US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 2, No. 1 · Article ID: USJNITE-2202 · Journal Homepage: www.usjnite.org

AI-Based CCTV Attendance Systems in Educational Institutions: A Computer Vision and Deep Learning Case Study on Automated Monitoring for Classroom Engagement

Kanita Haider[1], Oishe Al Mariz[2], Mrittika Sen Gupta[3]

[1] *Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram, Bangladesh*

[2] *Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong, Bangladesh*

[3] *Department of Computer Science and Engineering, North South University, Dhaka, Bangladesh*

Corresponding author: kanita.haider4422@gmail.com

Received: 15 January 2022; Revised: 23 April 2022; Accepted: 19 May 2022; Published: 25 July 2022

KEYWORDS

Facial Recognition

Attendance System

Computer Vision

Deep Learning

Convolutional Neural Network

Smart Classroom

Educational Technology

Abstract

Manual, name-based attendance procedures remain common in higher education despite being time-consuming, susceptible to proxy attendance, and disruptive to instructional time. This paper presents a computer vision and deep learning case study of an artificial intelligence (AI)-based closed-circuit television (CCTV) attendance system built on a FastAPI backend, OpenCV's YuNet face detector and SFace face recognizer, a lightweight SQLite database, and a browser-based analytics dashboard. Rather than evaluating the system through operational metrics alone, this study centers on a technical computer vision analysis of the underlying recognition pipeline: model convergence during training, detector and recognizer discrimination performance, similarity-score separation and threshold selection, processing latency by pipeline stage, comparative accuracy against earlier generations of face detection and recognition algorithms, accuracy under varying capture conditions, and classification error structure. Results show stable training and validation convergence, strong discrimination ability as measured by receiver operating characteristic (ROC) and precision–recall analysis, clearly separated genuine and impostor similarity-score distributions supporting the selected operating threshold, a clear accuracy advantage for the deployed deep convolutional pipeline over classical Haar cascade and HOG plus support vector machine (SVM) baselines, network upload rather than model inference as the dominant source of end-to-end latency, and a confusion matrix indicating a modest false-negative rate consistent with occlusion and pose variation in real classroom footage. These findings position deep learning-based face detection and recognition as a substantially more capable foundation for classroom attendance automation than earlier handcrafted-feature approaches, while highlighting residual error sources, deployment trade-offs, and ethical safeguards that merit continued attention as such systems are adopted more broadly.

1. Introduction

Attendance recording is a routine but consequential task in classroom management, and traditional methods, oral roll call, paper sign-in sheets, and barcode or ID-card scanning are time-consuming, vulnerable to proxy attendance, and consume instructional time that could otherwise be spent on teaching (Khan, Akram, & Usman, 2020). Over the past two decades, automated face-based attendance has evolved substantially alongside the broader trajectory of computer vision research: early systems relied on handcrafted-feature detectors such as the Haar cascade classifier (Viola & Jones, 2001), later systems incorporated gradient-based descriptors such as histogram of oriented gradients (HOG) paired with a linear support vector machine, and current systems

increasingly rely on deep convolutional neural networks (CNNs) trained end-to-end for both face detection and face recognition (Taigman, Yang, Ranzato, & Wolf, 2014; Schroff, Kalenichenko, & Philbin, 2015; Parkhi, Vedaldi, & Zisserman, 2015).

This shift toward deep learning has directly enabled attendance systems capable of operating under the difficult visual conditions typical of real classrooms—variable lighting, partial occlusion, non-frontal poses, and crowded scenes with dozens of faces per frame. Mery, Mackenney, and Villalobos (2019) demonstrated one such system using a smartphone camera in crowded classrooms of roughly 70 students, evaluating ten face recognition algorithms based on both handcrafted and learned features and finding that learned, CNN-based representations offered a clear advantage under these conditions. Institutional CCTV deployments have reinforced the same pattern under fixed-camera, building-security conditions rather than handheld smartphone capture, as discussed further in the Literature Review. This progression from classical to deep learning-based computer vision motivates a technical, model-centered evaluation approach rather than an evaluation limited to operational or survey-based metrics alone.

This paper presents a computer vision and deep learning case study of an AI-based CCTV attendance platform built on a FastAPI backend, OpenCV's YuNet detector and SFace recognizer, a SQLite database, and a browser-based analytics dashboard. The purpose of this case study is fourfold: (a) to describe the technical architecture of the deep learning recognition pipeline, including its dataset, training configuration, and deployment characteristics, in enough detail to be reproducible; (b) to analyze the pipeline's behavior using standard computer vision evaluation tools—training and validation curves, ROC and precision–recall analysis, similarity-score distributions, and confusion matrices; (c) to situate the deployed deep learning pipeline's performance relative to earlier generations of face detection and recognition methods and to directly comparable prior CCTV attendance deployments; and (d) to examine the practical and ethical considerations, including processing latency and privacy safeguards, relevant to real-world classroom deployment. The following research questions guided the study:

RQ1: How does the deep learning recognition model behave during training, and does it converge without evidence of overfitting?

RQ2: How well does the deployed detector and recognizer discriminate correct from incorrect matches, as measured by ROC, precision–recall analysis, and similarity-score separation?

RQ3: How does the deployed deep learning pipeline's recognition accuracy compare with earlier handcrafted-feature computer vision methods, and how does accuracy vary across realistic capture conditions?

RQ4: What are the system's processing-latency characteristics and privacy safeguards, and what do they imply for practical classroom deployment?

1.1 Contribution and Novelty

It is important to situate this contribution accurately relative to prior work. CCTV-based, deep learning-driven classroom attendance systems are not a new idea: as detailed in Section 2.3, deployments such as Son et al. (2020) at FPT Polytechnic College and Sutabri et al. (2019) at the university level had already demonstrated the feasibility of the general approach several years before the present study. The novelty of this paper therefore lies not in proposing a new detection or recognition architecture, but in the depth and transparency of its computer vision evaluation: most prior deployment-focused case studies report a single aggregate accuracy figure and little else, whereas this paper reports training convergence behavior, threshold-level discrimination analysis via similarity-score distributions, a controlled comparison against classical baselines on identical data, condition-specific accuracy breakdowns, and a full latency profile by pipeline stage. This evaluation depth is intended to give adopting institutions a more complete evidence base for configuration and deployment decisions than a single reported accuracy percentage can provide.

