edge-AI Framework for Early Warning of Classroom Engagement, Safety, and Student Well-Being

Mostafa Elshahat ¹ , Ahmed Sami ¹ , Mustafa Mohamed ¹ , Mahmoud Atef ¹ , Mohamed ElSayed ¹ , Ibrahim Nabil ¹ , and Ahmed Hossam ^{1,*} 

¹ Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt;
Emails: mostafa.mahmoud.elshahat1@gmail.com, ahmedsamielnaggar@gmail.com, sasa82823@gmail.com,
Mahmoudatef1404@gmail.com, abrahymmsy515@gmail.com, mohameddiavolo@gmail.com, adarwish@fci.zu.edu.eg.

Received: 15 May 2026

Revised: 02 Aug 2026

Accepted: 03 Oct 2026

Published: 07 Oct 2026

Abstract

Schools increasingly use digital tools to support learning, yet continuous observation of individual engagement and safety remains difficult in large classrooms, while continuous cloud video analytics raises substantial privacy concerns. This paper presents BrightPath AI, an edge-first educational early-warning framework that combines open-set face identity routing, classroom-behavior classification, facial-expression category detection, violence detection, a questionnaire-based well-being risk screen, and role-specific mobile reporting. Raw classroom video is processed on a local edge unit and discarded after inference; downstream services receive only derived metadata and alerts. The framework was evaluated using the project datasets and implementation logs supplied for five analytical modules. At a cosine-similarity threshold of 0.5, the face-verification module produced 0 false acceptances among 46,548 impostor comparisons and 27 false rejections among 862 genuine comparisons, corresponding to 99.94% pairwise accuracy, 96.87% recall, and an observed false-acceptance rate of 0%. The classroom-behavior model achieved 98.69% test accuracy across seven classes. The YOLOv8m facial-expression detector obtained mAP@0.50 of 0.877 and mAP@0.50:0.95 of 0.697. The MobileNet-BiLSTM violence detector achieved 90% accuracy on the reported 40-video test subset, with no false-positive violence alerts in that subset and 110.55 ms inference time per 15-frame sequence. The binary well-being screening classifier achieved 95% accuracy and 0.92 recall for the needs-support class. The scientific contribution is the integration of heterogeneous school-safety and well-being signals within a privacy-minimizing, human-in-the-loop architecture, together with explicit operational boundaries that prevent facial-expression or questionnaire outputs from being treated as autonomous clinical diagnoses or disciplinary decisions.

Keywords: Edge AI; Smart Classroom; Student Behavior Recognition; Privacy-preserving Analytics; Open-Set Face Recognition; Violence Detection; Well-Being Screening; Educational Technology.

1 | Introduction

Schools are increasingly expected to respond not only to academic performance but also to engagement, safety, and student well-being. The need is substantial: the World Health Organization reports that one in



Corresponding Author: adarwish@fci.zu.edu.eg



<https://doi.org/10.61356/j.scin.2026.3678>



Licensee SciNexuses. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

seven people aged 10–19 years experiences a mental disorder, while depression, anxiety, and behavioral disorders are leading contributors to illness and disability during adolescence [1]. At the same time, teachers in large classrooms cannot continuously observe every student, and parents have only episodic visibility into what occurs during the school day. These constraints motivate tools that can provide timely, interpretable signals without turning the classroom into a continuously recorded surveillance environment.

Computer-vision systems have already been investigated for attention monitoring, posture recognition, and classroom behavior analysis [3-7]. Privacy-preserving and edge-computing approaches are also emerging, including federated student-status monitoring and edge-deployable behavior detectors [5-7]. These developments show that local inference can reduce latency and limit the transfer of raw visual data. However, most studies remain task-specific: they typically focus on behavior recognition, posture, attendance, or privacy-preserving model training rather than integrating identity-safe event routing, classroom safety alerts, parent-facing reporting, and a separate questionnaire-based well-being screen within one operational architecture.

BrightPath AI is designed around that systems-level gap. The framework does not treat any single model output as a complete explanation of a child's state. Instead, it separates observable classroom events from higher-stakes interpretation. Visual models identify dataset-defined behaviors, facial-expression categories, and candidate safety incidents; a separate questionnaire module generates a screening risk indicator; and all outputs are routed to authorized adults for review. This separation is important because facial movements alone do not provide a scientifically sufficient basis for inferring an individual's internal emotional or psychiatric state [18].

The principal contribution of this work is therefore not a claim that every underlying model is novel. The contribution is a privacy-minimizing, modular early-warning architecture that integrates heterogeneous signals while preserving explicit decision boundaries. Specifically, the work contributes: (i) an edge-first data path in which classroom frames are processed locally and discarded after inference; (ii) an open-set identity gate that can reject unregistered faces before individualized metadata are routed; (iii) a multi-module pipeline spanning engagement, facial-expression categories, physical-violence alerts, and questionnaire-based risk screening; (iv) a parent/teacher/administrator reporting layer; and (v) a human-in-the-loop governance model in which AI outputs are evidence for review rather than autonomous clinical or disciplinary judgments. These contributions align with child-centered AI principles emphasizing privacy, safety, transparency, non-discrimination, and human oversight [2].

