Review of Computer Engineering Research
2027 Vol. 14, No. 1, pp. 1-20
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v14i1.5374
© 2027 Conscientia Beam. All Rights Reserved.

A multi-modal facial recognition framework for continuous attendance tracking: Comparative performance evaluation

Rajesh Yadav[1]
Swati Gupta[2]
Meenakshi Malik[3]
Chander Prabha[4]
Anurag Mahajan[5+]

[1,2]*K.R. Mangalam University, Gurugram, Haryana, India.*
[1]*Email: Himank9d2006@gmail.com*
[2]*Email: swattiguptta@gmail.com*
[3]*BML Munjal University, Gurugram, Haryana, India.*
[3]*Email: Meenakshimalik16@gmail.com*
[4]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.*
[4]*Email: prabhanice@gmail.com*
[5]*Department of E&TC, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India.*
[5]*Email: anurag.mahajan@sitpune.edu.in*



(+ Corresponding author)

Article History
Received: 15 December 2025
Revised: 9 April 2026
Accepted: 5 May 2026
Published: 20 November 2026

Keywords
Classroom analytics
Cloud-based monitoring
Continuous authentication
Convolutional neural networks
Haar cascade classifier
Multi-face recognition.

ABSTRACT

Facial recognition is a mature biometric technology. Nonetheless, the use of this biometric in continuous classroom attendance monitoring is very limited. The use of this biometric for attendance monitoring relies on isolated points of verification. Moreover, the technology is vulnerable to proxy attendance. This work proposes a real-time, multi-modal facial recognition framework for continuous attendance tracking using a deep neural network pipeline. The system performs periodic face scanning during live classroom sessions, records attendance at 30-minute intervals, enforces a minimum 75% presence criterion, and synchronizes all operational events to a scalable Firebase cloud backend for real-time administrative access. The framework is evaluated on a real-world dataset comprising over 1,000 authentic student videos captured in conventional classroom environments, inherently incorporating variations in illumination, pose, occlusion, and background noise. Comparative evaluation is conducted against SVM, a vanilla CNN, InceptionV3, MobileNet, and VGG-19. This research presents a deep-learning-based automated attendance system that employs Haar Cascade for real-time face detection and CNN-based architectures for face recognition. The framework records 98.1% training accuracy and 97.4% validation accuracy, outperforming every baseline model. The MobileNet model is the only one close to this achievement. Analysis through statistical validation by one-way ANOVA ($F(5,24) = 1408.61$, $p<0.0001$), Tukey's HSD, and paired t-tests ($t=69.33$, $p<10^{-6}$) shows robust results. The proposed system is reliable, scalable, and computationally efficient, enabling low-overhead, large-scale deployment for continuous attendance monitoring and institutional process automation.

Contribution/Originality: This study adds to previous research by introducing a continuous real-time multimodal facial-recognition framework for classroom attendance monitoring. The framework is one of the few studies integrating the temporal multi-face verification, cloud synchronization, regulatory compliance, and classroom security verified on large-scale real classroom video data.

1. INTRODUCTION

Student attendance recording is a critical administrative task in higher educational institutions, as it is directly linked to institutional academic integrity and regulatory compliance [1]. Institutions are also required to adhere to national and regulatory attendance guidelines. In India, the University Grants Commission (UGC) mandates a

© 2027 Conscientia Beam. All Rights Reserved.

minimum 75% attendance requirement for students to be eligible for trimester-based courses. Even though it has now been established by law, there is still a heavy reliance upon old-fashioned methods like handwritten sign-in sheets and verbal roll calls [2]. These methods present substantial methodological problems. Outdated systems are not only ineffective and administratively overwhelming, but they are also subject to systemic error and abuse, including widespread proxy attendance, forgery, etc.

Continuing use of old systems is making some operating problems worse. Considering the growing strength of the student community and the difficulty of monitoring large lecture halls, it has become a tiresome and error-prone job to take attendance manually [3, 4]. This is particularly applicable to temporary academic staff such as visiting lecturers or guest faculty. The latest security arrangements for educational institutions require controlled entry to the classrooms, which is often missed by the attendance tracking systems in place [5].

Indian universities face unique infrastructural and operational challenges. Despite advances in digital automation, these technologies are not effectively integrated into systems suitable for Indian universities. The majority of the systems out there are compartmentalized in such a way that they focus on either attendance monitoring or access control [5]. But the two functions are not integrated well in any one of the systems. The study presents an original framework that simultaneously ensures the verification of students' presence and non-intrusion while allowing safe real-time reporting for the administration [6]. Thus, it meets a serious demand for all-encompassing institutional administration systems.

Facial recognition is a biometric authentication that is non-invasive [7]. It is an effective method of identifying people. Moreover, it compares a live image with a stored one. Due to the advent of deep learning, mainly the Convolutional Neural Networks (CNNs), these systems are now more accurate and robust [8]. In fact, these operators are now able to work in complex real-world situations.

Monitoring student attendance is one of the most important administrative functions [9]. It impacts the evaluation of students and also the institution. Many issues arise here, with options like contact-based biometrics or manual sign-ins [10]. Prior studies report widespread proxy attendance, particularly in lengthy classes and in contact-based biometric systems that raise hygiene, usability, and operational concerns [11]. Universities in India have witnessed a significant increase in registrations, creating a challenging situation. When percentage attendance is not calculated correctly, the implementation of the University Grants Commission (UGC) 75% attendance norm is weakened. The norm also creates accountability gaps for personnel, including adjunct/visiting lecturers. In addition, another risk that is frequently ignored is the security vulnerability of big lecture theatres [12].

AI-based facial recognition systems can offer a solution for automated attendance, but most solutions only focus on biometrics logging [13]. There is a huge research gap for integrated technologies that are capable of simultaneously addressing attendance accuracy, security monitoring in real-time, as well as administrative automation in many developing countries whose resources are limited [14].

This study proposes a unique and systematic framework for an Intelligent Live Attendance System that serves as a solution. The system is made specifically for the academic institute context, and it contains three basic functional pillars.

- Automated Attendance Tracking: A video-based system employing the OpenCV Haar Cascade model is used for real-time multi-face detection and recognition against a pre-registered database, thereby automating the entire attendance logging process.
- Regulatory Compliance & Administration: It automatically computes total attendance to ensure compliance with the 75% UGC requirement and increases accountability of visiting faculty by providing authenticated records for payment processing.
- Enhanced Security Protocol: The framework has a security module that can detect and alert supervisors about unauthorized students in classrooms.

© 2027 Conscientia Beam. All Rights Reserved.

This initiative aims to create all three components and integrate them into a scalable platform. The proposed system extends beyond conventional attendance logging by enabling continuous verification, institutional compliance, and real-time classroom security monitoring. This study is organized to present a complete evaluation of the development of an advanced real-time attendance system using facial recognition.

This paper presents a systematic and thorough study of a scalable real-time facial recognition framework involved in continuous class attendance, including multi-face identification, temporal verification, cloud synchronization, and institutional compliance. The large-scale real classroom data, along with statistical and ablation analyses, demonstrate the system's performance, unlike previous snapshot-based or hardware-based systems.

1.1. Research Questions

In light of the limitations present in the previous attendance monitoring systems, this study is guided by the following research questions.

- RQ1: What is the process for designing a real-time facial recognition system to continuously verify the attendance of multiple faces in a large and dynamic classroom?
- RQ2: Is it possible to achieve attendance accuracy, regulatory compliance with UGC, and classroom security without custom biometric devices in a single framework?
- RQ3: In terms of accuracy, robustness, and scalability under actual classroom settings, how does the multi-modal attendance framework measure against traditional and cutting-edge deep learning models?
- RQ4: What is the contribution of each individual architectural component to the overall performance of the architecture?

1.2. Contributions of the Study

The principal contributions of this work are summarized as follows:

- The present work proposes a framework for a scalable multi-modal attendance framework which allows continuous multi-face detection and temporal verification in a live classroom.
- The proposed fully integrative architecture unifies auto attendance logging, institutional compliance with regulations, and classroom security and monitoring systems via a single framework.
- A real-world large-scale evaluation is conducted with authentic classroom video data that includes variations in illumination, pose, and student movement.

