Automatically Predicting Peer Satisfaction During Collaborative Learning with Linguistic, Acoustic, and Visual Features

Yingbo Ma
University of Florida
yingbo.ma@ufl.edu

Gloria Ashiya Katuka
University of Florida
gkatuka@ufl.edu

Mehmet Celepkolu
University of Florida
mckolu@ufl.edu

Kristy Elizabeth Boyer
University of Florida
keboyer@ufl.edu

Collaborative learning has numerous benefits such as enhancing learners' critical thinking, developing social skills, and improving learning gains. While engaging in this interactive process, learners' satisfaction toward their partners plays a crucial role in defining the success of the collaboration. However, detecting learners' satisfaction during an ongoing collaboration remains challenging, and there are no automatic techniques to predict learners' satisfaction. In this paper, we propose a multimodal approach to automatically predict peer satisfaction for co-located collaboration with features extracted from 44 middle school learners' collaborative dialogues. We investigated three types of features extracted from learners' dialogues: 1) *linguistic* features indicating semantics and sentiment; 2) *acoustic-prosodic* features including energy and pitch; and 3) *visual* features including eye gaze, head pose, facial action units, and body pose. We then trained several regression models with each of those features to predict the peer satisfaction scores that learners received from their partners. The results revealed that head position and body location were significant indicators of peer satisfaction: lower head and body distances between partners were associated with more positive peer satisfaction. Next, we investigated the influence of multimodal feature fusion methods on peer satisfaction prediction accuracy: early fusion versus late fusion. We report the comparison results between models trained with (1) best-performing unimodal features, (2) multimodal features combined by early fusion, and (3) multimodal features combined by late fusion. This line of research reveals how multimodal features from collaborative dialogues are associated with peer satisfaction, and represents a step toward the development of real-time intelligent systems that support collaborative learning.

Keywords: collaborative learning, peer satisfaction, pair programming, multimodal modeling, multimodal data fusion

1. Introduction

Collaborative learning benefits learners in numerous ways, such as enhancing critical thinking (Loes and Pascarella, 2017), developing social skills (Law et al., 2017), and improving learning gains (Madaio et al., 2017). During collaborative learning, partners may bring different

ideas to solve a problem, defend and evaluate their perspectives, and have a dynamic interaction with each other to produce a shared solution (Forsyth et al., 2020). Previous literature has highlighted the importance of this relationship between partners, as it can be a decisive factor for the success of the collaboration and positive team experience (Chan et al., 2008). Partners' satisfaction toward each other is positively associated with their perceived collaborative learning experience (Dewiyanti et al., 2007; Zhu, 2012), which could have a significant impact on their task performance (Zeitun et al., 2013) and learning outcomes (Goud et al., 2017). For example, Tseng et al. (2009) revealed that "trust among team members" is an effective factor in explaining learners' teamwork satisfaction and their perceptions of a collaborative experience; in a more recent study, Chen (2018) investigated college students' perceptions toward collaborative learning, finding that students who reported higher satisfaction with their collaborative learning experience produced stronger academic performance.

However, previous literature also suggests that students' interactions may not always be productive, and they may face challenges with their partners (Kapp, 2009), which could discourage them from working with partners in the future (Schultz et al., 2010). For example, conflicts related to interpersonal relationships during upper elementary students' collaborative learning tasks can negatively impact team members (Tsan et al., 2021). Therefore, it is important for teachers to monitor for downtrending peer satisfaction during a collaborative learning activity, and provide guidance to foster appropriate attitudes for learners to support their performance (Chan and Chen, 2010). However, in a classroom setting, teachers are responsible for supporting many groups of students, and they have limited resources to detect whether the partners in any one team have positive attitudes toward each other and enjoy working together.

One solution to this challenge is to build adaptive and intelligent systems to support collaborative learning (Magnisalis et al., 2011; D'Mello et al., 2019). Adaptive and intelligent collaborative learning support systems provide tailored learning scaffoldings to students working in teams by analyzing the dynamics of their group interaction in real time (Walker et al., 2014). If intelligent systems could predict peer satisfaction early during collaboration, they could intervene with adaptive support. However, despite the increase in the development of techniques and models to analyze students' interactions during collaborative learning (Stewart et al., 2018; Sinclair and Schneider, 2021), there is no research on automatically predicting peer satisfaction during collaboration. Current studies that analyzed learners' satisfaction during collaboration have revealed important factors such as social presence (sense of being with each other (So and Brush, 2008; Kim et al., 2011)), frequency and quality of team communication (Ku et al., 2013), and mutual trust between group members (Zhang et al., 2019)). However, most of these post-hoc studies relied on manual approaches (e.g., analyzing post-study attitude survey (Hasler-Waters and Napier, 2002) or open-ended questions (Ku et al., 2013)).

On the other hand, multimodal learning analytics (MMLA) research has created new opportunities to automatically analyze learners' interactions from multiple modalities (e.g., speech, facial expressions, body gestures), and provide insights into the learning process from different dimensions (Blikstein, 2013; Worsley and Martinez-Maldonado, 2018). Previous MMLA research explored data-driven approaches and multimodal modeling in attempts to understand students' learning (Andrade, 2017), predict learning performance (Di Mitri et al., 2017), and construct models of students' interactions (Sharma et al., 2020). In recent years, there have been increased research efforts toward using MMLA to investigate collaborative learning. For example, Worsley (2018) collected gesture, speech, and video data as college students collaborated in pairs to complete engineering design tasks, and utilized MMLA to analyze how these

multimodal data relate to students' learning gains. In a more recent study Stewart et al. (2021) successfully classified critical facets of the collaborative problem solving process with multimodal features (linguistic, acoustic-prosodic, facial expressions, and task context) derived from groups of learners' collaborative dialogues. However, multimodal learning analytics has not yet been used to automatically predict peer satisfaction from learners' interactions.

Aligned with this motivation, our goal in this paper is to investigate the automatic prediction of peer satisfaction during collaborative learning. We specifically address the following two research questions (RQs):

- **RQ 1**: What are the most predictive unimodal features of peer satisfaction during collaboration?
- **RQ 2**: Does multimodal feature fusion improve peer satisfaction prediction compared to the best-performing unimodal model?

To answer these research questions, we analyzed audio and video data collected from 44 middle school learners who worked in pairs on a series of collaborative coding activities. After participating in coding activities, each learner reported their overall satisfaction with their partners. To answer RQ 1, we examined the performance of the following features extracted from learners' collaborative dialogues, including: 1) linguistic features relating to word counts and speech rate (Stewart et al., 2018), semantic indicators from Word2Vec (Mikolov et al., 2013) and pre-trained BERT (Devlin et al., 2019), and sentiment indicators (Vivian et al., 2016); 2) acoustic features such as energy and pitch extracted with openSMILE (Eyben et al., 2010); 3) eye gaze, head pose, and facial AUs extracted with OpenFace (Baltrusaitis et al., 2018); and 4) body pose extracted with OpenPose (Cao et al., 2017). We followed a state-of-the-art methodology (Wei et al., 2021) that preserves the sequential nature of the features across the collaborative session.

The experimental results revealed two significant predictors. The first significant predictor was head position (x-axis), generated from OpenFace, which was the horizontal distance of a learner's head from the camera (located in the middle of two learners to collect video recordings). The second significant predictor was body key points (x-axis), generated from OpenPose, which was the horizontal pixel location of a learner's eight upper body key points (e.g., nose, neck, and shoulders). These results indicated that learners who had lower head and body distances from their partners were more likely to receive higher peer satisfaction scores. To answer RQ 2, we evaluated model performance when trained with different sets of multimodal features and while following both early and late fusion strategies (Khaleghi et al., 2013). The results showed that late fusion of Head Distance (x-axis) and Pre-trained BERT yielded the highest satisfaction prediction accuracy; however, we did not find a significant difference between multimodal models trained with early fusion versus late fusion.

This study provides two main contributions: 1) we present the results from extensive experiments evaluating both a variety of predictive features and a selection of sequential models; 2) and we identify two interpretable and meaningful learner behaviors that can be predictive of peer satisfaction. To the best of our knowledge, this is the first study to investigate the automatic prediction of peer satisfaction with multimodal features extracted from learners' interactions.

The rest of the paper is organized as follows: Section 2 presents the related work; Section 3 describes the dataset used for this study; Section 4 details the features we investigated; Section 5elaborates on the peer satisfaction prediction models; Section 6 presents the experimental settings and results; Section 7 discusses the implications of experimental results; and finally, section 8 concludes the paper and discusses future work.

2. Related Work

2.1. Manual Approaches to Analyzing Peer Satisfaction

Interpersonal interactions and soft skills play an important role in students' learning experiences and teams' success during collaboration (Cimatti, 2016). Previous research has emphasized that partners may have trouble collaborating on a task together for a variety of reasons, and many social factors can have an impact on peer satisfaction. For example, So and Brush (2008) recruited 48 graduate students who worked on a collaborative group project related to healthcare. The authors collected peer satisfaction data through a questionnaire developed to measure students' overall satisfaction with the experience. Sample questions included "I felt part of a learning community in my group", " I actively exchanged my ideas with group members", and "I was able to develop new skills and knowledge from other members in my group". They found that learners' perceptions of social presence and emotional bonding were statistically positive in relation to peer satisfaction, which means students, who reported to be more satisfied with their collaborative learning experience, also tended to perceive higher levels of social presence from other team members. Similarly, Zeitun et al. (2013) examined the relationship between team satisfaction and course project performance among 65 groups of students. In that study, data related to students' satisfaction levels were also collected through a satisfaction questionnaire. Sample questions included "I enjoyed working with my team members", "Our team members worked well together", and "Our team had a clear communication plan". The authors found that team satisfaction (toward partners and their collaborative work) was positively related to group performance only for American students, and there was no significant difference in learners' satisfaction or performance by gender. Despite the insights on peer satisfaction provided by the aforementioned studies, most of these studies relied on manual approaches (e.g., post-study attitude survey or open-ended questions), which do not enable the automatic prediction of peer satisfaction.

