2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00055

ExCEDA: Unlocking Attention Paradigms in Extended Duration E-classrooms by leveraging Attention-mechanism models

1st Avinash Anand, *IIIT, New Delhi*
2nd Avni Mittal, *IIT, Mandi*
3rd Laavanaya Dhawan, *NSUT, New Delhi*
4th Juhi Krishnamurthy, *Adobe, Noida*
5th Mahisha Ramesh, *IIIT, New Delhi*
6th Naman Lal, *IIIT, New Delhi*
7th Astha Verma, *IIIT, New Delhi*
8th Pijush Bhuyan, *IIIT, New Delhi*
9th Himani, *IIIT, New Delhi*
10th Rajiv Ratn Shah, *IIIT, New Delhi*
11th Roger Zimmermann, *NUS, Singapore*
12th Shin'ichi Satoh, *NII, Tokyo*

***Abstract*—Learner engagement wields considerable influence over educational outcomes. However, quantifying engagement within online learning contexts presents a persistent challenge, often exacerbated by the limitations of existing datasets. Existing datasets, with their focus on short-duration videos, hinder the comprehensive assessment of student attention over the extended periods characteristic of typical classroom settings. To address this research gap, we present the ExCEDA –Extended Classroom Engagement Dataset for Assessment. ExCEDA encompasses data from 224 participants, consisting of 6143 images annotated for four crucial affective states: boredom, engagement, confusion, and frustration. Collected during a 50-minute online lecture, ExCEDA uniquely facilitates in-depth analysis of how learner attention fluctuates throughout extended learning periods. We benchmark ExCEDA utilizing state-of-the-art classification models (EfficientNet, Residual Attention Network, and GLAMOR-net). Notably, combining GLAMOR-Net with Facial Action Unit features yields results on par with video-centric datasets. We demonstrate the effectiveness of our approach of combining GLAMOR-Net with Facial features by surpassing the engagement prediction baseline for the EngageNet dataset. Additionally, we introduce CG-ViT, a novel content-guided ViT model, which exhibits substantial performance gains (21.45%, 9.37%, 15.18%, and 13.93% accuracy increases for engagement, boredom, confusion, and frustration, respectively) over the baseline ViT model on our dataset. We also introduce the concept of dataset personalization that enhanced the predictive accuracy of multi-class classification of Confusion, Boredom, Engagement, and Frustration by 17%, 15%, 6%, and 20% respectively. Our novel dataset will be made publicly available.**

***Index Terms*—Engagement Classification, Affect recognition, User Engagement in the Wild**

I. Introduction

Student engagement is a critical determinant of successful learning outcomes in both traditional and online educational settings [13]. However, the online environment creates challenges for educators in gauging student engagement due to the limited visibility of non-verbal cues [31]. Automating engagement assessment in online classrooms offers the potential to augment educators' ability to identify and address disengagement. While video-based datasets like DAiSEE [17], AffectNet [21], EngageNet [25], and the Belfast Database [26] facilitate engagement prediction, accurately identifying disengagement remains a challenge. Image-based analysis can be computationally less demanding than video, enhancing real-time applicability. To address these limitations, we introduce ExCEDA (Extended Classroom Engagement Dataset for Assessment). ExCEDA comprises of images of 224 participants captured over a 50-minute online lecture, annotated for boredom, engagement, confusion, and frustration. Designed to enhance disengagement detection, ExCEDA uniquely includes a substantial representation of low-engagement states, which are often underrepresented in existing short-duration datasets. Our study investigates single-frame engagement recognition using ExCEDA. We benchmark contemporary models (GLAMOR-Net [19], EfficientNet [27], Residual Attention Networks [28]), and leverage facial landmarks through OpenFace [8]. We integrate eye gaze, head pose, and facial action units as crucial indicators. We also introduce CG-ViT, a content-guided ViT model. Our novel approach for multi-class engagement classification demonstrably surpasses the current state-of-the-art performance established by Singh et al. [25]

Key Contributions of our paper include:

1) We present ExCEDA, a unique dataset specifically designed to facilitate the identification of disengaged learners within online environments. ExCEDA captures engagement patterns over a longer duration and represents low-engagement states addressing the limitations in existing datasets.
2) Benchmarking State-of-the-Art Models including Efficient- Net, Residual Attention Networks and GLAMOR-Net on the ExCEDA dataset for single frame student affect recognition.Additionally, we introduce the concept of dataset personalization to enhance predictive accuracy potentially.

