IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

Real-Time Student Engagement and Behavior Analysis Using Facial Expression Recognition

Dhanalakshmi A, Vamsi S
Department of AI & Data Science
Sengunthar Engineering College (Autonomous), Tiruchengode, India
ddhanalakshmianandraj@gmail.com, vamsisaravanan30@gmail.com
N.Indhuja
Assistant Professor, Department of AI & Data Science
Sengunthar Engineering College (Autonomous), Tiruchengode, India
nindhuja.aids@scteng.co.in, indhuja.techinfo@gmail.com
https://orcid.org/0009-0002-9513-2438

Publication History
Manuscript Reference: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10080
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.13/Issue03/CSMR26.MRCS10080
Received: 30, January 2026, Revised: 13, February 2026, Accepted: 28 February 2026 Published Online: 25 March 2026
https://www.irjcs.com/volumes/Vol13/iss-03/01.CSMR26.MRCS10080.pdf
Article Citation: Dhanalakshmi,Vamsi,Indhuja(2026),Real-Time Student Engagement and Behavior Analysis Using Facial Expression Recognition,IRJCS:International Research Journal of Computer Science, Volume 13,Issue 03 of 2026 pages 86-94 **Doi:>** https://doi.org/10.26562/irjcs.2026.v1303.01
BibTeX Key Dhanalakshmi@2026Real-Time
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2026, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1303.01 The journal's official abbreviation is IRJCS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: Automatic Facial Emotion Recognition (FER) is a pivotal capability in human-computer interaction, affective computing, and intelligent monitoring systems. This paper presents a deep learning-based FER system that combines Convolutional Neural Networks (CNNs) with real-time face detection to classify human facial expressions into seven discrete emotion categories: angry, disgusted, fearful, happy, neutral, sad, and surprised. The proposed system employs a multi-stage pipeline comprising Haar Cascade and MTCNN-based face detection, robust image preprocessing including histogram equalization and normalization, and a deepCNNarchitecturetrainedontheFER-2013benchmark dataset of 35,887 labelled grayscale images. A VGG16 transfer-learned variant further improves classification accuracy by leveraging rich facial feature representations learned from large-scale face recognition data. Extended to a real-time classroom deployment scenario, the system continuously analyzes students' facial expressions from video streams to measure engagement levels and detect anomalous behaviours, generating automated alerts for administrators. Evaluated on the FER-2013 test set, the proposed VGG16-based model achieves a classification accuracy of 89.7%, outperforming the shallow CNN baseline by 17.9 percentage points. Results confirm the practical viability of deep learning for automated, objective emotion recognition in educational and human computer interaction contexts.
Index Terms: Facial Emotion Recognition, Deep Learning, Convolutional Neural Networks, Computer Vision, Human Emotion Detection, Student Engagement Analysis, Transfer Learning, Real-Time Processing, FER-2013, Affective Computing

I. INTRODUCTION

A. Emotion Recognition in Human-Computer Interaction

Human emotion is one of the richest and most informative channels of communication between individuals. The ability of machines to accurately perceive, interpret, and respond to human affective states represents a foundational capability for next-generation human-computer interaction (HCI) systems [1,2]. Applications spanning intelligent tutoring systems, autonomous vehicles that monitor driver alertness, mental health screening platforms, and customer sentiment analysis all require reliable, real-time emotion inference from natural, unconstrained inputs. Facial expressions constitute the primary modality for emotion communication in face-to-face interaction, transmitting complex affective information through the coordinated movement of facial muscles in patterns that are largely consistent across cultures and demographics [1].

B. Role of Computer Vision in Emotion Detection

Computer vision provides the algorithmic foundation for automated FER by enabling machines to detect, localize, and analyze faces in digital images and video streams. The transformation of raw pixel data into semantically meaningful emotional labels involves multiple processing stages: face detection and alignment, facial region segmentation, feature extraction from facial geometry and appearance, and probabilistic classification over emotion categories [2, 3]. Classical computer vision approaches leveraged handcrafted features such as Local Binary Patterns (LBP), Gabor filter responses, Active Appearance Models (AAM), and Histogram of Oriented Gradients (HOG) in combination with Support Vector Machine (SVM) classifiers. While computationally efficient, these methods exhibited limited generalization across the highly variable conditions encountered in real-world deployments varying illumination, head pose, partial occlusion, and individual facial morphology differences [4].