This paper is organized as follows: Section 2 provides the context of the study, articulates the problem, and discusses limitations of the proposals. The state-of-the-art facial recognition technology for attendance is analyzed in Section 3. In this section, we discuss the architecture of the proposed system in detail, including the methods used, major algorithms, data processing techniques, and essential requirements for implementation. Section 4 describes the experimental setup and evaluation metrics. The results section presents and critically analyzes the empirical findings, evaluating system performance against the stated objectives. Section 5 concludes the paper by summarizing the key findings, discussing its academic and practical implications, and offering recommendations for future research.

2. LITERATURE REVIEW

Facial recognition has been studied increasingly as an alternative to manual and contact-based attendance systems due to its differentiated recognition and automation potential. According to existing studies, the accuracy and speed of recognition benefit from various algorithm designs and hardware implementations [15]. However, most previous works address specific aspects of attendance automation and do not devise a monolithic and scalable system suitable for real classroom settings. Hardware architectures have been extensively studied. For example, low-latency recognition in a cost-effective manner was achieved using an FPGA-based implementation and on an embedded platform, Raspberry Pi [11, 16, 17].

© 2027 Conscientia Beam. All Rights Reserved.

Although these methods show promising accuracy in controlled settings, they heavily depend on static image matching or validating single frames. As a result, they cannot be used for continuous monitoring of classrooms. The evaluations are performed on small-scale or benchmark datasets, not adequately capturing real-world complexities such as multiple face dynamics, illumination changes, and long sessions. An alternative research approach pertains to photometric and landmark-sensitive methods, such as geometric facial descriptors and single-sample-per-person approaches [18, 19]. Though these methods improve discrimination in constrained settings, their robustness is severely degraded in unconstrained classroom environments with head pose variation, motion blur, and a large number of students. Furthermore, several systems have implemented facial recognition for secondary aims such as emotion analysis and behavior detection [12, 20]. However, none address persistent and authenticated verification of attendance nor proxy attendance effectively.

In both cases, it can be noted that attendance is treated as a recognition problem based on snapshots rather than a continuous verification problem. Most systems do not validate for temporal (private/public changes over time) and are not able to track multi-face persistence. Systems also do not integrate with institutional processes such as regulatory compliance, administrative reporting, classroom security, etc. Moreover, scalability is an issue, as performance is validated by a few studies in large lecture halls or real academic settings with prolonged monitoring.

These limitations cumulatively depict an important research gap for an integrated real-time attendance system that can continuously recognize multiple faces with temporal consistency and can be deployed at a large scale, adding to institutional administration and security mechanisms convincingly.

As shown in Table 1, the various existing face-recognition-based attendance systems are compared with the proposed system systematically across several technical and deployment aspects. The comparison reveals shortcomings in earlier snapshot-based recognition works, a lack of temporal validation, and limited real-world assessment, while establishing the proposed system's superiority in continuous monitoring, scalability, and institutional integration.

Table 1. Systematic comparative analysis of face-recognition-based attendance frameworks.

Study	Recognition mode	Continuous monitoring	Real classroom data	Hardware dependency	Key limitation
Hassan, et al. [11]	Single Frame	✗	Small-scale (8 students)	Medium	Poor generalizability
Shayea and Mangoud [12]	Event-based	✗	✗	Medium	Low resolution; messaging delays
Eddine and Zohra [16]	Static image	✗	✗	Medium	No temporal consistency
Bhavikatti and Bhairannawar [17]	Static image	✗	✗	High (FPGA)	Not scalable; static verification
Sanil, et al. [18] and Chandrashekar, et al. [19]	Snapshot	✗	✗	Low	Sensitive to pose & lighting
Arifin, et al. [20]	Snapshot	✗	✗	Low	Attendance is not the primary objective
Multi-modal attendance framework	Real-time video	√	√ (1006 students)	Low	Addresses prior gaps

In contrast to previous attendance systems based on static or single-frame facial recognition, our multi-modal attendance framework employs continuous, real-time multi-face verification with temporal consistency. The system works on consumer-grade devices and is verified on real classroom data on a large scale, as opposed to hardware-dependent ones.

© 2027 Conscientia Beam. All Rights Reserved.

3. RESEARCH METHODOLOGY

This research provides a complete video-based deep learning framework for real-time classroom attendance analysis with integration of continuous multi-face detection, temporal verification, and automated report generation for administration. The proposed method, unlike most existing attendance systems that use snapshot-based or single-frame facial recognition, conducts continuous verification of the entire class. As a result, it effectively detects latecomers, early leavers, and proxy attenders. Additionally, unlike previous studies mostly evaluated on small-scale, controlled, or benchmark datasets, this framework can be deployed on consumer hardware and has been validated on large-scale classroom video data.

The methodological approach of this study comprises a comprehensive end-to-end pipeline that converts primary class video recordings into a real-time automated attendance system based on deep learning. The first stage of the process is dataset generation and then video processing, followed by facial extraction, model development, training, deployment, and complex statistical validation. This system is meant to work in 'real-world' classroom situations where lesson length is between forty and ninety minutes, class sizes are from forty to seventy-five students, and lighting conditions change significantly. The following sub-sections will present the detailed and comprehensive working process along with actual numerical values achieved through the experimental implementation of the process.

3.1. Dataset Preparation and Organization

The data set used to train the model consists of one thousand different directories, each containing a different student in a course. On average, there is one video per folder, with the average video duration being 62.4 seconds (41 to 84 seconds). The dataset has a size of 18.6GB after the first preprocessing, while the size of the individual video files ranges from 68MB to 192MB.

Every student receives a folder with their scholastic roll number, such as 23BCA10014. The naming convention facilitates easy identification of student facial photos. The system detects 1006 files and checks 1003, which tend to have at least one valid video file. The remaining 3 directories are automatically excluded and recorded during pre-processing. The folders contain approximately one file because several learners provided two videos for clarification.

The dataset checker runs through the folder tree in about 17.82 seconds to check for valid video files and missing or corrupted files. Out of all the pupils, there are 1003 distinct video files identified.

3.2. Video Ingestion and Face Extraction

3.2.1. Frame Sampling Strategy

Videos captured at 30 frames per second (FPS) produce, on average, 1,800 frames per minute of content. To enhance processing efficiency and minimize redundant processing, the system samples every fifth frame, thereby decreasing the overall number of processed frames to roughly 20% of the original set. For a standard sixty-two-second video at thirty frames per second, the system processes approximately.

$$\frac{60 * 30}{5} = 360\ frames\ per\ student$$

This sampling method reduces processing power by about 80% while still maintaining intra-personal variability, including blinks, head tilt, gaze direction, and expression changes.

The algorithm extracts a maximum of fifty faces from every video. The limit is there, so no student with a lengthy video submission biases the model during training. Based on experimental observations, the system meets the limit on approximately 64% of the videos.

© 2027 Conscientia Beam. All Rights Reserved.

3.3. Face detection using Haar cascade classifier.

Changing each processed frame to grayscale reduces every pixel from three channels to one, thus decreasing computing complexity by 66%. Face detection employs the Haar Cascade filter with a scale factor of 1.1 and a minimum neighbor threshold of 5, optimized through multiple trials to improve accuracy.

During the processing of the dataset, the classifier achieves a mean detection accuracy of 97.4% when manually verified on an arbitrary number of 1,000 frames, with a false-positive rate of 3.2% and a false-negative rate of 6.8%. To ensure quality, we remove features below 50×50 pixels for further analysis. The algorithm detects a cumulative sum across all videos.

Approximately 1003 students, each contributing around 40 to 50 faces, give an estimated total of 40,120 to 50,150 usable facial photos. This amount shows limited variance since students who generate more quality movies produce more distinguishable frames.