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00055

3) We introduce CG-ViT, a novel architecture that integrates content guidance into the Vision Transformer (ViT) framework. This approach leverages contextual cues from the video lecture content to augment ViT's image-based engagement classification, demonstrating a significant improvement over baseline ViT models.

II. Related Work

While early research in affect detection primarily focused on basic emotions [32], recent efforts have shifted towards detecting nuanced mental states such as attention and engagement [10], [20], [32]. Recognizing the need for extensive, labeled, and publicly accessible datasets for training and evaluating applications, various initiatives have emerged in recent years. Concurrently, techniques have been developed to improve educational content quality through curriculum development [1], [3], [6], [7], assessment techniques, and AI-powered educational tools [2], [4], [5], [16].

The DAiSEE dataset, developed by Gupta et al. [17], is an example of a dataset that aims to measure students' involvement in e-learning courses. This collection includes videos of 112 people, with 80 being male and 32 being female. Video frames were annotated at four levels - engagement, boredom, confusion, and frustration. The ratings for the annotations ranged from 0 to 3, allowing for customization to various degrees of involvement.

A separate data collection called HBCU [30] consists of information from 34 people taken from two separate groups of 9 men and 25 women. Participants participated in a Cognitive Skills Training research project by Historically Black Colleges/Universities (HBCUs) and the University of California (UC).

Kaur et al. [18] introduced the EngageWild dataset, comprising videos captured from 78 individuals, including 25 females and 53 males. Crowd-sourced annotations categorized the engagement levels into four levels: disengaged, barely engaged, normally engaged, and highly engaged.

Sathayanarayana et al. [22] presented the SDMATH dataset, featuring videos from one-to-one mathematics tutoring sessions. This dataset offers richly labeled data with video and audio modalities, including deictic gestures, speech, eye gaze, and facial expressions.

The EngageNet Dataset [25], focuses solely on classifying Engagement into Highly Engaged, Engaged, Barely Engaged, and Not Engaged. The dataset includes 31 hours of video capturing 127 participants aged 18-37. Originally 5-10 minutes long, videos are divided into 10-second clips, yielding over 10,000 samples for engagement annotation and modeling.. Table I provides a concise comparison of existing datasets relevant to our work. Student engagement detection is a growing research area, with diverse machine and deep learning algorithms being explored. Zhang et al. [33] introduced an Inflated 3D Convolutional Network (I3D) for automated engagement detection, achieving 52.35% accuracy using the DAiSEE dataset augmented with OpenFace and AlphaPose features. Selim et al. [23] proposed a hybrid model combining pre-trained EfficientNetB7 with temporal modeling techniques , achieving a higher accuracy of 67.48% by emplying EfficientNetB7 combined with LSTM.

TABLE I
COMPARISON OF ENGAGEMENT LABELLED DATASETS

Dataset	Number of participants	Setting	Parameters Detected	Annotation Technique
DAiSEE	112	Diverse Setting, virtual classroom	Engagement, Confusion, Frustration, Boredom	Wisdom of Crowd
HBCU	34	Cognitive skills, training study	Engagement	Manual labeling by experts
EngageWild	78	In the wild	Engagement	Crowdsourced
SDMATH	20	In Person Tutoring	Deictic Gestures	Manual Labeling
EngageNet	127	Virtual Classroom	Engagement	Manual labeling by experts
ExCEDA (Ours)	72	Diverse Setting, Virtual Classroom	Engagement, Confusion, Frustration, Boredom	Manual & Semi Supervised Labeling

III. Our Dataset: ExCEDA

A. Data Collection

In this study, participants, consisting of first-year undergraduate students aged 17-18, were recorded while viewing a 50-minute online classroom lecture on human-computer interaction. We captured a typical online classroom setting which captured the organic emotion and behavior of students in real world scenarios. The lecture was delivered via a dedicated platform, with participants' snapshots captured at 40-second intervals throughout the session. This approach was chosen to accommodate fluctuations in engagement levels over extended video durations. Snapshots were preferred over video capture due to lower computational demands for individual image analysis. The decision to extract snapshots at 40-second intervals was guided by literature indicating that shorter durations lack contextual depth. It was found that meaningful shifts in facial cues and affective states typically occur over one minute, suggesting that shorter intervals may not adequately capture the temporal dynamics of student affect [14], [30].