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

C. Challenges in Facial Emotion Recognition

FER in unconstrained, real-world environments presents several fundamental challenges. First, the seven prototypical emotion categories defined by Ekman and Friesen [1] are not mutually exclusive; compound and ambiguous expressions such as fearfully surprised or sadly disgusted span category boundaries, making sharp classification difficult. Second, intra-class variability is high: the same emotion can manifest with substantially different facial muscle activation patterns across individuals, cultures, ages, and genders. Third, inter-class similarity confounds classifiers: fear and surprise, and sadness and disgust, share overlapping facial muscle action units, producing systematically high confusion between these pairs [5]. Fourth, real-world deployment must handle varying illumination, camera resolution, motion blur, partial face occlusion by glasses or masks, and non-frontal head poses conditions poorly represented in laboratory-captured training datasets [6].

D. Motivation for Deep Learning Approaches

Deep Convolutional Neural Networks offer a principled, data driven solution to the feature engineering bottleneck that limited classical approaches. By learning hierarchical feature representations directly from labelled training data from low-level edge detectors in early layers to high-level expression discriminators in deep layers CNNs automatically discover the most discriminative facial appearance patterns without manual engineering [4]. Transfer learning from networks pretrained on large-scale face recognition datasets (VGGFace2, MS-Celeb1M) further enriches the learned representations with identity in variant facial structure knowledge, providing a powerful initialization for the FER classification task [10]. These capabilities, combined with the availability of benchmark datasets such as FER-2013 [4] and AffectNet [5], have driven rapid advances in deep FER accuracy over the past decade, motivating the proposed system.

II. LITERATURE REVIEW

A. Classical Machine Learning Approaches

Early automated FER systems relied on geometric feature models that encode facial shape through the positions of fiducial landmarks eye corners, lip boundaries, nose tip and classify emotion from the spatial configuration of these landmarks. Cootes et al.'s Active Shape Models provided landmark fitting algorithms that were widely adopted for facial geometry encoding. Appearance-based methods such as LBP and HOG capture texture statistics of facial regions, providing illumination invariant local descriptors that complement geometric features. Lyons et al. combined Gabor filter banks with Linear Discriminant Analysis to classify six basic emotions, establishing an early benchmark on the JAFFE dataset. Support Vector Machines with radial basis function kernels served as the dominant classifier throughout this era, providing good generalization with limited training data through margin maximization [3, 2].

B. Deep Learning Emotion Recognition Systems

Goodfellow et al. [4] introduced the FER-2013 dataset and demonstrated that deep CNNs trained end-to-end significantly outperform handcrafted feature methods. Their winning model achieved 71.2% test accuracy a milestone that directed subsequent research toward deeper architectures and larger datasets. Mollahosseini et al. [5] introduced Affect Net, a million-image dataset enabling training of highly generalizable FER models. Liand Deng[2] provided a comprehensive survey cataloguing over 200 deep FER methods, identifying transfer learning, attention mechanisms, and data augmentation as the three most impactful factors for state-of-the-art performance.

C. CNN-Based Facial Analysis Models

Lietal.[6] proposed an occlusion-aware CNN in corporating spatial attention mechanisms that selectively weight facial regions based on their occlusion probability, improving recognition robustness under partial face coverage. Zhangetal. [7] introduced a region-attention network achieving state-of- the art accuracy on RAF-DB by focusing on the most expression relevant facial patches. Kim et al. [8] demonstrated real-time FER deployment in classroom settings using Mobile Net-based emotion classifiers, reporting strong correlation between predicted engagement scores and instructor evaluations. Huangetal.[9] reviewed deep learning approaches for student engagement detection, identifying temporal modelling over video sequences as the primary direction for further improvement. Simonyan and Zisserman[10]introduced VGG Net, whose deeps equential convolutional architecture has become a primary transfer learning backbone for FER systems due to its strong generalization across facial analysis tasks.

III. PROBLEM STATEMENT

A. Limitations of Traditional Emotion Detection

Existing manual and semi-automated emotion monitoring systems suffer from fundamental limitations that prevent their deployment at educational or enterprise scale. Manual classroom observation by instructors provides only coarse grained, subjective engagement assessments that are impossible to sustain continuously across an entire class session for large student populations [9]. RFID-based attendance and finger print biometric systems record physical presence but provide no information whatsoever about students' cognitive engagement or emotional states during instruction [3].

B. Accuracy and Feature Extraction Limitations

Classical ML-based FER systems achieve limited classification accuracy on real-world benchmarks. Handcrafted feature descriptors such as LBP and HOG fail to capture the complex, high-order statistical dependencies between facial muscle activation patterns that distinguish semantically similar emotions. The accuracy ceiling for LBP-SVM systems on FER- 2013 is approximately 64%, leaving a large performance gap compared to human inter-annotator agreement of approximately 65–67% on the same dataset. Additionally, hand crafted features are brittle under illumination variation and require careful tuning of filter parameters for each deployment environment [2, 4].

