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

Hadi Mostafa
Dept. of Electrical and Computer Engineering
Beirut Arab University
Beirut, Lebanon
Email: hmm377@student.bau.edu.lb

Rein Ghattas
Dept. of Electrical and Computer Engineering
Beirut Arab University
Beirut, Lebanon
Email: rmg181@student.bau.edu.lb

***Abstract*—Monitoring student engagement in virtual learning environments is essential for sustaining attention and improving academic outcomes [5], [13]. Many existing approaches rely on emotion-based deep learning models [15], [19] that require large datasets, lack interpretability, and do not necessarily reflect actual behavioral attention [9]. This paper introduces EngageSense AI, an interpretable, rule-based system that estimates engagement using head pose and eye gaze extracted with MediaPipe, OpenCV, and Dlib [6], [7]. The system classifies engagement into Engaged, Partially Engaged, and Not Engaged based on geometric cues. A dataset of 1,000 manually annotated frames was collected for evaluation. EngageSense AI achieved an accuracy of 82.2% while operating at approximately 25 FPS on CPU hardware. Results demonstrate that lightweight, transparent rule-based models can effectively detect engagement without the computational demands of deep learning [1], [2].**

***Index Terms*—Student engagement, gaze tracking, head pose estimation, real-time systems, rule-based models, behavioral analytics**

I. Introduction

Student engagement is widely recognized as a key predictor of learning outcomes and academic success [5], [13]. In traditional classrooms, teachers can rely on observable behavioral indicators such as gaze direction, posture, and responsiveness. In virtual environments, however, these cues become difficult to monitor consistently, motivating the need for automated engagement detection systems.

Most current AI-based engagement systems focus on affective or emotional cues, using facial expressions and related signals to infer boredom, confusion, or frustration [11], [15], [19]. While affective information is important, emotional expressions do not always correlate with attentional focus, and labels in such datasets are often subjective [9], [18]. Deep learning architectures applied to these datasets [1], [2], [14] typically require large training sets and powerful hardware, and they act as black-box models with limited interpretability.

Recent advances in computer vision frameworks such as MediaPipe FaceMesh, OpenFace, and Dlib [6], [7], [12] enable robust, real-time tracking of head pose and gaze direction in standard webcam settings. These behavioral cues are more directly related to attention than emotional labels [3], [10].

This paper proposes EngageSense AI, a lightweight and interpretable rule-based system that infers engagement from head pose and gaze direction alone. The main contributions are:

- A fully rule-based framework relying solely on geometric behavioral cues, with no deep learning component.
- A real-time implementation using commodity hardware (CPU only) and widely available computer vision libraries.
- An empirical evaluation using manually annotated frame-level engagement labels, demonstrating competitive performance.

II. Literature Review

Engagement detection methods can broadly be grouped into emotion-based, hybrid, and purely behavioral approaches.

A. Emotion-Based Methods

Emotion-based systems typically rely on large-scale affective datasets such as AffectNet and DAiSEE, which include labels for valence, arousal, or engagement-related affective states like boredom and confusion [15], [19]. Deep neural networks, including CNNs and CNN-LSTM hybrids, are trained on these datasets for engagement recognition [2], [11]. Although these models achieve good quantitative performance, they suffer from several limitations:

- Labels are inherently subjective and may not reflect true attention levels [9].
- Models require extensive training data and GPU resources [16].
- Interpretability is low, making it difficult for educators to understand why a given prediction was produced.

B. Hybrid and Multimodal Methods

Hybrid systems combine facial expressions, audio, gaze, and contextual signals to improve robustness [4], [8], [14]. Some works use multimodal fusion strategies or advanced ordinal regression techniques to better capture engagement levels [1], [17]. While hybrid models can improve accuracy, they further increase complexity and reduce transparency.

C. Behavioral and Gaze-Based Methods

Behavioral approaches focus on markers that more directly correlate with attention, such as gaze direction and head orientation. Surveys on gaze estimation using deep learning [3] and studies on gaze-based attention recognition in online environments [10] show that eye gaze is a strong predictor of engagement. Toolkits like OpenFace and MediaPipe FaceMesh provide robust, real-time facial landmark extraction [6], [7], enabling head pose estimation and gaze tracking on standard webcams [12].

Despite these advances, many behavioral systems still integrate deep neural networks or complex probabilistic models, which reintroduce opacity and computational overhead. EngageSense AI addresses this gap by using a fully deterministic, rule-based approach grounded in behavioral cues alone.