B. Data Annotations

Our dataset labels four affective states relevant to user engagement: engagement, frustration, confusion, and boredom similar to DAiSEE. Each state is defined on a four-level scale: (1) very low, (2) low, (3) high, and (4) very high, mirroring the approach of [30]. This labeling strategy intentionally omits a "neutral" state. Early experiments revealed crowd annotators' tendency to select "neutral" when uncertain [17], hindering the creation of a robust dataset. The defined levels encourage annotators to make a specific assessment of the affective state, enhancing dataset reliability. The four-level scale was designed to compel specific affective state selections, improving dataset reliability. A panel of three annotators worked collaboratively to assign engagement classifications to each frame of the

TABLE II
EXCEDA DATASET: AFFECTIVE STATE LABEL COMPOSITION

Affective State	Very low	Low	High	Very High
Engagement	28.61%	21.53%	34.79%	15.05%
Boredom	32.96%	16.46%	28.78%	21.78%
Confusion	62.40%	14.38%	17.73%	5.47%
Frustration	63.38%	18.68%	14.13%	3.8%

participants. Nuanced affective states, like user frustration, are inherently subjective constructs. Their interpretation can differ significantly based on the viewer's unique perspectives and biases. Hence, we used majority voting to assign the final labels and employed a cross-labeling technique to verify and correct annotations. Even if the annotators lack expertise, studies have demonstrated that consistent labeling of examples by numerous annotators yields high-quality labels. [24], [29]. We employed weighted Cohen's Kappa [12] with quadratic weights as the performance metric to evaluate the consistency between annotators.

C. Dataset Statistics

The dataset encompasses a cohort of 224 participants, each represented by snapshots obtained at 40 second intervals throughout the entirety of the 50 minute lecture session. The dataset covers a span exceeding 20 minutes. So throughout the span of the 50 minute lecture, we extracted snapshots at 40-second intervals. Then we annotated each of these snapshots for the four affective states of confusion,boredom, frustration and engagement. Hence each frame(snapshot) of the video was annotated. Specifically, data snapshots are available for 101 participants up to the 13th minute (20th frame), for 84 participants up to the 20th minute (30th frame), and for 72 participants up to the 25th minute (40th frame). Moreover, data collection extends to the 33rd minute (50th frame) for 56 participants, to the 40th minute (60th frame) for 39 participants, and to the 50th minute (75th frame) for 17 participants. The images of the students that leave the lecture early on contribute highly to the barely engaged and not engaged classes, thus eliminating the problem of class imbalance in existing datasets. However, for the purpose of conducting thorough analysis and establishing baseline metrics, the dataset subset utilized comprises imagery from 72 students, each spanning beyond the 25-minute mark, denoted by the 40-frame threshold.

We created subject-specific data splits for generalization with 80:10:10 participants in the training, validation, and test sets. For personalization, we split the images of each participant across Train, Test, and Validation in an 80:10:10 split, respectively.

The class distribution of labels is as follows: Highly Engaged (15.05), Engaged (34.79), Barely Engaged (21.53), and Not Engaged (28.61) as shown in II. Unlike previous datasets with pronounced class imbalance [17], [25] hindering accurate prediction of disengagement, our dataset exhibits improved class balance with a substantial representation of "Not Engaged" and "Barely Engaged" classes. This is crucial for enhancing the detection of disengaged students and achieving greater accuracy in predicting the "Not Engaged" class. Moreover, our dataset breaks new ground by capturing student affective states over durations exceeding 20 minutes. This extended timeframe aligns with research indicating attention decline beyond the 20-minute threshold [11]. To the best of our knowledge, no publicly available engagement assessment dataset exists comprising of images captured over a timespan exceeding 20 minutes [11].

