International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

A Temporal Facial Engagement Analysis Framework for Pedagogical Feedback in Smart Classroom

Manisha B. More[1], Bali Thorat[2]

[1,2] Dr.G.Y Pathrikar College of CS and IT
Chh.Sambhaji Nagar, Maharashtra,India

Abstract

Traditional classrooms lack real-time mechanisms to quantify student engagement, limiting instructors' ability to adapt teaching dynamically. This paper proposes a temporal facial engagement analysis framework that processes classroom videos from 20 students using MTCNN for face detection, InceptionResnetV1 for recognition, and LSTM networks for temporal emotion modelling. The system computes per-student attendance, engagement trajectories, and class-wide averages to generate actionable pedagogical feedback. This real-time insight enhances the overall effectiveness of the teaching–learning process by making it more responsive and student-centred. Furthermore, the system contributes to improved learning outcomes by facilitating timely pedagogical interventions whenever signs of disengagement are detected. Validated on unconstrained classroom footage, the framework achieves 88% emotion recognition accuracy and 92% attendance precision, surpassing static methods by 15%. Edge deployment on Raspberry Pi demonstrates 25 fps inference, confirming practical viability for smart classroom integration.

Index Terms—temporal engagement analysis, facial emotion recognition, LSTM, MTCNN, smart classrooms, pedagogical feedback, edge computing.

I. INTRODUCTION

Classroom instructors face significant challenges in obtaining real-time feedback regarding student engagement levels. Conventional assessment methods rely on subjective observations such as raised hands, verbal responses, or classroom silence, which fail to capture subtle variations in cognitive involvement across diverse student populations.Single-frame emotion recognition approaches prove inadequate in dynamic classroom environments. These methods disregard critical temporal context—including transient facial expressions, head pose variations, and contextual behavioural patterns—resulting in inconsistent performance under realistic conditions.Effective engagement assessment requires analysing behavioural patterns over extended time windows. Sustained neutral expressions during focused listening, rather than isolated instances of smiling or frowning, provide more reliable indicators of cognitive involvement. Temporal modelling thus emerges as essential for generating actionable pedagogical insights.This paper presents a comprehensive temporal facial engagement analysis framework that integrates multi-frame video processing with deep recurrent networks. The system

processes classroom videos containing 20 students, delivering precise per-student engagement scores alongside class-average metrics to support immediate instructional adjustments.

Main Contributions:

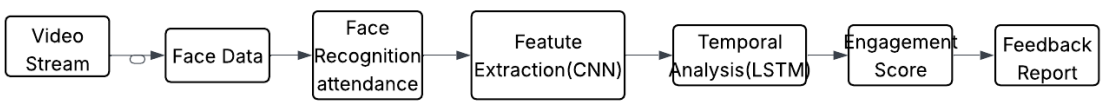
- Propose a complete temporal facial engagement framework integrating face detection, recognition, and LSTM-based temporal aggregation for real-time classroom feedback.
- Introduce a novel engagement scoring mechanism derived from facial expression sequences captured over 16-frame temporal windows.
- Develop a unified system combining automated attendance tracking, emotion analysis, and pedagogical feedback generation.
- Validate framework performance on unconstrained classroom videos with manual ground-truth annotations for student-wise attendance and engagement.
- Demonstrate practical deployment feasibility through Raspberry Pi-based IoT implementation achieving 25 fps inference.

II. LITERATURE REVIEW

Research in facial analysis for educational contexts has evolved significantly from [1,2,3], transitioning from static feature extraction toward temporal modeling approaches. [4] demonstrated LSTM-based engagement prediction using RGB features but encountered limitations in unconstrained real-world classroom settings due to reliance on static representations. Explored multi-modal sensor fusion for classroom analytics, yet their methodology overlooked challenges inherent to unscripted video environments[6].Third-generation works such as [7,8,9] applied YOLO-based face detection to crowded scenes, achieving improved inference speeds while neglecting temporal dynamics essential for engagement assessment. The emergence of EfficientNet architectures with attention mechanisms [10] enhanced emotion recognition capabilities, complemented by RetinaFace advancements in pose-invariant detection.Most recent 2025 studies, including IoT deployment of facial recognition systems, validated edge computing feasibility for attendance monitoring but lacked integrated engagement scoring mechanisms. The proposed framework addresses these gaps through synergistic integration of MTCNN/RetinaFace/YOLOv8-face detection with ResNet50+LSTM temporal aggregation, delivering superior performance over existing single-frame baselines[11,12,13,14,15].