2.2. Peer Satisfaction and Collaborative Dialogue

In recent years, researchers have further investigated the relationship between peer satisfaction and collaborative dialogue during collaborative learning activities. Katuka et al. (2021) examined the relationship between dialogue act patterns and partner satisfaction, which is a measure of learners' perceptions of their partner. The authors analyzed a dialogue corpus from 18 pairs of middle school students (aging 11-13) collaborating on a series of science-simulation coding activities. They found that specific dialogue act patterns, such as learners asking their partner a question followed by their partner seeking clarification, were positively associated with partner satisfaction; conversely, dialogue act patterns such as off-task utterances between collaborators (indicating mutual distraction) were negatively associated with partner satisfaction.

Researchers have also studied the relationship between dialogue act patterns and peer satisfaction in co-creative domains (Katuka et al., 2022; Griffith et al., 2022). Griffith et al. (2022) analyzed the collaborative dialogue and coding actions of 136 high school students (aging 15-18) while they were engaged in co-creative tasks to develop computational music using the EarSketch learning environment (Magerko et al., 2016). Using a hidden Markov model, the authors identified seven hidden states, of which three were related to conversation and four were related to task actions. They found that partner satisfaction was also associated with the relative frequency of transitions from "Aesthetic Dialogue" to "Technical Dialogue". For instance, rel-

ative frequency of transitions from "Curriculum Browsing" to "Coding Editing" was negatively associated with the combined partner satisfaction.

These recent studies highlight the relationship between partner satisfaction and the dialogue interactions between learners while engaging in collaborative learning tasks. In addition to the dialogue act patterns analyzed in the above works, the collaborative learning process generates multiple modalities of data from learners' natural interactions, which provide rich sources of information that can be employed to understand collaborative learning processes from different perspectives. Building upon the promising findings of recent studies, our work aims to find the predictors of peer satisfaction from multimodal data extracted from collaborative dialogue—including speech, facial behaviors and body pose.

2.3. Automatic Peer Satisfaction Prediction with Multimodal Approaches

The goal of our study is to investigate the relationship between peer satisfaction and multimodal predictors extracted from collaborative dialogue. To achieve this goal, we built regression models to automatically predict satisfaction with multimodal data.

There has been a large body of research applying multimodal learning analytics (MMLA) that combine multiple data streams (e.g., speech and spoken words (Reilly and Schneider, 2019; Pugh et al., 2021), eye gaze patterns (Nakano et al., 2015), text messages and facial expressions (Daoudi et al., 2020)) to analyze group interactions, with the aim of investigating the individual learning process as well as understanding group dynamics during collaborative learning tasks. Studies investigating individual learning processes during collaboration mainly include predicting leadership and expertise (Scherer et al., 2012; Ochoa et al., 2013), predicting affective state (Daoudi et al., 2020; Mangaroska et al., 2022), and estimating engagement level (Worsley, 2018; Vrzakova et al., 2020). For example, Liu et al. (Liu et al., 2016) used MMLA to understand learners' knowledge model refinement processes during collaboration. They were able to better predict learners' knowledge models when they combined multiple data streams (i.e., audio, screen video, webcam video, and log files), which convey important contextual information about student learning. In another study, Ochoa et al. (2013) predicted which individuals were group leaders and experts through manual annotation of individual contributions and manual calculation of expert performance. The results indicated that group leaders' uninterrupted speech intervals (as audio-recorded) and writing intervals (logged from digital pens) were both significantly higher than those of non-leaders. In work focusing on group dynamics, research goals mainly include understanding the group cognition process (Schneider, 2019), assessing the quality of the collaborative learning process (Spikol et al., 2017), and predicting group performance (Echeverria et al., 2019; Olsen et al., 2020). For example, Spikol et al. (Spikol et al., 2017) used MMLA to estimate the success of a collaboration with face tracking, hand tracking, and audio recording. They found that distances between learners' hands and faces were two strong indicators of group performance, and lower distances indicated a higher likelihood that successful collaboration occurred among students. Echeverria et al. (Echeverria et al., 2019) applied MMLA in a healthcare setting in which nurses collaborated in groups, with their audio, movement, and physiological data analyzed. However, multimodal learning analytics has not yet been used to automatically predict peer satisfaction from learners' interactions. Our study extends this body of MMLA research on learners' interactions. We investigate different modalities (linguistic, acoustic-prosodic, and visual) for automatically predicting peer satisfaction.

3. Corpus

3.1. Participants and Collaborative Activities

This study is part of a larger project aimed at developing computer science knowledge and deepening understanding of science concepts through computationally rich science activities for middle school students (Celepkolu et al., 2020). Our dataset was collected from 44 learners in 7th grade classrooms in a middle school in the southeastern United States during two semesters (Spring and Fall 2019). Out of 44 learners, 29 (65.9%) identified themselves as females and 15 (34.1%) as males. The distribution of race/ethnicity of these learners included 41.3% self-reporting as White, 26.1% Asian/Pacific Islander, 19.5% Multiracial, 8.7% Hispanic/Latino, 4.3% Black/African American, and 1.9% Other. The mean age was 12.1 with ages ranging from 11 to 13.

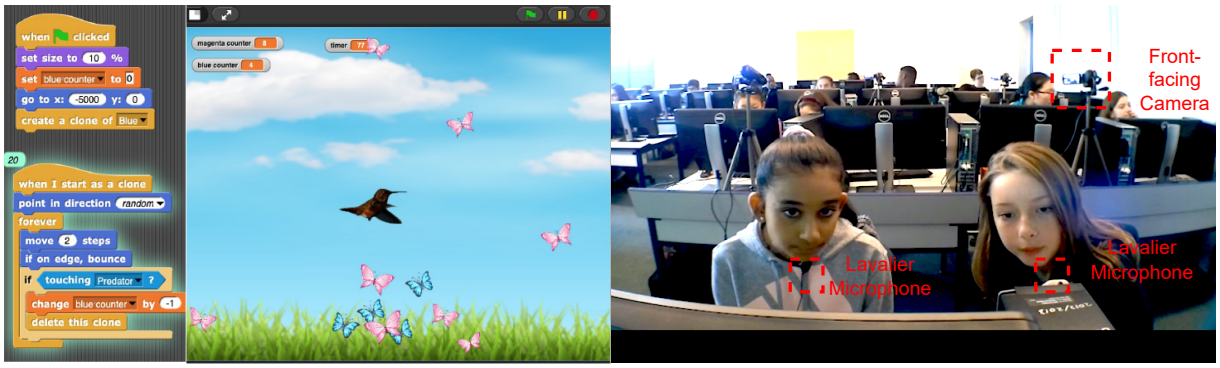


Figure 1: *Left*: A sample evolution script created with Snap!. *Right*: Two middle school learners collaborating on a pair programming task. In the captured moment, the learner in the left side of the frame is the *driver* and the learner on the right is the *navigator*; their collaborative interaction is video-recorded with a front-facing camera and audio-recorded with each learner wearing a lavalier microphone.

The learners collaborated on a series of coding activities in which they practiced fundamental CS concepts such as variables, conditionals, and loops, and applied their coding knowledge to create science models and simulations, such as homeostasis and evolution, using Snap! block-based programming environment[1]. Our learners followed the pair programming paradigm, in which each pair shared one computer and switched roles between the *driver* and the *navigator* during the science-simulation coding activity (See Figure 1-*Left*). The *driver* is responsible for writing the code and implementing the solution, while the *navigator* provides support by catching mistakes and providing feedback on the in-progress solution (Celepkolu and Boyer, 2018a). At the beginning of each collaborative session, the researchers explained both roles and the expectations for each one; during collaboration sessions, researchers reminded students to switch roles (and seats) with their partners regularly. In every class, researchers assisted the teacher by presenting an introduction to the science topics and providing students with a copy of the written instructions. Next, students worked on activities for 35-40 minutes with a randomly assigned partner. During these activities, the teacher and researchers were available to help students with their questions.

[1] https://snap.berkeley.edu/

3.2. Data Collection and Text Transcription

The collaborative coding session of each pair was video-recorded at 30 fps in 720p through a front-facing detached camera, and each child wore a lavalier microphone without active noise canceling (See Figure 1-*Right*). The audio was recorded by digital sound recorders with a sample rate of 48KHz. After the audio/video data collection process was finished, an online manual transcription service[2] generated the textual transcript for each pair (See Figure 2).

Timestamp	Speaker	Text
00:09	S2	Variable of each name B does not exist.
00:12	S1	Wait wait wait put, press that.
00:14	S1	Yeah okay so you're in here.
00:17	S2	Oh so now I need to sign in.
00:18	S1	Yeah in there.

Figure 2: A sample transcript excerpts manually generated by the online transcription service.

As shown in Figure 2, the transcripts included three pieces of information for each spoken utterance: (1) *Starting Time*, in the form of *min*:*sec*, which indicating the exact starting timestamp for each spoken utterance in the audio recording; (2) *Speaker*, in the form of *S1* (the learner sitting on the left of the video) or *S2* (the learner sitting on the right); and (3) *Transcribed Text*. Each collaborative coding session took around 30 minutes. In total, the corpus included 12 hours and 18 minutes of audio and video recordings, with 10,265 transcribed utterances. We used the timestamp from each spoken utterance to segment the audio and video recordings, generating an audio and corresponding video clip of each spoken utterance.

3.3. Peer Satisfaction Post Survey

After participating in the collaborative coding sessions, each learner completed a peer satisfaction post survey. We adopted a peer satisfaction survey from (Katuka et al., 2021), which includes the following six questions for analysis:

1. *My partner answered my questions well.*
2. *My partner listened to my suggestions.*
3. *My partner often cut my speech.*
4. *My partner was comfortable asking me questions.*
5. *My partner asking questions helped me think about things differently.*
6. *Overall, my partner was a good partner.*

[2] www.rev.com

Each of these six items in the survey was measured on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Figure 3 shows the distribution of the peer satisfaction post survey responses from 44 learners.

