Relating Students Cognitive Processes and Learner-Centered Emotions: An Advanced Deep Learning Approach

Ashwin T S
Gautam Biswas
ashwindixit9@gmail.com
gautam.biswas@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

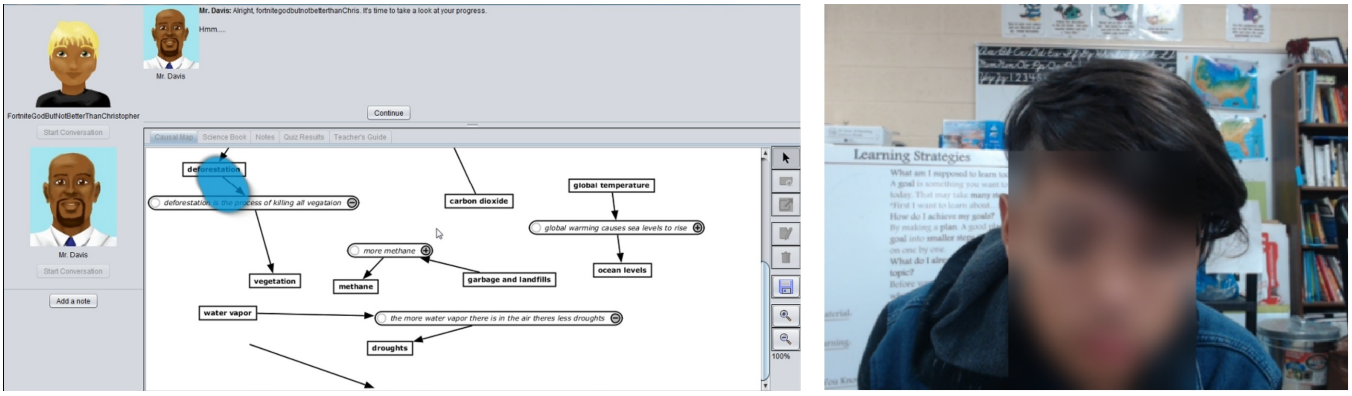

Figure 1: Screenshot from Betty's Brain Open Ended Learning Environment (OELE) and Student's Web Camera.

Abstract

While understanding Self-Regulated Learning (SRL) in Open-Ended Learning Environments (OELEs), it is crucial to examine the interplay between students' cognitive processes and affective states, especially learning centered emotions like delight, engagement, boredom, frustration and confusion. These affective states are particularly challenging to detect using facial expressions in middle school students, primarily due to the scarcity of relevant databases. This study introduces a novel approach that utilizes the EmoNet framework, enhanced with self-attention networks, to detect and analyze learning-centered emotions. We investigated the emotional and cognitive dynamics of 41 middle school students within an OELE. Our findings demonstrate distinct emotional patterns that significantly correlate with students' performance levels across various cognitive processes. By creating and analyzing a dataset from ten students, the proposed model achieved a test accuracy of 85%, indicating a substantial improvement over existing state-of-the-art models. These results lay the groundwork for future educational tools capable of adapting to a combination of students' affective and cognitive states thus enhancing their overall learning experiences that influence their educational outcomes.

CCS Concepts

• **Applied computing** → **Education**; **Computer-assisted instruction**; E-learning; • **Computing methodologies** → *Computer vision*; Machine learning.

Keywords

Affective States, Cognitive Processes, Self-regulated Learning, Learning-Centered Emotion, Open-ended Learning Environments, Facial Expressions, Deep Learning, Multimodal Learning Analytics, Vision Transformers, Convolutional Neural Networks, Academic Emotions

ACM Reference Format:
Ashwin T S and Gautam Biswas. 2024. Relating Students Cognitive Processes and Learner-Centered Emotions: An Advanced Deep Learning Approach. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685751

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685751

1 Introduction

Self-regulated learning (SRL) requires students to effectively integrate and coordinate their Cognitive, Affective, Metacognitive, and Motivation (CAMM) processes [9]. These processes are crucial in determining students' engagement with the learning environment and their learning performance in Computer-Based Learning Environments (CBLEs) [41]. Effective regulation of CAMM processes

helps students maintain a productive learning path, navigate challenges, and optimize their academic outcomes. The complexity of these tasks demands that students develop strategies to monitor their progress and adapt accordingly [46].

Students' emotions (affective states) and cognitive processes are closely intertwined within the framework of SRL. Studies have noted that emotions like confusion, frustration, and boredom significantly impact students' engagement, learning behaviors, persistence, and performance [14, 16]. While emotions like delight and flow (engagement) can positively influence performance, others can hinder it. Understanding these relationships can help develop more supportive and adaptive learning environments [8, 15].

Cognitive disequilibrium arises when students face uncertainty and difficulties during learning, leading to affective states like confusion, frustration, and boredom. Confusion is often the initial response to challenging tasks, which can either propel deeper engagement or escalate to frustration if unresolved. Prolonged frustration can diminish engagement and lead to boredom. This dynamic interplay underscores the need to map these affective states to students' learning performance [14, 16]. Research shows unresolved confusion and frustration can negatively impact learning outcomes, particularly when not effectively managed [10].

While traditional methods for recognizing students' emotions in learning environments often rely on manual observation or self-reports, both approaches have biases [31, 32]. Round-robin observation can miss key moments, and self-reports can lead to inaccuracies because students may not feel comfortable indicating negative emotions like frustration [42]. Moreover, current state-of-the-art deep learning methods in computer vision are primarily trained on limited demographics (adults and undergraduate students), making it challenging to analyze the interplay of cognitive processes and emotions for younger students at a fine-grained level [7, 18]. No comprehensive database exists containing class labels of children's learning-centered emotions. Therefore, a novel methodology is required to detect and understand learning-centered emotions, which will allow educators to tailor interventions and ensure effective emotional and cognitive regulation.

