IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2635-2642
Journal homepage: http://iieta.org/journals/isi

A Dual-Layer, Content-Aware Framework to Validate Online Student Engagement via ML-Based Comprehension Assessment

Parinita Chate[1,2], Vishal A. Meshram[3], Kailas Patil[1*]

[1] Department of Computer Engineering, Vishwakarma University, Pune 411048, India
[2] Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering Lavale, Pune 412115, India
[3] Department of Computer Engineering, Vishwakarma Institute of Technology, Pune 411037, India

Corresponding Author Email: kailas.patil@vupune.ac.in

https://doi.org/10.18280/isi.301010

Received: 19 August 2025
Revised: 11 October 2025
Accepted: 18 October 2025
Available online: 31 October 2025

Keywords:
online learning, cognitive engagement, validation framework, dual-layer assessment, content-aware analytics, comprehension assessment, MCQ generation, machine learning

ABSTRACT

Online student engagement monitoring tools based on computer vision and artificial intelligence are increasingly used in virtual classrooms to assess attentiveness through facial orientation, gaze, and posture. However, these systems largely capture superficial visual cues and fail to validate actual cognitive engagement or learning comprehension. This paper proposes a dual-layer, content-aware framework that verifies behavioural engagement scores (E-scores) using comprehension-based validation through lecture-specific quizzes. The framework integrates Whisper for real-time transcription and T5 for automatic generation of concise, content-aligned multiple-choice questions. Experiments were conducted across ten live lectures involving one hundred undergraduate students. Continuous engagement scores (0-100) received through a commercial system of engagement monitoring were in comparison with comprehension scores (C-scores) based on post-lecture quizzes using categorical thresholds: Low (0-49), Moderate (50-74), and High (75-100). The analysis discloses a low correlation between both E-, C-scores, and the prevalence of cases of mismatch (e.g., High-E/Low-C), where the visual attention was dissenting with the real. Some of the factors include cognitive overload, off-camera activities and partial occlusion. The modular pipeline runs more efficiently on a GPU-enabled workstation with CPU fallback support. Combining the understanding-based evaluation with the behavioural analytics, the proposed system enhances the validity of engagement measurement and enables the adaptive pedagogical practices in the online learning setting.

1. INTRODUCTION

The shift to online and hybrid education models has significantly increased reliance on AI-based engagement monitoring tools in virtual classrooms. During live sessions, most engagement tools infer attentiveness from webcam-based signals—face recognition, gaze trajectories, and coarse posture cues [1-3]. Commercial platforms such as GoGuardian convert these streams into live dashboards by analyzing camera feeds alongside interaction activity [2, 3]. Yet prior work shows that such indicators describe what is visible rather than what is understood [4, 5]. Learners can look focused and still miss core ideas [6, 7], whereas others may grasp the material without displaying the expected on-screen cues (e.g., sustained screen-directed gaze) [5, 8]. Routine conditions such as uneven lighting, camera angle, and network latency also decrease measurement and may confuse both algorithms and observers [9, 10]. Collectively, these problems reduce the validity and instructional worth of scores calculated based on vision alone, which explains why multimodal research has become popular in recent times that combines audio, linguistic content, and behavioral traces to enhance reliability [11, 12]. However, the majority of the existing solutions do not provide a formalized system of checking the translation of engagement to comprehension, which is an essential educational result. Although webcams record the behavioral cues, including gaze, head position, and posture, they cannot be used to predict cognitive engagement, the mental effort to interpret and combine new information. A learner may seem to be listening, but not to be processing the information or may seem to be minimally expressive on camera, but be busy taking notes. This disparity drives a validation layer that connects visible behavior with content comprehension estimated as an immediate result of instruction.

A new dual-layer validation model is suggested in the given context that would complement the traditional engagement monitoring scheme with the machine-learning-based evaluation of lecture understanding. It uses the pre-trained models of Whisper as a live speech transcription [13], multimodal transformer as a summary [14], and T5 as a generative model of real-time multiple-choice questions [15, 16]. Students will then be given a comprehension quiz right after the lecture and the performance compared with the software results of the engagement score is calculated.

The AI/ML has had a significant impact on various industries such as agriculture [17, 18], healthcare [19], security

[20], and finance [21]. The domain-specific datasets that are of high quality have been provided by several authors [22-24], which form the basis of the models used in this paper. Continuing this momentum, our research incorporates machine learning into the sphere of online learning validation that is a scalable.

In pedagogical studies, the engagement is usually classified into behavioural, emotional, and cognitive levels. The behavioural engagements refer to the observable activities; emotional part to interest and affect; and cognitive to the extent of mental processing. Since classroom video is primarily a manifestation of behavioural clues, this paper operationalises comprehension (MCQ-based) as a performance proxy of cognitive engagement. Unlike sustained-attention heuristics or self-reports, comprehension outcomes directly indicate whether instruction was mentally understood.

