Effects of Modalities in Detecting Behavioral Engagement in Collaborative Game-Based Learning

Fahmid Morshed Fahid
North Carolina State University
ffahid@ncsu.edu

Seung Lee
North Carolina State University
sylee@ncsu.edu

Bradford Mott
North Carolina State University
bwmott@ncsu.edu

Jessica Vandenberg
North Carolina State University
jvanden2@ncsu.edu

Halim Acosta
North Carolina State University
hacosta@ncsu.edu

Thomas Brush
Indiana University
tbrush@indiana.edu

Krista Glazewski
Indiana University
glaze@indiana.edu

Cindy Hmelo-Silver
Indiana University
chmelosi@indiana.edu

James Lester
North Carolina State University
lester@ncsu.edu

ABSTRACT

Collaborative game-based learning environments have significant potential for creating effective and engaging group learning experiences. These environments offer rich interactions between small groups of students by embedding collaborative problem solving within immersive virtual worlds. Students often share information, ask questions, negotiate, and construct explanations between themselves towards solving a common goal. However, students sometimes disengage from the learning activities, and due to the nature of collaboration, their disengagement can propagate and negatively impact others within the group. From a teacher's perspective, it can be challenging to identify disengaged students within different groups in a classroom as they need to spend a significant amount of time orchestrating the classroom. Prior work has explored automated frameworks for identifying behavioral disengagement. However, most prior work relies on a single modality for identifying disengagement. In this work, we investigate the effects of using multiple modalities to detect disengagement behaviors of students in a collaborative game-based learning environment. For that, we utilized facial video recordings and group chat messages of 26 middle school students while they were interacting with Crystal Island: EcoJourneys, a game-based learning environment for ecosystem science. Our study shows that the predictive accuracy of a unimodal model heavily relies on the modality of the ground truth, whereas multimodal models surpass the unimodal models, trading resources for accuracy. Our findings can benefit future researchers in designing behavioral engagement detection frameworks for assisting teachers in using collaborative game-based learning within their classrooms.

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576079

CCS CONCEPTS

• **Applied computing** → Education; Collaborative learning; Education; Interactive learning environments; • **Computing methodologies** → Machine learning.

KEYWORDS

Multimodal learning analytics, Collaborative game-based learning, Behavioral engagement, K-12 education

ACM Reference Format:
Fahmid Morshed Fahid, Seung Lee, Bradford Mott, Jessica Vandenberg, Halim Acosta, Thomas Brush, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. 2023. Effects of Modalities in Detecting Behavioral Engagement in Collaborative Game-Based Learning. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576050.3576079

1 INTRODUCTION

Recent years have seen significant growth in game-based learning environments for K-12 education [38]. These types of environments embed curricular content in gameplay to enhance students' learning experience. Studies show that game-based learning environments can increase motivation and engagement and promote positive cognitive and affective outcomes in students [13, 14]. A promising addition to the field is collaborative game-based learning environments where the gameplay elements are specifically designed to incorporate collaborative problem solving by introducing group goals, integrating group chat, and promoting group progressions [15, 16, 27]. When engaging in collaborative game-based learning, students interact with each other in groups to make progress, help others, negotiate, provide feedback, share knowledge, discuss strategies, and overcome obstacles as a group. By providing the benefits of collaborative learning, these environments can potentially increase engagement in learning. However, students sometimes disengage from productive behaviors, which may negatively affect the group and undermine the benefits of collaborative learning [18].

In game-based learning environments, disengagement might occur from time to time during the learning process. The definition of disengagement varies wildly in the literature [1]. In our

case, we refer to disengagement as behaviors that disrupt or impede learning flow [28], observed through in-game chat or facial expressions. Such behaviors have been discovered to be associated with boredom, frustration, or mind-wandering and can negatively affect learning outcomes [10, 35]. The impact is even higher in a collaborative learning space as students' behavioral disengagement can distract other students within the group leading to ineffective discussion [6], starting disruptive talk [25], or demotivating others in the classroom, impeding the learning process and engendering negative attitudes within the group and in the classroom [23]. In a classroom environment, teachers are often busy orchestrating the learning session, and as such, it is often difficult to identify individual students that are disengaged from the learning activity [26]. In the context of a collaborative game-based learning environment, detecting disengagement behaviors is even more difficult, as the disengagement can happen within and outside the game, and what constitutes disengagement behavior is often dependent on the context and modality of the learning environment [22, 29]. For example, in a game about aquatic ecosystems, chatting about fish would be relevant to the learning activities, but chatting about airplanes would likely be off-topic and, therefore, indicative of disengagement behavior. Similarly, speaking to a student who belongs to the same group may be reasonable, but speaking to a student who belongs to a different group may indicate disengagement. Limited work has explored methods for automatically identifying disengagement behaviors in collaborative game-based learning environments. Most prior work either relies on a single modality [6, 19, 25] or uses multimodal data streams while only looking at a single data stream for the ground truths [10, 33, 36].

In this paper, we have used group chat logs and facial video recordings to examine the efficacy of various modalities in identifying disengagement behaviors in collaborative game-based learning environments across multiple data streams, leveraging a multimodal disengagement detection framework from our prior work [12]. For our study, we captured middle school students' chat messages and facial recordings while they interacted with a collaborative game-based learning environment, Crystal Island: EcoJourneys. We separately labeled the chat and facial recording data to create distinct sets of ground truths for disengagement behaviors and investigated the following research questions:

- Do performance of predictive models using unimodal features vary by the modalities of the ground truths?
- Do models using multimodal features outperform models using unimodal features?

Results show that multimodal models incorporating both chat and facial features can achieve higher levels of predictive accuracy when automatically detecting disengagement behaviors among students compared to unimodal baselines, irrespective of the data stream used to define disengagement behaviors. Results also suggest that unimodal features are only relevant for detecting disengagement behaviors from the same data stream, are often insufficient for detecting disengagement behaviors from another data stream, and in some cases, multimodal features may not be necessary. Our findings can help inform decisions of collaborative game-based learning environments, provide insight into multimodal frameworks for detecting behavioral disengagement, support teachers in classroom orchestration, and provide opportunities to adaptively scaffold students to improve engagement in their learning process.

2 RELATED WORK

Game-based learning environments have been shown to have positive effects on learning. Greipl et al. [13] outlined the benefits of game-based learning and proposed a three-dimensional framework based on cognitive, emotional, and social factors. According to the study, game-based learning environments are excellent at complementing and enhancing traditional learning. To gain insights into teachers' perceptions of game-based learning in the classroom, Huizenga et al. [14] conducted semi-structured interviews with 43 secondary education teachers and found that game-based learning environments positively impact student engagement, make the students competitive, and thus motivate them in learning content and knowledge, and positively influences their learning outcomes. Similar positive effects of game-based learning have been seen in other studies [38]. Collaborative tools are often integrated in such game-based learning environments to enhance group collaboration, internal combination, and collaborative problem solving towards solving collaborative goals. Recent meta-analyses show that students in collaborative learning with a game-based learning environment produce more positive effects, in terms of knowledge acquisition, in terms of positive attitude and motivation, in terms of self-satisfaction, and self-efficacy, than individual learning [7, 15]. De Jesus and Silveira developed a framework for enhancing students' computational thinking skills in a collaborative game-based environment and show that their method can stimulate student interactions and problem-solving strategies [16]. Saleh et al. showed that collaborative game-based learning could promote students' knowledge acquisition and negotiation by leveraging common communication mediums in middle school [27].