2 | Related Work and Research Gap

Deep-learning approaches to classroom analysis have progressed from basic attention estimation toward multi-scene posture and behavior recognition. Trabelsi et al. developed a real-time classroom attention-monitoring system based on deep learning and student-behavior recognition [3]. Liu and Yang later proposed a meta-learning-based multi-scene posture detector intended to improve robustness across classroom settings [4]. These studies demonstrate the feasibility of computer vision for classroom observation, but their primary research objective is perception performance rather than an end-to-end family-facing early-warning workflow.

Privacy has become a parallel research direction. Wu et al. proposed a federated semi-supervised system for student-status monitoring, allowing clients to train collaboratively without centrally collecting classroom face data [5]. More recently, StudentNet was introduced as an edge-deployable student-behavior detector [6], and Valdes-Ramirez et al. reported a hybrid edge model capable of processing classroom video locally while substantially reducing transmitted bandwidth [7]. A systematic review of fog and edge computing for learning analytics likewise concluded that bringing analytics closer to users can improve privacy and reduce dependence on centralized cloud storage [9]. These works establish that edge computing is not, by itself, a novel contribution; accordingly, the novelty claim in this paper is limited to the particular system integration and governance structure evaluated here.

Educational platforms also illustrate the value and risks of behavioral data outside computer vision. ClassDojo, for example, demonstrates how behavior-management data can become part of routine communication between schools and families, while also raising concerns about psychological surveillance and persuasive behavioral technologies [8]. BrightPath therefore treats parent-facing visibility as useful only when paired with data minimization, restricted purposes, and human review. The system intentionally avoids presenting facial-expression categories as mental-health diagnoses and avoids allowing a violence classifier to trigger automatic punishment.

Table 8 later provides a structured comparison with representative classroom-AI systems. The comparison is qualitative because the cited studies use different datasets, population definitions, and evaluation protocols; direct numerical ranking would therefore be methodologically inappropriate.

3 | System Architecture and Scientific Contribution

3.1 | Edge-First Data Flow

The architecture in Figure 1 separates the school-local inference boundary from the authorized application layer. Classroom IP cameras supply frames to an on-premise edge unit. Face identity, classroom-behavior, facial-expression category, and violence models execute locally. Frames are retained only long enough to support inference and are then discarded. The external application layer receives derived metadata rather than raw classroom video. Questionnaire responses follow a separate path into the well-being risk-screening service. Both pathways terminate in human review and support rather than autonomous diagnosis or discipline.

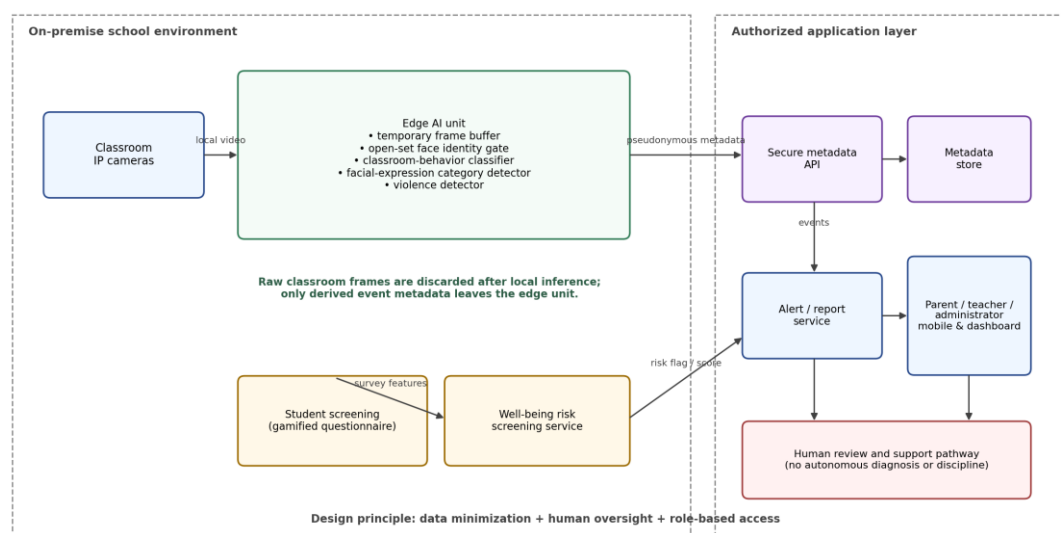


Figure 1. BrightPath edge-first architecture and data flow. Raw classroom frames remain within the on-premise inference boundary; downstream services receive derived metadata and alerts.

The architectural design reduces the attack surface associated with continuous video transfer and central storage, but it does not eliminate privacy risk. Biometric identity templates, event histories, and well-being indicators remain sensitive data. Consequently, practical deployment requires role-based access, encrypted communication, protected storage, retention limits, audit logging, and institutional authorization. The privacy claim of this paper is therefore specific: raw classroom video is not intended to leave the edge node or be persistently stored by the application layer.

3.2 | Functional Modules and Decision Boundaries

A key design decision is to constrain each model to the type of inference it can reasonably support. Table 1 summarizes these operational boundaries. This is scientifically important because a high classification score does not justify a stronger semantic claim than the training target itself.

Table 1. Functional modules, outputs, and permitted operational interpretation.