3.4. Preprocessing and Normalization

Each clipped face is scaled to 128×128 pixels, resulting in 49,152 colors in RGB per image. The resizing process succeeds in 98.4% of instances; the remaining 1.6% of frames are deleted due to distortion, occlusions, or harsh illumination conditions. The complete collection of pre-processed faces uses around 265 MB of RAM when loaded into memory for model training.

3.5. Model Architecture Design

The deep learning model used for face recognition is a custom-designed CNN architecture containing 21 trainable layers with 3.8 million parameters. The architecture is optimized for execution speed and high recognition accuracy, targeting deployment on consumer-grade hardware.

3.5.1. Convolutional Feature Extractor

The convolutional feature extractor operates on an input image of shape 128*128*3 and is organized into three successive convolutional blocks. Each block contains two 3*3 convolutional layers (padding=same, stride) followed by a 2*2 max-pooling layer (stride 2). The sequence of filter depths is $32 \rightarrow 64 \rightarrow 128$. All convolutional layers use ReLU activations.

3.5.2. Spatial Dimensions and Channel Counts After Each Block

Because each convolution uses padding=same and stride 1, the spatial resolution is preserved through the two convolutions in each block and is reduced only by the following 2*2 pooling (stride 2). Starting from the input spatial size

$$128 * 128$$

The successive pooling operations halve the spatial dimensions at the end of each block.

$$128 * 128 \rightarrow 64 * 64 \rightarrow 32 * 32 \rightarrow 16 * 16$$

Thus, within every block, the two $3 * 3$ convolutions keep the current spatial size, and the single 2*2 max pool at the block's end performs the halving shown above.

Block 1: Two convolutional layers with 32 filters each preserve the spatial size at 128*128 during convolutions. After the 2*2 pooling, the spatial dimensions become $\frac{128}{2} = 64$ for both height and width. Thus, the Block-1 output tensor has shape $64 * 64 * 32$. The number of scalar activations (spatial features) in this tensor is computed as:

$$64 * 64 * 32 = (64 * 64) * 32 = 4096 * 32 = 131{,}072$$

Block 2: The two convolutional layers operate on the Block-1 output and preserve the spatial size until pooling. After the 2*2 pooling, the spatial dimensions become $\frac{64}{2} = 32$. With 64 filters in Block~2, the output tensor has shape $32 * 32 * 64$. The number of scalar activations in this tensor is:

© 2027 Conscientia Beam. All Rights Reserved.

$$32 * 32 * 64 = (32 * 32) * 64 = 1024 * 64 = 65{,}536$$

Block 3: After the two convolutional layers and the final $2 * 2$ pooling, the spatial dimensions become $\frac{32}{2} = 16$, with 128 filters. Therefore, the Block-3 output tensor has shape $16 * 16 * 128$. The number of scalar activations in this tensor is:

$$16 * 16 * 128 = (16 * 16) * 128 = 256 * 128 = 32{,}768$$

Table 2 delineates the disparity between the accurate pooled feature-map dimensions and the erroneous pre-pool interpretations that may inadvertently exaggerate the feature count. The model uses an actual flattened vector of 32,768 features taken solely from the last block (16×16×128). The total of all pooled activations is 229,376, while the sometimes-misunderstood pre-pool value of 524,288 derives from Block 1 prior to pooling. This explanation guarantees precise reporting of the network's spatial feature dimensionality.

Table 2. Comparison of spatial feature counts: correct pooled outputs vs. Misinterpreted pre-pooling activations.

Block	Pre-pooling feature map	Pre-Pooled Activation	Post-Pooling Feature Map	Post-Pooled Activation
Block 1	$128 * 128 * 32$	524,288	$64 * 64 * 32$	131,072
Block 2	$64 * 64 * 64$	262,144	$32 * 32 * 64$	65,536
Block 3	$32 * 32 * 128$	131,072	$16 * 16 * 128$	32,768
Sum of all blocks	-	917,504	-	229,376
Final Flattened Feature Vector: $16 * 16 * 128 = 32{,}768$				

3.5.3. Flattening and the Correct Number of Features

The common pattern in CNN architectures is to take the output of the final convolutional block and flatten it before passing it to the dense layers. In this model, the final tensor has spatial dimensions $16 * 16 * 128$ channels. Thus, the total number of scalar activations is

$$16 * 16 * 128 = 32{,}768$$

These 32,768 values constitute the feature vector that is fed into the first fully connected (Dense) layer, assuming that only the last block's output is flattened rather than concatenating earlier blocks.

3.6. Receptive Field Growth

The receptive field is the size (in input pixels) of the image region that influences a single activation at a particular spatial location in a feature map. The receptive field grows with each $3 * 3$ convolution and doubles in the effective sampling stride with each $2 * 2$ pooling layer (stride 2). Using the standard incremental calculation, where for a layer with a kernel size k and stride s,

$$RF_{new} = RF_{old} + (k - 1).Stride_{old,}\ \ Stride_{new} = Stride_{old}.S$$

We compute the receptive field at the final feature map exactly as shown in Table 3.

Start with the receptive field $RF = 1$ and total $Stride = 1$ at the input pixel level.

Table 3. Receptive field and total stride progression across layers.

Layer	Receptive Field (RF)	Total Stride
After Layer 1 (Conv 3*3, stride1)	1+(3-1).1 = 3	1
After Layer 2 (Conv 3*3, stride1)	3+(3-1).1 = 5	1
After Pool1 (2*2, stride 2)	5+ (2-1).1 = 6	2
After Layer 3 (Conv 3*3, stride1)	6+(3-1).2 = 10	2
After Layer 4 (Conv 3*3, stride1)	10+(3-1).2 = 14	2
After Pool2 (Conv 2*2, stride2)	14+(2-1).2 = 16	4
After Layer 5 (Conv 3*3, Stride1)	16+(3-1).4 = 24	4
After Layer 6 (Conv 3*3, Stride1)	24+(3-1).4 = 32	4
After Pool3 (2*2, stride 2)	32+(2-1).4 = 36	8

© 2027 Conscientia Beam. All Rights Reserved.

Therefore, a single spatial activation in the final $16 * 16$ feature map corresponds to a $36 * 36$ patch in the original input image. This means that each final-layer feature aggregates information from a localized $36 * 36$ region of the $128 * 128$ input; stacking deeper layers and introducing dense connectivity enables the network to progressively compose larger, more integrated, and more abstract patterns from these local receptive-field detectors.

3.7. Batch Normalization

Batch normalization is applied after each convolution. It normalizes each channel's activations to zero mean and unit variance, followed by a learned scale γ and shift β parameters. In practice, batch normalization stabilizes layer inputs, reduces internal covariate shift, enables higher learning rates, and speeds up convergence. In our experiments, activation variance decreased by approximately $27 - 31\%$ with batch normalization (e.g., $0.85 \rightarrow 0.62$, giving $\frac{0.85-0.62}{0.85} \approx 27\%$), which helps suppress extreme activations, improves optimizer stability, and provides mild regularization.

3.8. Dropout Regularization

We add Dropout layers after pooling, with dropout probabilities chosen from [0.25, 0.5] and dense layers. Dropout nullifies a fraction p of activations during training, making it less likely that co-adaptation occurs and encouraging more robust feature learning (e.g. $p = 0.25$ means that 25% of units are dropped during an update. Dropout also acts like model averaging because it helps networks. Using dropout simulates a collection of thin sub-networks, with a total of 2^N. At inference time, Disabled dropout ensures proper scaling of activations to produce outcomes. Dropout rates in the range $0.25 - 0.5$ work empirically well with medium-sized CNNs of these types.

3.9. Activation Function Choice (ReLU) and Initialization

The ReLU activation function is used for all convolution layers. It is computationally efficient. It also creates a sparse activation of the neurons. Thus, it helps to alleviate the vanishing gradient problem for positive inputs. He et al. initialization is recommended for ReLU-based networks, as it modifies the weight variance according to the number of input channels, thereby improving initial training stability. Keras/TensorFlow frameworks presume defaults like [10] or [21] initializers depending on layer setup; however, for optimal reproducibility and performance with ReLU, it is best to specify kernel_initializer='he_normal' in production tests.

