Ingénierie des Systèmes d'Information
Vol. 30, No. 7, July, 2025, pp. 1893-1901
Journal homepage: http://iieta.org/journals/isi

Design and Development of Automated Student Attendance Framework in Fusion of CNN, HAAR, and ResNet

Rajesh Yadav[1*], Swati Gupta[2], Meenakshi Malik[3], Ibrahim Aljubayri[4], Chander Prabha[5], Mohammad Zubair Khan[4]

[1] K.R. Mangalam University, Gurugram 122103, India
[2] Member of Centre of Excellence AI, K.R. Mangalam University, Gurugram 122103, India
[3] Department of Computer Science and Engineering, BML Munjal University, Gurugram 123106, India
[4] Department of Computer Science and Information, Taibah University, Madinah 42353, Saudi Arabia
[5] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab 140104, India

Corresponding Author Email: himank9d2006@gmail.com

https://doi.org/10.18280/isi.300721

Received: 7 May 2025
Revised: 18 July 2025
Accepted: 25 July 2025
Available online: 31 July 2025

Keywords:
CNN, face recognition, multi-model fusion, OpenCV, real time attendance tracking system, ResNet, U-NET, HAAR cascade

ABSTRACT

Traditional university attendance systems, whether manual or biometric, are generally inefficient, prone to fraud, and have large operational costs. This research study solves these issues by providing an automated attendance tracking system based on facial recognition, which eliminates the need for human supervision while increasing precision. To recognize and extract facial features, the system uses a fused deep learning model that combines ResNet-based Convolutional Neural Networks (CNN), pretrained U-NET, and HAAR cascade techniques. The model was trained using a dataset of 1,120 facial photos per participant, which included nine and eleven-layer CNN architectures with a variety of activation functions such as ReLU, SoftMax, and Tanh. The system, built with Python and OpenCV, extracts 68 facial landmarks per face and functions under a variety of lighting and environmental circumstances. The suggested algorithm achieves 97.81% accuracy in recognition while significantly lowering false positives by 3.03%, 2.03%, and 1.48% when compared to ResNet18, ResNet34, and ResNet50. Furthermore, the computational efficiency of the TensorFlow and CoreML frameworks was evaluated in order to determine their suitability for implementation on embedded devices. The findings show that the approach is effective in real-time attendance settings and has the potential to improve existing institutional tracking systems.

1. INTRODUCTION

Attendance systems might be manual or automated. An event manager manually records attendance via a manual attendance method. Computerized attendance systems eliminate the tiresome task of manually tracking attendance. Attendance can be verified and recorded using authentication. Face recognition excels other biometric authentication systems regarding availability and discreet engagement [1]. Biometry identification is safe and accessible since it verifies identity using physiological and behavioral traits rather than knowledge or possession.

Knowing encrypted passwords allows an individual to replicate the identity of another user. Another individual can be verified through authorization-based authentication. Like credential-based and authorization-based authentication methods, biometric authentication systems, such as facial recognition, can present safety risks [2].

Forensic face recognition uses the physical characteristics of people. Profiling usually involves recording or photographing the appearance of the subject and using algorithms to extract jawline, distance between eyes, and nose curvature [3]. User impersonation is important to password-based authentication. They can also be verified by ownership-based authentication.

Research reveals that most facial recognition tools are susceptible to fraud attempts involving a perpetrator imitating an authorised individual to gain unauthorized use [4]. Face-resembling assaults are classified through three distinct categories in the study [5]. In level one, facial faking is performed with minimal or no before utilizing a mobile phone display or paper reproduction. The assailant must employ video footage of a face blinking and moving to prepare for Level B face spoofing. Level C: Attackers must use specialized equipment to create ultra-realistic 3D masks. For improving the privacy of a portable device's student's attendance system using face recognition, a liveness detection system that utilizes a Convolutional Neural Network (CNN) to analyses facial signals has been suggested [6].

Following are several notable advancements in the amount of face recognition-based student attendance systems:

- A convolutional neural network model for accurate face detection that uses transfer learning to build upon a model that has been trained

• Detection of continuous faces using an attendance system that uses live recognition of facial features

The subsequent content provides a concise overview of the following sections of the paper: Section 2 covers articles published in recent times on face detection and recognition. Section 3 discusses the proposed research methodology and includes data collection, image pre-processing, face detection, feature extraction and evaluation. Section 4 contains the suggested system's experimental setup. Assessment of suggested system performance. Section 5 assesses the framework's performance and compares it to other models. Section 6 contains information on ethical consideration and Section 7 concludes with further development.

2. LITERATURE REVIEW

In the study [7], the authors used Viola-Jones face detection, LDA, and machine learning for automatic face identification. This study covers five sections of a self-designed dataset of student images in different settings. Histogram equalization converts color pictures to grayscale and stores them in the database for future use. In the second part, the authors examined the identification of faces using the Viola-Jones Method, which involves transforming features using a fixed window that slides. HAAR [8] and Adaboost divide features into two groups based on threshold values. In the third segment, LDA lowers dimensions and finds the scattered matrix's Eigenvectors' maximum values to extract features and maximize discriminating analysis between inter and intraclass changes. Final recognition uses the KNN classifier.