The article is further divided into Section 2, which is comprised of related work; Section 3 which is comprised of the framework; Section 4, which is comprised of the experimental setup; Section 5, which is comprised of results and conducts a mismatch analysis; Section 6, which discusses implications and limitations; and finally, Section 7 provides the conclusion and future scope of this research.

2. RELATED WORK

The sphere of artificial intelligence-based surveillance of student engagement has been evolving quickly along with the popularity of online studies. The first systems mainly used visual cues (eye gaze and facial expressions and the position of a head) to determine the presence of attention during virtual meetings. Hossen and Uddin [1] used models that were based on convolutional neural networks to identify attentiveness using webcam feeds, despite the fact that they tended to have low accuracy in uncontrolled settings. There were also commercial visual engagement analytics at scale tools like GoGuardian [2] which provided attentiveness dashboards on live lectures in real-time. However, such systems mostly focus on behavioural cues at the surface level and cannot detect if learners are processing information or retaining whatever they are receiving.

The restrictions of engagement tools, which are based on the visual input only, are widely documented in existing literature. According to Ma and Shi [6], there had been a continued discrepancy between objective performance of students and how students self-rated their attentiveness in virtual classroom settings. Wang and Zheng [5] made an argument in a parallel study that visual cues are not reliable proxies of actual learning outcomes, particularly in cases where learners are looking at each other but failing to process learning material. All these observations echo the larger concerns about the legitimacy of web-based analytics, especially when such environmental factors as distractions at work, unsuitable lighting conditions, and privacy limitations undermine fidelity of data [9]. As a result, even the largest videoconferencing tools, such as Zoom, have rolled out their attention-tracking tools due to the reliability and ethical issues associated with them [4].

Multimodal interaction identification Multimodal interaction detection in order to overcome these shortcomings, recent research has sought to adopt multimodal interaction detection, combining visual, audio, and affective techniques to enhance predictive power. The evidence presented by Angeline and Nithya [12] indicates that multimodal fusion leads to increased engagement detection accuracy; however, the literature available does not include much information on the ways to determine whether the understanding or actual learning is achieved. Simultaneously, progress in natural language processing, particularly transformer-based models, has led to a variety of novel tasks, such as automated summarization and automatic question generation. Zhu et al. [14] used multimodal transformer to derive brief summaries of educational content to teach people, but Dhanya et al. [15] used the T5 model to get multiple-choice questions based on the lecture transcripts.

Despite these technological advances, a critical gap remains:

(1) Engagement monitoring systems are rarely cross-validated against actual student understanding.

(2) Most tools continue to assume that attention equates to comprehension, leading to potential misclassifications. For instance, students who appear disengaged due to camera issues or personal habits may score highly on assessments, while others who maintain screen focus may perform poorly.

This highlights the need for systems capable of verifying engagement claims through performance-based validation.

Novelty relative to prior work: Prior studies primarily (i) estimate engagement from visual behavior, (ii) combine multimodal signals (e.g., audio, context), or (iii) assess learning outcomes in isolation. To the best of our knowledge, no prior classroom study has positioned content-aligned comprehension as a systematic validation criterion for vision-based engagement dashboards in live instruction. The proposed framework operationalizes this link and reports when and why behavioral and cognitive indicators diverge. This work makes three contributions:

(1) Introduced a content-aware comprehension layer that validates commercial behavioral E-scores using immediate post-lecture C-scores.

(2) Implements a Whisper–BERTSUM–T5 workflow, optimized for GPU acceleration yet deployable on a single workstation in classroom settings.

(3) Across ten lectures involving one hundred students, demonstrates a weak E–C correlation and provides mismatch analyses showing when behavioral attention diverges from learning comprehension, thereby informing pedagogical feedback.

Table 1. Comparison of existing engagement monitoring methods

Feature	Traditional Visual Monitoring	Multimodal Systems
Type of input	Facial expressions, gaze, posture	Visual + audio + sensors (e.g., heart rate)
Real-time capability	Yes	Yes
Measures cognitive understanding	No	Partial
Scalability	High	Medium
Hardware dependency	Low	High
Assessment integration	No	No
Accuracy under diverse conditions	Low to Medium	Medium to High

Table 1 summarizes the key characteristics of two widely adopted engagement monitoring methods: Traditional Visual Monitoring and Multimodal Systems. Earlier engagement systems mainly relied on what they could see — a student's face or eye movement — to judge attentiveness. These methods are easy to scale and need little hardware, but they fail to show how much the learner actually understands and often give mixed results when lighting or background changes. Some newer models combine different types of data, such as sound or body signals, to make the predictions more stable. However, these systems need additional sensors and are difficult to apply in large online classes. More importantly, both approaches overlook what the student actually understands during the lesson, showing why a validation layer like the one proposed in this study is necessary.

