Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 31/20/2024; Camera ready: 20/02/2025; 1st round notif.: 22/01/2025; Edition review: 16/05/2025; New version: 05/02/2025; Available online: 16/05/2025; 2nd round notif.: 26/02/2025; Published: 16/05/2025;

End-to-end automated student attendance recording system using surveillance camera

John Davi Dutra Canuto Pires
Universidade Federal de Alagoas
ORCID: 0009-0009-3236-497X
jddcp@ic.ufal.br

Marcelo Costa Oliveira
Universidade Federal de Alagoas
ORCID: 0000-0002-0825-6217
oliveiramc@ic.ufal.br

Baldoino Fonseca dos Santos Neto
Universidade Federal de Alagoas
ORCID: 0000-0002-0730-0319
baldoino@ic.ufal.br

Márcio de Medeiros Ribeiro
Universidade Federal de Alagoas
ORCID: 0000-0002-4293-4261
marcio@ic.ufal.br

Rafael Sampaio de Melo Fragoso
Sistema FIEA
ORCID: 0009-0002-1807-1758
rafael.fragoso@sistemafiea.com.br

Resumo

A prática de registrar e monitorar a frequência dos alunos é uma ação fundamental em vários contextos, especialmente no ambiente escolar. No entanto, devido ao seu processo manual, muitas vezes consome uma parte significativa do tempo de aula. Este artigo apresenta um sistema de registro automatizado de frequência de alunos de ponta a ponta utilizando câmeras de vigilância presentes na sala de aula. O sistema usou tecnologias avançadas, como visão computacional, reconhecimento facial e técnicas de Rede Neural Convolucional (CNN) para simplificar o processo de frequência e melhorar a distribuição do tempo em sala de aula. O estudo avalia o desempenho de cinco sessões de aula de um conjunto de dados público e três sessões de aula coletadas por nós na escola estudada neste artigo em diferentes cenários, com qualidade de imagem e posicionamento do aluno variados. Nossos resultados destacam a superioridade do ResNet29 na detecção e no reconhecimento de alunos, especialmente em imagens de baixa resolução. Em comparação com os modelos Facenet512, Facenet e ArcFace, em termos de precisão final da marcação de frequência, o modelo apresentou resultados métricos superiores, com um aumento de 50% da precisão em comparação com os outros, atingindo 80% de precisão, além de demonstrar ser superior nas métricas de precisão, recuperação, F1-score e AUC-ROC. A implantação do sistema em um ambiente escolar apresentou resultados promissores, o que levou a planos de expansão para outras salas de aula. A natureza leve e não intrusiva do sistema está alinhada com o conceito de salas de aula inteligentes de última geração, enfatizando seu potencial para revolucionar o gerenciamento de presença em instituições educacionais.

Palavras-chave: *Infraestrutura tecnológica e conectividade; análise visual; visão computacional; reconhecimento de faces; detecção de faces; redes neurais profundas; sistemas automatizados; vigilância por vídeo; sistemas em tempo real*

Abstract

The practice of recording and monitoring student attendance is a fundamental action in various contexts, especially in the school environment. However, due to its manual process, it often consumes a significant part of class time. This paper presents an end-to-end automated student attendance recording system utilizing classroom surveillance cameras. The system used advanced technologies such as computer vision, face recognition, and Convolutional Neural Network (CNN) techniques to streamline the attendance process and enhance the time distribution in the classroom. The study evaluates the performance of five class sessions from a public dataset and three class sessions

Cite as: Pires, J. D. D. C., Oliveira, M. C., Santos Neto, B. F., Ribeiro, M. M. & Fragoso, R. S. M. (2025). End-to-end automated student attendance recording system using surveillance camera. Revista Brasileira de Informática na Educação, vol. 33, 371–393. https://doi.org/10.5753/rbie.2025.5125.

collected by us at the school studied in this paper through different scenarios with varying image quality and student positioning. Our results highlight the superiority of ResNet29 in detecting and recognizing students, especially in lower-resolution images. Compared to Facenet512, Facenet, and ArcFace models, in terms of final frequency marking accuracy, the model showed superior metric results by having an increase of 50% of accuracy compared to the others, reaching 80% accuracy, as well as demonstrating being superior in precision, recall, F1-score, and AUC-ROC metrics. The system's deployment in a school setting has shown promising results, prompting plans for expansion to additional classrooms. The lightweight and non-intrusive nature of the system aligns with the concept of Next-Generation Smart Classrooms, emphasizing its potential to revolutionize attendance management in educational institutions.

Keywords: *Technological infrastructure and connectivity; Visual Analysis; Computer Vision; Face Recognition; Face Detection; Deep Neural Networks; Automated Systems; Video Surveillance; Real-time Systems*

1 Introduction

School attendance is a fundamental aspect of daily routines for teachers and students, serving as a record of presence and a crucial indicator of student performance. Regular attendance often correlates with academic success, engagement, and commitment, making it a valuable criterion in educational assessment. Patterns in attendance data can reveal insights into student engagement levels and potential external factors affecting learning, thus supporting informed interventions to enhance educational outcomes. Therefore, attendance is no longer just a record but one of the most accurate pieces of information for analyzing students' performance (Adelman, 2006). Although still widely used, manual attendance recording presents challenges such as time consumption, the potential for human error, and difficulties in immediately integrating this data with other relevant data information, like student performance and attention span (Ashfaq et al., 2023).

Automating the attendance process offers numerous advantages, such as minimizing errors caused by misrecording or proxy attendance—when someone marks attendance on behalf of another—ensuring authenticity in attendance records, and freeing up time for teachers to focus on pedagogical activities (Decoito & Richardson, 2018). Additionally, it supports the vision of technology-integrated education, fostering the concept of Next-Generation Smart Classrooms presented by (Uskov et al., 2015) and applied in systems like CODEX (Norton et al., 2024) and the work from (Costa & Guedes, 2022).

Nowadays, many schools are implementing additional security measures to ensure the safety of their students. One such measure is installing surveillance cameras, which monitor the school's perimeter and classrooms for suspicious activities, potential threats, or issues of student misconduct (Ghimire & and, 2023). Moreover, surveillance cameras installed in the classroom open new opportunities to automate school attendance. Leveraging an existing classroom tool, such as the surveillance camera, aims to enhance its role in supporting school operations. Allowing cameras to manage attendance streamlines the process, helping to maintain an environment of trust while also saving class time and giving teachers more opportunities to focus on their lesson plans. For this type of automation to be effective, however, it is essential to implement it ethically and with respect for privacy, particularly when handling sensitive personal data, in compliance with regulations such as Brazil's General Data Protection Law (LGPD).

Machine learning (ML) is a promising solution for automating school attendance. Previous works, such as those by Gupta et al. (Gupta et al., 2018) and Sreedevi and Ranjith Ram (Sreedevi & Ram, 2019), have demonstrated the potential of integrating ML solutions with surveillance cameras in educational settings for more efficient monitoring and attendance control. Furthermore, Filho et al. (Lavareda Filho et al., 2022) showed a practical application of face recognition systems in a school environment. However, these solutions still require improvements, including challenges in dealing with low-quality video sources, unsuitable lighting conditions, privacy issues, and ethical concerns.

In this context, the main goal of our work was to address these limitations by developing an end-to-end automated student attendance system using Deep Neural Networks and surveillance cameras. Our system is designed to operate in practical environments with raw lighting and low-resolution conditions, handling detection and recognition tasks while generating attendance reports and managing class schedules. At the same time, build the idea of a future classroom, the Next-Generation Smart Classrooms (Uskov et al., 2015).