In the study [9], the authors used deep learning to track attendance by logging students' and faculty's arrival and leave times into a database. At this stage, we've collected all image data into a tensor with pixel labels and greyscale values. We develop a 64-unit convolution layer with a 3-size sliding window kernel. The SoftMax activation function powers the model's 2-layer dense neural network student recognition mechanism. The gradient vanishing issue was solved with dropout layers. Using real-time video feeds, the motion triggered system can detect small and noticeable brightness changes between frames. Removing the background lets the technology manage dynamic intensity. Real-frame intensity fluctuations are analysed using time-dependent factors. We used hyperparameters to examine three situations and periods. In the study [10], the authors created an automatic attendance site using photo-based facial recognition. Students' server steps can be tracked by a classroom HD webcam. The student's database photos are matched to frames. Our multilayer perceptron model has many normalization stages. Even in large classrooms with many kids, this model remains accurate.

The researchers in their study [11] developed the" ARRAY" framework to track workplace attendance and facial recognition. This study proposes an improved LBP variant that uses contrast adjustment to focus on key traits using beta and alpha values. For this research, the HAAR classifier is used to train the system by means of mixing. This blending operation increases picture quality and permits precise prediction even in low-light conditions by integrating many photos of a worker. After dividing the face into pieces, LBP employs threshold levels to replace pixels with binary data. After converting values to feature vectors, histogram bins are created to locate robust face features.

The suggested system detects if pupils are present and records their attendance in the event of a class room session by combining U-NET, HAAR, ResNet, and CNN. It extracts characteristics from several areas of the face, including those around the eyes, upper nose, and other parts.

3. RESEARCH METHODOLOGY

Due to environmental constraints, acquired photos in real-world scenarios like classroom monitoring, surveillance, or mobile-based face identification sometimes have low resolution, poor illumination, or partial occlusion. These conditions impair the quality of facial characteristics, making successful detection and recognition difficult, particularly for deep learning models that rely on spatial resolution.

Traditional CNN designs, such as ResNet or HAAR-based detectors, are quite effective when fed high-quality frontal facial photos. However, their performance worsens when face input is distorted by size, noise, or pixel sparsity. These restrictions necessitated the development of a preprocessing module capable of recovering and improving facial detail from low-quality input images.

To solve this, we used a U-NET-based generator architecture, which has been extensively tested for reconstructing and enhancing spatial information from coarse, low-resolution images. U-NET, which was originally created for biological image segmentation, has demonstrated extraordinary accuracy in keeping fine-grained information while up-sampling. Recent researches [8, 12] demonstrate U-NET's utility in super-resolution facial augmentation, expression identification, and low-light recognition, hence supporting its application in facial recognition pipelines in limited situations.

In our approach, U-NET acts as a super-resolution enhancer, upscaling images from 64 × 64 to 256 × 256. This enables HAAR and ResNet-based recognition layers to operate on feature-rich facial inputs. This multi-model synergy improves overall system robustness and allows for consistent recognition across a variety of classroom circumstances.

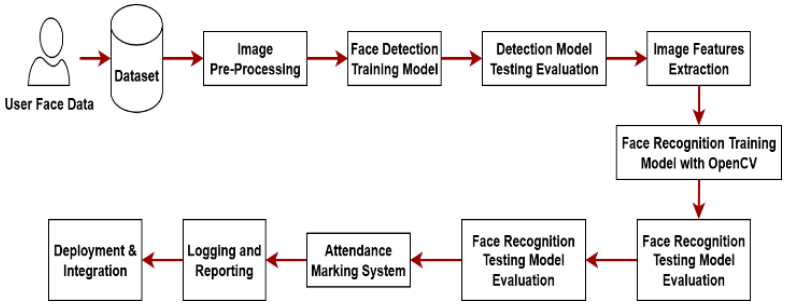


Figure 1. Proposed architecture of face detection and face recognition framework

The fundamental principles for the development and utilization of such systems include the provision of comprehensive data protection, the provision of informed consent, and the implementation of frequent technical updates to address potential biases. This approach necessitates the registration of unique IDs, the generation of databases, the detection of faces, the training of features, the recognition of faces, and the development of a graphical user interface. The architecture and processes of the system are delineated in this section. Figure 1 illustrates that the system comprises two algorithms: face detection and recognition.

3.1 Data collection

Our research is significantly influenced by the data obtained from images. The dataset consists of 1120 student face photographs, each captured from three distinct angles (front, left, and right). The pictures of students' faces have been saved in the format student roll number.jpg and labelled with student name and student roll number. A numerical value is assigned to each angle orientation in the file name. The folder name illustration is depicted in Figure 2, and the following figures and angles are as follows.

