ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240551

Unveiling Learner Dynamics: The ECLIPSE Dataset and NeuralGaze Framework for Prolonged Engagement Assessment in Online Learning

Avinash Anand[a,1], Avni Mittal[b], Laavanaya Dhawan[c], Mahisha Ramesh[a], Juhi Krishnamurthy[d], Naman Lal[a,2], Raj Jaiswal [a], Pijush Bhuyan[a], Himani [a], Astha Verma[a], Rajiv Ratn Shah[a], Roger Zimmermann[e] and Shin'ichi Satoh[f]

[a]IIIT, New Delhi
[b]IIT, Mandi
[c]NSUT, New Delhi
[d]Adobe, Noida
[e]NUS, Singapore
[f]NII, Tokyo

ORCID (Avinash Anand): https://orcid.org/0009-0003-2479-0342, ORCID (Avni Mittal): https://orcid.org/0009-0006-9543-9295, ORCID (Laavanaya Dhawan): https://orcid.org/0009-0009-4217-5400, ORCID (Mahisha Ramesh): https://orcid.org/0009-0000-1737-4058, ORCID (Juhi Krishnamurthy): https://orcid.org/0009-0006-2850-8799, ORCID (Naman Lal): https://orcid.org/0009-0008-2914-5509, ORCID (Raj Jaiswal): https://orcid.org/0009-0008-5044-1409, ORCID (Pijush Bhuyan): https://orcid.org/0009-0002-2132-0882, ORCID (Himani): https://orcid.org/0009-0006-2661-5822, ORCID (Astha Verma): https://orcid.org/0000-0003-3615-5373, ORCID (Rajiv Ratn Shah): https://orcid.org/0000-0003-1028-9373, ORCID (Roger Zimmermann): https://orcid.org/0000-0002-7410-2590, ORCID (Shin'ichi Satoh): https://orcid.org/0000-0001-6995-6447

Abstract. Understanding student engagement in online education is crucial for optimizing learning outcomes. This paper introduces ECLIPSE dataset (Extended Classroom Learning Insights via Prolonged Student Engagement), comprising 10,110 annotated images from a 55-minutes , 30-minutes and 20-minutes online lecture. Annotations include four affective states: engagement, boredom, confusion, and frustration. ECLIPSE enables the investigation of learner attention dynamics over extended periods, overcoming the limitations of short-duration datasets. We establish benchmarks for ECLIPSE using models such as EfficientNet, Vision Transformer, Residual Attention Network, and GLAMOR-Net. We propose NeuralGaze, a novel framework integrating Neural Cellular Automata (NCA) with self-attention mechanisms, demonstrating superior accuracy in engagement level assessment compared to basic single-frame models. Furthermore, we introduce CG-SwT, a content-guided Swin Transformer model, which significantly outperforms the baseline ViT model on the ECLIPSE dataset (with F1-score improvements of 21.12%, 12.5%, 16.77%, and 15.41% for engagement, boredom, frustration, and confusion respectively). Our methods surpass existing single-frame engagement prediction baselines for both EngageNet and DAiSEE datasets by significant margins (7.4% and 6.2%, respectively). The code and dataset will be made publicly available.

[1] Equal contribution.
[2] Equal contribution.

1 Introduction

Ensuring effective content delivery and retention in online education is a critical concern[8]. Student engagement, reflecting active involvement and interest, is traditionally gauged through non-verbal cues such as body language and facial expressions[16]. These cues enable real-time adjustments to lesson delivery[32, 19].

However, the shift to digital education complicates the assessment of engagement due to the absence of physical cues, making it difficult for educators to monitor understanding and enthusiasm. While digital platforms offer accessibility, they lack the immediacy of traditional settings.

Video-based datasets like DAiSEE[12], AffectNet[21], EngageNet[26], and the Belfast Database[27] have been crucial for predicting engagement levels. Although these datasets use temporal data for engagement evaluation[30, 13, 18, 1], video-based analysis is computationally intensive, challenging real-time application. Single-image analysis offers a computationally efficient alternative, particularly in low-resource environments, though detecting disengagement remains difficult due to dataset imbalances favoring high engagement.

To address this, we introduce the ECLIPSE (Extended Classroom Learning Insights via Prolonged Student Engagement) dataset, capturing 250 participants' images at 45-second intervals during online lectures. Each image is annotated with four emotional states—boredom, engagement, confusion, and frustration—rated on

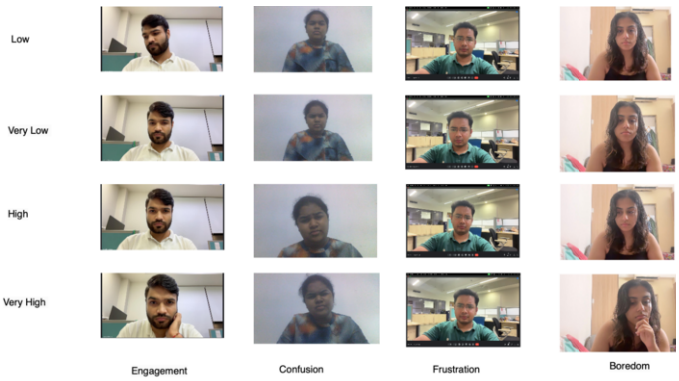

Figure 1. Sample of images in ECLIPSE for different affects at different levels

a scale from 0 to 3[31]. Unlike existing datasets, ECLIPSE includes prolonged engagement data, reflecting the decline in engagement typically observed after 20 minutes of instruction[4].