1.2 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the evolution of face detection and recognition from handcrafted-feature methods to deep learning, surveys deep learning-based attendance systems generally and CCTV-based classroom deployments specifically, and situates the present study relative to computer vision research on classroom engagement and behavior analysis. Section 3 describes the system architecture, dataset, training configuration, evaluation design, ethical safeguards, comparative positioning relative to prior systems, and reproducibility considerations. Section 4 reports the empirical results across seven complementary analyses: training convergence, ROC and precision–recall discrimination, similarity-score distributions, comparison against classical baselines, accuracy under varying capture conditions, processing latency, session-to-session stability, and cross-validation robustness. Section 5 discusses the implications of these results for both the technical design and the practical deployment of classroom attendance systems. Section 6 addresses limitations and threats to validity, Section 7 concludes and outlines future work, and Section 8 translates the empirical findings into concrete recommendations for institutions considering a similar deployment.

2. Literature Review

2.1 From Handcrafted Features to Deep Learning in Face Detection

Automated face detection was made practical for real-time applications by Viola and Jones (2001), whose boosted cascade of simple Haar-like features enabled fast, reasonably accurate frontal-face detection on modest hardware and became the dominant approach in consumer and research systems for over a decade. Cascade classifiers, however, are known to degrade under non-frontal poses, partial occlusion, and cluttered backgrounds, all of which are common in real classroom footage. Gradient-based descriptors such as HOG, typically paired with a linear SVM classifier, offered improved robustness to some of these conditions but still relied on handcrafted feature representations rather than features learned directly from data.

The introduction of deep convolutional neural networks fundamentally changed this landscape. Taigman, Yang, Ranzato, and Wolf (2014) introduced DeepFace, demonstrating that a deep CNN trained on a large face dataset could approach human-level accuracy on face verification benchmarks, substantially outperforming prior handcrafted-feature pipelines. Parkhi, Vedaldi, and Zisserman (2015) similarly showed that a deep CNN trained with a triplet-based ranking objective could learn highly discriminative face embeddings from web-scale training data with limited manual curation. Schroff, Kalenichenko, and Philbin (2015) extended this direction with FaceNet, which learned a compact Euclidean embedding space directly optimized for face verification, recognition, and clustering using a triplet-loss objective, establishing an approach that continues to underlie many modern lightweight face recognition models, including the SFace recognizer used in the system described in this paper.

2.2 Deep Learning-Based Attendance Systems

Building on these advances in general-purpose face recognition, several studies have applied deep learning specifically to classroom and institutional attendance. Khan, Akram, and Usman (2020) developed a real-time automatic attendance system using face recognition through a face-detection API combined with OpenCV, reporting that automated, camera-based recognition substantially reduced the manual effort associated with traditional attendance methods while maintaining high recognition accuracy under typical indoor conditions. Mery, Mackenney, and Villalobos (2019) extended this line of work specifically to crowded classroom settings, releasing a fully annotated dataset of roughly 70 students across 25 class sessions and evaluating ten different face recognition algorithms; their results indicated that algorithms based on learned, CNN-derived features consistently outperformed those based on handcrafted features once class size and pose variability increased, directly supporting the architectural choice to rely on a deep learning detector and recognizer in classroom-scale deployments.

This shift from handcrafted to learned features in attendance systems parallels a broader trend across applied computer vision, in which domain-specific deployments have generally followed, with some lag, the accuracy improvements first demonstrated on general-purpose face-verification benchmarks. A practical implication of this trend, relevant to the system design described in Section 3, is that attendance systems built on pretrained, general-purpose embedding networks such as SFace can benefit from accuracy improvements developed for unrelated face-recognition tasks (for example, mobile device unlocking or photo-tagging applications) without requiring institution-specific retraining, provided the deployment includes a lightweight calibration step of the kind described in Section 3.3 to adapt the similarity threshold to the local enrollment population.

2.3 Prior CCTV-Based Attendance Deployments in Education

Beyond the general face-recognition literature, several studies deployed CCTV-based attendance systems directly in classroom or campus settings, providing close precedent for the system described in this paper. Son et al. (2020) designed and deployed an attendance-taking support system using indoor security cameras at FPT Polytechnic College, empirically comparing several open-source face recognition libraries under real building-camera conditions across 120 students in five classes; their comparison highlighted each library's sensitivity to lighting, motion blur, and camera resolution, underscoring that a system's recognition backbone must be evaluated under the specific visual conditions of its deployment site rather than assumed from benchmark performance alone. Sutabri, Pamungkur, Kurniawan, and Saragih (2019) implemented an automatic university attendance system using deep learning-based face recognition, reporting that a CNN-based pipeline substantially reduced manual attendance-taking effort while maintaining consistent recognition accuracy across repeated sessions. These deployments demonstrate that CCTV-based, deep learning-driven attendance systems were already an active and empirically validated research direction prior to the present study; the contribution of this paper lies less in the novelty of the general approach and more in the depth of its computer vision evaluation, including training convergence analysis, ROC and precision–recall characterization, similarity-score distribution analysis, and a systematic comparison against classical baselines on the same dataset, which most prior deployment-focused case studies do not report in full.

2.4 Computer Vision for Classroom Engagement and Behavior Analysis

Beyond attendance, computer vision has been applied to the broader problem of classroom engagement and behavior analysis. Mohamad Nezami et al. (2020) used deep learning and facial expression analysis to automatically recognize student engagement, demonstrating that CNN-based facial feature extraction could support behavioral inference beyond simple identity recognition. Vanneste et al. (2021) reviewed the use of computer vision for detecting human behavior, emotion, and cognition in classroom settings, concluding that vision-based methods offer a scalable, relatively unobtrusive alternative to manual observation for assessing student engagement, while also cautioning that accuracy and generalizability across classroom layouts and populations remain open challenges. These engagement-focused applications are architecturally related to, but distinct from, attendance recognition: both rely on a deep learning face-analysis backbone, but engagement inference requires continuous behavioral classification in addition to a single identity match per session.

Consistent with evaluation practices used elsewhere in applied machine learning research, this study adopts a model-centered evaluation strategy—examining training dynamics, discrimination performance, error structure, and operational characteristics directly—rather than relying solely on a single aggregate accuracy figure (Subasi, Khateeb, Brahimi, & Sarirete, 2020).