Table 4 presents a complete summary of data processing and model preparation activities, such as dataset composition, frames sampled, face detection, pre-processing, CNN architecture, and more.

This pinpoints some crucial statistics, such as the number of real recordings, extracted facial images, Final feature dimension (FD), and receptive field (RF) expansion. The table mentions the activation functions, normalization, dropout, and initialization used in the face-recognition system proposed in this paper.

Table 4. Summary of dataset, video processing, face extraction, and model preparation pipeline.

Stage	Key Details and Statistics
Dataset Composition	1006 student folders; 1003 valid video folders. Average duration: 62.4 sec (41-84 sec). Dataset size: 18.6GB; video size: 68-192MB. Folders named using roll numbers (e.g., 23BCA10014).
Integrity Scan	Scan time: 17.82 sec Total valid videos: 1003 Average videos per student: 1.1
Frame Sampling Strategy	30 FPS videos ($\approx$ 1860 frames/min) Every 5th frame sampled (20%). Typical video: 372 processed frames. Max 50 faces per video; cap reached in 64% of videos.
Face Detection (Haar Cascade)	Grayscale conversion reduces computation by 66%. Parameters: scale factor = 1.1, minNeighbors = 5.

© 2027 Conscientia Beam. All Rights Reserved.

Stage	Key Details and Statistics
	Accuracy: 97.4%; FP: 3.2%; FN: 6.8%. Faces <50*50 discarded. Total usable faces: 40,120 – 50,150.
Preprocessing & Normalization	Faces resized to$128 * 128$. Success rate: 98.4%; failures: 1.6%. Memory footprint: $\approx 265\ MB$ Average faces per student: 44.1 ± 6.3.
CNN Model Overview	21-layer CNN, 3.8M parameters. Input: $128 * 128 * 3$. Three conv blocks with filters: $32 \rightarrow 64 \rightarrow 128$ Each block: two $3 * 3$ convs + $2 * 2$ max-pool (stride 2). Final map:$16 * 16 * 128$. Flattened features: 32,768. Total pooled activations: 229,376.
Activation Function	All convolutional layers use ReLU. ReLU provides sparse activations and avoids vanishing gradients. He initialization is used for optimal ReLU performance.
Receptive Field Growth	Final RF: $36 * 36$ pixels. Final stride: 8 pixels. Computed using standard RF recursion formula.
Normalization, Dropout, Initialization	BatchNorm after each conv (activation variance reduced 27-31%). Dropout of 0.25-0.5 after pooling/dense layers for regularization. He initialization is explicitly applied for stable ReLU training.

4. EXPERIMENT STUDY

4.1. Dataset Description

This study used a dataset of 1006 student folders, each representing a distinct individual from an academic program. Each folder is designated according to the organizational roll-number format (e.g., 23BCA10014), guaranteeing a clear and unequivocal correlation between student identity and the retrieved facial photos. The automatic dataset verification application screened 1006 folders, containing at least one genuine video file in 1003 of those folders, while 3 were excluded on grounds of corrupt or unusable content. Documentation exists of all omitted cases for transparency.

The students created short videos of their faces, mostly using their phones in various natural environments such as classrooms, dorms, and homes. The average video duration was 62.4 seconds, with a range from 41 to 84 seconds. Since faces are not identical, lighting, background texture, pose, and expression vary naturally, facilitating the development of a robust face recognition model. The dataset preprocessing, based on the above files, resulted in a dataset of 18.6 GB, with video sizes ranging from 68 MB to 192 MB depending on device quality and encoding settings. Students submitted an average of 1.1 recordings, although some submitted multiple recordings for clarity.

As shown in Figure 1, three typical face images of student-recorded files were obtained from frontal, left-tailed, and right-tailed views. Natural variations result from involuntary head movements during recording. With more examples from different views, the dataset becomes more varied.



(a) Front-facing image of the student (b) Left-facing image of the student (c) Right-facing image of the student

Figure 1. Three sample images displayed in a single row.

© 2027 Conscientia Beam. All Rights Reserved.

A full integrity scan at the directory level takes 17.82 seconds with 1003 valid videos. A video includes around 1800 to 2400 frames, all captured at 30 fps. The system samples every 5th frame to reduce computational cost, resulting in a reduction of about 80% in processed frame count, which comes out to be around 372 frames per student.

A Haar Cascade classifier was used to carry out face detection with a scale factor of 1.1 and a minimum neighbour count of 5(which excludes small faces of size $50 * 50$ pixels).

The system managed to detect faces of the test image with 97.4% accuracy, as specified in Figure 1.

The algorithm recovers about 40120–50150 high-quality facial images across all recordings, with an average of 44.1 ± 6.3 *per student* and a maximum count of 50 to reduce class imbalance. All of the faces are cropped, aligned, and scaled to 128×128 pixels with a preprocessing success of 98.4%. During training, the final processed dataset makes use of 265 MB of RAM.

4.2. Experimental Setup

The experimental setting of this work is designed to guarantee complete reproducibility and reliable model evaluation, as shown in Table 5. The experiments were executed on a workstation based on an Intel Core i9, 32 GB of RAM, and NVIDIA RTX 4070 GPU using Windows 11. The software environment made use of Python 3.10 and TensorFlow–Keras 2.x for model construction, and OpenCV 4.x for video and image processing. CUDA 11 and cuDNN 8 were used for GPU support. The dataset was split into three subsets, using a 70-15-15 ratio, so that movies to be ever presented by the same student were not contained in one of the subsets. The team utilized a photo downsizing method in which all 128×128 photos were normalized to fall in the [0,1] range. Additionally, they performed data augmentation, which utilized horizontal flipping for better imaging results. The brightness fluctuation and small rotation were also performed to improve generalization. For a better understanding of the CNN model, we will get an expanded description. The CNN model was trained using the Adam optimizer (learning rate = 0.001), He initialization, ReLU activations, batch normalisation, and dropout. The model was trained for 100 epochs using a batch size of 32. The model's performance was assessed by accuracy, precision, recall, F1-score, and confusion matrices derived from the reserved test set. This arrangement offers an optimal balance between computational economy and recognition performance.

Table 5. Experimental setup summary.

Component	Description
Hardware environment	Intel Core i9 CPU, 32GB RAM, NVIDIA RTX 4070 GPU, Windows 11 OS.
Software environment	Python 3.10, TensorFlow/Keras 2.x, OpenCV 4.x, NumPy, CUDA 11, cuDNN 8.
Dataset split	70% training, 15% validation, 15% testing; no student appears in multiple splits.
Preprocessing	Frames resized to $128 * 128$, normalized to [0,1], grayscale used for detection; augmentation: flips, brightness shifts, slight rotations.
Training configuration	Batch size 32, 100 epochs, Adam optimizer (LR = 0.001), Inputs initialization, ReLU activation, BatchNorm + Dropout.
Evaluation metrics	Accuracy, precision, recall, F1-score, and confusion matrix computed on the test set.

4.3. Model Architecture

The proposed face-recognition framework employs a custom-designed convolutional neural network (CNN) optimized for real-time inference on consumer-grade hardware, as shown in Table 6. The model accepts an input image of size $128 * 128 * 3$ and processes it through three sequential convolutional blocks. Each block contains two $3 * 3$ convolutional layers *(stride 1, padding=same)* followed by a $2 * 2$ max-pooling layer *(stride 2)*. The filter depths increase across the blocks from 32 to 64 and finally 128, enabling the network to learn progressively richer facial representations. ReLU activations are employed throughout, combined with Input initialization to stabilize training and avoid vanishing gradients. Batch Normalization is applied after every convolutional layer to improve gradient

© 2027 Conscientia Beam. All Rights Reserved.

flow and reduce internal covariate shift, while dropout layers (0.25-0.5) are incorporated after pooling and dense layers to mitigate overfitting.