We evaluate single-frame engagement recognition using ECLIPSE, DAiSEE, and EngageNet datasets, employing baselines such as EfficientNet[28], GLAMOR-Net[15], Residual Attention Networks[29], and Vision Transformers[10]. We also introduce two new techniques: CG-SwT, integrating watched content to predict engagement, and NeuralGaze, which combines image embeddings, facial landmarks, and gaze vectors with Neural Cellular Automata (NCA)[22]. These methods address class imbalance with weighted sampling and focal loss, and demonstrate a 7.4% accuracy improvement over the current best transformer-based approach[26].

Our contributions are:

1. Introduction of the ECLIPSE dataset, including 250 participants and 10,110 images, to assist in identifying disengaged students.
2. Benchmarking State-of-the-Art Models for single-frame engagement recognition across DAiSEE, EngageNet, and ECLIPSE datasets.
3. Development of CG-SwT, enhancing classification performance using content guidance with Swin Transformers.
4. Introduction of NeuralGaze, a framework integrating facial features, gaze vectors, and NCAs for competitive performance with reduced computational overhead.

2 Related Work

Student Engagement Prediction: Identifying student engagement in classroom settings is a critical challenge. Janez and Andrej et al. [33] developed a system using Kinect sensors to estimate attention levels, leveraging visual features like gaze, head motion, and body posture, applying machine learning models to predict attentiveness. Goldberg [11] found that gaze features and facial action units were highly indicative of engagement in seminar videos, achieving a 0.44 correlation with manual annotations.

Recent efforts underscore the need for extensive, labeled datasets for training and evaluating engagement prediction models. However, despite numerous studies, only a few publicly accessible datasets are available.

The DAiSEE dataset by Gupta et al. [12] measures students' involvement in e-learning courses with videos from 112 participants, annotated for engagement, boredom, confusion, and frustration using crowd-sourced ratings from 0 to 3. The HBCU dataset [31] includes data from 34 individuals, manually labeled for different engagement levels. Kaur et al. [14] introduced the EngageWild dataset, featuring videos from 78 individuals with crowd-sourced annotations classifying engagement into four levels. Sathayanarayana et al. [24] presented the SDMATH dataset, containing videos from one-to-one tutoring sessions, offering richly labeled data with both video and audio modalities. The EngageNet dataset [26] focuses on classifying engagement into four levels, with 31 hours of video from 127 participants, aged 18-37, with original videos divided into 10-second clips.

Table 1. Comparison of engagement labeled datasets

Dataset	Students	Setting	Parameters	Labelling Method
DAiSEE	112	Virtual classroom	Engagement, Confusion, Frustration, Boredom	Wisdom of Crowd
HBCU	34	Training study	Engagement	Manual labeling by experts
EngageWild	78	In the wild	Engagement	Crowd sourced
SDMATH	20	In Person Tutoring	Deictic Gestures	Manual Labeling
ECLIPSE (Ours)	250	Virtual Classroom	Engagement, Confusion, Frustration, Boredom	Manual & Semi Supervised Labeling

Table 1 compares existing datasets relevant to our work. Beyond these datasets, research into student engagement detection has leveraged various resources, including the EmotiW datasets [17, 6], the Engagement Recognition (ER) database [20], and other video-based collections [25, 34].

Neural Cellular Automata: Cellular automata (CA) [5] generate complex behaviors through rule-based interactions. Neural cellular automata (NCA) [22] extend this concept by using neural networks to define update rules, making them adaptable for various tasks. Due to their efficient architecture, NCAs have gained traction in image generation [23]. Unlike conventional deep learning models analyzing entire images simultaneously, NCAs focus on individual pixels, making them lightweight and efficient.

Limitations: A significant challenge in predicting student engagement is the lack of appropriate public datasets and the issue of class imbalance, particularly affecting the analysis of disengaged students. Existing datasets often focus on short video segments, limiting the assessment of engagement over extended durations. Previous studies rely heavily on video-based methods, which are impractical for real-time assessment in low-resource environments. A more feasible alternative involves analyzing engagement at specific time intervals using single-frame prediction methods. However, there is a noticeable gap in research on gauging engagement levels through single frames.

