EngageME: Exploring Neuropsychological Tests for Assessing Attention in Online Learning

Saumya Yadav[1][0000−0003−3747−961], Momin Naushad Siddiqui[2][0000−0003−1874−7789], Yash Vats[1], and Jainendra Shukla[1][0000−0002−6526−0087]

[1] HMI Lab, IIIT-Delhi, India
{saumya,yash18204,jainendra}@iiitd.ac.in
[2] Jamia Millia Islamia, Delhi, India
mominnaushad902@gmail.com

Abstract. During the pandemic, online learning has gained immense popularity. However, assessing student engagement in online settings remains challenging due to reliance on potentially biased and logistically complex self-reporting methods. This study investigates the use of neuropsychological tests, originally designed for attention assessment, to measure cognitive engagement in online learning. To conduct our analysis, we initially correlated clinical models of attention with pedagogical approaches in online learning. Subsequently, we pinpointed the three most crucial attention types in online learning—selective, sustained, and alternating attention. We used three neuropsychological assessments to evaluate attention in a cohort of 73 students contributing to the *EngageME* dataset. We manually annotated students' facial videos during neuropsychological assessments, revealing substantial agreement (Krippendorff's Alpha: 0.864) and a strong correlation (Spearman's Rank Correlation: 0.673) with neuropsychological test scores. This confirms the convergent validity of our approach in measuring attention during online learning. We further propose Nuanced Attention Labeling using neuropsychological test scores-based models in online learning attention assessment, enhancing sensitivity to nuanced cognitive engagement. To assess the reliability of our approach to online learning, we performed a user study in online settings. This work implies the potential for a more accurate and nuanced assessment of students' cognitive engagement in online learning, contributing to the refinement of personalized and effective educational interventions.

Keywords: Affective Computing · Education Technology · Cognitive Engagement · Interactive learning environments

1 Introduction

In the context of the recent and unprecedented COVID-19 pandemic, characterized by stringent social isolation measures, a profound transformation has occurred in the education system [13]. This transformation from traditional

classrooms to predominantly online learning relies on cutting-edge technologies, specifically webcams and microphones. The adoption of this technology-driven approach offers advantages such as cost-effectiveness and improved accessibility, contributing to societal sustainability. However, these approaches create a challenge for instructors to assess student engagement during online classes [13]. Both instructors and students significantly benefit from accurate and timely information on student engagement, impacting learning and retention [13, 4]. Ongoing studies explore modalities like audio and video to quantify student engagement in online learning [14], with webcams emerging as a prevalent tool due to their ubiquity and ability to capture essential visual cues.

Previous research exhibits inconsistencies in examining the relationship between engagement and attention [12, 30]. Engagement in education is a multifaceted concept involving emotional, behavioral, and cognitive dimensions [4]. Here, emotional pertains to effective emotional responses, behavioral involves observable actions like asking questions, and cognitive involves actively seeking to understand new information as a psychological process that requires attention and investment [4, 18]. Consequently, attention emerges as a fundamental element of cognitive engagement, encompassing various cognitive states, processes, and abilities. Student's cognitive engagement is considered more reliable learning than behavioral or emotional engagement [25]. Hence, we try to asses cognitive engagement in the proposed work and take attention as a latent variable for student engagement in the proposed work. It is distinguished mainly by its internal nature, which differs from the externally observable aspects of emotional and behavioral activity. Theorists recommend two measurement methods for measuring all facets of engagement: by observing external factors like body language and posture and by assessing internal factors related to the learner's cognitive and affective functioning [23].

Our approach to quantifying cognitive engagement draws upon a synthesis of neuroanatomical theories, factor analysis of psychometric assessments, cognitive processing frameworks, and clinically derived models, incorporating internal factors to assess attention [28]. This clinical-based model consists of five components for attention: alternating attention enables seamless shifts in focus, divided attention adeptly responds to multiple tasks, focused attention responds to sensory stimuli, selective attention manages distractions, and sustained attention maintains focus during repetitive activities, revealing enduring aspects over time. This model can be used to map students' attentional dynamics, summarized in Table 1, with each facet playing a unique role in the educational context. We excluded focused attention, as our focus lay on prolonged attention span [1]. Further previous research also showed that divided attention diminishes focus on a single task when multiple focuses occur simultaneously [11, 7]. Thus, our experiment finally assessed three key attention facets—alternating, sustained, and selective—essential for evaluating cognitive engagement in our research context.