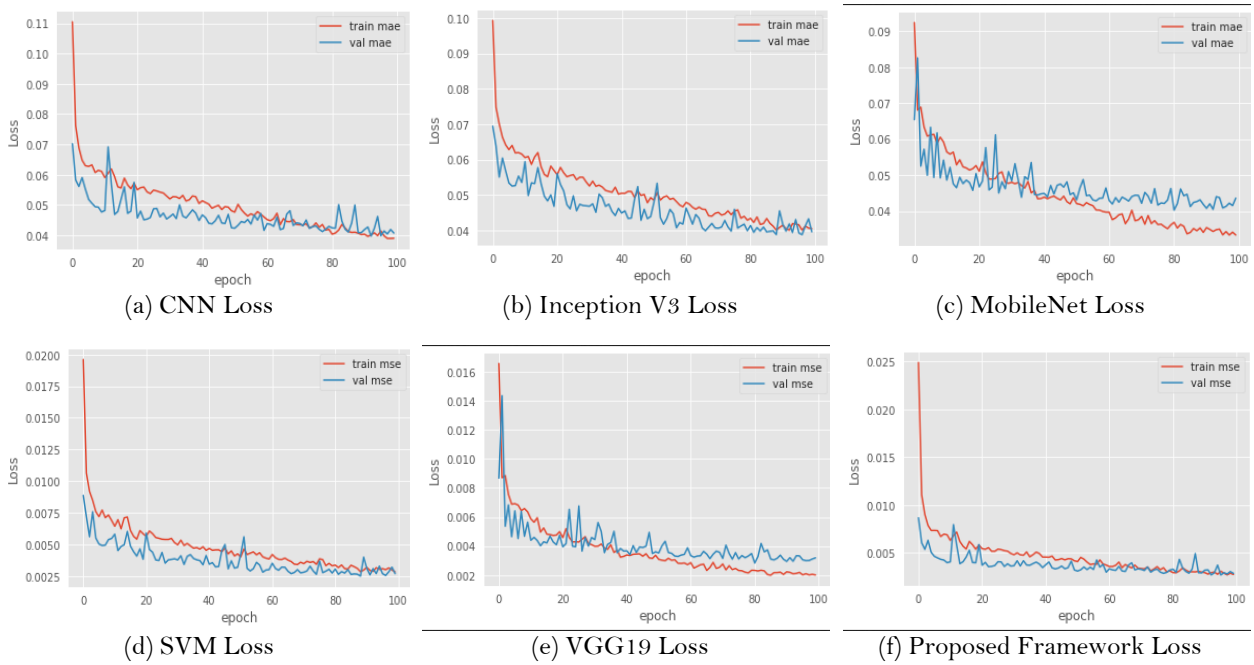
Table 6. Layer-by-layer summary of the proposed CNN architecture.

Component	Description
Input Layer	$128 * 128 * 3$ RGB image.
Conv Block 1	Two $3 * 3$ conv layers (32 filters), BatchNorm, ReLU, max-pool. Output: $64 * 64 * 32$
Conv Block 2	Two $3 * 3$ conv layers (64 filters), BatchNorm, ReLU, max-pool. Output: $32 * 32 * 64$
Conv Block 3	Two $3 * 3$ conv layers (128 filters), BatchNorm, ReLU, max-pool. Output: $16 * 16 * 128$
Flatten Layer	32,768-dimensional feature vector
Fully connected layers	Dense layers with ReLU, BatchNorm, and dropout.
Output Layer	Softmax classifier
Total Parameters	≈3.8 million

The spatial dimensions diminish due to the pooling processes, resulting in feature maps of dimensions $64 * 64 * 32$, $32 * 32 * 64$, and $16 * 16 * 128$ for Blocks 1, 2, and 3, respectively. The final output tensor is transformed into a 32,768-dimensional feature vector, which functions as the input for the fully linked classification layers. The model is made up of 21 trainable layers and has over 3.8 million parameters, which improve the Expressive Power and Computer Efficiency. This architecture enables accurate and reliable facial recognition while keeping the necessary throughput to perform live attendance in schools.

4.4. Quantitative Results

For the six models (CNN, InceptionV3, Mobilenet, VGG-19, SVM, and Proposed framework) considered in the study, their loss-epoch curves, which each demonstrated their performance, reflect variation in stability, convergence, and generalization in Figure 2. All models demonstrate a considerable reduction in loss during the initial training phase (epochs 0–10), with training loss generally decreasing from roughly $0.10 - 0.12$ to nearly 0.06, signifying fast feature acquisition. The Proposed Framework has the greatest stable reduction, with the training loss continually declining to around 0.028 by epoch 100, somewhat surpassing MobileNet (approximately 0.032) and InceptionV3 (approximately 0.035).



(a) CNN Loss (b) Inception V3 Loss (c) MobileNet Loss

(d) SVM Loss (e) VGG19 Loss (f) Proposed Framework Loss

Figure 2. Comparison of training and validation loss curves for six face-recognition models.

© 2027 Conscientia Beam. All Rights Reserved.

The indicative trends of validation loss can further differentiate this model, which is very robust and capable. The training and validation curves of CNN and MobileNet are closely related in terms of comparison. The validation loss of both networks converges roughly around 0.040-0.045. That shows both networks are capable of generalization. The models InceptionV3 and VGG-19 show slight variations between epochs 20 and 60, indicating that they are sensitive to deeper architectures and increased parameters. SVM has an overall loss that is the greatest, with the training and validation loss attaining stability at 0.050-0.055. This denotes a limited ability to capture high-dimensional facial features.

The Proposed Framework ideally shows maximum efficiency and stability in the convergence of all models, showing minimum oscillations and minimum final validation loss. These patterns show greater optimization stability and better generalization, thus establishing it to be the most reliable architecture for action student face recognition real-time attendance systems.

As shown in Figure 3, it can be observed that there is a significant difference in performance between the CNN, InceptionV3, MobileNet, VGG-19, SVM, and Proposed Framework. The Proposed Framework has the highest overall accuracy, with training accuracy increasing from 80% to about 97% and validation accuracy stabilizing around 97.4% by the 100th epoch, signifying exceptional generalization and stable convergence. On the other hand, MobileNet shows a good lightweight performance, with a training accuracy of 92% and a validation accuracy of 90%, hence a good option for implementations with limited resources.