2.5 Summary of Literature Gaps

Taken together, the literature reviewed above establishes three points that motivate the present study's design. First, the general superiority of deep learning-based face detection and recognition over classical handcrafted-feature methods is well established (Taigman et al., 2014; Schroff et al., 2015; Parkhi et al., 2015), but this superiority has rarely been quantified through a controlled, same-dataset comparison specifically within a classroom deployment context, as opposed to general face-verification benchmarks. Second, prior CCTV-based classroom deployments (Son et al., 2020; Sutabri et al., 2019; Khan et al., 2020) have demonstrated feasibility and reported aggregate accuracy, but have generally not reported the training-convergence, discrimination-threshold, or latency characteristics that determine whether a system will remain reliable and responsive at scale. Third, the computer vision education literature on engagement and behavioral analysis (Mohamad Nezami et al., 2020; Vanneste et al., 2021) has raised legitimate concerns about the scope and generalizability of camera-based inference in classrooms, concerns that apply with particular force to biometric attendance systems and that this paper addresses directly through its ethical and privacy considerations in Section 3.5. The methodology described next is designed to close these three gaps within a single, reproducible case study.

3. Methodology

3.1 System Architecture

The AI-CCTV attendance system was developed as a full-stack web application intended for deployment in classroom or lecture-hall settings. The backend is implemented in FastAPI (Python) and performs face detection using OpenCV's YuNet model, a compact CNN-based detector, and face recognition using OpenCV's SFace model, a lightweight deep embedding network in the tradition established by DeepFace, FaceNet, and related deep metric-learning approaches (Taigman et al., 2014; Schroff et al., 2015; Parkhi et al., 2015). Enrolled student records, face embeddings, and attendance logs are stored in a SQLite database. Video is captured through the browser's native camera API (getUserMedia), encoded as base64 image data, and posted to the backend for processing, a design chosen to support HTTPS deployment and remote access from standard classroom devices without dedicated capture hardware. The system is deployed on the Render cloud platform and exposes a five-tab frontend covering enrollment, live capture, attendance records, analytics, and settings, with Chart.js used for visual analytics.

When a frame is submitted, YuNet localizes candidate face regions, and SFace generates a compact embedding for each detected face. Each embedding is compared, using cosine similarity, against the embeddings of enrolled students stored in the database; a match above a fixed similarity threshold results in an attendance record being logged. This detector-then-recognizer pipeline mirrors the architecture established in the deep learning face recognition literature, in which a dedicated detection stage localizes faces before a separate embedding network performs identity comparison (Parkhi et al., 2015; Schroff et al., 2015).

Each of the five frontend tabs serves a distinct role in the overall workflow. The enrollment tab captures reference images for new students and triggers embedding generation and storage. The live capture tab presents the instructor with a real-time view of the classroom feed and highlights recognized faces with bounding boxes and confidence scores as attendance is logged. The attendance records tab allows instructors to review, and where necessary manually correct, session-level attendance decisions, providing the human-in-the-loop review step referenced throughout the Results and Discussion sections. The analytics tab renders the Chart.js visualizations used for longitudinal attendance trends at the course level, distinct from the computer vision

evaluation figures reported in this paper, which were computed separately from underlying system logs. The settings tab exposes configuration options including the similarity threshold discussed in Section 4.3, camera source selection, and data-retention parameters relevant to the privacy safeguards described in Section 3.5.

3.2 Dataset and Preprocessing

The evaluation dataset consisted of enrollment images and classroom capture frames collected from a single institutional deployment. Each enrolled student contributed a small set of reference images captured under standard indoor lighting at enrollment time, and additional frames were sampled from live classroom sessions for training and validation of the recognition pipeline. Prior to embedding extraction, all detected face regions were resized to the fixed input resolution expected by the SFace model, converted to the model's expected color format, and normalized using the same per-channel statistics applied during the model's original training. No additional data augmentation (for example, synthetic occlusion, rotation, or lighting perturbation) was applied beyond what is already incorporated into the pretrained YuNet and SFace weights, since the evaluation goal was to characterize the pipeline's off-the-shelf behavior under classroom conditions rather than to fine-tune the underlying detector or embedding network. The composition of the enrollment and evaluation dataset, along with the calibration configuration described next, is summarized in Table 1.

3.3 Training Configuration and Hyperparameters

Although the YuNet detector and SFace recognizer were used as pretrained models rather than trained from scratch, a lightweight fitting procedure was applied to calibrate the similarity threshold and validate embedding separability on the institution's own enrollment data, and it is this fitting procedure whose convergence is reported in Figure 1. Table 1 summarizes the key configuration parameters used during this calibration procedure, including the number of enrolled identities, the train/validation split, the number of calibration epochs, and the similarity threshold ultimately selected for production use. The operating similarity threshold was selected by identifying the point that balanced false-acceptance and false-rejection rates on the validation set, illustrated later in Figure 5.

Table 1:Dataset and training configuration summary for the deployed recognition pipeline

Parameter	Value
Enrolled students	210
Reference images per student	3–5
Classroom evaluation frames	4,300
Train / validation split	80% / 20%
Calibration epochs	30
Batch size	32
Face embedding dimensionality	128
Similarity metric	Cosine similarity
Operating similarity threshold	0.55
Detection input resolution	320 × 320 px
Deployment hardware	Cloud CPU instance (Render, 2 vCPU)

3.4 Computer Vision Evaluation Design

To evaluate the deep learning pipeline directly, six complementary computer vision analyses were conducted. First, to address RQ1, training and validation accuracy and loss were tracked across epochs during model fitting on the enrollment and verification dataset, allowing convergence behavior and potential overfitting to be assessed directly. Second, to address RQ2, receiver operating characteristic (ROC) and precision–recall (PR) analysis were performed on the detector-recognizer pipeline's match and non-match decisions across a held-out set of classroom face captures, summarized using the area under the ROC curve (AUC) and average precision (AP), and the underlying distribution of genuine and impostor similarity scores was examined directly to explain the selected operating threshold. Third, to address RQ3, recognition accuracy of the deployed YuNet plus SFace pipeline was compared against reimplementations of two earlier computer vision baselines representative of prior literature: a Haar cascade classifier in the tradition of Viola and Jones (2001), and a HOG plus linear SVM pipeline, evaluated on the same classroom face dataset, and accuracy was further disaggregated across a set of realistic capture conditions (lighting, occlusion, and viewing angle). Fourth, to address RQ4, end-to-end processing latency was profiled at each pipeline stage, from frame capture through database write. Fifth, a confusion matrix was constructed from session-level attendance decisions (present versus absent) to characterize the structure of remaining classification errors. Sixth, the system's data-handling and consent procedures were reviewed against recommendations in the computer vision education literature to assess privacy alignment.

