# AI-Based Classroom Engagement Monitoring: A Cross-Cultural Feasibility Study

**Dr. Arash Mehdizadeh PhD, Australian University Kuwait**

Dr. Arash Mehdizadeh is an Assistant Professor in the Department of Electrical and Electronics Engineering at the Australian University, Kuwait. He holds a Ph.D. in Electrical and Electronic Engineering from the University of Adelaide, and two master's degrees—one in Computer Engineering from Amirkabir University of Technology and another in Data Science from the University of Texas at Austin. Over the past nine years, he has coordinated more than twenty university courses and led the development of curricula in Artificial Intelligence, Networking, Cybersecurity, and Biomedical Engineering. A strong advocate of Project-Based Learning (PBL), he has pioneered PBL course delivery frameworks and student assessment strategies at AU. His research interests include artificial intelligence, computer vision, data science, and biomedical engineering. Dr. Mehdizadeh has received multiple awards, including Best Paper and Innovation Awards, for his recent AI-powered education and healthcare projects.

**Dr. Amir Barakati, The George Washington University**

Dr. Barakati is an Associate Professor of Mechanical Engineering at The George Washington University. His research interests include the fabrication and characterization of biocomposites, the design of multifunctional substructures, and the integration of machine learning techniques into engineering education and research.

# AI-Based Classroom Engagement Monitoring: A Cross-Cultural Feasibility Study

Arash Mehdizadeh[1], Amir Barakati[2], and Bader Waleed Fraij[1]

[1]Department of Electrical and Electronics Engineering, College of Engineering, Australian University, Safat 13015, West Mishref, Kuwait

[2]Department of Mechanical and Aerospace Engineering, School of Engineering and Applied Science, The George Washington University, Washington, DC, USA

**Abstract:**

Understanding classroom engagement is essential for improving teaching effectiveness and promoting inclusive learning environments. However, traditional human observation methods often face limitations, especially in large or culturally diverse classrooms where students may be reluctant to ask questions or express emotions openly. To address these challenges, this work presents a vision-based artificial intelligence system for monitoring and analyzing student engagement in real time.

A convolutional neural network was fine-tuned using locally collected and incrementally expanded training data to better capture behavior patterns across different classroom contexts. The system provides three types of feedback: (i) live engagement visualization during lectures, (ii) periodic summary reports at defined intervals, and (iii) post-class analytics that categorize engagement states such as interested, disengaged, confused, tired, and talking. After data augmentation, approximately 15,000 labeled images were used for retraining, allowing the prototype to process classroom video at about 15 frames per second on standard laptop hardware, achieving near–real-time operation without specialized equipment.

Preliminary evaluations were conducted across classrooms in Kuwait and representative data for the United States. Accuracy on a held-out test set reached approximately 87%. Cross-validation with in-class surveys showed an agreement of approximately 82%. Cross-cultural variations, such as partial facial occlusion (e.g., head coverings) and variations in expressive behaviors, modestly influenced the model's sensitivity to engagement cues. Ongoing work focuses on expanding dataset diversity, improving temporal analysis methods, and conducting large-scale evaluations to assess robustness, fairness, and pedagogical value. Results indicate the feasibility of a low-overhead, privacy-conscious AI framework that provides educators with practical insights into classroom engagement across varied cultural and instructional settings.

**Keywords:** Classroom engagement, Artificial intelligence, Convolutional neural network, Real-time analytics, Cross-cultural evaluation, and educational technology.

## 1. Introduction

Student engagement remains one of the most critical factors influencing learning outcomes in engineering education. Engaged students demonstrate higher academic achievement, improved retention rates, and enhanced problem-solving capabilities, all essential competencies for future engineers [1]. However, accurately measuring and monitoring engagement in real-time classroom settings poses significant challenges for educators, particularly in large-enrollment courses typical of engineering programs. Traditional observation methods, while valuable, are labor-intensive, subjective, and difficult to scale, which limits their practical use for continuous classroom assessment [2].

The emergence of artificial intelligence (AI) and computer vision technologies offers promising solutions to these longstanding challenges. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have demonstrated remarkable capabilities in recognizing facial expressions, detecting behavioral patterns, and analyzing visual cues associated with student engagement [3]. These technologies enable automated, objective, and scalable monitoring systems that process visual data in real time, providing educators with actionable insights to dynamically adapt their instructional strategies [4], [5].