Researchers have identified facial features as robust indicators of cognitive engagement [13, 34, 31, 23, 21, 6]. Hence, we used external factors, which include facial landmarks, head poses, Facial Action Units (FAUs), gaze tracking, and

Table 1: Assessing Cognitive Engagement in Education

Attention Type	Pedagogical Implications	Attention Assessment Tool Example
Alternating Attention [20]	Enhances adaptability in shifting focus between tasks, subjects, and classroom activities, optimizing versatility in educational contexts.	Trail Making Test.
Divided Attention [11]	Handling multiple tasks simultaneously, like taking notes and participating in an online discussion	Dual-Task Paradigm.
Focused Attention [15]	Enhances response to specific sensory stimuli and understanding of educational material.	Attention Network Test.
Selective Attention [29]	Vital for concentrating on specific educational tasks, filtering distractions, and absorbing relevant information.	Stroop Test.
Sustained Attention [27]	Essential for long lectures, extended study sessions, and completing assignments efficiently.	Continuous Performance Test.

facial feature extractions for attention assessment. We acknowledge that screen-based interactions in online learning are intricate [13]. However, it's crucial to note that standardized neuropsychological tests are crafted to assess various facets of attention despite their seemingly elementary nature. Exploring these tests can deepen our understanding of the cognitive processes involved in students' attention and performance in virtual learning environments [18].

Previous literature heavily relies on self-reporting methods to establish its ground truth [23, 8, 21]. However, this practice can introduce potential distractions and may not provide presenters with truly objective information. Moreover, self-reporting methods and manual annotations tend to yield discrete annotations, which fail to capture nuanced attention during online learning [22, 8]. In light of these challenges, this paper introduces *EngageME*, an innovative dataset utilizing neuropsychological test scores to establish a consistent attention label. Additionally, we propose the *Nuanced Attention Labeling method*, finely assigning attention weights to capture subtle variations. We aim to build a foundational understanding of attention using standardized tests that tap into fundamental cognitive mechanisms associated with engagement. The approach integrates unobtrusive behavioral cues from a webcam, offering a holistic insight into cognitive engagement. To address these gaps, we outline our key contributions as:

- We introduce the EngageME dataset, which represents a pioneering in-the-wild data collection from 73 students undergoing neuropsychological tests online, alongside their attention scores.
- We posit that neuropsychological tests will detect attention during the performance of cognitive engagement tasks in an online learning environment.
- We propose the Nuanced Attention Labeling method, an effective approach for assessing attention by capturing subtle changes in cognitive engagement.

2 Related Work

Engagement is a complex and multifaceted phenomenon that is important in learning and academic success. It is often confused with attention, despite their distinct roles in cognitive functioning [12]. Some studies interchangeably use attention and engagement, limiting the understanding of these phenomena [30,

31]. Equating sustained attention with engagement and relying solely on self-reported methods, as seen in previous research [23, 34], fails to capture the full spectrum of attention and may inaccurately reflect performance. Some studies focus on predicting disengagement using specific facets [32] or analyzing behaviors like mouse movements and keyboard keystroke behaviors [3], but a precise assessment requires considering various engagement and attention facets, not just single attention or engagement.

Previous studies commonly rely on self-evaluation [8, 21, 14, 34] to establish ground-truth labels for attention, but this method introduces vulnerability to reporter biases [13]. Learners' self-assessment of engagement may lead to inaccuracies [9], potentially influenced by their capacity for accurate self-assessment or a desire to avoid repercussions [10]. Additionally, self-reporting or manual annotation methods yield discrete annotations [22, 8]. which can be unreliable for attention assessment. Nuanced annotation is essential to capture subtle variations in attention levels, ensuring unbiased evaluation.

Similarly, previous studies have recognized the benefits of using rating scales and checklists throughout the annotating process to reduce self-report biases. However, these methods are not without their challenges. This annotation method needs a lot of effort and time [16, 33, 6]. For instance, accurately tracking each learner's engagement in large classrooms or online settings becomes a major challenge for teachers. Learners may feign engagement but are not genuinely involved in the assigned activities. Despite their potential biases, effort requirements, and subjectivity, these annotation methods are unreliable. Real-time annotation measures may also impact the state of attention of the instructor and students while actively participating in a class, further compromising their dependability.