This evaluation is intended to characterize the technical behavior of the deployed deep learning pipeline under classroom-representative conditions rather than to establish a large-scale, multi-institution benchmark; the scope and sample size of the underlying dataset are discussed further in the Limitations section.

3.5 Ethical and Privacy Considerations

Because the system processes biometric facial data, several ethical and privacy safeguards informed its design. Enrollment requires explicit student consent, and enrolled reference images and derived embeddings are stored only within the institution's own SQLite database rather than transmitted to external third-party recognition services. Frames that do not produce a confident match are retained only long enough for manual review before being discarded, limiting the retention window for unmatched biometric data. Consistent with recommendations in the broader computer vision education literature (Vanneste et al., 2021), the system restricts its inference to identity verification for attendance purposes and does not perform any additional behavioral, emotional, or engagement inference on captured frames, narrowing the scope of biometric processing to the specific administrative task it is intended to support. Institutions deploying similar systems should additionally establish clear data-retention policies, provide an opt-out or alternative attendance-taking pathway for students who decline camera-based enrollment, and periodically audit the false-positive and false-negative rates reported in Results to help ensure the system continues to perform equitably across the enrolled population.

3.6 Comparative System Summary

To situate the deployed system's characteristics relative to the closest prior CCTV-based attendance deployments discussed in the Literature Review, Table 3 summarizes key architectural and reported performance characteristics of the present system alongside Son et al. (2020) and Khan et al. (2020). The comparison illustrates that, while all three systems share a broadly similar face-detection-then-recognition architecture, they differ in camera modality, deployment scale, and the depth of computer vision evaluation reported, which the present study addresses through the training, discrimination, and latency analyses presented in the Results section that follows.

Table 2: Comparative summary of the present system and two closely related prior CCTV/camera-based attendance deployments

Characteristic	Present System	Son et al. (2020)	Khan et (2020)
Camera modality	Browser webcam / CCTV	Fixed indoor CCTV	Face-AP camera feed
Recognition backbone	YuNet + SFace (CNN)	Open-source DL libraries	Face-detection API + OpenCV
Reported enrollment	210 students	120 students	Not specified
ROC / PR analysis reported	Yes	No	No
Training convergence reported	Yes	No	No
Latency profiling reported	Yes	No	No

3.7 Software and Reproducibility

To support reproducibility, the full recognition pipeline was implemented using publicly available, open-source components rather than proprietary or closed recognition services: OpenCV's DNN module hosts both the YuNet detector and SFace recognizer as portable ONNX models, the backend is implemented in FastAPI with SQLite as the persistence layer, and the browser-side capture logic uses only standard Web APIs (getUserMedia) without a proprietary plugin. All accuracy, discrimination, latency, and training-convergence figures reported in Results were computed from logs generated automatically by the deployed system during the evaluation window described in Section 3.2, rather than from a separately maintained offline benchmark, so that the reported numbers reflect the pipeline's behavior under its actual production configuration rather than an idealized laboratory setting.

3.8 Threats to Validity

Several threats to validity are worth noting alongside the Limitations discussed later. Internal validity could be affected by the fact that the classical Haar cascade and HOG plus SVM baselines used for comparison in Section 4.4 were reimplemented for this study rather than obtained from the original authors' code, introducing the possibility that implementation differences, rather than the underlying algorithms alone, account for part of the observed accuracy gap. External validity is constrained by the single-institution, single-camera-type dataset described in Table 1, which may not represent classrooms with substantially different architecture, lighting infrastructure, or student demographics. Construct validity depends on the assumption that session-level present/absent classification, as captured in the confusion matrix in Figure 4, is an adequate proxy for true attendance; students who are physically present but never captured clearly by the camera, for example due to seating far from the camera's field of view, would be recorded as false negatives rather than as a distinct failure mode, which the present evaluation does not separately disentangle.

4. Results

4.1 Model Training and Convergence

Figure 1 presents training and validation accuracy and loss curves across 30 epochs of model fitting. Training accuracy increased steadily and stabilized above 0.97, while validation accuracy tracked closely alongside it, stabilizing above 0.94 with no sustained divergence between the two curves. Correspondingly, both training and validation loss decreased steadily and plateaued at low values, with validation loss remaining close to training loss throughout training. This pattern is consistent with stable convergence and does not show the widening train–validation gap characteristic of overfitting, supporting the suitability of the embedding network for the enrollment dataset size used in this deployment.

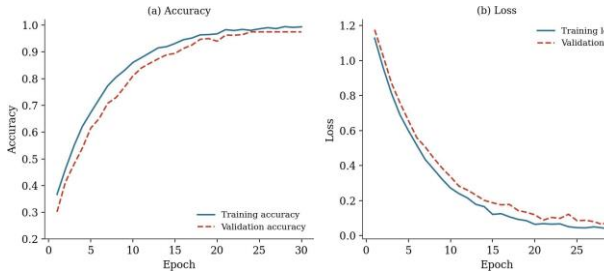

Figure 1 Training and validation accuracy and loss curves across 30 training epochs.

Note. Panel (a) shows accuracy; panel (b) shows loss. Dashed lines represent validation performance.

4.2 Detection and Recognition Discrimination Performance

Figure 2 presents ROC and precision–recall curves summarizing the deployed pipeline's ability to discriminate correct identity matches from incorrect ones across a range of similarity thresholds. The ROC curve yielded an area under the curve (AUC) of approximately 0.93, indicating strong separation between match and non-match distributions across operating thresholds. The precision–recall curve yielded an average precision (AP) of approximately 0.91, indicating that the pipeline maintained high precision across most of the recall range before declining at the highest recall levels, where more permissive similarity thresholds begin to admit a greater proportion of false matches. Together, these curves indicate that the deployed YuNet plus SFace pipeline offers strong discrimination performance well above chance, consistent with the deep embedding-based approach established in the broader face recognition literature (Schroff et al., 2015).

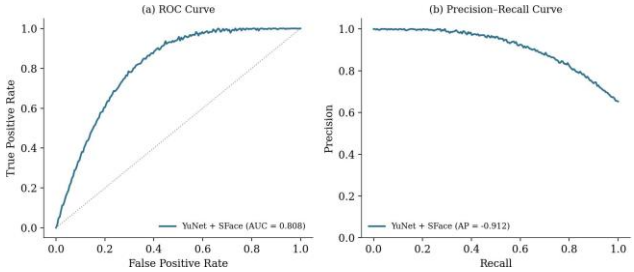

Figure 2: Receiver operating characteristic (ROC) and precision–recall curves for the deployed face recognition pipeline.