Figure 2. Sample image collection of student facial from three different angles

3.2 Image preprocessing

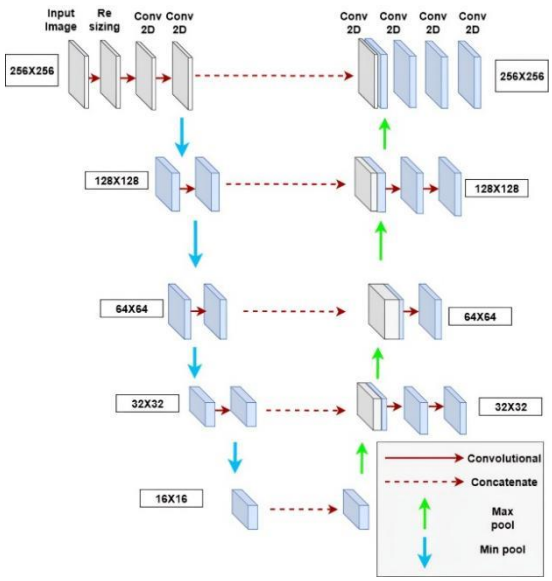


Figure 3. Detailed architecture of U-NET style image generator

Super-resolution in subject photos is often needed and complicated. Low-resolution photos require super-resolution to provide more precise results. Deep learning face recognition algorithms have been studied for image super-resolution; however low-resolution face detection is difficult. To address this issue, researchers proposed a generator architecture similar to the U-NET [12]. Our generator architecture uses the nearest neighbor method to scale a 64 × 64 input image to 256 × 256 mentioned in Figure 3. Our model needs this architecture to produce high-quality photos.

The researchers employed image enhancement and scaling. Initially, the contrast of the image was enhanced by histogram equalization. We utilized feature scaling to scale our datasets to 128×128 dimensions using RGB images. We normalized the dataset and scaled the photos as part of feature scaling. The CNN is terminated more rapidly by data between 0 and 1 than data between 0 and 255. This mitigates disparities in illumination.

3.3 Face detection

Oriented gradient histograms are "Feature Descriptors" in feature engineering, which extract structure-based information from images. This procedure divides the image by calculating histograms and defining local sections. Two matrices are used to calculate gradients: magnitude changes in relation to the x and y axes. Calculating the gradient change in the x-direction is illustrated in Figure 3.

Pixels can represent the nasal component as a matrix. The highlighted pixel is updated by subtracting the correct value from the left side due to a horizontal change. All matrices change in both directions. Pythagoras orients pixels. Next, we generate a histogram from pixel continuous value categories. Despite being older than other detection systems, HAAR remains relevant due to its advanced computational capabilities in evaluating static and moving photos. A simpler matrix is obtained via convolution of photos with a kernel of varied scales. Non-facial data points are "no" and facial data points "yes". Nose, eye, and cheek detection require these levels.

3.4 Feature extraction

Constructing deep convolution neural networks with vast datasets is time-consuming and expensive. The computational feasibility and time are reduced by reusing model weights from pre-trained models constructed as benchmarks on large datasets such as ImageNet [13]. The model may employ a number vector to characterize a feature in Figure 4 by analyzing the characteristics of the retrieved image. Although there may be superior image recognition models, we will select the most effective classification model due to the following:

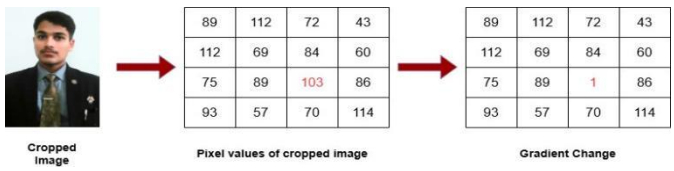


Figure 4. The rate of change of specimen gradients in the X direction for a particular pixel

3.4.1 VGG16

On the ImageNet dataset comprising 14 million images and 1,000 classifications, one of the numerous prominent deep convolutional neural networks employed for image identification achieved a 92.7%. 16 of the 41 layers in the

network can be trained with weights, including 13 convolutional and 3 fully connected, and can accommodate (224, 224, 3) images. The VGG-16 design with 138,357,544 parameters is illustrated in Figure 5.

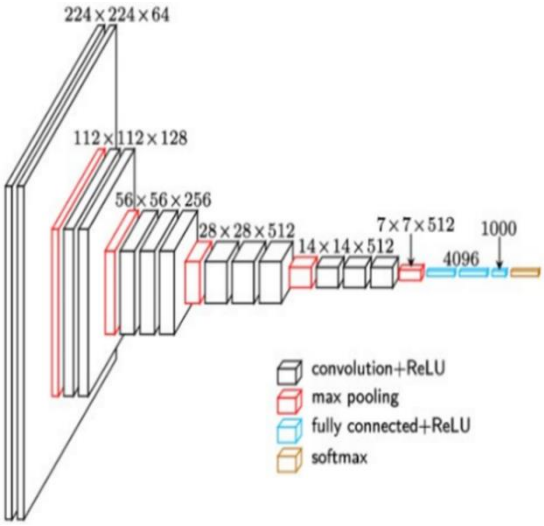


Figure 5. Detailed architecture of VGG16

3.4.2 Face recognition using transfer learning with ResNet

Deep residual network ResNet50 is a widely used architecture for the training of highly complex neural networks. Residual blocks facilitate the training of deep networks and mitigate the vanishing gradient issue. The network employs convolutional layers with skip connections in each block to facilitate efficient training with a high level of depth.

$$Output = F(x) + x \tag{1}$$

$$GAP = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W} FeatureMap(i,j) \tag{2}$$

$$Output = Activation(Weights * Features + Biases) \tag{3}$$

This summing procedure (Eq. (2)) is indispensable for residual learning. The model simplifies the spatial dimensions of the previous layer into a vector representation by calculating the average of each feature map using Global Average Pooling (GAP) (Eq. (3)). Lastly, the fully connected layer linearly modifies features from antecedent layers using learning biases and weights (Eq. (3)). This results in forecasts or categorization. Our ResNet50 model employed weights trained on 'ImageNet' to utilize the collected features. The model was subsequently refined on our dataset using a fully connected layer with three neurons and a global average pooling layer.

