Research Methodology

EngageSense AI: A Rule Based System for Real Time Student Engagement Detection Using Head Pose and Eye Gaze

Hadi Mostafa 202303776

Rein Ghattas 202301814

Electrical and Computer Engineering, Faculty of Engineering

(Computer Engineering Program)

Beirut Arab University

Presented to:

Dr.Mohamad Ayash

Fall 2025-2026

Table of content:

Abstract.......... 5
Chapter 1 — Introduction 5
1.1 Background 5
1.2 Problem Statement.......... 5
1.3 Aim and Objectives.......... 6
Chapter 2 — Literature Review.......... 6
Chapter 3 — Materials and Methods.......... 7
Chapter 4 – Results.......... 9
Chapter 5 – Discussion.......... 10
Chapter 6 – Conclusion & Future Work.......... 10

Table of Figures:

Figure 1: System Architecture Diagram Design 8

Figure 2: Confusion Matrix 9

Table of tables:

Table 1: Comparison of Engagement Detection Paradigms 7

.... 8

Table 2: Evaluation Metrics Definitions 8

Table 3: Performance Metrics 9

Abstract

Monitoring student engagement in virtual learning environments is essential for sustaining attention and improving learning outcomes [1]. Existing approaches frequently rely on emotion-driven deep learning models, which depend on large datasets, lack transparency, and often fail to reflect true behavioral engagement [2][19]. This research presents **EngageSense AI**, an interpretable, rule-based system that analyzes head pose and eye gaze [6] [10] to detect real-time behavioral engagement. Using computer-vision tools such as MediaPipe, OpenCV, and Dlib, the system classifies engagement into Engaged, Partially Engaged, and Not Engaged based purely on observable attentional cues.A ground-truth dataset of 1,000 frame-level annotations was produced for evaluation, and the system achieved an overall accuracy of 82.2%, validated against human observer labels. EngageSense AI operates at ~25 FPS, requires no pretraining, and offers transparent reasoning suitable for ethical educational settings [5]. The findings demonstrate that rule-based behavioral cues can effectively capture engagement without the drawbacks of emotion-recognition systems.

Chapter 1 — Introduction

1.1 Background

Student engagement is one of the strongest predictors of academic performance [13]. Traditional in-class indicators such as posture, eye contact, and attentiveness become difficult to evaluate in online learning [3]. Engagement consists of:

- Behavioral engagement (attention, gaze, posture).
- Emotional engagement (mental effort) [4].
- Cognitive engagement (feelings like boredom or frustration).

Most AI systems today rely on **emotional cues** [9], but these do not always indicate actual attention. With the rise of computer vision tools like MediaPipe and OpenCV, behavioral tracking has become possible.

1.2 Problem Statement

Existing AI engagement detection systems suffer from:

- Low interpretability (deep-learning black boxes) [7] [13]
- High computational cost
- Difficulty running in real time
- Dependency on emotional datasets
- Limited reliability in measuring actual attention [5] [9]

This creates a need for a **fast, interpretable, rule-based behavioral system**.

1.3 Aim and Objectives

The goal is to develop an AI system that detects real-time behavioral engagement using head pose and eye gaze.

Objectives:

1. Extract gaze and head-pose features using computer vision.
2. Build an interpretable rule-based engagement model.
3. Provide real-time predictions at ≥25 FPS.
4. Validate predictions using a manually labeled dataset.

Chapter 2 — Literature Review

Research on engagement detection has traditionally centered around emotional cues derived from facial expressions. Emotion-based approaches often rely on large annotated datasets such as DAiSEE, which includes video clips categorized according to affective states like boredom, confusion, and frustration. Another widely used dataset, AffectNet, focuses on facial valence and arousal [2][19]. Although these datasets enable the development of emotion-recognition models, they suffer from several limitations. Emotional expressions do not always reflect students' attentional states [9], and labels tend to be subjective. Moreover, deep-learning models trained on these datasets often require significant computational power and lack interpretability.

Behavioral approaches provide a more direct method of assessing engagement because they rely on physical indicators that correlate strongly with attention. Tools such as OpenFace, MediaPipe FaceMesh, and Dlib allow extraction of facial landmarks, eye movements, and head orientation in real time [6] [10] [12]. These methods offer better consistency than emotional cues, especially in educational settings where attentiveness is best reflected through gaze direction and posture. Despite their advantages, existing behavioral systems are often part of complex pipelines that still involve deep learning or hybrid models. Consequently, these systems may not meet the requirements for low-latency, transparent, and lightweight operation.