In conclusion, prior research has exhibited a certain degree of inconsistency when probing the relationship between engagement and attention, thereby casting doubts on the validity of the established ground truth used for assessing attention. Additionally, prevalent literature tends to rely heavily on explicit methods for collecting ground truth, a practice that not only introduces potential distractions but also escalates the cognitive workload for users. Moreover, these methods may fall short of furnishing the nuanced annotation that attention requires. To address these challenges, we propose standardized neuropsychological assessments administered by trained professionals. These tests, covering alternate, selective, and sustained attention, provide a more objective measure than conventional self-reporting, enhancing our understanding of attention dynamics without relying on subjective reports. Above all, currently, there is no dataset offering ground-truth labels for different attention components.

3 Hypothesis Development

To address the proposed research objectives, we conducted research under the following hypothesis:

- **H1:** Neuropsychological tests will detect attention while performing cognitive engagement tasks that occur in an online learning environment.

– **H2:** Nuanced Attention Labeling in online learning attention assessment will result in heightened sensitivity to intricate cognitive engagement.

For our first hypothesis, we aim to establish a baseline understanding of attention through these standardized tests, capturing core cognitive mechanisms underlying cognitive engagement, as included in Table 1. While neuropsychological tests traditionally assess attention within clinically derived models, they aren't commonly used for evaluating attention during cognitive engagement tasks in online learning environments. To validate this hypothesis, we establish the convergent validity of our new annotation method. We also performed a user study to validate the assessment for the online learning environment. In our second hypothesis, we suggest that using Nuanced Attention Labeling, as opposed to traditional discrete labeling, will offer instructors a more detailed insight into attention granularity. This shift is expected to enhance accuracy in representing cognitive engagement. The finer granularity of continuous labels is anticipated to boost model performance by reducing Mean Squared Error (MSE) and capturing subtle variations in attention dynamics.

4 Methodology

The EngageME dataset was meticulously gathered from willing participants following the receipt of ethical clearance from the Institutional Review Board (IRB). Participants engaged in an experimental procedure detailed in Section 4.1, submitting webcam recordings and cognitive task responses in CSV format to a web server. Both video recordings and CSV data were segmented into approximately 60-second intervals inspired by [17], aligning with timestamps in the CSV file to prevent data loss during splitting. Participant performance was annotated using CSV data, and features were extracted from each video segment.

4.1 Data Collection

73 participants (26 females, 47 males), aged 19 to 21 years (M = 20.644, SD = 2.628), completed our study. All were undergraduate students invited to participate in an online data collection experiment from the comfort of their homes without any compensation. Informed consent was obtained, and participants were guided about the test, including webcam calibration, without the need for specialized software or hardware installations closely resembling their typical study environment. Participants were unaware of the attention level assessment but were briefed about the tests and informed about the recording. The online data collection framework, outlined in Figure 1, utilized a javascript library and lasted approximately 20 minutes for each participant. The experiment included eye-tracker calibration and three cognitive tasks targeting specific attention types. Importantly, upon acceptance, we commit to releasing the dataset at the researcher's request for transparency and contribution to the research community.

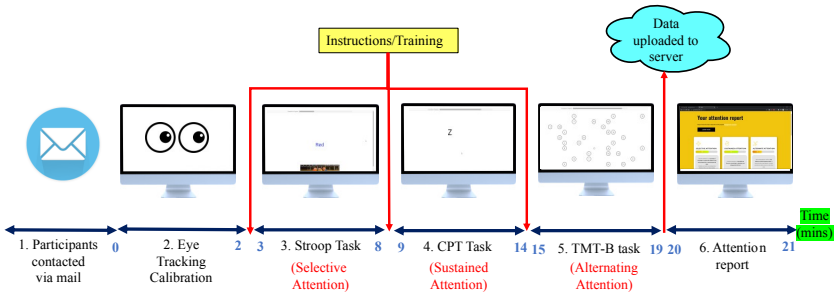


Fig. 1: Timeline of the EngageMe Data Collection Experiment.

In our experiment, we used the webcam-based eye-tracking system [24], performing a crucial calibration to ensure accuracy exceeding 60%. The decision to utilize this JavaScript library is a widely accepted and cost-effective approach, aligning with practical considerations for potential classroom use. The Stroop Test, a widely used measure of selective attention [29], evaluates individuals' accuracy in verbally articulating the font color of a word representing an incongruent color. In our study, 'congruent stimuli' denote matching text color and color name (e.g., "red" written in red), while 'incongruent stimuli' involve a mismatch (e.g., "red" written in blue ink). Conducted through a computerized version, participants used arrow keys with a fixed 4000-millisecond stimulus presentation time and a variable post-trial gap (100-500 milliseconds). Training trials mirrored experimental test blocks, with feedback for errors and correct responses to aid learning. The test was structured to align with reaction coordinates obtained from Webgazer in post-processing.