Module	Primary output	Operational use	Explicit boundary
Open-set face identity	Registered ID or Unknown	Route metadata to the correct authorized profile	Not a general-purpose identity adjudication system
Classroom behavior	One of seven observable behavior labels	Build engagement timelines and repeated-event summaries	Does not infer motivation, intent, or learning ability
Facial-expression category	Dataset-defined facial-expression class	Optional contextual cue for human review	Not a direct measure of internal emotion or mental health
Violence detection	Violence / non-violence candidate event	Generate a safety alert for staff review	No automatic disciplinary action
Well-being risk screening	Needs-support / lower-risk screening class	Prioritize follow-up by counselor or authorized adult	Screening only; not a clinical diagnosis

As Table 1 shows, BrightPath is structured as an early-warning system rather than a decision-making authority. The system's scientific value lies partly in this separation of sensing, risk indication, and human interpretation. For children, this boundary is especially important because AI errors can produce lasting consequences if they are converted directly into discipline, stigma, or clinical labels [2].

3.3 | Software Stack and Modular Deployment

The implementation is divided into three principal software layers. The machine-learning service is implemented in Python and exposed through FastAPI endpoints, with NumPy and scikit-learn supporting numerical processing and classical machine-learning components. The application backend uses .NET Core Web API for authentication, metadata routing, and service integration. Flutter provides the cross-platform mobile interface. Individual inference services are containerized so that failure or update of one analytical module does not require replacement of the entire application stack. Model weights and preprocessing logic are loaded at service start-up to reduce per-request overhead.

This modularization supports two distinct deployment properties. First, compute-intensive vision models can remain on the school edge unit, while mobile clients receive only authorized summaries and alerts. Second, modules can be independently validated, replaced, or disabled when local policy does not permit a particular form of analysis. This is particularly relevant for facial-expression inference, which should be considered an optional low-stakes contextual signal rather than a mandatory component of the system.

4 | Data and Preprocessing

The implementation uses heterogeneous data because the five analytical modules solve different tasks. Table 2 summarizes the data sources and their roles. Public repository names and links are retained where they were available in the project record. The datasets are not merged into a single multimodal cohort; each module is evaluated on its own task-specific data.

Table 2. Datasets and task-specific roles in the BrightPath prototype.

Dataset / source	Data type	Role in the system	Preprocessing retained in the project
Face enrollment/evaluation image set	Images	Open-set identity registration and verification	Face detection/alignment; embedding extraction; cosine similarity
Real-Life Violence Situations [15]	Video	Violence / non-violence sequence classification	Frame sampling, resizing, normalization, fixed-length sequences
Classroom Student Behaviors [23]	Images	Seven-class observable classroom behavior classification	Class balancing to 6,000 samples/class; embedding extraction; integer labels
Emotion Detector, Roboflow [24]	Images	Five dataset-defined facial-expression categories	Resizing; grayscale/contrast preprocessing; noise and augmentation
Children Mental Health, OSF [25]	Tabular survey data	Questionnaire-derived well-being risk screening	Feature reduction from 190 to 42 variables; composite screening features

For the classroom-behavior dataset, the seven classes were balanced by limiting each category to 6,000 samples. The reported evaluation output contains 1,500 test samples per class, or 10,500 samples in total. DeepFace/FaceNet-derived 128-dimensional representations were used as the input to the behavior MLP, followed by integer encoding of the seven class labels [12, 21]. This design is retained because it reflects the implemented prototype; however, the discussion section explicitly limits generalization claims because face-centered embeddings can carry identity information and are not an ideal representation for body-centric behavior recognition.

The facial-expression dataset was organized into five classes: Angry, Fear, Happy, Neutral, and Sad. Preprocessing included intensity normalization, controlled perturbations, and augmentation intended to reduce sensitivity to lighting and local image noise. The labels are treated exactly as dataset categories; they are not interpreted as verified internal emotional states. This distinction follows evidence showing that facial movements alone do not support one-to-one inference of emotion across people and contexts [18].

The well-being dataset originally contained 190 variables. The project reduced this set to 42 variables derived from questionnaire and contextual features, including PHQ-9-, GAD-7-, adverse-experience-, and bullying-related items. A project-defined binary Risk_Flag was generated using a threshold of 10 on the aggregated anxiety/depression screening construct. Because the target is derived from questionnaire variables, the resulting machine-learning score should be interpreted as consistency with an operational screening rule, not independent clinical validation. PHQ-9 and GAD-7 are established screening instruments [16, 17], but screening thresholds are not equivalent to diagnosis, particularly when instruments are combined or adapted.

5 | Model Development

5.1 | Open-Set Face Identity Routing

The identity module uses RetinaFace for face localization and alignment and ArcFace for discriminative face embedding [10, 11]. For a detected face, the system computes a normalized embedding and compares it with enrolled templates using cosine similarity. A similarity threshold of 0.5 was used in the reported experiments. Scores below the threshold are rejected as Unknown rather than forced to the nearest registered identity. This open-set rejection behavior is central to the privacy objective because it reduces the risk of associating observations with the wrong student profile.

Figure 2 shows the empirical score distributions used to evaluate the threshold. The distribution contains 46,548 impostor comparisons and 862 genuine comparisons. The highest impostor similarity is 0.457, below

the selected threshold, while the genuine distribution has a mean similarity of 0.929. The presence of a small genuine tail below 0.5 explains the observed false rejections.