The leading contributions of this work include developing an automated attendance system integrated with existing surveillance cameras to minimize additional hardware costs; besides this, the system is designed with a practical approach, focusing on real-world classroom environments with not ideal video quality conditions. The system achieves 80% in accuracy and F1-Score performing the attendance in challenging video conditions. With the entirety of the structure around the model being built by us, the organization and speed achieved are improved, allowing attendance to be realized in real-time. Together with all this, we care about privacy and security challenges associated with using AI for student monitoring, ensuring that ethical considerations are considered throughout the implementation. The study lays a foundation for future advancements in creating Next-Generation Smart Classrooms by addressing these practical and ethical challenges.

2 Related Works

Kaliappan et al. (Kaliappan et al., 2019) have shown a system that resignifies surveillance cameras inside a school perimeter to a smart camera that can perform face detection and face recognition. Their approach uses three technologies: Haarcascades, Triplet Loss, and FaceNet (Schroff et al., 2015). The detection and recognition are realized to register attendance and improve energy saving by turning off the power supply when no student is detected. The counterpart of this approach, even though it uses Deep Learning, is the need for more than 50 images per student, making it difficult to apply in an entire school because of the exhaustive process aiming to collect 50 images to get good results.

Gupta et al. (Gupta et al., 2018) presented the idea of a system where surveillance cameras perform attendance through face recognition using neural networks leveraging insights from Next-Generation Smart Classrooms to enhance scalability and adaptability. At the same time, Sreedevi and Ranjith Ram (Sreedevi & Ram, 2019) proposed a non-intrusive system based on face recognition using the Viola-Jones algorithm and Principal Component Analysis (PCA) with Eigenvalues; both systems focus on saving time and reducing administrative burden and have shown the same flaws: applying the solution in a prepared environment with a high-definition camera and a smaller classroom, with fewer students in a small space, not addressing the major situation in most classrooms.

According to Serengil and Ozpinar (2020) (Serengil & Ozpinar, 2020) work, the most accurate model involving face recognition is FaceNet512 (Firmansyah et al., 2023), a variation of the model FaceNet, developed by Google Inc. FaceNet (Schroff et al., 2015) proposes a solution based on CNN and in the Inception architecture (Szegedy et al., 2015). The original Facenet study creates 128-dimensional vectors, but the FaceNet512 approach, developed by David Sandberg, is an extended version of Facenet, creating 512 dimensions. Its neural network consists of a CNN that uses Adam as the optimizer, Triplet Loss as a facial verification power, and Euclidean distance calculation as a measure of face similarity. Its most expressive results were obtained using the Labeled Faces in the Wild (LFW) database (Huang et al., 2008). The main disadvantage of this model is its application in a context; as a robust model, it can return optimal results but will consume more machine resources, making it very difficult to apply in an entire school system without spending significant capital to improve its capacity.

Sawal et al. (Sawall et al., 2021) presented a student attendance recording system using students' smartphones and the school's Wi-Fi network. This method establishes a system of zero effort and interaction, where the student's presence is confirmed through their connection to the school's access point. However, the system requires the student to maintain a constant connection to the school's Wi-Fi network, which can result in distractions during class. Furthermore, the system is vulnerable to cheating, as the smartphone only needs to be at the school to register attendance rather than the student who owns the smartphone.

In 2023, Chetty and Sharma (**attendance_fisherface**) proposed the FisherFace algorithm, among different tested algorithms, in addition to Fisher's Linear Discriminant (FLD) (Kar et al., 2012) approach to perform the attendance through a computer's webcam. When it comes to face recognition, using the framework of the pattern classification paradigm, the Fisherface algorithm tackles this problem by thinking of each pixel value in a sample image as a coordinate in a space with a high dimension. At the same time, the FLD works by optimizing the total scatter across all classes through linear discrimination. Under ideal conditions, this algorithm has a low error rate, but it relies on ideal scenarios to perform the attendance. The face detection algorithms this work uses are likely to fail under extreme illumination conditions and perform poorly when there is a preponderance of shadowing, indicating that the system cannot model or hide the shadowed parts.

The main problem of all these systems is to focus on a specific solution, not making them general enough by not dealing with questions like computational resources' consumption, use of robust techniques, and fail-safe logistics simultaneously.

3 Materials and Methods

3.1 System Overview

The pipeline of this work is shown in Figure 1. Each of its parts is described below:

- First, (A) an API deals with the attendance scheduling and classrooms, cameras, and data that should be analyzed by the model (Section 3.2).
- Next, (B) the student's data from *(school name - blinded)* is collected, including the students' faces (Section 3.3).
- Then, (C) the data from the surveillance cameras in classrooms are collected, and the faces of the students attending in the classroom are filtered using the face detection model (Section 3.4).
- Finally, step (D) represents the last stage, where student recognition and attendance recording occur. The camera captures images of students' faces, which are then processed by the face recognition model, ResNet29, converting each face into an embedding. These student embeddings are stored and used to compare with new face embeddings from each detected

frame. Ultimately, a comparison is made between the embeddings of known students and any newly detected faces to mark attendance (Section 3.5).

