Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Combining multimodal analyses of students' emotional and cognitive states to understand their learning behaviors

Ashwin T S[a*], Caitlin SNYDER[a], Celestine E. AKPANOKO[a], Srigowri M P[a] & Gautam BISWAS[a]

[a]*Computer Science, Vanderbilt University, USA*

*ashwindixit9@gmail.com

Abstract: The incorporation of technology into primary and secondary education has facilitated the creation of curricula that utilize computational tools for problem-solving. In Open-Ended Learning Environments (OELEs), students participate in learning-by-modeling activities that enhance their understanding of (Science, technology, engineering, and mathematics) STEM and computational concepts. This research presents an innovative multimodal emotion recognition approach that analyzes facial expressions and speech data to identify pertinent learning-centered emotions, such as engagement, delight, confusion, frustration, and boredom. Utilizing sophisticated machine learning algorithms, including High-Speed Face Emotion Recognition (HSEmotion) model for visual data and wav2vec 2.0 for auditory data, our method is refined with a modality verification step and a fusion layer for accurate emotion classification. The multimodal technique significantly increases emotion detection accuracy, with an overall accuracy of 87%, and an F1-score of 84%. The study also correlates these emotions with model building strategies in collaborative settings, with statistical analyses indicating distinct emotional patterns associated with effective and ineffective strategy use for tasks model construction and debugging tasks. These findings underscore the role of adaptive learning environments in fostering students' emotional and cognitive development.

Keywords: STEM Education, Collaborative Learning, Affective States, Learning-Centered Emotion, Open-ended Learning Environments, Facial Expressions, Multimodal Learning Analytics

1. Introduction

The increased focus on incorporating and leveraging technology into K-12 classrooms has been actualized by designing curricula where students develop and solve problems using computational tools (Wing, 2006). As there are innate connections between science, technology, engineering, and mathematics (STEM) and computing (Grover & Pea, 2018), learning-by-modeling tasks have been designed within open-ended learning environments (OELEs) to support students' conceptual knowledge construction of STEM and computing through authentic, real-world scientific problems (Hutchins, et al., 2020). The design goals of OELEs often target both conceptual understanding as well as the development of problem-solving skills that can be adapted and used in other problem spaces (Land, 2000). During learning-by-modeling, students develop and employ problem-solving skills such as abstraction, decomposition, and debugging (Grover and Pea, 2018; Hutchins, et al., 2020).

While these skills have been identified as important for task success and student learning, developing and applying them is innately complex. This is particularly true within the learning-by-modeling paradigm where students are applying these skills while constructing and applying conceptual knowledge in multiple domains (i.e., STEM and computing). As such, students are often assigned to work collaboratively to mitigate some of the difficulties associated with these complex tasks. Understanding how students use problem-solving strategies to solve such tasks can provide insights into their learning processes and help educators design interventions to support effective problem-solving (Järvelä, Nguyen, & Hadwin, 2023). Advances in technology enable the automatic understanding of problem-solving strategies using sequence mining, Markov chains, and other machine learning analytics. These computational methods can analyze students' interactions within OELEs to identify patterns and sequences of actions that characterize different problem-solving approaches. This automated analysis provides a scalable way to monitor and support student learning as they work in these learning environments (Järvelä, Nguyen, & Hadwin, 2023).

Emotions play a significant role in problem-solving and learning. Learning-centered emotions like confusion, frustration, and boredom can hinder progress, while engagement and delight can enhance it (D'Mello & Graesser, 2012). Understanding the impact of emotions on problem-solving is crucial for designing adaptive learning environments that provide timely interventions to maintain motivation and engagement (Jordan & Troth, 2021). Despite their importance, there is a gap in research on multimodal emotion recognition in educational data mining. Existing studies focus on basic emotions or valence differences in collaborative learning, with limited exploration of learning-centered emotions using multimodal approaches (Törmänen et al., 2023; Dindar et al., 2020). Nonverbal speech cues are underutilized in OELEs for emotion recognition due to their computer-based nature. In collaborative learning, however, speech is crucial for understanding learning processes through interaction. Therefore, our study integrates speech and image data to detect and analyze learning-centered emotions comprehensively. There is a critical need for novel methodologies to accurately detect learning-centered emotions in educational contexts. Traditional methods like manual observation and self-reports have biases that impact accuracy. Single modality (vision-based) models often fail in computer-based collaborative environments, where webcams can lose track of moving students (Järvelä, Nguyen, & Hadwin, 2023; Hutchins, et al., 2020). Our proposed multimodal approach leverages advances in machine learning and computer vision to provide a robust understanding of students' emotional states. The key contributions of the paper are:

1. *A Novel Multimodal Emotion Detection Methodology*: We introduce a methodology for detecting learning-centered emotions (engagement, frustration, confusion, delight, and boredom) in high school students using multimodal data from facial expressions and speech.
2. *Establishing the relations between Emotions and Strategies*: We analyze the relationship between detected emotions and effective/ineffective problem-solving strategies.