Fig. 1. Sample of images in LEEDA for different affects at different levels

IV. METHODOLOGY

A. Feature Extraction

To uncover more nuanced variations in engagement levels displayed in people's faces, facial features can assist in analyzing raw image frames and improving engagement classification.

Facial Feature Extraction and Analysis

To quantify facial behavior indicative of engagement during e-learning sessions, we utilized Action Units (AUs). We integrated these with head pose and eye gaze features, as head and eye movements offer valuable insights into student interest levels within traditional classroom settings. Following the experiment, we analyzed the student's facial features using the OpenFace software [8]. We extracted 49 features from each image.

B. Dataset Splits

We introduce a novel data-splitting approach that has resulted in higher accuracy and reduced loss.

Generalization: To ensure our model's generalizability and its accurate prediction of engagement across diverse classroom settings, we maintain the original data splits. This involves dividing the dataset into training, validation, and testing sets, each containing different students. This methodology fosters robust model development by realistically evaluating its

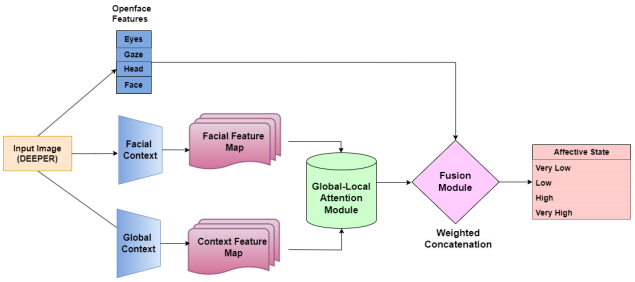

Fig. 2. Model architecture: GLAMOR-Net with Facial features combined

performance. The unseen testing set provides an unbiased assessment of the model's ability to perform well on entirely new student data, reflecting real-world scenarios where the model encounters previously unseen students.

Personalization: In our experimental design, we introduce subject-specific personalization to enhance the model's engagement prediction capabilities. Each participant's 50-minute image sequence is partitioned into training, validation, and testing sets. Unlike the generalization approach, here, some frames for each student are present in each train, test, and validation set. The model is trained to recognize each student's unique patterns and temporal evolution of affective state expression. This personalization strategy lays the groundwork for future model fine-tuning with minimal additional classroom data, enabling rapid adaptation to new students and evolving engagement patterns. These methods prevent the model from becoming overly reliant on specific features or patterns present in the training data. Also by utilizing pre-trained models like ViT [15] and feature extractors, like OpenFace features [8] which are trained on large, diverse datasets, we leverage knowledge learned from a broader set of data, enhancing its generalization ability to new students.

To evaluate the impact of dataset generalization and personalization specifically on the EngageNet dataset, a balanced subset comprising 2800 images was sampled, ensuring equal representation of engagement levels (Table VII).

C. Models

EfficientNet

EfficientNet [27] redefines CNN design with its compound scaling method, simultaneously adjusting depth, width, and resolution via a fixed set of scaling coefficients. This yields superior performance-to-complexity ratios. Its architecture leverages inverted residual blocks featuring depthwise and pointwise convolutions for efficient computation.

Residual Attention Network

The Residual Attention Network (RAN) [28] extends the ResNet architecture by strategically integrating attention modules within its residual blocks. This integration enables the model to selectively focus on salient image regions, facilitating the capture of both fine-grained (local) and broader (global) contextual features.

Global Local Attention (GLAMOR-Net)

The GLAMOR-Net model introduced by Le et al. [19] leverages a multi-modal approach for emotion recognition, combining facial expressions with contextual cues. Feature extraction is performed via separate convolutional neural networks (CNNs) dedicated to the face and background regions, respectively. A global-local attention mechanism selectively emphasizes salient regions within both facial and contextual feature maps, facilitating the learning of inter-dependencies between these modalities.