3 ECLIPSE Dataset

3.1 Data Collection

A 55-minute online classroom lecture on human-computer interaction was viewed by participants in this study, who were first-year undergraduate students between the ages of 17 and 18. For the purpose

of accommodating changes in participant attention across longer video durations, participant snapshots were taken at 45-second intervals during the session. The lower processing demands for individual image analysis made snapshots the favored method of capturing images. Since meaningful shifts in facial cues and affective states usually occur over one minute, the literature suggested that shorter durations lack contextual depth and may not adequately capture the temporal dynamics of student affect [9, 31]. This led to the decision to extract snapshots at 45-second intervals. Furthermore, our dataset pioneers the capture of student affective states for durations longer than twenty minutes, consistent with studies showing a drop in attention beyond twenty minutes [4]. To the best of our knowledge, no engagement evaluation dataset that is made accessible to the public includes photos taken over longer than 20 minutes.

3.2 Data Annotations

The dataset we have includes four affective states that are important for user involvement: engagement, annoyance, bewilderment, and boredom, similar to the DAiSEE dataset. Each state is categorized using a four-level scale: (1) extremely low, (2) low, (3) high, and (4) very high. This labeling method deliberately excludes a "neutral" state. The initial trials demonstrated that crowd annotators have a proclivity to choose the label "neutral" when they are unsure, which hampers the development of a strong and reliable dataset. The purpose of the four-level scale was to ensure that participants make precise selections regarding their affective state, hence enhancing the dependability of the dataset. A group of three annotators collaborated to assign engagement classifications to each frame of the participants. We implemented a cross-labeling strategy to verify and rectify any errors in the annotations. Three annotators categorized participants in images into classes based on their level of engagement, boredom, confusion. and frustration.

Table 2. Weighted Cohen's kappa for the three annotators

State	Labeler 1 vs 2	Labeler 1 vs 3	Labeler 2 vs 3
Boredom	0.839	0.735	0.749
Engagement	0.891	0.981	0.875
Confusion	0.855	0.719	0.702
Frustration	0.813	0.805	0.708

We utilized weighted Cohen's Kappa with quadratic weights as the performance metric to assess the consistency among annotators. This metric measures the degree of agreement across numerous annotators who categorize the same data points, thereby evaluating the inter-rater reliability. A Kappa value ranging from 0.70 to 0.80 is considered to be satisfactory. The Kappa coefficients for our dataset are located in Table 2.

3.3 Dataset statistics

The dataset consists of photos from 250 students, with 100 students contributing exclusive photographs over 20 frames, each captured at 45-second intervals. These images significantly enhance the representation of the "Barely Engaged" and "Not Engaged" classes, addressing the class imbalance seen in prior datasets. The class distribution is: Highly Engaged (15.05%), Engaged (34.79%), Barely Engaged (21.53%), and Not Engaged (28.61%) as shown in Figure 3. Unlike previous datasets such as Gupta et al. (2016) and Singh et al. (2023), which struggled with class imbalance, our dataset offers a more balanced distribution, particularly for the lower engagement classes, as depicted in Figure 2.

Table 3. ECLIPSE 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%

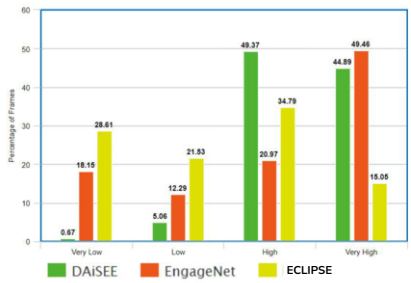

Figure 2. Engagement affect's class distribution for comparison between EngageNet, DAiSEE & ECLIPSE

3.4 Research Ethics and Participant Protections

Prior to participating, all subjects were required to offer informed consent, which was documented by their signature. Participation was plainly optional, and participants maintained the right to withdraw at any time without providing a reason. In order to protect the privacy of the participants, a distinct identifier was allocated to each individual, ensuring that it had no connection to their personal information.

4 Proposed Methodology

We aim to fuse local facial features, such as expressions, facial landmarks, and gaze vectors, with global image details like surrounding context, body pose, and hand movements. Facial expressions and gaze direction provide emotional and focus cues, while body pose and hand movements offer broader contextual insights into engagement. Integrating these elements, our analysis encompasses subtle facial expressions and broader behavioral cues. To achieve this, we devise a model architecture and methodology inspired by GLAMOR-Net[15], tailored for capturing engagement prediction data. Figure 3 gives an overview of our model with the NCA-based feature extraction module for engagement classification.