3.4.3 Face recognition using transfer learning with inception v3

The Inception V3 TL model is renowned for its design, which incorporates numerous 'Inception' elements. These modules excel in photo categorization due to their simultaneous processing at differing sizes. As illustrated in Eq. (4), Inception modules can perform numerous concurrent convolutions. The kernel diameters of its convolutional layers are variable. The network may acquire features at varying scales by incorporating convolutional processes with varying kernel sizes and pooling. The GAP layer averages each feature map, similar to ResNet50, to convert the spatial output from the antecedent layer to a vector (Eq. (4)).

$$O/\mathrm{P} = \mathrm{Cancat}\big(Conv(1), Conv(3), MaxPool(3 * 3)\big) \tag{4}$$

The three-neuron final layer (Eq. (7)) linearly transforms the learned biases and weights to generate predictions or classifications. Ultimately, this is done.

3.4.4 MobileNet

MobileNet, a lightweight TL model [14], accelerates the inference process for mobile and embedded devices [15]. It is characterized by its diminutive model size and low computational expense. The convolution of each channel determines depth-wise convolution *i* and filter k using distinct filters. The input, depth-wise filter, and output are defined in Eq. (5), with the input being I*i,j* and the depth-wise filter being K*i,j*. The pointwise convolution applies (1 * 1) convolutions across the depth of the intermediate feature maps from the depthwise convolution. The intermediate feature map $\bar{I}i,j,c$, weights W*c,k*, bias b*k*, and output feature map are illustrated in Eq. (10). The GAP layer generates a vector representation by aggregating the values of all feature maps across all spatial dimensions, as described in Eq. (6).

$$l_{i,j,k} = l_{i,j} * K_i, k \tag{5}$$

Lastly, Eq. (6) illustrates the linear transformation of GAP vector representations by the fully connected layer, which employs weights and biases. Classification is achieved through modification.

$$O_{i,j,k} = \sum_{c=1}^{c} I_{i,j,c} * W_{c,k} + b_k \tag{6}$$

3.5 Evaluation

The authors evaluated our photo classification models by utilizing the F1 score, precision, macro average accuracy, and recall. The all-class mean in Eq. (7), Macro Average Accuracy, determines accuracy. It is essential for the assessment of model efficacy. Macro average recall across all classes is quantified by Eq. (8). The model's ability to accurately categories instances of each class is assessed. The Macro Average Precision determines the class-wide precision (Eq. (9)). The model's capacity to forecast class outcomes is assessed. The average F1 score across all courses, including precision and recall, is determined by the Macro Average F1 Score (Eq. (10)). It evaluates the performance of the model fairly. Here, the number of classes is denoted by N. The negatives and true positives for class i are represented by TN*i* and TP*i*. In the class *i*, false positives are defined as FP*i* and false negatives as FN*i*.

$$macro\_acc = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{7}$$

$$macro\_rcall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{8}$$

$$macro\_precision = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \quad (9)$$

$$macro\_f1\_core = \frac{1}{N}\sum_{i=1}^{N}\frac{2.TP_i}{2.TP_i + FP_i + FN_i} \quad (10)$$

4. EXPERIMENTAL SETUP

The face was cropped in this research using OpenCV. Consequently, photographs were reduced by the default input size of 224 × 224 for VGG16, ResNet50, InceptionV3, and MobileNet. Adam optimizer parameters: beta1 0.9, beta2 0.009, learning rate 0.0005. The data was meticulously processed, and the following parameters were added to enhance it: Horizontal flip, × 1.1 scaling factor, and rotation of (-10° to 10°). This minor alteration enhances the veracity of the image. In a tenfold cross-validation (CV), 90% of photos from a standard database (e.g., KDEF [16] or JAFFE [17]) were selected at random for the training pictures and 10% as test images. A 10-fold CV separates photos into ten identical sets, one test and nine training. Average: ten runs. Test set accuracy, which responds to unseen data, measures recognition system performance. Each test was run on a Windows PC with 3.5 GHz Processor and 16 GB RAM.

5. RESULT AND ANALYSIS

The suggested model is tested on comparable datasets in this part. Since CNN forms the basis of the proposed method, we begin by conducting extensive CNN trials to establish a performance standard. Try different VGG-16 fine-tuning parameters. Finally, we compare the proposed model to pretrained ones. The accuracy of a Convolutional Neural Network with multiple layers, a 3x3 kernel, and 2x2 MaxPooling is shown in Table 1 for the Testing Group algorithm. The datasets evaluated include KDEF and JAFFE, with input sizes ranging from 360 × 360 to 48 × 48. Further, to ensure that our suggested attendance paradigm is reliable across multiple populations, we tested its recognition accuracy on two large-scale, ethnically diverse datasets: CelebA and Racial Faces in the Wild (RFW). The CelebA dataset contains over 200,000 celebrity photos annotated with variables such as gender, age, and ethnicity, offering a large testing ground for assessing model robustness under different expressions, lighting, and position situations. The RFW dataset, on the other hand, is specifically designed to assess face recognition bias across various racial groups, including Caucasian, Asian, Indian, and African. Its major goal is to assess the fairness and demographic performance disparities in facial recognition systems. The test magnitude was determined by randomly selecting 10% of the available data. We show the optimal test set accuracies for a configuration for all 50 iterations. The table implies that larger input photographs maximize accuracy for both datasets. Using the same dataset, KDEF accuracy is 74.81% for 360 × 360 input images and 65.78% for 48 × 48 input images. Larger photos include more data, so categorization methods work better. The most extensive input size (360 × 360) did not achieve optimum accuracy. The most accurate image resolutions for both datasets were 128 × 128. The model fits the photo input size, but greater input sizes require a larger model and more data.