Disengagement behaviors in learning often play an important role in learning outcomes [35]. Studies show that behaviors associated with positively valenced emotions (i.e., flow) are often associated with improved learning outcomes and engagement [24]. Similarly, negative behaviors associated with emotions like boredom or frustration often result in disengagement or disinterest [4]. Under the lens of activity theory, Maimaiti et al. [20] examined how student disengagement is affected by student-teacher interactions in a video conferencing scenario and suggested implementing more opportunities for student-student interactions, designing methods to reduce daydreaming (mind-wandering), and providing more incentives for online interactions will help to reduce the intensity of student disengagement. Student affective states can be leveraged to understand students' choice to disengage as well as to engage in on-task conversation. D'Mello et al. [10] found that students' behavioral posture (leaning forward or backward) is a direct predictor of their engagement in the learning activity. Another study showed that students' engagement could be observed by their interaction with the system (like keystrokes) [2]. Similarly, Baker et al. [11] examined the affective states that precede, co-occur, and follow student disengagement behavior. Findings suggest that bored students are more likely to disengage and are subsequently not likely to be bored in the next observation. They also found that frustrated students tend not to become disengaged, and disengagement behaviors are unlikely to co-occur with frustration.

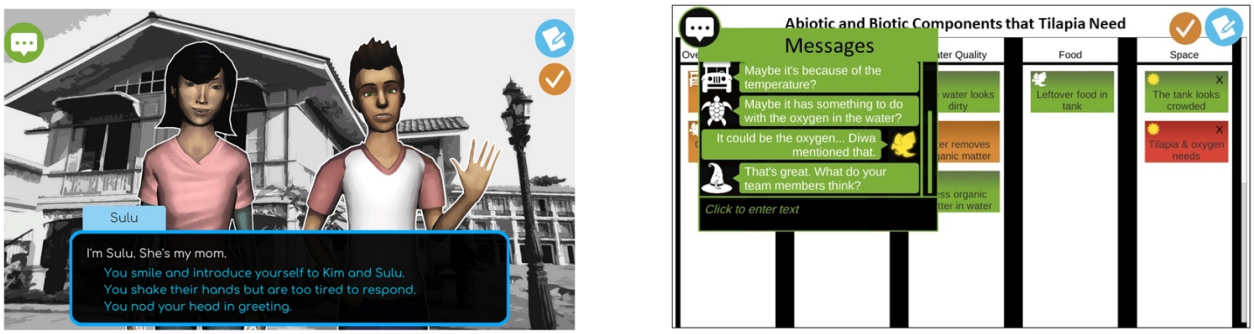

Figure 1: Students in-game activities while (left) interacting with NPCs and (right) chatting during white board interactions in the Crystal Island: EcoJourneys collaborative game-based learning environment.

Recent studies have used different modalities to identify students' learning behaviors automatically. Nikiforos et al. [23] used students' conversation data to explore the automatic detection of aggressive behavior (i.e., bullying) in two K-12 computer-supported collaborative learning environments. In their study, they found that using students' text responses as unigrams on shallow neural networks can outperform traditional machine learning models. More recently, deep learning-based embeddings have been used in students' group chat data to detect off-task conversation (disengagement behaviors) in a collaborative learning environment [6]. They found that Long Short-Term Memory-based models with word embeddings learned from language models like ELMo or BERT are suitable for finding off-task behaviors in group chats. Another study by Park et al. [25] used a user-aware attention mechanism in neural networks to detect disruptive talk in multi-party dialogue of middle school students using a game-based learning environment. Facial video recording data has been used to understand student behaviors, such as moments of arousal in collaborative learning [21]. Lee et al. [19] used facial action units from videos of students to identify students' disengagement behaviors (mind-wandering) in online learning. Bixler and D'Mello [5] used eye trackers for gaze information to detect students' disengagement behaviors in a learning activity. Other modalities that have been studied include audio, gesture, and motion to detect different types of behavioral disengagement among students [30].

Multimodal learning analytics can leverage features from different modalities to identify key aspects of learning behaviors during game-based learning. Recent studies show that multimodal fusion of data can improve over unimodal data and can help to better understand the complex learning processes that students engage in during game-based learning [22, 29]. Sharma et al. employed hidden Markov models in game-based learning environments to predict students' effortful behavior using game logs, physiological data, and self-assessment tests [31]. Worsley and Blikstein [37] used multiple data streams (human-annotated video data, automated annotation of gesture, audio, and bio-physiological data) to show that multimodal learning analytics for student behaviors can vary wildly based on different modalities and can have a large impact on the outcomes. Multimodal learning analytics have also been used to find productive engagement and disengagement behaviors in the context of collaborative problem solving, showing a possible alternative to purely qualitative or machine learning approaches [36]. But limited work has used multimodal data in collaborative game-based learning to understand the impact of modalities on predicting student disengagement behaviors. In this work, we utilized multiple data streams to observe, identify and compare student disengagement behaviors and introduced a framework that leverages these modalities to detect behavioral disengagement of middle school students in a collaborative game-based learning environment.

3 METHODS

This section describes the game-based learning environment we used, our study design and data collection, our annotation process of creating multiple sets of ground truths capturing different aspects of disengagement behaviors, their synchronization process, and finally, our framework for analysis.

3.1 Collaborative Game-Based Learning Environment

For our study, we used Crystal Island: EcoJourneys, a collaborative game-based learning environment, designed to teach ecosystem science to middle school students. In the game, students in groups of three or four are placed on a remote island where fish on the island are getting sick. The students are tasked to investigate the cause of such sudden sickness as a group by talking to different non-playable characters (NPCs) around the island, reading relevant materials found around the island, observing symptoms by roaming around, collecting samples of different elements, and taking notes of relevant facts. In the game, students can use a virtual school research center (game location) to investigate and analyze different samples that they have collected and find relevant evidence. Throughout the journey, the students can discuss with each other, share information, negotiate, and help each other by using an in-game chat tool. The chat window is always available to the students, irrespective of their current location in the game. The students also use a collaborative whiteboard to share and organize their notes and evidence into different categories to narrow down the cause of the sickness (see Figure 1). The whiteboard uses a voting mechanic where all the students within the group are required to agree on the current organization of the evidence and notes. The game ends