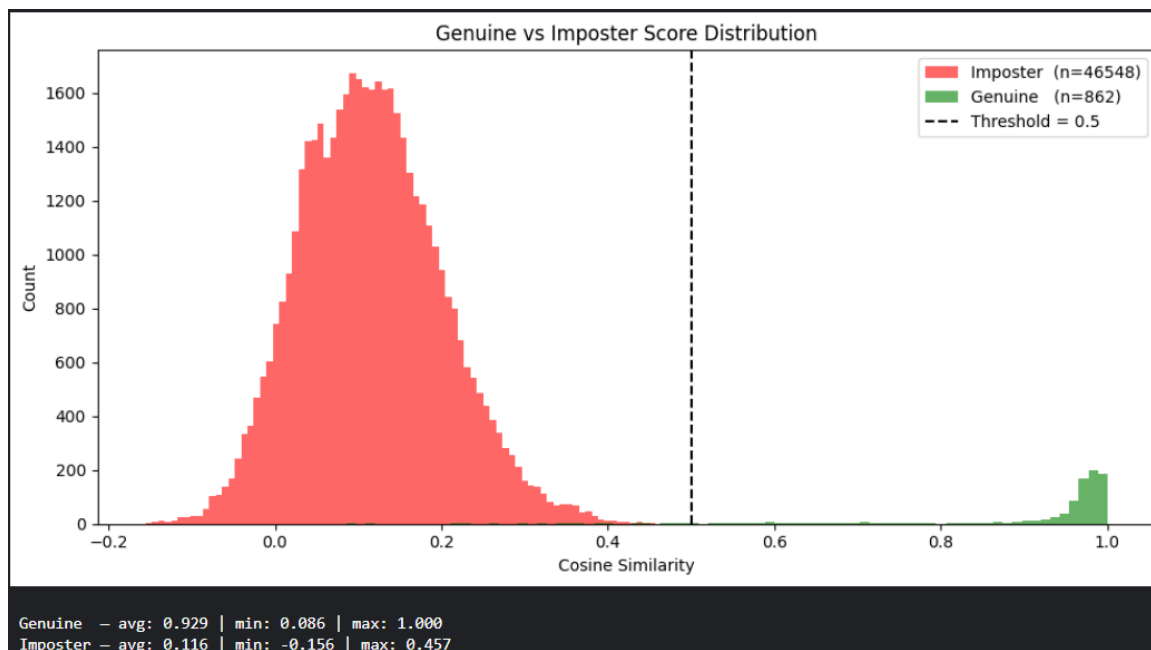


Figure 2. Genuine and imposter cosine-similarity distributions for the face-verification module at the 0.5 decision threshold.

5.2 | Classroom Behavior Classifier

The classroom-behavior module uses a multilayer perceptron with four hidden layers of 512, 256, 128, and 64 units on the FaceNet-derived representation [12]. ReLU activations are used throughout, with dropout rates between 0.3 and 0.5 to reduce overfitting. The model was trained with Adam and cross-entropy loss for up to 100 epochs, with early stopping after 10 epochs without improvement. The output classes are Turning Around, Looking Forward, Standing, Reading, Raising Hand, Writing, and Sleeping. In the system workflow, these labels are aggregated over time before being surfaced to adults; isolated frame-level predictions are not treated as evidence of chronic disengagement.

5.3 | Facial-Expression Category Detector

The facial-expression module uses YOLOv8m at 640×640 input resolution [22]. The model was trained for 80 epochs with Adam on five dataset-defined categories. The detector performs both localization and category prediction, so evaluation uses mean Average Precision (mAP) in addition to precision and recall. The classifier is intentionally described as a facial-expression category detector rather than an emotion-state estimator because the dataset labels do not establish the student's internal psychological state [18].

5.4 | Violence Detection

The safety module combines MobileNet spatial features with a bidirectional LSTM temporal layer [13], [14]. Individual frames are processed by MobileNet inside a time-distributed wrapper; the resulting feature sequence is passed to a Bidirectional LSTM with 32 units and then to dense layers of 256, 128, 64, and 32 units with ReLU activations and 0.25 dropout. Training used binary cross-entropy and Adam with a learning rate of 5×10^{-4} for 20 epochs with a 20% validation split. Real-time inference uses a fixed-length sliding queue of 15 frames, allowing each new frame to update the current sequence without reallocating the full temporal buffer.

5.5 | Well-Being Risk Screening Classifier

The tabular screening experiment compared Logistic Regression, Random Forest, and Gradient Boosting formulations under multi-class and binary targets. The binary formulation was retained for reporting because it achieved the highest operational performance in the supplied results, including 95% overall accuracy and 0.92 recall for the needs-support class. Importantly, the output is a screening flag for human follow-up. It is not used to assign a psychiatric diagnosis, and it should not be combined mechanically with facial-expression categories to infer a mental-health condition.

6 | Experimental Evaluation

6.1 | Evaluation Principles

Evaluation was performed module by module using the test partitions and inference logs available in the project record. Because the datasets, units of analysis, and class structures differ across modules, the paper does not average their scores into a single “system accuracy.” Instead, each module is evaluated with metrics appropriate to its task. Classification modules use precision, recall, F1-score, and accuracy; the object-detection module uses mAP; the biometric module additionally reports false acceptance and false rejection; and the violence module reports sequence latency. Table 3 provides a concise cross-module summary, while Tables 4–7 provide detailed module results.

Table 3. Cross-module evaluation summary.