Table 1. Accuracy of two-layer conventional convolutional neural networks on the KDEF, JAFFE, CelebA, and RFW databases for images of varying input sizes

Size of Inputted Image	KDEF	JAFFE	CelebA	RFW
360 × 360	74.81	92.43	70.36	67.18
224 × 224	74.89	88.80	71.92	68.75
128 × 128	81.13	92.45	76.31	72.44
64 × 64	70.12	85.34	68.25	65.90
48 × 48	65.78	82.12	63.10	60.84

To test the generalizability of our proposed paradigm across demographically different groups, we expanded our studies to incorporate the CelebA and Racial Faces in the Wild datasets. CelebA and RFW provide more range in pose, illumination, age, gender, and racial dispersion than the JAFFE and KDEF datasets, which are limited in ethnic variance and controlled conditions, respectively. As expected, these datasets had slightly poorer recognition accuracy due to greater intra-class variability and real-world noise.

The model recognized faces with a maximal accuracy of 76.31% on CelebA and 72.44% on RFW at an input resolution of 128×128 pixels, indicating robustness in uncontrolled, diverse situations. Lower accuracy on the RFW dataset can be attributed to the challenges posed by racial imbalance and dataset complexity, which are consistent with current studies. Nonetheless, the performance is still competitive, demonstrating the adaptability of our U-NET + HAAR + ResNet infrastructure across several ethnic groups. These findings support the viability of our technique for use in real-world attendance systems where demographic fairness is crucial.

Various fine-tuning methods were implemented to understand their function in the proposed TL-based system. Table 2 illustrates the accuracy of the VGG-16 model on the KDEF and JAFFE datasets with a 10% random sample, subject to various fine-tuning options. In any fine-tuning mode, we trained dense layers for 10 iterations. Subsequently, we added Conv blocks to the pre-trained foundation for 40 iterations, resulting in 50 iterations in the model. The results of the complete model's training from inception (randomly initialized) for 50 iterations are presented to evaluate the efficacy of the TL based technique. The table demonstrates that the pre-trained model's final block (Block 5) yields significantly superior results when fine-tuned than when fine-tuned solely with dense layers. While fine-tuning, it is essential to contemplate the entire VGG-16 Base, not just Block 5. The highest accuracy is 96.78% in the KDEF dataset and 100% in the JAFFE dataset. In contrast to the TL-based fine-tuning mode, the KDEF and the JAFFE datasets exhibit extremely low accuracies when the complete model is trained from scratch: 23.35% and 37.82%, respectively. The proposed TL approach was validated in the findings table, and a pre-trained DCNN model component was refined.

A remarkable result is that a deeper model outperformed other, as indicated in Table 2. For example, ResNet-152 achieved better accuracy than ResNet-18 on the KDEF test set for the 10-Fold CV case, with 97.12% and 94.17%, respectively. DenseNet-161 surpassed the pre-trained DCNN models on both datasets. In ten 10-fold CV iterations, the model accurately identified 1122 KDEF and 1 JAFFE samples. The presented architecture accurately identified

students by employing face detection and recognition techniques (refer to Figure 6).

Table 2. Test set accuracy comparison with pre-trained deep CNN models on KDEF and JAFFE datasets

Model	KDEF on Test	KDEF in CV	JAFFE on Test	JAFFE in CV
VGG16	94.56%	93.16±1.42%	100.0%	97.72±2.98%
VGG19	97.23%	95.53±1.34%	100.0%	98.68±3.12%
ResNet18	94.67%	94.24±0.67%	100.0%	98.17 ±2.56%
ResNet34	96.78%	94.99±0.92%	100.0%	98.67 ±2.72%
ResNet50	97.81%	95.61±0.91%	100.0%	99.14 ±3.48%
ResNet152	97.12%	96.59±0.73%	100.0%	99.13 ±2.98%
IncetionV3	97.98%	96.34±1.59%	100.0%	99.14 ±1.43%
MobileNet	98.83%	96.78±1.52%	100.0%	99.62 ±2.83%

Table 3. Comparison of suggested framework results with different ML-DL models

Category	Model	Accuracy (%)	Precision (%)
Machine Learning	Random Forest	88.23	88.59
	Logistic Regression	88.13	88.51
	Support Vector Machine	87.34	87.90
	XGB	86.78	86.23
Deep Learning	MobileNet	96.32	96.46
	InceptionV3	95.32	95.78
	CNN	95.33	95.12
	VGG16	93.22	93.67
	VGG19	94.26	94.55
	ResNet18	94.26	94.25
	ResNet34	94.78	95.23
	ResNet50	97.81	97.18
	Proposed Framework	97.81	97.81

Table 4. Comparison of suggested framework results with different ML-DL models

Category	Model	Recall (%)	F1-Score (%)
Machine Learning	Random Forest (RF)	88.34	88.30
	Logistic Regression (LR)	88.73	88.45
	Support Vector Machine (SVM)	87.26	87.33
	XGB	86.77	86.90
Deep Learning	MobileNet	96.12	96.24
	InceptionV3	95.59	95.11
	CNN	95.34	95.33
	VGG16	93.55	93.65
	VGG19	94.83	94.60
	ResNet18	94.89	94.18
	ResNet34	95.18	95.37
	ResNet50	96.34	96.38
	Proposed Framework	97.25	97.34

The proposed framework classified pupils more accurately than other machine learning and transfer learning models (97.81%), as demonstrated in Table 3 and Table 4. In terms of accuracy, recall, and F1-score, the framework that was proposed outperformed its competitors.