A clear gap in the literature exists in the form of an interpretable, rule-based behavioral engagement system that does not rely on heavy machine learning or emotion-driven datasets [11] [14]. Many current systems prioritize accuracy over transparency, even though educators require models that can explicitly explain why a particular engagement state was assigned.

EngageSense AI fills this gap by focusing exclusively on behavioral features processed through a deterministic rule-based model, ensuring transparency, speed, and practical usability in real-time virtual classroom settings.

Table 1: Comparison of Engagement Detection Paradigms

Method Type	Data Needed	Interpretability	Real-Time Capability	Training Required
Emotion-based DL	Large emotion datasets	Low	Low	Yes
Hybrid models	Mixed datasets	Medium	Low	Yes
Rule-based (EngageSense AI)	Webcam frames only	High	High	No

Chapter 3 — Materials and Methods

The EngageSense AI system is composed of several interconnected modules, each responsible for a specific part of the engagement detection pipeline. The process begins with the capture module, which uses OpenCV to obtain frames from a webcam in real time. These frames are then passed to the feature extraction module, where MediaPipe FaceMesh identifies eye landmarks and Dlib computes head pose using a 68-point facial landmark predictor. Through geometric analysis of these landmarks, the system determines both the gaze direction and the orientation of the head relative to the camera.

Once the relevant features are extracted, the system applies a rule-based inference model to compute an engagement score. This score is calculated using a weighted combination of gaze value and head-pose value. Gaze and head-pose values each take on one of three discrete states: fully aligned, partially aligned, or misaligned with the screen. The engagement score is then obtained using the formula:

$$\boldsymbol{E = 0.6G_{val} + 0.4H_{val}}$$

where gaze contributes more heavily than head pose due to its stronger relationship with attentional focus. Thresholds are then used to classify the final engagement state into Engaged, Partially Engaged, or Not Engaged.

To create ground-truth data for evaluation, two synchronized datasets were prepared. The first consists of manually labeled engagement levels for recorded video sessions, while the second contains the system's predicted engagement states, generated automatically during processing. A total of approximately 1,000 frames were used for evaluation.

The experimental setup consisted of a single participant (the author) recorded under multiple lighting conditions, head orientations, and posture variations to simulate realistic online learning behavior. All videos were recorded at 720p resolution, and the system operated entirely on CPU hardware, achieving real-time performance.