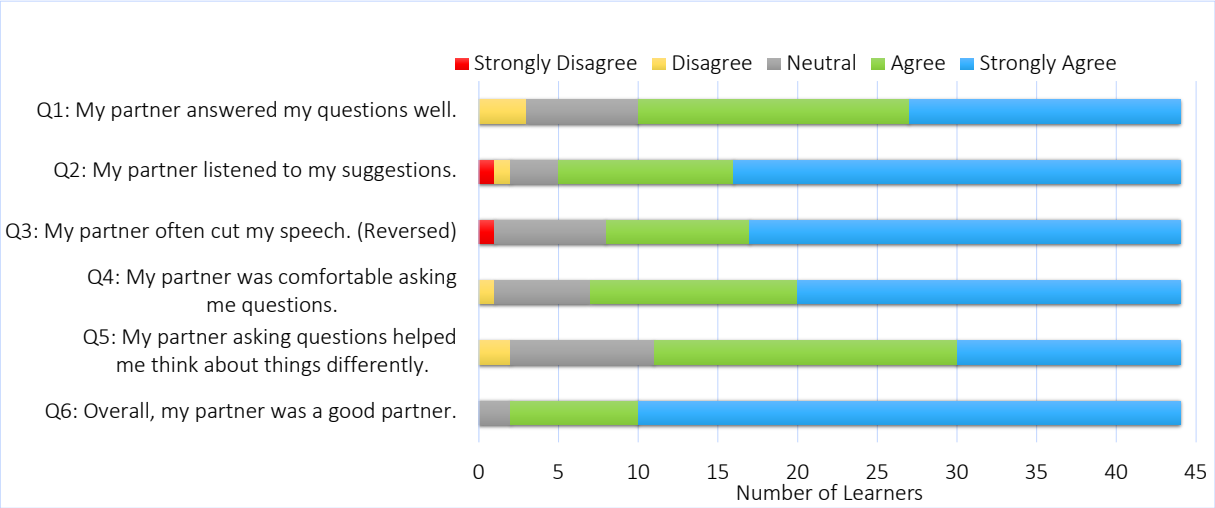


Figure 3: Distribution of peer satisfaction post-survey items from from 44 learners.

The distribution of the satisfaction scores shows that most learners agreed or strongly agreed that they were satisfied with the overall interaction with their partner. To determine whether to treat the six post-survey items as a single item or multiple items, we conducted a principal component analysis. The results of PCA suggested proceeding with only one derived outcome variable, which we refer to as *satisfaction* score (the average score of six items). This derived outcome explains 52% of the variation across all six survey items, with an eigenvalue of 3.15. The mean value of the *satisfaction* score is 4.3 (*SD*=0.6) out of 5, with a maximum value of 5.0, and a minimum value of 2.2.

4. Multimodal Features

In this study, we extracted linguistic, acoustic, and visual features for each spoken utterance during middle school learners' dialogues for collaborative learning. The organization of this section is as follows: we first describe the preprocessing of the multimodal data for feature extraction in subsection 4.1; next, we introduce the feature extraction process from language (section 4.2), audio (section 4.3), and video (section 4.4) modalities. Last, we describe the feature padding process (section 4.5) that prepared the extracted features for model training.

4.1. Multimodal Data Preprocessing

After audio and video data of collaborative coding sessions were collected from 22 pairs of middle school learners, we manually generated textual transcripts of each spoken utterance, with the additional information of the corresponding starting timestamp in the audio data and the speaker. With the help of the starting timestamp of each utterance, we used *pydub.AudioSegment*, a function in the pydub[3] library, to read audio files and extract audio segments with given starting

[3] https://github.com/jiaaro/pydub

timestamps; similarly, we used *cv2.VideoWriter*, a function in the OpenCV[4] library, to automatically extract video segments of each spoken utterance.

Before extracting linguistic features from the textual transcript of each spoken utterance, we used NLTK[5], a natural language processing tookit, to apply a few text preprocessing steps. These steps included (1) removing extra white spaces and spacial characters (e.g., "[]","{}"), (2) expanding contractions, (3) tokenization, and (4) lower casing.

To prepare to extract acoustic features from the audio segment of each spoken utterance, we used *pydub.detect_silence*, a function in the pydub library to detect the time of end-of-utterance in a given audio segment. Because we only obtained the *Starting Time* of each utterance from the online transcription service, and not the *Ending Time*, the raw audio segments in our corpus also contain silence (background noise when a learner stops talking) that elapsed before the next utterance started. It was necessary to detect and remove such silences, as they were not relevant to this analysis. The *pydub.detect_silence* function required a pre-defined parameter: silence removal threshold (any segment of audio quieter than this threshold will be considered silence). For each raw audio segment, we used three different silence thresholds to produce three different preprocessed audio segments: -6 dBFS (half of the audio's maximum level), -16 dBFS (default setting of the function), and -30 dBFS (low enough to avoid losing any speech at all).

4.2. Language-based Features

Linguistic features related to semantics extracted from spoken utterances have been used to model collaborative problem solving skills and predict collaborative task performance (Reilly and Schneider, 2019; Vrzakova et al., 2020). In addition, Literature has highlighted the positive association between sentiments detected from dialogues and collaboration performance when groups of learners engaged in solving problems (Zheng and Huang, 2016). In our study, multiple commonly used linguistic features were extracted from each spoken utterance, indicating semantics and sentiment. The five categories of language-derived features are as follows:

[4] https://github.com/opencv/opencv-python
[5] https://www.nltk.org/

1. **Word Count** For each utterance, word count was calculated as the number of words.

2. **Speech Rate** For each utterance, speech rate was calculated as the number of words divided by the number of elapsed seconds in the utterance, to produce words per second.

3. **Word2Vec** is a semantic method which learns word associations from the text, and groups similar words together in a vector space based on their semantics. We train our Word2Vec model with *gensim*[6], an open-source natural language processing library. The default settings of parameters were used, in which the dimension of each word embedding was set to 100, with a sliding window size of 5.

4. **Pre-trained BERT** is a language model trained on a large amount of data (e.g., texts from Wikipedia and books) in a self-supervised way. Similar to Word2Vec, BERT represents the semantics of words in a vector space. In this study, we used the BERT-base-uncased[7] model, which is a publicly available BERT model trained only on English texts with a hidden size of 768. With this pre-trained BERT, we generated one 768-dimensional vector for each utterance.

5. **Sentiment** indicates the attitudes of speakers. In this study, we extracted speakers' sentiment features with two commonly used libraries: (1) *nltk.SentimentAnalyzer*, which is a built-in, pretrained sentiment analyzer from the NLTK library; and (2) *analyze_sentiment*, which is a method from the Google Cloud Natural Language API[8]. Both methods take an utterance as input, and automatically generate the corresponding sentiment feature with regard to three possible sentiment categories: "positive", "neutral", or "negative". The final one-hot encoded sentiment feature was decided based on the sentiment category that was assigned with the highest confidence score.

4.3. Audio-based Features

Simple acoustic-prosodic features (e.g., sound level, synchrony in the rise and fall of the pitch) derived from audio have proven to be effective in predicting learners' engagement level (Stewart et al., 2018) and estimating group performance on solving open-ended tasks (Spikol et al., 2017). In our study, we extracted audio-derived features on the corresponding audio segment for each utterance. After removing the potential silence contained in raw audio segments, we used openSMILE[9] v2.2, an open-source toolkit for automatic acoustic feature extraction. We used the default feature set *eGeMAPSv02* (Eyben et al., 2015), which is a widely adopted acoustic feature set used for automatic voice analysis and speech emotion recognition. We used openSMILE for the automatic extraction of following five types of audio-based features within a 20-ms frame and a 10-ms window shift. The five categories of audio-based features are as follows:

1. **Loudness (Intensity)** measures the energy level of the signal. For each audio frame, 11 loudness-related features were extracted, including the mean and the standard deviation loudness values, the number of loudness peaks per second, etc.

[6] https://radimrehurek.com/gensim/
[7] https://github.com/google-research/bert
[8] https://github.com/googleapis
[9] https://audeering.github.io/opensmile-python

2. **Pitch** measures the frequency scale of a signal. For each audio frame, 10 pitch-related features were extracted, including the fundamental frequency, the mean and the standard deviation pitch values, etc.

3. **Shimmer** measures how quickly the loudness of the signal is changing, computed as the average of the relative peak amplitude differences over frames. For each audio frame, 2 shimmer-related features were extracted, including the mean and the standard deviation shimmer values.

4. **Jitter** measures how quickly the frequency of the signal is changing, computed as the average of the absolute pitch differences over frames. For each audio frame, 2 jitter-related features were extracted, including the mean and the standard deviation jitter values.

5. **MFCCs** (Mel-Frequency Cepstral Coefficients) measures the shape of the signal's short-term spectrum. For each audio frame, 16 MFCCs-related features were extracted, including the mean and the standard deviation values of lower order MFCC 1-4.

4.4. Video-based Features

A variety of features generated from video modality have been investigated in prior literature modeling collaborative problem solving. For example, eye gaze has proven effective in evaluating learners' attentiveness (Schneider et al., 2018; Huang et al., 2019) and learning performance (Celepkolu and Boyer, 2018b; Rajendran et al., 2018); head pose has also been used for assessing learners' collaborative problem solving competence (Cukurova et al., 2020); facial action units (AUs) have been used to measure both individual learners' tutoring outcomes (Grafsgaard et al., 2013) and interaction level during collaborative learning (Malmberg et al., 2019; Ma et al., 2022). Body pose has been used for analyzing learners' engagement level (Radu et al., 2020) and modeling collaborative problem solving competence (Cukurova et al., 2020). In our study, video-derived features (Figure 4) were extracted from the corresponding raw video segment of each utterance. We used the OpenFace[10] v2.0 facial behavior analysis toolkit and OpenPose[11] v1.7 body key points detection toolkit to automatically extract video-based features. For OpenFace, we used the *multiple faces* mode since both learners' faces and upper bodies were captured by one camera. The OpenPose toolkit automatically detected the body locations of both learners by default. Both OpenFace and OpenPose were run through command line arguments. Features were automatically extracted for each detected face over every video frame. The extracted four categories of video-based features are as follows (See Figure 4):

1. **Eye Gaze Direction** refers to the direction in which an eye looks. For each detected face per video frame, 8 eye gaze direction-related features were extracted. They included 3 eye gaze direction vectors (x direction, y direction, and z direction) for each eye, and 2 eye gaze directions in radians averaged for both eyes. For example, a learner looking from left to right will change the value of x direction from negative to positive; a learner looking from up to bottom will change the value of y direction from negative to positive.

