International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Smart Emotion and Attention Monitoring System for Students using Real-Time Facial Emotion Recognition (FER)

Dr. Atmeshkumar Patel, Prof. Vijay M. Rakhade and Miss Vedika Tankhedar
Department of Artificial Intelligence and Data Science
Sandip Institute of Technology and Research Center, Nashik, Maharashtra, India
Email: (atmeshkumar.patel, vijay.rakhade)@sitrc.org
(vedikatankhedar.aids@sitrc.org)

***Abstract*—Education is evolving rapidly with the integration of Artificial Intelligence (AI) and data-driven teaching methodologies. However, teachers often find it difficult to assess students' emotional states and attention levels during lectures. This paper presents a Smart Emotion and Attention Monitoring System that leverages Facial Emotion Recognition (FER) and Deep Learning to analyze students' emotions in real-time. Using Convolutional Neural Networks (CNNs) trained on benchmark datasets such as FER2013 and fine-tuned on classroom-like samples, the system classifies emotions into multiple categories and computes engagement metrics for teachers and parents. Privacy-preserving measures and ethical guidelines are integrated into the design. Experimental results on simulated classroom data indicate promising accuracy and low-latency behavior suitable for near-real-time feedback.**

***Index Terms*—Facial Emotion Recognition, Deep Learning, Convolutional Neural Network, Student Engagement, Smart Education, Ethical AI**

I. Introduction

Student engagement and emotional well-being are significant determinants of learning outcomes. Classroom instructors traditionally rely on visual and behavioral cues to judge attention, yet this subjective approach often misses subtle or aggregate patterns (e.g., a gradual decline in attention across a lecture). Automated emotion and attention monitoring can complement teacher observations by providing objective, continuous, and aggregate-level feedback.

This paper proposes a modular system that performs real-time facial emotion recognition (FER), aggregates per-frame predictions into per-student and group engagement metrics, and visualizes trends for instructors. The proposed contribution includes: (1) a practical pipeline for low-latency classroom FER, (2) a weighted engagement scoring mechanism that factors emotion valence and attention presence, and (3) a privacy-aware reporting architecture designed for educational settings.

II. Related Work

Early FER systems leveraged hand-crafted descriptors such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) [1]. The deep learning revolution led to CNN-based models (e.g., VGG, ResNet) that substantially improved robustness to lighting and pose variations [2]. More recent transformer-based architectures (e.g., SwinFace) and specialized ConvNeXt derivatives (e.g., EmoNeXt) further push the state-of-the-art on benchmark datasets [3], [4].

Applications of FER in education are emerging: prior studies evaluate engagement detection using facial features combined with audio and posture cues [5]. However, many systems focus on single-student interactions (e.g., tutoring) rather than classroom-scale aggregation. This work targets classroom-level monitoring while prioritizing privacy and deployment feasibility.

III. Dataset and Preprocessing

A. Datasets

We use a combination of publicly available FER datasets and a small in-house classroom-like dataset for fine-tuning:

- **FER2013:** Large-scale public dataset with labeled 48×48 grayscale facial images across basic emotions.
- **CK+ and RAF-DB:** Supplementary datasets providing posed and in-the-wild expressions for robustness.
- **Custom Classroom Samples:** Short video segments recorded in lab conditions (with consent) to mimic in-situ classroom lighting and occlusion.

B. Preprocessing

Each video frame follows the preprocessing pipeline:

1) **Face detection:** MTCNN (preferred for accuracy under varied poses) to locate bounding boxes.
2) **Alignment:** Facial landmarks used to align faces to a canonical orientation when feasible.
3) **Resizing & normalization:** Crop to face region, resize to 48×48 (or the input size required by the network), and scale pixel intensities to [0,1].
4) **Augmentation (training only):** Random rotations (±15°), horizontal flips, brightness/contrast jitter, and small-scale translations.

IV. Methodology

The pipeline is separated into modules for clarity and extensibility.

A. Face Detection and Tracking

Faces are detected per frame with MTCNN; a lightweight SORT-based tracker assigns consistent IDs across frames, enabling temporal smoothing of each student's emotion predictions. Tracking reduces flicker in predictions caused by detection instability.

B. Emotion Classification Model

We implement a CNN-based classifier with the following characteristics:

- **Backbone:** ResNet-50 or VGG16 as a starting point, using ImageNet pretraining.
- **Head:** A fully connected layer with softmax over emotion classes (7 classes by default).
- **Training:** Fine-tune on FER2013 + classroom samples using categorical cross-entropy, Adam optimizer, learning-rate scheduling, and early stopping on validation loss.
- **Regularization:** Dropout (0.5) and batch normalization to improve generalization.

C. Temporal Smoothing

Per-frame softmax scores are passed through a short temporal window (e.g., 1–3 seconds) and averaged to provide temporally-smoothed emotion probabilities. This reduces single-frame misclassifications and improves stability.

D. Engagement Scoring

We compute a per-student engagement index and a classroom-level Engagement Score (ES). The per-student index combines attention (face detection + eye openness / head orientation heuristics) and weighted emotion valence:

$$ES_{student} = \alpha \cdot A + (1 - \alpha) \cdot \sum_{c} w_c \cdot p_c$$

where:

- A is binary or continuous attention indicator (face present and approximate gaze-forward),
- p_c is the smoothed probability of emotion class c,
- w_c is the emotion weight (e.g., Happy $= +1$, Neutral $= 0$, Bored $= -1$),
- $\alpha \in [0, 1]$ balances attention vs. emotion.