3. PROPOSED METHODOLOGY

To address the boundaries of systems that decides engagement only by appearance, our method works in two connected parts. The first, called the behavioural layer, takes an Engagement Score (E-score) from an existing vision tool, which measures how attentive a student appears on camera. The second, the comprehension layer, converts the lecture's audio into short summaries and builds quick, topic-based MCQs to calculate a Comprehension Score (C-score). These two layers run together for every lecture and every student, helping us check whether visible attention truly matches actual understanding.

The framework is built for live online classes, where most engagement tools watch eye direction, facial expressions, and neck position to find an E-score. Yet these visible signs do not always show what a student actually understands. To overcome this gap, the model adds a second layer that measures C-scores, taken from short quizzes created automatically with natural language processing (NLP) techniques after each lecture. The system's dashboard shows a continuous E-score from 0 to 100. For grouped results, engagement levels are marked as i) Low (0-49) ii) Moderate (50-74), and iii) High (75-100). For simple yes-or-no analysis, students with E ≥ 75 are counted as *Engaged*, and those with E < 75 as *Disengaged*. Tests with slightly different cut-off values (± 5 points) showed that the overall patterns stayed the same.

Whisper speech-to-text model is used to covert lecture audio into text. The transcript is then shortened into a summary with BERTSUM, and then T5 model uses that summary to create five to ten multiple-choice questions focused on key ideas. Each question has the same weight, and the comprehension score is calculated as follows:

$$C-score = \frac{Correct\ Resonses}{Total\ Question} \times 100 \qquad (1)$$

Score C ≥ 70 represents satisfactory comprehension, consistent with standard mastery thresholds in classroom assessment.

Three pretrained machine learning models were used in the framework:

(1) Whisper (OpenAI) – To transcribes live lecture audio [13],

(2) BERTSUM – To summarizes transcribed content [14],

(3) T5 – To generates content aligned MCQs from the summary [15].

The quizzes are given right after each lecture, and the students' results are then compared with their E-scores to check how closely observed attention matches actual understanding.

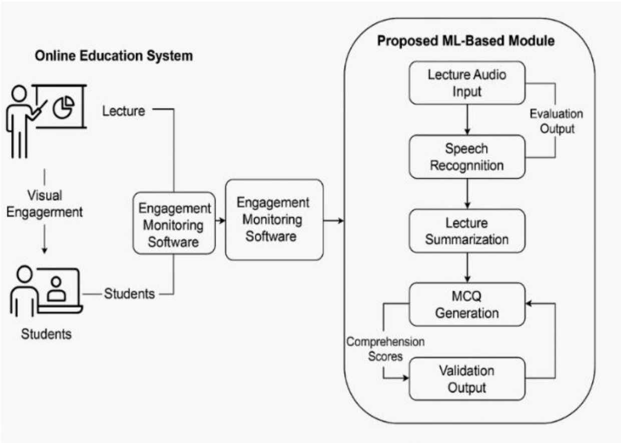

Figure 1. Overall system architecture integrating the proposed ML-based engagement validation module

Figure 1 shows how the system works as a whole, connecting the live lecture, webcam-based engagement tracking, and comprehension checking. While the class is in progress, the lecture audio passes through the ASR–NLP pipeline, and the vision module records visual engagement. The NLP component creates a short summary and a quick quiz, and the students' answers generate a C-score. This score is then compared with the related E-score to see how well the two align.

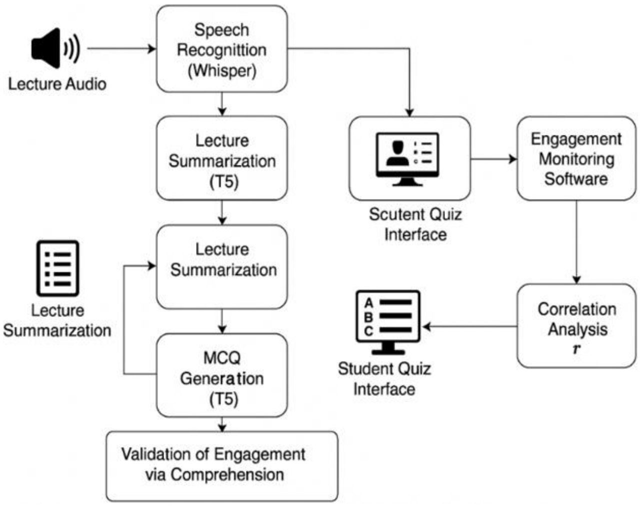

Figure 2. System architecture for engagement validation via lecture comprehension

Figure 2 enlarges the ML pipeline, showing its add-in stages: audio capture -> transcription ->summarization -> MCQ generation -> quiz delivery and scoring -> statistical comparison between E-score and C-score. After scoring, the Pearson correlation coefficient (*r*) is calculated to measure the relationship between the two scores.