To address these challenges, our work employs a multimodal approach that integrates video data (capturing affective states through facial expressions) and interaction log data (capturing cognitive processes through mouse and keyboard interactions). By time-aligning these data sources, we can comprehensively analyze the interplay between cognitive and affective states, leading to more accurate and meaningful insights. Our overall contributions in this paper include:

- *A Novel Emotion Detection Methodology*: This paper presents a novel methodology for detecting learning-centered emotions (engagement/flow, frustration, confusion, delight, and boredom) in middle school students using facial expressions.
- *Distribution of Emotions Across Cognitive Processes*: The paper also analyzes the distribution of emotions between high and low performers for each cognitive process, offering deeper insights into how these emotions influence performance.

The rest of this paper is organized as follows. We first present a literature review of the theoretical background for understanding the relationship between affective states and cognitive processes in SRL. Next, the methodology section details the proposed emotion recognition framework and data analysis techniques. This is followed by the results section, where findings on the distribution of emotions across cognitive processes are presented. Finally, the paper discusses implications for designing more adaptive learning environments and concludes with suggestions for future research.

2 Literature Review

In Self-Regulated Learning (SRL), cognitive processes can involve the systematic acquisition and processing of information to meet learning objectives. Winne's Information Processing Theory identifies five core cognitive operations – searching, monitoring, assembling, rehearsing, and translating – which learners apply sequentially in their learning tasks [45]. The CAMM framework broadens SRL's definition by considering the interplay between Cognitive, Affective, Metacognitive, and Motivation processes [9]. Open-ended Learning Environments (OELEs), such as Betty's Brain and others, provide inquiry-based learning experiences that challenge students to set goals, devise plans, and engage in active reflection. These cognitive strategies are analyzed to understand how students seek information, refine their knowledge, and evaluate progress [22]. However, learners often struggle to simultaneously acquire knowledge and regulate their learning, highlighting the need for scaffolding in OELEs [3, 9].

As explained by Pekrun's control-value theory, academic emotions are linked to perceived control and value, which in turn are linked to positive and negative affective states and students' learning performance. The theory underscores how emotions impact motivation, engagement, and achievement [34]. D'Mello and Graesser's affective dynamics in learning environments mainly focus on the importance of confusion, frustration, and boredom [16]. Therefore, state-of-the-art emotion recognition and tracking methods based on deep learning techniques and computer vision can play an important role in understanding student learning. However, their application in OELEs remains limited due to training biases and data scarcity [5, 43].

Recent advances in deep learning-based affect models have primarily targeted adult learners, with a notable deficiency in extensive datasets for training models to recognize emotions pertinent to children's learning. Research efforts have aimed to generate emotion-related datasets across various educational settings, including open-ended learning environments [29], traditional classrooms [40], embodied learning contexts [44], and x-reality or game-based educational platforms [30]. When data is collected using Closed-Circuit Television (CCTV) footage in computer science labs [6, 36], the distribution of this data diverges from that obtained via a laptop webcam in a computer-based study setting [2, 3], where the focus is on individual students. Moreover, the spectrum of emotional states exhibited in game-based learning is different from that in conventional educational contexts, with the former displaying more physical engagement in conjunction with facial expressions [8, 17]. Conversely, learners remain relatively stationary when working on computer-based OELEs, with the upper body playing a more significant role in inferring emotional states. To date, no existing methodologies or research has concentrated on analyzing children's facial expressions as they work in OELEs. While emotions can also

be detected through alternative modalities, such as speech and physiological signals, these approaches are less viable in OELEs due to minimal verbal interaction. Physiological measurements may also be deemed to be intrusive in e-learning contexts.

The relationship between cognitive processes and emotions has been studied to understand how emotions shape learning behaviors using interaction log data combined with the human observation protocol Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) protocol [11]. Recognizing affective states at a fine-grained level is essential to understand students' emotions more accurately and tailor interventions accordingly. Current methods are based on manual observations and self-reports, which are prone to biases and inaccuracies [27, 32].

Emotion recognition for basic emotions is well-explored in the literature, with several state-of-the-art methods demonstrating high accuracy on standard facial expression databases. Some architectures that perform well include EmoFAN, AffectNet, Face Single Shot MultiBox Detector (Face SSD), and High-speed emotion recognition (HSEMotion) [13, 24, 26, 28, 39]. Additionally, methods like bidirectional Long Short-Term Memory (LSTM) can utilize temporal data for emotion prediction. However, these databases primarily focus on adults, and some, like Acted Facial Expressions In The Wild (AFEW), consist of acted images [24].

While these models are well-trained to extract low-level features from faces for emotion recognition, they cannot be directly used for recognizing emotions in children. There are several methods specifically aimed at recognizing basic emotions in children, but they are not trained on large databases, nor are their weights available for further training [1, 13, 21, 26]. None of these models, however, are applied for learning-centered emotion recognition.

In the education domain, several methods rely on Support Vector Machines (SVMs) or basic Convolutional Neural Networks (CNN) models for recognizing children's faces and have not utilized transfer learning to make the models more robust [19]. State-of-the-art methods like Inception or transfer learning from well-trained models are used for learning-centered emotion recognition, and there are some databases for adults focused on learning-centered emotions, but these are predominantly for undergraduate or graduate students [1, 7, 18].

Hence, to address these gaps, we propose a novel methodology that accurately detects learning-centered emotions like engagement, frustration, confusion, delight, and boredom using facial expressions. In addition, we explore the distribution of emotions among high- and low-performing students, leveraging this fine-grained emotional data to enhance our understanding of cognitive-affective transitions. Ultimately, this will contribute to developing adaptive, affect-aware learning environments.

3 Learning Environment and Data

Betty's Brain, an OELE designed for middle school students, uses a learning-by-teaching approach to help students learn their science by building a causal model of a scientific process and, at the same time, by developing their cognitive and metacognitive abilities to become better learners [22, 25].