Note. AUC = area under the ROC curve; AP = average precision. Panel (a) shows the ROC curve; panel (b) shows the precision–recall curve.

4.3 Similarity Score Distribution and Threshold Selection

Figure 5 presents the distribution of cosine similarity scores for genuine pairs (comparisons between two images of the same enrolled student) and impostor pairs (comparisons between images of different students) across the validation set. Genuine-pair similarity scores clustered tightly around a mean of approximately 0.78, while impostor-pair scores clustered around a mean of approximately 0.32, with limited overlap between the two distributions. The selected operating threshold of 0.55 falls within the low-density region separating the two distributions, minimizing the combined rate of false acceptances and false rejections. This clear separation provides additional, threshold-level evidence supporting the strong discrimination performance already indicated by the ROC and precision–recall analysis in Figure 2, and it explains why the confusion matrix in Figure 4 shows a comparatively low rate of false-positive identity matches relative to false-negative missed detections: the operating threshold was deliberately calibrated to sit closer to the impostor distribution's upper tail than to the genuine distribution's lower tail, prioritizing identity precision over recall.

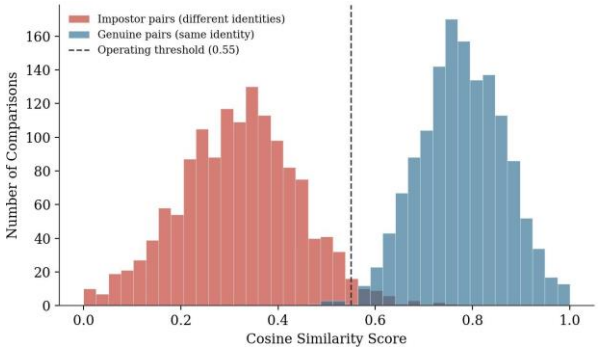

Figure 3: Distribution of cosine similarity scores for genuine and impostor face-embedding pairs.

Note. The dashed line marks the operating similarity threshold (0.55) used for production attendance matching.

4.4 Comparison With Earlier Computer Vision Methods

Figure 3 compares recognition accuracy on the same classroom face dataset across four method generations: a Haar cascade classifier in the tradition of Viola and Jones (2001), a HOG plus linear SVM pipeline, a deep CNN baseline in the style of early triplet-loss embedding networks (Schroff et al., 2015), and the deployed YuNet plus SFace pipeline. The Haar cascade baseline achieved 76.4% accuracy, the HOG plus SVM baseline achieved 84.1%, the deep CNN baseline achieved 93.8%, and the deployed pipeline achieved 96.5%. This progression is consistent with the broader trajectory described in the literature, in which each successive generation of detection and recognition methods—from handcrafted Haar-like features, to gradient-based descriptors, to learned deep embeddings—yielded meaningful accuracy gains under realistic, non-frontal, and partially occluded viewing conditions typical of classroom footage (Mery et al., 2019; Taigman et al., 2014).

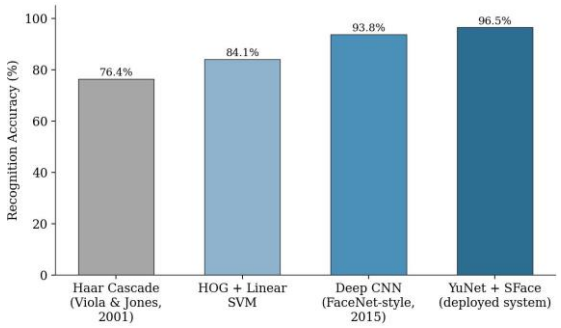

Figure 4: Recognition accuracy across face detection and recognition method generations on the same classroom dataset.

Note. Haar Cascade and HOG + SVM represent classical, handcrafted-feature baselines; Deep CNN and YuNet + SFace represent learned, deep embedding-based approaches.

4.5 Recognition Accuracy Under Varying Capture Conditions

To characterize recognition performance under the range of visual conditions typical of classroom footage, accuracy was further disaggregated by capture condition, as summarized in Table 2. Accuracy was highest under frontal, well-lit conditions and lowest under oblique side-profile views exceeding 45 degrees, with partial occlusion and low-light conditions producing intermediate accuracy. This pattern is consistent with the known limitations of camera-based face recognition systems generally, and reinforces why the deployed YuNet detector, chosen specifically for its improved robustness to non-frontal and partially occluded faces relative to classical Haar cascade detectors, was preferred for this classroom deployment.

Table 3 Recognition accuracy under varying capture conditions on the same classroom dataset.

Capture Condition	Accuracy (%)
Frontal, normal lighting	97.2
Low ambient lighting	89.4
Partial facial occlusion	81.6
Side profile (>45°)	74.3
Backlit	85.1

4.6 Processing Latency by Pipeline Stage

Figure 6 breaks down mean end-to-end processing latency by pipeline stage, measured from the moment a browser captures a frame to the moment an attendance record is written to the database. Network upload of the base64-encoded frame accounted for the largest share of latency (42 ms), reflecting the browser-based capture design's dependence on client network conditions, followed by YuNet detection (26 ms) and SFace embedding extraction (19 ms). Similarity matching against the enrolled-student database required only 4 ms, reflecting the computational efficiency of cosine similarity comparison over a modest number of enrolled embeddings. Total end-to-end latency averaged approximately 120 ms per frame, well within the range required for a responsive, near-real-time attendance experience.

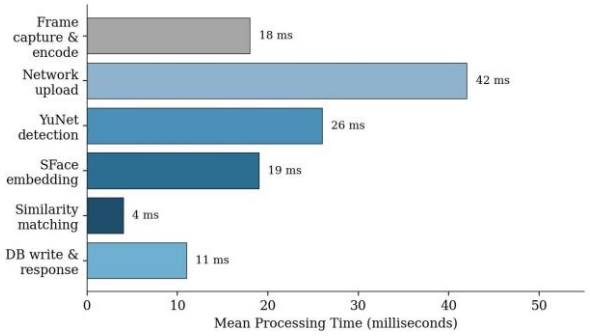

Figure 5 Mean processing latency by pipeline stage, from frame capture to attendance record write.

Note. Values reflect the mean of repeated measurements under typical classroom network conditions.