4.1 NCA-based Feature Extraction Module

We develop a Neural Cellular Automata[22](NCA) based encoding module to extract meaningful image embedding. The hidden channels per pixel capture global information learned by propagating local information surrounding a pixel over multiple time steps. Each pixel in an image is treated as a cell, with each cell containing a set of learned channels. These channels capture local pixel interactions and are updated using learned rules. Instead of generating new images, we use the final state of these channels as rich feature representations for downstream classification tasks. This shared update rule, which leads to extensive parameter sharing, reduces the overall parameter count ensuring parameter efficiency, fast training, and

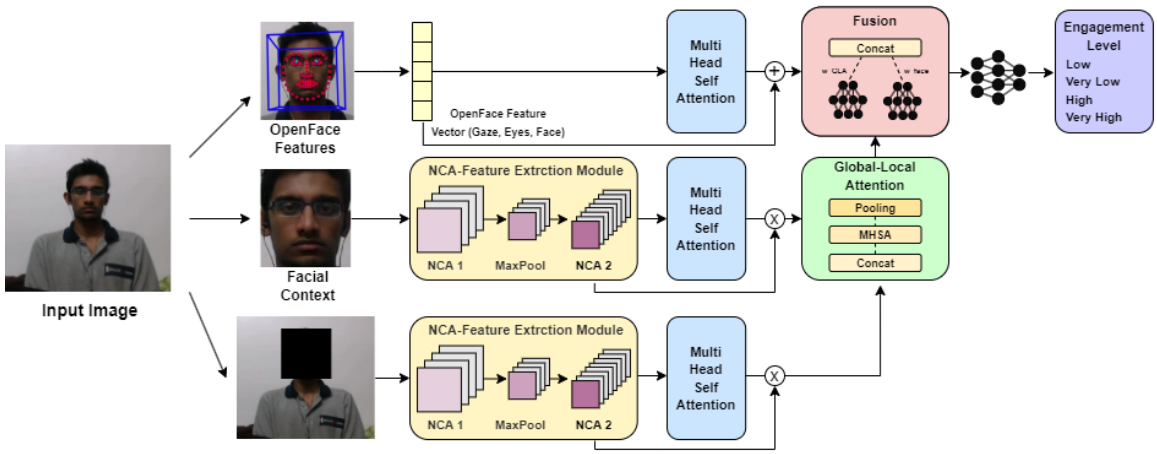

Figure 3. Model architecture with NCA-based feature embedding module combined with Multi-head self-attention and facial landmarks

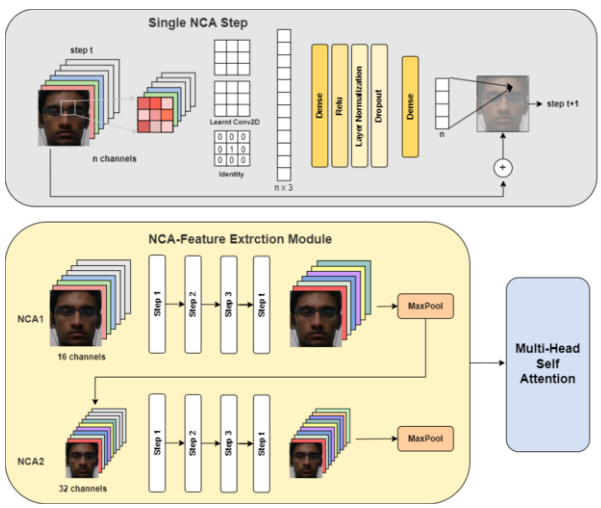

Figure 4. NCA-based feature extraction module consists of 2 NCA models. For NCA1, the original image is concatenated with 16 additional hidden channels. The output of NCA1 is downsampled using MaxPooling and given as input to NCA2 after the addition of 32 additional channels

convergence, making the model lightweight and deployable with low computational resources. Thus, NCAs provides a valuable approach for creating image embeddings because they efficiently capture local and global information.

The complete encoding module utilizes two NCA models. The output of the first NCA model undergoes a MaxPool with a pool size and stride of (4,4). Subsequently, the reduced embedding is passed to a smaller NCA model, NCA2, with increased channel length to capture more information. Empty channels are concatenated to the downsampled embedding from NCA1 and sent to NCA2. The final output embeddings learned from NCA2 undergo downsampling using a MaxPool layer with a pool size and stride (4,4). For the feature extraction module of Global content, an additional MaxPool operation is applied with a pool size of (2,2) and stride 1. This ensures that both facial and global context features are dimensionally aligned for subsequent processing. Refer to Figure 4 for a detailed illustration of the NCA feature extraction module and a single NCA step.

The NCA model captures feature embeddings by considering the pixel's important neighborhood locations and propagating that knowledge across the image. MaxPooling layers extract the most significant features, while the second NCA model refines the embeddings through additional iterations.

Multi-head self-attention (MHSA) is then applied to the output of the second NCA, enhancing the embeddings with attention information. This attention mechanism determines the relevance of specific features. This NCA-based attentive feature extraction module is repeated for both Facial and Global Context to extract their respective encodings. The student's face is cropped out for the facial context, and the facial region is blacked out for the global context before using the complete image. Taking cues from GLAMOR-Net, the context module aims to actively acquire valuable data from the environment rather than redundant facial details.