2. **Head Pose** refers to head position and direction. For each detected face per video frame, 6 head-related features were extracted with OpenFace, including three head position vectors

[10] https://github.com/TadasBaltrusaitis/OpenFace
[11] https://github.com/CMU-Perceptual-Computing-Lab/openpose

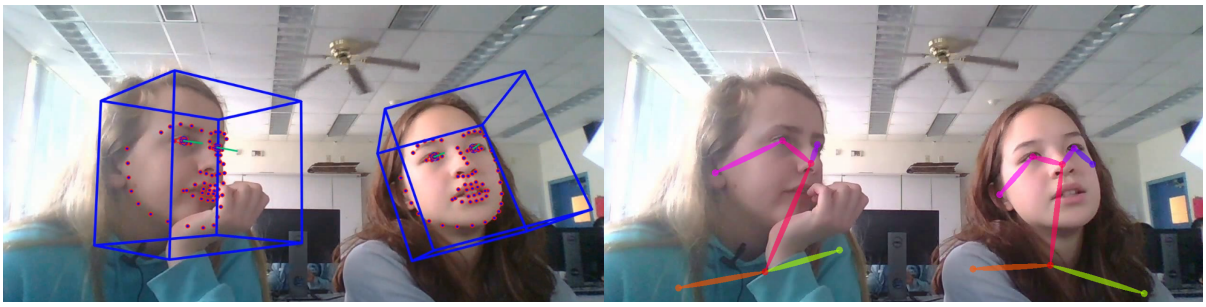


Figure 4: An example of the video-derived feature extraction process for both learners in a specific video frame. *Left*: eye gaze direction (green vectors), head pose (blue 3D bounding boxes), and facial AUs (recognized later) extracted with OpenFace. *Right*: upper body key points (e.g., nose, neck, and shoulders) extracted with OpenPose.

(x direction, y direction, and z direction) representing the location of the head with respect to the camera in millimeters, and three head direction vectors in radius with respect to the camera. Since the front-facing camera was located in the middle of two learners during the data collection process, positive values of the x direction vector and the z direction vector indicate that the learner is sitting on the right side of the video and away from the camera, and vice versa. The head position features used in our study were the absolute values of the x, y, and z direction vectors, representing the spatial location of each learner's head from the camera.

3. **Facial AUs** refer to the movements of an individual's facial muscles, and are commonly used to describe human facial expressions. Movements of facial muscles are taxonomized according to their appearance on the face by Facial Action Coding System[12]. Research has shown that specific types of AUs could be related to learners' different cognitive states (e.g., Brow Lower (AU4) and eyelid tightening (AU7) are associated with confusion (Grafsgaard et al., 2013)). In our study, for each detected face per video frame, 35 facial AU-related features were automatically extracted with OpenFace, including 17 facial AU intensity features (how intense is the AU, ranging from 0 to 5), and 18 facial AU presence features (if the AU is visible in the face, 0-absence or 1-presence).

4. **Body Pose** refers to the location of each joint (e.g., neck, shoulders) of the human body, which are known as key points that can describe a person's pose. For each learner appearing in each video frame, the 2D locations (x direction and y direction) of 8 body key points (*pose_keypoints_2d* from the output of OpenPose), measured in pixels, were extracted with OpenPose. These included the locations of each learner's eyes, nose, neck, and shoulders. OpenPose supports real-time detection of 25 full body key points (hand, facial, and foot key points); however, since our video recordings only captured learners' upper bodies, OpenPose was not able to detect the locations of some body points such as hand and foot. Therefore, only 8 body key points related to learners' upper bodies were extracted and used in this study. Because the resolution of our cameras was 720p, the maximum pixel value of body key points generated from OpenPose was 1280 pixels in the x direction, and 720 pixels in the y direction.

[12] https://www.cs.cmu.edu/~face/facs.htm

4.5. Multimodal Feature Postprocessing

After we used OpenFace and OpenPose to automatically generate video-based features, we manually corrected the output features if the following two cases happened:

- More than two faces were detected in the video frame. For example, other non-related faces could be captured by the camera when students from other pairs were seated in the background, or when teachers or researchers passed through the frame as they moved around the classroom. In this case, we manually removed the features related to these detected non-related faces after the automatic video-based feature extraction process.

- If two learners switched seats during a session. During collaboration sessions, researchers reminded learners to switch roles (*driver* and *navigator*) and seats with their partners regularly. Since neither OpenFace nor OpenPose supports identity tracking, the detected visual features and learner identities would be mismatched after each switch. In this case, we manually corrected the detected visual features to maintain learner identities after seat switching happened to ensure the detected visual features were matched with the correct learner.

Another postprocessing step was multimodal feature padding. Spoken utterances naturally vary in duration, and feature padding is an important step for ensuring the uniform size of model inputs before training machine learning models. We averaged the audio-based and video-based features across a small non-overlapping time window because they were extracted on the frame level. Following the feature aggregation strategy used in prior work (Spikol et al., 2017; Stewart et al., 2018), in which the average time windows of 500 ms and 1000 ms were chosen respectively, we selected the time window of 500 ms. We did not choose a longer window because audio-based features (e.g., pitch) could vary over a longer duration, which would lead to losing fine-grained details. Finally, post padding (adding zeros to the end of vectors) was applied on each averaged feature vector with the maximum time length (29s, 32s, and 45s) for different silence removal thresholds. For the Word2Vec-based feature, word embeddings were concatenated to form one feature vector for each utterance. Then, post padding was applied to the Word2Vec-based feature vector and the BERT-based feature vector with the maximum number (42) of spoken words. Table 1 lists the details of the multimodal features extracted and investigated in this study, and their corresponding vector dimensions after feature postprocessing.

Table 1: Utterance-level Features and the Details of Their Final Feature Dimensions

Modality	**Feature Name**	**Vector Dimensions***
Language	Word Count	1
	Speech Rate	1
	Word2Vec	4,200
	Pre-trained BERT	768
	Sentiment	3
Audio	Loudness	638, 704, 990
	Pitch	580, 640, 900
	Shimmer	116, 128, 180
	Jitter	116, 128, 180
	MFCCs	928, 1024, 1440
Video	Eye Gaze Directions	784
	Head Directions	294
	Head Position (x-axis)	98
	Head Position (y-axis)	98
	Head Position (z-axis)	98
	Facial AUs	3,430
	Body Key Points (x-axis)	784
	Body Key Points (y-axis)	784

* For every audio-based feature, the three vector dimensions resulted from different speech lengths (29s, 32s, and 45s) after applying three different silence removal thresholds (-6, -16, and -30 dBFS) respectively. For language-derived features, the maximum number of spoken words was 42. For video-derived features, the maximum time length of video segments was 49s.

5. AUTOMATIC PEER SATISFACTION PREDICTION MODELS

For a collaborative coding session of a given pair (See Figure 5), their spoken utterances were denoted by red circles (learner A) and blue circles (learner B). The peer satisfaction prediction problem can be described as follows: "Given the dialogues of learner B during the collaboration session, what predictive features can we found to best approximate the satisfaction score y that learner B received from the partner learner A?"

5.1. MODEL INPUT

As shown in Figure 5, the input of our peer satisfaction prediction model is a session-level feature sequence $X = [x_0, x_1, ..., x_{N-1}]$ extracted from learner B's dialogues, in which N is the number of spoken utterances from learner B. For each element in the feature sequence X (each spoken utterance from learner B), we used the early fusion method to generate utterance-level multimodal feature $x_t = [a_t, v_t, l_t]$ by concatenating unimodal audio-derived feature a_t, video-derived feature v_t, and language-derived feature l_t. Before concatenating unimodal feature vectors into a single multimodal feature vector, we applied z-score normalization to all the features by subtracting their mean value and dividing by their standard deviation.

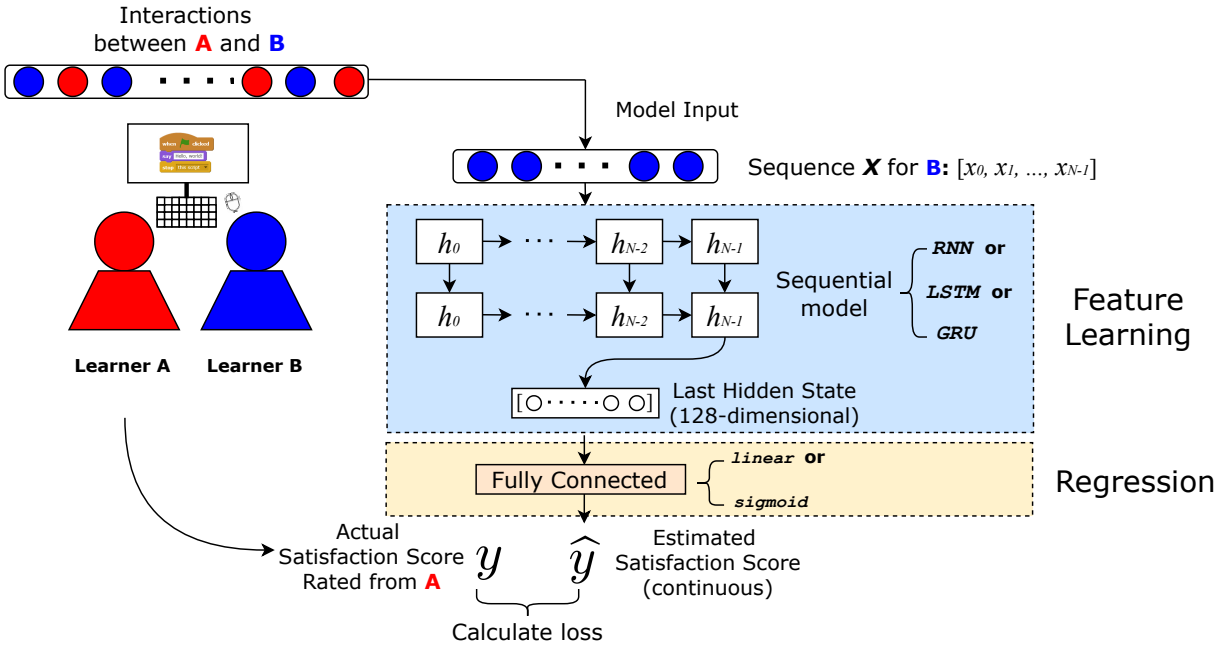


Figure 5: Architecture of the early-fusion based prediction model. The model takes multimodal features extracted from learner B's dialogues, to predict the satisfaction score learner B received from the partner learner A. For unimodal modeling, x_t ($0 \leqslant t \leqslant N-1$) is a unimodal feature vector (audio a_t, video v_t, or language l_t). For multimodal modeling, x_t is a subset of an early-fused vector $[a_t, v_t, l_t]$ (normalized).