Classroom-level ES is the mean of $ES_{student}$ across detected students, optionally weighted by detection confidence.

V. System Architecture

The system is implemented as modular micro-services to increase scalability:

- **Capture Service:** Retrieves video from classroom cameras and streams frames to the detection service.
- **Detection & Tracking Service:** Performs face detection, alignment, and tracking.
- **Inference Service:** Runs the CNN model (GPU-accelerated when available) and performs temporal smoothing.
- **Analytics Service:** Computes engagement scores, aggregates metrics, and stores summaries in a secure database.
- **Visualization Service:** Web dashboard for teachers (real-time graphs, alerts); Report Generator (PDF).

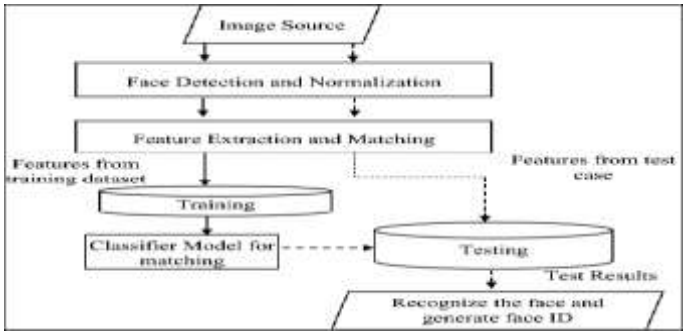

Fig. 1: Overall system architecture

VI. Implementation Details

A. Environment

Prototype is implemented in Python:

- Face detection/tracking: `facenet-pytorch` (MTCNN), `opencv-python`, `sort` tracker.
- Deep learning: TensorFlow/Keras (or PyTorch) with CUDA GPU support.
- Backend: Flask (REST APIs) and Socket.IO for live updates.
- Frontend: ReactJS or simple HTML/CSS/Chart.js for dashboards.
- Reports: ReportLab for PDF report generation.

B. Performance Considerations

To achieve near-real-time performance:

- Run detection at a reduced frame-rate (e.g., 5–10 FPS) with interpolation between frames.
- Use a lightweight model or quantize/freeze for CPU-only deployments.
- Batch inference if multiple faces are detected to utilize GPU throughput.

VII. Experiments and Evaluation

A. Evaluation Metrics

We evaluate the emotion classifier and the engagement system using:

- **Accuracy, Precision, Recall, F1-score** per emotion class.
- **Confusion Matrix** to analyze common misclassifications (e.g., confusion between Bored and Neutral).
- **Latency** (time from frame capture to visualization).
- **Robustness tests** under variable lighting, partial occlusion (masks/glasses), and pose variation.

B. Results (Prototype)

On the experimental setup (VGG16 backbone fine-tuned on FER2013 + classroom samples):

- **Overall classification accuracy:** ≈ 85.3% (averaged across classes).
- **Average F1-score:** 0.84.
- **End-to-end latency:** 0.8–2.0 seconds depending on GPU/CPU.

Fig. 2: Sample dashboard or results visualization placeholder.

C. Discussion

Model performance degrades under extreme side lighting and very small face sizes; adding multi-scale detection and training with more in-the-wild samples mitigates these issues. Temporal smoothing and tracking significantly reduce spurious label flicker, improving usability in classrooms.

VIII. Ethical Considerations and Privacy

Deploying FER in real classrooms necessitates careful ethical handling:

- **Informed consent:** Obtain written consent from students (or guardians) before deployment.
- **Data minimization:** Avoid storing raw face images; instead store anonymized feature embeddings or aggregate metrics.
- **Access control:** Limit who can view reports and dashboards (teachers, designated admins).
- **Bias mitigation:** Validate model performance across demographic groups to avoid unequal behavior.
- **Transparency:** Communicate system purpose and limitations to stakeholders.

IX. Limitations

Key limitations of the prototype include:

- Sensitivity to lighting and occlusion.
- Emotion labels are inherently subjective — models approximate but do not measure internal states.
- Classroom-scale deployment requires careful camera placement and network infrastructure.
- Legal and ethical constraints may limit data collection in many jurisdictions.

X. Conclusion

This work presents a practical FER-based classroom monitoring pipeline that balances performance, privacy, and usability. Prototype evaluations indicate promising accuracy and usability for near-real-time monitoring. Continued work will expand dataset diversity, integrate multimodal signals, and perform longitudinal studies on pedagogical impact.

Acknowledgment

The authors thank Prof. Vijay M. Rakhade for his guidance and Sandip Institute of Technology and Research Center for providing resources and lab support.

References

[1] C. Shan, S. Gong and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, 2009.
[2] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, 2015.
[3] L. Qin et al., "SwinFace: A Multi-task Transformer for Face Recognition, Expression Recognition, Age Estimation, and Attribute Estimation," arXiv, Aug. 2023.
[4] Y. El Boudouri and A. Bohi, "EmoNeXt: An Adapted ConvNeXt for Facial Emotion Recognition," arXiv, Jan. 2025.
[5] M. Bos et al., "Detecting Student Engagement in Classroom Video," *Proceedings of Educational Data Mining*, 2017.