This modular and platform-independent design allows whole integration with any vision-based tracker, linking its behavioral outputs with performance-based comprehension rather than presence alone.

3.1 Research contributions

The approach aims to close the gap between what students seem to pay attention to and what they truly understand. It also brings a few useful improvements, both in method and in practice, to the field of AI-supported learning. The main ideas and contributions of this work are outlined below:

(1) A dual-layer workflow pairing behavioral engagement estimates with post-lecture comprehension test [7, 8, 11].

(2) Replaces subjective surveys or manual rubrics [6] with automatically generated quizzes, providing a quantifiable reference for student understanding.

(3) The correlation module enables comparisons across commercial or academic engagement tools without modifying their internal algorithms.

(4) Inspired by ML use cases in healthcare, agriculture, finance and IoT domains [17-24], the framework is optimized for real classroom integration with minimal hardware and software dependencies.

Together, these components form a practical and scalable, tested framework that introduces accountability and cognitive depth into online learning engagement systems.

4. EXPERIMENTAL SETUP

This section explains the setup, data, and steps used to test the dual-layer engagement validation framework in real classroom sessions. The configuration was designed to replicate realistic online-learning scenarios where computer-vision-based engagement monitoring runs in parallel with automated comprehension assessment.

4.1 Experimental environment

The experiments were conducted in a controlled virtual-classroom setup integrating a standard video-conferencing platform (Google Meet) with a commercial vision-based engagement monitor (GoGuardian) [2]. Ten independent 45-minute lectures were delivered to a cohort of 100 undergraduate students from the Computer Science programme. During each live session, facial-orientation, gaze-direction, and posture data were continuously analyzed to compute real-time engagement scores (E-scores) ranging from 0 to 100.

Table 2. Reviews the components employed in the study

Component	Tool / Model Used	Purpose
Virtual classroom	Google Meet	Live lecture delivery
Engagement monitoring	GoGuardian / Engageli	Real-time visual engagement tracking
Speech recognition	Whisper (OpenAI)	Live ASR for lecture audio
Summarization & MCQ generation	T5 (fine-tuned) / BERTSUM (pilot)	Content summarization -> quiz creation
Quiz interface	Flask Web App	Secure post-lecture testing
Statistical analysis	NumPy, SciPy	Pearson *r*, *p*-values

In parallel, lecture audio was routed to the Whisper automatic speech-recognition (ASR) model [13] for live transcription. The transcripts were summarized by the T5 transformer [15], and concise multiple-choice questions (MCQs) were generated automatically and delivered immediately after the lecture through a secure Flask-based quiz interface. All models were built using Python 3.10 SDK (Software Development Kit) using the Hugging Face Transformers library and executed locally within a Docker environment on an Intel i7 workstation (32 GB RAM, RTX 3060 GPU). This setup ensured real-time operation while emulating the bandwidth and latency constraints typical of remote teaching in Table 2.

4.2 Dataset and participants

To test the dual-layer engagement validation framework, we created our own dataset that records both visual engagement signals and students' quiz responses during live online classes. The data was gathered only for research and comparison under controlled classroom conditions. It includes matched E-scores from a commercial vision-based tool and C-scores from short quizzes given right after each lecture.

A separate set of 16,000 labelled facial expression images was also created to fine-tune a local vision model, used only for testing and comparing results with the commercial E-scores. The live classroom study itself depended entirely on the commercial engagement tool and the new comprehension layer, while this extra dataset was included to help others reproduce the work and run future comparisons. Images were captured using a OnePlus 7 Profor high-resolution input, a Realme UI 5.0 or mid-range performance, and a Lenovo 300 FHD Webcam. This variety of devices represents the mixed hardware available to students, which can influence both image quality and engagement detection accuracy.

In the comprehension layer, the same group of students took automatically generated MCQ tests right after each lecture. For every student, a Comprehension Score (C-score) was worked out by dividing the number of correct answers by the total questions, which helped show how well each learner understood the topic. The study involved students aged 10 to 32 years, covering both school and undergraduate levels, so that the framework could be tested across different age groups and learning abilities. All data were collected under the institute's ethical approval process, and personal details were removed before analysis to maintain privacy.