Figure 1 depicts the system interface, which provides learners with a variety of resources and tools for knowledge acquisition, model construction, and model evaluation. The system components include a science book, which is a collection of hypermedia resource pages that provide the subject knowledge for constructing the causal model. Learners read relevant pages of the science book to learn about relevant science concepts and the causal (cause-and-effect) relationships between these concepts. Students then teach their agent by using a visual causal map construction and viewing tool.

Table 1: Cognitive Processes and Description

Cognitive Process	Actions/Description
Information Acquisition	Reading hypertext resource pages (Read)
Solution Construction	Building and refining causal maps (Build)
Solution Assessment	Engaging in quiz-related activities (Quiz)

The system provides additional tools to assess the causal map's correctness using the query and quiz features. Quiz results help students check the correctness of their causal maps by uncovering errors and omissions in their current causal map. Proficient learners use this feedback to find and correct errors, but other students often have difficulties in translating their quiz results into productive, actionable information [22]. Students may also choose to find pages in the science book that help them review the knowledge corresponding to incorrect or incomplete answers before they continue to build their causal map. In essence, the quizzes allow students to track their learning progress and, therefore, their understanding of the required science knowledge.

We analyzed the cognitive processes of 41 students, consisting of 12 males and 29 females, learning about climate change topics in our study using interactive log data extracted from the Betty's Brain environment in Comma-Separated Value (CSV) format. Additionally, we collected web camera data to analyze facial expressions and emotions. The inclusion criteria, such as excluding students with corrupted data or those absent on any day of the study, resulted in a final sample size of 41 students. Each student, aged between 10 and 12, worked approximately 40 minutes per day on Betty's Brain, generating a total of around 5000 minutes of screen-recording videos over the three days of the study. The video data, captured through Open Broadcaster Software (OBS) using the laptop's webcam, maintained a resolution of 1092*614 at a frame rate of 30 frames per second. The students' final map scores, calculated by subtracting incorrect causal links from correct ones, were collected and documented. We performed a median split of the final map scores, resulting in 16 high performers and 17 low performers. The Institutional Review Board's approval was obtained, and all necessary participant consent procedures and formalities were diligently followed. A Sample image screenshot of a student using the Betty's Brain learning environment is shown in Figure 1.

3.1 Cognitive Processes

As discussed earlier, we considered students' *Read*ing the science book, *Build*ing the causal map, and *Quiz*ing to check the correctness of their map as the primary activities students conducted in the Betty's Brain environment. All of these actions are collected with timestamps and additional contextual information in the system

logs. Kinnebrew et al. [22] mapped these actions to higher-level cognitive processes within the Betty's Brain learning environment, (1) Information Acquisition (IA), (2) Solution Construction (SC), and (3) Solution Assessment (SA). The description of cognitive processes and their mapping are provided in Table 1.

3.2 Affective States

In general, basic emotion recognition is well-defined and frequently used, but in education, we focus on emotions that are more specifically known as academic or learning-centered emotions. The most dominant learning-centered emotions observed during the learning process are engagement (flow), confusion, frustration, delight, and boredom. These represent affective states compared to mere emotions. Confusion and boredom are epistemic affective states, while engagement is a cognitive-affective state. Frustration and delight are achievement emotions [14, 15]. The definitions as per the literature [15, 33] are provided below for a better understanding of the classification model.

- Engagement/Flow: A state of high concentration and positive valence where learners are fully immersed in their task and find the challenge both stimulating and manageable.
- Confusion: A state of moderate arousal and negative emotions where learners experience uncertainty or a lack of clarity in comprehending the material.
- Frustration: A state of high arousal and negative emotions where learners feel blocked or overwhelmed by the task at hand.
- Delight: High arousal and positive emotions resulting from the joy or satisfaction of overcoming challenges and successfully grasping the material.
- Boredom: A state characterized by low arousal and negative emotions where learners feel disengaged or uninterested in the learning material.

4 Methodology

Figure 2 shows the complete methodology. The EmoNet model [24] is well-trained on various databases and exhibits state-of-the-art performance. This model includes valence and arousal class labels along with basic emotions, which are used in the literature to map to learning-centered emotions [15, 16]. Since the initial layers are adept at recognizing low-level features, we used this model as a backbone and trained the final layers with annotated images of learning-centered emotions. Additionally, facial action units play a crucial role in identifying peak emotions, so we incorporated a facial action unit detector with transformers [20] as a self-attention network to enhance classification accuracy further. The complete details of the method and annotation process are presented below.

4.1 EmoNet Backbone

The EmoNet model is pre-trained on a large dataset of basic facial emotions and is designed for robust emotion classification. Its architecture is a deep convolutional neural network (EmoFAN) [24], and the feature extractor layers are denoted as: *Feature Extractor:* $f(x, \theta_F)$ where x represents a facial image from the learning-centered emotion dataset, and θ_F are the weights learned from the pre-trained EmoNet model. By retaining these weights, the feature extractor captures valuable features (like facial landmarks and expressions) relevant to basic emotion detection, which can then be mapped to learning-centered emotions.

Remove Final Layers: The final classification layers of EmoFAN (EmoNet) are specifically designed for basic emotions. Removing these layers and replacing them with new fully connected layers enables the model to specialize in learning-centered emotions like confusion, frustration, boredom, engagement, and delight. The new layers are defined as: *Classifier:* $g(z, \theta_C)$, where z represents the output of the high-level features by the feature extractor, and θ_C are the weights of the new classifier layers. These weights are explicitly learned for the classification of learning-centered emotions.

Transfer Learning: For feature extraction, we froze the feature extractor layers $f(x, \theta_F)$ to retain their pre-trained weights and prevent them from being updated during training.