4.2 *Global-Local Attention module*

The GLAMOR-Net-inspired idea of having a global-local attention(GLA) module is adapted to combine the global and local feature embeddings. The outputs of the facial and context NCA-based feature extraction module are concatenated and sent to another MHSA block. The output from the MHSA block is then added to the combined concatenated features to get an attention-aware representation. This attention-aware representation is then passed through layer normalization and global average pooling to get the final vector embedding, which combines the global and local features. This GLA module implementation is entirely different from that used in GLAMOR-Net.

4.3 *Head Pose and Facial Features*

The extracted head pose and facial features, such as facial landmarks and gaze vectors, extracted using OpenFace[2], are combined with the GLA feature vector in the fusion module before performing the final classification. Multi-head self-attention is applied to the extracted features. The output is added to the original feature vector to enhance representation, which is used further.

Motivated by experiments [26] conducted using GLAMOR-Net with and without OpenFace features and getting better results in the former, we include the OpenFace facial features in the final proposed architecture.

4.4 Fusion Module and Classification Network

Combining outputs from the Global-Local Attention module, which has attention-aware combined and condensed embeddings of the Facial and Context information, with the improved OpenFace Facial Features, we get the final features for downstream classification. Separate neural networks calculate the score for GLA and OpenFace module outputs. The score is then normalized using the softmax function to obtain the corresponding weights. The weighted vectors are concatenated to be sent to a Fully Connected Neural Network for final classification.

5 Experiments

5.1 Dataset Splits

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

Generalization: In order to ensure the applicability of our model in various classroom environments, we partitioned the dataset into several training, validation, and testing sets, each comprising different groups of students. This approach promotes the creation of strong models by accurately assessing their performance. The testing set offers an impartial evaluation of the model's capacity to excel with completely novel student data. We allocated participants in the training, validation, and test sets using an 80:10:10 ratio.

Personalization: We implement subject-specific personalization by dividing each participant's 55-minutes , 30-minutes , 20-minutes video image sequence into separate training, validation, and testing sets. Contrary to the generality technique, each train, test, and validation set contains specific frames for every student. The model is trained to identify the distinct patterns and temporal changes in how each student expresses their emotional state. This technique of personalization establishes the foundation for future refinement of models with minimal more data from the classroom. It allows for quick adjustment to new students and changing patterns of interaction. The photos of each participant were divided into training, testing, and validation sets in an 80:10:10 ratio, respectively.

We used the complete EngageNet dataset, as shown in Table 4, to evaluate and compare the effectiveness of different models in predicting engagement levels across multiple classes and binary classes. The results reported correspond to the validation split of the original dataset. In order to assess the effects of dataset generalization and customization on the EngageNet dataset, a subset of 3000 photos was selected. This subset was carefully chosen to ensure that each degree of interaction was equally represented, as shown in Table 8.

5.2 Baseline Models

We compare the results of the benchmarks created using EfficientNet[28], Vision Transformers[10], Residual Attention Networks[29], and GLAMOR-Net[15] with the original baselines for DAiSEE and EngageNet. We also compare the results of our proposed Content-guided Swin Transformer and NCA-based Self-Attention framework, NeuralGaze, to capture Global and Local information.

EfficientNet: EfficientNet[28] is a compact yet accurate convolutional neural network designed for efficiency at scale through compound scaling. We fine-tune the final classification layer of the EfficientNet-B0 model, initially pre-trained on ImageNet[7].

Residual Attention Networks: Residual Attention Networks[29] combine residual connections for gradient flow optimization with attention modules to enhance model optimization and representation learning in computer vision. Key facial features like eyes and mouth are emphasized in tasks like engagement assessment. We train the end-to-end RAN on the datasets.

Global-Local Attention (GLAMOR-Net): GLAMOR-Net[15] integrates facial expressions and contextual cues for emotion recognition, utilizing separate convolutional neural networks (CNNs) for feature extraction from the face and surrounding context. These features undergo a global-local attention mechanism to highlight salient aspects. Finally, the fused features are used for emotion prediction. Our extension of GLAMOR-Net incorporates additional features such as OpenFace head pose, eye gaze, and Facial Action Units (FAU) into the fusion module, alongside the utilization of Focal Loss (FL). Through ablation studies, detailed in Table 9, we demonstrate the efficacy of these enhancements in enriching contextual information and improving model performance.

Vision Transformer (ViT): The ViT [10] model represents images as sequences of patches, converted to embeddings capturing appearance and spatial information. Self-attention in a transformer encoder analyzes these embeddings for local and global contexts, followed by classification via an MLP head. We fine-tune its final classification layer for engagement-level prediction by utilizing the pre-trained ViT model on the ImageNet-21k [7] dataset.

Content-guided Swin Transformer model (CG-SwT)

We extended the SwT model to incorporate the content the user had viewed for 45 seconds. 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 Swin Encoder(E_1) as feature vector(f_i^1); 2) Given video segment $v_i \in V$ where V:={ set of 45s video segments }, we take the output of the final layer of TimeSformer[3] 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 pre-train the Swin Transformer model on the DAiSEE dataset, freeze the pre-trained Swin Encoder and TimeSformer layers, and only train the classification layer on the ECLIPSE dataset. Results can be found in Table 7, where we observe significant improvement in the trained ViT on adding content information.