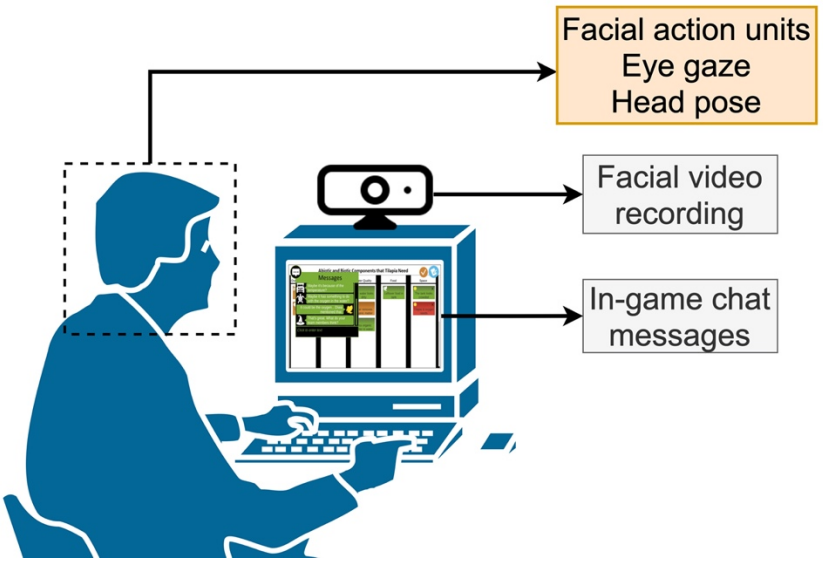

Figure 2: Collecting multiple data streams while students interact with the learning environment.

when all the students in the group agree on the cause of the sickness. Typically, it takes approximately 2-3 hours to complete the game as a group.

3.2 Study Design and Dataset

To understand students' disengagement behaviors in collaborative game-based learning environments, we conducted an IRB-approved study with 28 middle school students, among which, a total of 26 students assented to participate (along with a signed consent form from their parents) in the study and completed all the study activities (18 males and 8 females). The students consisted of sixth graders (11 students) and seventh graders (15 students). All students were aged between 11 to 13 years old (M=12.08, SD=0.75).

The students interacted with the game in groups of three or four (total 7 groups) for three hours divided into two sessions (1.5 hours each). Each group was accompanied by a facilitator who assisted students during their gameplay by communicating with them via the in-game chat tool. The role of the facilitators was to ensure that each group was making progress by providing hints, asking reflection questions, and providing positive feedback. The facilitators also intervened when conversations within the group were unproductive. Within each group, students assumed different roles in the game for solving the ecosystem science problem. To ensure maximum communication occurred through the in-game chat, students within the same group were physically seated apart from each other, but in the same classroom. Throughout the learning session, we collected facial video recordings of each student using webcams and in-game chat messages using game trace logs (see Figure 2). A total of approximately 48 hours of video recordings were collected during the study after removing segments with technical issues (such as the learning session being interrupted due to technical issues). On average, 111 minutes of video were recorded per student (M=111 minutes, SD=29 minutes, Min=57 minutes, Max=140 minutes). On the other hand, a total of 3,650 chat messages were collected after removing facilitators' chat messages. An average of 140 chat messages were sent by individual students (M=140.38, SD=101.54, Min=24, Max=438), and on average, 521 chat messages were sent by individual groups (M=521.43, SD=224.33, Min=272, Max=905).

3.3 Annotation of Ground Truths

We formulated our multimodal disengagement detection as a supervised binary classification task. In our study, we wanted to capture students' disengagement from multiple perspectives as a single data stream may not always reflect the complete story [29]. For example, looking at engagement using only chat messages shows partial truth of the situation, or similarly, looking at engagement using video recordings does not show if the student is truly engaged in the game or not. To achieve a holistic understanding of disengagement behaviors, we tagged group chat messages and segments of video recordings separately as *engaged* and *disengaged* and merged them together using different heuristics.

All of our chat messages were divided into four categories, namely, (1) *content related*, (2) *task related*, (3) *socio-emotional related*, and (4) *others* by two raters[1]. Looking into each type of chat, we found that *content-related and task-related* chat messages show high engagement in the learning process, whereas chat messages marked as *others* are typically off-topic conversations or unrelated discussions. Therefore, we marked all messages that are *content related* or *task related* as *engaged* and *others* as *disengaged*. The *socio-emotional* messages were a mixed bag containing messages

[1]The complete set of chat messages contained a total of 8,938 messages that combined data from another study with a similar setting. Two raters tagged the combined data with Cohen's Kappa (inter-rater reliability) of 0.925, indicating very strong agreement. We are using a subset of the combined data that contains a total of 3,650 chat messages sent by students that also included facial video recordings of the students. The same subset was also used in [27].

Table 1: Engaged and disengaged behaviors across two modalities and their coding scheme.

Modality	Coding Scheme
Group Chat Message	**Engaged (1):** messages related to content or task or socio-emotional messages that are positive and related to learning or task. For example, "*where would organic matter go?*", or "*kmon guys we got this*". **Disengaged (0):** messages unrelated to the learning activity or are spams or socio-emotional messages that are negative or not related to the learning or task. For example, "*you're a wizard Harry*", or "*REEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE*".
Facial Video Recording	**Engaged (1):** within a 10-second window, showing engaged behaviors (looking at the monitor, keyboard or taking notes, leaning forward, focused on monitor, etc.) for more than 6 seconds. **Disengaged (0):** within a 10-second window, showing disengaged behaviors (talking to someone, distracted, looking around, leaning back, sleeping, eating, playing with toys, using a mobile phone, etc.) for at least 4 seconds.

that positively impact the learning (such as motivating some teammates) as well as messages that negatively impact the learning (such as trying to guess the identity of each other). Prior studies also found that positive socio-emotional interactions can lead to positive engagement, whereas negative socio-emotional messages tend to disrupt learning [17]. Following that, two raters (first and second authors) separately labeled these *socio-emotional* messages as *engaged* or *disengaged* by looking into the context of such messages and identifying messages that positively correlate with the learning (*engaged*) and messages that hamper the learning process (*disengaged*). With 25% overlapping data, the two raters achieved a kappa of 0.716, showing significant agreement. Finally, the first author tagged the rest of the messages following the agreed-upon coding scheme (see Table 1). In our dataset, a total of 2,584 (70.79) chat messages were marked as *engaged,* and 1,066 (29.21%) chat messages were marked as *disengaged*, indicating imbalanced data.

Next, we tagged individual segments of the video recordings as *engaged* or *disengaged*. For that, we used a window size of 10 seconds following previous work [33, 34]. For our tagging process, we investigated the video recordings and identified some common behavioral patterns in these videos, such as looking at the monitor, leaning forward, leaning backward, looking around, eating, talking to another student, using cell phones, etc. Next, we created a coding scheme to mark certain behaviors as *disengaged* following previous works [8, 10, 19]. Looking into the video recordings, we found that students were periodically disengaged from the learning activity, but they rarely disengaged for a significant time in a single stretch. The reason could be that, during our study, the classroom had multiple facilitators, teachers, and researchers around, and thus students were conscious of their surroundings and never disengaged for lengthy periods of time. With that observation, we decided to mark a segment *disengaged* if the student showed disengaged behaviors for at least 4 seconds within a 10-second window, else we mark the segment as *engaged*. Two raters (first and fifth author) labeled 25% of the video data using the HELP framework [3] and reached a Cohen's Kappa of 0.763, suggesting significant agreement. The first author labeled the remaining data following the coding scheme (see Table 1). The final data contained a total of 12,533 segments, of which 9,684 (77.27%) were marked as *engaged,* and 2,849 (22.73%) were marked as *disengaged*, again indicating imbalanced data.