Custom Layers Training: Trained only the new classifier layers $g(z, \theta_C)$ on the learning-centered emotion dataset. This allowed the model to adapt to new emotion labels without modifying the foundational feature extraction layers (Equation 1):

$$\min_{\theta_C} \mathcal{L}\left(g(f(x, \theta_F), \theta_C), y\right) \tag{1}$$

where y represents the target labels for learning-centered emotions, and $\mathcal{L}$ is the loss function which is categorical cross-entropy. This is used to measure the difference between the predicted and actual emotion classes.

4.2 Incorporating Facial Action Units (FAUs)

ROI Attention Module: In this we have Region-of-Interest (ROI) maps that focus on specific facial regions relevant to the FAUs, comparing the predicted attention maps against ground truth maps derived from facial landmarks. The ground truth maps are manually annotated based on facial landmarks.

$$E_{\text{att map}} = L_\delta\left(F_M(f(x)) - A_m(x)\right) \tag{2}$$

The Huber loss (L_δ) measures discrepancies between predicted maps ($F_M(f(x))$) and ground truth maps ($A_m(x)$).

Self-Attention Network: Using a Vision Transformer (ViT) model [20], we applied a self-attention mechanism to identify relevant facial action units. Each attention head is trained to focus on important facial regions. The attention output A can be described as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where:

- Q, K, V: Query (Q): Represents the features that need to be evaluated for their importance in predicting learning-centered emotions. It allows the model to query specific facial regions. Key (K): Contains information about potential attention regions (e.g., facial action units) that the model should focus on. Value (V): Holds the actual data from the facial image that the attention mechanism ultimately uses to refine predictions.
- d_k: The dimensionality of the key vectors is used to scale the dot product, helping stabilize gradients.

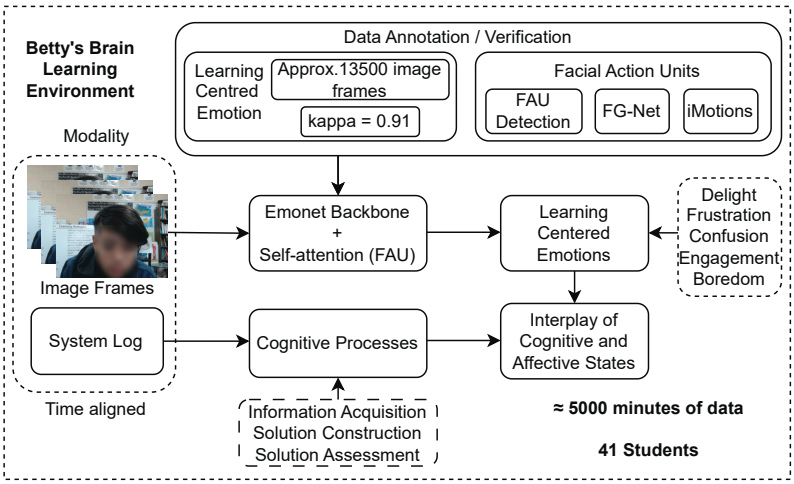

Figure 2: Complete flow of the methodology.

Fine-Tune with FAUs: The attention features (weights) from the ViT model was added to the new fully connected layers of EmoNet. This involves concatenating or combining the attention features with the original feature map, yielding a richer representation:

$$z' = \text{concat}\,(z, A) \tag{4}$$

Then, fine-tuned the model on the created learning-centered emotion dataset by optimizing both the facial action unit model and the final classification layers of EmoNet:

$$\min_{\theta_C, \theta_{\text{ViT}}} \mathcal{L}\left(g(z', \theta_C), y\right) \tag{5}$$

where θ_{ViT} represents the weights of the ViT attention model. The loss function now optimizes both the ViT weights θ_{ViT} and classifier weights θ_C to accurately classify learning-centered emotions. This combination of features from the ViT attention mechanism and EmoNet's original feature extractor is to improve classification accuracy.

We used the 12th Gen Intel® Core™ i9-12900F processor clocked at 2.40 GHz, coupled with the NVIDIA GeForce RTX 4070 graphics card and a total of 32GB RAM for our emotion recognition architecture.

4.3 Annotation

Annotating learning-centered emotions is a labor-intensive task. To partially alleviate the task, much like D'Mello et al. [15], we used Russell's circumplex [37] of emotions to identify discrete emotions on a valence-arousal scale in the education domain; these methods are not specifically applied to OELEs involving children of the age group considered in this study. Therefore, we manually annotated 10 instances of each learning-centered emotion and ran the HSEmotion model [38, 39] to find the valence and arousal values. From this, we used the same model for the same person with the same valence arousal values to find out all other instances across the days. This way we generated many instances for each class label that we collected into separate folders. We manually verified it and deleted the ambiguous image frames from the folders. This semi-automated emotion annotation process reduced the annotation time. We did it separately for each student because the valence arousal values varied as the expression of these emotions varied among the students.

This semi-automated annotation for 10 students' data produced ≈ 13500 image frames, ensuring an equal division of gender (5 males and 5 females) and diverse demographic backgrounds, including 3 white Americans, 3 African Americans, 2 Hispanics, and 2 Asians. We intentionally selected participants through purposeful sampling to address the imbalance in the training data, particularly concerning demographic diversity, and to ensure a more balanced representation. Two different annotators independently verified each emotion using facial expressions, and the inter-rater reliability, measured by Cohen's Kappa, was 0.91. For training the model we used the instances where both annotators were in full agreement ($\kappa = 1$) for each learning-centered emotion. We also performed manual verification for facial action units using a semi-automated approach. We employed three different methods to identify the action units: iMotions [12], Facial Action Unit (FAU) Detector with transformer [20], and FG-Net [47]. While iMotions detects 22 AUs, FAU detector and FG-Net both considered 12 AUs. To ensure consistency, we used the subset of 12 AUs mentioned in [20] and [47] for classification. The detection threshold for these AUs was set at 0.8 instead of the standard 0.4, and classifications were made accordingly. For image frames where all three methods consistently detected the same AUs, these results were considered as ground truth for the self-attention model. These frames were then manually verified by two different annotators for 500 random instances before using them as ground truth, and both agreed that the classifications were accurate.