N-back Continuous Performance Test (CPT), a widely utilized tool for assessing sustained attention, was used in our experiment [27]. It requires participants to respond to a series of stimuli by determining whether each stimulus matched the one presented N-items earlier. In our experiment, participants pressed "M" if the presented letter matched the one shown two items earlier. Participants were instructed that omission occurs when failing to press the designated key for a matching stimulus, and commission happens with an incorrect key press. These measures facilitate an assessment of sustained attention, capturing instances of inattention (omission) and response errors (commission) during the N-back CPT.

The Trail Making Test (TMT), a widely used neuropsychological tool, assesses alternating attention [20]. It comprises two versions: TMT-A focuses on number sequencing (1 to 15), while TMT-B assesses set-shifting abilities by alternating between numerical and alphabetic sequences (1–A–2–B–3...). In our study, we exclusively used TMT-B to analyze participants' alternating attention. They selected alphanumeric sequences, which changed color to green upon selection. Three TMT-B trials were conducted.

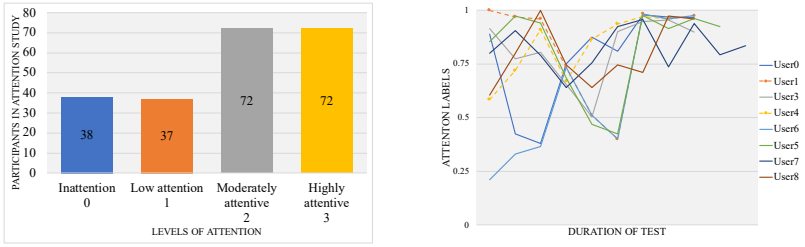
4.2 Label generation from generated CSVs

After collecting videos and CSV data from 98 participants, we addressed issues such as unclear videos, cameras not capturing faces, and excessive video duration. Subsequently, 73 participants' data was selected for further processing. Before task-specific analysis, distributions for reaction times, accuracy, and total time across the three tests were obtained. The Stroop test revealed a significant difference in reaction times between congruent and incongruent stimuli ($t(2,73) = 3.058$, $p < 0.05$), with average times of 892.952 ms and 937.497 ms, respectively. Scoring was based on response accuracy, with equal weight assigned to congruent and incongruent Stroop test types. Accuracy scores, calculated as a percentage of correct responses and normalized on a 0 to 1 scale, indicate attention levels (1 for highest, 0 for lowest). The average congruent accuracy surpassed incongruent accuracy by 1.97%.

In CPT, a weak negative correlation was found between the number of false alarms and median reaction time (Pearson's correlation of -0.196, $p < 0.05$). Only accuracy was considered for both omission and commission tests. Each CPT test type is given 50% weightage. The accuracy scores were scaled to generate a final CPT score ranging from 0 to 1, with 0 indicating least attentive and one indicating highly attentive. Additionally, the average accuracy for non-match stimuli was 4.883% higher than that for matching stimuli.

In the TMT test, response times decline as competing stimuli decrease. Specifically, average response times for three consecutive trials were 3148 ms, 2580 ms, and 3110 ms, with the first half having a higher mean than the second. We standardized labels using min-max normalization and inverted attention scores by subtracting them from 1. The resulting nuanced attention label scores for TMT are scaled between 0-1, indicating performance levels. A higher score signifies superior performance, while a lower score implies the opposite.

Manual Annotation: We adopted a comparative approach to support our first hypothesis by establishing convergent validity [19] for our new attention test. Convergent validity assesses the correlation between different measures theoretically expected to be related. We collected scores from neuropsychological tests and manual annotation scores generated by three annotators who manually reviewed recorded participant clips. This approach demonstrates the degree of agreement between neuropsychological test scores and manual annotations, providing compelling evidence for the convergent validity of our attention test. Annotators labeled video clips on a scale of 0 to 3, inspired by [34], to indicate the level of cognitive engagement. A rating of 3 signifies high attentiveness, where participants continuously look at the screen, while a rating of 2 indicates moderate attentiveness, including occasional glances away. A rating of 1 suggests low attentiveness, characterized by gaze shifts, yawning, fidgeting, and frequent body movements. A rating of 0 denotes inattention and encompasses behaviours like frequent gaze shifts, resting the head on the table, or skipping the test. Clear instructions and examples were provided to distinguish between all the attentions and ensure consistency across datasets. For TMT clip annotation, the video's duration was considered.