3.4 Synchronizing Ground Truth Signals

In both of our ground truth sets (chat-based and video-based), the indications of being engaged or disengaged were only available at a particular moment (for chat-based ground truth) or a particular segment (for video-based ground truth). To have a comparable measurement of both sets of ground truths, we first synchronized the two modalities by using timestamps of the video recordings with timestamps from the in-game logs. We validated the synchronization by observing the timestamps of 10 random chat message-sending moments in the game logs with the corresponding segment of the video. Next, we assumed that every chat message is also segmented and has a "window of chat", in other words, we assumed that a student remains *engaged* (or *disengaged*) if the majority of the chat messages sent by that student within a 10-second window were *engaged* (or *disengaged*) chat messages. In short, we made a 10-second window for our unit of measurement across both sets of ground truths. To understand how students were engaged with the learning activity throughout the session, we decided to propagate students' current engagement state (*engaged* or *disengaged*) until the next ground truth was available. For video-based ground truth, there were only a handful of segments where these ground truths were missing (i.e., recording error) and needed to propagate, but for chat-based ground truth, there were many moments where no chat was sent (no ground truth available) and needed to propagate.

An example time-series engagement flow is shown in Figure 3. Here, we can see that chat-based engagement and video-based engagement sometimes disagree with each other (see *A-E* in Figure 3). Moments like *A* or *B* are cases where the chat message shows disengagement, but in the facial video recording, the student appeared engaged. This can happen when students are chatting about other topics but are focused (for example, one student was chatting about COVID, which is *disengaged*, but his expression was very serious). Also, note that the chat messages sent by the students prior to *A* and *B* were mostly *disengaged* chat messages from other group members (red triangle). This might indicate that group members were already having a disengaged conversation that the student participated in. On the other hand, cases like *D* can happen when a student loses focus for a short period, but it is not apparent through the chat messages as their next chat-based interaction is still engaged. Looking at the group chat, note that there was a *disengaged* chat message (red triangle) just prior to *D*, which might trigger the

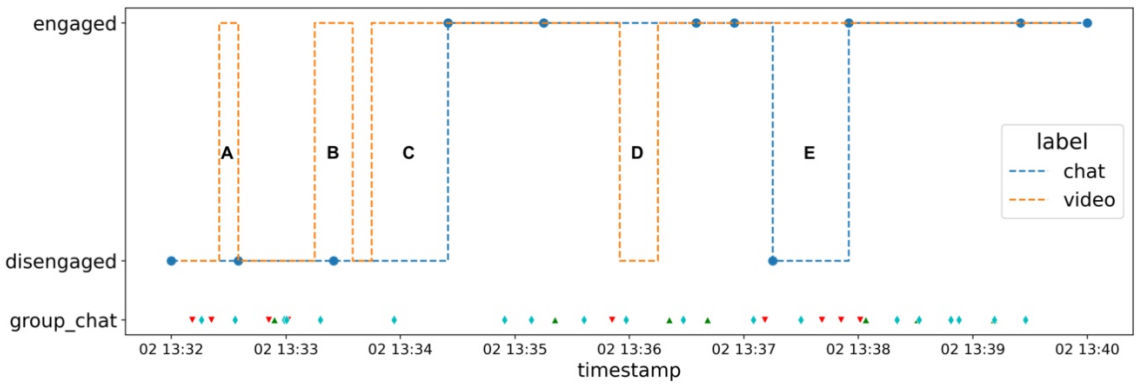

Figure 3: Example of a flow of engagement for two sets of ground truths (video-based and chat-based). The blue dots show the moment when a chat message was sent by a student. The red and green triangle shows when teammates send chat messages that are disengaged or engaged, respectively. The cyan diamond is the moment when facilitators intervened. Mark A-E shows different disagreement points between the two sets of ground truths.

momentary disengagement of the student, and a follow-up message from the facilitator (cyan diamond) that re-established the engagement. For example, one student was asking about how to vote and was momentarily distracted by another student's frustrated response ("BRUH KMON!!!"), but soon regained focus due to the facilitators' intervention. Such momentary disengagements are only reflected in facial recordings but are not visible in the chat messages and convey the core concept of data fusion for understanding student behaviors using different data streams. Cases like *E* show that students sometimes write sudden *disengaged* chat messages, but that never gets reflected in the video, as the students are seen focused on the video recording. This might also be the case when the student participated in a *disengaged* conversation, as just before E, another student in the group sent a *disengaged* message. Finally, cases like *C* show that sometimes students become *engaged* (or *disengaged*), but the moment of disengagement can be delayed when observing one data stream (chat message) while another data stream can capture that in real-time (facial video). For example, in one case, the facilitator asked the group to look into their notes, and so one of the students started looking into her notes (*engaged* seen from facial video) but replied to the chat after her observation was finished (*engaged* seen from the chat messages).

After removing segments with no students, segments when the game was paused (due to technical issues), and other sanitation (removing segments after the game ends), a total of 12,912 segments (10-seconds each) were created, of which 12,533 segments had video-based ground truths without propagation (few missing were technical errors or blocks in the video), and 2,620 segments had chat-based ground truths without propagation. Note that all 12,912 segments had both sets of ground truths after propagating the corresponding engagement tags. Among 12,912 segments, 10,575 (81.90%) segments were *engaged* when considering chat-based ground truths, and 9,926 (76.87%) segments were *engaged* when considering video-based ground truths. Across both sets of ground truth, 9,149 (70.85%) tags agree with each other. More specifically, 780 segments agree on *disengaged,* and 8,369 segments agree on *engaged.*

3.5 Behavioral Disengagement Detection Analysis

For our analysis, we designed a behavioral disengagement detection framework where we utilized two modalities of features, namely (1) *video-only-features,* where features are from the video modality, such as facial action units, pose and gaze estimations, etc., and (2) *chat-only-features,* where features are from the chat modality, such as word embeddings, for predicting students' disengagement behaviors using off-the-shelf binary classifiers (e.g., random forest, decision tree, etc.). We also combined the two modalities of features together as *multimodal-features* by concatenating *chat-only-features* with *video-only-features* for each segment. A complete flow of our framework is shown in Figure 4.