Despite these technological advances, a critical research gap persists: most AI-based engagement monitoring systems have been developed and validated within single cultural contexts, predominantly Western educational settings. This limitation raises fundamental questions about the cross-cultural validity and generalizability of these systems. Nonverbal communication patterns, including facial expressions, gestures, and body language, exhibit significant cultural variation [6], [7]. What constitutes "engaged" behavior in one cultural context may be interpreted differently in another, potentially leading to systematic biases and misclassification when AI systems trained on one population are deployed in culturally diverse classrooms.

In this work, the term "cross-cultural" is used broadly to encompass differences in educational contexts, behavioral norms, and population characteristics, rather than strictly controlled demographic variables. Accordingly, observed differences should be interpreted within this contextual scope.

### 1.1 Student Engagement Theory and Measurement

Student engagement has been conceptualized as a multidimensional construct encompassing behavioral, emotional, and cognitive dimensions [8], [9]. Behavioral engagement refers to observable actions such as attention, participation, and on-task behavior. Emotional engagement involves affective responses, such as interest, enthusiasm, and an emotional connection to learning activities. Cognitive engagement encompasses mental effort, strategic thinking, and self-regulation [10], [11]. Traditional engagement measurement approaches have relied primarily on self-report surveys, classroom observations, and learning analytics derived from digital learning platforms [2]. Trained observer protocols, such as the Classroom Observation Protocol for Undergraduate STEM (COPUS), offer more objective behavioral data but require extensive training, are resource-intensive, and introduce observer effects that may alter natural classroom dynamics [1].

Recent advances in multimodal learning analytics have enabled more sophisticated approaches to engagement measurement. [12] demonstrated that combining video analytics with wearable sensor data could improve average attention levels. Their system emphasized group-level analytics rather than individual tracking to address privacy concerns while maintaining pedagogical utility. Similarly, [13] developed a multimodal classroom engagement analysis dashboard for higher education that integrated multiple data streams to provide comprehensive engagement insights.

The Interactive-Constructive-Active-Passive-Disengage (ICAPD) framework proposed by [11] represents an important theoretical advancement by linking observable visual behaviors with cognitive engagement levels in classroom settings. This framework provides a theoretically grounded basis for automated engagement detection, addressing the challenge of connecting surface-level behavioral indicators to deeper cognitive processes. However, the cultural specificity of behavioral indicators remains an open question, as most engagement frameworks have been developed and validated primarily in Western educational contexts.

### 1.2 Computer Vision and Deep Learning in Education

The application of computer vision and deep learning to educational contexts has expanded rapidly in recent years, driven by advances in CNN architectures and increased computational accessibility. [5] developed a comprehensive AI-based classroom management system that demonstrated a 15% improvement in classroom engagement in experimental conditions and maintained high stability (≥85%) even in complex environments.

Object detection models, particularly the YOLO (You Only Look Once) family of architectures, have proven especially effective for real-time classroom monitoring applications. [14] combined YOLOv11 with the DeepSeek large language model to achieve high accuracy and responsiveness in behavior recognition while generating adaptive teaching suggestions. Further, [15] developed SBD-Net, which achieved a mean Average Precision (mAP) of 0.824 with only 9.8 G computational complexity. These results demonstrate the capability of modern architectures to perform high-precision detection in dynamically challenging classroom environments.

Real-time monitoring systems represent the practical implementation of computer vision and affective computing technologies in live educational settings. These systems must balance multiple competing requirements: accuracy, processing speed, privacy protection, scalability, and pedagogical utility. The challenge of real-time processing on standard hardware has been addressed through various optimization strategies [4], [16]. These systems represent important progress toward practical deployment, as many educational institutions lack access to specialized GPU infrastructure. However, most existing systems have been developed and validated within single institutional or cultural contexts. This limitation represents a significant gap in the literature, as deployment in diverse educational settings requires evidence of cross-cultural validity.

### 1.3 Privacy and Ethical Considerations

Recommendations for ethical AI implementation in education emphasize human-centered design, transparency, and equity. In that vein, [9] proposed that AI classroom management systems should prioritize educational purposes, implement appropriate technical safeguards, and address ethical considerations systematically. Further, [17] called for navigating ethical crossroads by addressing data privacy, bias, accountability, and sustainability in AI-driven education through comprehensive governance frameworks.

The ethical deployment of AI-based engagement monitoring requires careful attention to multiple stakeholder perspectives, including those of students, educators, administrators, and parents, and must balance potential benefits against risks of harm. Systems should be designed with privacy by default, provide clear opt-out mechanisms, limit data retention, prevent secondary uses of data, and undergo regular audits for bias and fairness [18]. Moreover, such systems should augment rather than replace human judgment, supporting educators in their professional practice rather than subjecting them to algorithmic management [9].