(a) (b)

Fig. 2: Distribution of attention labels based on neuropsychological test scores (a) and Visualization of attention during a complete test of random users (b).

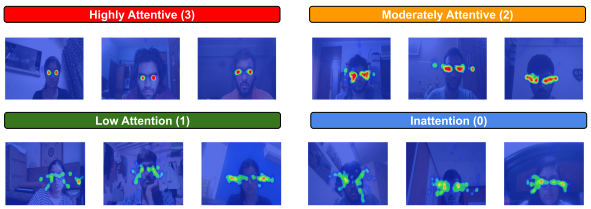


Fig. 3: Heatmap for different types of attention labels.

We categorized the continuous spectrum into four classes for the neuropsychological test by segmenting the distribution within 0.5 standard deviations. This discretization, tailored for left-skewed data, ensures a balanced representation of attention levels. Labels were assigned based on performance scores, streamlining the creation of a multi-labeled classification model for attention. Fig 2a categorizes participants' attention levels during neuropsychological testing. A significant number fall into 'Moderately attentive' and 'Highly attentive' categories, highlighting prevalent attention. The variability in participants labeled 'Inattention' (38) and 'Low attention' (37) provides insights into attention levels, showcasing the distribution across different categories during neuropsychological tests. Fig 2b illustrates attention scores for eight random users (User0 to User7) at distinct time points, capturing individual attentional dynamics. Variability is observed in completion times, with User6 taking longer while User3 finishes first. User4 begins with the lowest attention score, and User1 achieves the highest (100%). Temporal fluctuations in user attention reveal dynamic patterns—some consistent, others varied. This variation forms a foundation for nuanced insights valuable to educators.

We used the OpenFace framework [5] for extracting facial behavior. It generates 712 features per frame, including gaze, pose, FAUs, landmarks in 2D/3D, and shape parameters. Fig 3 displays a heatmap of eye-gaze locations for different attention labels, emphasizing distinctions between highly attentive and less attentive students during the test.

4.3 Regression Approach

Sequential data from the EngageME consists of approximately 1800 rows for each clip extracted from OpenFace. We used a Bidirectional-Long short-term memory (Bi-LSTM) [6] model to generate embeddings to capture temporal dependencies and behavioral patterns, aligning with our goal of analyzing attention fluctuations. Using a basic Bi-LSTM encoder with 355 units and a Dense layer with 200 units, we produced embeddings from the sequential data, normalized through Min-Max scaling ranging from 0 to 1. The resulting $200 \times N$ embeddings, where N is the total clip count, served as input for five established Machine Learning (ML) models: Support Vector Machine for regression (SVR) [13, 6], Light Gradient-Boosting Machine (LGBM) [2], K-Nearest Neighbors (KNN) [6], eXtreme Gradient Boosting (XGBoost) [2], and Random Forest (RF)[13, 6].

We used these regression models as they are recognized for their effectiveness in similar research. SVR was utilized for its effectiveness in capturing complex relationships. LGBM provides high performance and efficiency. KNN offers a versatile approach based on proximity to neighbors. XGBOOST demonstrated powerful ensemble learning capabilities, while RF uses decision tree ensembles to enhance accuracy and reduce overfitting. Each algorithm assessed attention dynamics, leveraging its strengths for the online learning context. We used the MSE metric for attention labels between 0 and 1. MSE quantifies squared differences between predicted and actual attention scores, ensuring accurate evaluation within the 0 to 1 scale.

5 Results and Discussion

In this section, we present the outcomes of two experiments. The first experiment validates our novel attention annotation method using neuropsychological test scores by correlating it with manual annotations. We achieved high inter-annotator reliability, with a Krippendorff's Alpha (α) value of 0.869, indicating substantial agreement between annotators. Additionally, Spearman's correlation coefficient demonstrated strong correlations (82-93%) across three independent annotators, affirming the reliability of our annotation approach. This assessment gauges the strength and direction of the relationship between annotators' ratings, highlighting strong agreement and high correlation in their annotations. To establish the convergent validity of our new annotation method, we computed Spearman's Rank correlation between discretized neuropsychological test scores and manual annotation, yielding a correlation value of 0.673. This significant correlation confirms the reliability of neuropsychological tests for assessing students' attention in online learning settings. The outcomes of our experiments contribute significantly to Artificial Intelligence (AI) in education, specifically addressing challenges in assessing student attention in online learning.