The paper is organized as follows: Section 2 reviews the relationship between affective states and problem-solving strategies. Next, we describe the emotion recognition framework. We then present our findings on emotions and strategies. Finally, we discuss implications for adaptive learning environments and future research.

2. Literature Review

Previous research has identified strategies in block-based environments, such as the leveraging of environmental data tools to evaluate computational models, multiple reviews of output and code, and forward reasoning i.e. examining the program line by line (Hutchins, et al., 2021; Kim, et al. 2018; McCauley, 2008). For example, Hutchins, et al (2021) identified the following strategies students employed when building computational models of scientific processes using a block-based environment: (1) *Depth-First*, i.e., multiple code construction actions without assessment actions, can corresponds to a lack of insight for breaking down a complex task into its subparts (Grover & Pea, 2018); (2) *Tinkering*, i.e., trying small changes in the blocks making up the executable model, can be used to gain some understanding of code prior to making changes; (3) *Multi-Visual Feedback*, represented by sequence of simulation executions, can represent a lack of understanding if the simulations were run in rapid succession; and (4) *Simulation-based Assessment*, which typically involves using tools, such as plots, to understand and analyze model behavior, and in past work has been observed to represent a decomposition process, i.e., a build and test behavior (Basu, et al, 2017; Hutchins, et al, 2021).

As students' emotional states impact their engagement and effectiveness during collaborative problem-solving, we aim to examine the relationship between their problem-solving strategies and affective states (Lazarus, 1982). Learning-centered emotions like confusion, frustration, boredom, engagement, and delight play a crucial role in socially shared regulated learning (SSRL) processes (Pekrun, 2006; Panadero & Järvelä, 2015; Xu & Lou, 2023). However, research has not adequately explored how these emotions relate to specific strategies used in building computational models. Understanding this relationship can provide deeper insights into how emotions influence learning behaviors and outcomes in collaborative settings (D'Mello et al., 2008). Current methods for emotion recognition in SSRL processes rely on self-reports or vision-based emotion recognition (Järvelä, Nguyen, & Hadwin, 2023; Nguyen et al., 2023; Xu & Lou, 2023).

State-of-the-art emotion recognition often relies on single modalities like vision or speech. Vision-based approaches use facial expressions from webcams, while speech-based methods analyze vocal cues (Ahmed, Al Aghbari, & Girija, 2023; Mejbri et al., 2022). In OELEs, webcams can lose track of faces if students move, and relying solely on speech is insufficient as students do not always verbalize their thoughts (Nguyen et al., 2023; Järvelä, Nguyen, & Hadwin, 2023; Emara et al., 2020). This necessitates a multimodal approach combining speech and image data for accurate emotion detection. Multimodal recognition is crucial in collaborative learning with rich interactions.

Initially, classifiers like Support Vector Machines and regression were used, but state-of-the-art methods now employ convolutional neural network (CNN) based models for facial expressions and Mel-frequency cepstral coefficients (MFCC) features for speech, analyzing cues like pitch, intensity, and frequency for valence and arousal. These modalities are combined using decision-level or feature-level fusion. Advanced methods use transformers and self-attention models, including bi-directional long short-term memory (BLSTM) networks, Deep Belief Networks (DBNs), and hierarchical networks (Geetha et al., 2024; Mittal et al., 2020; Tzirakis et al., 2017). However, these methods have not been widely explored for valence-arousal detection or learning-centered emotion classification.

Research on multimodal learning-centered emotion recognition in collaborative learning environments is lacking. Existing studies focus on basic emotions or valence differences but do not integrate multiple modalities to detect learning-centered emotions (Ahmed, Al Aghbari, & Girija, 2023). This study addresses this gap by combining speech and image data to recognize emotions like engagement, frustration, confusion, delight, and boredom. This multimodal approach aims to provide a more accurate understanding of students' emotional states during collaborative learning tasks.

3. Learning Environment and Data

In our curriculum, designed to support the synergistic learning of science and computing by leveraging the innate connections between these two domains (Grover & Pea 2018), students work in the Collaborative, Computational STEM (C2STEM) block-based learning environment (Hutchins, et al., 2020). In C2STEM, students create models of scientific phenomena and simulate the motion of objects using variables within the computational model. In the environment, students have access to an animation to view the objects' motion as well as variable inspection, graph, and table tools that are updated dynamically as the simulation runs. Students can leverage these tools to investigate the connection between scientific variables (e.g., the relationship between velocity and acceleration) as well as develop key computing practices such as debugging (Grover & Pea, 2018).