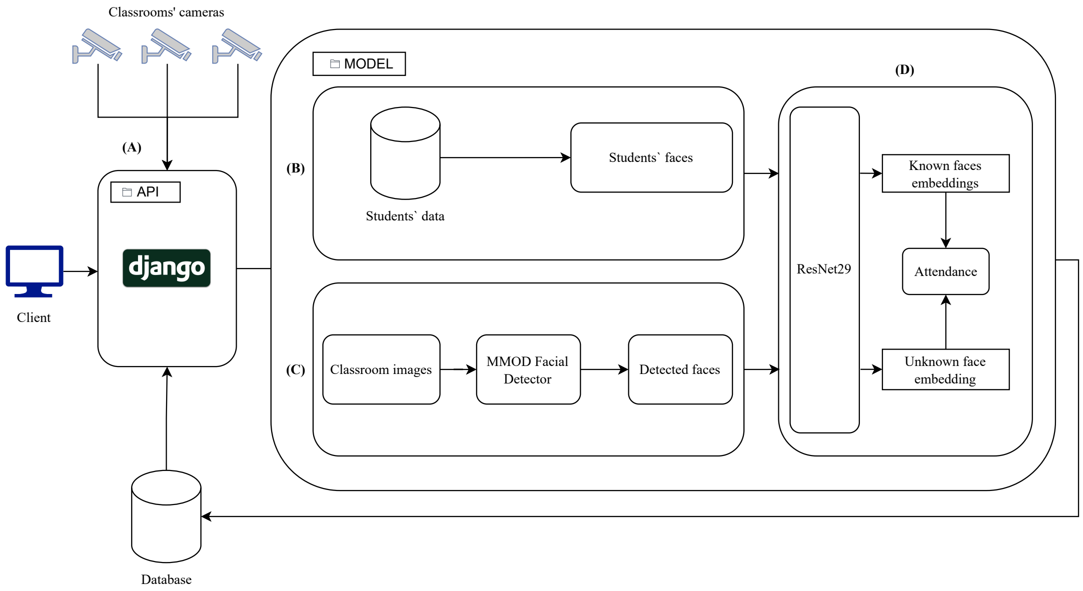


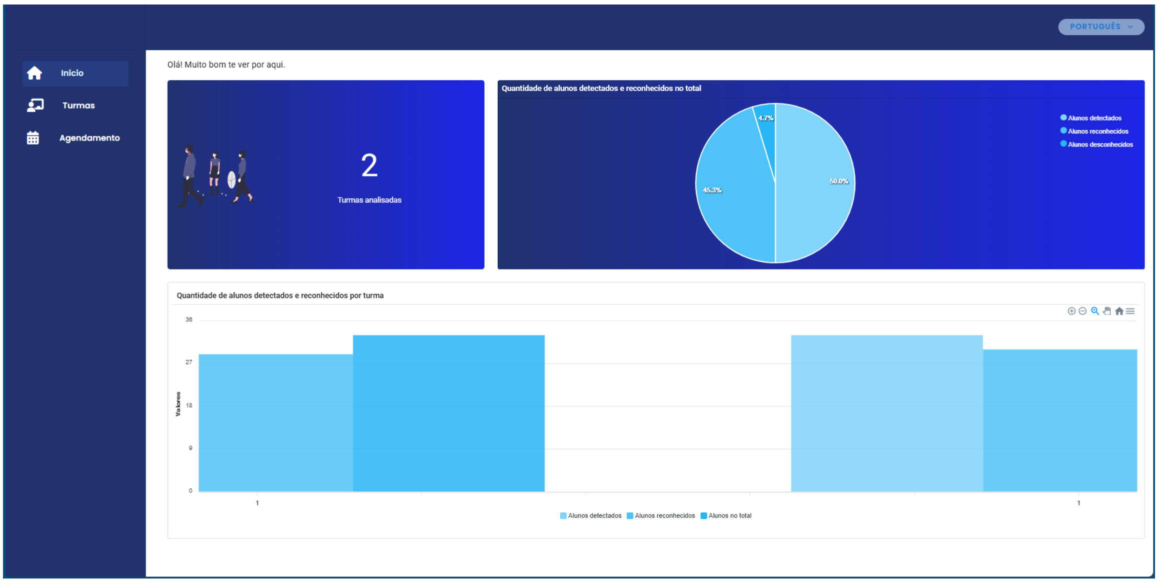
Figure 1: Schematic containing all the processes required to run the automatic frequency system.

3.1.1 *Ethical Background*

The first consideration when developing a system that utilizes facial recognition technology is obtaining explicit consent to use the images of the individuals involved. In this project, which was conducted in partnership with an educational institution, *(school name - blinded)*, using student data and images was strictly governed by the Brazilian General Data Protection Law (LGPD). Compliance was ensured through a consent clause, whereby parents or guardians formally authorized the collection and use of student's data and images. The entire application's structure was supported and integrated into the school's internal system and database, making it only accessible to the school's employees and internal collaborators, avoiding transferring sensible data and keeping the workflow internal.

3.2 API and Components Outline

One or two CCTV cameras have been installed in the classroom to ensure all students are visible without blind spots. These cameras are Clear CCTV L42 IP models with 2 megapixels. The API was connected to all the school cameras and was available through a Network Video Recorder (NVR). It can access every student's essential information, such as registration number, registration photo, name, and class number. The API's interface works as an attendance scheduler based on the input video captured in different classrooms. Once the classroom code is registered in our system, the API starts to track the classroom's image cameras. It retrieves the attendance predicted by the system to the *(school name - blinded)*'s database and returns statistical data to be displayed for the users, as can be seen in Figure 2.



Olá! Muito bom te ver por aqui.

Figure 2: Main page of the system in initial stages, showing the distribution of students and its states of recognition.

We worked with two classrooms: first, we used the *(school name - blinded)* database composed of 3 class sessions from 6th grade of elementary school, each with 60 frames and 34 students, resulting in 180 classroom images and 34 student images, all students with ages between 10 and 12. As a second database, we used the public data from Mery et al. (Mery et al., 2019), which comprises 5 class sessions, resulting in 45 classroom images with 67 students distributed in ages between 20 and 30. Both databases were verified regarding accuracy, precision, recall, F1-Score, and Area Under the Curve with Receiver Operating Characteristic (AUC-ROC).

Table 1: Table containing the data composition from the two sources analyzed in this work.

Source	Sessions	Number of students	Camera input size	Student input size
(school name - blinded)	3	34	2,304×1,296	200x200
Mery et al., 2019	5	67	2,000×2,000	150x150

3.3 Data composition

All the data was acquired by the pedagogical sector of the *(school name - blinded)*, which took care of the contractual procedures with the parents, registering the students' faces in the school's database at the time of enrollment and providing access to the surveillance cameras. The students' face images are captured at the start of each school year and linked to the students' registration records; all of it is stored in JPG format, RGB, 240 DPI, and have 5,472×3,648 of resolution, to be posteriorly cropped to 200×200, keeping all the original properties. In contrast, the classroom images were captured using the surveillance cameras present in the classrooms, returning images in 2,304×1,296 resolution, in JPG format, RGB, and having 96 DPI.

The data source is not limited to the RTSP protocol; it can also include raw video files or a sequence of individual frames. Figure 3 illustrates the camera placement in the classroom. The two images highlight the dual data streams captured, as two cameras are positioned to eliminate blind spots.

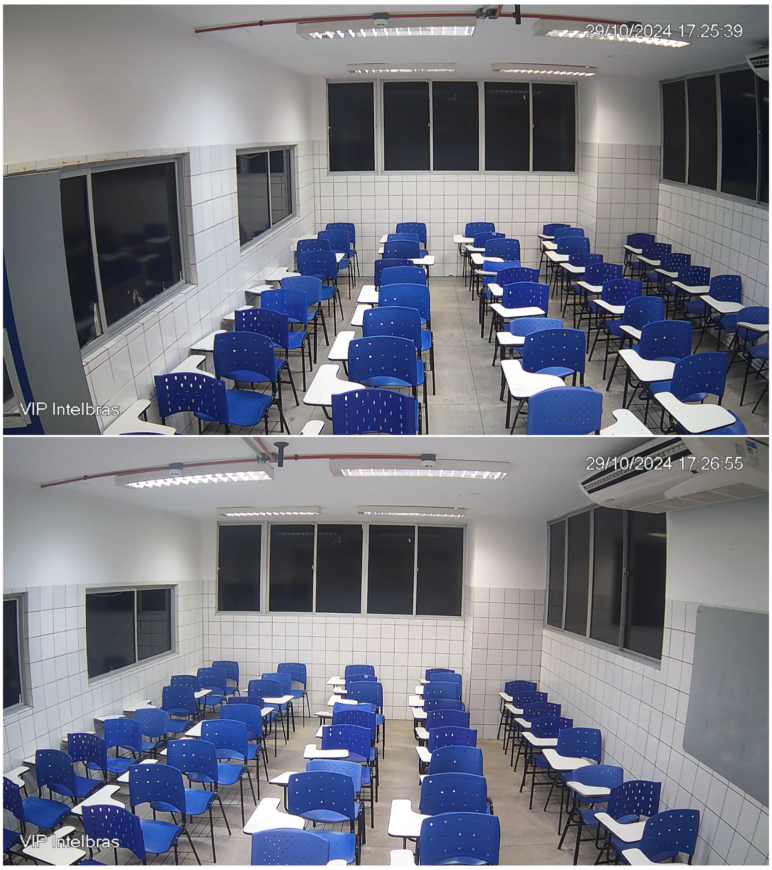


Figure 3: Classroom captures using a the surveillance cameras with 2,304×1,296 resolution present in *(school name - blinded)*.