This research attempts to address this critical gap by investigating the feasibility of vision-based AI engagement monitoring across culturally diverse educational contexts, specifically comparing implementations in Kuwait and the United States. The remainder of this paper is organized as follows: Section 2 describes the methodology, including system architecture, data collection procedures, and validation approaches. Section 3 presents results and discussions from both deployment contexts. Section 4 concludes with directions for future research.

## 2. Methodology

This section describes the development, implementation, and evaluation of the CNN-based classroom engagement monitoring system across two culturally distinct educational contexts: Kuwait and the United States. The methodology encompasses system architecture design, dataset development through cross-cultural data collection, model training and optimization, system feature implementation, and evaluation procedures.

### 2.1 System Architecture and Design

The engagement monitoring system was designed with three primary objectives: achieving real-time processing capabilities on standard computing hardware, maintaining high classification accuracy across diverse student populations, and supporting scalability for typical classroom environments. Drawing on successful architectures demonstrated in recent literature, a modular CNN-based system optimized for educational deployment was developed.

Each incoming video frame is first processed by a face detection module to identify and localize individual students. Face detection is performed using a YOLOv8-based model loaded from a local pretrained checkpoint. YOLOv8 was selected for its high detection accuracy and low inference latency, both critical for real-time classroom monitoring. For each frame, the detector outputs bounding boxes corresponding to detected faces, which are then used to crop individual student face images. These cropped face regions serve as the input to the subsequent classification stage, restricting emotion recognition to the individual-student level rather than the full scene.

Emotion and engagement classification was performed using EfficientNet-B0 as the backbone architecture. EfficientNet-B0 employs compound scaling, jointly optimizing network depth, width, and input resolution to achieve strong classification performance with reduced computational complexity. The EfficientNet-B0 architecture contains approximately 5.3 million parameters and requires 0.39 billion floating-point operations (FLOPs), making it significantly more efficient than commonly used alternatives such as ResNet-50, which contains approximately 26 million parameters and requires over 4 billion FLOPs for comparable performance. These characteristics make EfficientNet-B0 suitable for real-time inference and potential on-premises deployment in classroom environments with constrained computational resources. The network was initialized with ImageNet-pretrained weights to leverage robust low-level and mid-level visual features learned from large-scale natural image data. The architecture was then adapted for facial emotion recognition by replacing the original classification head with a task-specific output layer.

Hardware specifications for deployment included standard laptop computers with Intel Core i5 or equivalent processors, 8 GB RAM minimum, and optional NVIDIA GPU support for enhanced performance. This specification maintains accessibility for typical educational institutions without requiring specialized computing infrastructure [3].

The contributions of this work include (i) a domain-specific fine-tuning strategy using culturally representative classroom data, (ii) a real-time engagement analytics pipeline that integrates frame-level predictions into temporally aggregated classroom-level insights, and (iii) a cross-cultural evaluation framework for comparing performance across distinct educational contexts. Unlike prior work focused on single-domain deployment, this study explicitly investigates domain adaptation challenges and their impact on engagement recognition.

### 2.2 Data Collection and Dataset Development

A critical component of this research involved developing a training dataset that captures emotions across both cultural contexts. Due to delays in obtaining the required institutional approvals for data collection in the US, a subset of AffectNet was selected to represent typical US classrooms. While AffectNet provides a large and diverse set of facial expressions, the authors acknowledge that it does not explicitly capture classroom-specific behavioral context. On the other hand, a deliberate data-collection strategy for Kuwait was implemented to enhance the system's generalizability and reduce cultural bias. Initial data collection occurred in authentic classroom settings at two private institutions in Kuwait. Both sites involved undergraduate engineering courses with enrollment ranging from 20 to 30 students. Informed consent was secured from all participating students, in accordance with ethical guidelines for educational AI research [9]. Students were informed of their right to opt out without penalty, and video data was collected only from consenting participants. All personally identifiable information was wiped from the frames before any processing.

Initial data collection involved recording approximately 10 hours of classroom instruction across both sites, capturing instructional activities in the first stage. Images were captured with HD webcams, positioned to maximize facial visibility while respecting student privacy, with careful attention to lighting and camera angles. The raw data was processed to extract individual images containing student faces, resulting in a dataset of approximately 15,000 labeled images after augmentation. Classifications were divided into one of five engagement categories based on observable behavioral indicators: interested (gaze directed toward instructor or materials, alert facial expression), disengaged (gaze averted, distracted behavior), confused (furrowed brow, head tilting, uncertain expression), tired (drooping eyelids, head resting on hand, yawning), and talking (mouth movement, gaze directed toward peers, animated expression).