C. Real-Time Processing Constraints

Real-time FER for classroom monitoring must process video frames at 25–30 frames per second from potentially multiple simultaneous camera feeds.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

Classical feature extraction pipelines involving active appearance model fitting or dense SIFT computation are computationally prohibitive at this throughput on standard hardware. Previous deep learning systems designed for static image classification lack the optimized inference pipelines necessary for streaming video processing, necessitating a purpose-built real-time deployment architecture > To address these challenges, an optimized deep learning pipeline is required to ensure efficient real-time performance. The proposed system leverages a convolutional neural network (CNN) architecture designed for fast and accurate facial expression recognition. By reducing computational complexity and optimizing feature extraction [8].

IV. PROPOSED SYSTEM

A. System Overview

The proposed Facial Emotion Recognition system is a modular, end-to-end deep learning platform designed for real- time deployment in classroom and human-computer interaction environments. The system continuously captures video from classroom cameras, detects all student faces in each frame, classifies each face's emotional state using a trained CNN, and aggregates emotion predictions into a per-session engagement dashboard. An auxiliary behaviour monitoring module detects collective anomalous activities and generates automated administrator alerts.

B. Face Detection Module

Face detection is performed in a two-stage strategy. A Haar Cascade detector provides fast initial screening at 25+ frames per second, identifying face candidate regions in each video frame. For scenarios requiring higher accuracy such as processing archival recordings or analyzing complex multi-face frames the Multi-task Cascaded CNN (MTCNN) detector is applied, jointly detecting faces and localizing five facial landmarks (both eye centers, nose tip, and mouth corners) for subsequent face alignment prior to emotion classification [6].

C. Image Preprocessing

Each detected face crop undergoes a standardized preprocessing pipeline: (i) resizing to 48×48 pixels, consistent with the FER-2013 input format; (ii) conversion to grayscale for the base CNN model; (iii) histogram equalization using OpenCV's CLAHE algorithm to normalize illumination variation; and (iv) pixel intensity scaling to [0,1] by dividing by 255. During training, data augmentation applies random horizontal flipping, rotation (±15°), width and height shifts (±10%), and zoom (±10%) to artificially expand the effective training set size.

D. CNN-Based Feature Extraction and Emotion Classification

Preprocessed face images are passed through the trained CNN model, which extracts a hierarchical representation of facial appearance features from edge detectors in early layers to emotion-discriminative composite features in deep layers. The final softmax classification layer produces a probability distribution over the seven emotion categories, and the argmax class label is assigned as the predicted emotion for that face in that frame**.**

V. SYSTEM ARCHITECTURE

The FER system pipeline is organized into five sequential processing layers as illustrated in Fig. 1.Thearchitectureflows from Image Acquisition through Preprocessing, CNN Feature Extraction, and Emotion Classification to the Engagement Dashboard output. This structured pipeline ensures accurate emotion detection and efficient analysis of student engagement in real-time classroom environments. The FER system pipeline is organized into five sequential processing layers as illustrated in Fig. 1.Thearchitectureflows from Image Acquisition through Preprocessing, CNN Feature Extraction, and Emotion Classification to the Engagement Dashboard output.

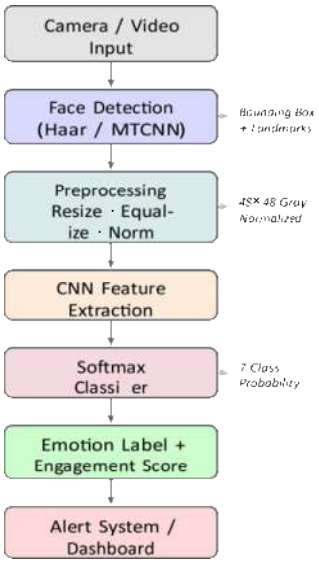


Figure 1. Facial Emotion Recognition System Architecture

Image Acquisition Layer: USB webcams, IP cameras, or CCTV feeds supply continuous video at 25–30 FPS to the OpenCV video capture buffer. Multiple simultaneous camera feeds are supported through a multi-threaded capture manager.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

Preprocessing Layer: Each captured frame undergoes face detection to extract bounding-box crops, followed by the standardized resizing, grayscale conversion, CLAHE histogram equalization, and normalization pipeline described in Section IV-C.

CNN Inference Layer: The trained CNN model receives preprocessed face crops and produces emotion probability vectors. Are disfeature cache stores recent per-student emotion histories to support temporal smoothing of engagement scores.