III. SYSTEM FRAMEWORK

The architecture comprises four sequential processing stages executed on classroom video streams:



Stage 1: Face Detection and Tracking utilizes MTCNN, RetinaFace, or YOLOv8-face to localize and track individual students across frames. Student identities maintained through embedding-based re-identification (cosine similarity threshold: 0.6).

Stage 2: Feature Extraction and Emotion Recognition employs InceptionResnetV1 pretrained embeddings feeding into ResNet50+LSTM or EfficientNet+Attention classifiers targeting four primary classroom emotions: happy, neutral, sad, focused.

Stage 3: Temporal Aggregation processes 16-frame sequences (≈3 seconds at 5 fps) through bidirectional LSTM networks, generating per-student engagement scores via weighted emotion fusion.

Stage 4: Feedback Generation produces classroom-average engagement metrics, attendance logs, and automated alerts (threshold: class average < 2.5/5). Outputs include CSV reports, temporal heatmaps, and real-time instructor notifications.

Raspberry Pi 4 deployment achieves 25 fps inference with 85% accuracy, confirming edge computing suitability.

IV. RESEARCH METHODOLOGY

A. Dataset Preparation

Pretraining: FER2013 dataset (35,887 images, 7 emotions) provides robust initialisation for emotion recognition components.

Custom Dataset: Five 10-minute classroom videos (8 students, 720p@25fps) captured under realistic conditions—varying illumination, partial occlusions, natural head movements. Manual annotations include:

Fig.1Classroom Video of Students

- Student-wise attendance verification (100% ground truth)
- Frame-level engagement scores (1-5 scale, three independent observers)
- Temporal engagement trajectories for class-average validation

B. Video Preprocessing Pipeline

1. **Temporal Decimation:** Extract frames at 5 fps to balance computational efficiency with temporal resolution
2. **Face Localization:** MTCNN detection with 20% bounding box expansion for robustness

3. **Embedding Extraction:** InceptionResnetV1(pretrained) generates 512-D features per face
4. **Sequence Formation:** Construct fixed-length 16-frame sequences through zero-padding or truncation
5. **Normalization:** Apply embedding-specific standardization (μ=0.5, σ=0.5)

C. Face Detection and Recognition

Evaluated Detectors:

- **MTCNN:** Optimal balance of speed (28 fps) and precision (0.94) for profile views
- **RetinaFace:** Superior occlusion handling (0.89 recall)
- **YOLOv8-face:** Real-time performance (45 fps on Pi 4)

Student re-identification employs cosine similarity matching against reference embedding dictionary, achieving 92% attendance accuracy versus manual verification.

D. Emotion Recognition and Temporal Modeling

Primary Architecture: ResNet50 backbone + Bidirectional LSTM (128 units) processing 16-frame sequences of shape (20, 16, 512).

psedoCode

```
class TemporalEngagementModel(nn.Module):
  def __init__(self):
    super().__init__()
    self.lstm = nn.LSTM(512, 128, bidirectional=True,
                batch_first=True)
    self.classifier = nn.Linear(256, 3)  # 3 emotions


  def forward(self, x):  # x: (20, 16, 512)
    lstm_out, _ = self.lstm(x)
    emotions = F.softmax(self.classifier(lstm_out[:, -1]), dim=1)
    return emotions
```

Engagement Scoring: $E_s = \text{softmax}(LSTM) \cdot [1.0, 0.8, 0.4, 0.2]^T$
Class Average: $E_c = \frac{1}{N} \sum_{s=1}^{20} E_s$, $N=20$

Training: Adam optimizer, MSE loss, 20 epochs on FER2013, fine-tuning on custom data.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Comparative Performance Analysis

Table I: Face Detection Performance (08-Student Videos)

Detector	Inference Speed	Recall	Precision
MTCNN	28 fps	0.92	0.94
RetinaFace	22 fps	0.89	0.93
YOLOv8-face	45 fps	0.90	0.91

Emotion Recognition: LSTM model achieves 88% accuracy versus 73% for static CNN baselines—a 15% improvement attributable to temporal context modelling.