This cognitive process-related data is then cross-referenced with the CSV file generated from emotion recognition. The latter contains frame numbers, timestamps, and corresponding 5 class labels based on the system timestamp. The objective is to amalgamate

these files, resulting in a comprehensive dataset that provides insights into the total instances of each emotion label for every instance of the cognitive process during its specific duration. Here, each instance of emotion refers to each frame.

5 Results and Inferences

We discuss two sets of results. Our first set of results provide a comparative analysis of the performance of the different approaches we use for emotion recognition and classification. We tested our methods on the annotated data because ground truth was needed for training and testing. We used approximately 13,500 annotated images, of which 73% were engagement (9,855 images), 9% confusion (1,215), 8% boredom (1,080), 7% frustration (945), and 3% delight (405). As the data was imbalanced, we used data augmentation to increase the number of instances of each emotion to match the amount for engagement. Specifically, we applied nine different data augmentation techniques, to address the class imbalance [43]. Furthermore, we performed student-independent 10-fold cross-validation (ensuring students in the training set were not present in the test set) to evaluate the model's generalizability [6, 43]. For the second part of the results, the proposed model was tested on data from all 41 students, without applying any data augmentation. For the second set of analyses, we looked at the differences between the affect distributions of the high and low performers across the three primary cognitive processes: Information Acquisition, Solution Construction, and Solution Assessment.

Learning centered Emotion Recognition: The performance metrics of the proposed model are given in Table 2. The overall accuracy of the model is 85%. Engagement is the most accurately recognized learning-centered emotion with an F1-score of 0.90. Misclassification primarily occurred between confusion and frustration (10% of confusion instances are predicted as frustration) and between frustration and boredom (4% of frustration instances are misclassified as boredom). Boredom's performance is relatively balanced, but 8% of boredom instances are misclassified under engagement. The overall precision (0.83), recall (0.80), and F1-score (0.78) demonstrate reasonable classification accuracy, but there's room for improvement, particularly in distinguishing closely related emotions like frustration and confusion. In our ablation study, we placed greater emphasis on precision over recall, particularly in the educational context, where providing incorrect feedback based on misclassified emotions like frustration can be more disruptive to the learning process than missing an instance of frustration. This approach aligns with the literature, which underscores the importance of precision in educational interventions [8, 15, 23, 35].

Table 2: Performance Metrics for learning-centered emotion

Emotion	Precision	Recall	F1-score
Confusion	0.81	0.78	0.74
Frustration	0.78	0.75	0.71
Boredom	0.76	0.75	0.73
Engagement	0.91	0.88	0.90
Delight	0.89	0.85	0.81
Overall	0.83	0.80	0.78

Valence-Arousal Mapping: For the same set of images that were tested for learning-centered emotions, we ran the analysis for valence-arousal and classified the quadrants they fall into according to Russell's circumplex model [37]. We used the specific ranges of valence and arousal values defined in the literature to accurately identify each affective state [3, 4, 15, 17]. For instance, engagement, as a learning-centered emotion, has a small positive arousal and a positive valence value, indicating that engagement falls into the first quadrant. We mapped all five learning-centered emotions to valence-arousal quadrants as defined in the literature and found that 92% of learning-centered emotion instances align correctly. Deviations were found in engagement (3%) and confusion (4%), where confusion instances were near neutral or in the third quadrant. Engagement expressions occasionally fell slightly below the neutral arousal line but remained primarily positive valence.

Facial Action Unit: Among the open-source action unit recognition methods (weights) available, FAU Detection [20] and FG-Net [47] use larger datasets compared to other methods. FAU Detection leverages three datasets: Denver Intensity of Spontaneous Facial Action (DISFA), EmotioNet, and Binghamton-Pittsburgh 3D Dynamic Spontaneous Facial Expression Database (BP4D), whereas FG-Net only uses DISFA and BP4D. iMotions, a commercially available software that we purchased, uses AFFDEX 2.0 [12], which has been trained and tested on multiple databases, including AffectNet.

We experimented with different thresholds for facial action unit recognition using iMotions, FAU Detection, and FG-Net. We tested thresholds of 0.4 (the standard used by iMotions), 0.6, and 0.8. Overall, the highest rate of misclassification or disagreement was observed in FG-Net (37%), whereas both iMotions and FAU Detection had 21% disagreements across all thresholds.

Although we selected a threshold of 0.8 with complete agreement across image frames from the three models for training the self-attention model, we chose FAU Detection with transformers over FG-Net for the self-attention network. iMotions performed well, likely because it is trained on several databases compared to other models, but we did not have access to AFFDEX weights to include them in this methodology.

Model Comparison: The HSEmotion model [38, 39], which is open-source and trained on comprehensive emotion recognition datasets like AffectNet, EmotiW challenges (AFEW, Video level Group AFfect (VGAF), and EngageWild), and Affective Behavior Analysis in-the-wild (ABAW) challenges (Learning from Synthetic Data and Multi-task Learning), served as a comparison for the proposed EmoNet with Self-Attention model. HSEmotion achieved a mean Average Precision (mAP) of 0.71, while EmoNet with Self-Attention achieved a significantly higher mAP of 0.83.

Ablation: To evaluate the effectiveness of the Self-Attention mechanism, we also compared EmoNet with and without Self-Attention, where the latter resulted in an mAP of 0.78, while the inclusion of Self-Attention increased it to 0.83.

Testing different thresholds for facial action unit detection further impacted classification accuracy, with average detection accuracies of 76%, 81%, and 85% at thresholds of 0.4, 0.6, and 0.8 respectively.