For the model to achieve the detection and recognition actions for the frequency, the student images were organized to include at least one image containing the student's frontal face. This value is related to the attendance precision since more facial variations (frontal, profile, etc.) allow a wider recognition range. At the same time, the model receives camera images in video format via the RTSP protocol or from video files of the classroom surveillance cameras.

3.4 Face detection

In this stage, the students' faces from the surveillance camera are detected and gathered to be extracted, preceding the face recognition stage (Section 3.5). This stage is intrinsically connected to the surveillance camera flow since it is used in each frame coming from the cameras, aiming to find every student in the classroom frame. Not only this, but the detection has to be precise to allow the crop to capture the entire face of all students. The method used was Max-margin Object Detection (King, 2015), a CNN model which returns a matrix containing the bounding boxes delimiting the faces found, the classes detected, and the confidence score (Figure 4).

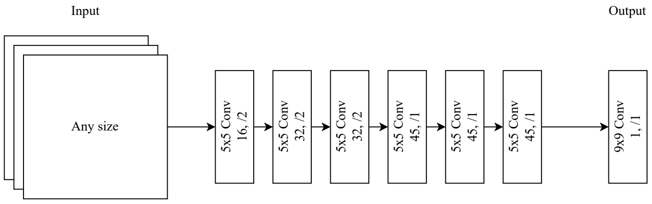


Figure 4: Structure of the convolutional neural network that performs face detection.

The Max-margin Object Detection, present in the dlib (King, 2009) library, was written in C++, trained and tested by Davis King (King, 2015) using the Face Detection Dataset and Benchmark (FDDB), which also serves as the ground truth for face detection, reaching 98.2% accuracy. The model contains seven convolutional layers, three of which are geared towards downsampling to reduce the dimensionality of the input images, and four common convolution layers that culminate in the delivery of the location coordinates in the image for the class found and an auxiliary metric of accuracy when detecting.

3.5 Face recognition

Face recognition models identify individuals in an image by pinpointing their unique characteristics. This makes face identification straightforward when the relevant features are accurately discerned. In this work, we utilized Deep Metric Learning to assess the similarity between samples, specifically using the Euclidean distance metric to measure feature vector similarity. Our approach leverages a ResNet-29 architecture trained with Triplet Loss, which ensures that embeddings of the same individual are closer together while pushing apart embeddings of different individuals. The model converts images into 128-dimensional feature vectors, embedding each face into a unique representation. To optimize training, we used the Adam optimizer and batch normalization. This process maps the facial characteristics of the person in the input data.

To enhance the model's ability to differentiate between individuals, training is performed by simultaneously analyzing triplets of face images composed of:

1. A training face image of a known person;
2. Another image of the same known person;
3. An image of a completely different person.

Next, the algorithm analyzes the measurements it is currently generating for each of these three images and slightly adjusts the neural network to ensure that the measurements generated for No. 1 and No. 2 are slightly closer and that the measurements for No. 2 and No. 3 are somewhat further apart.

For this to occur, we modified the standard ResNet-34 architecture (He et al., 2016), reducing its depth to 29 layers to balance accuracy and computational efficiency. The network was trained using Triplet Loss, which ensures that embeddings of the same individual are closer in feature space while pushing apart embeddings of different individuals.

The model processes images as follows:

- **Feature Extraction:** The input image (student's face) is passed through the modified ResNet-29 network, which extracts hierarchical facial features across multiple convolutional layers. While the final layer outputs a 128-dimensional embedding vector, which uniquely represents the student's face in a high-dimensional space.
- **Triplet Loss Training:** During training, the model optimizes embedding distances by analyzing triplets of images: Anchor (A), a known image of a student, Positive (P), another image of the same student, and Negative (N), an image of a different student. The network is trained to minimize the Euclidean distance $d(A,P)$ (similar faces) and maximize $d(A,N)$ (different faces), reinforcing intra-class compactness and inter-class separation.
- **Recognition Process:** At inference time, the system extracts feature vectors from: known images (enrollment photos, ground truth) stored in the database, and unknown images (faces detected in classroom surveillance footage). The extracted embeddings are compared using Euclidean distance. If the distance between a detected face and a stored embedding is below a set threshold τ, the system considers it a match, marking the student as present.

3.5.1 Model implementation and Loss function

Our model used the ResNet34 backbone, but we removed the final residual block, which consisted of six convolutional layers. Additionally, we reduced the number of filters per layer to decrease both training and inference costs. The resulting architecture contains 14 residual blocks (as can be seen in Figure 5), maintaining the skip connections to leverage residual learning while optimizing computational efficiency.

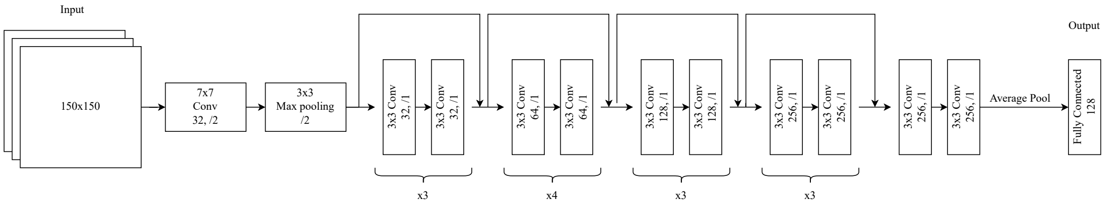


Figure 5: Structure of the convolutional neural network that performs facial recognition (ResNet29).

The model was trained and tested with the LFW (Huang et al., 2008) dataset, which encompasses 13,233 target face images containing 5,749 individual faces and is widely used in face recognition processes, serving as the ground truth for this task. The training stage of this model was run with its weights initialized randomly and a learning rate of 0.1 using the Stochastic Gradient Descent (SGD) optimizer. For each iteration, this learning rate decreases by 10% until it reaches a learning rate of 0.0001.

To achieve this accuracy value, triplet loss was used to minimize the Euclidean distance between the embeddings of the same person and, at the same time, maximize the same distance between the embeddings of different persons. This configuration obtained an accuracy value of 99.38%. The triplet loss is defined by the following structure presented in the equation 1:

$$Loss = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right] \quad (1)$$

where:

N : Total number of triplets in the batch

x_i^a : Anchor sample

x_i^p : Positive sample (same class as anchor)

x_i^n : Negative sample (different class from anchor)

$f(x)$: Feature extraction function (e.g., neural network)

$\|\cdot\|_2^2$: Squared L2 norm (Euclidean distance squared)

α : Margin that ensures separation between positive and negative pairs

Based on the Euclidean distance of the faces' embeddings, a threshold was used to check whether two faces are the same person. If the distance is below the threshold, the embeddings belong to the same person. Otherwise, they belong to different people. This threshold was initially defined as 0.62 based on experiments with the LFW (Huang et al., 2008) dataset.

Comparing distances with only one example per class works similarly to a *K-Nearest Neighbor* classifier, in which K is set to 1 (Fix & Hodges, 1989). Faced with a specific case where two individuals are below the established threshold, a method that can be applied to the context of face verification using this architecture is that if we have a situation where more than one student has a Euclidean distance less than the lower threshold for a specific identity, we assume that the student's real identity is when the distance is minimal (closer to zero).