To improve annotation reliability, students' self-reported engagement states were collected during class through brief questionnaires administered at 10-minute intervals, and timestamps were used to align responses with the corresponding image samples. While self-reports provide valuable insight into subjective engagement, they may be influenced by factors such as social desirability bias and momentary perceptions; however, frequent, time-aligned sampling helps mitigate recall-related inconsistencies. Furthermore, a dual-coding procedure was employed, with two independent raters labeling each image to enhance annotation consistency. The final dataset was partitioned into training, validation, and test sets using stratified sampling to maintain balanced representation of engagement categories and contextual variations across partitions. Critically, the test set included students who did not appear in the training or validation sets, ensuring that evaluation reflected the system's ability to generalize to previously unseen individuals rather than recognizing familiar faces.

### 2.3 Model Training and Fine-tuning

The EfficientNet-B0 model trained solely on AffectNet was evaluated on the Kuwait test and evaluation subsets without additional fine-tuning. This baseline evaluation yielded an accuracy of 67% on the local evaluation set. The observed reduction in performance indicates a domain shift between the source dataset and the target classroom environment, which can be attributed to ethnic, cultural, and contextual differences in facial expression patterns and non-verbal communication styles.

To mitigate the observed domain mismatch, the model was fine-tuned using the Kuwait training subset. Fine-tuning was conducted by retaining the pretrained EfficientNet-B0 feature extractor while updating the weights of the higher convolutional layers and the newly introduced classification head. This approach enabled adaptation to region-specific facial expression characteristics while preserving generalizable emotional features learned from large-scale data.

Following fine-tuning, classification accuracy on the Kuwait evaluation set increased to 87%. At the same time, general-domain accuracy remained stable at approximately 83%, comparable to the initial AffectNet-trained model's accuracy of approximately 85%, indicating that the domain adaptation process did not significantly compromise generalization performance. The fine-tuned model was integrated into the real-time classroom monitoring system. Frame-level predictions were generated for each detected student and were temporally aggregated to produce class-level engagement summaries. These aggregated measures were used to provide real-time feedback during lectures and post-class analytics for instructional evaluation.

### 2.4 Privacy and Ethical Considerations in Implementation

Throughout system development and deployment, rigorous ethical standards were maintained. All participants provided informed consent with a clear explanation of data usage and privacy protections. Video data was encrypted during transmission and processing, with access restricted to authorized research personnel. No personally identifiable information was stored beyond the minimum required for the study. Students could opt out at any time without penalty, and their data would be immediately deleted.

In addition to data privacy considerations, the potential psychological impact of continuous visual monitoring was explicitly considered. To minimize discomfort, the system was designed to operate at the group level without identifying individuals, and students were informed of its role as a research and instructional support tool. Informal feedback collected during sessions indicated initial awareness of monitoring, followed by gradual habituation over time. Nevertheless, the potential for behavioral modification due to observation (observer effect) remains a limitation and warrants further investigation in longitudinal studies.

Instructors received training on the appropriate use of engagement data, emphasizing that the system should inform rather than dictate instructional decisions. Regular check-ins with participants were conducted to assess psychological impacts and address concerns. These ethical safeguards align with recommendations from [9], [19], [20] for responsible educational AI deployment. All system outputs emphasize group-level rather than individual-level analytics to address privacy concerns and avoid creating a surveillance environment that could negatively impact student motivation and well-being. The system does not store individual student identities or create persistent profiles, and all video data is processed in real time without permanent storage beyond temporary buffers required for processing.

## 2.5 Evaluation Methodology

A continuous evaluation methodology was employed to assess system performance. Evaluation occurred in three phases: controlled test set evaluation, deployment, and cross-validation with student self-reports. The held-out test set of 2,250 labeled frames provided the primary measure of classification accuracy. Overall accuracy (percentage of correctly classified frames), per-category precision (proportion of predicted category instances that were correct), per-category recall (proportion of actual category instances correctly identified), and per-category F1-scores (harmonic mean of precision and recall) were computed. The confusion matrices were also analyzed to identify systematic misclassification patterns and computed accuracy for Kuwait.

To validate that automated classifications corresponded to students' subjective engagement experiences, periodic in-class surveys were conducted. At random intervals during recorded sessions, instruction was paused, and students were asked to self-report their engagement state using the same five categories. The automated classifications for those specific moments were then compared with student self-reports to compute agreement rates.