Global-Local Attention Module

The Global-Local Attention Module [19] seeks to elucidate latent associations between a subject's facial expression and distinctive regions within the image background by assessing their degree of similarity. To facilitate context-aware attention inference, feature maps derived from both the face and the context are concatenated. The ultimate output is a linear combination of context region vectors, weighted according to their respective attention weights.

Fusion Module

The Fusion Module [19] integrates information from two branches using distinct sub-networks dedicated to calculating fusion weights for the two branches, respectively. In our implementation firstly, the Global-Local Attention module merges facial and contextual information, condensing it into attention-aware embeddings that highlight the most relevant elements. After that, OpenFace facial features are incorporated, adding further refinement. These combined sources yield a robust feature set that supports the subsequent emotion classification process.

GLAMOR-Net with Facial features

We employed the GLAMOR-Net model as originally introduced by Le et al. [19], integrating it with facial features such as action units (AUs), gaze, and other relevant cues. Figure I illustrates the model architecture. Our implementation strictly adheres to the Facial and Contextual Encoding modules, as well as the Fusion module, described in the original work by Le et al. [19]. The primary modification lies in the incorporation of facial features as input to the Fusion module.

Content-guided ViT model (CG-ViT)

ViT model [15] is a transformer-based model that matches or even surpasses CNNs for the Image Classification task. We extended the ViT model to incorporate the content the user had viewed for 40 seconds(40s). We achieve this in the following steps: 1) Given an image $x_i \in I$ where I:={ set of images }, we take the output of the final layer of the ViT Encoder(E_1) as feature vector(f_i^1); 2) Given video segment $v_i \in V$ where V:={ set of 40s video segments }, we take the output of the final layer of TimeSformer [9] Encoder(E_2) as feature vector(f_i^2); 3) We concatenate both these vectors and pass it through an MLP head(MLP) for engagement classification(o).

$$f_i^1 = E_1^{freeze}(x_i); f_i^2 = E_2^{freeze}(v_i)$$

$$f_i = \{f_i^1 || f_i^2\}; o_i = MLP(f_i)$$

We have utilized not only the images of the participant but also the video that they are viewing for classifying their

TABLE III
AFFECTIVE STATE CLASSIFICATION RESULTS FOR EXCEDA DATASET

Method	EfficientNet-B0	RAN	GLAMOR-Net
Boredom	0.382	**0.392**	0.365
Engagement	0.394	0.387	**0.412**
Confusion	0.578	0.561	**0.586**
Frustration	0.571	0.485	**0.602**

engagement. We had a hypothesis that every participants' engagement is dependent on the type of content being viewed, like if an example based explanation of the topic is being viewed or the mathematical formulation of the topic in the video is being shown then these two scenarios can have different engagement across participants. Thus we added the previous 40 seconds of the video before the image is taken as a guide to the network to classify the engagement, we do so by taking the participant image features from ViT and the content's video features using the TimeSformer network and concatenate these features and pass it through a classification layer(MLP head) to get whether the participant is engaged or not.

We pre-train the ViT model on the DAiSEE dataset, freeze the layers of both pre-trained ViT and TimeSformer, and only train the classification layer on the ExCEDA dataset. Results can be found in Table V, where we observe significant improvement in the pre-trained ViT on adding content information.

V. RESULTS

Table III summarizes multi-class classification benchmark results on the ExCEDA dataset for four affective states, comparing EfficientNet [27], Local-Global Attention (GLAMOR-Net) [19], and Residual Attention model [28]. On the ExCEDA dataset, our findings indicate that the GLAMOR-Net model, attained the highest accuracies in classifying affective states such as confusion, engagement, and frustration. While Residual Attention Network yielded the highest accuracy for Boredom affective state.

For binary classification, we consolidated engagement levels into two distinct classes. Instances labeled "highly-engaged" and "engaged" were combined into a single class, while "not-engaged" and "barely-engaged" constituted the second class. Our extended analysis of EfficientNet for binary classification of student engagement yielded a noteworthy 78.98% accuracy in identifying the 'Not Engaged' class. This finding is particularly relevant within classroom settings, where accurately detecting disengagement is of paramount importance for effective instruction.