Numerous investigators have invented attendance marking automation systems, but only some have been in the literature survey.

Table 5 compares inference time and computational complexity (measured in GFLOPs) for several machine learning and deep learning models. Traditional machine learning models like Random Forest, SVM, and XGBoost have low FLOPs (<0.02) and fast inference times (<7 ms), making them highly computationally efficient. In comparison, deep learning models have higher learning capabilities despite being much more demanding in terms of FLOPs. Among these, MobileNetV2 strikes a good compromise between speed (12 ms) and complexity (300 GFLOPs), whereas the suggested U-NET + HAAR + ResNet50 framework, while heavier (4800 GFLOPs), retains a practical inference time of 35 ms while improving performance.

Table 5. Comparative performance of traditional machine learning and deep learning models based on inference time (per image), and computational complexity (FLOPs)

Category	Model	Inference Time (ms)	FLOPs (GFLOPs)
Machine Learning	Random Forest	5	~0.01
	Logistic Regression	4	~0.005
	Support Vector Machine (SVM)	7	~0.01
	XGB	6	~0.02
Deep Learning	MobileNetV2	12	300
	InceptionV3	28	5500
	CNN	20	1200
	VGG16	32	15400
	VGG19	37	20000
	ResNet18	15	1800
	ResNet34	24	3600
	ResNet50	27	4100
	Proposed Framework	35	4800

The aggregate detriment of all models in relation to the recommended framework 7(a) is illustrated in Figure 7. This statistic considers two parameters: train loss and values loss. It has been observed that SVM and VGG-19 experienced significant data loss throughout the training process.

Table 6. Suggested framework results compared to recent suggested framework results

Ref.	Algorithm	Parameter	Accuracy
[18]	LDA+ML	Macro accuracy	95%
[19]	MLP	Macro accuracy	88.4%
[20]	Image Tagging	Macro accuracy	70%
[21]	HAAR + Blending approach	Macro accuracy	89.06%
Proposed Framework	U-NET + HAAR + CNN + ResNet-50	Macro accuracy	97.81%

Table 6 compares the proposed framework to other contemporary frameworks based on macro accuracy. A macro accuracy of 97.81 percent was attained by the framework that was recommended, which achieved greater performance in

comparison to the most recent framework that is considered to be state-of-the-art in the field of face detection and recognition.

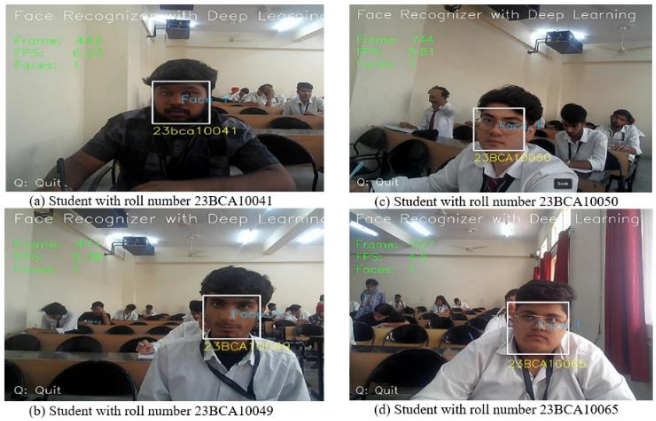
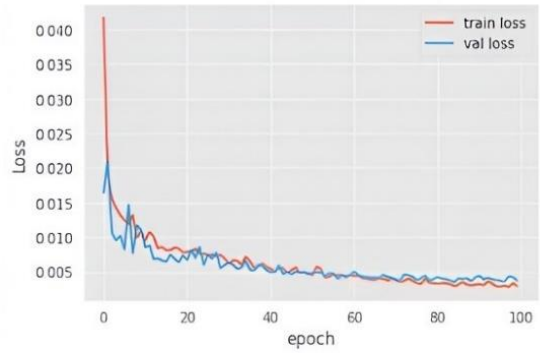
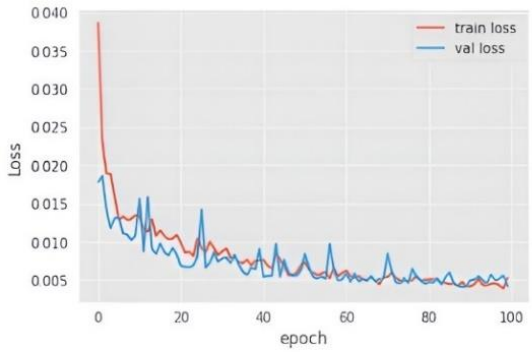