Twenty-four 10th grade high school students completed a 6-week kinematics curriculum in which they worked on 1D and 2D kinematic modules for two hours every week. The curriculum was designed as a sequence of computational modeling tasks with increasing complexity (Hutchins, et al., 2020). Each of the 1D and 2D modules consisted of a scaffolded instructional task, a model building task, a hands-on activity, and a complex challenge task. The instructional task was completed individually while the hands-on activity was completed in larger groups (~6 students). In this work, we analyze four tasks that were completed by students working in pairs: *1D model building task* (modeling the motion of a truck speeding up from rest to a speed limit and cruising at that speed limit before coming to a stop at a stop sign), *1D challenge task* (modeling the motion of an autonomous truck that adapts its speed whenever the car in front of it does), *2D model building task* (modeling the motion of a drone dropping a package off at a target), and *2D challenge task* (modeling the motion of a drone dropping off two packages at two different targets). Students were first divided into dyads randomly for the 1D module tasks and then assigned into different dyads for the 2D module tasks. Procedures for this study were approved by the University Institutional Review Board and included the collection of summative assessment data, logged actions in the C2STEM environment, and video and audio data collected using Open Broadcaster Software (OBS) and lapel microphones. Student actions were recorded in log files with timestamps and aligned with the video and audio data using system time stamps.

Data: 1,718 problem-solving strategy instances were extracted from the log data. Video data was collected at 30 fps, and audio was recorded at 572 kbps bit rate with 2 stereo channels and a 44.100 kHz sampling rate. Due to hardware errors, student absences, or accidental recording interruptions, we standardized the data and removed unwanted or corrupt files, resulting in a total of 51 dyad datasets over 6 weeks. Initially, there were 12 groups for 6 weeks, with 2 hours of data per week, expected to yield 60 dyad datasets, but the first week was used for explaining the study and preprocessing reduced it to 51 dyads. This preprocessing resulted in approximately 15 million image frames. The image frames were fed into Multi-task Cascaded Convolutional Networks (MTCNN) for face recognition and then passed to the valence-arousal prediction model. For audio, recordings ranged between 60 to 120 minutes per day over five weeks. Using Deepgram, a commercial speech-to-text Application Programming Interface (API) service, we generated utterances with start and end timestamps and transcripts. Some groups had a little over 400 segments, while others exceeded 1000 segments. The shortest segment lasted just a second, while the longest extended up to 17 seconds. Overall, we had approximately 39,335 utterances. These utterances were used to segment each large audio file into shorter segments, ranging from just a second to 17 seconds. Each audio segment was standardized to the consistent sampling rate and format mentioned above before being fed into the valence-arousal recognition model.

4. Methodology

In this paper, we aim to better understand the relationship between students' problem-solving skills and emotional states. Specifically, we study students' problem-solving skills through students' effective and ineffective construction and assessment strategies. We identified these strategies by processing the logged data utilizing a hierarchical task-oriented structure adapted from Emara et al. (2020) to abstract student actions such that the calculated action patterns can be mapped to students' problem-solving strategies. The logged actions were categorized into five abstracted categories: (1) *Build,* where students are adding blocks to the executable model; (2) *Adjust,* where students are editing blocks in the executable model; (3) *Draft*, where students are moving or editing blocks not connected to the executable model (similar to commenting code); (4) *Execute*, where students are executing the model; and (5) *Visualize,* where students are using the data tools or variable inspection tool.

Leveraging previous research that identified common computational modeling strategies (Hutchins, et al., 2021), we used action log patterns to identify the time intervals during which students employed construction strategies: (1) *Depth-first* (BUILD-> BUILD -> BUILD -> BUILD) and (2) *Tinkering* (PLAY -> ADJUST -> PLAY) and assessment strategies: (3) *Multi-visual Feedback* (PLAY -> PLAY -> PLAY) and (4) *Simulation-based Assessment* (VISUALIZE -> PLAY or PLAY -> VISUALIZE). As discussed in Section 2, *Tinkering* and *Simulation-based Assessment* have been identified as more effective strategies while *Depth-first* and *Multi-visual feedback* are ineffective strategies. This detailed analysis allows us to map students' problem-solving strategies and their effectiveness, providing insights into their relationship with students' affect and learning outcomes. The complete flow of the methodology is shown in Figure 1.