Table IV presents a comparative analysis of dataset generalization and personalization results by emplying the Local Global Attention (GLAMOR-Net) network, both independently and in combination with OpenFace features, across multi-class and binary engagement classification tasks. A consistent and striking finding is the substantial performance improvement achieved through dataset personalization across all model configurations. Notably, the GLAMOR-Net model

TABLE IV
ENGAGEMENT CLASSIFICATION RESULTS OF GLAMOR-NET AND GLAMOR-NET WITH OPENFACE FEATURES ON GENERALIZED AND PERSONALIZED EXCEDA DATASET

Method	Generalization	Personalization
GLAMOR-Net (4 class)	41.20%	52.07%
GLAMOR-Net (2 class)	60.36%	71.90%
GLAMOR-Net + Openface(4 class)	33.0%	43.81%
GLAMOR-Net + Openface(2 class)	54.66%	67.18%

TABLE V
RESULTS FOR BINARY CLASSIFICATION OF CONTENT-GUIDED VIT MODEL ON THE EXCEDA DATASET VS VIT MODEL WITHOUT VIDEO FEATURES

Affective State	F1-score		Improvement(%)
	with content	without content	
Confusion	56.22%	48.81%	15.18%
Boredom	79.00%	72.23%	9.37%
Frustration	63.14%	55.42%	13.93%
Engagement	82.53%	64.83%	21.45%

trained on a personalized dataset yields the highest accuracy for multi-class classification. In each model configuration, there is an improvement of about 10% in the accuracy after dataset personalization.

The content-guided ViT model explores whether there is a correlation between the content being viewed and the engagement level of the participant. Our CG-ViT model results are presented in Table V, revealing that incorporating content information into the model trained on participant expression leads to a significant enhancement in model performance. These findings with CG-ViT offer compelling evidence for the influence of instructional content on student behavior. This emphasizes the potential for educators to strategically tailor lecture content, thereby promoting a more engaged classroom environment.

VI. ABLATION STUDY

Personalization of Data

Across all evaluated models, personalization of student data yielded substantial improvements in final accuracy. Figure 3 illustrates the increase in accuracy achieved in multi-class engagement classification after dataset personalization on ExCEDA across EfficientNet, Residual Attention, GLAMOR-Net, and GLAMOR-Net+OpenFace models. Table VII presents a comprehensive assessment of single-frame engagement recognition on EngageNet and EngageNet combined with LEEDA by employing dataset generalization and personalization. We employ GLAMOR-Net and GLAMOR-Net combined with OpenFace features for binary and multi-class classification tasks. It can be seen that dataset personalization improves the predictive accuracy. Moreover, Table VII shows that the personalization of EngageNet yields an impressive 92.12% accuracy for binary engagement classification. The EngageNet dataset has a significant imbalance towards the "engaged" class. Combining the highly-engaged and engaged classes as "engaged" and the not-engaged and barely-engaged classes as "not engaged" when mapping from 4-level to 2-level

TABLE VI
ENGAGEMENT LEVEL CLASSIFICATION RESULTS OF EFFICIENTNET ON PERSON- ALIZED AND GENERALIZED EXCEDA DATASET

Affective State	EfficientNet Generalization	EfficientNet Personalization
Confusion	57.80	75.25
Boredom	38.00	53.03
Engagement	39.40	45.96
Frustration	57.14	77.78

classification improves the results because the classes become more balanced.

We compared EfficientNet's performance on a generalized and personalized version of ExCEDA dataset as shown in Table VI. There was an improvement in accuracy of multi class classification of Confusion, Boredom, Engagement and Frustration by 17%, 15%, 6% and 20% respectively after dataset personalization. These results demonstrate the efficacy of personalization in enhancing accuracy by training models fine-tuned on a particular batch of people.