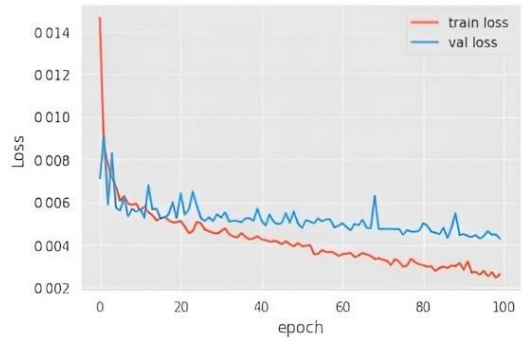
Figure 6. Results of proposed U-NET + HAAR +CNN + ResNet 50 face recognition framework



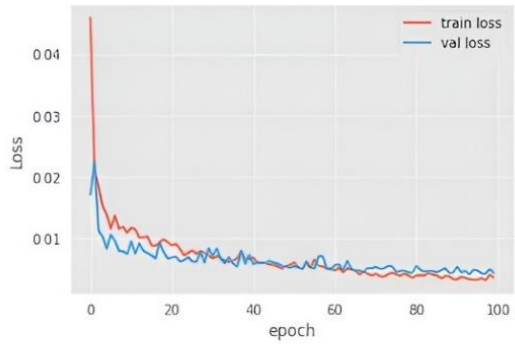
(a) Proposed Framework



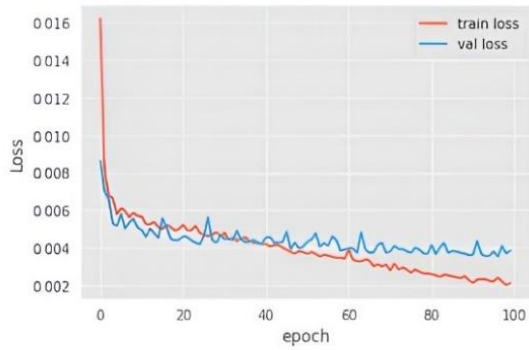
(b) MobileNet



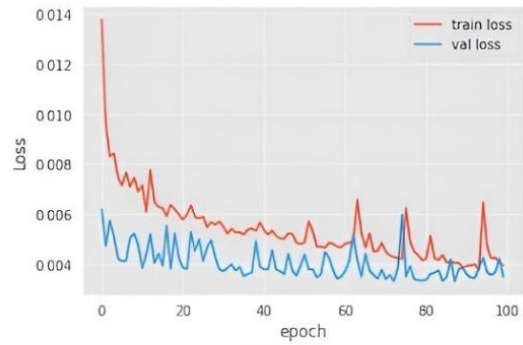
(c) InceptionV3



(d) CNN



(e) VGG19



(f) SVM

Figure 7. Training and test loss of all the models compared with the proposed framework

The following are two essential elements of the suggested system that contribute to an increase in the precision of attendance recordings. An automatic attendance authentication system that is based on face recognition is the best choice because it is straightforward and requires only a small number of acquisition appliances. In the second step, it is necessary to train models with limited datasets and resources and to include advanced mobile devices that have limited RAM and storage space in highly mobile environments. When it comes to machine learning systems, selecting parameters is quite necessary. The hyperparameters were chosen after many pipeline training cycles were completed. This set of hyperparameters includes the number of neurons, dense layers, and optimization learning parameters. By optimizing the parameters of the suggested model for each dataset, it is possible to improve the performance of the approach that has been proposed.

6. ETHICAL CONSIDERATION: PRIVACY AND BIAS

Facial recognition techniques naturally generate ethical questions about privacy, data security, and algorithmic fairness. To ensure compliance with privacy regulations such as the General Data Protection Regulation (GDPR), all facial data acquired during this study was anonymized and saved using safe encryption technologies. Before using the data, participants provided explicit agreement, assuring transparency and legal compliance.

Furthermore, like with many deep learning models, the proposed system may show performance discrepancies between different skin tones and ethnic groups, particularly when trained on datasets with insufficient demographic balance. While the initial examination of the RFW dataset reveals insights into model generalizability, further work will concentrate on bias mitigation, demographic-aware training, and the incorporation of fairness auditing systems to assure equitable performance across all user groups.

7. CONCLUSION

The proposed research aims to develop a method for automated student attendance systems using face recognition. Consequently, the proposed system implemented the technique to automatically identify and document the attendance of all pupils in a specific classroom. The mask intended to conceal most of the mouth, nostrils, and cheeks complicates the process of identifying facial features. However, the U-NET and HAAR cascade enables the model to extract more information from fewer attributes. The ResNet assistance in parameter setting and neighbors selection facilitates the identification of pupils with high degrees of certainty. The proposed method can be deployed to find the footfall in numerous public locations, such as supermarkets, offices, businesses, train stations, malls, colleges, and airports. The authors aspire to eventually be able to autonomously mark attendance in Excel and identify individuals' faces in images containing multiple faces.