3.6 Attendance algorithm

Automating attendance primarily aims to identify the most similar face embedding between the detected face and a pre-existing database of face embeddings. As shown in figure 1, the system operates in two stages of embedding creation.

The first stage involves processing and storing a sample of each student's face data by cropping, resizing the images, and aligning (rotating the faces to keep the eyes' line straight) them using facial landmark detection from Dlib (King, 2009). To optimize computational resources, the facial data of all students is collected, converted into embeddings, and stored statically, allowing it to be reused for subsequent class sessions.

With the students' face embeddings already stored, the system focuses solely on analyzing the live data stream from the classroom. It compares the detected faces in real-time with the stored embeddings, using the Euclidean distance algorithm, since it preserves a direct relationship with the difference in specific facial features that define a face. In the latent space, it effectively reduces the number of features and removes redundant information between similarity features (Zhou & Jiang, 2019).

The Euclidean distance algorithm is set with a predefined threshold acting as the recognition boundary, as outlined in equation 2.

$$Threshold \geq d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2}$$

where:

$$\begin{aligned} d(p,q) &: \text{Euclidean distance between two points } p \text{ and } q \\ n &: \text{Number of dimensions in the feature space} \\ p_i, q_i &: \text{Components of vectors (embeddings) } p \text{ and } q \text{ in the } i\text{-th dimension} \\ \sum_{i=1}^{n} &: \text{Summation over all } n \text{ dimensions} \end{aligned}$$

The Euclidean distance is calculated as the square root of the sum of squared differences between corresponding elements of the two embedding vectors. Mathematically, for two embeddings $e_{1,i}$ and $e_{2,i}$ (each with 128 dimensions) are the elements of the embeddings for the two faces. If the calculated distance is below the threshold, the embeddings are considered to represent the same person; otherwise, they are considered to represent different individuals. This threshold was initially set to 0.62 based on experiments with the LFW (Huang et al., 2008) dataset made by Adam Geitgey (Geitgey, 2017).

3.7 Experiment environment

The experiments were conducted in two distinct environments: personal and school computers, as outlined in table 2. The code was implemented in Python version 3.9 and executed within Docker containers to ensure consistency in the testing environment.

To automate the attendance process, we utilized Django with ApScheduler, a library that enables the scheduling of tasks. This setup allowed us to efficiently create and manage attendance schedules by automating key actions, ensuring a streamlined and reliable attendance monitoring process.

Table 2: Configuration of the environments used for the experiments.

Computer parts	Personal computer	School computer
Central Processing Unit (CPU)	3.9GHz, 6 cores	2.3GHz, 4 cores
Graphics Processing Unit (GPU)	12GB	None
RAM Memory	32GB	16GB

The choice of hardware configurations reflects the need to test the system under different performance constraints. The personal computer setup represents a scenario with higher processing power, while the school computer setup demonstrates the feasibility of running the system on lower-end hardware. This confirms that the system can operate effectively even with limited computational resources, making it suitable for institutions with budget constraints. Additionally, the use of Docker containers enhances portability and scalability, allowing easy deployment across multiple school environments without requiring extensive reconfiguration or high-end infrastructure.

3.8 Evaluation metrics

The following metrics were used to evaluate the models' performance on attendance execution: Accuracy (equation 3), Precision (equation 4), Recall, F1-Score (equation 5), and AUC-ROC.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4}$$

$$\text{F1-score} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = 2\times\frac{\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \tag{5}$$

These metrics were chosen to comprehensively assess the system's behavior in different classification scenarios, considering the ability to correctly identify students present and minimize false positives and negatives. The focus of these evaluation metrics is the AUC-ROC. Curves, which are generated to give a more detailed view of performance in terms of trade-offs between sensitivity and specificity, i.e., information about the true positive, equation 6, and false positive rates, equation 7.

$$TPR = Recall = \frac{TP}{TP+FN} = \frac{TP}{P} \tag{6}$$

$$FPR = \frac{FP}{TN+FP} = \frac{FP}{N} \tag{7}$$

True Positive Rate (TPR) is the rate of present subjects correctly detected, while False Positive Rate (FPR) is the rate of present subjects incorrectly detected. During discussions with the school board, it became clear that incorrectly marking a student as absent when they are present represents the most critical error. Therefore, TPR emerged as a significant metric, which can be seen by analyzing the y-axis of the AUC-ROC graph. Its significance arises from the fact that TPR is the complement of the False Negative Rate (FNR), as shown in equation 8. This highlights the importance of TPR, as it represents students who are present but marked as absent, making it the most important metric.

$$FNR = \frac{FN}{FN+TP} = \frac{FN}{P} = 1-TPR \tag{8}$$

Since there is no ground truth for each frame in a classroom video stream, we base our metrics on a comparison between the final attendance report generated by our system and the teacher's attendance report. All metrics are computed using this approach.

3.9 Execution Procedures

Two experiments were conducted: one using the dataset from (Mery et al., 2019) and other using the data we got from *(school name blinded)*. Three sessions were run using the *(school name* *blinded)* dataset, and five sessions were run using the data from (Mery et al., 2019). Each session involved testing the models Facenet512 (Firmansyah et al., 2023), Facenet (Schroff et al., 2015), Arcface (Deng et al., 2019), and our solution.

For each session, the models were tested on these datasets. The total experimental setup involved evaluating the models with approximately 101 students—67 from (Mery et al., 2019) and 34 from *(school name blinded)*—each with one photo sample. The execution time for each session was one hour, using both surveillance camera feedback and smartphone capture data as described in Mery et al. (Mery et al., 2019). That's why the results are divided in five sessions from (Mery et al., 2019) and three sessions for *(school name blinded)* . To ensure consistency and minimize random variability, the experiments were run multiple times, with the mean and standard deviations of the performance metrics calculated for robust assessment.

4 Results and Discussion

Once the system was ready and able to perform the frequency action, we compared our model with state-of-the-art models: Facenet512 (Firmansyah et al., 2023), Facenet (Schroff et al., 2015), and Arcface (Deng et al., 2019). A comparison was made for each detected face, and the result was limited to a threshold of 0.65. This threshold, determined empirically through experiments with the system, indicates that a distance below this value confirms face recognition.

Table 3: Table containing the accuracy, precision, recall and F1-Score values for the system evaluation using Mery et al. data (Mery et al., 2019).

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Facenet512	46.86 ± 0.17	90.67 ± 0.03	43.54 ± 0.22	55.45 ± 0.23
Facenet	29.55 ± 0.05	87.66 ± 0.06	25.08 ± 0.10	37.36 ± 0.14
Arcface	41.79 ± 0.13	90.51 ± 0.05	39.15 ± 0.19	50.84 ± 0.23
ResNet29	**76.71 ± 0.05**	**90.19 ± 0.10**	**83.00 ± 0.07**	**85.64 ± 0.03**

Our work has shown an average percentage improvement in accuracy of around 1.5 times compared to the other models in delivering a correct frequency, as seen in table 3. The superiority of the ResNet29 model, especially in recall, also called sensitivity, with a 40% higher confidence rate in identifying when a particular student is present, can be attributed to its residual architecture, which likely enhances its ability to capture fine-grained facial features in the training stage, particularly in a controlled classroom environment.

The higher recall value shows that the model is better at minimizing false negatives, i.e., it avoids missing students who are present, which is critical in an attendance monitoring system. Consequently, ResNet29's F1-Score also showed better values, with an average of 55% higher than the other models, making it 1.5 times better. It presented a good trade-off between precision and recall, making it more reliable for real-world applications, where both missed detections and false positives could have practical consequences.