The dataset was created with two clear purposes. One was to recognize engagement levels from visible behaviors, and the other was to see if those behaviors actually reflected real understanding. The combined record of E-scores and C-scores provided the key information for measuring how closely observed attention matched genuine comprehension, which forms the core of this framework's validation.

4.3 Result and analysis

Figure 3 describes the working of online student engagement tool. During the live session, the engagement tool continuously captured visual cues while the lecture audio was processed in parallel. The Whisper model converted the audio to text, which was then summarized by T5 and turned into a ten-question, content-based MCQ quiz. Students completed the quiz right after the lecture, and the system calculated their Comprehension Score (C-score) as the percentage of correct answers, as shown in Eq. (1).

A C-score of 70 or above was taken as a sign of good understanding, which matches the usual mastery level used in education. The summaries and quizzes created by the system were reviewed by two senior teachers to make sure the content was accurate and matched the learning goals. Their agreement was strong, with reliability values of κ = 0.82 for relevance and κ = 0.79 for difficulty, showing that the generated material was of good quality.

Student engagement was divided into three groups — Low (0–49), Moderate (50–74), and High (75–100). The E-scores and C-scores were compared using Pearson's correlation (r) to find how closely visible attention matched actual learning. The whole process ran automatically in the background, without interrupting the lecture or disturbing the students.

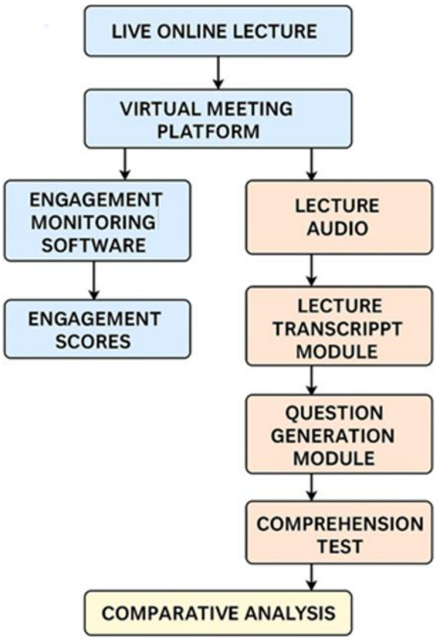

Figure 3. Experimental setup for validating online student engagement

Table 3. Sample results mapping: Comprehension vs engagement scores

Student Id	Lecture Id	E-Score	E-Level	C-Score	C-Level
S001	L1	82	High	84	High
S002	L1	76	High	78	High
S003	L1	63	Moderate	72	High
S004	L1	91	High	62	Moderate
S005	L1	58	Moderate	45	Low
S006	L2	69	Moderate	81	High
S007	L2	49	Moderate	39	Low
S008	L2	80	High	52	Moderate
S009	L2	54	Moderate	77	High
S010	L2	44	Low	48	Low

Table 3 shows sample results for ten students chosen at random. In several cases, the level of visible engagement did not match the actual understanding. For example, student S008 had an E-score of 80 (High) but a C-score of 52 (Moderate), while student S006 showed the opposite pattern with E = 69 (Moderate) and C = 81 (High). These differences make it clear that paying attention on camera does not always mean the content was fully understood.

When all records were analysed, the Pearson correlation between E-scores and C-scores was r = 0.31, showing only a weak link between the two. As seen in Figure 4, the scatter plot spreads widely with just a slight upward trend, which means that webcam-based attention does not consistently reflect real understanding. These results highlight why a content-aware and performance-based layer is important for making sense of engagement data.

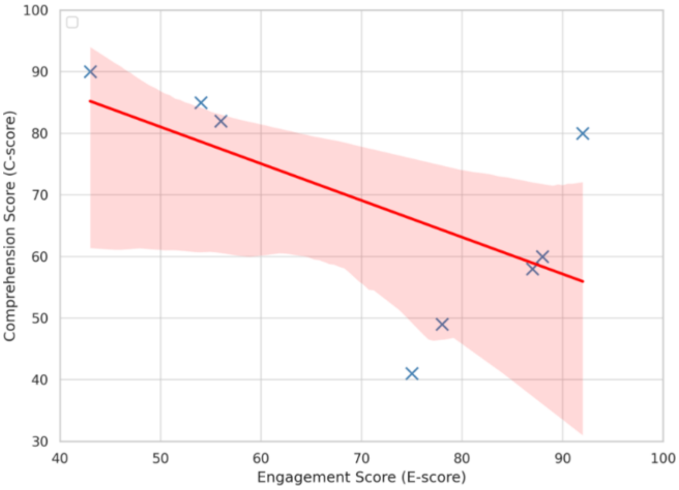

Figure 4. Correlation scatter plot – E-score vs C-score

5. PERFORMANCE EVALUATION

This section shows the numerical results comparing the E-scores from visual tracking with the C-scores produced by the proposed comprehension layer. To check how closely the two measures matched, both standard classification methods and correlation analysis were used to test the accuracy and reliability of the framework.