For *chat-only-features*, we first pre-processed students' chat messages with *NLTK*[2] by tokenizing and removing white space, punctuations, and stop words. Next, we transformed the tokenized chat messages into distributed vector representations using ELMo, a pre-trained language model that produces deep contextualized word embeddings. For each chat message, we computed a 256-dimension embedding for each token and then averaged over all tokens to calculate a single mean word embedding. The ELMo model was pre-trained using 5.5B tokens from Wikipedia and 3.6B tokens from the WMT 2008-2012 datasets. The *allennlp*[3] Python library was used for creating the embedding vectors. If multiple chat messages were sent within a single segment, we averaged the embedding vectors to represent the *chat-only-feature* for the segment.

For *video-only-features*, we used the *OpenFace*[4] behavior analysis toolkit that uses a convolutional neural network model to process facial videos and outputs multiple action unit features (35), pose features (6), gaze features (8), and facial landmark features (67). For our analysis, we used the action unit features, pose features, and gaze features (total 49 video-related features). For a single segment (10 seconds), we kept the average of all 10 seconds for the 49 features to represent the *video-only-feature* for that segment. Note that, like

[2]https://www.nltk.org
[3]https://allenai.org/allennlp/software/allennlp-library
[4]http://cmusatyalab.github.io/openface/

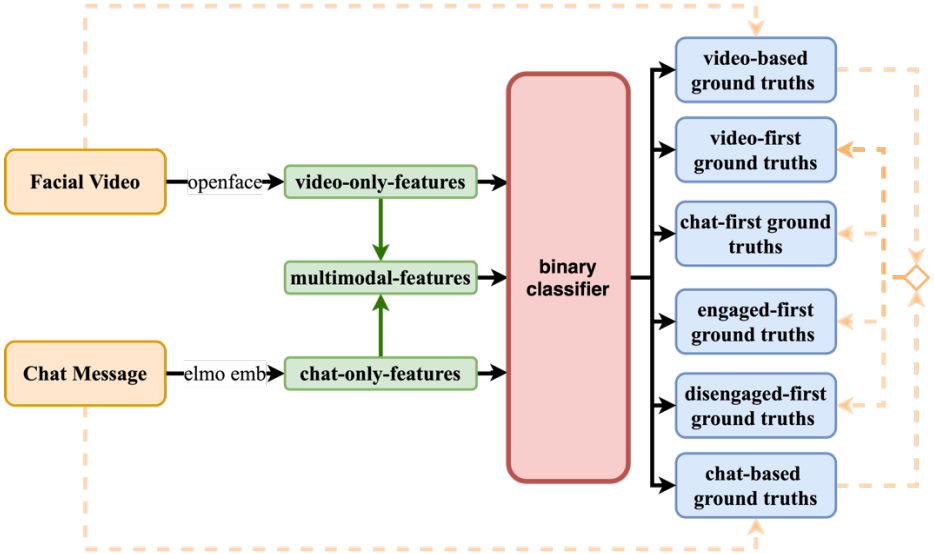

Figure 4: Behavioral disengagement detection framework. Here the black-solid lines show prediction flow. The orange-dashed lines show how the ground truths were generated. And the green-solid line shows the concatenation of features.

ground truths, the *chat-only-features* and *video-only-features* were both propagated when some segments were missing corresponding features.

Besides our two sets of ground truths (video-based ground truths and chat-based ground truths), we also create four mixed sets of ground truths by combining these two sets using different heuristics: (1) the *chat-first* ground truths prioritize chat-based ground truths in segments where a chat message was available, else relies on video-based ground truths; (2) the *video-first* ground truths prioritize video-based ground truths in segments where the facial video was available, else relies on chat-based ground truths; (3) the *engaged-first* ground truths marks a segment engaged whenever there is a disagreement; (4) similarly, the *disengaged-first* ground truths marks a segment disengaged whenever there is a disagreement. All four heuristics have their own assumptions that are easy to see. For example, for heuristic chat-first, our assumption is that chat messages are better at detecting behavioral engagement, but when not available, we can still rely on video data for detecting engagement. Similarly, disengaged-first ground truth is an aggressive approach in detecting engagement where we assume the worst. The distribution of these ground truth sets can be seen in Table 2.

Next, to compare our feature groups across different ground truth sets, we have utilized four off-the-shelf classifiers using scikit-learn Python library, namely, Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). All classifiers were trained for binary classification to predict each segment of 10 seconds as *engaged (1)* or *disengaged (0)*. The reason behind using these classifiers is somewhat arbitrary and based on popular usage, as we are not investigating classifiers or comparing them. Rather, our goal is to investigate the impact of unimodal features (*chat-only features* and *video-only features*) and *multimodal features* across different ground truth sets. As our dataset is imbalanced, we only modified the classifiers to handle class weight to balance. The rest of the parameters were in their default settings. For robustness, we have repeated all experiments with three random seeds, each with five-fold cross validation, and only reported the mean scores. As the goal of the framework is to detect segments of disengagement, we use precision, recall, and F1 for disengagement as our evaluation metrics.

4 RESULTS

Results from the unimodal and multimodal features for detecting behavioral engagement of students across each segment (10-second window) of the learning session is shown in Table 3. The columns of the table show two unimodal feature sets, namely, (1) *chat-only features* that contain 256-dimension vectors representing a single chat message, and (2) *video-only features* that contain 49-dimension vectors representing features such as action units, gaze, and post estimations, etc. (see Section 3.5), and one multimodal feature set, namely, (3) *multimodal-features* that combine both unimodal feature sets by concatenating them. The rows of Table 3 are divided into six groups, each containing a set of ground truths (see Table 2). As mentioned before, the first two groups (chat-based and video-based) are directly annotated from chat message stream and video recording stream, respectively. The other four are different combinations of these two sets of ground truths (see Section 3.5). The highest F1 scores among the three feature sets are marked in bold for ease of understanding.

First, note that chat-based ground truths have significantly higher performance when using *chat-only features*. This shows that the vector embedding from chat messages can sufficiently capture the difference between an engaged and disengaged chat message. Looking further, we can see that chat-based ground truths have significantly lower performance when using *video-only features*. The reason could be that chat message instances are few and far between (see Section 3.3 and Section 3.4), and we propagated the current

Table 2: List of different ground truth sets and their distributions.

Ground Truth Set	Heuristic for combining data streams	Total *disengaged* (0)	Total *engaged* (1)
Chat-based	Annotated from chat messages	2,337 (18.10%)	10,575 (81.90%)
Video-based	Annotated from video recordings	2,986 (23.13%)	9,926 (76.87%)
Chat-first	Prioritize chat-based ground truths in segments where a chat message was available, else relies on video-based ground truth	3,148 (24.38%)	9,764 (75.62%)
Video-first	Prioritize video-based ground truths in segments where facial video was available, else rely on chat-based ground truth	2,915 (22.57%)	9,997 (77.42%)
Engaged-first	Marks a segment *engaged* whenever there is a disagreement	780 (6.04%)	12,132 (93.96%)
Disengaged-first	Marks a segment *disengaged* whenever there is a disagreement	4,543 (35.18%)	8,369 (64.82%)

Table 3: Precision, Recall, and F1 scores (in %) for models using unimodal and multimodal features across multiple ground truth sets. Highest F1 scores are bolded and best F1 scores are given asterisk (*) for each classifier across each ground truth sets. (rows).