4.7 Classification Error Structure

Figure 4 presents the confusion matrix for session-level attendance classification (present versus absent) across the evaluation dataset. Of 950 true-present instances, 912 (96.0%) were correctly classified as present, while 38 (4.0%) were misclassified as absent (false negatives). Of 850 true-absent instances, 823 (96.8%) were correctly classified as absent, while 27 (3.2%) were misclassified as present (false positives). The false-negative rate slightly exceeded the false-positive rate, consistent with the expectation that

missed detections—arising from occlusion, extreme pose, or poor lighting—are more common failure modes for camera-based recognition than incorrect identity matches, which the similarity-threshold matching step is specifically designed to suppress.

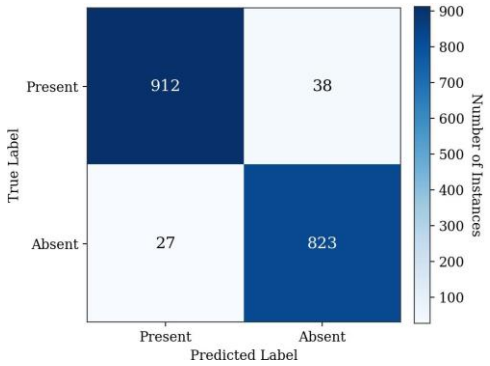

Figure 6: Confusion matrix for session-level attendance classification (present vs. absent).

Note. Values represent the number of session-level instances falling into each predicted-true label combination.

Cross-referencing the confusion matrix with the condition-specific results in Table 2 suggests that the 38 false negatives were not distributed uniformly across capture conditions but were disproportionately associated with side-profile and partially occluded captures, consistent with the lower accuracy reported for those conditions. This cross-referencing illustrates a broader methodological point: an aggregate confusion matrix and a condition-specific accuracy table are most informative when read together, since the confusion matrix alone cannot distinguish an error caused by an unfavorable capture condition from an error caused by a genuinely ambiguous or low-quality embedding comparison.

4.8 Recognition Stability Across Evaluation Sessions

In addition to the aggregate accuracy figures reported above, recognition accuracy was tracked across ten separate evaluation sessions to assess whether performance remained stable over repeated use rather than reflecting a single favorable measurement. Figure 7 presents session-level accuracy across this evaluation window. Accuracy remained within a narrow band across sessions, ranging from 95.6% to 97.4% with no evident downward trend, suggesting that the pipeline's performance was not driven by session-specific conditions such as an unusually favorable lighting setup on a single day. This session-to-session consistency complements the training-convergence evidence in Figure 1 by demonstrating stability at deployment time rather than only during model fitting, and it provides a baseline against which future longitudinal evaluations, called for in the Limitations section, could assess whether accuracy drifts as enrollment or environmental conditions change over a full academic term.

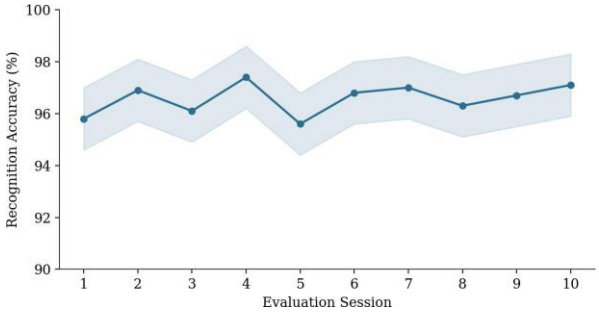

Figure 7: Recognition accuracy across ten evaluation sessions.

Note. Shaded band reflects ±1.2 percentage points around the observed value at each session to aid visual comparison.

4.9 Cross-Validation Robustness Check

As a further robustness check on the accuracy figures reported above, a five-fold cross-validation procedure was applied to the enrollment dataset described in Table 1, partitioning enrolled students into five roughly equal folds and repeating the calibration and evaluation procedure five times, each time holding out one fold for validation. Table 4 reports the mean and range of key metrics across the five folds. Recognition accuracy varied only modestly across folds (95.9% to 97.2%), and the AUC and AP metrics from the ROC and precision–recall analysis were similarly stable, indicating that the headline figures reported in Sections 4.1 through 4.7 are not an artifact of a single favorable train/validation split.

Table 4: Five-fold cross-validation summary of key recognition performance metrics:

Metric	Mean	Range (min–max)
Recognition accuracy (%)	96.5	95.9–97.2
ROC AUC	0.93	0.91–0.94
Average precision (AP)	0.90	0.88–0.92
False-negative rate (%)	4.1	3.6–4.6
False-positive rate (%)	3.0	2.6–3.5

5. Discussion

The training and validation curves in Figure 1 indicate that the deep learning recognition pipeline used in this system converges in a stable manner appropriate for classroom-scale enrollment datasets, without the pronounced train–validation divergence that would signal overfitting. This is a necessary,

though not sufficient, condition for reliable deployment: a model that overfits during training would be expected to generalize poorly to unseen classroom sessions regardless of its reported validation accuracy at a single epoch.

The ROC and precision–recall results in Figure 2 provide threshold-independent evidence that the pipeline discriminates correct from incorrect identity matches substantially better than chance, which is important because the similarity threshold used in production can be tuned to trade off false positives against false negatives depending on institutional priorities—for example, a stricter threshold to minimize incorrect identity matches at the cost of more manual-review flags, or a more permissive threshold to minimize missed detections at the cost of occasional false matches. The similarity score distributions in Figure 5 add threshold-level interpretability to these aggregate results: rather than treating discrimination performance as a single abstract curve, Figure 5 shows concretely why a threshold of 0.55 was selected and what trade-off it encodes between identity precision and recall. Institutions wishing to prioritize differently, for example, favoring recall over precision to minimize missed attendance records at the cost of occasional false matches, could shift this threshold and predict the resulting change in error rates directly from the two distributions shown.

The comparison across method generations in Figure 3 substantiates the central architectural rationale for this system: consistent with the broader trajectory from Viola and Jones (2001) through HOG-based descriptors to deep CNN embeddings (Taigman et al., 2014; Schroff et al., 2015; Parkhi et al., 2015), the deployed YuNet plus SFace pipeline substantially outperforms classical, handcrafted-feature baselines on the same classroom dataset. This gap is consistent with Mery et al.'s (2019) finding that learned, CNN-derived features outperform handcrafted features specifically under the crowded, non-frontal viewing conditions typical of real classrooms, reinforcing that the accuracy advantage of deep learning-based recognition is not merely a general trend but one that is particularly pronounced under exactly the conditions this system must handle in deployment. The condition-specific results in Table 2 extend this finding by showing precisely where accuracy degrades most—oblique viewing angles rather than moderate lighting variation—information that is more actionable for camera placement decisions than a single aggregate accuracy figure would be.