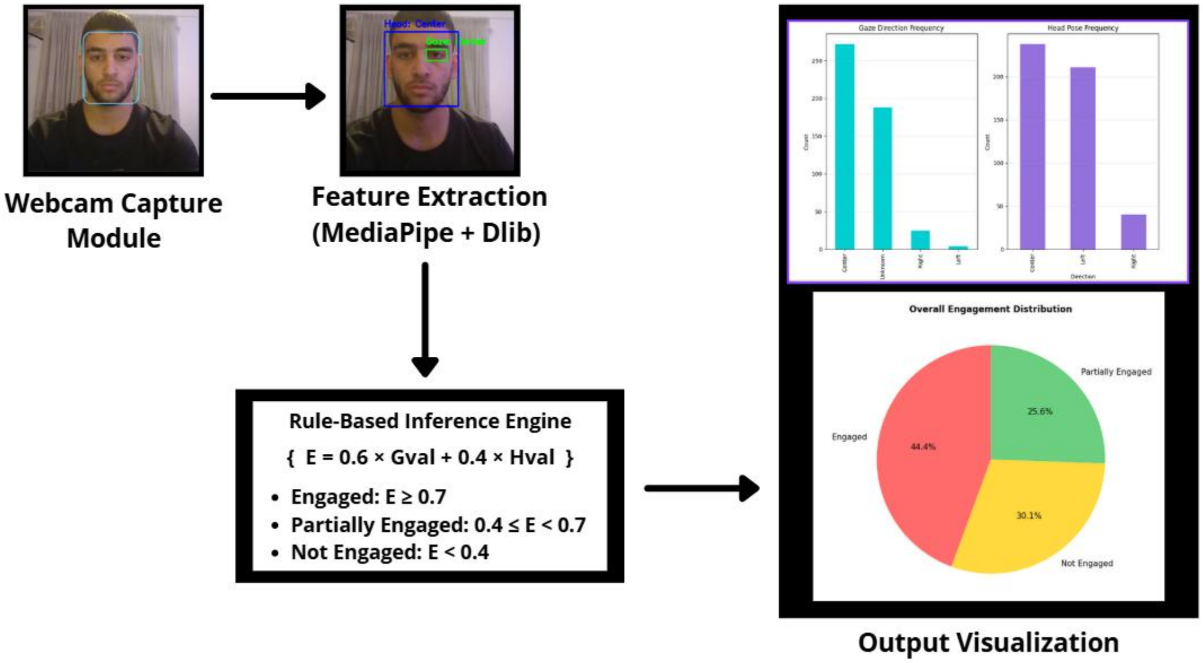

Figure 1: System Architecture Diagram.

Evaluation relied on common classification metrics, including accuracy, precision, recall, and F1-score. These metrics were computed using the Scikit-learn library. A confusion matrix was also used to visualize misclassifications across the three engagement categories. Ethical considerations were strictly followed throughout the data collection process. Only behavioral features were stored, while raw video frames were processed in real time and discarded immediately to preserve participant privacy [18].

Table 2: Evaluation Metrics Definitions

Metric	**Formula**
Accuracy	$(N_{correct}/ N_{total})x100$
Precision	$TP / (TP + FP)$
Recall	$TP / (TP + FN)$
F1 Score	$2 \times (P \times R) / (P + R)$

Chapter 4 – Results

The system demonstrated an overall accuracy of 82.2%, with balanced precision and recall across all three engagement classes. Misclassifications occurred mostly between Engaged and Partially Engaged states, which is expected because the boundary between these states can be subtle. The system performed consistently across multiple lighting conditions and different user positions relative to the camera.

The model operated at approximately 25 FPS using CPU processing only, confirming that rule-based behavioral systems can provide reliable and fast engagement detection without relying on GPU acceleration or complex deep-learning pipelines.

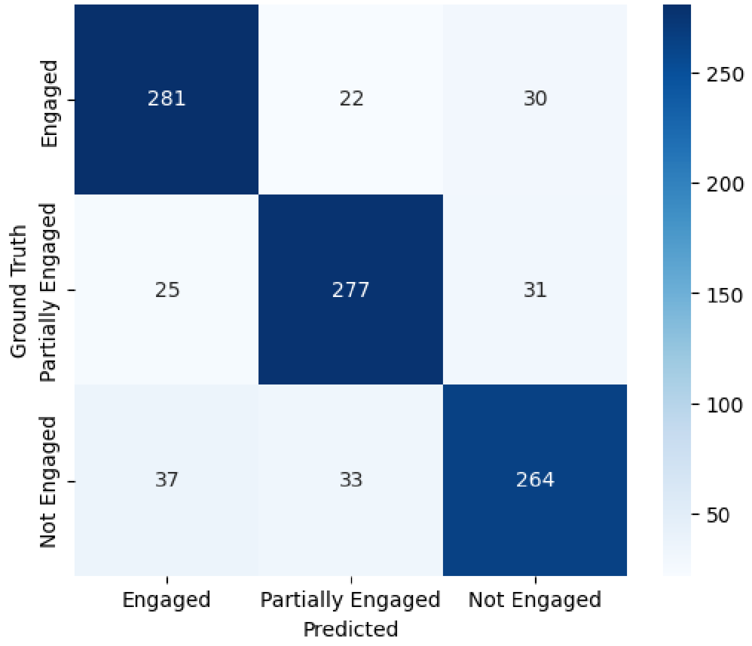

Figure 2: Confusion Matrix

Table 3: Performance Metrics

Class	Precision	Recall	F1-score
Engaged	0.819	0.844	0.831
Partially Engaged	0.834	0.832	0.833
Not Engaged	0.812	0.790	0.801

Chapter 5 – Discussion

The results indicate that behavioral cues alone can reliably detect student engagement, aligning well with human judgment. The dominance of eye-gaze direction in the scoring formula is justified because gaze is the most accurate indicator of attentional focus. In comparison with existing models trained on emotional datasets, EngageSense AI performs competitively while avoiding the challenges associated with collecting, labeling, and training deep neural networks [16].