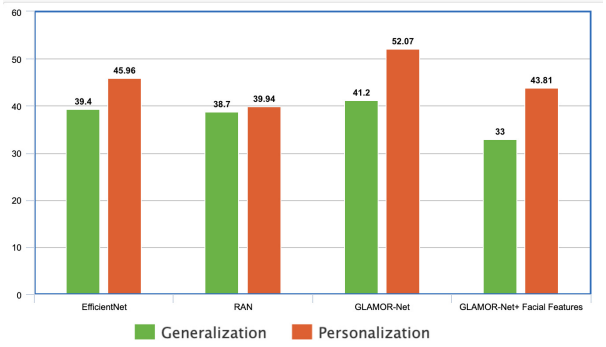

Fig. 3. Accuracy comparison of RAN, EfficientNet, LGA and LGA combined with OpenFace features for generalized and personalized ExCEDA dataset

Employing OpenFace and GLAMOR-Net

Important findings in table VII demonstrate that the integration of OpenFace features improves the model's accuracy. This underscores the powerful interpretive capabilities of OpenFace features (eye gaze, head pose, and facial action units) for affective state recognition.

For the personalized EngageNet dataset, our model integrating GLAMOR-Net with facial features achieves an accuracy of 68.72, surpasses the benchmark of 67.61 established by [25]. Unlike Singh et al.'s transformer-based approach to analyzing videos, our methodology focuses on single-frame analysis for classification. This not only reduces computational requirements but also enhances efficiency.

VII. CONCLUSION AND FUTURE WORKS

In this study, we present ExCEDA, a novel dataset that focuses on four affective states over extended durations and is the first dataset to provide large-scale engagement data exceeding 20 minutes in duration. ExCEDA's balanced representation of boredom and engagement levels improves its effectiveness in

TABLE VII
ENGAGEMENT LEVEL CLASSIFICATION RESULTS FOR ENGAGENET AND ENGAGENET & EXCEDA COMBINED RESPECTIVELY

	EngageNet		ExCEDA+EngageNet	
Models	Gen.	Pers.	Gen.	Pers.
GLAMOR-Net (4 class)	54.94	64.23	53.29	56.32
GLAMOR-Net (2 class)	78.83	86.25	74.34	76.75
GLAMOR-Net+ OpenFace (4 class)	56.64	**68.72**	55.68	56.81
GLAMOR-Net+ OpenFace (2 class)	84.15	**92.12**	77.46	77.72
Transformer (Baseline)	67.61	-	-	-

real-time disengagement detection, providing a more computationally efficient solution than video-based datasets. Through extensive baseline analyses GLAMOR-Net emerged as a top performer, especially on incorporating OpenFace features. Furthermore, we introduce CG-ViT, a groundbreaking model that incorporates the impact of video lectures and student behavior to greatly enhance the predictive capabilities of the engagement model solely trained on student behavior.

Moreover, dataset personalized significantly improved the accuracy across both GLAMOR-Net and EfficientNet. Moreover, after personalizing the EngageNet dataset and then employing GLAMOR-Net and OpenFace features, we surpassed the benchmark by Singh et al. [25]. ExCEDA's exceptional class balance for boredom and engagement enhances its utility for real-time disengagement detection, offering a computationally efficient alternative to video-based datasets. We anticipate that both the ExCEDA dataset and the extensive baseline analyses will serve as valuable benchmarking resources.

VIII. ACKNOWLEDGEMENT

This work is supported by the Advanced Research and Technology Innovation Centre (ARTIC), the National University of Singapore under Grant (project number: A-8000969-00-00). Rajiv Ratn Shah is partly supported by the Infosys Center for AI, the Center of Design and New Media, and the Center of Excellence in Healthcare at IIIT Delhi.

REFERENCES

[1] Anand, A., Addala, K., Baghel, K., Goel, A., Hira, M., Gupta, R., Shah, R.R.: Revolutionizing high school physics education: A novel dataset. In: International Conference on Big Data Analytics. pp. 64–79. Springer (2023)

[2] Anand, A., Goel, A., Hira, M., Buldeo, S., Kumar, J., Verma, A., Gupta, R., Shah, R.R.: Sciphyrag-retrieval augmentation to improve llms on physics q &a. In: International Conference on Big Data Analytics. pp. 50–63. Springer (2023)

[3] Anand, A., Gupta, M., Prasad, K., Singla, N., Sanjeev, S., Kumar, J., Shivam, A.R., Shah, R.R.: Mathify: Evaluating large language models on mathematical problem solving tasks