5.3 Implementation Details

The models are trained for a maximum of 60 epochs with a batch size of 8, the best validation model is saved, and the corresponding results on the test dataset are recorded. Adam optimizer uses a learning rate scheduler that decays the learning rate after regular intervals starting from 5×10^{-4} with a decay of 0.1. The hyperparameters chosen for the NCA feature extraction module include NCA channel size for the hidden feature of 16 channels for NCA1 and 32 additional channels for NCA2, a fire rate of 0.5, and a dense layer hidden dimension of 128. The NCA model initializes the additional hidden channels with a 0 value. The exact architecture can be seen in Figure 3. All the Multi-head Attention modules use eight heads of attention. All the dense neural networks are single-layer Neural Networks with hidden layer dimensions 128. Focal Loss has been used. Classification accuracy is chosen as the evaluation metric due to its widespread use

within the domain. In our four-class classification, we use the original levels of "very low," "low," "high," and "very high." However, in our two-class classification, we combine the levels of "very low" and "low" into one category labeled "low," and the levels of "high" and "very high" into another category labeled "high."

6 Results

Table 4 summarizes multi-class and binary classification results for three datasets: DAiSEE, EngageNet, and ECLIPSE, focusing on different levels of the engagement affective state. The results highlight GLAMOR-Net as the top performer across all datasets, especially when combined with OpenFace facial features and focal loss for handling class imbalance. GLAMOR-Net combined with focal loss surpassed the other models by a significant margin. For the EngageNet dataset, our models integrating GLAMOR-Net with focal loss and GLAMOR-Net with both facial features and focal loss achieve an accuracy of 0.750, surpassing Singh et al.'s benchmark of 0.676 [26]. 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.

Table 5 reveals that for confusion, boredom, and frustration, GLAMOR-Net coupled with OpenFace features and focal loss demonstrates superior performance, followed by our proposed NeuralGaze method for the DAiSEE dataset. Notably, in two-class classifications, especially for categories such as Boredom and Frustration, NeuralGaze surpasses GLAMOR-Net. These categories showcase a more equitable distribution of data points across various engagement levels. This achievement underscores the efficacy of NeuralGaze in situations where a balanced dataset is accessible.

On the ECLIPSE dataset, our findings indicate that the GLAMOR-Net model, when combined with OpenFace features and focal loss, consistently attained the highest accuracies in classifying affective states such as confusion, boredom, and frustration, both in four-class and two-class classification tasks. This suggests the effectiveness of OpenFace features, including eye gaze, head pose, and facial action units, in interpreting affective states. Furthermore, the incorporation of weighted loss functions such as focal loss led to improved accuracy by mitigating the disproportionate influence of majority classes.

Our analysis of various video-based models on DAiSEE, detailed in Table 6, underscores the potential for extracting valuable insights from video data. However, computational feasibility and cost considerations are critical factors that must be addressed through methodological enhancements.

The results in Table 6 compare the performance of our model with existing single-frame-based and video-based benchmarks. Our proposed architecture, NeuralGaze, and modification to GLAMOR-Net to incorporate Facial Action Units and Head Pose surpass the single-frame classification benchmark and have competitive performance with other video-based benchmarks that use multiple frames per video for analysis. We observed increased accuracy with increasing frames processed for a single video classification. This suggests the potential for further exploration of temporal domain classification methods, with a specific interest in incorporating Local-Global Attention embeddings and computationally effective video data analysis methods in future studies.

The content-guided Swin Transformer model investigates the relationship between the content being viewed and the participant's

Table 4. Engagement level Classification Results for DAiSEE, EngageNet, and our dataset ECLIPSE.

Method	EngageNet	DAiSEE	ECLIPSE
EfficientNet-B0	0.606	0.526	0.394
ViT	0.604	0.456	0.382
RAN	0.604	0.541	0.387
GLAMOR-Net (4 class)	0.586	0.568	0.412
GLAMOR-Net + FL (4 class)	**0.75**	0.569	**0.75**
GLAMOR-Net + FA + FL (4 class)	**0.75**	**0.572**	0.389
NeuralGaze (4 class)	0.604	0.551	0.366
Transformer (benchmark for EngageNet)	0.6761	-	-
GLAMOR-Net (2 class)	0.871	0.95	0.532
GLAMOR-Net + FL (2 class)	0.878	0.95	0.5
GLAMOR-Net + FA + FL (2 class)	0.882	0.9	0.561
NeuralGaze (2 class)	0.871	0.953	0.471

Table 5. Confusion, Frustration, and Boredom results for DAiSEE and ECLIPSE dataset for various models