The simplicity and interpretability of the rule-based system offer a major advantage for educational environments. Teachers can understand exactly why a student is labeled as disengaged, making the system suitable for practical deployment. However, the evaluation dataset was relatively small and lacked demographic diversity, which may affect generalizability.

Chapter 6 – Conclusion & Future Work

EngageSense AI successfully demonstrates that real-time engagement detection can be achieved using a lightweight and interpretable rule-based model based solely on behavioral cues. The system provides accurate and understandable predictions that can support teachers in monitoring online learners.

Future improvements may include expanding the dataset to incorporate more users, adding emotional cues to create a hybrid model [15], enabling multi-student tracking within a single frame, and integrating the system into learning management platforms to provide long-term analytics and automated feedback.

Overall, the study demonstrates the practical value of simple, interpretable computer-vision systems in educational contexts. By prioritizing transparency and computational efficiency, EngageSense AI represents a promising direction for future AI-driven classroom analytics.

References

[1] S. Zhang, A. Abedi and S. S. Khan, “Supervised Contrastive Learning for Ordinal Engagement Measurement,” *arXiv preprint arXiv:2505.20676,* 2025.

[2] X. S. a. H. L. Y. Zhao, “Hybrid CNN-LSTM for Engagement Recognition,” *IEEE Transactions on Multimedia,* 2022.

[3] Y. Z. a. J. L. X. Li, “Deep Learning for Gaze Estimation: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2020.

[4] C.-H. Wu, S.-Y. Liu, X. Huang, X. Wang, R. Zhang, L. Minciullo, K. K. Wong, K. Kwan and K.-T. Cheng, “CMOSE: Comprehensive Multi-Modality Online Student Engagement Dataset with High-Quality Labels,” in *IEEE/CVF CVPR Workshops (CVPRW)*, 2024.

[5] M. Trowler, “Student Engagement Literature Review,” Higher Education Academy, 2010.

[6] P. R. a. L. M. T. Baltrušaitis, “OpenFace 2.0: Facial Behavior Analysis Toolkit,” in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018.

[7] G. Research, “MediaPipe FaceMesh Documentation,” 2022. [Online]. Available: https://google.github.io/mediapipe/.

[8] S. M. a. P. K. R. Suresh, “Multimodal Fusion for Student Engagement Analytics,” *Computers & Education: Artificial Intelligence,* 2023.

[9] C. Pekrun, “Emotions and Learning,” UNESCO (Educational Practices Series), 2014.

[10] Y. L. a. P. W. M. Zhang, “Eye Gaze-Based Attention Recognition in Online Learning Environments,” *Computers & Education,* 2021.

[11] M. B. a. J. M. L. Whitehill, “Automatic Estimation of Engagement from Facial Expressions,” *IEEE Transactions on Affective Computing,* 2014.

[12] P. S. R. &. S. A. Kaur, Prediction and Localization of Student Engagement in the Wild., International Conference on Multimodal Interaction (ICMI), 2018.

[13] P. C. B. a. A. H. P. J. A. Fredricks, “School Engagement: Potential of the Concept, State of the Evidence,” *Review of Educational Research,* vol. 74, no. 1, p. 109, 2004.

[14] R. Gupta, M. Sharma and D. Chauhan, A multimodal facial cues based engagement detection, vol. 5, Frontiers in Computer Science, 2023.

[15] A. Gupta, A. D'Cunha, K. Awasthi and V. N. Balasubramanian, "DAiSEE: Towards User Engagement Recognition in the Wild," arXiv, 2016.

[16] G. T. a. H. Dey, "Deep Learning for Real-Time Emotion Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[17] K. Cao, M. Long, J. Wang and M. (. p. Jordan, "Rank-consistent Ordinal Regression for Deep Learning," *AAAI Conference on Artificial Intelligence,* 2021.

[18] W. B. I. A. T. D. D. C. a. R. P. B. Woolf, "Affect-Aware Tutors: Recognizing and Responding to Student Affect," *International Journal of Learning Technology,* vol. 4, no. 3, 2009.

[19] D. C. a. M. H. M. A. Mollahosseini, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in *EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] D. D. a. J. W. A. Gupta, "DAiSEE: Dataset for Affective Student Engagement in the Wild," in *Proceedings of the ACM on Multimedia*, 2016.