		Chat-only features			**Video-only features**			**Multimodal features**		
Ground Truth	**Classifier**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Chat-based	**RF**	99.37	91.23	**95.16***	28.88	13.45	18.34	99.33	91.10	95.03
	DT	95.64	93.14	**94.36**	27.89	36.20	31.50	95.14	92.44	93.76
	LR	89.46	88.38	**75.23**	24.77	61.50	35.32	65.32	87.51	74.79
	SVM	96.37	93.78	**90.34**	23.52	38.40	29.15	25.86	40.19	31.45
Video-based	**RF**	51.60	65.81	57.82	67.99	59.04	63.18	84.58	54.30	**66.11***
	DT	46.94	72.59	57.00	52.82	62.43	57.36	60.14	59.99	**60.05**
	LR	35.13	58.83	43.98	51.51	79.90	62.63	55.23	79.44	**65.15**
	SVM	45.49	60.86	**52.03**	43.23	49.72	46.24	43.39	49.59	46.27
Chat-first	**RF**	49.53	59.91	54.22	65.23	52.77	58.33	80.98	48.74	**60.84**
	DT	44.76	67.79	53.91	48.31	55.10	51.47	55.43	56.21	**55.80**
	LR	37.80	59.25	46.15	48.98	76.32	59.66	53.34	76.30	**62.78***
	SVM	46.06	60.31	**52.22**	42.07	46.71	44.24	42.23	46.53	44.25
Video-first	**RF**	49.82	65.88	56.70	64.64	57.21	60.67	83.20	52.31	**64.20***
	DT	45.48	72.75	55.95	49.63	58.43	53.65	58.47	58.54	**58.47**
	LR	34.78	58.26	43.55	49.51	78.62	50.75	53.73	78.61	**63.82**
	SVM	44.50	61.12	**51.47**	40.66	47.16	43.64	40.88	47.23	43.80
Engaged-first	**RF**	52.05	82.69	63.84	15.99	19.91	17.72	85.25	60.64	**70.78***
	DT	47.96	84.23	61.07	16.51	33.55	22.12	66.71	63.63	**65.08**
	LR	32.38	85.51	46.94	15.15	73.42	25.12	38.21	85.86	**52.86**
	SVM	41.72	87.69	**56.51**	13.41	48.46	21.00	14.54	51.03	22.62
Disengaged-first	**RF**	75.77	75.16	75.45	70.89	57.31	63.37	89.50	73.83	**80.91***
	DT	73.00	72.57	72.77	54.33	59.71	56.88	72.32	77.34	**74.74**
	LR	60.20	65.05	62.52	56.30	69.49	62.20	68.40	76.43	**72.19**
	SVM	76.12	67.92	**71.77**	52.39	40.17	45.45	53.14	40.70	46.07

state of engagement as well as the *chat-only features* across the learning session. Thus, although *video-only features* were presumably changing over time, *chat-only features* and their corresponding ground truths remained the same for all segments between two chat messages. This might cause the classifier using *video-only features* to get confused. For the same reason, using *multimodal features* does not help and the performance remains somewhat similar.

Next, looking at unimodal feature sets, we see that *video-only features* have very low performance for chat-based ground truths,

and similarly, *chat-only features* have low performance for video-based ground truths. But *video-only features* perform better in video-based ground truths, and *chat-only features* perform better in chat-based ground truths. In other words, unimodal features perform better if the ground truths are coming from the same data stream. For mixed ground truths (chat-first, video-first, engaged-first, and disengaged-first), both unimodal feature sets perform similarly, apart from engaged-first. The reason that classifiers using *video-only features* performed poorly when predicting engaged-first ground truths (17%-25% F1) could be related to the large class imbalance seen in engaged-first ground truths (see Table 2). Overall, the results show that the predictive performances of models using unimodal features change when ground truths are different (RQ1).

As we can see, apart from the chat-based ground truths, *multimodal features* are always outperforming their unimodal counterparts (2-13% F1 increased) for Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). The highest improvement is seen when using Logistic Regression in the video-first ground truths (17%) when using *multimodal features*. Also, note that, apart from one case (Logistic Regression for chat-first ground truths), Random Forest classifier always performs the best (64-95% F1). This shows that, in general, Random Forest classifiers are better at predicting disengagement behaviors among students. In other words, models using multimodal features are outperforming unimodal models in most cases (RQ2).

Another key point to notice is that SVMs always perform better with chat-only features and are never improved with video-only features or multimodal features. This is somewhat expected as SVMs are traditionally good at using small text-based data mining [32] and using multimodal features becomes a curse of dimensionality. But notice that, in all the cases, Random Forest and Decision Trees outperform SVMs for unimodal as well as for multimodal feature sets.

5 DISCUSSION AND LIMITATIONS

Unimodal features can capture student behaviors and are often sufficient for designing models that capture partial behavior. But student behaviors can only be partially observed using a single data stream. For example, understanding if a student is engaged in a learning activity can be partially observed by their interaction within the environment (chat messages data), but also can be partially observed by looking at their facial expression (video recording data). Thus, a more complete understanding can only happen when we consider different data streams together [29]. To address this, we used combined ground truths using different heuristics (chat-first, video-first, engaged-first, disengaged-first) while designing our predictive task. The intuitions of using such heuristics are given in Section 3.5.

To answer RQ1, our results show that the predictive performances of our unimodal models vary wildly across different ground truth sets. Furthermore, models using unimodal features are great at predicting behavioral disengagement if these disengagements are defined using similar data streams. For example, *video-only features* are good at predicting video-based ground truths (rows 5-8 in Table 3) of disengagement behaviors (46% to 63% F1 scores, depending on classifiers). Whenever we are using features from a different data stream, the predictive accuracy usually decreases. For example, when *video-only features* are used for predicting chat-based ground truths (rows 1-4 of Table 3), the F1 scores fall from 40% to 77%, depending on the classifier. Similarly, when *chat-only features* are used for predicting video-based ground truths (rows 5-8 of Table 3), the F1 scores decrease by 1% to 7%, except SVMs.

For RQ2, our results show that no matter which ground truth sets are used (including combined ground truths), models using *multimodal features* can at least perform as good as the unimodal counterpart (for chat-based ground truths), if not better (for the rest of the ground truth sets). A simple solution thus is to use *multimodal features* that include all the different data streams together. In such a case, no matter how the behavioral disengagements are defined, the classifiers would have sufficient knowledge to leverage the features accordingly. Our finding also supports previous work that shows *multimodal features* can generally improve predictive accuracy of student behaviors [9].

Multimodal data is resource intensive and often requires expensive equipment [22]. Understanding the trade-off between predictive accuracy and the resource cost thus becomes a crucial consideration. Our analysis shows that, in a chat-based disengagement detection, using *multimodal features* (or using features from *video-only features*) is unnecessary and does not improve performance. In other words, it is better to use unimodal features from chat messages if we are concerned about identifying disengagement behaviors from chat messages and forgo expensive data collection such as video-recordings. On the other hand, the same is not true of video-based disengagement detection or a hybrid disengagement detection (like chat-first, video-first, engaged-first, or disengaged-first ground truths). In such cases, there is a significant improvement in predictive accuracy when using *multimodal features*.