Method	Dataset	Confused	Bored	Frustrated
EfficientNet-B0	DAiSEE	0.671	0.462	0.777
	ECLIPSE	0.578	0.382	0.571
ViT	DAiSEE	0.672	0.481	0.777
	ECLIPSE	0.576	0.227	0.626
RAN	DAiSEE	0.672	0.468	0.777
	ECLIPSE	0.561	0.392	0.485
GLAMOR-Net	DAiSEE	0.682	0.484	0.782
	ECLIPSE	**0.586**	0.365	0.602
GLAMOR-Net +FL (4 class)	DAiSEE	0.688	**0.492**	0.789
	ECLIPSE	0.572	0.360	**0.657**
GLAMOR-Net +FL+FA (4 class)	DAiSEE	**0.691**	0.491	**0.793**
	ECLIPSE	0.584	**0.413**	0.636
NeuralGaze (4 class)	DAiSEE	0.688	0.479	0.790
	ECLIPSE	0.582	0.300	0.629
GLAMOR-Net (2 class)	DAiSEE	0.91	0.74	0.742
	ECLIPSE	0.719	0.564	0.784
GLAMOR-Net (2 class) + FL	DAiSEE	0.92	0.73	0.751
	ECLIPSE	**0.726**	0.596	0.783
GLAMOR-Net +FA + FL (2 class)	DAiSEE	**0.94**	0.74	0.755
	ECLIPSE	0.725	**0.654**	**0.786**
NeuralGaze (2 class)	DAiSEE	0.916	**0.790**	**0.967**
	ECLIPSE	0.718	0.420	0.784

Table 6. Engagement level prediction comparison with single-frame and video-based benchmarks. Our model gives competitive performance to video-based models and surpasses existing single-frame benchmarks. We set new benchmarks with GLAMOR-Net+FA+FL and NeuralGaze

Configuration	Frames	Accuracy
EmotionNet (DAiSEE)	1	0.5107
NeuralGaze (Ours)	1	**0.554**
GLAMOR-Net + FA + FL	1	**0.572**
DFSTN	20	0.5884
Marlin + FA + BodyPose Features	30	0.59
ResNet + TCN	50	0.639
LRCN (DAiSEE)	250	0.579
BERN	300	0.60

engagement level. Our CG-SwT model results, as shown in Table 7, demonstrate that integrating content information into the model trained on participant expression significantly improves performance. These findings provide strong evidence of the impact of instructional content on student behavior, highlighting the potential for educators to strategically design lecture content to foster a more engaged classroom environment.

Table 7. Results for binary classification of Content-guided Swin Transformer model on the ECLIPSE dataset vs ViT model without video features

Classes	F1-score		Improvement
	with content	without content	
Confusion	60.82%	52.70%	15.41%
Boredom	82.76%	73.57%	12.5%
Frustration	71.01%	60.82%	16.75%
Engagement	81.92%	67.64%	21.12%

6.1 Ablation Study

6.1.1 Personalization of Data

Table 8 presents a comprehensive assessment of single-frame engagement recognition on ECLIPSE and EngageNet with dataset generalization and personalization. We employ GLAMOR-Net and GLAMOR-Net combined with OpenFace features for binary and multi-class classification tasks. Key findings demonstrate that dataset personalization and integration of OpenFace features improve the model's accuracy.Furthermore, as shown in Table 8, personalizing EngageNet achieves a remarkable 92.12% accuracy for binary engagement classification. The EngageNet dataset is notably imbalanced, favoring the "engaged" class. By merging the highly-engaged and engaged categories into a single "engaged" class, and the not-engaged and barely-engaged categories into a "not engaged" class, when transitioning from a 4-level to a 2-level classification, the class distribution becomes more balanced, which leads to improved classification results. The most notable improvements were seen in Boredom and Frustration, with accuracy increasing by around 15% after dataset personalization. Utilizing the GLAMOR-Net model in conjunction with OpenFace features trained on a personalized dataset resulted in the highest accuracy for multi-class classification. These results demonstrate the efficacy of personalization in enhancing accuracy by training models fine-tuned on a particular batch of people.

Table 8. Engagement level Classification Results for personalized and generalized EngageNet and ECLIPSE respectively.

Model	EngageNet		ECLIPSE	
	Gen.	Pers.	Gen.	Pers.
GLAMOR-Net (4 class)	54.94	64.23	40.92	**52.07**
GLAMOR-Net + OpenFace (4 class)	56.64	**68.72**	33	43.81
GLAMOR-Net (2 class)	78.83	86.25	53.27	71.90
GLAMOR-Net + OpenFace (2 class)	84.15	**92.12**	54.66	**67.18**

6.1.2 Employing OpenFace and GLAMOR-Net

We aimed to assess the interpretability of OpenFace features by performing classification tasks using these features exclusively. A simple Artificial Neural Network (ANN) was trained on extracted openface features for binary and multi-class classification. The outcomes of these classification tasks on the DAiSEE dataset are summarized in Table 9. Our findings indicate that the extracted OpenFace features, including eye gaze, head pose, and facial action units, significantly interpret an individual's affective state.