In our second experiment, we aimed to validate our second hypothesis by demonstrating the efficacy of Nuanced Attention Labeling in understanding attention dynamics. Our aim was to highlight how fine-grained labeling enables a detailed examination of cognitive engagement, anticipating reduced errors in

Table 2: Model performance using different annotation methods

Models	Discretized Attention Labelling	Nuanced Attention Labelling
XGBoost	0.0558	0.0131
SVR	0.0469	0.0111
LGBM	0.0493	0.0138
KNN	0.0601	0.0166
RF	0.05018	0.0143

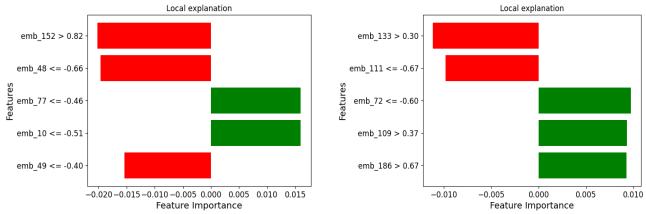


Fig. 4: LIME explanation using (a) Discretized Attention Labelling (b) Nuanced Attention Labelling

capturing subtle attention variations. To ensure model robustness, we divided the EngageME dataset (80:20 ratio) at the user level for training and testing—58 users for training and 15 for testing. We then created regression models using training data and labeled them with neuropsychological test scores for Nuanced Attention Labeling alongside discretized annotations from manual annotation. The discretized annotations averaged across annotators and normalized to a 0 to 1 scale, the same as we did with neuropsychological test scores. Additionally, clips where annotator labels differed by more than two values were ignored. The results of these regression models, assessed using five-fold cross-validation on unseen test data, are presented in Table 2. It shows that the Nuanced Attention Labeling method outperforms the discretized attention labeling method across multiple models on the same test data, as reflected in lower MSE values. This indicates superior accuracy in capturing subtle variations in attention, highlighting the method's effectiveness in detailed and accurate assessment of cognitive engagement. Its finer granularity scale reduces errors, providing a more nuanced representation of attention dynamics in the given context.

Further, we used Local Interpretable Model-agnostic Explanations (LIME) [26] to enhance the interpretability of our ML models, especially in attention labeling. LIME explanations elucidate how embeddings influence predictions, providing insights into key features. This enhances transparency in model behavior. Fig. 4 displays LIME results for one random instance using Discretized and Nuanced Attention Labeling from the SVR model, where the y-axis represents the top-5 extracted features for that instance. In the discrete label approach, binary conditions negatively influence the prediction, like embeddings emb_152 surpassing 0.82 or emb_48 falling below -0.66. These rigid cut-offs suggest a reliance on specific threshold values for decision-making. Conversely, the nuanced label

Table 3: Model performance on Online Learning Test Data

Models	EngagME Data	Online Learning Data
XGBoost	0.0727	0.0705
SVR	0.0637	0.0630
LGBM	0.0679	0.0665
KNN	0.0758	0.0857
RF	0.0578	0.0597

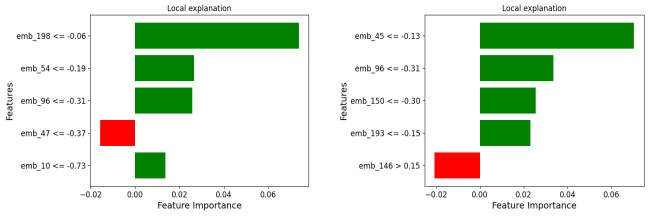


Fig. 5: LIME explanation on (a) EngagME Data (b) Online Learning Data

method reveals a more nuanced understanding, with embeddings like emb_133 exceeding 0.30 or emb_72 being less than or equal to -0.60, contributing positively or negatively. These LIME interpretations highlight differences in interpretability for attention assessment methods. The discrete approach, similar to strict grading, uses binary conditions. In contrast, the nuanced method considers a continuous impact, offering a more subtle understanding of attention dynamics.