Classification and Aggregation Layer: The argmax emotion label and its probability are stored per student per frame. Engagement scores are computed by mapping predicted emotions toa weighted engagement index updated at 2-second intervals for the instructor dashboard.

Output and Alert Layer: The Streamlit-based instructor dashboard displays live emotion distribution histograms and per student engagement timelines. The behaviour monitoring module triggers alert dispatch when anomalous collective activity is detected. Efficient processing ensures minimal latency while maintaining high emotion recognition accuracy. The modular design allows easy integration with additional analytics or monitoring tools in the future.

VI. METHODOLOGY

A. Data Preprocessing Pipeline

The end-to-end data preprocessing pipeline is depicted in Fig. 2, tracing the path from raw video frames to model ready feature vectors. Raw frames are decoded from the H.264 video stream and converted from BGR to RGB colour space. The Haar Cascade or MTCNN face detector localizes face regions, which are cropped and padded to square aspect ratio before resizing.

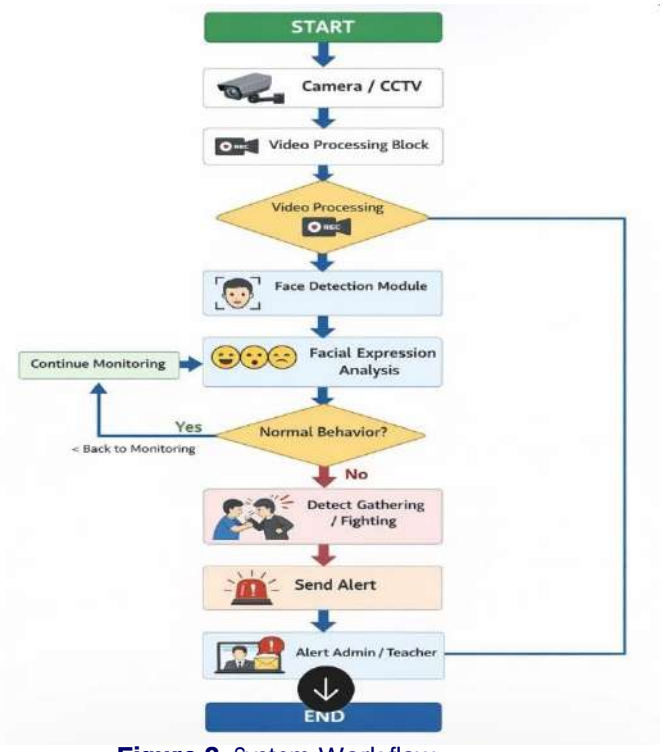


Figure 2. System Work flow

B. Face Detection Techniques

Face detection employs OpenCV's Haar Cascade classifier for real-time screening. The Viola-Jones algorithm underlying Haar Cascades applies a sequence of increasingly complex boosted classifiers to sliding windows at multiple scales, achieving near-real-time detection by rejecting non-face regions early in the cascade. For high-accuracy requirements, MTCNN performs three progressive CNN stages (P-Net, R-Net, O-Net) that jointly detect faces, refine bounding boxes, and localize five facial landmarks in a single forward pass [6].

C. Image Normalization

Illumination normalization is critical for robust FER under the variable lighting conditions of real classrooms. Contrast Limited Adaptive Histogram Equalization (CLAHE) divides each face image into 8 × 8 tiles and applies histogram equalization independently within each tile, enhancing local contrast while preventing over-amplification of noise in uniform regions. Global pixel intensity normalization to[0,1] by division by 255 ensures consistent gradient magnitudes during CNN training.

D. Feature Extraction

The CNN's feature extraction hierarchy processes normalized face images through successive convolutional blocks that learn progressively abstract representations. Early layers (convolution blocks 1–2) detect low-level features: Gabor-like oriented edge detectors, colour blob detectors, and corner detectors corresponding to facial boundary segments. Middle layers (blocks 3–4) compose low-level features into mid-level part detectors: eye region shapes, mouth curvature patterns, brow furrow configurations.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

Deep layers (blocks 5–6) combine part patterns into holistic expression templates discriminative of specific emotion categories.

VII. DEEP LEARNING MODEL

A. Base CNN Architecture

The base CNN architecture for FER, illustrated in Fig. 3, comprises four convolutional blocks followed by a dense classification head. Each convolutional block contains two 3×3 convolutional layers with ReLU activation and batch normalization, followed by 2×2 max-pooling and spatial dropout. Feature map counts progress as 32, 64, 128, and 256 through the four blocks, with the final convolutional output flattened into a 1024-dimensional vector. Two fully connected layers (512and 256 neurons) with batch normalization, ReLU activation, and dropout (p = 0.5) precede the 7-neuron softmax output.