5.1 Evaluation metrics

To estimate the relationship between engagement and comprehension, five commonly used metrics were adopted: Accuracy, Precision, Recall, F1-Score, and the Pearson Correlation Coefficient (r). These provide both categorical and continuous perspectives on model alignment.

Accuracy (Acc):

Accuracy represents the proportion of correctly predicted engagement–comprehension alignments among all predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

where,

TP = True Positives (High E-score and High C-score)
TN = True Negatives (Low E-score and Low C-score)
FP = False Positives (High E-score but Low C-score)
FN = False Negatives (Low E-score but High C-score)

Precision (P):

Accuracy represents the proportion of correctly predicted engagement–comprehension alignments among all predictions:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

A higher precision indicates fewer false identifications of engagement.

Recall (R):

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

This measures how effectively the system detects all genuinely engaged students.

F1-Score:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + \mathrm{Re}call} \tag{5}$$

The F1-score provides a balanced evaluation when data across engagement categories are unevenly distributed.

Pearson Correlation Coefficient (r):
To quantify the linear association between E-scores and C-scores, the Pearson correlation coefficient was calculated as:

$$r = \frac{\sum\left(Ei - \bar{E}\right)\left(Ci - \bar{C}\right)}{\sqrt{\sum\left(Ei - \bar{E}\right)^2} \cdot \sqrt{\sum\left(Ci - \bar{C}\right)^2}} \tag{6}$$

where Ei and Ci represent individual engagement and comprehension scores, and $\bar{E}\ and\ \bar{C}$ denote their means.

In this study, the correlation value was r = 0.31, showing only a weak positive relationship. This means that students who appeared more attentive on camera did not always score higher in comprehension.

5.2 Quantitative results and confusion metrics

To further analyze classification reliability, a confusion matrix was produced using comprehension outcomes as the performance-based ground truth. Engagement was classified as *Engaged* ($E \geq 80$) or *Disengaged* ($E < 80$), while comprehension was considered *Actually Engaged* if $C \geq 70$.

Table 4. Confusion matrix for binary classification

	Predicted Engaged	**Predicted Disengaged**
Actually engaged	36	14
Actually disengaged	19	31

Based on this matrix, the following metrics were calculated:
Accuracy: (36 + 31) / 100 = 0.67
Precision: 36 / (36 + 19) = 0.65
Recall: 36 / (36 + 14) = 0.72
F1-Score: 2 × (0.65 × 0.72) / (0.65 + 0.72) = 0.685

The results show that the system performed moderately well, with the engagement software matching actual comprehension in about 67% of the cases. In the remaining third, it either rated students as attentive when they weren't or missed signs of genuine understanding in Table 4.

These differences highlight the restrictions of relying only on visual signals to judge engagement. To overcome this, the proposed dual-layer framework links visible behaviour with real learning outcomes through a comprehension-based check, offering a more trustworthy way to measure true engagement.

6. RESULT AND ANALYSIS

This section compares the Engagement Scores (E-scores) from visual tracking with the Comprehension Scores (C-scores) produced by the dual-layer framework. The analysis covers basic statistics, correlation results, both multi-class and binary classifications, and cases where visible attention and actual understanding do not match. All findings are based on data from 100 students who took part in ten live lectures conducted under the setup described earlier.

6.1 Comprehension–engagement mapping

Each lecture was tracked by an automatically generated post-lecture quiz using the T5 model. Continuous E-scores (0 to 100) were obtained from commercial vision-based tools such as GoGuardian and Engageli, while C-scores were derived from the students' performance in quiz.

For consistency with engagement software outputs, three categorical bands were adopted for both metrics i) Low (0 to 49), ii) Moderate (50 to 74), and iii) High (75 to 100). For binary evaluation, thresholds were defined as Engaged ($E \geq 75$) and Actually Engaged ($C \geq 70$).

Representative examples demonstrate the deviation between behavioral and cognitive indicators. Student S007 achieved a high E-score of 82 yet obtained a C-score of 49, reflecting weak comprehension despite seeming focus. Conversely, Student S004 recorded moderate engagement ($E = 60$) but a C-score of 91, indicating deep understanding despite limited visible cues. Such discrepancies validate the core hypothesis that visual attentiveness alone is an untrustworthy factor for cognitive engagement.