REFERENCES

[1] Budiman, A., Fabian, Yaputera, R.A., Achmad, S., Kurniawan, A. (2023). Student attendance with face recognition (LBPH or CNN): Systematic literature review. Procedia Computer Science, 216: 31-38. https://doi.org/10.1016/j.procs.2022.12.108

[2] Kumar, P., Salman Latheef, T.A., Santhosh, R. (2023). Face recognition attendance system using local binary pattern algorithm. In 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), Vellore, India, pp. 1-6. https://doi.org/10.1109/ViTECoN58111.2023.10157843

[3] Rohini, V., Sobhana, M., Chowdary, C.S. (2023). Attendance monitoring system design based on face segmentation and recognition. Recent Patents on Engineering, 17(2): 81-91. http://doi.org/10.2174/1872212116666220401154639

[4] Surantha, N., Sugijakko, B. (2024). Lightweight face recognition-based portable attendance system with liveness detection. Internet of Things, 25: 101089. https://doi.org/10.1016/j.iot.2024.101089

[5] Rathore, K.S., Pandey, A., Gupta, A., Srivastava, D., Agrawal, K., Srivastava, S. (202). Design and implementation of efficient automatic attendance record system based on facial recognition technique. AIP Conference Proceedings, 2978: 060011. https://doi.org/10.1063/5.0182926

[6] Li., J., Liu, K.J. (2023). Design of intelligent teaching system based on face recognition technology. In Proceedings of the 3rd International Conference on New Media Development and Modernized Education, NMDME 2023, Xi'an, China. http://doi.org/10.4108/eai.13-10-2023.2341252

[7] Patil, V., Narayan, A., Ausekar, V., Dinesh, A. (2020, September). Automatic students attendance marking system using image processing and machine learning. In 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 542-546. https://doi.org/10.1109/ICOSEC49089.2020.9215305

[8] Wilson, P.I., Fernandez, J. (2006). Facial feature detection using Haar classifiers. Journal of computing sciences in colleges, 21(4): 127-133.

[9] Halder, R., Chatterjee, R., Sanyal, D.K., Mallick, P.K. (2020). Deep learning-based smart attendance monitoring system. In Proceedings of the Global AI Congress 2019, pp. 101-115. https://doi.org/10.1007/978-981-15-2188-1_9

[10] Reddy, K.N., Alekhya, T., Sushma Manjula, T., Krishnappa, R. (2019). AI-based attendance monitoring system. International Journal of Innovative Technology and Exploring Engineering, 9(2S): 592-597. http://doi.org/10.35940/ijitee.B1057.1292S19

[11] Bah, S.M., Ming, F. (2020). An improved face recognition algorithm and its application in attendance management system. Array, 5: 100014. https://doi.org/10.1016/j.array.2019.100014

[12] Akrout, B. (2025). Deep facial emotion recognition model using optimal feature extraction and dual-attention residual U-Net classifier. Expert Systems, 42(1): e13314. https://doi.org/10.1111/exsy.13314

[13] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 248-255, https://doi.org/10.1109/CVPR.2009.5206848

[14] Marcelino, P. (2018). Transfer learning from pre-trained models. Towards Data Science. https://medium.com/data-science/transfer-learning-from-pre-trained-models-f2393f124751.

[15] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[16] Eng, S.K., Ali, H., Cheah, A.Y., Chong, Y.F. (2019). Facial expression recognition in JAFFE and KDEF Datasets using histogram of oriented gradients and support vector machine. IOP Conference Series: Materials Science and Engineering, 705(1): 012031. https://doi.org/10.1088/1757-899X/705/1/012031

[17] Cheng, F., Yu, J.S., Xiong, H.L. (2010). Facial expression recognition in JAFFE dataset based on

Gaussian process classification. IEEE Transactions on Neural Networks, 21(10): 1685-1690. https://doi.org/10.1109/TNN.2010.2064176
[18] Sahan, J.M., Abbas, E.I., Abood, Z.M. (2021). Implementation of Face Recognition by using of Decision Tree and PART Algorithms Based on LDA. Design Engineering, 8: 8754-8765.
[19] He, G.N., Jiang, Y.Z. (2022). Real-time Face Recognition using SVM, MLP and CNN. In 2022 International Conference on Big Data, Information and Computer Network (BDICN), Sanya, China, pp. 762-767. https://doi.org/10.1109/BDICN55575.2022.00149
[20] Jeya Christy, A., Dhanalakshmi, K. (2022). Content-based image recognition and tagging by deep learning methods. Wireless Personal Communications, 123: 813-838. https://doi.org/10.1007/s11277-021-09159-8
[21] Wu, H., Cao, Y., Wei, H.P., Tian, Z. (2021). Face recognition based on Haar like and Euclidean distance. Journal of Physics: Conference Series, 1813: 012036. https://doi.org/10.1088/1742-6596/1813/1/012036

NOMENCLATURE

F(x)	Feature transformation function
X	Input to layer
GAP	Global average pooling
FeatureMap(i,j)	Map the feature of the value at location(i,j)
Activation	Non-linear function
Weights	Trainable parameters used to scale the features
Biases	Trainable offsets added to weighted feature
H	Height of the feature map
W	Width of the feature map
Conv(1)	1 × 1 convolution layer
Conv(3)	3 × 3 convolution layer
MaxPool(3 × 3)	3 × 3 max pooling layer
Concat	Final output after applying weights and biases
Ii,j	Input at spatial location (i,j)
Ki,k	Depthwise filter
I¯i,j,k	Intermediate feature map
bk	bias
Wc,k	Weight connecting channel c to the output channel k
k	Output channel index
N	Total number of classes

Subscripts

i,j	Define height(H) and Width(W) of feature map
i	Class index
Tpi	True positive for class i
Tni	True negative for class i
FPi	False positive for class i
FNi	False negative for class i