6 User Study

Validating neuropsychological test scores for online learning attention, we now aim to confirm Hypothesis 1 using student data from online learning. For this, we used an additional dataset, *Online Learning dataset*, which includes 21 video datasets featuring individual participants taking online classes. Here, participants voluntarily chose subjects like "Linear Algebra" and "Introduction to ML" to watch during data collection. The video recordings capture students' interactions with course materials. Out of 21 videos, we extracted 651 clips (around 60 seconds each, excluding segments less than 30 seconds). However, 25 inadvertently truncated clips were omitted from the dataset. We applied the same rigorous feature extraction and annotation method as we used for the EngageME dataset for discretized attention labeling. The final dataset comprises 606 clips with annotations labeled on a scale from 0 to 1. We then split the Online Learning data in 80:20 ratio, where 16 users were designated for training and 5 for testing. We used the same models above with the discretized labels and generated the new models using Online Learning training data. We tested these models on unseen Online Learning data, comparing their outputs in Table 3.

The MSE outcomes for the EngageME and Online Learning models are nearly identical, highlighting the consistency and reliability of our attention assessment

methodology. LIME explanations in Fig. 5 spotlight facial features influencing attention predictions. We utilized the same pipeline to generate embeddings from both datasets, capturing temporal dependencies. Notably, among the top 5 features of both models, emb_96 consistently emerges as a crucial indicator. The significance of emb_96 is evident in both the EngageME and Online Learning models, as supported by their respective feature importance scores. Additionally, the proximity of other top features in both models, such as emb_45, emb_150, emb_193, emb_198, and emb_54, underlines the robustness and consistency of our extraction methods, as these features exhibit proximity to each other. These features collectively contribute to the model's proficiency in understanding nuanced temporal variations in facial behavior. The aligned patterns in emb_198 (EngageME) and emb_193 (Online Learning) further reinforce the model's ability to predict attention dynamics precisely and establish meaningful connections with specific facial features. By using Nuanced Attention Labeling in online learning, our approach offers granular insights into student attention, enabling tailored interventions and adaptive educational experiences, addressing concerns about real-world applicability, and contributing novel perspectives to artificial intelligence in education.

7 Conclusion and Future Work

The study proposes using standardized neuropsychological tests to assess students' attention in online learning. This involved a cohort of 73 students undergoing three relevant neuropsychological assessments. Our findings revealed a notable inter-annotator reliability and convergent validity between manual annotations and neuropsychological test scores. By introducing the Nuanced Attention Labeling method, the assessment of attention significantly advances with heightened sensitivity. It captures subtle variations and intricate patterns in attention dynamics and offers a more detailed and precise assessment than traditional methods. The efficacy of the Nuanced Attention Labeling method was validated through a user study conducted in online settings. The study verified the precision of Nuanced Attention Labeling in assessing cognitive engagement subtleties. In future work, we aim to refine our approach by integrating a deep learning model for enhanced analysis, addressing concerns about state-of-the-art methodologies. Additionally, we will explore multiple explainability models or acknowledge the limitations of single explainers to ensure robust interpretation. Moreover, we plan to enhance the nuanced attention labeling scheme to detect various types of attention, advancing the novelty and relevance of our contribution.

Acknowledgements. This research work is funded by a research grant (Ref. ID.: IHUB Anubhuti/Project Grant/03) of IHUB Anubhuti-IIITD Foundation and is partly supported by the Infosys Center for AI and the Center for Design and New Media (A TCS Foundation Initiative supported by Tata Consultancy Services) at IIIT-Delhi, India. We are grateful to all our participants and to Harshit Chauhan for his contribution to designing the data collection pipeline.