5.2. Model Architecture

Our prediction model contains two stages: feature learning (blue box) and regression (yellow box). In the feature learning stage, we followed the current state-of-the-art methodology (Wei et al., 2021) that preserves the sequential nature of dialogue to learn the input feature sequence X. The sequential model is a two-layer LSTM network with 128 units. We obtained a final 128-dimensional hidden state h_T from the sequential model. During the regression stage of the model, we used h_T as input, and fully connected layers to output a continuous estimated satisfaction score $\widehat{y}$, in order to approximate the actual satisfaction score y rated by Learner A.

Recent research (Wei et al., 2021) has shown that the type of sequential model can play an important role for feature learning. Therefore, we also evaluated the performance of recurrent neural network (RNN) and gated recurrent unit (GRU) models to understand the influence of different sequential model architectures during feature learning. In addition, we evaluated the performance of two different output units, *sigmoid* and *linear* functions, to compare between linear and non-linear regression.

5.3. Early Fusion vs. Late Fusion

There are two critical factors that could potentially influence the performance of our satisfaction prediction model. First, different combinations of multimodal features play a fundamental role in discovering predictive features that are highly associated with peer satisfaction. Second, empirical evidence has suggested that effective feature fusion methods, which exploit the underlying relationships and mutual information among modalities (Lahat et al., 2015), could lead to significant performance improvement in downstream learning tasks such as semantic video analysis (Snoek et al., 2005), multimodal text summarization (Zhu et al., 2018) and human activity recognition in videos (Gadzicki et al., 2020). Generally, there are two commonly used feature fusion methods: early fusion and late fusion. Early fusion, which is also referred to as feature-level fusion, first concatenates various unimodal features into a single multimodal feature, then uses this multimodal feature to train learning models and output the final decision. Late fusion, which is also called decision-level fusion, first trains separate learning models with the inputs of separate unimodal features, then jointly considers the outputs from these separate models to output the final decision. There is no firm conclusion on which feature fusion is better, as literature has highlighted both advantages and disadvantages for each of these two methods.

For early fusion, the advantage lies in being able to exploit correlations across modalities (Gadzicki et al., 2020). However, it is challenging to combine multimodal features as each data source is usually collected at a different frame rate; therefore, jointly training features with different dimensions is not always effective, and could easily lead to over-fitting and poor learning performance (Gogate et al., 2017). Furthermore, the performance of early fusion could be greatly affected by using features that have low contribution (Lan et al., 2014). For late fusion, the advantage is that it focuses on the strengths of individual modalities, and then the information from each separate modality is jointly considered at the decision-level. Particularly when the different modalities vary significantly in terms of data dimensionalities and sampling rates, late fusion often results in improved performance compared to early fusion (Ramachandram and Taylor, 2017). However, the major disadvantage of late fusion is its very limited potential for exploitation of the cross-correlations among different unimodal features (Gadzicki et al., 2020). In addition, late fusion is more computationally demanding on hardware compared to early fusion, since each modality requires a separate learning stage.

In this study we experimented with both of these feature fusion methods (i.e., early fusion in Figure 5 and late fusion in Figure 6). As shown in Figure 6, after predictive unimodal features were identified, the late-fusion model takes these predictive unimodal features as input, and trains separate learning models to output the predicted satisfaction scores $\widehat{y}_1, \widehat{y}_2, ...$ The final estimated satisfaction score $\widehat{y}$ is calculated using fusion satisfaction scores generated from each learning model. A large body of research has investigated the optimal way of fusing such scores, and in this study we experimented with Average Late Fusion (Ye et al., 2012), which is a commonly used late fusion method that directly averages the scores from all learning models as the final estimated satisfaction score.

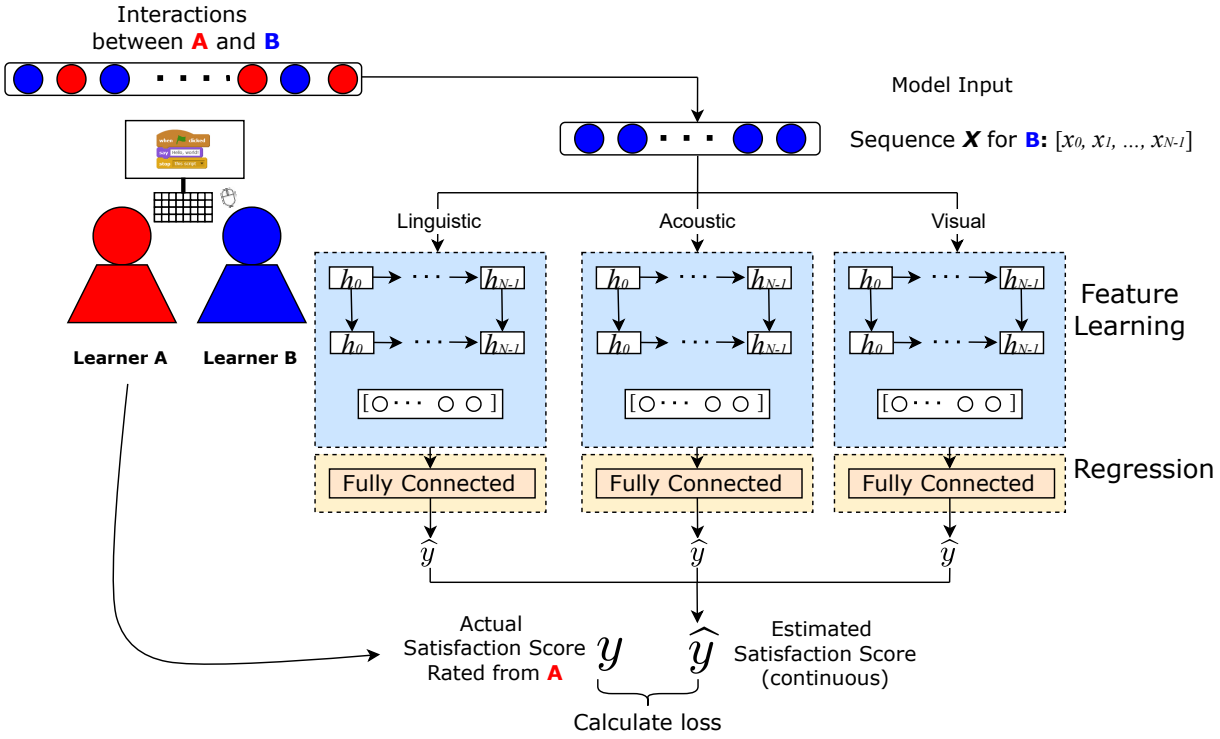


Figure 6: Architecture of the late-fusion multimodal peer satisfaction prediction model. The model takes predictive unimodal (linguistic, acoustic, or visual) features extracted from learner B's dialogue and trains learning models separately, to predict the satisfaction score learner B received from the partner learner A.

6. Experiments and Results

6.1. Experimental Setups

We implemented the Python code[13] for our prediction models in Keras with a Tensorflow backend. We conducted five-fold cross-validation to train and validate the models. All labels (y) were normalized (ranging from 0 to 1) before the model training process because the *sigmoid* activation function was used to produce the predicted satisfaction scores $\widehat{y}$. We used Adam optimizer with the learning rate of $1 \times e^{-3}$ to train the prediction model, which was trained for up to 100 epochs. The mean absolute error (MAE) was calculated for the loss function. After five rounds of cross-validation, we aggregated the MAE of each fold during the model testing process.

6.2. Investigating Unimodal Features

To identify predictive unimodal features, we compared the prediction accuracy of each unimodal feature with a randomly generated baseline feature. Followed a common method of generating uniform random baselines (Clarke and Tatler, 2014; Gambäck and Sikdar, 2017), we used the Python function *random.uniform*(0, 1), which can be interpreted as white noise without any meaningful content. We then trained the model with the white noise to generate the random baseline results (error MAE_{base} and predicted scores $\widehat{y}_{base}$). This low baseline allows us to measure the extent to which each feature predicts the outcome better than random chance. Next, we trained the model with each of the unimodal features from Table 1, and generated corresponding MAE and $\widehat{y}$. A paired-samples *t*-test (Mee and Chua, 1991) between $\widehat{y}_{base}$ and $\widehat{y}$ checked whether adding that unimodal feature significantly reduced error compared to the random baseline. Table 2 shows the regression results of peer satisfaction prediction models trained on unimodal features.

For audio-derived features, the three values in each column (from left to right) resulted from different silence removal thresholds (-6, -16, and -30 dBFS). Although time-domain features (e.g., Loudness, Shimmer) performed better than frequency-domain features (Pitch, Jitter), as indicated by lower *MAEs*, the associated *p-values* showed that none of the acoustic and prosodic features significantly outperformed the baseline. Models trained on language-based features yielded similar *MAEs* compared to the baseline model; therefore, none of the language-based features evaluated in this study were predictive for this task. For video-derived features, we identified two predictive unimodal features: learners' head positions in the *x* direction (*p-value* = 0.03), and the locations of their body key points in the *x* direction (*p-value* = 0.03).