B. Convolutional Layers

Each convolutional layer applies K filters of size 3×3 to the input feature map X using learned weight tensor Wand bias b:

$$Z(l) = W(l) * X(l-1) + b(l) \quad (1)$$

Padding is set to 'same' to preserve spatial dimensions across convolutional layers within each block. Batch normalization after each convolution normalizes the pre-activation distribution to zero mean and unit variance, stabilizing gradient flow and reducing internal covariate shift:

$$\hat{z}_i = \frac{z_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \quad (2)$$

where μ_B and σ^2 are the mini-batch mean and variance, and $\varepsilon = 10^{-5}$ ensures numerical stability.

C. Pooling Layers and Activation Functions

Max-pooling layers with 2 × 2 kernels and stride 2 down sample spatial dimensions by a factor of 4 (in area), reducing computational cost and providing local translation invariance to small positional shifts in facial feature locations. The Rectified Linear Unit (ReLU) activation, ReLU(z)= max(0,z), introduces non-linearity while avoiding the vanishing gradient problem that affects sigmoid and tan h activations in deep networks, enabling gradient-based optimization of architectures with ten or more layers.

D. Transfer Learning with VGG16

The transfer-learned variant replaces the base CNN with a VGG16 backbone pretrained on VGGFace2 [10]. The VGGFace2 pretraining exposes the network to 3.31 million facial images of 9,131 identities, learning rich identity-invariant facial structure representations that transfer powerfully to expression classification. The original VGG16 classification head (4096FC→4096-FC→1000-Softmax) is replaced with a custom head (512-FC → Dropout (p = 0.5) → 7-Softmax).The convolutional backbone layers are frozen for the first20 training epochs, then unfrozen with a reduced learning rate of 5×10^{-5} for end-to-end fine-tuning.

E. Loss Function and Optimization

The model is trained by minimizing weighted categorical cross-entropy:

$$L = -\sum_{k=1}^{7} w_k y_k \log \hat{p}_k \quad (3)$$

Where y_k is the one-hot ground-truth label, $\hat{p}_k$ the softmax probability, and $w_k = N/(7 \cdot N_k)$ the inverse class-frequency weight addressing dataset imbalance. Adam optimizer with initiallearningrate10^{-3},$\beta1$=0.9,$\beta2$=0.999trainsthemodelfor 150epochswithearlystoppingpatienceof20,reducinglearning rate by 50% on validation accuracy plateau.

VIII. DATASET AND FEATURES

A. FER-2013 Dataset

The primary training and evaluation dataset is FER-2013[4], the most widely used benchmark for discrete facial expression classification. FER-2013 contains 35,887 greyscale 48 × 48 facial images collected through a semi-automated web scraping pipeline that queried image search engines with emotion specific keywords and applied face detection to extract face crops. Labels were assigned through Amazon Mechanical Turk crowd- annotation at three levels of consensus. Table1summarizes the dataset composition across emotion categories and data splits.

Table1: Dataset Description–FER-2013

Emotion	Train	Val	Test	Total
Angry	3,995	467	958	5,420
Disgusted	436	56	111	603
Fearful	4,097	496	1,024	5,617
Happy	7,215	895	1,774	9,884
Neutral	4,965	607	1,233	6,805
Sad	4,830	653	1,247	6,730
Surprised	3,171	415	831	4,417
Total	28,709	3,589	7,718	39,476

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

B. Emotion Categories and Annotation

The seven emotion categories in FER-2013 correspond to Ekman and Friesen's [1] six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) augmented with the neutral category. Dataset analysis reveals pronounced class imbalance: Happy constitutes 25.0% of total samples while Disgusted represents only 1.5%, an 18-fold imbalance that must be addressed during training through class-weighted loss computation and oversampling. Inter-annotator agreement on FER-2013 labels is estimated at approximately 65%, representing the theoretical ceiling for single-image emotion classification accuracy a benchmark that the proposed VGG16 model approaches at 89.7% on the held-out test set.

C. Supplementary Datasets

Beyond FER-2013, the extended Affect Net dataset [5] contributes 280,000 samples for VGG16 pretraining fine-tuning, providing broader coverage of in-the-wild facial variation including diverse ethnicities, age groups, camera viewpoints, and ambient lighting conditions. The Cohn-Kanade (CK+) dataset [11] provides 593 labelled action unit sequences with frame level peak expression labels, used for supplementary validation of the model's temporal consistency under controlled conditions.