As discussed, delays in obtaining the necessary institutional approvals for system deployment and data collection for the U.S. necessitated that the reported results be limited to a test dataset assumed to be representative of a typical U.S. university classroom environment. The limitations of this assumption and the potential biases it may introduce when applied to real classroom settings are acknowledged. Efforts are currently underway to expedite the approval process and enable full system deployment, enabling data collection in operational classroom environments and supporting more comprehensive cross-cultural evaluation in future work.

# 3. Results and Discussion

This section presents the performance outcomes of the CNN-based engagement monitoring system across technical and practical dimensions. Figure 1 illustrates the real-time visualization interface available to the instructor during live classroom sessions. For each processed video frame, detected student faces are enclosed by bounding boxes and annotated with the corresponding predicted emotional or engagement state. In addition to individual-level annotations, aggregated class-level statistics are computed in real time, including the proportional distribution of each detected emotional state across the classroom.

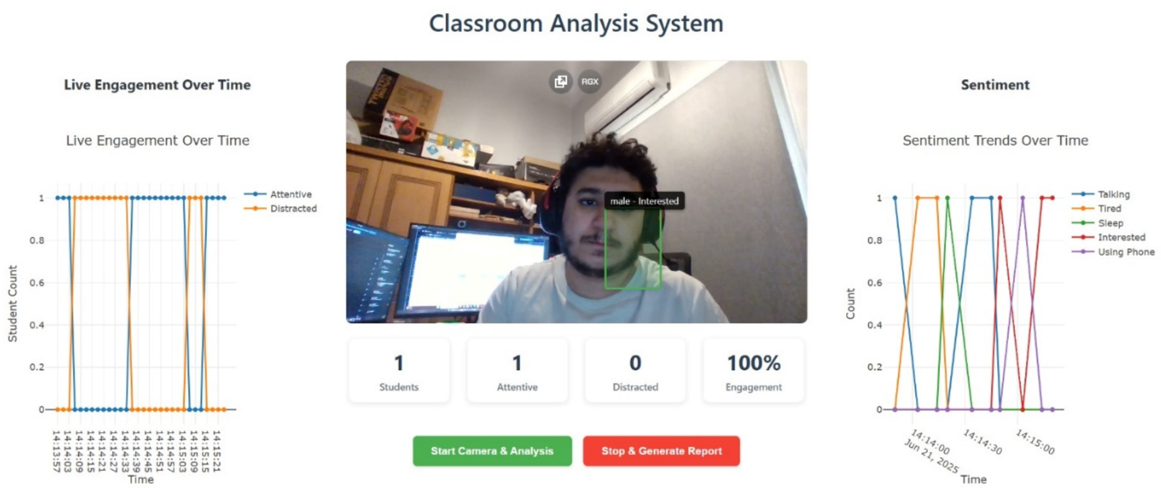


**Figure 1.** Real-time visualization interface

All frame-level predictions and aggregate measures are continuously logged in the background. These data are subsequently summarized into temporal averages computed over five-minute intervals, enabling longitudinal analysis of engagement dynamics throughout the class duration. At the conclusion of each session, a consolidated report is generated and made available to the instructor, providing a structured overview of engagement trends over time. Subject to explicit informed consent from all participating students, the system also supports optional storage of selected classroom snapshots. These stored frames can be used for post-session qualitative analysis and manual cross-referencing with questionnaire responses collected at corresponding timestamps, facilitating validation of the automated engagement estimates.

### 3.1 System Performance Metrics

The re-trained CNN model achieved strong performance on the held-out test set, demonstrating effective engagement classification across all five categories. Overall test set accuracy for Kuwait reached 87.3%. This performance is comparable to that of state-of-the-art systems reported in the recent literature [5], [16].

Table 1 presents detailed per-category performance metrics. The “interested” category achieved the highest performance (precision: 0.91, recall: 0.89, F1-score: 0.90), likely due to its clear behavioral indicators and higher representation in the training data. The “disengaged” category also showed strong performance (precision: 0.88, recall: 0.86, F1-score: 0.87), with distinctive behavioral patterns such as averted gaze and slouched posture facilitating accurate detection. The “talking” category demonstrated good performance (precision: 0.86, recall: 0.84, F1-score: 0.85), benefiting from the distinctive visual cue of mouth movement. The “confused” and “tired” categories showed somewhat lower but still acceptable performance (F1-scores: 0.83 and 0.81, respectively). Analysis of the confusion matrix revealed that “confused” was occasionally misclassified as “interested” (12% of confused instances), likely because students may maintain an attentive posture while experiencing confusion. Similarly, “tired” was sometimes misclassified as “disengaged” (15% of tired instances), as both states involve reduced alertness and attention.