References

1. Abdelrahman, Y., Khan, A.A., Newn, J., Velloso, E., Safwat, S.A., Bailey, J., Bulling, A., Vetere, F., Schmidt, A.: Classifying attention types with thermal imaging and eye tracking. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**(3), 1–27 (2019)
2. Alruwais, N., Zakariah, M.: Student-engagement detection in classroom using machine learning algorithm. Electronics **12**(3), 731 (2023)
3. Altuwairqi, K., Jarraya, S.K., Allinjawi, A., Hammami, M.: Student behavior analysis to measure engagement levels in online learning environments. Signal, image and video processing **15**(7), 1387–1395 (2021)
4. Alyuz, N., Aslan, S., D'Mello, S.K., Nachman, L., Esme, A.A.: Annotating student engagement across grades 1–12: associations with demographics and expressivity. In: International Conference on Artificial Intelligence in Education. pp. 42–51. Springer (2021)
5. Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al.: Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science **6**(2), 20 (2016)
6. Booth, B.M., Ali, A.M., Narayanan, S.S., Bennett, I., Farag, A.A.: Toward active and unobtrusive engagement assessment of distance learners. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 470–476. IEEE (2017)
7. Cherry, E.C.: Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America **25**(5), 975–979 (1953)
8. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 653–656 (2018)
9. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. Learning and Instruction **29**, 153–170 (2014)
10. Fuller, K.A., Karunaratne, N.S., Naidu, S., Exintaris, B., Short, J.L., Wolcott, M.D., Singleton, S., White, P.J.: Development of a self-report instrument for measuring in-class student engagement reveals that pretending to engage is a significant unrecognized problem. PloS one **13**(10), e0205828 (2018)
11. Glass, A.L., Kang, M.: Dividing attention in the classroom reduces exam performance. Educational Psychology **39**(3), 395–408 (2019)
12. Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., Kasneci, E., Trautwein, U.: Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. Educational Psychology Review **33**, 27–49 (2021)
13. Gorgun, G., Yildirim-Erbasli, S.N., Epp, C.D.: Predicting cognitive engagement in online course discussion forums. International Educational Data Mining Society
14. Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., Alt, F.: Engagemeter: A system for implicit audience engagement sensing using electroencephalography. In: Proceedings of the 2017 Chi conference on human factors in computing systems. pp. 5114–5119 (2017)
15. Herpich, F., Guarese, R.L., Cassola, A.T., Tarouco, L.M.: Mobile augmented reality impact in student engagement: An analysis of the focused attention dimension. In: 2018 International Conference on Computational Science and Computational Intelligence (CSCI). pp. 562–567. IEEE (2018)

16. Khan, S.S., Abedi, A., Colella, T.: Inconsistencies in the definition and annotation of student engagement in virtual learning datasets: A critical review. arXiv preprint arXiv:2208.04548 (2022)
17. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing **3**(1), 18–31 (2011)
18. Lackmann, S., Léger, P.M., Charland, P., Aubé, C., Talbot, J.: The influence of video format on engagement and performance in online learning. Brain Sciences **11**(2), 128 (2021)
19. Lekwa, A.J., Reddy, L.A., Shernoff, E.S.: Measuring teacher practices and student academic engagement: A convergent validity study. School Psychology **34**(1), 109 (2019)
20. Linari, I., Juantorena, G.E., Ibáñez, A., Petroni, A., Kamienkowski, J.E.: Unveiling trail making test: Visual and manual trajectories indexing multiple executive processes. Scientific Reports **12**(1), 14265 (2022)
21. Linson, A., Xu, Y., English, A.R., Fisher, R.B.: Identifying student struggle by analyzing facial movement during asynchronous video lecture viewing: Towards an automated tool to support instructors. In: Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I. pp. 53–65. Springer (2022)
22. Ma, J., Jiang, X., Xu, S., Qin, X.: Hierarchical temporal multi-instance learning for video-based student learning engagement assessment. In: IJCAI. pp. 2782–2789
23. Monkaresi, H., Bosch, N., Calvo, R.A., D'Mello, S.K.: Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing **8**(1), 15–28 (2016)
24. Papoutsaki, A.: Scalable webcam eye tracking by learning from user interactions. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. pp. 219–222 (2015)
25. Pickering, J.D.: Cognitive engagement: A more reliable proxy for learning? Medical Science Educator **27**(4), 821–823 (2017)
26. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
27. Roebuck, H., Freigang, C., Barry, J.G.: Continuous performance tasks: Not just about sustaining attention. Journal of speech, language, and hearing research **59**(3), 501–510 (2016)
28. Sohlberg, M.M., Mateer, C.A.: Effectiveness of an attention-training program. Journal of clinical and experimental neuropsychology **9**(2), 117–130 (1987)
29. Stevens, C., Bavelier, D.: The role of selective attention on academic foundations: A cognitive neuroscience perspective. Developmental cognitive neuroscience **2**
30. Szafir, D., Mutlu, B.: Pay attention! designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 11–20 (2012)
31. Thomas, C., Jayagopi, D.B.: Predicting student engagement in classrooms using facial behavioral cues. In: Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education. pp. 33–40 (2017)
32. Verma, M., Nakashima, Y., Takemura, N., Nagahara, H.: Multi-label disengagement and behavior prediction in online learning. In: International Conference on Artificial Intelligence in Education. pp. 633–639. Springer (2022)

33. Wang, X., Wen, M., Rosé, C.P.: Towards triggering higher-order thinking behaviors in moocs. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 398–407 (2016)
34. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing **5**(1), 86–98 (2014)