Module	Evaluation unit	Primary reported result	Interpretive note
Face identity	47,410 pair comparisons	99.94% accuracy; FAR 0%; FRR 3.13%	Observed FAR is zero in the test set, not a universal guarantee
Classroom behavior	10,500 held-out images	98.69% accuracy	Seven balanced classes; internal held-out performance
Facial-expression detector	288 evaluation images / 1,672 instances	mAP@0.50 = 0.877; mAP@0.50:0.95 = 0.697	Dataset-defined expression categories only
Violence detector	40 test video sequences	90% accuracy; 110.55 ms/15 frames	Pilot-scale test subset; human review required
Well-being screen	569 tabular test records	95% accuracy; needs-support recall 0.92	Agreement with operational screening target, not diagnosis

Table 3 demonstrates why the framework should be evaluated as a collection of specialized modules rather than as one monolithic predictor. The strongest numerical performance is observed in face routing and classroom behavior classification, while the violence module is evaluated on a much smaller test subset and therefore carries substantially greater statistical uncertainty. The well-being model’s 95% accuracy likewise reflects a project-defined screening label rather than an independently adjudicated clinical endpoint.

6.2 | Face Verification Results

At the 0.5 cosine-similarity threshold shown in Figure 2, the face-verification test included 46,548 impostor pairs and 862 genuine pairs. No impostor pair exceeded the threshold. Twenty-seven genuine pairs fell below the threshold, leaving 835 correctly accepted genuine comparisons. These counts reproduce the reported accuracy, recall, F1-score, FAR, and FRR values in Table 4.

Table 4. Face-verification performance at cosine-similarity threshold 0.5.

Metric	Value	Evidence / interpretation
Pairwise accuracy	99.94%	47,383 correct decisions out of 47,410 comparisons
Precision	100.00%	No false acceptance in the reported impostor comparisons

Recall / TAR	96.87%	835 of 862 genuine comparisons accepted
F1-score	98.41%	Harmonic mean of precision and recall
Observed FAR	0.00%	0/46,548 impostor comparisons; exact 95% upper bound $\approx$ 0.0079%
FRR	3.13%	27/862 genuine comparisons; exact 95% CI $\approx$ 2.07%–4.52%

Table 4 is more informative than reporting accuracy alone. The observed zero false-acceptance count is favorable for privacy-sensitive routing, but a finite test set cannot prove a true FAR of exactly zero. The exact binomial upper confidence bound is approximately 0.0079% at 95% confidence. Conversely, the 3.13% FRR indicates that the conservative threshold sometimes rejects a genuine registered face. In a school deployment, this error mode is preferable to silently assigning a student's data to the wrong family profile, provided that missed recognition is handled as "unknown" rather than as absence or misconduct.

6.3 | Classroom Behavior Results

The classroom-behavior evaluation contains 1,500 held-out examples for each of seven classes. The project training pipeline reports 98.69% overall accuracy, while the rounded classification report shows precision, recall, and F1-score values between 0.97 and 1.00. Table 5 reproduces the per-class results from the evaluation report.

Table 5. Seven-class classroom-behavior classification results.

Behavior class	Precision	Recall	F1-score	Support
Turning Around	0.97	0.99	0.98	1,500
Looking Forward	0.99	0.99	0.99	1,500
Standing	0.99	0.99	0.99	1,500
Reading	0.99	0.99	0.99	1,500
Raising Hand	0.98	0.97	0.98	1,500
Writing	0.99	1.00	0.99	1,500
Sleeping	0.99	0.97	0.98	1,500

As shown in Table 5, the lowest recall is 0.97 for Raising Hand and Sleeping, and the lowest precision is 0.97 for Turning Around. The uniform support avoids class-frequency dominance in the reported test metric. Nevertheless, these results should be interpreted as performance on the project's held-out image partition. Because the behavior representation is derived through a face-centered embedding pipeline, the result does not establish that the same accuracy will hold under subject-disjoint, school-disjoint, or camera-disjoint evaluation.

6.4 | Facial-Expression Category Results

The YOLOv8m evaluation contains 288 images and 1,672 annotated instances. Table 6 reports the detector metrics by class. The aggregate precision is 0.861, recall is 0.816, mAP@0.50 is 0.877, and mAP@0.50:0.95 is 0.697.

Table 6. YOLOv8m facial-expression category detection metrics.

Class	Images	Instances	Precision	Recall	mAP@0.50	mAP@0.50:0.95
All	288	1,672	0.861	0.816	0.877	0.697
Angry	138	322	0.888	0.761	0.869	0.698
Fear	128	308	0.879	0.834	0.915	0.728
Happy	147	431	0.949	0.906	0.961	0.768
Sad	124	263	0.829	0.772	0.816	0.648
Neutral	92	348	0.763	0.807	0.824	0.640

Table 6 shows that Happy is the easiest category in this evaluation ($\text{mAP}@0.50 = 0.961$), whereas Neutral and Sad are comparatively more difficult under the stricter $\text{mAP}@0.50:0.95$ metric. This pattern reinforces the decision not to interpret the classifier as a psychological state estimator. The detector may provide a contextual visual cue, but the output should be combined with temporal context and human judgment and should not be used to diagnose anxiety, depression, or other conditions.