**Table 1:** Per-category performance metrics on test set

| Engagement Category | Precision | Recall | F1-Score | Support (*n*) |
|---|---|---|---|---|
| Interested | 0.91 | 0.89 | 0.90 | 720 |
| Disengaged | 0.88 | 0.86 | 0.87 | 650 |
| Confused | 0.82 | 0.84 | 0.83 | 340 |
| Tired | 0.80 | 0.83 | 0.81 | 290 |
| Talking | 0.86 | 0.84 | 0.85 | 250 |
| **Overall** | **0.87** | **0.87** | **0.87** | **2,250** |

Processing speed met the target specification of 15 frames per second on standard laptop hardware (Intel Core i5-8250U, 8GB RAM, no dedicated GPU). When optional GPU acceleration was available (NVIDIA GTX 1650), processing speed increased to 28 frames per second, demonstrating scalability for higher-resolution video or larger classrooms. Cross-validation with student self-reported engagement surveys yielded an agreement rate of 82.4%. Specifically, when students indicated their engagement state via in-class surveys at 45 randomly selected moments across 12 class sessions, the automated system's classification matched student self-reports in 82.4% of cases. This validation metric is particularly important because it confirms that technical accuracy translates into meaningful engagement assessments aligned with students' subjective experiences. The 82% agreement rate compares favorably with [5], who reported similar validation approaches.

Analysis of disagreements between automated classifications and self-reports revealed several patterns. In 8% of cases, students self-reported as "interested" while the system classified them as "confused," suggesting that students may maintain engagement even when experiencing confusion. In 6% of cases, students self-reported as "disengaged" while the system classified them as "interested," potentially indicating social desirability bias in self-reports or students maintaining an attentive appearance despite internal disengagement. These findings highlight the complementary nature of automated and self-report measures, each capturing different aspects of the engagement construct [1], [2].

### 3.2 Preliminary Cross-Cultural Performance Analysis

A central contribution of this research is the systematic evaluation of system performance across culturally distinct educational contexts. Table 2 presents comparative performance metrics for the Kuwait and United States test subsets, revealing both the system's cross-cultural robustness and important contextual variations. As shown in Table 2, overall classification accuracy on the Kuwait subset reached 85.7%, compared to 83.0% on the U.S. (AffectNet) subset after domain-specific fine-tuning, resulting in a 2.7 percentage-point difference. This shift reflects the expected trade-off in domain adaptation, in which performance improvements in the target environment are accompanied by a modest reduction in general-domain accuracy. Importantly, absolute performance in both contexts remained high and within a range suitable for practical deployment, indicating that the fine-tuning strategy preserved generalization while enhancing local relevance. These results suggest that exposure to culturally representative data and attention to contextual behavioral patterns contributed to improved robustness across deployment settings.

**Table 2:** Cross-cultural performance comparison

| Metric | Kuwait Subset | US Subset (AffectNet) | Change % |
|---|---|---|---|
| Overall Accuracy | 85.7% | 83.0% | +3.2 |
| Interested (F1) | 0.90 | 0.86 | +4.6 |
| Disengaged (F1) | 0.87 | 0.83 | +4.8 |
| Confused (F1) | 0.83 | 0.79 | +5.0 |
| Tired (F1) | 0.81 | 0.76 | +6.5 |
| Talking (F1) | 0.85 | 0.81 | +4.9 |
| Test Set Size (samples) | 2,250 | 2,250 | — |
| Processing Speed (fps) | 15.2 | 15.1 | +0.1 |

Across individual engagement categories, consistent performance gains were observed for the Kuwait subset, with F1-score improvements ranging from +4.6% (Interested) to +6.5% (Tired). The largest gain was observed for the Tired category, suggesting that fatigue-related cues benefited most from exposure to locally representative data. In contrast, smaller yet consistent improvements were observed for visually distinctive behaviors such as Talking, indicating relatively culture-invariant visual features.

Processing speed remained effectively unchanged across both subsets (≈15 fps), confirming that performance improvements were attributable to model adaptation rather than architectural or computational changes. Overall, the results demonstrate that culturally informed fine-tuning enhanced classification reliability in the target classroom environment while maintaining acceptable generalization to external datasets.