While the results demonstrate the high performance of the multimodal framework for disengagement detection in collaborative game-based learning, the work has limitations. We defined disengagement behaviors using 10-second segments for video-based ground truths. Previous work also used different sizes of segmented windows to define or observe student behavior [33, 34], but such behavior, in general, is continuous. Another limitation is the use of 10-second segments for chat-based annotations, as chat messages are event-driven and do not have a "window of engagement." This assumption, however, is reasonable as student engagement behaviors are known to persist for a short period of time [34], and it enables the possibility of combining video-based ground truths with chat-based ground truths. This work also assumes that engagement and disengagement are binary values, as is done in previous work analyzing similar student behaviors [6, 25]. Furthermore, our study involves multimodal data that has inherent challenges due to privacy and ethical concerns. Leveraging such data streams for building machine learning models that can be used in actual classroom settings calls for additional investigation.

6 CONCLUSION AND FUTURE WORKS

A vast amount of engaging and effective learning opportunities exist in collaborative game-based learning environments. These environments are designed to enhance the collaborative experiences of a group of students by embedding collective goals in the game

that requires communication, negotiation, knowledge sharing, and discussion to collaboratively resolve obstacles. However, a wide variety of negative student behaviors, such as mind-wandering, off-task conversation, disruptive chats, etc., can be observed during such learning sessions, which can lead to disengagement from the learning activity. These disengagement behaviors can be detrimental to individual learners as well as their groups and often result in negative learning outcomes. It is often overwhelming for a teacher to orchestrate a classroom while also observing individual students' engagement in the activity. Thus, assisting teachers by automatically detecting engagement behaviors offer potential for creating effective learning sessions. However, it is very challenging to automatically identify disengagement behaviors as the behaviors are scattered across different data streams. Most prior work has used a single modality to partially identify such disengagement behaviors (for example, chat-based analysis to identify off-task conversation [6]). Previous work used multimodal approaches to identify disengagement behaviors as well [36]. Limited work has investigated the impact of modalities on detecting disengagement behaviors across different data streams. In our work, we have used a multimodal behavioral disengagement detection framework that leverages in-game chat messages and facial video recordings to detect and compare the effect of different modalities on identifying disengagement behaviors across multiple data streams. Our study with middle school students interacting with a collaborative game-based learning environment shows that the predictive accuracy of unimodal features at predicting disengagement behaviors can vary significantly based on how the behaviors were observed. Our results suggest that using multimodal features can ensure maximum predictive accuracy in predicting disengagement behaviors of students, irrespective of their data streams. We also found that unimodal features are suitable (and often sufficient, in the case of text-based disengagement) for detecting disengagement behaviors from the same data streams but are inefficient for detecting disengagement behaviors across different data streams. Our findings show that when designing behavioral disengagement detection models for collaborative learning environments, it is important to consider the relationship between the modalities involved and the data streams used for observing disengagement to ensure maximum resource utilization. Furthermore, our findings can inform the design of better orchestration assistance to support teachers. AI-supported dashboards can integrate our multimodal disengagement detection framework to provide real-time information to teachers on disengaged students in different groups. Furthermore, our framework can be used for providing adaptive scaffolding for collaborative game-based learning environments and engage students in the learning activity by providing motivational instructions or feedback. Our findings can also give insight into designing multimodal models for predicting students' engagement behaviors with respect to their cost and gain trade-offs. Given these results, a promising direction for future work is investigating the integration of additional modalities, including game interaction logs and in-game collaborative processes, to further increase the accuracy of disengagement detection to improve collaborative game-based learning. Another direction of this work is to investigate deep learning models for improving the disengagement detection framework's predictive accuracy by leveraging embeddings from different modalities. Finally, further investigations are required to address privacy and ethical concerns of using multimodal models in actual classrooms.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation through grants IIS-1839966 and SES-1840120. Any opinions, findings, 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] Atezaz Ahmad, Jan Schneider, Dai Griffiths, Daniel Biedermann, Daniel Schiffner, Wolfgang Greller, and Hendrik Drachsler. 2022. Connecting the dots–A literature review on learning analytics indicators from a learning design perspective. Journal of Computer Assisted Learning (2022). DOI:https://doi.org/10.1111/jcal.12716

[2] Laura K Allen, Caitlin Mills, Matthew E Jacovina, Scott Crossley, Sidney D'mello, and Danielle S McNamara. 2016. Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, 114–123. DOI:https://doi.org/10.1145/2883851.2883939

[3] Sinem Aslan, Sinem Emine Mete, Eda Okur, Ece Oktay, Nese Alyuz, Utku Ergin Genc, David Stanhill, and Asli Arslan Esme. 2017. Human expert labeling process (HELP): Towards a reliable higher-order user state labeling process and tool to assess student engagement. Educational Technology (2017), 53–59. Retrieved from https://www.jstor.org/stable/44430540

[4] 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. DOI:https://doi.org/10.1016/j.ijhcs.2009.12.003

[5] Robert Bixler and Sidney D'Mello. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. User Modeling and User-Adapted Interaction 26, 1 (2016), 33–68. DOI:https://doi.org/10.1007/s11257-015-9167-1

[6] Dan Carpenter, Andrew Emerson, Bradford W Mott, Asmalina Saleh, Krista D Glazewski, Cindy E Hmelo-Silver, and James C Lester. 2020. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, Springer, 55–66. DOI:https://doi.org/10.1007/978-3-030-52237-7_5

[7] Juanjuan Chen, Minhong Wang, Paul A Kirschner, and Chin-Chung Tsai. 2018. The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis. Review of Educational Research 88, 6 (2018), 799–843. DOI:https://doi.org/10.3102/0034654318791584

[8] Sidney D'Mello and Art Graesser. 2007. Monitoring affective trajectories during complex learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 203–208. Retrieved from https://escholarship.org/uc/item/6p18v65q

[9] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. ACM computing surveys (CSUR) 47, 3 (2015), 1–36. DOI:https://doi.org/10.1145/2682899

[10] Sidney S D'Mello, Patrick Chipman, and Art Graesser. 2007. Posture as a predictor of learner's affective engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Retrieved from https://escholarship.org/uc/item/7hs9v2hr

[11] Ryan S J d Baker, Gregory R Moore, Angela Z Wagner, Jessica Kalka, Aatish Salvi, Michael Karabinos, Colin A Ashe, and David Yaron. 2011. The dynamics between student affect and behavior occurring outside of educational software. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, Springer, 14–24. DOI:https://doi.org/10.1007/978-3-642-24600-5_5