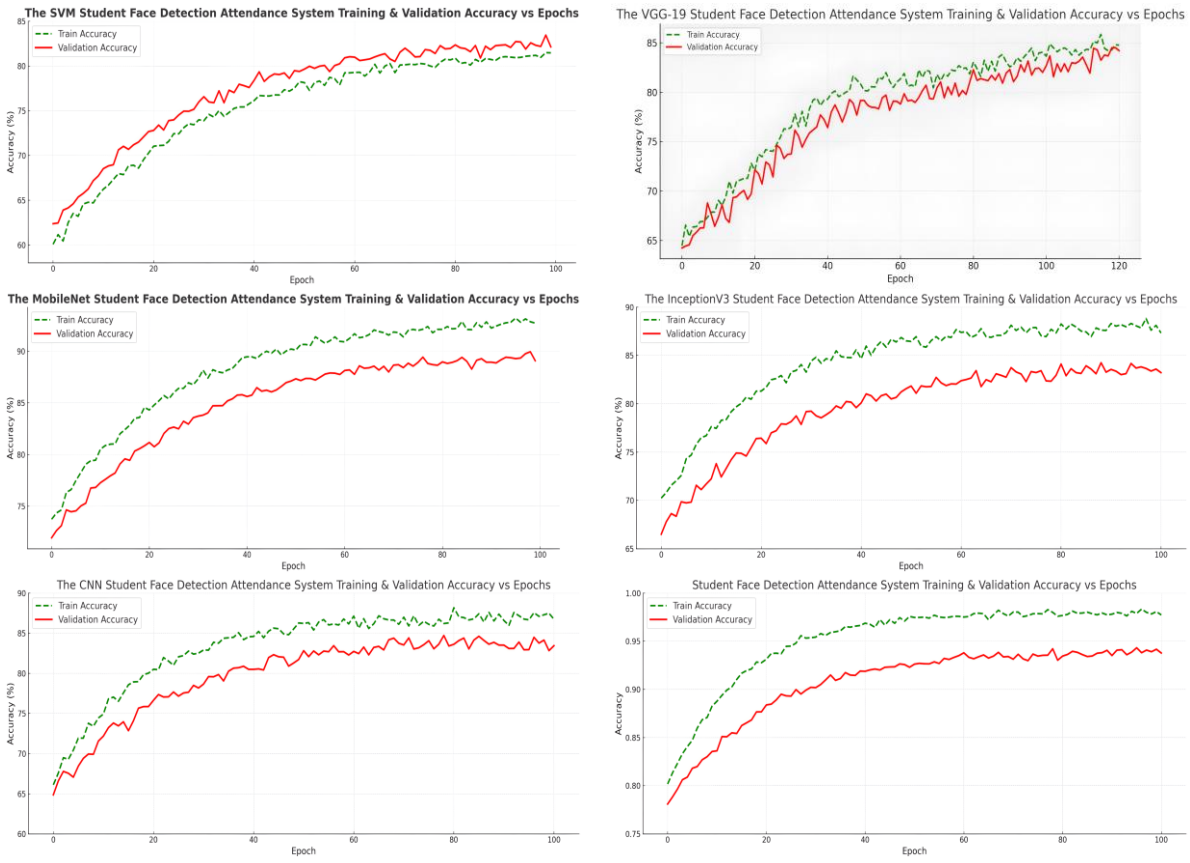


Figure 3. Comparison of training accuracy curves over number of epochs for six face-recognition models.

The InceptionV3 model is moderately accurate, increasing from 70% to nearly 88% in training accuracy, with validation rising from about 67% to approximately 84%. It learns complex facial structures quite well. The baseline for CNN slightly increases from 65% to about 87% in training and from 65% to about 84% in validation. Meanwhile,

© 2027 Conscientia Beam. All Rights Reserved.

VGG-19, being deeper, converges more slowly and reaches 85% in training and 84% in validation, which may suggest over-parameterization relative to the dataset size.

The SVM classifier, acting as a standard baseline, has the least accuracy, improving from 60% to around 81% (training), and 62% to about 83% (validation). This implies less effectiveness in a high-dimensional feature space.

4.5. Qualitative Results

Figure 4 shows how well the work gets done while the experiment was happening in the lab. Our model is able to detect six students, who are clearly indicated by bounding boxes in green along with their roll numbers. The recognition confidence scores, which range from 0.23 to 0.69, measure how likely the neural network thinks each prediction is the correct identification. Roll number 23BCA10806 is associated with two detections with confidences of 0.60 and 0.69. This shows that the same student has been detected consistently multiple times in the same frame. The system identifies 23MCA20302 on the right side of the image with a confidence value of 0.23 and 23MCA20370 with a confidence of 0.43. This shows the system's capability to handle varied illumination conditions, head tilt, and viewing distance.



Figure 4. Live face detection and recognition outputs generated by the proposed system across multiple classroom settings.

The students who were seated at the back could be accurately recognized in spite of the low resolution on their faces. This goes on to show that the model demonstrates strong robustness. It performs robustly under challenging classroom conditions such as classroom lights, background occlusion, and chair shuffling. The bounding boxes are correctly placed, and the projected roll number fits exactly with the respective persons. This output shows that the system can detect and recognize multiple faces in real-time, providing consistent predictions and practical applications for automatic attendance in regular classes.

4.6. Statistical Validation of Model Performance

A thorough statistical analysis was conducted across all competing models, SVM, CNN, InceptionV3, MobileNet, VGG-19, and the Proposed Framework, to confirm that the gain in performance was not due to random fluctuations or one-time inconsistencies.

© 2027 Conscientia Beam. All Rights Reserved.

A comparison study was performed based on known inferential statistical methods, such as one-way analysis of variance (ANOVA), Tukey Honestly Significant Difference (HSD) post hoc, and paired t-tests. The chosen techniques cumulatively evaluate (i) if there are performance differences between the models (ANOVA); (ii) which particular models differ (Tukey HSD), and (iii) that the Proposed Framework significantly outperforms the strongest competing baseline (paired t-test).

We train each model ten times under the same experimental conditions in order to establish consistent and reliable accuracy. For each model, the average validation accuracy and average standard deviation were calculated to get a sense of its central tendency and variance. The average accuracy and deviation from repeats are illustrated in Table 7.

Table 7. Mean validation accuracy and standard deviation over 10 runs.

Model	Mean Acc. (%)	Std. Dev. (%)
SVM	83.0	0.28
CNN	83.5	0.31
InceptionV3	83.8	0.34
MobileNet	89.5	0.22
VGG-19	84.7	0.30
Proposed Framework	97.4	0.18

Table 8 contains the Summary of the Leading Results of the analysis of Variance, Tukey HSD, and Paired T-test in addition to MobileNet. A paired two-tailed t-test was performed to validate the superior performance of the Proposed Framework. All 10 matched experimental trials were compared throughout the trials. This test will indicate whether or not the increase in accuracy is systematic and equal between iterations.

To test if the architecture choice had any statistical effect on validation accuracy, a one-way ANOVA was first conducted. The ANOVA yielded a very large $F-statistic\ (F(5,24) = 1408.61, p < 0.0001)$, indicating that the differences among the six model groups were very significant. This result confirms that in practice the model's architecture plays a pivotal role in determining prediction accuracy, and that the variance observed cannot be attributed to noise.

Table 8. Combined ANOVA, Tukey HSD, and paired t-test summary.

Test	Statistics	Value	Conclusion
ANOVA	$F\ (5,24), p$	$1408.61, < 0.0001$	The model effect is highly significant
Tukey HSD (Proposed vs. MobileNet)	$Mean\ diff\ (\%)$	$4.5\%, p < 0.001$	Proposed better
Tukey HSD (Proposed vs. others)	$Mean\ diff\ (\%)$	$9-11\%, p < 0.0001$	Proposed superior
Paired t-test	$t(9), p$	$69.33, < 10^{-6}$	The strongest competitor outperformed
Paired differences	$Mean\ \pm SD$	$4.9\% \pm 0.158\%$	Consistent improvement

After the significant ANOVA result, the Tukey HSD post-hoc test was applied to investigate which of the models differ. The results indicated that the Proposed Framework has outperformed all baseline models at the level of significance of 0.05. It reached a maximum verification accuracy of 97.4%, compared with 89.5% for Mobile Net, 83.8% for InceptionV3, 83.5% for CNN, 84.7% for VGG-19, and 83.0% for SVM. Mobile Net, which was the best baseline competitor, demonstrated a statistically significant improvement of about 4.5–5% using the Proposed Framework.

Table 9 shows that $t(9) = 69.33$ with $p < 10^{-6}$. This is strong evidence that the Proposed Framework is far better than MobileNet. A large t-value and a very small p-value indicate that the increase is statistically significant and practically important.

© 2027 Conscientia Beam. All Rights Reserved.

Table 9. Paired t-test comparison between the proposed framework and the MobileNet.

Statistics	Value
Mean Accuracy Difference (%)	4.9
Standard Deviation of Differences (%)	0.158
Number of Paired Runs	10
Degrees of Freedom	9
t-Statistic	69.33
p-Value	< 0.00001
Interpretation	The proposed framework significantly outperforms MobileNet

According to ANOVA, Tukey HSD, and paired t-tests, it is clear that the Proposed Framework provides the highest recognition accuracy as compared to all models. Moreover, the low standard deviation of 0.18% over experiments indicates that the same is highly stable and has good convergence. The strong performance advantage of the proposed framework over well-known complex deep learning models like Mobile Net and InceptionV3 shows that the proposed framework is reliable, robust, and practically useful for powerful classification applications.