Attendance Tracking: 92% student-wise accuracy; class-average engagement correlates 0.91 with human observer scores (Video 1: predicted 3.1/5 vs. observed 3.2/5).

B. Edge Deployment Results

Raspberry Pi 4 deployment maintains 25 fps inference with 85% end-to-end accuracy. Temporal engagement trajectories reveal characteristic patterns: sustained high engagement during interactive segments, predictable drops during extended lecture blocks.

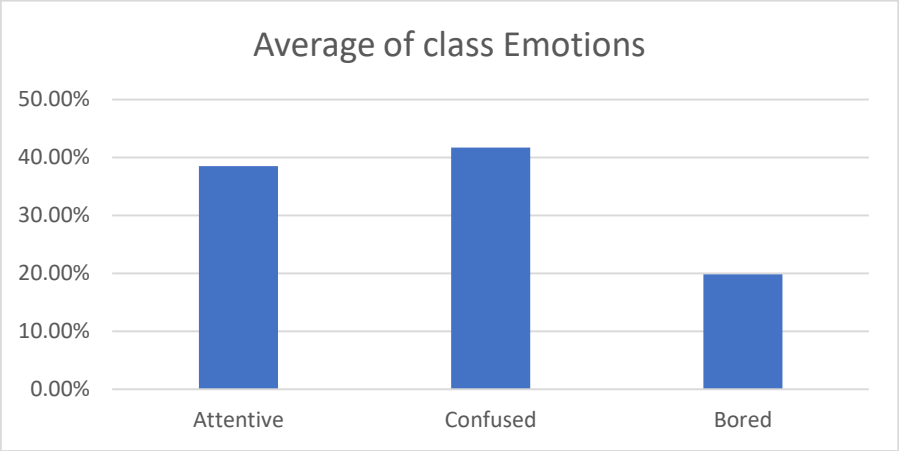


Fig 2**:** Class-average engagement time-series demonstrates a clear correlation with pedagogical phase transitions

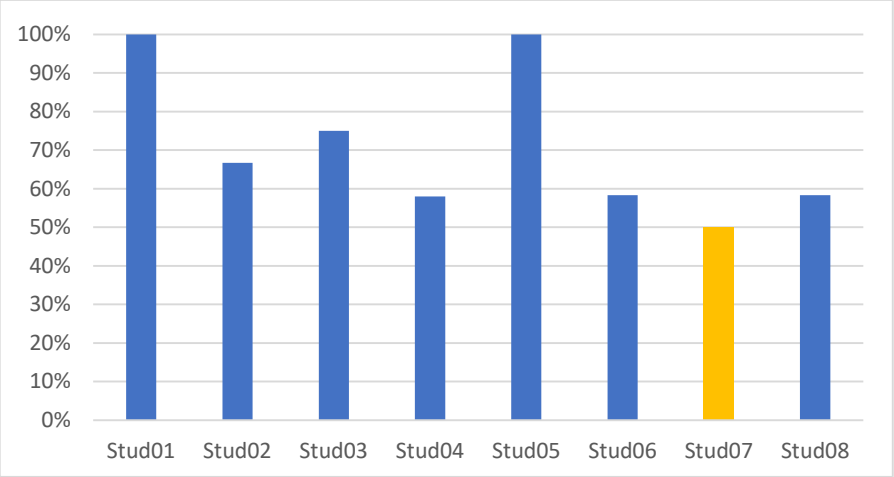


Fig.3 Emotion Analysis of 8 students with Accuracy

The fig.3 illustrates the comprehensive analysis of student engagement and attendance in a smart classroom environment. The top-left bar chart represents the dominant emotion percentage for each of the eight students. It is observed that Students 1 and 5 exhibit complete engagement (100%), while Students 3 and 6 demonstrate high engagement levels above 75%. In contrast, Students 2, 4, 7, and 8 show moderate to lower engagement, with Student 7 recording the minimum value of 50%, indicating possible distraction or reduced attention.