Learning-centered emotion annotation: While D'Mello et al. (2007) utilized Russell's circumplex model (Russell, 1980) to identify discrete emotions on a valence-arousal scale in the education domain, these methods have not been specifically applied to multimodal emotion recognition. However, several studies have used valence-arousal values obtained from image frames to map learning-centered emotions (Fonteles, 2024). We used this conversion to map the annotated emotions to valence-arousal values during our annotation process. We manually annotated 500 utterances and their corresponding image frames for each learning-centered emotion. These were randomly selected from various timestamps during the study. Additionally, we annotated 5,000 image frames that lacked accompanying audio and 500 utterances where the student's image was not visible on the webcam. These data were chosen from the data of 10 students, ensuring an equal division of gender (5 males and five females) and diverse demographic backgrounds, including 5 White Americans and 5 African Americans. Two different annotators independently annotated each emotion using facial expressions, and the inter-rater reliability, measured by Cohen's Kappa, was 0.89.

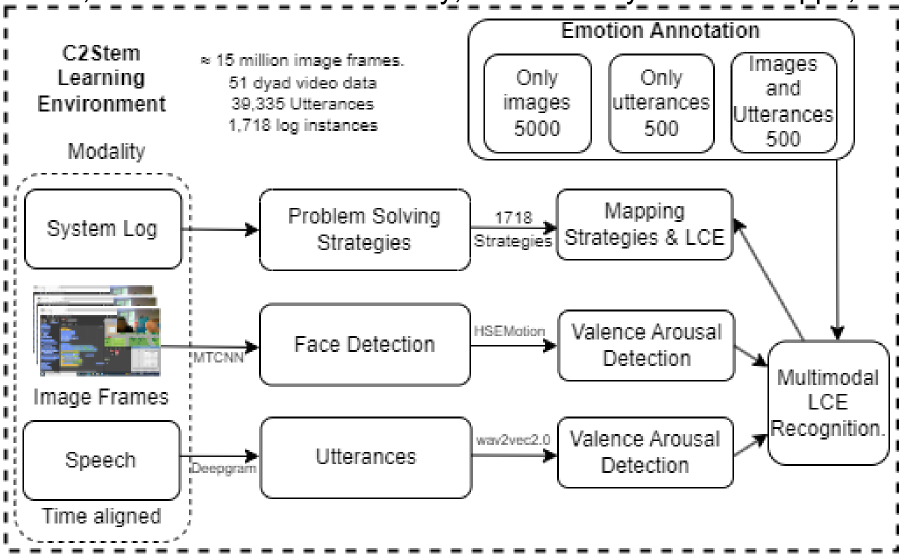

Figure 1: Overview of the complete methodology

Vision and Speech Emotion Recognition: Facial regions within images were detected using MTCNN. HSEmotion (High-Speed Face Emotion Recognition) is used in this study to predict the valence and arousal values (Fonteles, 2024). The HSEmotion architecture is trained for face identification leveraging the Visual Geometry Group Face2 (VGGFace2) dataset. The VGGFace2 dataset, comprising 3.31 million images across 9131 subjects (identities), presents diverse variations in pose, age, illumination, ethnicity, and profession. This dataset includes facial images of children within the age group relevant to our study's focus. Consequently, the models were trained and validated on images representative of the specific age group considered in this study. Similarly, for audio, we used the wav2vec 2.0 large robust (w2v2-L-robust-12) model, which outperforms state-of-the-art methods for valence and arousal recognition from speech. Facial regions within images were detected using MTCNN. HSEmotion is used in this study to predict the valence and arousal values. The HSEmotion architecture is trained for face identification leveraging the VGGFace2 dataset. The VGGFace2 dataset, comprising 3.31 million images across 9131 subjects (identities), presents diverse variations in pose, age, illumination, ethnicity, and profession. This dataset includes facial images of children within the age group relevant to our study's focus. Consequently, the models were trained and validated on images representative of the specific age group considered in this study. Similarly, for audio, we used the wav2vec 2.0 large robust (w2v2-L-robust-12) model (Wagner et al., 2023), which outperforms state-of-the-art methods for valence and arousal recognition from speech.

Multimodal Emotion Recognition: We denote the set of modalities as M= {image, speech}. The feature vectors for each modality are denoted as f_{image} and f_{speech}, respectively. We denote the set of predicted emotions as E= {confusion, frustration, delight, engagement, boredom}. The proxy feature vectors generated for image and speech modalities are represented by p_{image} and p_{speech}, respectively. Finally, we define an indicator function $I_e(f)$ that outputs either a vector of zeroes or ones of the same dimension as f, depending on the conditions of the function definition.