Layer ablation studies showed that reducing the EmoNet backbone by one layer decreased the average F1-score to 0.74, whereas

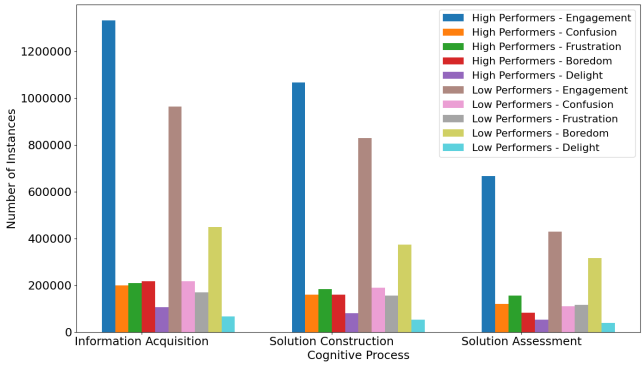

Figure 3: Distribution of Learning-Centered Emotions.

adding two layers maintained it at 0.78. Regarding ground truth annotation, using a single annotator resulted in an average detection accuracy of 81%, but having two annotators improved this to 85%. These comparisons demonstrate that the proposed EmoNet with Self-Attention consistently outperforms other configurations and models, highlighting the importance of incorporating Self-Attention and optimizing thresholds and annotation processes.

Distribution of High and Low Performers' Cognitive Processes:

Figure 3 illustrates the distribution of learning-centered emotions across three cognitive processes: Information Acquisition (IA), Solution Construction (SC), and Solution Assessment (SA). The bar chart shows clear differences between the affect state distributions of the high- and low-performing students within each category. As discussed earlier, we used a median split of the students' final map scores to define the high and low performer split. These map scores were derived by comparing the students' progress against an expert map, as detailed in the system log files generated during the map-building stage. High Performers exhibited a consistent pattern of high engagement across all cognitive processes, underscoring their ability to remain focused and on task. For instance, during the Information Acquisition phase, engagement peaked significantly at 1,333,300 instances, overshadowing other emotions such as boredom, which registered 216,660 instances, and confusion at 199,995 instances. This trend of predominant engagement continued in the Solution Construction and Solution Assessment phases, although the gap narrowed with other emotions, such as frustration and boredom. Low performers, also showed high levels of engagement. However, they did experience notably longer instances of boredom, especially during the Information Acquisition phase, where it nearly doubled compared to high performers. This indicated a disengagement from the learning task, presumably because the students had difficulty in understanding and keeping their interest in the learning material. However, this difference narrowed for the Solution Assessment phase. The data underscored a crucial contrast between high and low performers: although both groups showed high levels of engagement, high performers had greater periods of engagement than low performers. This difference translated to proportionally greater periods of boredom among the low performers, and this may have adversely affected their learning outcomes. We conducted a Chi-square test of independence to show the difference in the affect distributions of high- and low-performers. The test resulted in a significant difference with a p-value of < 0.001. Table 3 summarizes these results.

Table 3: Consolidated Chi-Square Test Results for Cognitive Processes

CP	Chi-square	Dof	Note on Expected Frequencies
IA	145587.85	4	Significant variances observed
SC	124215.49	4	Variations across categories
SA	193449.75	4	Highest discrepancies noted
All	442167.21	4	Consistent significant differences

CP: Cognitive Processes; DoF: Degrees of freedom

Significance of Results Across Processes: The Chi-Square tests conducted for the "IA," "SC," and "SA" processes, as well as the overall aggregation, consistently revealed significant disparities in emotional distributions between high and low performers, each with a p-value of < 0.001. These results emphasize a strong association between students' performance levels and their emotional responses during different cognitive processes.

IA (Read), SC (Build), and SA (Quiz) Processes: In each specific process—Read, Build, and Quiz—the high Chi-square statistics indicate that the actual emotional distributions deviate significantly from what would be expected if there were no association between performance levels and emotional outcomes. The Read process showed notable differences, particularly with a Chi-square statistic of 145587.85, suggesting that reading activities might elicit strong emotional disparities. Similarly, the Build and Quiz processes demonstrated significant emotional variance, with the Quiz process exhibiting the most pronounced differences, indicated by a Chi-square value of 193449.75.

Overall Emotional Distribution: The overall test, combining all processes, produced a Chi-square statistic of 442167.21, affirming that the observed patterns are not isolated to specific types of cognitive tasks but are pervasive across all examined educational interactions.

5.1 Discussion

Emotion Transition: Both high and low performers experience boredom, frustration, and confusion during the learning process, though the frequency of these emotions, especially boredom, which represents disengagement varying significantly between the groups. According to the literature on the dynamics of learning-centered emotions, an engaged student encountering cognitive disequilibrium typically transitions into a state of confusion, which can escalate to frustration and, if unaddressed, lead to boredom [16]. Our analysis of the CSV file containing emotion data with timestamps reveals that *low performers tend to progress from frustration to boredom more swiftly compared to high performers.* In contrast, high performers actively engage in coping mechanisms such as laughter and peer interaction, which help them quickly return to an engaged state.

This resilience was particularly evident during the 45-minute sessions involving extensive reading and model building, which are likely to induce confusion and frustration when students cannot understand the material they are reading or successfully translate

that reading to correct map links. Overall, our data shows that while confusion and frustration were prominent in both groups, low performers tended to transition into boredom and remain in a state of boredom more often than the high performers. Despite these challenges, moments of delight were also observed for both groups; however, high performers, who were more successful in building correct causal maps and, therefore, got better results on the quizzes, experienced more frequent instances of delight.

Emotion Detection: In emotion recognition, particularly in learning environments, confusion and frustration are commonly misclassified due to overlapping facial action units (AUs) that exhibit similar physical manifestations. Literature suggests a range of AUs associated with these emotions, which may contribute to their frequent misclassification.