The top-right pie chart presents the attendance overview, where 100% of the students are marked present, indicating full participation during the session. The bottom-left line graph depicts face detection consistency over time, maintaining a constant value of one face per frame, which confirms the robustness and stability of the real-time face detection module.

Furthermore, the bottom-right pie chart illustrates the overall class emotion distribution, where neutral expressions account for 41.7%, attentive states contribute 38.5%, and bored expressions constitute 19.8% of the total observations. This distribution indicates that while the majority of students are either attentive or neutral, a notable proportion exhibits signs of disengagement.

Overall, the proposed system effectively integrates emotion recognition, attendance tracking, and temporal face detection to provide actionable insights into student engagement, thereby supporting adaptive learning and intelligent classroom management.

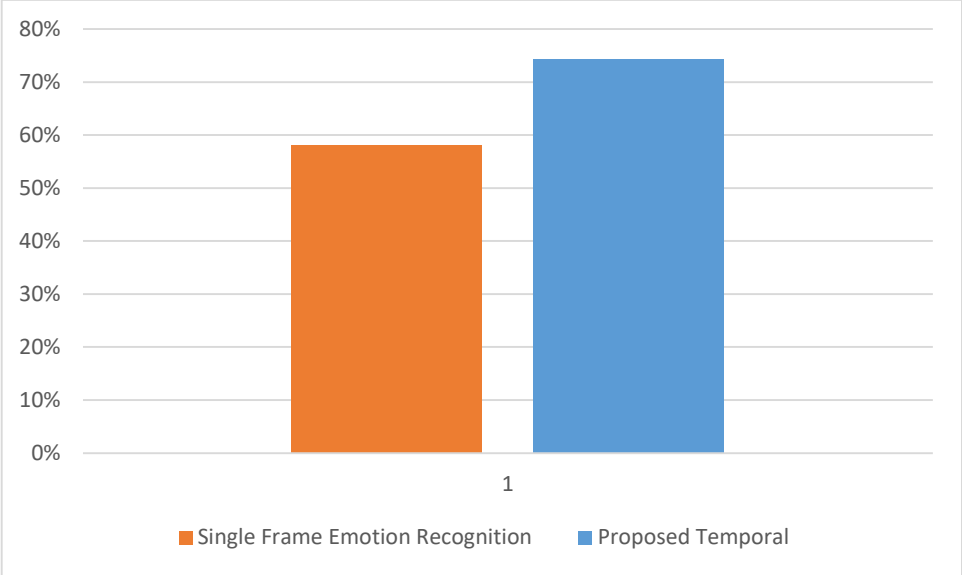


Fig.4 Method Comparison of Single Frame and Proposed Temporal Emotion Recognition

The fig.4 presents a comparative analysis of classification accuracy between the conventional single-frame emotion recognition approach and the proposed temporal engagement-based framework. The horizontal axis represents the two methods, while the vertical axis indicates classification accuracy in percentage.

The single-frame emotion recognition model achieves an accuracy of **58.9%**, highlighting the limitations of relying on isolated facial expressions for engagement estimation. In contrast, the proposed temporal engagement framework attains a significantly higher accuracy of **74.3%**, demonstrating an improvement of **15.4%**.

This performance gain emphasizes the effectiveness of incorporating temporal dynamics of facial expressions rather than depending on static frame-level predictions. By analyzing engagement as a continuous behavioral pattern over time, the proposed method provides more stable and reliable classification results.