We present an overview of our multimodal perceived emotion recognition model. During training, we first extract feature vectors f_{image} and f_{speech} from raw inputs (video frames and audio segments, respectively). For image data, the HSEmotion model uses EfficientNet-B0 to process facial images of resolution 224 × 224, outputting embeddings of dimensionality

1280, or EfficientNet-B2 with 260 × 260 input images and 1408 output features. For speech data, we use the wav2vec2.0 model which processes raw waveforms normalized to have zero mean and unit variance, extracting features through a series of convolutional and transformer layers. The valence-arousal values obtained from HSEmotion for images and wav2vec 2.0 for audio were converted to learning-centered emotions using a predefined mapping function obtained from the annotation. These features are then passed through the modality check step to distinguish between effective and ineffectual signals, discarding the latter if any. The effective feature vectors are then processed through three deep-layered feed-forward neural network channels. Finally, we combine the modalities using a multiplicative fusion layer. During testing, the data point goes through the modality check step again, and if a modality is deemed ineffectual, we regenerate a proxy feature vector which is then passed to the network for emotion classification.

Modality Check Step: To enable perceived emotion recognition in real-world scenarios where sensor noise is inevitable, we introduce the Modality Check step, which filters ineffectual data. Previous studies in emotion prediction have shown that effective emotion recognition relies on the correlation between modality signals. We exploit this by using Canonical Correlation Analysis (CCA) to compute the correlation score, ρ, of every pair of input modalities. Given feature vectors f_{image} and f_{speech}, we compute the projective transformations, H_{image} and H_{speech}, and obtain projected vectors. The correlation score is calculated as (Equation 1):

$$\rho(f_{image}, f_{speech}) = \frac{\mathrm{cov}(H_{image} f_{image}, H_{speech} f_{speech})}{\sigma H_{image} f_{image} \ \sigma \ H_{speech} f_{speech}} \qquad \text{Equation (1)}$$

If the correlation score for a pair of modalities is below an empirically chosen threshold *τ*, the modality is considered ineffectual. The indicator function $I_e(f)$ is used to filter out these ineffectual features.

Regenerating Proxy Feature Vectors: When one or more modalities are deemed ineffectual at test time, we generate proxy feature vectors for these modalities. Generating exact feature vectors for missing modalities is challenging due to the non-linear relationship between modalities. However, by relaxing the non-linear constraint, we approximate the feature vectors using a linear algorithm. For instance, if the speech modality is corrupt, we regenerate a proxy speech vector p_{speech} using the effective face modality vector f_{image}. This involves preprocessing the inputs to construct bases from the observed face and speech vectors and applying a linear transformation to approximate the missing feature vector.

Multiplicative Modality Fusion: In our approach, we use a multiplicative fusion layer to combine the effective feature vectors from each modality. This fusion method explicitly boosts the stronger modalities in the combination network, enhancing the overall emotion classification accuracy. The modified loss function for the i-th modality is defined as:

$$c(y) = \prod_{i=1}^{M} (p(y)i)^{\frac{\beta}{M-1}} \log p(y)i \qquad \text{Equation (2)}$$

where y is the true class label, MM is the number of modalities, β is a hyperparameter, and $p(y)_i$ is the prediction for class yy given by the network for the i-th modality (Equation 2).

This methodology leverages robust feature extraction techniques from both image and speech modalities, applies effective noise filtering through the modality check step, and combines the modalities using advanced fusion techniques to achieve accurate emotion recognition.

5. Results and Inference

<u>*A Novel Multimodal Emotion Detection Methodology*</u>*:* Since the annotated data is limited, we performed data augmentation, as mentioned in (TS and Guddeti, 2020), this increased the annotated data (mentioned in Section 4) by 10-fold. The dataset was then split into training (70%), validation (10%), and testing (20%) sets. We used the Adam optimizer with a learning rate of 0.01 to train our models. The training process utilized a batch size of 256 and was conducted over 500 epochs. The HSEmotion model achieved an accuracy of 91%, a precision of 89%, a recall of 86%, and an F1-score of 88% for valence-arousal prediction. The wav2vec2.0 model achieved an accuracy of 81%, a precision of 79%, a recall of 76%, and an F1-score of 76% for valence-arousal prediction. We performed student-independent cross-

validation, and the results are in line with existing results that use HSEmotion for student data within the OELE learning environment.

The overall results of our fusion model (Learning-centered Multimodal Emotion Recognition) show significant improvements over single-modality approaches. The overall accuracy for emotion recognition using the fused modalities was 87%, with a precision of 85%, a recall of 83%, and an F1-score of 84%. These results demonstrate the effectiveness of our fusion approach in improving the robustness and accuracy of emotion recognition in collaborative learning environments.