The feature space in our study is large compared to the relatively small corpus size. Therefore, identifying predictive unimodal features helped with filtering out noisy features that are not useful in predicting satisfaction scores. Next, we examined the performance of multimodal models by combining the unimodal features that were useful.

6.3. Examining Multimodal Features

For testing the performance of combining multiple features, we selected the two significant (*p-value* $< .05$) unimodal features (Head Position x-axis and Body Key Points x-axis). In addition,

[13] https://github.com/yingbo-ma/Predicting-Peer-Satisfaction-EDM2022

Table 2: Regression results of unimodal models. Two highlighted features: Head Position (x-axis) and Body Key Points (x-axis), significantly reduced the *MAE* compared to the baseline feature (*p-value* $< .05$).

Modality	**Unimodal Feature**	***MAE***	***p-value*** ($\widehat{y}_{base}$ ***and*** $\widehat{y}$)	R^2 (y *and* $\widehat{y}$)
	Baseline	0.1953	—	0.07
Language	Word Count	0.1794	0.43	0.07
	Speech Rate	0.1790	0.19	0.29
	Word2Vec	0.1751	0.09	0.06
	Pre-trained BERT	0.1789	0.06	0.10
	Sentiment (NLTK-based)	0.1796	0.08	0.09
	Sentiment (Google API-based)	0.2073	0.14	0.06
Audio	Loudness	0.1981, 0.1790, 0.1796	0.31, 0.19, 0.12	0.02, 0.05, 0.07
	Pitch	0.2073, 0.1902, 0.1881	0.25, 0.14, 0.15	0.01, 0.15, 0.03
	Shimmer	0.1895, 0.1794, 0.1713	0.12, 0.19, 0.42	0.04, 0.12, 0.20
	Jitter	0.1983, 0.1896, 0.1853	0.14, 0.31, 0.31	0.01, 0.08, 0.06
	MFCCs	0.2341, 0.2405, 0.2318	0.19, 0.19, 0.24	0.01, 0.03, 0.03
Video	Eye Gaze Directions	0.1689	0.10	0.21
	Head Directions	0.1583	0.09	0.23
	Head Position (x-axis)	0.1402	0.03	0.68
	Head Position (y-axis)	0.1902	0.21	0.15
	Head Position (z-axis)	0.1640	0.09	0.25
	Facial AUs	0.1927	0.19	0.11
	Body Key Points (x-axis)	0.1376	0.03	0.64
	Body Key Points (y-axis)	0.1761	0.39	0.10

MAE: aggregated testing absolute error for all data samples. y: actual satisfaction scores. $\widehat{y}$: predicted satisfaction scores with each unimodal feature. $\widehat{y}_{base}$: predicted satisfaction scores with the baseline feature. R^2: another widely used metric to evaluate a regression task's level of goodness-of-fit.

Table 3: Regression results of multimodal models. None of the multimodal features significantly outperformed the baseline feature.

Multimodal Feature	***MAE***	***p-value*** ($\widehat{y}_{base}$ **and** $\widehat{y}$)	R^2 (y *and* $\widehat{y}$)
Baseline: Body Key Points (x-axis)	0.1376	—	0.64
Head Position (x-axis, z-axis)	0.1484	0.39	0.65
Head Position (x-axis), Head Directions	0.1355	0.17	0.68
Head Position (x-axis), Body Key Points (x-axis)	0.1367	0.10	0.68
Head Position (x-axis), Pre-trained BERT	0.1409	0.13	0.65
Head Position (x-axis), Sentiment (NLTK-based)	0.1467	0.15	0.65

Table 4: *MAEs* under different architecture settings.

Sequential Model	LSTM	RNN	GRU
MAE	0.1376	0.1401	0.1382
Output Unit	*sigmoid*	*linear*	
MAE	0.1376	0.1741	
# of Layers	1	2	3
MAE	0.1359	0.1376	0.1384

we also selected Head Direction, Pre-trained BERT, and Sentiment (NLTK-based), as their *p-values* are lower than 0.1 (a threshold that has been used to identify a weak trend or association (Houssami et al., 2010)). We used the best-performing unimodal model trained on Body Key Points (x-axis) as the baseline (predicted satisfaction scores $\widehat{y}_{base}$), and investigated the *p-values* of the paired-samples *t*-test between the predicted scores $\widehat{y}$ and the baseline results $\widehat{y}_{base}$. Table 3 shows the regression results of peer satisfaction trained on multimodal features.

The results shown in column 2 of table 3 indicated that combining Head Position (x-axis) and Head Directions yielded the lowest MAE. However, none of these multimodal features significantly improved the regression performance compared to the unimodal model.

6.4. Comparing Different Model Architectures

To understand the influence of different sequential models during feature learning, and compare the performance between linear and non-linear regression models, we selected the best-performing unimodal model and examined how prediction accuracy varied under different model architectures.

Table 4 shows the experimental results with different model architectures. For the selection of different sequential models, three models provided comparable performances, with LSTM yielding a slightly lower MAE. As for the selection of different activation functions, the model predicting satisfaction score with a *sigmoid* activation function performed better than with a *linear* function. In addition, although we observed faster convergence speed with *linear*, *sigmoid* provided more stable training and testing performance (see Figure 7). As for the selection of the number of layers, the one-layer LSTM performed similarly compared to two- or three-layer LSTM.

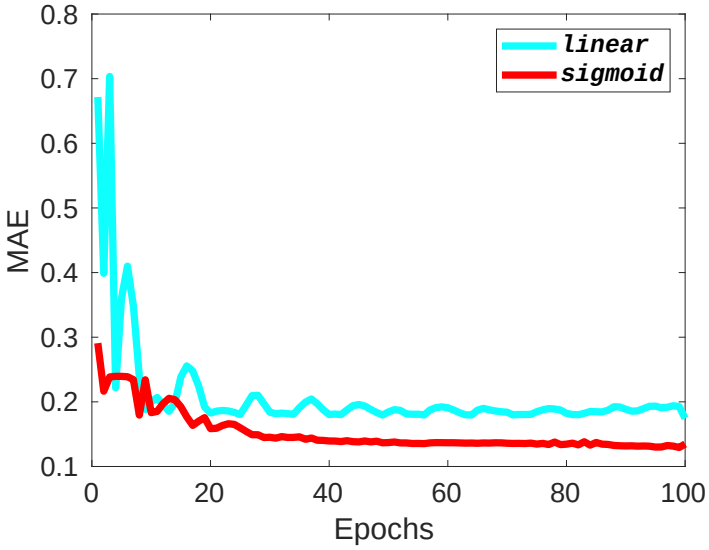


Figure 7: Testing *MAEs* under different activation functions (blue-*linear*, red-*sigmoid*). *linear* provided faster converge speed during training, while *sigmoid* provided lower MAE and more numerical stability during testing.

6.5. Comparing Between Early Fusion and Late Fusion

The method of feature combination could also potentially influence the accuracy of peer satisfaction prediction. In this study, we investigated two commonly adopted feature fusion methods, namely early fusion and late fusion. The same features were selected as in subsection 6.3: Body Key Points (x-axis), Head Position (x-axis), Head Direction, Pre-trained BERT, and Sentiment (NLTK-based). We used the best-performing unimodal model trained on Body Key Points (x-axis) as the baseline (predicted satisfaction scores $\widehat{y}_{base}$), and applied the Kruskal–Wallis test to compare three groups of satisfaction prediction scores: (1) baseline $\widehat{y}_{base}$, (2) satisfaction score from early fusion $\widehat{y}_{early}$, and (3) satisfaction score from late fusion $\widehat{y}_{late}$.

Table 5 shows the regression results of multimodal features fused by both early fusion and late fusion. The results showed that late fusion of Head Position (x-axis) and Pre-trained BERT yielded the lowest *MAE* score among different feature combinations and feature fusion methods that were experimented with. However, the *p-values* indicate that the differences among the satisfaction scores predicted with the best-performing unimodal feature, early fusion of multimodal features, and late fusion of multimodal features were not statistically significant.

Table 5: Regression results of multimodal models comparing early fusion and late fusion methods. The result show that early fusion tends to perform slightly better than late fusion, but *p-values* indicate that there was no significant difference among (1) baseline (best-performing unimodal features), (2) early fusion of multimodal features, and (3) late fusion of multimodal features.

Feature		***MAE***	***p-value*** **(**$\widehat{y}_{base}$**,** $\widehat{y}_{early}$**, and** $\widehat{y}_{late}$**)**	R^2 **(**y ***and*** $\widehat{y}$**)**
Baseline:	Body Key Points (x-axis)	0.1376		0.64
Early Fusion:	Body Key Points (x-axis) + Head Position (x-axis)	0.1367	0.50	0.65
Late Fusion:	Body Key Points (x-axis) + Head Position (x-axis)	0.1409		0.60
Baseline:	Body Key Points (x-axis)	0.1376		0.64
Early Fusion:	Body Key Points (x-axis) + Head Direction	0.1316	0.39	0.68
Late Fusion:	Body Key Points (x-axis) + Head Direction	0.1500		0.58
Baseline:	Body Key Points (x-axis)	0.1376		0.64
Early Fusion:	Body Key Points (x-axis) + Pre-trained BERT	0.1487	0.27	0.60
Late Fusion:	Body Key Points (x-axis) + Pre-trained BERT	0.1521		0.43
Baseline:	Body Key Points (x-axis)	0.1376		0.64
Early Fusion:	Body Key Points (x-axis) + Sentiment (NLTK-based)	0.1376	0.43	0.64
Late Fusion:	Body Key Points (x-axis) + Sentiment (NLTK-based)	0.1583		0.37
Baseline:	Head Position (x-axis)	0.1402		0.68
Early Fusion:	Head Position (x-axis) + Head Direction	0.1355	0.45	0.64
Late Fusion:	Head Position (x-axis) + Head Direction	0.1514		0.57
Baseline:	Head Position (x-axis)	0.1402		0.68
Early Fusion:	Head Position (x-axis) + Pre-trained BERT	0.1409	0.20	0.64
Late Fusion:	Head Position (x-axis) + Pre-trained BERT	0.1286		0.68
Baseline:	Head Position (x-axis)	0.1402		0.68
Early Fusion:	Head Position (x-axis) + Sentiment (NLTK-based)	0.1462	0.55	0.62
Late Fusion:	Head Position (x-axis) + Sentiment (NLTK-based)	0.1503		0.55

7. Discussion

We have reported on studies investigating the prediction of peer satisfaction using multimodal features from learners' interactions during collaborative learning activities. This section discusses the results with respect to our two research questions, as well as the implications of comparing the performance of different model architectures.