III. Materials and Methods

A. System Overview

EngageSense AI is composed of four main modules:

1) Real-time frame capture using OpenCV.
2) Feature extraction using MediaPipe FaceMesh and Dlib.
3) Rule-based inference for engagement classification.
4) Evaluation using standard classification metrics.

The overall system architecture is illustrated in Fig. 1. The design follows common engagement analytics pipelines [8], [12], but replaces deep learning stages with deterministic rules.

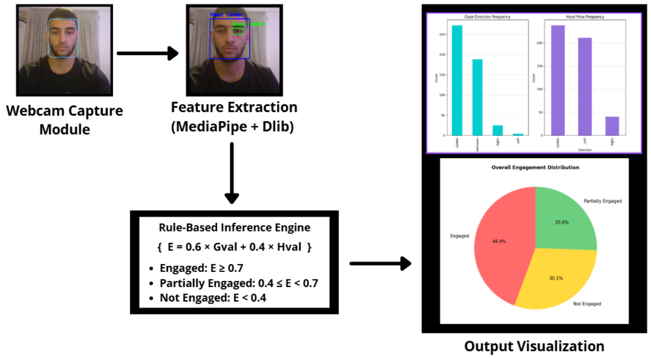

Fig. 1. System architecture of EngageSense AI. The pipeline includes frame capture, feature extraction (gaze and head pose), rule-based inference, and metric evaluation.

B. Feature Extraction

MediaPipe FaceMesh provides dense facial landmarks, including eye regions, at real-time speeds [7]. From these landmarks, the system computes gaze direction by measuring iris displacement relative to eye corners, a strategy consistent with prior gaze-estimation work [3], [10].

Head pose is estimated using Dlib's 68-point model and a PnP-based approach, following standard methods in facial behavior analysis [6], [12]. Both gaze and head pose signals are discretized into three states relative to the screen:

- Fully aligned (looking directly at the screen).
- Partially aligned (slight deviations).
- Misaligned (looking away from the screen).

C. Rule-Based Engagement Model

Rather than training a neural network, EngageSense AI combines gaze and head pose states using a weighted rule-based model. This design is inspired by rule-based affect-aware tutoring systems [18] but focuses exclusively on behavioral indicators. Gaze receives higher weight than head pose given its stronger correlation with moment-to-moment attentional focus [3], [10].

The final engagement score is mapped to three classes:

- Engaged
- Partially Engaged
- Not Engaged

This categorical mapping aligns with prior engagement detection work [11], [15].

D. Dataset and Experimental Setup

A dataset of 1,000 frames was collected from recorded online-learning scenarios. Each frame was manually annotated with one of the three engagement labels, following behavioral definitions consistent with the literature [11], [13]. The recordings were captured at 720p resolution under varying lighting and posture conditions. All experiments were carried out on CPU-only hardware, similar to realistic deployment environments in classrooms and home settings.

E. Evaluation Metrics

Classification performance is evaluated using accuracy, precision, recall, and F1-score, standard metrics in engagement and affective computing research [1], [11]. Table I summarizes the metric definitions.

TABLE I
Evaluation Metrics Definitions

Accuracy	$\frac{N_{\text{correct}}}{N_{\text{total}}}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1 Score	$2 \times \frac{PR}{P+R}$

Ethical and privacy considerations follow best practices from prior affect-aware education systems [18]: no raw video frames are stored, and only derived behavioral features are retained.

IV. Results

The proposed system achieved an overall accuracy of 82.2%, with balanced performance across the three engagement classes. This performance is competitive with more complex models while remaining fully interpretable [11], [14].

Most misclassifications occurred between Engaged and Partially Engaged, which is consistent with prior studies highlighting the subtle transitions between intermediate engagement states [1], [11]. The confusion matrix in Fig. 2 illustrates the distribution of predictions.