6.5 | Violence Detection Results

The reported violence test subset contains 40 sequences: 21 non-violent and 19 violent. The confusion matrix contains 21 true negatives, 0 false positives, 4 false negatives, and 15 true positives. The resulting 90% accuracy and class-level metrics are shown in Table 7. The model processed a 15-frame sequence in 110.55 ms in the supplied inference run.

Table 7. Violence-detection results on the reported 40-sequence test subset.

Class / metric	Precision	Recall	F1-score	Support / value
Non-violence	0.84	1.00	0.91	21
Violence	1.00	0.79	0.88	19
Overall accuracy	—	—	0.90	40 sequences
Inference latency	—	—	—	110.55 ms per 15 frames

The zero false-positive count in Table 7 is operationally attractive because unnecessary violence alerts can undermine trust. However, the subset includes only 21 non-violent test examples, so the result is statistically less certain than the face-verification result. The more important error in this run is the four missed violent sequences, which produce a violence recall of 0.79. For that reason, the model is positioned as an alerting aid, not as a substitute for supervision or incident investigation. The original Real-Life Violence Situations benchmark contains 2,000 videos and was introduced by Soliman et al. [15]; broader cross-dataset validation would be required before interpreting the 40-sequence result as a stable deployment-level estimate.

6.6 | Well-Being Screening Results

The binary well-being evaluation contains 569 records, including 384 lower-risk records and 185 needs-support records. The classification report is summarized in Table 8. The overall accuracy is 0.95; for the needs-support class, precision is 0.93, recall is 0.92, and F1-score is 0.93.

Table 8. Binary well-being screening classification report.

Screening class	Precision	Recall	F1-score	Support
Lower-risk (0)	0.96	0.97	0.97	384
Needs-support (1)	0.93	0.92	0.93	185
Macro average	0.95	0.95	0.95	569
Weighted average	0.95	0.95	0.95	569

The high recall for the needs-support class is consistent with the module's intended triage role: a screening system should avoid missing students who may warrant a conversation with an authorized professional. Nevertheless, Table 8 does not establish clinical sensitivity or specificity because the target is the project-defined Risk_Flag rather than an independent diagnostic interview. The result therefore supports implementation consistency, not a claim that the model can diagnose depression, anxiety, or any other disorder.

7 | Integrated Workflow and Application Layer

The application layer converts model outputs into role-specific information rather than exposing raw predictions indiscriminately. Figure 3 illustrates the mobile interaction flow. Users first select an authorized role and sign in. The overview screen then presents summarized trends and alerts, while the notification view

lists events such as repeated aggression, low engagement, smoking-related alerts, or sleeping-in-class events. In a production deployment, each alert should include timestamp, source module, confidence or persistence status, and a clear indication that human review is required.

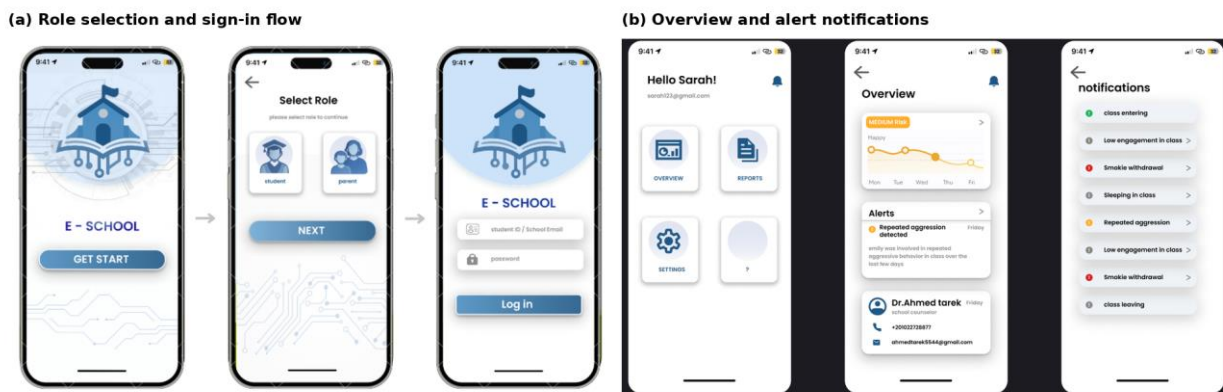


Figure 3. Mobile application workflow: (a) role selection and sign-in; (b) overview and alert-notification interfaces.

The end-to-end workflow is deliberately asymmetric. Low-stakes aggregate engagement summaries may be delivered periodically, whereas safety or well-being alerts should be routed only to the minimum set of authorized recipients. Parent access does not imply unrestricted access to all school observations; school policy must determine which events are appropriate to share, when a counselor should review a risk flag first, and how long event metadata should be retained. This prevents the communication layer from converting a privacy-preserving inference pipeline into excessive behavioral surveillance.

8 | Discussion

8.1 | Technical Interpretation

The results show that the prototype can support several distinct edge-AI tasks with useful internal test performance. Face routing is the most statistically well-supported module because its evaluation includes tens of thousands of impostor comparisons and hundreds of genuine comparisons. Classroom-behavior classification also reports a large held-out sample, although its image-level split and representation design limit claims about cross-student generalization. The facial-expression detector achieves moderate-to-strong object-detection performance, with the expected decline from $mAP@0.50$ to the stricter $mAP@0.50:0.95$. The violence detector is computationally compatible with near-real-time sequence analysis but has only a small reported test subset. The well-being classifier shows strong agreement with its operational binary label but should remain a screening aid because the target is derived from questionnaire features.