Table 9. DAiSEE Ablation on OpenFace Features(4 class)

Affective State	ANN	ANN with Focal Loss	ANN	ANN with Focal Loss
	F1 /Acc 4 class	F1 / Acc 4 class	F1 /Acc 2 class	F1 / Acc 2 class
Confusion	0.68 / 0.69	0.57 / 0.66	0.85 / 0.88	0.87 / 0.90
Boredom	0.42 / 0.45	0.40 / 0.44	0.70 / 0.74	0.73 / 0.77
Engagement	0.46 / 0.48	0.47 / 0.48	0.93 / 0.95	0.96 / 0.96
Frustration	0.64 / 0.62	0.69 / 0.77	0.92 / 0.93	0.93 / 0.95

6.1.3 Employing Focal Loss

We employed weighted loss functions such as cross-entropy and focal loss to address the class imbalance challenge within the dataset. These techniques aimed to mitigate biases towards the majority classes and improve the overall performance of the trained models. Comparisons in Table 5, 4 and Table 9 using Focal loss show that using it effectively reduced the impact of class imbalance on model training.

7 Conclusion

In this study, we present ECLIPSE, the first dataset to provide large-scale engagement data exceeding 20 minutes in duration. ECLIPSE's balanced representation of boredom and engagement levels improves its effectiveness in real-time disengagement detection, providing a more computationally efficient solution than video-based datasets. We also introduce CG-SwT, enhancing results obtained with ViT by integrating video lecture content. The particularly notable improvement with CG-SwT highlights the impact of content-driven tailoring on student learning outcomes. We also proposed NeuralGaze- a model combining local, global, and extracted facial features using Neural Cellular Automata (NCA). Through extensive baseline analyses for single-frame engagement recognition, GLAMOR-Net emerged as a top performer, particularly when incorporating OpenFace features and Focal Loss. Our study outperforms the baseline single-frame classification results on DAiSEE and EngageNet datasets and established competitive baselines for ECLIPSE, which will serve as valuable benchmarks for future research in this field.

Acknowledgements

This research is supported by the Advanced Research and Technology Innovation Centre (ARTIC) at the National University of Singapore under Grant A-8000969-00-00. Additionally, the resources for experimentation partially supported by the Infosys Center for AI, the Center of Design and New Media, and the Center of Excellence in Healthcare at Indraprastha Institute of Information Technology, Delhi.

References

[1] A. Abedi and S. S. Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *2021 18th Conference on Robots and Vision (CRV)*, pages 151–157. IEEE, 2021.

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

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

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

[5] B. Chopard and M. Droz. Cellular automata. *Modelling of Physical*, pages 6–13, 1998.

[6] O. Copur, M. Nakıp, S. Scardapane, and J. Slowack. Engagement detection with multi-task training in e-learning environments. In *International Conference on Image Analysis and Processing*, pages 411–422. Springer, 2022.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] M. Dewan, M. Murshed, and F. Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1–20, 2019.

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

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein. Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33:27–49, 2021.

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

[13] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang. Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pages 338–341. IEEE, 2019.

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

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

[16] Y. Li and R. M. Lerner. Interrelations of behavioral, emotional, and cognitive school engagement in high school students. *Journal of youth and adolescence*, 42:20–32, 2013.

[17] Y.-Y. Li and Y.-P. Hung. Feature fusion of face and body for engagement intensity detection. In *2019 IEEE international conference on image processing (ICIP)*, pages 3312–3316. IEEE, 2019.

[18] J. Liao, Y. Liang, and J. Pan. Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, 51:6609–6621, 2021.

[19] L. Mishra, T. Gupta, and A. Shree. Online teaching-learning in higher education during lockdown period of covid-19 pandemic. *International journal of educational research open*, 1:100012, 2020.

[20] O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan, and C. Paris. Automatic recognition of student engagement using deep learning and facial expression. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 273–289. Springer, 2020.

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

[22] A. Mordvintsev, E. Randazzo, E. Niklasson, and M. Levin. Growing neural cellular automata. *Distill*, 5(2):e23, 2020.

[23] R. B. Palm, M. González-Duque, S. Sudhakaran, and S. Risi. Variational neural cellular automata. *arXiv preprint arXiv:2201.12360*, 2022.

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

[25] P. Sharma, S. Joshi, S. Gautam, S. Maharjan, S. R. Khanal, M. C. Reis, J. Barroso, and V. M. de Jesus Filipe. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In *International Conference on Technology and Innovation in Learning, Teaching and Education*, pages 52–68. Springer, 2022.

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

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

[28] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

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

[30] Y. Wang, A. Kotha, P.-h. Hong, and M. Qiu. Automated student engagement monitoring and evaluation during learning in the wild. In *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)*, pages 270–275. IEEE, 2020.

[31] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagement-from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

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

[33] J. Zaletelj and A. Košir. Predicting students' attention in the classroom from kinect facial and body features. *EURASIP journal on image and video processing*, 2017(1):1–12, 2017.

[34] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu. Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research*, 58(1): 63–86, 2020.