[4] Anand, A., Jairath, A., Lal, N., Bangar, S., Sikka, J., Verma, A., Shah, R.R., Satoh, S.: Gec-dcl: Grammatical error correction model with dynamic context learning for paragraphs and scholarly papers. In: International Conference on Big Data Analytics. pp. 95–110. Springer (2023)

[5] Anand, A., Kapuriya, J., Kirtani, C., Singh, A., Saraf, J., Lal, N., Kumar, J., Shivam, A.R., Verma, A., Shah, R.R., et al.: Mm-phyrlhf: Reinforcement learning framework for multimodal physics question-answering. arXiv preprint arXiv:2404.12926 (2024)

[6] Anand, A., Kapuriya, J., Singh, A., Saraf, J., Lal, N., Verma, A., Gupta, R., Shah, R.: Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 53–64. Springer (2024)

[7] Anand, A., Prasad, K., Goel, U., Gupta, M., Lal, N., Verma, A., Shah, R.R.: Context-enhanced language models for generating multi-paper citations. In: International Conference on Big Data Analytics. pp. 80–94. Springer (2023)

[8] Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)

[9] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (ICML) (July 2021)

[10] Bohus, D., Horvitz, E.: Models for multiparty engagement in open-world dialog. In: Proceedings of the SIGDIAL 2009 Conference, The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue. p. 10 (2009)

[11] Bradbury, N.A.: Attention span during lectures: 8 seconds, 10 minutes, or more? (2016)

[12] Cohen, J.: Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)

[13] Dewan, M.A.A., Murshed, M., Lin, F.: Engagement detection in online learning: a review. smart learning environments, 6 (1), 1–20 (2019)

[14] Dhall, A., Sharma, G., Goecke, R., Gedeon, T.: Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 784–789 (2020)

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

[16] Goel, A., Hira, M., Anand, A., Bangar, S., Shah, D.R.R.: Advancements in scientific controllable text generation methods. arXiv preprint arXiv:2307.05538 (2023)

[17] Gupta, A., D'Cunha, A., Awasthi, K., Balasubramanian, V.: Daisee: Towards user engagement recognition in the wild. arXiv preprint arXiv:1609.01885 (2016)

[18] Kaur, A., Mustafa, A., Mehta, L., Dhall, A.: Prediction and localization of student engagement in the wild. In: 2018 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2018)

[19] Le, N., Nguyen, K., Nguyen, A., Le, B.: Global-local attention for emotion recognition. Neural Computing and Applications **34**(24), 21625–21639 (2022)

[20] McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., Graesser, A.: Facial features for affective state detection in learning environments. In: Proceedings of the annual meeting of the cognitive science society. vol. 29 (2007)

[21] Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)

[22] Sathayanarayana, S., Kumar Satzoda, R., Carini, A., Lee, M., Salamanca, L., Reilly, J., Forster, D., Bartlett, M., Littlewort, G.: Towards automated understanding of student-tutor interactions using visual deictic gestures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 474–481 (2014)

[23] Selim, T., Elkabani, I., Abdou, M.A.: Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm. IEEE Access **10**, 99573–99583 (2022)

[24] Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 614–622 (2008)

[25] Singh, M., Hoque, X., Zeng, D., Wang, Y., Ikeda, K., Dhall, A.: Do i have your attention: A large scale engagement prediction dataset and baselines. arXiv preprint arXiv:2302.00431 (2023)

[26] Sneddon, I., McRorie, M., McKeown, G., Hanratty, J.: The belfast induced natural emotion database. IEEE Transactions on Affective Computing **3**(1), 32–41 (2011)

[27] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

[28] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2017)

[29] Welinder, P., Perona, P.: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. pp. 25–32. IEEE (2010)

[30] 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)

[31] Xu, D., Xu, Y.: The promises and limits of online higher education: Understanding how distance education affects access, cost, and quality. American Enterprise Institute (2019)

[32] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual and spontaneous expressions. In: Proceedings of the 9th international conference on Multimodal interfaces. pp. 126–133 (2007)

[33] Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y., Li, J.: An novel end-to-end network for automatic student engagement recognition. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). pp. 342–345. IEEE (2019)