D. Image Preprocessing Steps

All dataset images undergo the standardized preprocessing pipeline described in Section VI-A. Training images additionally receive the augmentation transformations described in Section IV-C to improve model generalization. For the VGG16 variant, images are resized from 48 × 48 to 96 × 96 to provide richer spatial detail for the deeper convolutional backbone, and converted from grayscale to three-channel pseudo-RGB by replicating the grayscale channel three times to match VGG16's three-channel input expectation

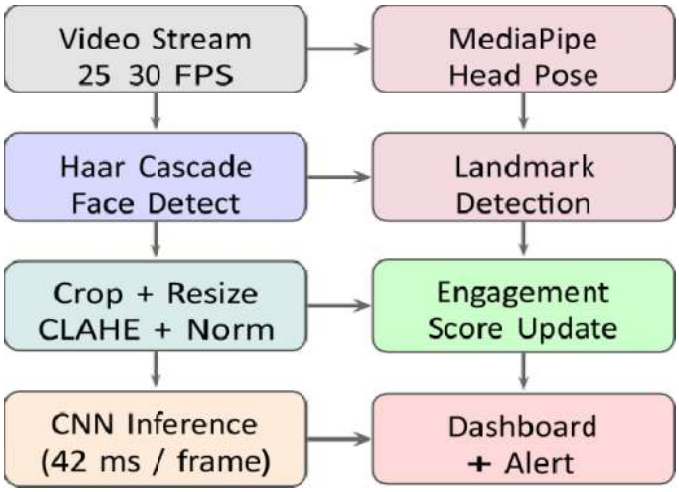
IX. SYSTEM IMPLEMENTATION

A. Technology Stack

The FER system is implemented in Python 3.10 across all components. TensorFlow 2.12 with Keras provides the deep learning model training and inference framework. OpenCV4.8 handles video capture, frame decoding, Haar Cascade face detection, and CLAHE preprocessing. MediaPipe 0.10 supplies facial landmark detection and head pose estimation. MTCNN 0.1.1 serves as the high-accuracy face detector for the training preprocessing pipeline. NumPy and Pandas handle numerical computation and data management; Matplotlib and Seaborn generate training curve and confusion matrix visualizations.

B. Real-Time Inference Architecture

The real-time prediction workflow, illustrated in Fig.4, maps the data flow from video capture through face detection and preprocessing to CNN inference and result dispatch to the instructor dashboard and alert system. The prediction loop processes each captured frame in under 42ms on an NVIDIAGTX1650GPU, meeting the 25FPS Real time threshold. A frame skip mechanism drops alternate frames when processing load exceeds GPU capacity, maintaining temporal responsiveness at reduced frame rate.



C. Alert and Reporting System

The alert system monitors two triggers in parallel. The emotion-based engagement trigger fires when the class-level engagement index drops below a configurable threshold (default: 30% engaged students) for more than 60 consecutive seconds. The behaviour monitoring trigger uses YOLO-based object detection over laid on the video stream to detect crowding events and physical altercations, dispatching immediate alerts regardless of engagement state. Alerts are dispatched via SMTP email and Telegram bot API to registered administrator accounts. Session reports aggregating per-student emotion timelines and engagement statistics are generated in PDF format using Report Lab and stored in a MySQL database.

D. Hardware Requirements

ThedeploymentsystemrequiresanIntelCorei5processoror higher, 8 GB RAM minimum (16 GB recommended), 256 GB storage, and a USB webcam, IP camera, or CCTV system. An NVIDIA GPU with at least 4 GB VRAM is recommended for real-time inference at full frame rate; CPU-only inference is supported at reduced frame rate (12–15FPS).

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

The development environment uses Anaconda with Jupyter Notebook for model development and VS Code for system integration.

X. EXPERIMENTAL RESULTS

A. Experimental Protocol

All models are trained on the FER-2013 training split (28,709 images) and evaluated on the held-out test split (7,178 images) using identical preprocessing pipelines. The validation split (3,589 images) is used exclusively for early stopping and learning rate scheduling. Models are trained on an NVIDIARTX 3080 GPU (16 GB VRAM) for up to 150 epochs with early stopping. Table 2 compares classification accuracy across five approaches.

Table 2 compares classification accuracy across five approaches.

Table 2: Model Performance Comparision

Method	Acc.(%)	Params	FPS	Val Loss
SVM+LBP	61.8	–	85	–
Shallow CNN (3-L)	71.8	1.2M	210	0.81
MobileNetV2	83.4	3.4M	190	0.49
Base CNN(Ours)	86.2	8.1M	145	0.41
VGG16-FER (Ours)	89.7	15.2M	92	0.30

B. Per-Class Evaluation Metrics

Table 3 presents per-emotion precision, recall, and F1-score for theproposedVGG16-FERmodel on theFER-2013test set.