Following the metrics selected and the results obtained, we acknowledge that the AUC-ROC value represents the balance of the model in terms of its classification. As shown in figure 6, the model presented in this work delivers the greatest balance overall, always surpassing the value of 0.6 in AUC value, achieving a favorable balance between sensitivity and specificity, in other words, a better balance dealing with false negatives and false positives, respectively.

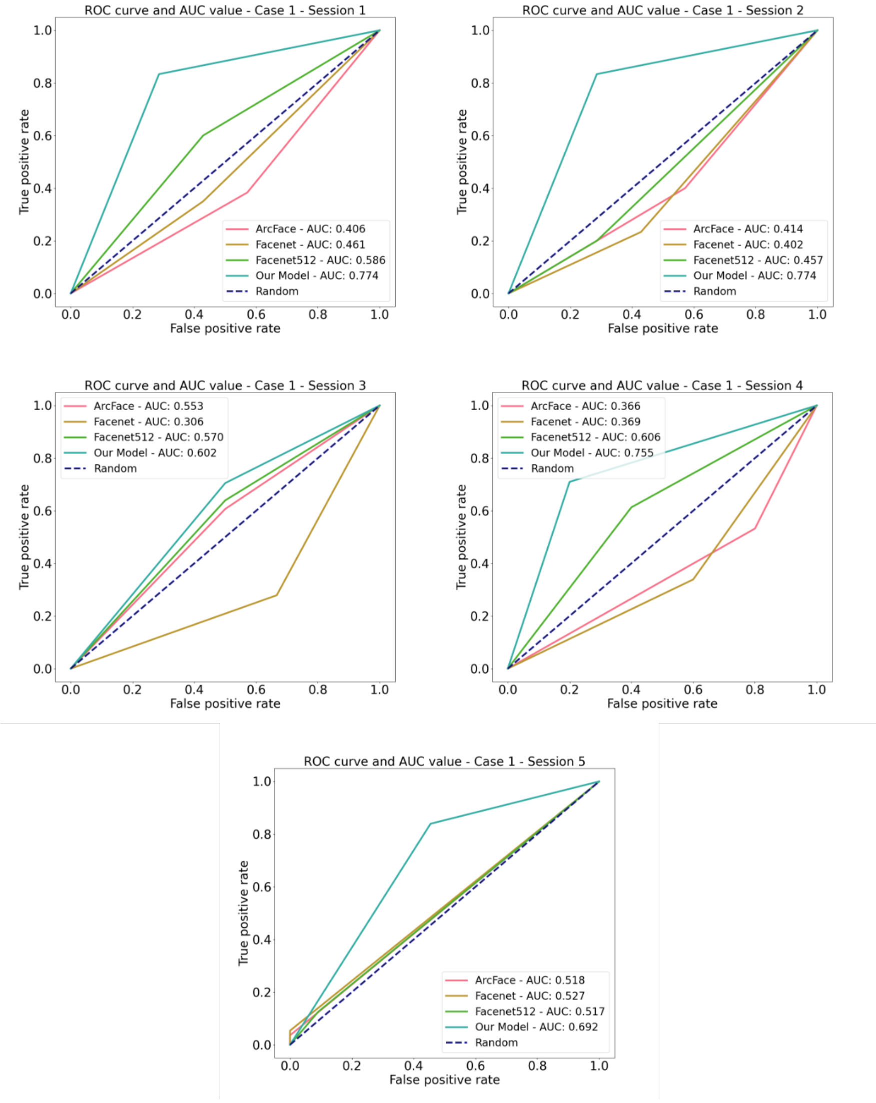


Figure 6: Graphs containing the AUC-ROC values for class sessions 1 to 5 from work by Mery et al., 2019.

Following on from the tests, the *(school name - blinded)* data is larger but has a poorer resolution because it uses ordinary surveillance cameras with 2 megapixels available for capture. The values shown in Table 4 result from a mean between the metrics from all three sessions captured. Even when dealing with less quality data, our solution showed better results than the other models, demonstrating values above 80% in all metrics. This indicates that our model is robust to variations in image quality and can effectively leverage the available data to achieve high performance. Furthermore, these findings suggest the potential for practical applications in environments with limited camera specifications, highlighting the adaptability of our approach.

Similar to the findings in (Mery et al., 2019), the *(school name - blinded)* data in Figure 7 also demonstrated better performance metrics, with ResNet29 showing approximately double the accuracy compared to other models. Furthermore, the model yielded a higher AUC-ROC value, indicating its superior capacity to distinguish between correct and incorrect classifications.

Table 4: Table containing the accuracy, precision, recall and F1-Score values for the system evaluation using the *(school name - blinded)* data.

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Facenet512	44.44 ± 0.08	84.44 ± 0.13	33.33 ± 0.12	46.23 ± 0.12
Facenet	37.37 ± 0.09	85.71 ± 0.11	19.99 ± 0.02	32.31 ± 0.03
Arcface	35.35 ± 0.07	87.50 ± 0.10	17.30 ± 0.06	28.17 ± 0.09
ResNet29	**80.80 ± 0.05**	**82.97 ± 0.07**	**93.16 ± 0.02**	**87.66 ± 0.04**

In scenarios where image resolution was lower, ResNet29 maintained strong performance, likely due to its ability to effectively capture key facial features despite the limited data quality. Its high AUC score and excellent recall rate of 93.16% highlight the model's robustness in identifying present students while minimizing false negatives. This is particularly crucial in attendance systems, where missing a student who is present could undermine the system's reliability. Furthermore, the model's strong decision power in managing false positives further strengthens its application in real-world classroom environments, ensuring that absent students are accurately distinguished from those present.

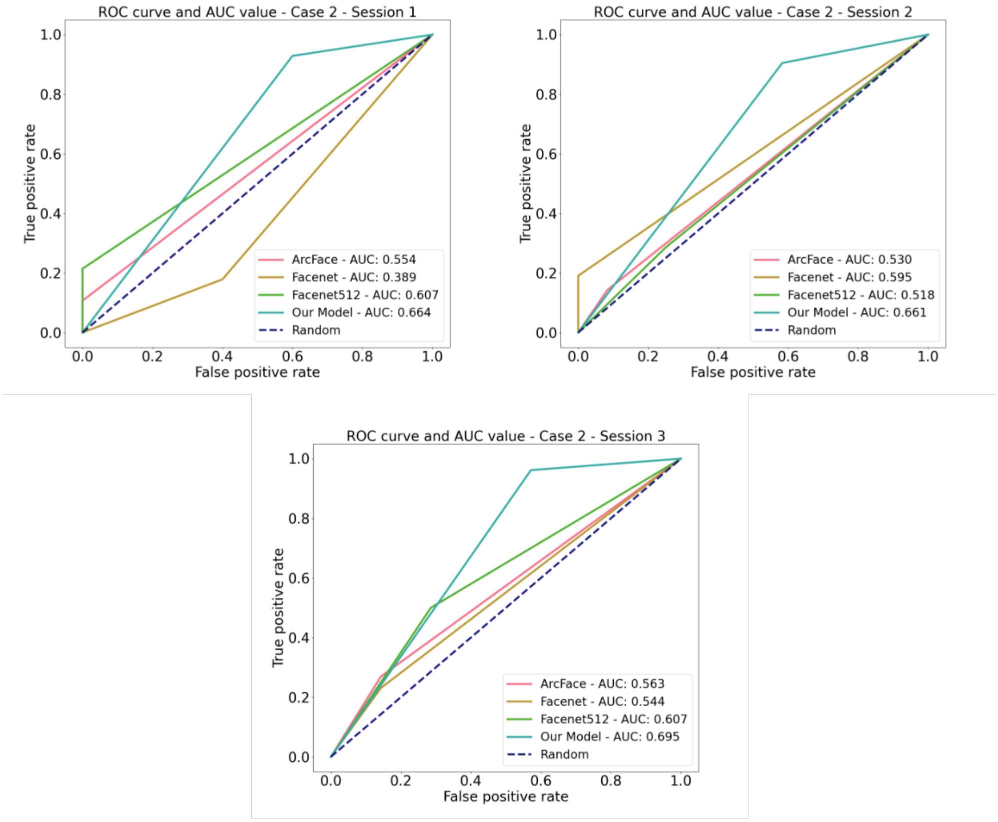


Figure 7: Graphs containing the AUC-ROC values for a *(school name - blinded)* class session number 1 to 3.