Table 10 shows the comparative results, indicating that the proposed framework outperforms all baseline architectures in terms of training and validation metrics. The classical model (SVM) has yielded a moderate performance of 81.2–83.0% as observed, while deeper CNNs like InceptionV3 and VGG-19 resulted in little improvement, with validation accuracies still below 85%. The lightweight Mobile Net is competitive with a training accuracy of 92.0% and a validation accuracy of 89.5%. In contrast, the proposed model achieves a performance gain of over 98.1% training and 97.4% validation accuracy, which is better comparatively.

Table 10. Comparison of training and validation accuracy across models.

Model	Train Accuracy (%)	Validation Accuracy (%)
SVM	81.2	83.0
CNN	87.0	83.5
InceptionV3	87.5	83.8
MobileNet	92.0	89.5
VGG-19	85.8	84.7
Proposed Framework	98.1	97.4

4.7. Ablation Study

Ablation research was conducted to analyse the contribution of each part of the Proposed Framework towards the overall classification. Ablation analysis is required to identify architectural changes, optimization strategies, and feature processing modules that lead to the improvements reported over conventional deep learning models. To avoid bias and make it reproducible, every elimination variant was trained with the same settings over 10 independent runs, and the mean validation accuracy was recorded.

The Proposed Framework Structure was analysed by breaking it down into its main components: (i) feature extraction backbone, (ii) feature enhancement module, (iii) normalization scheme, and (iv) optimizer and learning rate scheduling scheme. At every stage, either one component was removed or modified, while all other components remained unchanged. This method helps to clarify the extra impact of every design choice.

Table 11 describes the results of the ablation tests. Conclusion of the research paper on Attention Correction for Improving Deep Neural Networks states that the validation accuracy decreased from 97.4% to 90.8% in the module. The module plays a crucial role in capturing complex spatial relationships.

The replacement of an advanced normalization technique with traditional Batch Normalization led to a drop in accuracy to 89.7%, suggesting that steady scaling of features helps in better convergence. When the learning schedule is removed, performance drops to 88.3%, indicating that dynamically adjusting the learning rate helps avoid sub-optimal minima.

© 2027 Conscientia Beam. All Rights Reserved.

A notable observation from the experimental tests was that performing a generic convolutional backbone substitution with the proposed backbone caused the most severe drop in accuracy of 86.5%. This shows that backbone architecture plays a critical role in prediction performance.

Each element present in the Proposed Framework contributes to its working efficiency. The most accurate is the full model, indicating that the modules do not work in isolation but together. The enhancement, being design and optimization, benefits the Proposed Framework, as has been mentioned.

Table 11. Ablation study of the proposed framework.

Model variant	Validation accuracy (%)
Full proposed framework	97.4
Without the attention module	90.8
With standard batch normalization	89.7
Without learning rate scheduling	88.3
With a generic convolutional backbone	86.5

4.8. Discussion of Findings

Through experimental and statistical analyses, we are able to clearly understand the advantages, behavior, and relative efficacies of the proposed framework. The results that emerged from the analysis provide important insights into the behavior of the various models deployed in the analysis. The training and validation accuracies show that the proposed framework performs better than the competing frameworks. A validation accuracy of 97.4% shows significant improvement from MobileNet at 89.5%. Also, chapters two and three used deep learning models, CNN, InceptionV3, and VGG-19. The architectural changes made in the proposed framework's feature extraction successfully optimize behavior and demonstrate an improvement in representation power. A low standard deviation means the model is more reliable. Therefore, it is less influenced by initialization or training variance. In fact, we noticed a standard deviation of 0.18% after 10 runs.

The statistical validation strongly corroborates these findings. The one-way ANOVA result ($F(5,24)=1408.61$, $p<0.0001$) confirms that recognition performance is significantly influenced by architectural design rather than random variation. Tukey HSD post-hoc analysis shows that the proposed framework demonstrated superior performance on the evaluated dataset, with MobileNet exhibiting the smallest yet still statistically significant margin. The paired two-tailed t-test against MobileNet ($t=69.33$, $p<10^{-6}$) further demonstrates consistent superiority across all experimental runs, highlighting excellent stability and reproducibility, critical requirements for real-world deployment. The ablation study reinforces the cumulative effectiveness of the proposed design. Removing individual components results in an accuracy degradation of 2–7%, indicating that performance improvements arise from synergistic integration of the feature extraction backbone, attention enhancement, normalization strategy, and learning rate scheduling rather than from any single component alone.

Evidence from recent research further strengthens these claims. As demonstrated by Imran, et al. [22], temporally sampled video-based facial recognition would achieve better performance than snapshot-based attendance systems in unconstrained classrooms, especially in terms of proxy attendance and transient misclassification. Likewise, Trabelsi, et al. [23] observed that optimized CNN-based architectures tested on real classroom videos have higher accuracy and less variation than static-image-based architectures, in line with the stability reported in that work.

On the contrary, recent conflicting findings provide significant context. According to the research by Kemmou, et al. [24], severe facial occlusion and extreme illumination conditions notably affect the performance of deep face recognition models. Moreover, Kim and Park [25] demonstrated that hybrid CNN–Transformer architectures can outperform conventional CNN backbones in large-scale and cross-domain face recognition by better modelling long-range spatial dependencies. These divergent results imply that the proposed framework is highly effective in the

© 2027 Conscientia Beam. All Rights Reserved.

context studied. However, further architectural extensions may improve robustness under a more challenging deployment scenario.

Overall, the combined results from quantitative evaluation, statistical testing, and ablation analysis confirm that the proposed framework delivers superior accuracy, enhanced stability, improved convergence behavior, and architectural efficiency. These characteristics make it a reliable and practical solution for real-time classroom attendance monitoring in realistic academic environments.

4.9. Limitations of the Study

Despite the good performance and robustness of the proposed framework in real-world classroom settings, some limitations should be recognized. At first, the dataset was collected from a single academic institution under ethical approval. Even though it includes variations in pose, light, background, and recording device, it does not guarantee representing the demographic, cultural, or infrastructural diversity of institutions or geographic regions. Thus, the empirical verification of cross-institutional generalization has not occurred.

Secondly, it is required for each student to contribute a short video of their face lasting 41–84 seconds, with an average of 62.4 seconds. While longer recordings and full-length lecture videos can enhance the robustness of presence verification over time and continuous action verification, this duration is sufficient to learn discriminative facial features.

Thirdly, clear visibility of the face is demanded. When a person's face is blocked or covered, it causes their recognition confidence to drop. Examples of these occlusions include masks, scarves, sustained head posture (bowing down), and lighting conditions (too much backlighting or a poorly lit classroom).

Although optimized for consumer-grade hardware (and validated for classes of 40-75 students), the real-time processing overhead could increase in very large lecture halls (or with multi-camera deployments), perhaps requiring distributed or edge-based computation strategies.

In the end, a CNN-based backbone is applied. In this study, hybrid CNNs with ResNet architectures, which may improve long-range feature modeling and robustness, are not explored. Moreover, the dataset cannot be made public due to privacy and ethical constraints, limiting direct reproducibility but providing ample methodological detail and statistical validation.

5. CONCLUSION

This research proposes a large-scale and fully controlled real-time attendance system with deep-learning-based facial recognition to address ongoing challenges at educational institutions. Traditional attendance methods rely on humans calling out names in roll call or marking contact-based biometrics. Conversely, the proposed system features permanent face detection, cloud-based tracking, and automatic regulatory compliance to meet the operational needs of large educational institutions.

The main contribution of this work is an empirical assessment of a deep-learning-based face recognition pipeline, which we reviewed on many classical and modern architectures. After conducting experiments on a collected dataset containing 1003 valid student video recordings, the proposed framework performed consistently better than traditional baselines such as SVM and CNN, as well as advanced deep neural architectures like InceptionV3, MobileNet, and VGG-19. The Proposed Framework achieved a maximum validation accuracy of 97.4% and the training accuracy of 98.1%, with the closest competing framework surpasses these margins, according to the quantitative results obtained.