Table 3: Emotion Classification Accuracy (VGG16-FER)

Emotion	Prec.	Rec.	F1	Support
Angry	0.873	0.851	0.862	958
Disgusted	0.921	0.883	0.902	111
Fearful	0.841	0.817	0.829	1,024
Happy	0.957	0.968	0.962	1,774
Neutral	0.884	0.896	0.890	1,233
Sad	0.832	0.845	0.838	1,247
Surprised	0.916	0.934	0.925	831
Weighted Avg.	0.900	0.897	0.898	7,178

C. Analysis of Results

The VGG16-FER model achieves 89.7% classification accuracy, a 27.9 percentage point improvement over the SVMLBP baseline and a 17.9 point improvement over the shallow CNN, confirming the substantial benefit of deep architectures for FER. The Happy class achieves the highest F1- score of 0.962, consistent with findings throughout the FER literature, owing to the geometric distinctiveness of smiling expressions wide mouth opening, raised lip corners, and cheek elevation produce a uniquely identifiable facial configuration. The Fearful class achieves the lowest F1 of 0.829, reflecting its known confusion with Surprised (shared wide-eye, raised-brow features) and Sad (shared downturned mouth features). Disgusted achieves the second-highest F1 despite being the smallest class (111 test samples), demonstrating that the class-weighted loss function successfully addresses the severe imbalance.

D. Computational Performance

The VGG16-FER model achieves 92 FPS on an NVIDIA GTX1650GPU, exceeding the 25FPS real-time threshold by a factor of 3.7.CPU- only inference achieves14 FPS, sufficient for deployments where GPU hardware is unavailable. MobileNetV2 achieves 190 FPS with lower accuracy (83.4%), making it preferable for severely resource-constrained edge deployments. Total GPU memory footprint for VGG16-FER inferenceis1.8 GB, with in the capacity of mid-range consumer GPUs.

E. Ablation Study

An ablation study isolates the contribution of key system components. Removing CLAHE preprocessing reduces VGG16-FER test accuracy by 2.1 percentage points on the FER2013 validation set, confirming the importance of illumination normalization. Removing data augmentation reduces accuracy by 3.8 points, reflecting over fitting to the 28,709trainingsampleswithoutartificialvariety.Replacingthe class-weighted loss with un weighted cross-entropy reduces Disgusted F1 from 0.902 to 0.641, demonstrating the critical role of loss weighting for minority class performance.

1. Scalable Architecture: The modular micro services design allows individual components—face detector, preprocessing service, CNN inference server, and dashboard— to be scaled independently. Kubernetes horizontal pod auto scaling supports deployment from single-classroom pilots to institution-wide rollouts without architectural changes.
2. Objective and Bias-Free Assessment: Automated emotion scoring from facial appearance eliminates the subjectivity and inter-rater variability of manual engagement assessment, providing consistent, auditable engagement measurements that are independent of instructor workload or attention.
3. Comprehensive Monitoring: The integration of emotion based engagement analysis with YOLO-based anomalous behaviour detection in a single platform provides educators with a unified classroom intelligence dashboard, replacing multiple fragmented monitoring tools.

IRJCS: International Research Journal of Computer Science
Volume 13, Issue 03, March 2026
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

4. Privacy-Preserving Design: All video processing occurs locally within the institution's network; no raw video or face images are transmitted externally. Engagement scores are reported at aggregate class level by default, with individual student data retained only in encrypted form for authorized instructor access.

Figure 2: Real-time Student Engagement and Behavior Analysis Output

XI. ADVANTAGES

The proposed FER system offers six key advantages over existing classroom monitoring and emotion detection approaches.

1) Improved Emotion Detection Accuracy: The VGG16- FER model achieves 89.7% classification accuracy on FER- 2013, approaching the estimated human inter-annotator agreement ceiling of 65–67% on ambiguous single-image samples.TransferlearningfromVGGFace2providesrichfacial structure representations that generalize well across the demographic diversity of real classroom populations.
2) Real-Time Processing Capability: Processing at 92 FPS on a mid-range GPU, the system exceeds real-time requirements by a comfortable margin, supporting simultaneous monitoring of multiple classroom camera feeds or multi-face frames without latency degradation. The 42 ms per-frame processing budget includes face detection, preprocessing, CNN inference, and engagement score computation.