TABLE II
PERFORMANCE METRICS BY CLASS

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

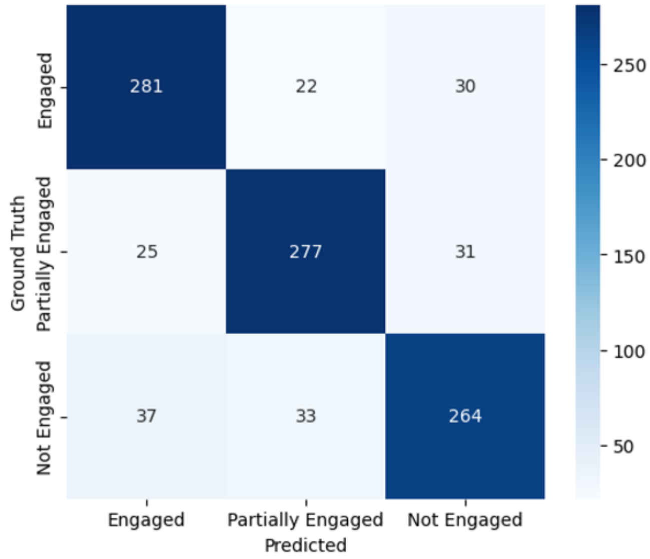

Fig. 2. Confusion matrix for the three engagement classes: Engaged, Partially Engaged, and Not Engaged.

Runtime measurements show that EngageSense AI operates at approximately 25 FPS on CPU-only hardware, meeting real-time constraints without GPU acceleration. This efficiency compares favorably to deep learning-based systems that typically require more computational resources [2], [16].

V. DISCUSSION

The experimental results indicate that behavioral cues alone—specifically head pose and eye gaze—are sufficient to achieve reliable engagement detection, confirming findings from gaze-based studies [3], [10]. The dominance of gaze in the weighted rule is supported by its strong correlation with attentional focus.

Compared with emotion-based and hybrid methods [8], [14], [15], [19], EngageSense AI offers several advantages:

- No need for large emotion-labeled datasets.
- Full transparency: every decision can be traced back to explicit rules.
- Low computational cost, enabling deployment in resource-constrained settings.

However, the current dataset is limited to a single primary subject and may lack demographic diversity, which can affect generalizability. Future work should include multiple participants and varied contexts, as done in larger engagement datasets [4], [20].

VI. CONCLUSION AND FUTURE WORK

This paper presented EngageSense AI, a lightweight rule-based system for student engagement detection using only head pose and eye gaze. Building on prior work in gaze estimation and engagement analytics [3], [10], [11], the system achieves 82.2% accuracy while operating in real time on CPU hardware, without deep learning components or large-scale emotion datasets.

Future research directions include:

- Expanding the dataset to more users and environments.
- Incorporating additional behavioral signals, such as body posture.
- Exploring hybrid models that combine interpretable rules with small, specialized neural components [1], [17].
- Integrating the system into learning management systems for longitudinal analytics and personalized feedback.

ACKNOWLEDGMENT

The authors thank Dr. Mohamad Ayash for guidance and feedback during the development of EngageSense AI.

REFERENCES

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

[2] X. Zhao and H. L. Y. Zhao, "Hybrid CNN-LSTM for Engagement Recognition," *IEEE Transactions on Multimedia*, 2022.

[3] Y. Li and J. Li, "Deep Learning for Gaze Estimation: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] C.-H. Wu et al., "CMOSE: Comprehensive Multi-Modality Online Student Engagement Dataset with High-Quality Labels," in *CVPR Workshops*, 2024.

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

[6] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018.

[7] Google Research, "MediaPipe FaceMesh Documentation," 2022. [Online]. Available: https://google.github.io/mediapipe/

[8] S. Suresh, M. Rao, and P. Kumar, "Multimodal Fusion for Student Engagement Analytics," *Computers & Education: Artificial Intelligence*, 2023.

[9] R. Pekrun, "Emotions and Learning," UNESCO Educational Practices Series, 2014.

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

[11] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. Movellan, "The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," *IEEE Transactions on Affective Computing*, 2014.

[12] P. S. Kaur and S. A. Kaur, "Prediction and Localization of Student Engagement in the Wild," in *Proc. International Conference on Multimodal Interaction*, 2018.

[13] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School Engagement: Potential of the Concept, State of the Evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004.

[14] R. Gupta, M. Sharma, and D. Chauhan, "A Multimodal Facial Cues Based Engagement Detection," *Frontiers in Computer Science*, vol. 5, 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. Dey and H. Dey, "Deep Learning for Real-Time Emotion Recognition," in *IEEE CVPR Workshops*, 2020.

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

[18] B. Woolf et al., "Affect-Aware Tutors: Recognizing and Responding to Student Affect," *International Journal of Learning Technology*, vol. 4, no. 3, 2009.

[19] A. Mollahosseini, D. Chan, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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