Table 1. *Ablation Study of proposed Multimodal Emotion Recognition*

Method	Accuracy	Precision	Recall	F1-Score
Image Only	81%	79%	76%	77%
Speech Only	75%	73%	71%	72%
Fusion without Modality Check	85%	83%	80%	82%
Full Model	87%	85%	83%	84%

The CCA analysis results indicated strong correlations between the modalities, validating the effectiveness of our modality check step. The mean correlation score for effective modalities was significantly higher than the threshold (τ = 0.7), while the ineffectual modalities showed correlation scores well below the threshold. This clear distinction allowed us to filter out noisy data effectively.

Ablation Study: Table 1 highlights the modality check step's importance and our fusion strategy's effectiveness. The full model, which includes all components, outperforms the other configurations, demonstrating the value of our approach in leveraging both image and speech data for accurate emotion recognition.

Establishing the relations between Emotions and Strategies: For the second part of the results where we mapped the strategies to learning-centered emotions, we considered the entire data and did not consider the data augmentation. We ran the multimodal emotion recognition with the entire data, including cases with only utterances and both utterances and image frames during the period of utterances. The utterances-based multimodal data suggests that effective construction strategies, such as Tinkering, lead to higher engagement (14983 instances) and more delight (180 instances) compared to ineffective construction strategies, such as Depth-first, which show higher confusion (1487 instances) and frustration (319 instances).

Effective assessment strategies, like Simulation-based Assessment, also show higher engagement (6881 instances) and moderate levels of delight (132 instances) compared to ineffective assessment strategies, like Multi-visual Feedback, which have significant engagement (6508 instances) but higher levels of confusion (1026 instances). The overall pattern indicates that effective strategies are associated with higher engagement and delight, whereas ineffective strategies correlate more with confusion and frustration.

When considering only image-based emotion classification, which was 67% of the time in the entire data (59% with only image frames and 8% with one student out of frame), we observed 54% engagement, 14% confusion, 10% frustration, 5% delight, and 17% boredom.

Effective Construction (Tinkering) shows the highest levels of engagement (60%) and a moderate amount of delight (8%). Confusion (12%) and boredom (10%) are relatively low, indicating that students are actively engaged and positively affected by this strategy. In contrast, Ineffective Construction (Depth-first) exhibits a significantly lower engagement level (40%) and higher boredom (30%) compared to Tinkering. Confusion (15%) is also slightly higher, while delight (5%) is lower, suggesting that students are less engaged and more likely to be bored when using less effective construction strategies. Effective Assessment (Simulation-based Assessment) also demonstrates high engagement (55%) and moderate levels of delight (7%). Confusion (14%) and boredom (14%) are balanced but relatively low, indicating a generally positive impact on student emotions. Conversely, Ineffective Assessment (Multi-visual Feedback) presents lower engagement (47%) and higher boredom (24%) compared to effective assessment strategies. Confusion (15%) remains consistent with ineffective construction, while delight (4%) is the lowest among all strategies.

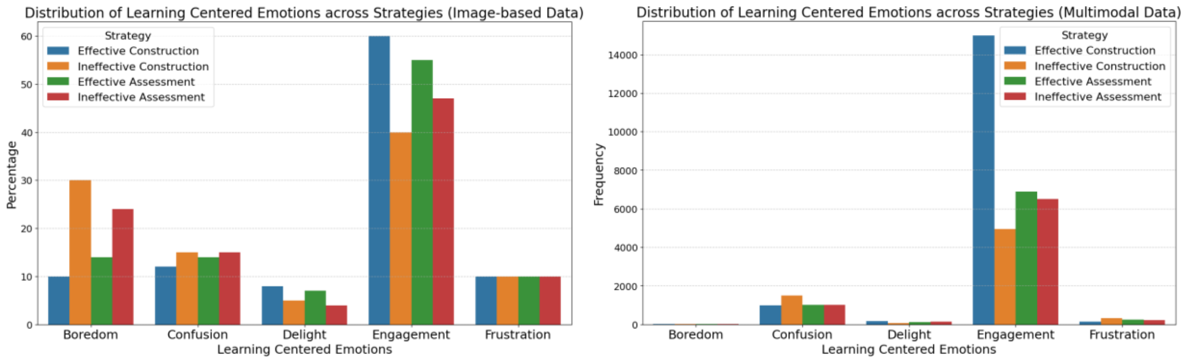

Figure 2. Distribution of learning centered emotions across strategies

Our findings indicate a strong correlation between learning-centered emotions and students' problem-solving strategies. Effective strategies such as Tinkering and Simulation-based Assessment are associated with higher levels of engagement and delight, while ineffective strategies like Depth-first and Multi-visual Feedback are linked with increased confusion and frustration. This correlation highlights the critical role of emotions in influencing problem-solving effectiveness, supporting the need for adaptive learning environments that can respond to students' emotional states in real-time.