From the manual annotation and also while verifying the semi-automatically annotated emotions, it was observed that both confusion and frustration share AUs, for example, brow lowerer and head forward. Confusion often involves AUs like the brow lowerer (indicative of deep thinking or difficulty understanding), which can be mistaken for frustration, which also utilizes brow lowerer but typically in a more intense form paired with other indicators like lip tightener or nose wrinkler, signaling annoyance or dissatisfaction. The facial action units present in the lower part of the face for confusion will have more positive valence than that of frustration, and the model that we used for the self-attention network did not use all 22 action units; hence, the chances of misclassification were high. Similarly, frustration is associated with AUs like hands on head or scratching head, which indicate a higher level of distress or exasperation. These AUs can occasionally appear during intense cognitive effort, typically in states of confusion, leading to misclassification.

Delight is characterized by unique AUs that are distinctly different from those associated with negative emotions like confusion or frustration. These typically include the cheek raiser and lip corner puller. These AUs are part of a genuine smile, commonly known as the Duchenne smile (AU6), which is difficult to confuse with expressions of confusion or frustration. However, a cheek raiser is also observed during the engagement, and hence, there is a slight misclassification of delight with engagement. Boredom can often be misclassified with engagement as sometimes boredom is just a resting face with a lack of strong facial movements. Boredom sometimes does not typically involve intense facial expressions, which makes its AUs subtle and easy to confuse with engagement class label. Also, cheek puffer is seen in both frustration and boredom, which causes a lot of misclassification when this is observed and requires temporal data to classify the emotion correctly.

Contextual Similarity: In educational settings, both emotions often arise from similar contexts (e.g., challenging tasks or obstacles), which makes distinguishing based purely on facial cues more challenging without considering contextual factors such as task difficulty or individual learner responses over time.

5.2 Limitations

There are several limitations in this study. Since the annotations are made for static images, fleeting facial expressions can sometimes be misclassified. For instance, a closed eye might be considered boredom (sleeping) when it could simply be a case of blinking. However, when analyzing a sequence of frames, such instances do not significantly impact the results. The self-attention mechanism uses facial action units that could capture mixed emotions, like a student laughing out of frustration, which may then be classified as a positive emotion. But this may not happen quite often. Our data is also limited to a specific demographic and a single learning environment. However, the literature indicates that even for 41 students, sharing classification weights based on facial expressions is rare. The database of primary and middle school students is understandably smaller than Affect databases due to privacy and security concerns. Finally, temporal data is not considered in this study. Although peak emotions may not last long, affective states remain longer compared to basic emotions. For instance, confusion may peak during cognitive disequilibrium, but the state likely begins before the peak and persists even after the expression diminishes.

6 Conclusion

In this study, we developed a method leveraging the EmoNet model as a backbone, enhanced by a self-attention network, to predict learning-centered emotions accurately. This approach was applied to data collected from an OELE where we analyzed the facial expressions of 41 middle school students engaged in cognitive tasks such as reading, building causal maps, and assessing their knowledge through quizzes. Our method utilized the robust feature extraction capabilities of EmoNet, integrating it with a self-attention mechanism that focuses on relevant facial action units, thereby improving the specificity of emotion recognition in educational contexts. We documented each student's emotional response, linking these affective states to their cognitive activities within the OELE, thereby mapping how emotions like engagement, confusion, frustration, boredom, and delight fluctuate during different learning phases.

The results of our study were statistically significant, indicating clear differences in the frequency of emotional instances between high and low performers across various cognitive states. For instance, high performers demonstrated consistently high engagement, particularly noted in the Solution Construction phase with 1,066,640 instances of engagement. In contrast, low performers showed a marked increase in boredom during the Information Acquisition phase, with boredom nearly doubling compared to high performers, reaching 449,990 instances.

In the future, the emotion recognition model will be refined by retraining it across diverse learning environments and demographic settings to enhance its robustness and applicability, while also incorporating temporal analysis into the methodology to more accurately capture and analyze fleeting emotions. These improvements aim to reduce misclassifications and enhance the model's ability to distinguish between quickly changing emotional states, providing more reliable and nuanced insights into students' learning experiences.

Acknowledgments

This research was supported by the National Science Foundation AI Institute Grant No. DRL-2112635. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications* 17 (2023), 200171.

[2] Celestine E Akpanoko, Gautam Biswas, et al. 2024. The Interplay of Affective States and Cognitive Processes in an Open-Ended Learning Environment: A Case Study. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 873-880*. International Society of the Learning Sciences.

[3] Celestine E. Akpanoko, Ashwin T. S., Grayson Cordell, and Gautam Biswas. 2024. Investigating the Relations between Studentsâ€™ Affective States and the Coherence in their Activities in Open-Ended Learning Environments. In *Proceedings of the 17th International Conference on Educational Data Mining*, Benjamin PaaÃŸen and Carrie Demmans Epp (Eds.). International Educational Data Mining Society, Atlanta, Georgia, USA, 511–517. https://doi.org/10.5281/zenodo.12729872

[4] Celestine E. Akpanoko, Ashwin T. S., Grayson Cordell, and Gautam Biswas. 2024. Investigating the Relations between Studentsâ€™ Affective States and the Coherence in their Activities in Open-Ended Learning Environments. In *Proceedings of the 17th International Conference on Educational Data Mining*, Benjamin PaaÃŸen and Carrie Demmans Epp (Eds.). International Educational Data Mining Society, Atlanta, Georgia, USA, 511–517. https://doi.org/10.5281/zenodo.12729872

[5] TS Ashwin and Gautam Biswas. 2024. Identifying and Mitigating Algorithmic Bias in Student Emotional Analysis. In *International Conference on Artificial Intelligence in Education*. Springer, 89–103.

[6] TS Ashwin and Ram Mohana Reddy Guddeti. 2018. Unobtrusive students' engagement analysis in computer science laboratory using deep learning techniques. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 436–440.

[7] TS Ashwin and Ram Mohana Reddy Guddeti. 2020. Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems* 108 (2020), 334–348.