Describing all three graphs in figure 7, one by one:

- The first session, as shown in figure 7, exhibits a clear upward trend with an initial dip, reflecting the model's performance across various thresholds. Initially, the curve ascends slowly, allowing FaceNet512 (Firmansyah et al., 2023) to outperform it momentarily. However, following this initial phase, the model demonstrates a steady improvement, marked by a gradual rise in performance. This progression indicates the model's ability to differentiate between positive and negative classes, even under varying threshold conditions. The steepness of the AUC-ROC curve highlights its high true positive rates while maintaining a relatively low false positive rate, which is a crucial indicator of its strong predictive capability. Additionally, an AUC score of 0.684 further validates the model's efficacy, significantly surpassing the random guess baseline of 0.5 and demonstrating its ability to provide reliable predictions well above the threshold of random classification.

- In the second session depicted in figure 7, a similar trend emerges. Initially, the model's performance declines, allowing FaceNet (Schroff et al., 2015) to surpass it briefly. However, following this phase, the model's performance improves significantly, ultimately outperforming the other models. The curve's steepness indicates that the model successfully maintained a high true positive rate while keeping the false positive rate low, reflecting a strong ability to discriminate between classes. This is particularly noteworthy, as maintaining a balance between true and false positives is crucial in minimizing incorrect detections in a real-world scenario. Moreover, an AUC score of 0.684 surpasses the random guess baseline of 0.5, demonstrating that the model provides meaningful and reliable predictions beyond chance.
- Moreover, the third session of figure 7 presented a different situation, where the competition between the AUC-ROC curves of the models was higher. However, at the same time, this session showed the biggest difference between the first and second-best values. The slope of this session also showed better capacity to deliver true positives than the other models and a better capacity of generalization due to its AUC value of 0.695. All these sessions showed results with a good capacity to recognize true positives but with a high rate of false positives, which is the main problem in this scenario of automatic attendance, showing that the model can be improved to differentiate better the positive results, resulting in a more balanced model.

Implementing the system eliminated the need for manual attendance-taking during class, which previously required 3 to 6 minutes of classroom time. Considering that each class lasts 60 minutes, this time savings translates into a significant increase in instructional time. As a result, the system provided an approximate 9% gain in classroom time, allowing teachers to focus more on delivering content and engaging with students rather than administrative tasks.

In terms of practical application results, the proposed system was tested in a real-world school environment with real situations, not requiring time dedicated to the capture of students' faces during class time; in situations like this, were where the system demonstrated significant efficiency improvements compared to traditional attendance methods, such as manual roll calls, token-based systems (Sawall et al., 2021) or even other ML-based systems (Mery et al., 2019) which requires a pause on the class to capture the students' faces. Besides this, the system comprises an API that integrates the school system, scheduling all classes and removing the need to create a system that revolves around the model since the system is already developed.

The use of the solution does not need personnel training since it uses only the cameras and the lightweight servers, nor higher initial investments in terms of hardware; it is shown that low-cost cameras could be used to achieve metrics above 80%, and servers with a lightweight configuration can also be used. Over time, this investment leads to savings by reducing the need for administrative oversight and improving the efficiency of attendance tracking, particularly in large institutions.

It is safe to say that the presented metrics demonstrated the system's effectiveness in recognizing students and showcased its lightweight architecture, allowing it to handle multiple tasks indirectly and simultaneously efficiently. The system was achieved without overburdening the available resources, even with a simple configuration like the one in the second column of Table 2, making it ideal for schools with limited technological infrastructure. Furthermore, the system's modular design and ease of integration emphasize its strong potential for replication in other educational institutions, particularly in environments with reduced computing capacity or minimal IT support. Its flexibility ensures it can be adapted to different classroom setups and operational needs, promoting broader adoption in diverse educational contexts.

4.1 Adaptability to Different Educational Contexts

Following the results, the proposed system demonstrates the capacity for adaptability across various educational levels and cultural contexts due to its lightweight structure and ability to function using common surveillance cameras, making it a versatile solution for attendance tracking in diverse learning environments.

The system can be characterized as flexible due to its passive operation, ensuring that classroom activities remain uninterrupted regardless of the students' academic level. Unlike traditional attendance methods, which require direct interaction, this system allows students to focus entirely on their lessons without disrupting the instructional flow. Its flexibility enables implementation in institutions ranging from primary schools to higher education. In primary and secondary schools, where younger students may be less disciplined in maintaining visibility for facial recognition, minor adjustments such as strategic camera placement and real-time feedback could be introduced to ensure optimal performance. Furthermore, in higher education settings with large lecture halls, the system could be optimized to handle larger student volumes while maintaining accuracy, similar to the approach in (Mery et al., 2019).

Different regions and educational institutions have varying policies regarding student privacy, data protection, and technological adoption. The system can be configured to comply with local regulations by implementing secure data encryption, anonymization techniques, and customizable access controls. In schools with strict privacy guidelines, alternative recognition methods, such as anonymized attendance tracking through unique identifiers, can be integrated while preserving student confidentiality. In areas where technological infrastructure is limited, the system's ability to function on low-cost cameras and lightweight servers while using any video streamline makes it a practical solution. Schools with budget constraints or minimal IT support can deploy the system without requiring significant financial investments, ensuring broader accessibility and adoption.

Scalability is a crucial factor when considering the system's applicability in large institutions or multi-campus environments. The modular nature of the system allows for easy expansion by incorporating additional cameras and processing units with minimal adjustments to the existing setup.

To scale the system, institutions can gradually increase the number of surveillance cameras while distributing computational workloads across multiple servers or cloud-based solutions, ensuring performance is maintained even with a higher number of students.

Regarding costs, the system's implementation primarily involves three key components:

- **Hardware:** Standard surveillance cameras, edge-computing devices, and storage solutions. The cost varies depending on the number of cameras required and whether the institution uses existing infrastructure.
- **Software:** The system relies on open-source frameworks and customizable machine learning models, reducing licensing costs compared to proprietary solutions.
- **Human Resources:** Initial setup requires technical expertise for deployment and integration into the institution's IT environment. However, long-term maintenance is minimal compared to other attendance solutions requiring frequent manual intervention.