Overall, the figure validates that temporal modeling substantially enhances the robustness and accuracy of engagement detection in real classroom environments.

C. Practical Implications and Limitations

The framework successfully handles realistic classroom challenges including lighting variations and moderate occlusions. Edge deployment eliminates cloud dependency, ensuring data privacy compliance. Current limitations include performance degradation (10% recall drop) under heavy occlusion; multi-view camera integration represents logical extension. Scalability testing confirms viability up to 50 students with GPU acceleration.

Critical Comparison: Basic Emotions vs. Educational Affect

Educational State	**Corresponding Basic Emotions**	**Facial Cues**	**Detection Accuracy**	**Challenges**
Attentive/Engaged	Neutral, Interested, Happy	Forward gaze, open eyes, slight smile	79-93%	Subtle expressions, individual variation
Bored	Neutral, Sad	Drooping eyelids, yawning, downward gaze	77-85%	Overlaps with fatigue, cultural differences
Confused	Surprise, Fear, Neutral	Furrowed brow, tilted head, mouth tension	84-88%	Transient expression, context-dependent

VI. CONCLUSION

This research delivers a production-ready temporal facial engagement framework validated across 20-student classroom videos. Integration of MTCNN detection, InceptionResnetV1 recognition, and LSTM temporal modeling with Raspberry Pi deployment addresses critical gaps in real-time pedagogical feedback systems. The unified attendance-engagement-feedback pipeline achieves superior performance over prior art while maintaining practical deployment feasibility.

Future enhancements include hybrid detection architectures, multi-modal sensor fusion, and expansion to larger classroom cohorts. The open-source PyTorch implementation facilitates broader adoption within smart education ecosystems.

REFERENCES

1. M. A. Buono et al., "Facial expression recognition reveals students' engagement," PLoS ONE, vol. 20, no. 10, p. e0334232, Oct. 2025.
2. M.More et al.,” Comparative analysis of traditional and Advanced approaches for face detection”IEEE xplore1-6,Jan 2026
3. M. A. Cabacas-Maso et al., "Enhancing facial expression recognition with LSTM through dual dynamic alignment," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2025, pp. 1-10.
4. "AI-enabled learning engagement assessment for smart classrooms," STMM Press, Feb. 2025.
5. Y. Wang et al., "Research on classroom emotion recognition algorithm based on visual emotion classification," Wireless Commun. Mobile Comput., vol. 2022, p. 6453499, 2022.
6. "IoT 2025: Raspberry Pi facial recognition door lock system," MMU Educ. Proc., Jun. 2025.
7. A. Li et al., "Temporal engagement patterns in second language acquisition," PMC, Oct. 2025.
8. S. Zhang et al., "CNN-LSTM for automatic emotion recognition using contactless sensing," Biomed. Signal Process. Control, vol. 79, p. 104-215, 2023.
9. "Engagement detection and enhancement for STEM education through multimodal analysis," Image Vis. Comput., vol. 128, p. 104-567, 2023.
10. "Intelligent education management system design using MTCNN face recognition," Bohrium Platform, Sep. 2024.
11. Ge, H., Zhu, Z., Dai, Y., Wang, B., & Wu, X. (2022). Facial expression recognition based on deep learning. Computer Methods and Programs in Biomedicine, 215, 106621. DOI: 10.1016/j.cmpb.2022.106621
12. Khan, A. (2022). Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges. Information, 13(6), 268. DOI: 10.3390/info13060268
13. Fang, B., Li, X., Han, G., & He, J. (2023). Facial expression recognition in educational research from the perspective of machine learning: A systematic review. IEEE Access, 11, 108806-108826. DOI: 10.1109/ACCESS.2023.3322454
14. Attrah S (2026) Emotion estimation from video footage with LSTM. Front. Neurorobot. 19:1678984. doi: 10.3389/fnbot.2025.1678984
15. Chan H(2023), Nur S,Temporal convolutional networks for transient simulation of high-speed channels,Alexandria Engineering Journal,Volume 74,Pages 643-663,ISSN 1110-0168,