7.1. RQ 1: What are the most predictive unimodal features of peer satisfaction during collaboration?

7.1.1. Language-based Features

We examined several statistical (word count and speech rate), semantic (Word2Vec and BERT), and sentiment (NLTK-based and Google Cloud Natural Language API-based) features from the language modality. Statistical features such as word count per utterance and speech rate have shown to be associated with learners' active participation and turn-taking during collaboration (Stewart et al., 2018). In our study, we did not find word count or speech rate significantly related to peer satisfaction scores.

The results from our study showed that there was a trend toward significance when more semantic information was added to the features (p-values for word count, Word2Vec, and BERT: 0.46, 0.16, 0.06); however, none of these models yielded statistically significant results for predicting peer satisfaction (Table 2). Previous literature on collaborative dialogues and peer satisfaction (Katuka et al., 2022; Griffith et al., 2022) have highlighted the promising relationships between dialogue acts and learners' satisfaction during collaborative learning. For example, the dialogue sequence of questions followed by clarifications was found to be positively related to peer satisfaction (Katuka et al., 2021), the relative frequency of transitions from "Aesthetic Dialogue" to "Technical Dialogue" was found to be negatively related to peer satisfaction (Griffith et al., 2022). However, our results did not show a direct correlation between semantics and peer satisfaction. One potential reason may be that the semantic representation methods used in our study did not have the same explanatory power as dialogue acts to directly indicate learners' intentions.

In addition to semantic features, we also experimented with the sentiment extracted from learners' dialogue. Sentiment analysis, which indicates learners' positive or negative attitude, has been combined with other linguistic features (e.g., Word Vectors, Part-of-Speech Tags) to predict group performance and yielded good performance on this task (Murray and Oertel, 2018). In addition, specific positive learning sentiments (e.g., insightful sentiment) have been highlighted as positively related to group performance (Zheng and Huang, 2016). However, our study did not find a significant association between sentiment and peer satisfaction. This result may imply that current state-of-the-art automatic sentiment analysis methods, which estimate the sentiment scores on only three categories (i.e., "positive', "neutral", and "negative") are insufficient to predict learners' satisfaction effectively; this falls short of the six-dimension learning-related sentiments (e.g., "confused", "joy") described in Zheng and Huang (2016). Another potential reason for sentiment features' low performance could be that a large proportion of corpus was identified as "neutral". Among all 10,265 transcribed utterances in the corpus, 74% of the utterances in our corpus was identified as "neutral"; in contrast, only 5% of the corpus was identified as "positive" and 16% was identified as "negative". It may be that as a result, the sentiment features were not informative enough for learning models to differentiate high sat-

isfaction pairs and low satisfaction pairs. We also noticed that some task-related dialogues that should have been identified as "neutral" were mistakenly identified as "negative". For example, for the utterance such as "*Alright, amplitude is less than 30.*" or "*You can just delete that, alright.*" in which learners were describing their coding tasks, both sentiment analysis methods investigated in this study identified these utterances as "negative" in sentiment. However, even after we manually inspected and corrected these cases, we did not find a significant relationship between sentiment features and learners' satisfaction scores.

7.1.2. **Audio-based Features**

In this study, we investigated several commonly used acoustic-prosodic features and the results showed that none of these features were significant predictors of peer satisfaction. Previous literature has associated learners' peer satisfaction with their emotional bonding (So and Brush, 2008). Acoustic-prosodic features have been widely used for detecting speaker emotion detection (positive, neutral, and negative) (Chowdhury et al., 2016) and predicting learners' task performance (Spikol et al., 2017). However, the results from this study indicate that the acoustic-prosodic features we tested may not have the explanatory power to predict peer satisfaction.

One potential reason for audio features not performing well in models is if there are periods of silence included in what the model thinks are periods of speech only. We examined several different silence removal thresholds (-6, -16, -30 dBFS) and the results indicated that this strategy did not help with peer satisfaction prediction. A higher silence removal threshold (e.g., -6 dBFS) could help reduce the negative influence of background noise; however, it is also more likely to remove learners' speech. While selecting between and qualitatively examining different thresholds, we determined that -30 dBFS was optimal for our corpus to balance between eliminating periods of silence without excessively cutting off speech. However, acoustic features under that threshold were not predictive of peer satisfaction.

7.1.3. **Video-based Features**

Among the several video-based features extracted in this study, head position and body location on the horizontal axis were the only two predictive unimodal features. To better understand how the patterns of these two predictive features varied among learners with different satisfaction scores, we selected three groups of five learners and examined their sessions in more detail. The groups are as follows:

- High satisfaction group: five learners who received the highest scores (5.0 / 5.0).
- Average satisfaction group: five learners who received the exact score of 4.3 / 5.0 (mean peer satisfaction score of our corpus).
- Low satisfaction group: five learners who received the lowest five scores (all below 3.7 / 5.0).

Figure 8 shows the patterns of horizontal (x-axis) head distance from the camera, in meters, for the three groups of learners. For each group, we calculated their averaged Head Distance (x-axis) over whole sessions. Since the camera was positioned horizontally in the middle between two learners, if learners had a lower head distance from the camera, this likely reflects that the learners were sitting closer to one another. From Figure 8 we could see that learners who received high satisfaction scores (green) had lower head distances over the collaborative coding sessions, compared to learners who received an average (red) and low (blue) satisfaction scores.

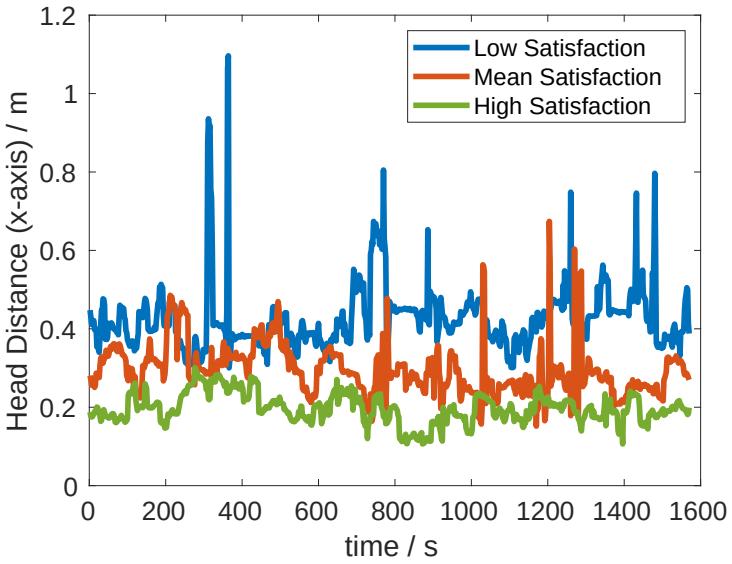


Figure 8: Head Distance (x-axis) in Meters from OpenFace

Figure 8 also depicts the difference in the head distance variance over time across the three groups of learners. Learners who received high satisfaction scores (green) had lower head distance variance and fewer numbers of sharp distance increase over time, compared to learners who received average (red) and low (blue) satisfaction scores. A sharp head distance increase could happen when the learner became disengaged in the collaborative coding tasks (e.g., talking to learners in other groups). In comparison, for learners in the high satisfaction group (green), only a small range of head distance variance over time was observed.

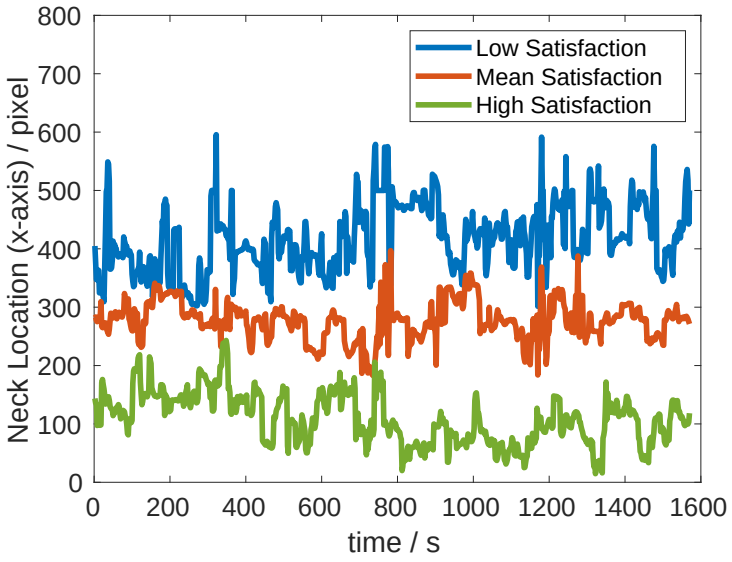


Figure 9: Neck Location (x-axis) in Pixels from OpenPose

In addition to head distance (x-axis), another predictive unimodal feature identified in our study was *body key points*, such as the location of the nose, neck, and shoulders. Figure 9 shows

the patterns of neck location (x-axis) in pixels toward the camera for three groups of learners; locations for other body key points followed relatively similar patterns. The maximum neck distance from the camera that could be detected was 640 pixels (half of 1280 pixels) because the resolution of our cameras was 720p. Figure 9 shows that learners who received high satisfaction scores (green) sit closer toward the camera (they had closer distances to their partners) over the collaborative coding sessions, compared to learners who received an average (red) and low (blue) satisfaction scores. Additionally, learners who received high satisfaction scores (green) had lower neck location variance over time, compared to learners who received an average (red) and low (blue) satisfaction scores.