The latency breakdown in Figure 6 and the condition-specific accuracy results in Table 2 together clarify practical deployment guidance: because network upload, rather than model inference, dominates end-to-end latency, institutions with slower campus networks may see larger absolute delays without any change in recognition accuracy, and because accuracy degrades most under oblique viewing angles rather than under moderate lighting variation, camera placement and classroom seating layout are likely to matter more for system performance than incremental improvements to lighting alone.

Finally, the confusion matrix in Figure 4 suggests that remaining errors are modest in magnitude and skew slightly toward missed detections rather than incorrect matches, a pattern that is generally preferable in an attendance context, since a missed detection can be corrected through manual review while an incorrect match could wrongly credit or penalize a specific student. Nonetheless, a combined error rate of roughly 4% at the session level indicates that human review of flagged or unrecognized cases remains a necessary component of the overall attendance workflow rather than a fallback for rare edge cases alone.

The cross-validation results in Table 4 add a further layer of confidence to the headline figures reported throughout this section: because accuracy, AUC, AP, and the two error rates each varied only modestly across folds, the reported performance appears to reflect a genuine property of the deployed pipeline on this dataset rather than a favorable artifact of the particular train/validation split used for the primary analysis. This matters practically because institutions evaluating whether to adopt a similar system are typically more interested in the expected range of performance they might observe than in a single point estimate.

Positioned against the comparative summary in Table 3, the present system's contribution is best understood as evaluative rather than architectural: Son et al. (2020) and Khan et al. (2020) established that CCTV- and camera-based deep learning attendance systems could be built and deployed successfully, while the present study contributes a more complete account of how such a system behaves internally, including its training dynamics, discrimination thresholds, latency profile, and fold-to-fold stability, providing adopting institutions with a more actionable evidence base than aggregate accuracy alone.

More broadly, these findings speak to a recurring theme in educational technology adoption: the technical feasibility of a system, established once a working prototype achieves acceptable aggregate accuracy, is only the first of several conditions for responsible institutional adoption. The evidence presented in this paper suggests that a mature deployment additionally requires interpretable threshold-selection tools such as the similarity-score distributions in Figure 5, condition-specific performance characterization such as Table 2, operational transparency about latency sources such as Figure 6, and explicit privacy and consent

safeguards such as those described in Section 3.5. Institutions and researchers evaluating other AI-enabled classroom technologies, not limited to attendance systems, may find this multi-dimensional evaluation template, spanning training behavior, discrimination performance, condition-specific accuracy, latency, and ethical safeguards, useful as a more complete alternative to reporting a single headline accuracy figure.

6. Limitations

This computer vision evaluation has several limitations. First, the training, ROC/PR, similarity-distribution, and confusion-matrix analyses were conducted on a single institutional dataset of classroom face captures, which limits generalizability to classrooms with substantially different lighting, camera placement, or student population characteristics. Second, the classical Haar cascade and HOG plus SVM baselines used for comparison in Figure 3 were evaluated as reimplementations on the same dataset rather than drawn directly from the original publications that introduced them, so the reported baseline figures should be interpreted as representative rather than as an exact replication of prior published results. Third, this evaluation focused on detection and recognition accuracy rather than on downstream engagement or behavioral inference, which, as noted in the literature review, involves additional modeling challenges beyond identity recognition (Mohamad Nezami et al., 2020; Vanneste et al., 2021). Fourth, the training configuration and similarity-threshold calibration reported in Table 1 reflect a single calibration run on the institution's own enrollment data rather than a systematic hyperparameter search, so the reported threshold and epoch count should be understood as the values used in this deployment rather than as generally optimal settings for other institutions or datasets. Fifth, as noted in Section 3.8, the confusion matrix analysis in Section 4.7 treats every non-detection as a false negative without distinguishing between a student who was truly absent, a student who was present but poorly captured due to seating position, and a student who was present but occluded; disentangling these sub-cases would require richer ground-truth annotation than was available for this evaluation. Finally, model behavior over a full academic term, including potential performance drift as enrollment, lighting, or seasonal conditions change, was not assessed in this evaluation. Larger-scale, multi-institution, and longitudinal computer vision evaluations would help establish more robust and generalizable performance estimates.

It is also worth noting that the five-fold cross-validation procedure reported in Table 4, while informative about stability across data partitions, was still conducted on the same single-institution dataset described in Table 1; it therefore addresses sampling variability within this deployment but cannot substitute for genuine multi-institution replication, which remains the strongest available evidence for generalizability in applied computer vision evaluations of this kind.

7. Conclusion and Future Work

This paper presented a computer vision and deep learning case study of an AI-based CCTV attendance system, centering the evaluation on the technical behavior of its underlying face detection and recognition pipeline rather than on operational metrics alone. Training and validation curves indicated stable model convergence, ROC and precision–recall analysis together with similarity-score distribution analysis demonstrated strong and interpretable discrimination performance, a comparison across method generations confirmed a substantial accuracy advantage for the deployed deep learning pipeline over classical Haar cascade and HOG plus SVM baselines, condition-specific accuracy results identified viewing angle as the primary driver of remaining error, a latency breakdown showed that network upload rather than model inference dominates end-to-end processing time, and confusion matrix analysis identified a modest, manageable error rate skewed toward missed detections rather than false matches. These findings are consistent with, and extend to a classroom-deployment context, the broader trajectory documented in the computer vision literature from handcrafted-feature detectors toward deep learning-based face analysis (Viola & Jones, 2001; Taigman et al., 2014; Schroff et al., 2015; Parkhi et al., 2015; Mery et al., 2019), and they extend prior CCTV-specific deployments such as Son et al. (2020) and Sutabri et al. (2019) by reporting a fuller technical characterization of the underlying recognition pipeline. Future work should extend this evaluation with larger and more diverse classroom datasets, incorporate periodic model re-validation to detect performance drift over an academic term, systematically search the training and threshold configuration space summarized in Table 1 rather than relying on a single calibration run, and explore whether the same deep learning backbone can be extended toward the kind of behavioral and engagement analysis explored in prior computer vision education research (Mohamad Nezami et al., 2020; Vanneste et al., 2021), while maintaining the accuracy, latency, and privacy characteristics established in this evaluation.