Discussion: We observe that boredom is significantly detected through images. When students are bored, they tend to speak less, displaying low valence and low arousal facial expressions. Confusion, characterized by action units like brow furrow and top-left quadrant valence-arousal values, is also more frequently detected in images than in utterances. The valence-arousal values distributed across all four quadrants align with Russell's circumplex model. For instance, delight had a valence of 0.9 and arousal of 0.8, while boredom had a valence of 0.3 and arousal of 0.2, showing clear distinctions. Engagement was mostly observed when both valence and arousal values were above 0.5, whereas confusion and frustration were identified when arousal was above 0.5 and valence was below 0.5. Several instances where students discussed topics had utterances that fell close to neutral; however, since we did not annotate for neutral, these instances were classified as either engagement or confusion. The distribution of valence-arousal values for each learning-centered emotion was more condensed in images compared to speech values. Additionally, frustration audio cues such as frequency, tone, and pitch were quite clear, making frustration detected in utterances more accurate than in facial expressions.

We conducted a Chi-square test of independence to examine the overall differences between effective and ineffective strategies, specifically focusing on the four strategies and the distribution of learning-centered emotions. The results revealed significant variances in confusion and frustration between effective and ineffective strategies (Chi-square = 145.59, $p < 0.001$). Additionally, for boredom, we observed variations across categories, with the highest discrepancies noted (Chi-square = 124.22, $p < 0.01$). Delight showed consistent significant differences between the strategies (Chi-square = 193.45, $p < 0.01$).

In ineffective strategies, we found that boredom was more prevalent than frustration. Students tended to become less frustrated and more bored, often moving from a state of frustration to boredom and remaining in that state. In contrast, for effective strategies, while students did experience frustration, they quickly transitioned out of boredom and resumed problem-solving. This pattern was confirmed through manual verification of several instances, where we observed that students engaged in effective strategies were able to recover from boredom more rapidly and continue their tasks.

Statistically, the analysis showed that confusion and frustration had significant differences between effective and ineffective strategies, indicating that these emotions are critical markers of the strategies' effectiveness. Boredom also varied significantly across different strategies, reflecting its role in disengagement and its higher prevalence in ineffective strategies. Delight, on the other hand, consistently differed, with effective strategies showing higher levels of this positive emotion. The analysis of Figure 2 further elucidates the differences in emotional states between effective and ineffective strategies. The data indicates that effective strategies not only improve problem-solving outcomes but also foster more positive emotional experiences, thereby enhancing overall learning. These findings suggest that adaptive learning environments should prioritize both cognitive and emotional support to optimize student outcomes.

Limitations: This study has several limitations. Firstly, sentiment analysis from the transcribed text was not performed; we only used non-verbal cues from image and audio data. Although we checked the transcribed data for one entire week and observed that most of the content was related to the topic of study, incorporating text-based sentiment analysis could provide additional insights. Secondly, individual-based, gender-based, or performance-based analyses were not conducted. Such analyses require person re-identification and speaker diarization within each webcam video dyad. Additionally, no fine-tuning was performed on the valence-arousal models for image and speech data, which might have improved the accuracy of emotion recognition. Lastly, the number of students considered in this study is relatively small and belongs to a single demographic group, which limits the generalizability of the findings.

6. Conclusion

This study demonstrates the effectiveness of a multimodal approach for detecting learning-centered emotions in a collaborative learning environment. By integrating facial expression and speech data, we achieved higher emotion recognition accuracy than single-modality approaches. The HSEmotion and wav2vec 2.0 models, enhanced by the modality check step and multiplicative fusion layer, proved effective in identifying emotions such as engagement, frustration, confusion, delight, and boredom. Our findings show that effective problem-solving strategies are associated with higher engagement and delight, while ineffective strategies correlate more with confusion and boredom. Statistical analysis confirmed significant differences in emotion distribution between effective and ineffective strategies, underscoring the importance of adaptive learning environments that respond to students' emotional states. In the future, we plan to address some of the limitations that are already mentioned and explore different collaborative learning environments with varying sizes of groups.

Acknowledgements

This work was supported by the National Science Foundation (NSF) AI Institute Grant No. DRL-2112635 and NSF Eager Award No. 2327708. 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

Ahmed, N., Al Aghbari, Z., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17, 200171.