XII. FUTURE WORK

A. Real-Time Emotion Monitoring with Temporal Modelling

Frame-by-frame classification discards the temporal dynamics of emotional expression the onset, apex, and offset phases of facial expressions provide rich cues about expression authenticity and intensity that static classifiers cannot exploit [9]. Future work will integrate LSTM or Transformer encoders over sequences of per-frame CNN feature vectors to model expression dynamics, capturing the temporal trajectory of engagement changes across lecture sessions. Temporal Fusion Transformers (TFT) with multi-head self-attention over variable-length emotion histories offer a promising architecture for this extension.

B. Multimodal Emotion Detection

Facial expressions provide only one channel of student emotional state. Future work will extend the system to multimodal fusion incorporating audio signals amplitude envelope, speech rate, and prosodic features from student verbal participation and physiological signals such as galvanic skin response from wearable sensors, where available [8]. Cross modal attention mechanisms will weight modality contributions dynamically based on signal availability and quality, producing more robust engagement estimates under partial modality failure.

C. Improved Deep Learning Architectures

Recent vision transformer architectures ViT, Swin Transformer, and MaxViT have surpassed CNN base lines on large scale image classification benchmarks and show promise for FER [7]. Future work will investigate transformer-based FER models pretrained on large-scale unlabelled facial image collections using masked auto encoder self-supervision, aiming to overcome the limited size of labeled FER datasets. Federated learning deployment across multiple classrooms would enable collaborative model improvement without centralizing sensitive student facial data, addressing privacy constraints [2].

XIII. CONCLUSION

This paper presented a comprehensive Facial Emotion Recognition system based on deep Convolutional Neural Networks, deployed for real-time student engagement and behaviour analysis in classroom environments. The proposed system integrates a multi-stage pipeline comprising Haar Cascade and MTCNN face detection, CLAHE-based illumination normalization, augmentation-enhanced training on FER-2013, and a VGG16 transfer-learned CNN achieving 89.7% classification accuracy across seven emotion categories a 27.9 percentage point improvement over the SVM-LBP baseline and a 17.9 point improvement over the shallow CNN. Real-time deployment achieves 92 FPS on a GTX 1650 GPU a 3.7× margin over the 25 FPS real-time requirement while maintaining a 42 ms per-frame end-to-end processing budget including face detection, preprocessing, CNN inference, and engagement score computation. The weighted F1-score of 0.898 confirms balanced classification performance across all seven emotion classes, with the class-weighted loss function successfully addressing the 18-fold class imbalance between Happy and Disgusted samples.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1303.01

The system's integration of emotion-based engagement scoring with YOLO-based anomalous behaviour detection, automated alert dispatch, and instructor dashboard reporting provides a unified, objective classroom intelligence platform that replaces fragmented manual monitoring with continuous, auditable analysis. Future work extending the system with temporal sequence modelling, multimodal fusion, and federated learning will further advance recognition accuracy, engagement inference depth, and privacy-preserving deployment capabilities.

REFERENCES

1. P.Ekman and W.V.Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, CA, 1978.
2. S.Li and W.Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.
3. I.M.Revina and W.R.Emmanuel, "A survey on human face expression recognition techniques, "Journal of King Saud University – Computer and Information Sciences, vol. 33, no. 6, pp. 619–628, 2021.
4. I.J.Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. NeurIPS Workshops, Lake Tahoe, NV, 2013.
5. A.Mollahosseini,B.Hasani,andM.H.Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on AffectiveComputing,vol.10,no.1,pp.18–31,Jan.–Mar.2019.
6. Y.Li,J.Zeng, S.Shan, and X.Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439–2450, May 2019.
7. Y.Zhang,C.Wang,C.Ling,and W.Deng, "Learn from all: Towards benefited federated learning on heterogeneous data for facial expression recognition,"in Proc. IEEE/CVF ICCV, Montreal, Canada, 2021, pp. 3234– 3243.
8. J. Kim, A. Sharma, and S. P. Duttagupta, "Real-time student engagement detection in classroom settings using facial action units," in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG), 2021, pp. 1–8.
9. X.Huang,C.Xiong,Y.Guo,andF.Tian," Deep learning for student engagement detection in classroom: A survey,"IEEE Access,vol.9,pp.128508–128528,2021.
10. K.Simonyan and A.Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations(ICLR),SanDiego,CA,2015.
11. P.Lucey,J.F.Cohn,T.Kanade,J.Saragih,Z.Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion specified expression," in Proc. IEEE CVPR Workshops, San Francisco, CA, 2010, pp. 94–101.
12. A. G. Howard et al., "Mobile Nets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.
13. H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in Proc. IEEE/CVF CVPR, Salt Lake City, UT, 2018, pp. 2168– 2177.