6.2 Correlation analysis

Across all lectures, the Pearson correlation between continuous E-scores and C-scores was r = 0.31 ($p < 0.05$), showing only a weak positive link between visual attention and actual comprehension. As seen in Figure 4, the scatter plot shows a wide spread of data points around the regression line, meaning that higher E-scores only slightly relate to higher C-scores. This pattern supports previous research suggesting that webcam-based measures mainly capture surface-level attention. In numerous cases, students who appeared highly engaged on camera performed poorly on comprehension tests, while others with moderate engagement showed strong understanding. These differences clearly highlight the gap between observed behaviour and real cognitive learning that still exists in many AI-driven monitoring systems.

6.3 Classification performance

To compare engagement and comprehension across clear categories, both E-scores and C-scores were grouped into three levels — High, Moderate, and Low. The summary of these classification results is shown in Table 5.

The findings show that the vision-based monitoring system was fairly reliable in matching engagement with comprehension results. Even so, about one-third of the cases showed mismatches, where students either appeared highly engaged but understood less, or seemed less attentive yet showed strong comprehension.

Using binary thresholds (E ≥ 75; C ≥ 70), the confusion matrix presented in Figure 5 and Table 5 quantifies this relationship. The model achieved an overall accuracy of 68.7%, with precision = 0.72, recall = 0.65, and F1 = 0.68.

Table 5. Sample results mapping: Comprehension vs engagement scores

Metric	Value
Accuracy	72.5%
Precision	71.3%
Recall	68.9%
F1-Score	70.1%
Pearson (r)	0.31

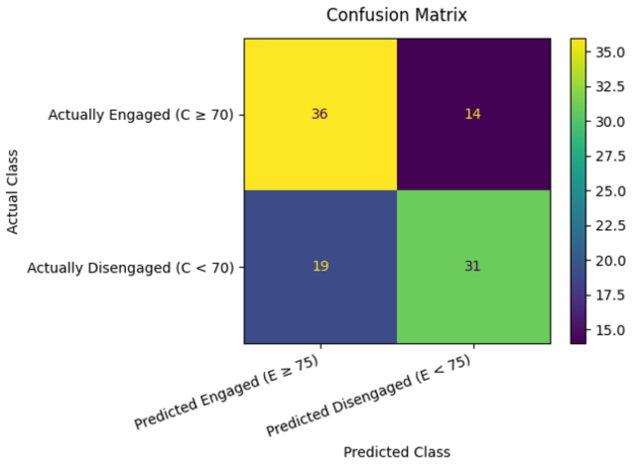

Figure 5. Confusion matrix – engagement classification vs. comprehension assessment

6.4 Discussion and findings

The weak correlation (r = 0.31) and reasonable accuracy suggest that behavioural E-scores on their own do not give a full picture of learning. A student may appear focused on camera but still fail to grasp the topic, while another who looks less attentive—or even works off camera—might understand the lesson very well.

Adding a comprehension-based check benefits cut down these wrong classifications and turns basic behavioural data into insights that teachers can actually use. This shows that the proposed dual-layer framework makes AI-based engagement analysis more meaningful and trustworthy by connecting what is seen on camera with what students truly understand.

7. CONCLUSION AND FUTURE SCOPE

This study proposed and validated a dual-layer, content-aware framework that cross-verifies vision-based behavioral engagement scores with comprehension outcomes derived from immediate, lecture-aligned assessments. Integrating Whisper for real-time transcription and T5 for summarization and quiz generation, experiments with 100 students across 10 lectures showed only a weak correlation ($r = 0.31$, $p < 0.05$) between behavioral attention (E-scores) and comprehension (C-scores). The model achieved 68.7% accuracy, confirming that nearly one-third of predictions from visual-only systems misrepresent true learning.

By linking engagement analytics with measurable comprehension, the framework offers a trustworthy and learner-centric alternative to conventional webcam-based monitoring. It provides an adaptable layer that can be embedded into existing EdTech platforms to support evidence-based feedback and adaptive instruction.

Future work will expand validation across larger and more diverse datasets, integrate multimodal inputs (gaze, speech, context), and apply explainable-AI techniques to enhance transparency. Extending comprehension analysis to open-ended responses will further improve the framework's ability to capture deeper learning outcomes.

REFERENCES

[1] Hossen, M.K., Uddin, M.S. (2023). Attention monitoring of students during online classes using XGBoost classifier. Computers and Education: Artificial Intelligence, 5: 100191. https://doi.org/10.1016/j.caeai.2023.100191

[2] Kurt, S. (2022). GoGuardian teacher: Real-time student monitoring. Educational Technology Insights. https://www.goguardian.com.