Basu, S., Biswas, G., Kinnebrew, J.S. (2017). Learner modeling for adaptive scaffolding in a Computational Thinking-based science learning environment. User Modeling and User-Adapted Interaction, 27(1), 5-53.

D'Mello, S., & Graesser, A. (2007). Monitoring affective trajectories during complex learning. In Proceedings of the annual meeting of the cognitive science society (Vol. 29, No. 29).

D’Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., ... & Graesser, A. (2008, June). AutoTutor detects and responds to learners' affective and cognitive states. In Workshop on emotional and cognitive issues at the international conference on intelligent tutoring systems (pp. 306-308).

Dindar, M., Järvelä, S., Ahola, S., Huang, X., & Zhao, G. (2020). Leaders and followers identified by emotional mimicry during collaborative learning: A facial expression recognition study on emotional valence. IEEE Transactions on Affective Computing, 13(3), 1390-1400.

Emara, M., Hutchins, N. M., Grover, S., Snyder, C., & Biswas, G. (2021). Examining Student Regulation of Collaborative, Computational, Problem-Solving Processes in Open-Ended Learning Environments. Journal of Learning Analytics, 8(1), 49-74.

Geetha, A. V., Mala, T., Priyanka, D., & Uma, E. (2024). Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions. Information Fusion, 105, 102218.

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. Computer science education: Perspectives on teaching and learning in school, 19(1), 19-38.

Hutchins, N. M., Snyder, C., Emara, M., Grover, S., & Biswas, G. (2021). Analyzing debugging processes during collaborative, computational modeling in science. In Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021. International Society of the Learning Sciences.

Hutchins, N. M., Biswas, G., Maróti, M., Lédeczi, Á., Grover, S., Wolf, R., ... & McElhaney, K. (2020). C2STEM: A system for synergistic learning of physics and computational thinking. Journal of Science Education and Technology, 29(1), 83-100.

J. A. Russell. A circumplex model of affect. Journal of personality and social psychology, 39(6):1161, 1980.

Järvelä, Sanna, Dragan Gašević, Tapio Seppänen, Mykola Pechenizkiy, and Paul A. Kirschner. "Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning." British Journal of Educational Technology 51, no. 6 (2020): 2391-2406.

Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. British Journal of Educational Technology, 54(5), 1057-1076.

Jordan, P. J., & Troth, A. C. (2021). Managing emotions during team problem solving: Emotional intelligence and conflict resolution. In Emotion and Performance (pp. 195-218). CRC Press.

Fonteles, J., Davalos, E., Zhang, Y., Zhou, M., Ayalon, E., Lane, A., ... & Biswas, G. (2024). A First Step in Using Machine Learning Methods to Enhance Interaction Analysis for Embodied Learning Environments. arXiv preprint arXiv:2405.06203.

Kim, C., Yuan, J., Vasconcelos, L., Shin, M., & Hill, R. B. (2018). Debugging during block-based programming. Instructional Science, 46(5), 767-787.

Land, S. M. (2000). Cognitive requirements for learning with open-ended learning environments. Educational Technology Research and Development, 48(3), 61-78.

Lazarus, R. S. (1982). Thoughts on the relations between emotion and cognition. American psychologist, 37(9), 1019.

McCauley, R., Fitzgerald, S., Lewandowski, G., Murphy, L., Simon, B., Thomas, L., & Zander, C. (2008). Debugging: a review of the literature from an educational perspective. Computer Science Education, 18(2), 67-92.

Mejbri, N., Essalmi, F., Jemni, M., & Alyoubi, B. A. (2022). Trends in the use of affective computing in e-learning environments. Education and Information Technologies, 1-23.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020, April). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 02, pp. 1359-1367).

Nguyen, A., Järvelä, S., Rosé, C., Järvenoja, H., & Malmberg, J. (2023). Examining socially shared regulation and shared physiological arousal events with multimodal learning analytics. British Journal of Educational Technology, 54(1), 293-312.

Panadero, E., & Järvelä, S. (2015). Socially shared regulation of learning: A review. European Psychologist.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational psychology review, 18, 315-341.

Törmänen, T., Järvenoja, H., Saqr, M., Malmberg, J., & Järvelä, S. (2023). Affective states and regulation of learning during socio-emotional interactions in secondary school collaborative groups. British Journal of Educational Psychology, 93, 48-70.

TS, A., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Education and information technologies, 25(2), 1387-1415.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing, 11(8), 1301-1309.

Wing, J. M. (2006). Computational thinking. Communications of the ACM, 49(3), 33-35.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Xu, W., & Lou, Y. F. (2023). Changes in the socially shared regulation, academic emotions, and product performance in venue-based collaborative learning. Active Learning in Higher Education, 14697874231167331.