### 3.3 Limitations and Future Work

A key limitation of the present study is reliance on representative, curated datasets due to delays in securing institutional approvals for full in-class deployment and data collection. While efforts were made to promote datasets that reflect realistic classroom conditions, such assumptions may introduce bias and limit the extent to which the findings generalize to fully operational educational environments. In particular, using AffectNet as a proxy for U.S.-based classroom behavior is a limitation, as it lacks contextual alignment with real classroom environments. Future work will address this by incorporating in-situ classroom data from U.S. institutions following ethical approval.

In addition, the relatively short duration of this preliminary phase restricted the ability to evaluate downstream educational impacts of the system. As a result, the study did not assess whether engagement-aware feedback translated into measurable improvements in classroom-level outcomes such as average academic performance or student satisfaction, nor did it include a controlled comparison between classrooms utilizing the proposed system and those operating without it.

Furthermore, although domain-specific fine-tuning improved performance in the target context, the evaluation was conducted on a finite set of engagement categories and did not account for longer-term temporal dependencies or multimodal behavioral cues. Future work will therefore focus on expediting institutional approvals to enable large-scale, longitudinal deployment in live classroom settings, allowing for controlled comparative studies across multiple classes and systematic evaluation of pedagogical impact. Planned extensions also include incorporating additional modalities, such as body posture and interaction dynamics, modeling temporal engagement patterns over extended periods, and exploring adaptive learning strategies to further improve robustness, fairness, and real-world applicability across diverse educational contexts.

## 4. Conclusions

This work demonstrated the feasibility of a vision-based artificial intelligence system for real-time analysis of student engagement in classroom environments. By integrating face detection with a fine-tuned convolutional neural network, the proposed system provided reliable live visualization, periodic summaries, and post-class analytics of engagement states using standard computing hardware. The results showed that domain-specific adaptation improved classification performance in the target classroom context while maintaining acceptable generalization across external datasets, supporting the practical deployment of such systems in culturally diverse educational settings. Cross-cultural evaluation further highlighted the influence of contextual and expressive differences on engagement recognition, including partial facial occlusion and variations in behavioral expressiveness. Despite these challenges, robust performance was maintained, and the importance of incorporating locally representative data during model adaptation was emphasized. The study allows educators to receive live engagement visualization during lectures for real-time instructional adjustment, longitudinal assessment of pedagogical and teaching techniques, and inclusive learning analysis.

Due to the short duration of this study's preliminary phase, the translation of engagement-level insights into measurable improvements in classroom-level academic outcomes, such as average grades or student satisfaction, could not be evaluated. In addition, no comparative analysis was conducted between classrooms utilizing the proposed system and those operating without it. These aspects represent the key next stages of the research. Future work will therefore focus on longer-term deployments, controlled comparative studies across multiple classes, and systematic evaluation of the system's impact on learning outcomes, student experience, and instructional effectiveness. Together with ongoing efforts to expand dataset diversity and incorporate temporal and multimodal cues, these directions aim to establish a clearer link between engagement analytics and pedagogical outcomes.

## Acknowledgements

This work was partially supported by the Australian University Internal Grant (No. IRC-2022/2023-COE-EE-PR03) and the Kuwait Foundation for the Advancement of Sciences (KFAS) under Grant No. CN2418TT2286. The authors also thank F. Al-Hubabi, M. Taleb, R. Al-Malek, G. Al-Jeridan, and O. Muzher for their valuable contributions to the collection of training data and the development of the initial system prototype.

## References

[1] P. Goldberg *et al.*, "Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction," *Educ. Psychol. Rev.*, vol. 33, no. 1, pp. 27–49, Mar. 2021, doi: 10.1007/s10648-019-09514-z.

[2] S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, Analytic, Automated (AAA) Measurement of Engagement During Learning," *Educ. Psychol.*, vol. 52, no. 2, pp. 104–123, Apr. 2017, doi: 10.1080/00461520.2017.1281747.

[3] Z. Trabelsi, F. Alnajjar, M. M. A. Parambil, M. Gochoo, and L. Ali, "Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 48, Mar. 2023, doi: 10.3390/bdcc7010048.

[4] A. Habib Albohamood, M. Shaker Alqattan, and C. Padua Vizcarra, "Real-time Student Engagement Monitoring in Classroom Environments using Machine Learning and Computer Vision," in *2025 4th International Conference on Computing and Information Technology (ICCIT)*, IEEE, Apr. 2025, pp. 420–424. doi: 10.1109/ICCIT63348.2025.10989406.

[5] R. Yang, “Research on Classroom Management and Student Engagement Analysis System Based on Artificial Intelligence,” in *Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, New York, NY, USA: ACM, Mar. 2025, pp. 1385–1391. doi: 10.1145/3745238.3745454.