Compared to other attendance-tracking solutions, such as RFID-based systems (Kurunthachalam et al., 2021; Rashmi et al., 2022) or biometric scanners (Wati et al., 2021), which require dedicated hardware and infrastructure investments, the proposed system operates with standard surveillance cameras and lightweight servers, significantly reducing costs. RFID systems involve purchasing and distributing individual student cards, as well as maintaining specialized scanners. Similarly, biometric systems require high-cost fingerprint or iris scanners, along with regular calibration and upkeep. Traditional machine learning-based attendance solutions also introduce high costs due to manual image collection and annotation. In contrast, our system eliminates classroom interruptions, automates attendance tracking, and minimizes operational expenses, making it a cost-effective and scalable solution.

5 Conclusion

In this work, we developed an end-to-end automated student attendance recording system using existing classroom security cameras, the school's database, and an API. We evaluated model performance in ideal scenarios with high-quality images and challenging conditions, such as low-quality images and students not facing the camera.

Our results show that ResNet29 outperformed other models with low-resolution images, achieving metrics at least twice as effective in practical applications. This highlights ResNet29's robustness in real-world settings, where image quality may vary. It proved 1.5 times more accurate than other solutions in distinguishing present from absent students, yielding a higher specificity and true negative rate, as reflected in AUC-ROC curves.

The automated system reduces attendance recording time by approximately 9%, freeing up classroom time for instruction rather than administrative tasks. Its accuracy and reliability and lower implementation costs provide long-term benefits in reduced labor and enhanced precision. However, challenges remain in ensuring system scalability, maintaining LGPD compliance, and integrating smoothly with existing school infrastructure.

Reframing surveillance cameras as a support tool for teachers highlights their role in freeing time for lesson plans. The system has been deployed at *(school name - blinded)* and is currently used by teachers across different periods in two classrooms. Given the positive results from the initial implementation, we are expanding to additional classrooms, further reinforcing the system's non-intrusive nature and contributing to the concept of Next-generation Smart Classrooms.

References

Adelman, C. (2006, February). *The toolbox revisited: Paths to degree completion from high school through college* (tech. rep.). U.S. Department of Education. https://eric.ed.gov/?id=ED490195 [GS Search].

Ashfaq, M., Amaan, M., Aalam, M., Ullah, M. A., Raghav, D., & Goel, A. (2023). Computer vision based attendance management system for students. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2023.48767 [GS Search].

Costa, J., & Guedes, L. (2022). Proposta de integração curricular com internet das coisas na educação profissional técnica de nível médio. *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, 244–254. https://doi.org/10.5753/sbie.2022.225171 [GS Search].

Decoito, I., & Richardson, T. (2018). Teachers and technology: Present practice and future directions. *Contemporary Issues in Technology and Teacher Education*, *18*(2), 362–378. https://www.learntechlib.org/primary/p/180395/ [GS Search].

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4685–4694. https://doi.org/10.1109/CVPR.2019.00482 [GS Search].

Firmansyah, A., Kusumasari, T. F., & Alam, E. N. (2023). Comparison of Face Recognition Accuracy of ArcFace, Facenet and Facenet512 Models on Deepface Framework. *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, 535–539. https://doi.org/10.1109/ICCoSITE57641.2023.10127799 [GS Search].

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, *57*(3), 238–247. http://www.jstor.org/stable/1403797 [GS Search].

Geitgey, A. (2017). Face_recognition. https://github.com/ageitgey/face_recognition

Ghimire, S. N., & and, K. R. (2023). CCTV in schools: An examination of perceived value of surveillance. *Journal of Education for Students Placed at Risk (JESPAR)*, *28*(4), 351–379. https://doi.org/10.1080/10824669.2022.2092110 [GS Search].

Gupta, S., T. S., A., & Reddy Guddeti, R. M. (2018). CVUCAMS: Computer vision based unobtrusive classroom attendance management system. *IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 101–102. https://doi.org/10.1109/ICALT.2018.00131 [GS Search].

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90 [GS Search].

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. https://inria.hal.science/inria-00321923 [GS Search].

Kaliappan, J., Shreyansh, J., P, S. S., & Singamsetti, M. S. (2019). Surveillance camera using face recognition for automatic attendance feeder and energy conservation in classroom. *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 1–5. https://doi.org/10.1109/ViTECoN.2019.8899369 [GS Search].

Kar, N., Deb Barma, D. M., Saha, A., & Pal, D. (2012). Study of implementing automated attendance system using face recognition technique. *International Journal of Computer and Communication Engineering*, 100–103. https://doi.org/10.7763/IJCCE.2012.V1.28 [GS Search].

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, *10*, 1755–1758. https://dl.acm.org/doi/abs/10.5555/1577069.1755843 [GS Search].

King, D. E. (2015). Max-margin object detection. *CoRR*, *abs/1502.00046*. http://arxiv.org/abs/1502.00046 [GS Search].

Kurunthachalam, A., Sangeetha, S., Periyakaruppan, K., Keerthana, K., SanjayGiridhar, V., & Shamaladevi, V. (2021). Design of attendance monitoring system using RFID. *7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, *1*, 1628–1631. https://doi.org/10.1109/ICACCS51430.2021.9441704 [GS Search].

Lavareda Filho, R., Colonna, J., Oliveira, D., Monteiro, E., & Gonçalves, P. (2022). Autenticação de alunos utilizando dinâmica de digitação e redes neurais profundas em sistemas juiz on-line. *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, 1222–1232. https://doi.org/10.5753/sbie.2022.225779 [GS Search].

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 [GS Search].

Norton, V., Honda, F., Pessoa, M., & Pires, F. (2024). CodeX: Ambiente virtual de aprendizagem em programação, integrado à LLM, para auxiliar estudantes com TEA. *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, 1770–1783. https://doi.org/10.5753/sbie.2024.242533 [GS Search].

Rashmi, Brindha, S., Srinithin, & Gnanasudharsan. (2022). Smart Attendance System Using RFID and Face ID. *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–5. https://doi.org/10.1109/IC3IOT53935.2022.9768003 [GS Search].

Sawall, E., Honnef, A., Mohamed, M., Alqahtani, A., & Alshayeb, T. (2021). Covid-19 zero-interaction school attendance system. *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 1–4. https://doi.org/10.1109/IEMTRONICS52119.2021.9422614 [GS Search].

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *CoRR*, *abs/1503.03832*. http://arxiv.org/abs/1503.03832 [GS Search].

Serengil, S. I., & Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 1–5. https://doi.org/10.1109/ASYU50717.2020.9259802 [GS Search].

Sreedevi & Ram, R. (2019). Smart attendance registration for future classrooms. *Proceedings of the International Conference on Systems, Energy & Environment (ICSEE)*. https://doi.org/10.2139/ssrn.3444043 [GS Search].

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594 [GS Search].

Uskov, V. L., Bakken, J. P., & Pandey, A. (2015). The ontology of next generation smart classrooms. In V. L. Uskov, R. J. Howlett, & L. C. Jain (Eds.), *Smart education and smart e-learning* (pp. 3–14, Vol. 41). Springer International Publishing. https://doi.org/10.1007/978-3-319-19875-0_1 [GS Search].

Wati, V., Kusrini, K., Fatta, H. A., & Kapoor, N. (2021). Security of facial biometric authentication for attendance system. *Multimedia Tools and Applications*, *80*, 23625–23646. https://doi.org/10.1007/s11042-020-10246-4 [GS Search].

Zhou, L., & Jiang, X. (2019). Face recognition based on weighted fusion of face similarity features. *3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, 1691–1694. https://doi.org/10.1109/eitce47263.2019.9094888 [GS Search].