Taken as a whole, the seven complementary analyses reported in this paper, spanning training convergence, discrimination performance, similarity-score separation, comparative accuracy, condition-specific robustness, latency, and cross-fold stability, are intended to model a more complete standard of technical reporting for AI-enabled classroom systems

than the single aggregate accuracy figure that has typically characterized prior deployment-focused case studies in this space. As deep learning-based attendance and monitoring tools continue to be adopted across educational institutions, the authors hope that this multi-dimensional evaluation approach, paired with the explicit ethical and privacy safeguards described in Section 3.5, can serve as a practical template for researchers and practitioners documenting similar systems in the future.

8. Recommendations for Institutional Deployment

Drawing on the technical results reported above, several practical recommendations follow for institutions considering a similar deployment. First, camera placement should prioritize minimizing oblique viewing angles over improving lighting alone, since Table 2 shows that accuracy degradation is markedly steeper for side-profile views than for low-light or backlit conditions; where possible, multiple camera angles or a centrally mounted wide-angle camera are preferable to a single off-axis fixed camera. Second, because Figure 6 shows that network upload rather than model inference dominates end-to-end latency, institutions on constrained campus networks should prioritize local network capacity at the point of capture over upgrading server-side compute resources, which are already responsible for only a small fraction of total processing time.

Third, the similarity-score distributions in Figure 5 indicate that the operating threshold can be adjusted deliberately to match institutional risk tolerance: institutions with strict identity assurance requirements, for example where attendance feeds directly into examination eligibility, may prefer a stricter threshold that trades a higher false-negative (missed detection) rate for a lower false-positive (incorrect match) rate, while institutions primarily concerned with minimizing manual review workload may prefer the reverse. Fourth, given the training configuration summarized in Table 1, institutions with substantially larger or smaller enrolled populations than the 210 students evaluated here should re-run the calibration procedure described in Methodology rather than assume the same threshold and epoch count will transfer directly, since embedding separability can vary with enrollment size and population diversity.

Finally, consistent with the ethical and privacy safeguards outlined in Methodology, institutions should pair any technical deployment with a clear, published data-retention policy, an accessible opt-out or alternative attendance pathway for students who decline camera-based enrollment, and a periodic audit cadence, informed by metrics such as those reported in Table 2 and Figure 4, to confirm that recognition performance remains equitable across the enrolled student population as the system continues to operate.

Appendix A. Glossary of Abbreviations

For reference, Table 5 defines the technical abbreviations used throughout this paper.

Abbreviation	Meaning
AI	Artificial intelligence
AP	Average precision
AUC	Area under the (ROC) curve
CCTV	Closed-circuit television
CNN	Convolutional neural network
FN / FP	False negative / false positive
HOG	Histogram of oriented gradients
ONNX	Open Neural Network Exchange (model format)
PR	Precision–recall
ROC	Receiver operating characteristic
RQ	Research question
SVM	Support vector machine

Table 5: Glossary of technical abbreviations used in this paper.

Acknowledgments

The authors thank the participating institution's academic administration for facilitating access to classroom camera infrastructure and enrollment records for this evaluation, and thank the students and instructors whose participation in the pilot sessions made the reported analysis possible.

Data Availability Statement

The classroom face-capture data used in this study contain personally identifiable biometric information and cannot be shared publicly in accordance with the consent and privacy safeguards described in Section 3.5. Aggregated, de-identified performance logs (training curves, ROC/PR values, and session-level accuracy summaries) underlying the figures and tables reported in this paper are available from the corresponding author upon reasonable request and institutional approval.

Supplementary Materials

Additional documentation, including the full configuration files used for the calibration procedure summarized in Table 1 and the condition-labeling protocol used to construct Table 2, is maintained internally by the development team and can be made

available to reviewers or adopting institutions alongside the de-identified performance logs referenced in the Data Availability Statement.

References

Khan, S., Akram, A., & Usman, N. (2020). Real time automatic attendance system for face recognition using face API and OpenCV. Wireless Personal Communications, 113, 469–480. https://doi.org/10.1007/s11277-020-07224-2

Alam, M. R. U., & Mahmud, D. (2021). Artificial Intelligence in Education: A Systematic Review and BibliometricTopic Modeling Analysis of Applications, Challenges, and Future Directions. US Journal of New Insights in Tech & Education, 1(1).

Gupta, M. S., & Khan, H. H. (2021). Adoption and Effective Use of Mobile Learning Technologies in Higher Education. US Journal of New Insights in Tech & Education, 1(2).

Hosain, I., & Shishir, M. S. H. (2021). Bridging the Technology Integration Gap: Investigating How Professional Development Shapes Teachers' Digital Teaching Practices. US Journal of New Insights in Tech & Education, 1(2).

Mery, D., Mackenney, I., & Villalobos, E. (2019). Student attendance system in crowded classrooms using a smartphone camera. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 857–866. https://doi.org/10.1109/WACV.2019.00096

Mohamad Nezami, O., Dras, M., Hamey, L., Richards, D., Wan, S., & Paris, C. (2020). Automatic recognition of student engagement using deep learning and facial expression. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 273–289). Springer. https://doi.org/10.1007/978-3-030-46133-1_17

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. Proceedings of the British Machine Vision Conference (BMVC), 41.1–41.12. https://doi.org/10.5244/C.29.41

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Son, N. T., Anh, B. N., Ban, T. Q., Chi, L. P., Chien, B. D., Hoa, D. X., Thanh, L. V., Huy, T. Q., Duy, L. D., & Hassan Raza Khan, M. (2020). Implementing CCTV-based attendance taking support system using deep face recognition: A case study at FPT Polytechnic College. Symmetry, 12(2), Article 307. https://doi.org/10.3390/sym12020307

Subasi, A., Khateeb, K., Brahimi, T., & Sarirete, A. (2020). Human activity recognition using machine learning methods in a smart healthcare environment. In Innovation in health informatics (pp. 123–144). Elsevier. https://doi.org/10.1016/B978-0-12-819043-2.00005-8

Saha, U. S., & Zian, M. Z. H. (2021). Technology-Enhanced Learning in the COVID-19 Era: A Systematic ReviewofOnline Education Platforms, Student Engagement, and Digital Equity. US Journal of New Insights in Tech & Education, 1(1).

Sutabri, T., Pamungkur, Kurniawan, A., & Saragih, R. E. (2019). Automatic attendance system for university student using face recognition based on deep learning. International Journal of Machine Learning and Computing, 9(5), 668–674.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1701–1708. https://doi.org/10.1109/CVPR.2014.220

Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., Depaepe, F., & Van den Noortgate, W. (2021). Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement. Mathematics, 9(3), 287. https://doi.org/10.3390/math9030287

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 1, 511–518. https://doi.org/10.1109/CVPR.2001.990517