[8] TS Ashwin and Ram Mohana Reddy Guddeti. 2020. Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. *User Modeling and User-Adapted Interaction* 30, 5 (2020), 759–801.

[9] Roger Azevedo, Michelle Taub, and Nicholas V Mudrick. 2017. Understanding and reasoning about real-time cognitive, affective, and metacognitive processes to foster self-regulation with advanced learning technologies. In *Handbook of self-regulation of learning and performance*. Routledge, 254–270.

[10] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[11] Ryan S Baker, Jaclyn L Ocumpaugh, and JMAL Andres. 2020. BROMP quantitative field observations: A review. *Learning Science: Theory, Research, and Practice. New York, NY: McGraw-Hill* (2020).

[12] Mina Bishay, Kenneth Preston, Matthew Strafuss, Graham Page, Jay Turcot, and Mohammad Mavadati. 2023. Affdex 2.0: A real-time facial expression analysis toolkit. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[13] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences* 582 (2022), 593–617.

[14] Sidney D'Mello and Art Graesser. 2014. Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta psychologica* 151 (2014), 106–116.

[15] Sidney D'Mello, Art Graesser, et al. 2007. Monitoring affective trajectories during complex learning. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 29.

[16] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[17] Joyce Fonteles, Eduardo Davalos, T. S. Ashwin, Yike Zhang, Mengxi Zhou, Efrat Ayalon, Alicia Lane, Selena Steinberg, Gabriella Anton, Joshua Danish, Noel Enyedy, and Gautam Biswas. 2024. A First Step in Using Machine Learning Methods to Enhance Interaction Analysis for Embodied Learning Environments. In *Artificial Intelligence in Education*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 3–16.

[18] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885* (2016).

[19] Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. *Journal of Network and Computer Applications* 147 (2019), 102423.

[20] Geethu Miriam Jacob and Bjorn Stenger. 2021. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7680–7689.

[21] Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. 2023. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion* (2023), 102019.

[22] John S Kinnebrew, James R Segedy, and Gautam Biswas. 2014. Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition and learning* 9 (2014), 187–215.

[23] John S Kinnebrew, James R Segedy, and Gautam Biswas. 2017. Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. *IEEE Transactions on Learning Technologies* 10, 2 (2017), 140–153.

[24] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[25] Krittaya Leelawong and Gautam Biswas. 2008. Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education* 18, 3 (2008), 181–208.

[26] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* 13, 3 (2020), 1195–1215.

[27] Jennifer Dodorico McDonald. 2008. Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire* 1, 1 (2008), 1–19.

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

[29] Anabil Munshi, Ramkumar Rajendran, Jaclyn Ocumpaugh, Gautam Biswas, Ryan S Baker, and Luc Paquette. 2018. Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In *Proceedings of the 26th conference on user modeling, adaptation and personalization*. 131–138.

[30] Manuel Ninaus, Simon Greipl, Kristian Kiili, Antero Lindstedt, Stefan Huber, Elise Klein, Hans-Otto Karnath, and Korbinian Moeller. 2019. Increased emotional engagement in game-based learning–A machine learning approach on facial emotion detection data. *Computers & Education* 142 (2019), 103641.

[31] Jaclyn Ocumpaugh et al. 2015. Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences* 60 (2015).

[32] Reinhard Pekrun and Markus Bühner. 2014. Self-report measures of academic emotions. In *International handbook of emotions in education*. Routledge, 561–579.

[33] Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P Perry. 2002. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist* 37, 2 (2002), 91–105.

[34] Reinhard Pekrun and Elizabeth J Stephens. 2012. Academic emotions. (2012).

[35] Ramkumar Rajendran, Sridhar Iyer, and Sahana Murthy. 2019. Personalized Affective Feedback to Address Students' Frustration in ITS. *IEEE Transactions on Learning Technologies* 12, 1 (2019), 87–97. https://doi.org/10.1109/TLT.2018.2807447

[36] M Rashmi, TS Ashwin, and Ram Mohana Reddy Guddeti. 2021. Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus. *Multimedia Tools and Applications* 80, 2 (2021), 2907–2929.

[37] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[38] Andrey Savchenko. 2023. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *International Conference on Machine Learning*. PMLR, 30119–30129.

[39] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. 2022. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing* (2022). https://ieeexplore.ieee.org/document/9815154

[40] Archana Sharma and Vibhakar Mansotra. 2019. Deep learning based student emotion recognition from facial expressions in classrooms. *International Journal of Engineering and Advanced Technology* 8, 6 (2019), 4691–4699.

[41] Michelle Taub, Roger Azevedo, Ramkumar Rajendran, Elizabeth B Cloude, Gautam Biswas, and Megan J Price. 2021. How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system? *Learning and Instruction* 72 (2021), 101200.

[42] Roger Tourangeau and Ting Yan. 2007. Sensitive questions in surveys. *Psychological bulletin* 133, 5 (2007), 859.

[43] Ashwin TS and Ram Mohana Reddy Guddeti. 2020. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and information technologies* 25, 2 (2020), 1387–1415.

[44] Kathryn F Whitmore, Christie Angleton, Jennifer Pruitt, and Shauntá Miller-Crumes. 2019. Putting a focus on social-emotional and embodied learning with the visual learning analysis (VLA). *Early Childhood Education Journal* 47 (2019),

549–558.
[45] Philip H Winne. 2018. Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology* 88, 1 (2018), 9–20.
[46] Fielding I Winters, Jeffrey A Greene, and Claudine M Costich. 2008. Self-regulation of learning within computer-based learning environments: A critical analysis. *Educational psychology review* 20 (2008), 429–444.
[47] Yufeng Yin, Di Chang, Guoxian Song, Shen Sang, Tiancheng Zhi, Jing Liu, Linjie Luo, and Mohammad Soleymani. 2024. FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6099–6108.