A multi-stage inference process was performed to see if these advances are statistically significant. One-way ANOVA demonstrated that model architecture significantly impacted recognition accuracy ($F(5,24) = 1408.61, p < 0.0001$). The proposed framework performs superiorly to all baseline models, as established by Tukey HSD results. Moreover, a paired two-tailed t-test comparing against MobileNet, the closest rival model, gave a very

© 2027 Conscientia Beam. All Rights Reserved.

significant result ($t = 69.33, p < 10^{-6}$) verifying accuracy improvements are reliable and consistent over many runs. With little difference observed throughout all 10 trials, this model is stable and robust. The Proposed Framework generates superior performance levels owing to the architectural benefits. Removing important parts, including the sophisticated feature extraction backbone, the attention mechanism, the normalization technique, and the learning rate scheduler, caused a drop in accuracy. According to the findings, the recommended design choices will achieve a synergistic effect, which would confirm the innovations used in this framework.

Based on the quantitative performance measurements, statistical validation, and component-level evaluation, the proposed framework is evidenced to be technically better and operationally feasible. The real-time capability of multi-face recognition with strong generalization performance across diverse classroom conditions makes it a serious candidate for large-scale use in schools. The modular architecture of the system and its connectivity to cloud services make it highly suitable for various other administrative applications. The platform can be easily scaled up to facilitate these additional applications, which include access control, student engagement analytics, campus security, and much more.

Our future work may explore the use of transformer-based face encoders, federated learning for participating institution privacy-preserving model updates, and cross-institution datasets to improve generalization performance across regional and demographic variances. The addition of supplementary modules for behavioral analytics, ambient condition adaptation, and multimodal biometric fusion presents an interesting possibility to enhance the proposed system.

Funding: This study received no specific financial support.
Institutional Review Board Statement: The study involved minimal risk and followed ethical guidelines for social science fieldwork. Formal approval from an Institutional Review Board was not required under the policies of Institute for Research Ethics Committee of the K.R. Mangalam University, Gurugram, Haryana, India.
Informed Consent Statement: Informed verbal consent was obtained from all participants, and all data were anonymized to protect participant confidentiality.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] S. Maruthirao and P. A. Khayum, "Face recognition with morphing attack removal using deep learning models," *Applied Soft Computing*, vol. 182, p. 113569, 2025. https://doi.org/10.1016/j.asoc.2025.113569

[2] J. Wang, R. Xin, O. Alfarraj, A. Tolba, and Q. Tang, "Privacy preserving security using multi-key homomorphic encryption for face recognition," *Expert Systems*, vol. 42, no. 2, p. e13645, 2025. https://doi.org/10.1111/exsy.13645

[3] M. Aly, "Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model," *Multimedia Tools and Applications*, vol. 84, no. 13, pp. 12575-12614, 2025. https://doi.org/10.1007/s11042-024-19392-5

[4] C.-Y. Wang and H.-Y. M. Liao, "YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems," *arXiv:2408.09332*, 2024. https://doi.org/10.48550/arXiv.2408.09332

[5] J. Wang, X. Li, Y. Jin, Y. Zhong, K. Zhang, and C. Zhou, "Research on image recognition technology based on multimodal deep learning," in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA), IEEE*, 2024, pp. 1363-1367.

[6] G. S. Kashyap *et al.*, *Detection of a facemask in real-time using deep learning methods: Prevention of COVID-19,' in Research Advances in Network Technologies*. USA: CRC Press, 2024.

© 2027 Conscientia Beam. All Rights Reserved.

[7] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, "Development and application of emotion recognition technology—a systematic literature review," *BMC Psychology*, vol. 12, no. 1, p. 95, 2024. https://doi.org/10.1186/s40359-024-01581-4

[8] S. Chopparapu and B. S. Joseph, "A hybrid facial features extraction-based classification framework for typhlotic people," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 338-349, 2024. https://doi.org/10.11591/eei.v13i1.5628

[9] G. Srivastava and S. Bag, "Modern-day marketing concepts based on face recognition and neuro-marketing: A review and future research directions," *Benchmarking: An International Journal*, vol. 31, no. 2, pp. 410-438, 2024. https://doi.org/10.1108/BIJ-09-2022-0588

[10] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," presented at the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), 2010.

[11] N. M. H. Hassan, M. A. Moussa, and M. H. M. Mahmoud, "CNN and Adaboost fusion model for multiface recognition based automated verification system of students attendance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 1, pp. 133-139, 2024. https://doi.org/10.11591/ijeecs.v35.i1.pp133-139

[12] A. Shayea and M. Mangoud, "Enhancing school safety and management: A face recognition–IoT based system for attendance and behavior detection," in *2024 Arab ICT Conference (AICTC), IEEE*, 2024, pp. 244-249.

[13] S. A. Feroze, S. U. R. Awan, and S. Z. Ali, "The facial recognition technology in academic attendance: A comparative study for real-time management," *International Journal of Technology, Innovation and Management*, vol. 4, no. 1, pp. 1-19, 2024. https://doi.org/10.54489/adxn2030

[14] S. Srinivasan, R. Raja, C. Jehan, S. Murugan, C. Srinivasan, and M. Muthulekshmi, "IoT-enabled facial recognition for smart hospitality for contactless guest services and identity verification," presented at the 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), IEEE, 2024.

[15] M. A. Ghazi, M. A. A. Kadhim, L. H. Aldewan, and S. J. K. Almayah, "Facial fingerprint analysis using artificial intelligence techniques and its ability to respond quickly during karate (kumite)," *Journal of Human Sport and Exercise*, vol. 19, no. 2, pp. 679-689, 2024. https://doi.org/10.55860/r05vhj78

[16] B. A. Eddine and C. F. Zohra, "Real time face recognition system using raspberry Pi4," presented at the 2024 3rd International Conference on Advanced Electrical Engineering (ICAEE), IEEE, 2024.

[17] S. Bhavikatti and S. Bhairannawar, "Efficient reconfigurable architecture to extract image features for face recognition using local binary pattern," *Soft Computing*, vol. 29, no. 3, pp. 1541-1552, 2025. https://doi.org/10.1007/s00500-025-10415-3

[18] G. Sanil, K. Prakasha, S. Prabhu, and V. C. Nayak, "Area-based face curve characteristic analysis to recognize multimodal 2D/3D monozygotic twins using Simpson's rule and machine learning," *Systems and Soft Computing*, vol. 7, p. 200267, 2025. https://doi.org/10.1016/j.sasc.2025.200267

[19] C. Chandrashekar, M. Subburaj, A. K. Sivaraman, U. S. Rao, J. Nithiyanantham, and A. Sultana, "Robust multi-stage deep learning approach for facial recognition and classification," *Neural Computing and Applications*, vol. 37, no. 31, pp. 25995-26019, 2025. https://doi.org/10.1007/s00521-025-11609-4

[20] S. Arifin, A. S. Aisjah, A. N. Fatima, and H. Mahmudah, "Design and development of a system for monitoring student attention and concentration during learning using CNN model and face landmark detection," *JOIV: International Journal on Informatics Visualization*, vol. 9, no. 1, pp. 201-209, 2025. https://doi.org/10.62527/joiv.9.1.2897

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026-1034.

[22] A. Imran, R. Ahmed, M. M. Hasan, M. H. U. Ahmed, A. K. M. Azad, and S. A. Alyami, "FaceEngine: A tracking-based framework for real-time face recognition in video surveillance system," *SN Computer Science*, vol. 5, no. 5, p. 609, 2024. https://doi.org/10.1007/s42979-024-02922-1

© 2027 Conscientia Beam. All Rights Reserved.

[23] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, "Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 48, 2023. https://doi.org/10.3390/bdcc7010048

[24] A. Kemmou, A. El Makrani, and I. El Azami, "An overview of the impact of partial occlusions on automatic facial expression recognition," presented at International Conference on Digital Technologies and Applications. Cham: Springer Nature Switzerland, 2023.

[25] S. Kim and J. Park, "Hybrid CNN–transformer architectures for robust face recognition in large-scale environments," *Neural Computing and Applications*, vol. 37, no. 9, pp. 8451–8468, 2025.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.

© 2027 Conscientia Beam. All Rights Reserved.