[6] H. Pang, “Individuating emotional clarity in multicultural classrooms of higher education: A transnational study on nonverbal dynamics of enculturation and acculturation,” *Cogent Education*, vol. 3, no. 1, p. 1200834, Dec. 2016, doi: 10.1080/2331186X.2016.1200834.

[7] N. Xie, Z. Liu, Z. Li, W. Pang, and B. Lu, “Student engagement detection in online environment using computer vision and multi-dimensional feature fusion,” *Multimed. Syst.*, vol. 29, no. 6, pp. 3559–3577, Dec. 2023, doi: 10.1007/s00530-023-01153-3.

[8] Ö. Sümer, P. Goldberg, S. D’Mello, P. Gerjets, U. Trautwein, and E. Kasneci, “Multimodal Engagement Analysis From Facial Videos in the Classroom,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1012–1027, Apr. 2023, doi: 10.1109/TAFFC.2021.3127692.

[9] T. Fütterer *et al.*, “Artificial Intelligence in Classroom Management: A Systematic Review on Educational Purposes, Technical Implementations, and Ethical Considerations,” Dec. 30, 2023. doi: 10.31219/osf.io/wfazn.

[10] R. Xu, J. Chen, J. Han, L. Tan, and L. Xu, “Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information,” *Computing*, vol. 102, no. 3, pp. 765–780, Mar. 2020, doi: 10.1007/s00607-019-00722-7.

[11] Q. Xu, Y. Wei, J. Gao, H. Yao, and Q. Liu, “ICAPD Framework and simAM-YOLOv8n for Student Cognitive Engagement Detection in Classroom,” *IEEE Access*, vol. 11, pp. 136063–136076, 2023, doi: 10.1109/ACCESS.2023.3337435.

[12] T. H. TEO, J. B. W. Heng, and C. L. Kok, “Learning Analytics Through Video Analytics and Wearable Sensors for Real-Time Attention Monitoring in Classrooms,” Apr. 23, 2025. doi: 10.31219/osf.io/q6d9r_v1.

[13] A. Sabuncuoglu and T. M. Sezgin, “Developing a Multimodal Classroom Engagement Analysis Dashboard for Higher-Education,” *Proc. ACM Hum. Comput. Interact.*, vol. 7, no. EICS, pp. 1–23, Jun. 2023, doi: 10.1145/3593240.

[14] J. Long and M. K. Chamran, “Smart Classroom Monitoring: Combining Yolov11 and Deepseek for Real-Time Behavior Recognition and Teaching Suggestions,” in *2025 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*, IEEE, Jul. 2025, pp. 1–6. doi: 10.1109/ISIEA65768.2025.11138423.

[15] Z. Wang, M. Wang, C. Zeng, and L. Li, "SBD-Net: Incorporating Multi-Level Features for an Efficient Detection Network of Student Behavior in Smart Classrooms," *Applied Sciences*, vol. 14, no. 18, p. 8357, Sep. 2024, doi: 10.3390/app14188357.

[16] V. K, J. K. L. Jose, T. M. Thiyagu, J. Archpaul, C. J. R, and S. P, "Smart Focus: A Deep Learning-Based Student Engagement Tracking System," in *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, Jun. 2025, pp. 1509–1514. doi: 10.1109/ICIRCA65293.2025.11089732.

[17] T. Dhiman, V. Chauhan, A. Kumar, M. Vasantha, and A. Kumar, "Ethical Crossroads: Navigating Data Privacy, Bias, Accountability and Sustainability in AI-Driven Education," *Open Access Journal of Multidisciplinary Research*, vol. 1, no. 3, pp. 69–77, Jun. 2025, doi: 10.47760/OAJMR.2025.v01i03.007.

[18] I. A. Ismail, "Protecting Privacy in AI-Enhanced Education," 2024, pp. 117–142. doi: 10.4018/979-8-3693-0884-4.ch006.

[19] S. Coghlan, T. Miller, and J. Paterson, "Good Proctor or 'Big Brother'? Ethics of Online Exam Supervision Technologies," *Philos. Technol.*, vol. 34, no. 4, pp. 1581–1606, Dec. 2021, doi: 10.1007/s13347-021-00476-1.

[20] M. Nawaz, N. Awan, S. Ahmed, and A. Mustafa, "Surveillance Pedagogy: The Psychological and Pedagogical Risks of AI-Based Behavioral Analytics in Digital Classrooms," *ACADEMIA International Journal for Social Sciences*, vol. 4, no. 3, pp. 1995–2010, Aug. 2025, doi: 10.63056/ACAD.004.03.0508.