From a systems perspective, the main result is that these modules can be separated into independent services and joined through a common metadata layer. This makes it possible to apply different thresholds, retention rules, and access controls to different types of evidence. It also reduces the temptation to fuse unrelated outputs into a single opaque “student risk score,” which would be difficult to validate and potentially harmful.

8.2 | Comparison with Existing Work

Table 9 positions BrightPath relative to representative classroom-AI research. The purpose is not to rank methods numerically, because the studies evaluate different tasks and populations. Instead, the table identifies architectural scope. Existing work demonstrates strong advances in attention detection, privacy-preserving federated learning, and edge behavior recognition [3, 5-7]. BrightPath adds value by combining an open-set identity-routing gate, safety detection, a separate questionnaire screen, and a family-facing application within one human-reviewed workflow.

Table 9. Qualitative comparison with representative classroom-AI systems.

System / study	Primary scope	Edge / privacy mechanism	Open-set identity routing	Safety / violence	Questionnaire well-being screen	Family-facing application
Trabelsi et al. [3]	Attention / classroom behavior	Not the primary contribution	Not reported	Not reported	No	No
Wu et al. [5]	Student status monitoring	Federated semi-supervised learning	Not reported	No	No	No
StudentNet [6]	Student behavior detection	Edge-deployable detector	No	No	No	No
Valdes-Ramirez et al. [7]	Behavior + environmental variables	Hybrid edge processing; low-bandwidth metadata	No	No	No	Dashboard-oriented
BrightPath AI	Identity routing + behavior + safety + screening	Local video inference; raw frames discarded after inference	Yes	Yes	Yes	Yes

Table 9 clarifies the contribution without overstating novelty. Edge processing and privacy-preserving classroom analytics already exist in the literature. BrightPath’s contribution is the integrated workflow and its explicit separation between observation, screening, and human action. That distinction is important for scientific reproducibility because it identifies exactly which parts of the framework are established components and which part is the system-level contribution.

8.3 | Responsible-Use Boundaries, Privacy, and Child Protection

AI systems involving children require a higher standard of governance than ordinary analytics. UNICEF guidance emphasizes safety, privacy, non-discrimination, transparency, accountability, and children’s best interests [2]. BrightPath addresses part of this requirement through local video processing and raw-frame deletion, but responsible deployment requires additional controls beyond model accuracy. These include purpose limitation, documented legal basis or consent as applicable, role-based access, retention schedules, security controls, procedures for contesting incorrect records, and independent review of harms and bias.

Biometric performance must also be evaluated across relevant demographic and environmental groups. NIST evaluations have repeatedly shown that face-recognition error rates can vary with image quality, age, sex, and demographic composition, depending on the algorithm and application [19]. The 99.94% pairwise accuracy reported here is therefore not evidence of demographic parity. Before use with minors, the identity module should be audited on the actual deployment population, with particular attention to false matches, false non-matches, camera position, lighting, occlusion, and age-related appearance changes.

The facial-expression model requires an even stricter semantic boundary. Research in affective science cautions against assuming a direct mapping between facial movements and internal emotional state [18]. Consequently, BrightPath must not use an Angry, Sad, or Fear category as evidence that a student is depressed, anxious, dangerous, or in need of discipline. Such outputs may at most serve as low-weight contextual cues for a trained human who considers broader information. The questionnaire module is likewise a screen, not a diagnosis; positive results should lead to appropriate human follow-up rather than automated labeling.

8.4 | Validity Considerations and Limitations

Several constraints define the scope of the reported results. First, the component models were evaluated on different datasets rather than on a single prospective cohort. The paper therefore establishes modular technical feasibility, not clinical or educational effectiveness in a longitudinal school trial. Second, the violence test subset is small, making its zero false-positive count uncertain. Third, the behavior classifier uses a representation pathway centered on FaceNet-derived embeddings; although the internal results are high, a body- or pose-specific representation would provide a stronger basis for cross-subject behavior generalization. Fourth, the screening target is constructed from questionnaire-derived variables, so the reported classification performance is not independent diagnostic validation. Fifth, the current evidence does not establish demographic fairness, school-to-school transfer, or robustness under all forms of occlusion and illumination.

These limitations do not negate the system-level contribution; rather, they define what the present evidence supports. The manuscript demonstrates a technically integrated, privacy-minimizing prototype and provides transparent module-level performance. Claims are intentionally restricted to that level. A publication or deployment decision should therefore evaluate the architecture, the measured component behavior, and the governance design separately from any future claim of improved mental-health or academic outcomes.

9 | Conclusion

BrightPath AI presents an integrated edge-AI framework for classroom engagement, safety, and student well-being that is designed around data minimization and human oversight. The system combines open-set face identity routing, seven-class classroom-behavior recognition, facial-expression category detection, temporal violence detection, and questionnaire-based well-being screening with a parent/teacher/administrator application layer. Raw classroom frames are processed locally and discarded after inference, while downstream services operate on derived metadata.

The reported experiments show strong internal performance for face routing and classroom-behavior classification, useful object-detection performance for facial-expression categories, real-time-capable temporal violence inference, and high agreement with an operational well-being screening label. Equally important, the paper defines what these results do not mean: facial-expression categories are not psychological diagnoses, screening outputs are not clinical determinations, and automated alerts are not disciplinary decisions. By making these boundaries explicit, the framework converts a collection of AI models into a more defensible early-warning architecture for child-centered educational settings.