[12] Fahmid Morshed Fahid, Halim Acosta, Seung Lee, Dan Carpenter, Bradford Mott, Haesol Bae, Asmalina Saleh, Thomas Brush, Krista Glazewski, Cindy E Hmelo-Silver, and James Lester. 2022. Multimodal behavioral disengagement detection for collaborative game-based learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Education*, Springer, 218–221. DOI:https://doi.org/10.1007/978-3-031-11647-6_38

[13] Simon Greipl, Korbinian Moeller, and Manuel Ninaus. 2020. Potential and limits of game-based learning. International Journal of Technology Enhanced Learning 12, 4 (2020), 363–389. DOI:https://doi.org/10.1504/IJTEL.2020.110047

[14] J C Huizenga, G T M Ten Dam, J M Voogt, and W F Admiraal. 2017. Teacher perceptions of the value of game-based learning in secondary education. Computers & Education 110, (2017), 105–115. DOI:https://doi.org/10.1016/j.compedu.2017.03.008

[15] Heisawn Jeong, Cindy E Hmelo-Silver, and Kihyun Jo. 2019. Ten years of computer-supported collaborative learning: A meta-analysis of CSCL in STEM education during 2005–2014. Educational Research Review 28, (2019), 100284. DOI:https://doi.org/10.1016/j.edurev.2019.100284

[16] Ângelo Magno de Jesus and Ismar Frango Silveira. 2019. A collaborative game-based learning framework to improve computational thinking skills. In *Proceedings of the 9th International Conference on Virtual Reality and Visualization*, IEEE, 161–166. DOI:https://doi.org/ICVRV47840.2019.00038

[17] Kyungbin Kwon, Ying-Hsiu Liu, and LaShaune P Johnson. 2014. Group regulation and social-emotional interactions observed in computer supported collaborative learning: Comparison between good vs. poor collaborators. Computers & Education 78, (2014), 185–200. DOI:https://doi.org/10.1016/j.compedu.2014.06.004

[18] Ha Le, Jeroen Janssen, and Theo Wubbels. 2018. Collaborative learning practices: teacher and student perceived obstacles to effective student collaboration. Cambridge Journal of Education 48, 1 (2018), 103–122. DOI:https://doi.org/10.1080/0305764X.2016.1259389

[19] Taeckyung Lee, Dain Kim, Sooyoung Park, Dongwhi Kim, and Sung-Ju Lee. 2022. Predicting mind-wandering with facial videos in online lectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2104–2113.

[20] Gulipari Maimaiti, Chengyuan Jia, and Khe Foon Hew. 2021. Student disengagement in web-based videoconferencing supported online learning: an activity theory perspective. Interactive Learning Environments (2021), 1–20. DOI:https://doi.org/10.1080/10494820.2021.1984949

[21] Jonna Malmberg, Sanna Järvelä, Jukka Holappa, Eetu Haataja, Xiaohua Huang, and Antti Siipo. 2019. Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning? Computers in Human Behavior 96, (2019), 235–245. DOI:https://doi.org/10.1016/j.chb.2018.06.030

[22] Su Mu, Meng Cui, and Xiaodi Huang. 2020. Multimodal data fusion in learning analytics: A systematic review. Sensors 20, 23 (2020), 6856. DOI:https://doi.org/10.3390/s20236856

[23] Stefanos Nikiforos, Spyros Tzanavaris, and Katia-Lida Kermanidis. 2020. Virtual learning communities (VLCs) rethinking: influence on behavior modification—bullying detection through machine learning and natural language processing. Journal of Computers in Education 7, 4 (2020), 531–551. DOI:https://doi.org/10.1007/s40692-020-00166-5

[24] Zachary A Pardos, Ryan S J D Baker, Maria O C Z San Pedro, Sujith M Gowda, and Supreeth M Gowda. 2013. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the third international conference on learning analytics and knowledge*, 117–124.

[25] Kyungjin Park, Hyunwoo Sohn, Bradford Mott, Wookhee Min, Asmalina Saleh, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. 2021. Detecting disruptive talk in student chat-based discussion within collaborative game-based learning environments. In *Proceedings of the 11th International Learning Analytics and Knowledge Conference*, 405–415. DOI:https://doi.org/10.1145/3448139.3448178

[26] Luis P Prieto, Kshitij Sharma, and Pierre Dillenbourg. 2015. Studying teacher orchestration load in technology-enhanced classrooms. In *European Conference on Technology Enhanced Learning*, Springer, 268–281. DOI:https://doi.org/10.1007/978-3-319-24258-3_20

[27] Asmalina Saleh, Chen Feng, Haesol Bae, Cindy E Hmelo-Silver, K Glazewski, Seung Lee, Bradford Mott, and James Lester. 2021. Negotiating accountability and epistemic stances in middle-school collaborative discourse. In *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning*, 197–200. DOI:https://doi.org/10.22318/cscl2021.197

[28] Yusuf Can Semerci and Dionysis Goularas. 2021. Evaluation of students' flow state in an e-learning environment through activity and performance using deep learning techniques. Journal of Educational Computing Research 59, 5 (2021), 960–987. DOI:https://doi.org/10.1177/0735633120979836

[29] Kshitij Sharma and Michail Giannakos. 2020. Multimodal data capabilities for learning: What can multimodal data tell us about learning? British Journal of Educational Technology 0, 0 (2020), 1–35. DOI:https://doi.org/10.1111/bjet.12993

[30] Kshitij Sharma and Michail Giannakos. 2020. Multimodal data capabilities for learning: What can multimodal data tell us about learning? British Journal of Educational Technology 51, 5 (2020), 1450–1484.

[31] Kshitij Sharma, Zacharoula Papamitsiou, Jennifer K Olsen, and Michail Giannakos. 2020. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge*, 480–489. DOI:https://doi.org/10.1145/3375462.3375498

[32] Aixin Sun, Ee-Peng Lim, and Ying Liu. 2009. On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems 48, 1 (2009), 191–201. DOI:https://doi.org/10.1016/j.dss.2009.07.011

[33] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, 33–40.

[34] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. IEEE Transactions on Affective Computing 5, 1 (2014), 86–98. DOI:https://doi.org/10.1109/TAFFC.2014.2316163

[35] Zi Yang Wong and Gregory Arief D Liem. 2021. Student engagement: Current state of the construct, conceptual refinement, and future research directions. Educational Psychology Review 34 (2021), 107–138. DOI:https://doi.org/10.1007/s10648-021-09628-3

[36] Marcelo Worsley. 2018. (Dis) engagement matters: Identifying efficacious learning practices with multimodal learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 365–369. DOI:https://doi.org/10.1145/3170358.3170420

[37] Marcelo Worsley and Paulo Blikstein. 2018. A multimodal analysis of making. International Journal of Artificial Intelligence in Education 28, 3 (2018), 385–419. DOI:https://doi.org/10.1007/s40593-017-0160-1

[38] Jialing Zeng, Sophie Parks, and Junjie Shang. 2020. To learn scientifically, effectively, and enjoyably: A review of educational games. Human Behavior and Emerging Technologies 2, 2 (2020), 186–195. DOI:https://doi.org/10.1002/hbe2.188