[3] Neuwirth, L.S., Jović, S., Mukherji, B.R. (2021). Reimagining higher education during and post-COVID-19: Challenges and opportunities. Journal of Adult and Continuing Education, 27(2): 141-156. https://doi.org/10.1177/1477971420947738

[4] Li, T.W., Arya, A., Jin, H. (2024). Redesigning privacy with user feedback: The case of zoom attendee attention tracking. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1-14. https://doi.org/10.1145/3613904.3642594

[5] Yang, Z. (2022). Digital transformation to advance high-quality development of higher education. Journal of Educational Technology Development and Exchange (JETDE), 15(2): 15-23. https://doi.org/10.18785/jetde.1502.02

[6] Ma, G., Shi, Y. (2025). Self-perceived vs. actual online engagement: Relationships with academic achievement among Chinese undergraduate English learners in a blended learning environment. Social Education Research, 6(1): 56-68. https://doi.org/10.37256/ser.6120255882

[7] Wakjira, A., Bhattacharya, S. (2021). Predicting student engagement in the online learning environment. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 16(6): 1-21. https://doi.org/10.4018/IJWLTT.287095

[8] Kolosov, O., Yadgar, G., Liram, M., Tamo, I., Barg, A. (2020). On fault tolerance, locality, and optimality in locally repairable codes. ACM Transactions on Storage (TOS), 16(2): 1-32. https://doi.org/10.1145/3381832

[9] Rani, S.S., Pournima, S., Aram, A., Shanmuganeethi, V., Thiruselvan, P., Rufus, N.H.A. (2024). Enhancing facial recognition accuracy in low-light conditions using convolutional neural networks. Journal of Electrical Systems, 20: 2140-2148.

[10] Xu, L. (2025). A study of human-computer interaction in evaluating students' emotional behavior and educational management under big data. Journal of Combinatorial Mathematics and Combinatorial Computing, 127: 8515-

8528. https://doi.org/10.61091/jcmcc127b-466

[11] Chen, L., Chen, P., Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8: 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

[12] Angeline, R., Nithya, A.A. (2023). A review on educational engagement recognition model based on multimodal features in online learning. AIP Conference Proceedings, 2581(1): 060002. https://doi.org/10.1063/5.0126227

[13] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning, 1182: 28492-28518. https://dl.acm.org/doi/10.5555/3618408.3619590

[14] Zhu, Y., Zhao, W., Hua, R., Wu, X. (2023). Topic-aware video summarization using multimodal transformer. Pattern Recognition, 140: 109578. https://doi.org/10.1016/j.patcog.2023.109578

[15] Dhanya, N.M., Balaji, R.K., Akash, S. (2022). AiXAM-AI assisted online MCQ generation platform using Google T5 and Sense2vec. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, pp. 38-44. https://doi.org/10.1109/ICAIS53314.2022.9743027

[16] Rodriguez-Torrealba, R., Garcia-Lopez, E., Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. Expert Systems with Applications, 208: 118258. https://doi.org/10.1016/j.eswa.2022.118258

[17] Jhajharia, K., Mathur, P. (2022). A comprehensive review on machine learning in agriculture domain. IAES International Journal of Artificial Intelligence, 11(2): 753-763. https://doi.org/10.11591/ijai.v11.i2.pp753-763

[18] Wang, W., Zhu, A., Wei, H., Yu, L. (2024). A novel method for vegetable and fruit classification based on using diffusion maps and machine learning. Current Research in Food Science, 8: 100737. https://doi.org/10.1016/j.crfs.2024.100737

[19] Bajwa, J., Munir, U., Nori, A., Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. Future Healthcare Journal, 8(2): e188-e194. https://doi.org/10.7861/fhj.2021-0095

[20] Meshram, V., Patil, K., Meshram, V., Dhumane, A. (2023). Addressing misclassification in deep learning: A Merged Net approach. Software Impacts, 17: 100525. https://doi.org/10.1016/j.simpa.2023.100525

[21] Cherukuri, B.R. (2024). AI-driven security solutions: Combating cyber threats with machine learning models. 6(5): 1-17.

[22] Cabani, A., Hammoudi, K., Benhabiles, H., Melkemi, M. (2021). MaskedFace-Net–A dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart Health, 19: 100144. https://doi.org/10.1016/j.smhl.2020.100144

[23] Meshram, V., Choudhary, C., Kale, A., Rajput, J., Dhumane, A. (2023). Dry fruit image dataset for machine learning applications. Data in Brief, 49: 109325. https://doi.org/10.1016/j.dib.2023.109325

[24] Chate, P., Patil, K. (2024). Student engagement and disengagement image dataset for educational research. Educational Administration Theory and Practice, 30(10): 2599-2604. https://doi.org/10.53555/kuey.v30i10.10736