The scientific contribution is therefore an integrated and auditable architecture rather than an attempt to replace teachers, counselors, or parents. The evidence supports the feasibility of combining multiple analytical services at the edge while preserving raw-video locality and routing only derived events to authorized users. This provides a practical foundation for subsequent prospective validation under institutionally approved, consent-aware, fairness-audited school deployments.

Author Contribution

All authors contributed equally to this study.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Ethics and Responsible-Deployment Statement

This manuscript reports a technical system prototype evaluated with project and public datasets; it does not report a clinical trial and does not claim diagnostic validation. Any real-world deployment involving minors or identifiable biometric data should be subject to the applicable institutional ethics process, legal and school-governance requirements, age-appropriate consent/assent procedures where required, data-protection controls, fairness assessment, and human oversight. The system should not be used for autonomous mental-health diagnosis, student ranking, or disciplinary action.

References

- [1] World Health Organization, “Mental health of adolescents,” 1 Sep. 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health>
- [2] UNICEF Office of Strategy and Evidence Innocenti, Guidance on AI and Children, Version 3.0. Florence, Italy: UNICEF, Dec. 2025.
- [3] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, “Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student’s Behavior Recognition,” *Big Data and Cognitive Computing*, vol. 7, no. 1, art. 48, 2023, doi: 10.3390/bdcc7010048.
- [4] L. Liu and J. Yang, “A Meta-Learning-Based Multi-Scene Student Posture Detection Method for Enhancing Learning Motivation and Engagement in Smart Classrooms,” *Journal of Applied Science and Engineering*, vol. 29, no. 3, 2026, doi: 10.6180/jase.202603_29(3).0021.
- [5] H. Wu, Z. Lu, and J. Zhang, “A privacy-preserving student status monitoring system,” *Complex & Intelligent Systems*, vol. 9, pp. 597–608, 2023, doi: 10.1007/s40747-022-00796-5.
- [6] L. Ai and H. Hu, “StudentNet: an edge-deployable approach for student behavior detection in smart classrooms,” *Frontiers in Computational Neuroscience*, vol. 20, art. 1828972, 2026, doi: 10.3389/fncom.2026.1828972.
- [7] D. Valdes-Ramirez, V. S. Caamal, J. I. Huertas Cardozo, and G. Zavala, “Edge Computing Model for Real-Time Classroom Behavior Analysis in Face-to-Face Lessons,” in *Distributed Computing and Artificial Intelligence, Special Sessions 2, Lecture Notes in Networks and Systems*, vol. 1630, 2026, pp. 495–505, doi: 10.1007/978-3-032-05748-8_42.
- [8] B. Williamson, “Decoding ClassDojo: psycho-policy, social-emotional learning and persuasive educational technologies,” *Learning, Media and Technology*, vol. 42, no. 4, pp. 440–453, 2017, doi: 10.1080/17439884.2017.1278020.
- [9] D. Amo-Filva, D. Fonseca, F. J. García-Peñalvo, M. Alier Forment, M. J. Casany Guerrero, and G. Godoy, “Exploring the landscape of learning analytics privacy in fog and edge computing: A systematic literature review,” *Computers in Human Behavior*, art. 108303, 2024, doi: 10.1016/j.chb.2024.108303.
- [10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.
- [11] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild,” in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212.
- [12] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
- [13] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *arXiv:1704.04861*, 2017.
- [14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [15] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, and D. Khattab, “Violence Recognition from Videos using Deep Learning Techniques,” in *Proc. 9th Int. Conf. Intelligent Computing and Information Systems (ICICIS)*, 2019, pp. 79–84, doi: 10.1109/ICICIS46948.2019.9014714.
- [16] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9: Validity of a brief depression severity measure,” *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.
- [17] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, “A brief measure for assessing generalized anxiety disorder: The GAD-7,” *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, 2006, doi: 10.1001/archinte.166.10.1092.
- [18] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements,” *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019, doi: 10.1177/1529100619832930.
- [19] P. Grother, M. Ngan, and K. Hanaoka, Face Recognition Vendor Test Part 3: Demographic Effects, NISTIR 8280. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2019, doi: 10.6028/NIST.IR.8280.
- [20] B. Pang, E. Nijkamp, and Y. N. Wu, “Deep Learning With TensorFlow: A Review,” *Journal of Educational and Behavioral Statistics*, vol. 45, no. 2, pp. 227–248, 2020, doi: 10.3102/1076998619872761.

- [21] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," in Proc. Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 1–5.
- [22] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [23] Kaggle, "Classroom Student Behaviors Dataset." [Online]. Available: <https://www.kaggle.com/datasets/phamluhyunhmai/classroom-student-behaviors>. Accessed: 13 Jun. 2026.
- [24] Roboflow Universe, "Emotion Detector Dataset." [Online]. Available: <https://universe.roboflow.com/img-project/emotion-detector-ngezz>. Accessed: 13 Jun. 2026.
- [25] Open Science Framework, "Children Mental Health Dataset." [Online]. Available: <https://osf.io/k3xtd/files/4nzkt>. Accessed: 13 Jun. 2026.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.