The findings from Figure 8 and Figure 9 were aligned with previous literature that found learners' perceived social presence and proximity significantly impacted their satisfaction, as well as group performance during collaborative learning (Chae, 2016; Molinillo et al., 2018). For example, a study conducted by So and Brush (2008) revealed that learners' perceptions of physical proximity and psychological aspects of distance were both important factors in their reported satisfaction with their partner. In another similar study conducted by Spikol et al. (2017), the authors found that the distances between learners' faces and between learners' hands were two strong indicators of task performance when groups of college students were engaged in open-ended collaborative tasks. Fig 10 (shown on the next page) depicts the two pairs (high / low satisfaction groups) of elementary learners' pair programming sessions. The learners in the high satisfaction group (top) rated each other with the highest satisfaction score (both 5.0 out of 5.0); the learners in the low satisfaction group (top) rated each other with the lowest satisfaction score (2.2 out of 5.0; 2.3 out of 5.0).

7.2. RQ 2: Does multimodal feature fusion improve peer satisfaction prediction compared to the best-performing unimodal model?

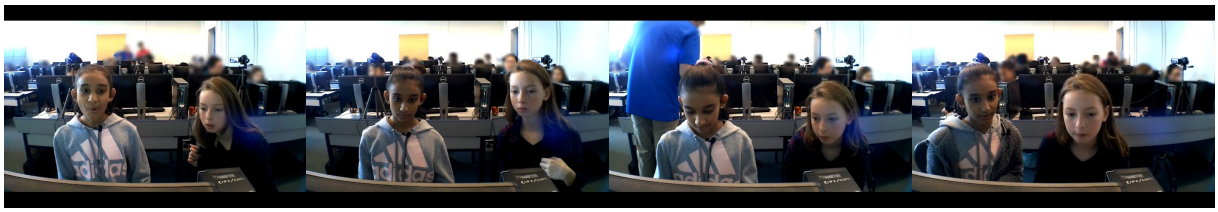
7.2.1. Unimodal vs. Multimodal

In this study, experiment results in Table 3 from several multimodal models indicated that although using multimodal features (Head Position (x-axis) combined with Body Key Points) yielded lower *MAE* than the best-performing unimodal feature, there was no significant performance advantage of using multimodal over unimodal features. The potential reason may be that both head position and body key points represented learners' spatial locations; therefore, combining these two unimodal features did not add extra useful information to predict peer satisfaction. Therefore, each of these two unimodal features alone could be used to predict peer satisfaction. However, the head pose feature extraction process with OpenFace was faster than OpenPose. OpenPose was computationally demanding, and required GPU acceleration to perform the body key points detection. Therefore, OpenFace may be a more practical feature extraction choice over OpenPose when deploying real-time learning support systems.

7.2.2. Early vs. Late Fusion

The method of combining different features could also potentially influence peer satisfaction prediction accuracy. In this study, we investigated two commonly adopted feature fusion methods: early fusion and late fusion. Experimental results shown in Table 5 indicated that the choice of feature fusion method could potentially influence accuracy: the lowest *MAE* score was achieved by late fusion of Head Distance (x-axis) and Pre-trained BERT. However, there were no significant differences between the satisfaction scores predicted by models trained with

High Satisfaction Group: S1 (5.0); S2 (5.0)



S1: What is that?
S2: It's the clone counter.

S2: Oh, just, the hint is here.
S1: I'm so stupid.

S2: We should go on the instruction.
S1: Oh, we are switching roles.

S1: I want to do like coding, and like apps and stuff.
S2: Yeah.

Low Satisfaction Group: S1 (2.2); S2 (2.3)

S1: No, that's just an example.
S2: I don't know. Hmm..

S2: Will you stop? Stop.
S1: We're working on this robotics thing but we don't know what we are doing.

S2: Are you doing the coding class?
S1: What is it?

S1: How did you guys do that? How do you..
S2: She is not being a good group member she is not paying attention.

Figure 10: Excerpts of two pairs (high / low satisfaction groups) of elementary learners' pair programming sessions. The figure includes the screenshots of their video recordings and the corresponding dialogues at the timestamp of 6, 9, 12, and 15 minutes. Over the whole collaboration session, learners in the high satisfaction group had relatively closer head distance compared to learners in the low satisfaction group. S1 represents the student who is on the left part of the video; S2 represents the student who is on the right.

the best-performing unimodal feature, multimodal features combined with early fusion, and multimodal features combined with late fusion.

7.3. Implications from Comparing Different Model Architectures

The experimental results comparing performance of different model architectures showed that the three different sequential models (RNN, LSTM, and GRU) had similar peer satisfaction prediction accuracy; in addition, non-linear regression models yielded lower MAE than linear regression models. These results have a few practical implications for researchers in the educational data mining community seeking to conduct similar studies with the methodology presented in this study.

Although sequential models were able to represent the sequential nature of utterance-level features, the comparison between different sequential models (RNN, LSTM, and GRU) did not

reflect significant performance differences. Given that GRU usually has a faster training speed than LSTM and RNN due to its simpler cell structure, GRU could be a better choice over RNN or LSTM for similar tasks. In addition, the comparison between different activation functions (*linear* and *sigmoid*) showed that the *sigmoid* regression model yielded lower *MAE* and provided more numerical stability during testing than the *linear* model. The reason may be that the satisfaction scores predicted in this study only ranged from 1 to 5, so the constrained output value range of the *sigmoid* function could better avoid large error values during training. On the contrary, there was no mechanism to prevent the *linear* activation function from predicting out-of-range satisfaction scores.

7.4. Limitations

The current work has several important limitations that need to be addressed in the future: (1) We only studied peer satisfaction in the context of co-located pair programming and analyzed recordings collected from a relatively small corpus with 44 middle school learners; therefore, the predictive features found in this paper may not generalize well to group collaboration involving three or more team members, or to learners in other populations or learning environments, such as adults or online learning. (2) The LSTM-based feature learning process was black-box, which makes it relatively difficult to interpret what predictive information was learned from each unimodal feature. (3) This study did not account for learners' prior relationships with their randomly assigned partners; however, learners' familiarity with, and perception of, their partner prior to interaction may have influenced their resulting satisfaction. (4) The effectiveness of video-derived features identified in this study relies heavily on the correct setup of the front-facing camera, in which the distance from the camera to each learner pair was kept the same during our video recording process. Generalizing the methodology of the current study to other datasets may require calibrating the coordinate values extracted by OpenFace and OpenPose. In addition, OpenFace sometimes failed to detect both learners' faces when they were not directly facing the camera or in the case of occlusion (Ahuja et al., 2019). Even though we used wide-angle camera lenses for video recording student interactions, there were some cases in which some students were sometimes out of the recording range. (5) Finally, while the peer satisfaction survey has been used in numerous prior projects with this age group, it has not been specifically tested with neurodiverse learners such as those with autism and ADHD, whose conversational patterns may differ from neurotypical norms.

8. Conclusion and Future Work

Learners' satisfaction toward their partners plays a crucial role in group performance and learning outcomes. Automatically predicting peer satisfaction during collaboration holds great promise for providing guidance to foster appropriate learning attitudes. This article has reported on the first attempt to automate the task of peer satisfaction prediction by analyzing 44 middle school learners' collaborative dialogues. We compared a set of state-of-the-art multimodal learning analytics techniques with linguistic, acoustic, and visual features extracted from students' interactions. Linguistic features included word count, speech rate, semantics, and sentiment; acoustic features included energy, pitch, and MFCCs; and visual features included eye gaze, head pose, facial AUs, and body pose. To further understand the influence of multimodal feature fusion methods on peer satisfaction prediction accuracy, we compare the performance between multi-

modal models trained with early fusion and late fusion.

The experimental results revealed two significant predictors: head position and body location. Learners who had shorter head and body distances from their partners were more likely to receive higher peer satisfaction scores. In addition, the late fusion of Head Distance (x-axis) and Pre-trained BERT yielded the highest satisfaction prediction accuracy; however, we did not find a significant difference between multimodal models trained with early fusion versus late fusion. The findings of this study, while preliminary, provide insights into how multimodal features from collaborative dialogues are associated with middle school learners' attitudes toward partners.

There are several important directions for future work. First, future work should examine the generalizability of the findings in this study using larger datasets, including data from online learning environments and multi-party interactions among groups of three or more learners. Second, although OpenFace and OpenPose support accurate detection of head pose and body pose, it remains challenging to integrate them into intelligent learning support systems for real-time analysis. Future work should investigate other methods and tools to detect learners' pose features accurately and time-efficiently. Third, the satisfaction survey was administered post-hoc in this study. Future work could investigate potential variations in students' attitudes that may occur during the ongoing collaborative process. Fourth, the current study only models partner satisfaction as its primary outcome. This intentional choice is due to the importance of learners' affective and motivational states during collaboration, for which satisfaction with a partner is an important component. Future work should also investigate the relationship between learners' peer satisfaction and their learning performance or other process-oriented collaboration metrics. Finally, more work is needed to investigate the effectiveness of feedback delivery strategies in fostering better learning attitudes during the collaborative learning process. As we move toward predicting peer satisfaction in real time, we will be able to build and investigate systems that can significantly improve learners' collaborative learning experiences in classrooms.

9. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation through grant DRL-1640141. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

REFERENCES

AHUJA, K., KIM, D., XHAKAJ, F., VARGA, V., XIE, A., ZHANG, S., TOWNSEND, J. E., HARRISON, C., OGAN, A., AND AGARWAL, Y. 2019. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3,* 3, 1–26.

ANDRADE, A. 2017. Understanding student learning trajectories using multimodal learning analytics within an embodied-interaction learning environment. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK 2017)*. LAK '17. Association for Computing Machinery, New York, NY, USA, 70–79.

BALTRUSAITIS, T., ZADEH, A., LIM, Y. C., AND MORENCY, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

BLIKSTEIN, P. 2013. Multimodal learning analytics. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK 2013)*. 102–106.

BOUKRICHA, H., WACHSMUTH, I., HOFSTÄTTER, A., AND GRAMMER, K. 2009. Pleasure-arousal-dominance driven facial expression simulation. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–7.

CAMERON, A. C. AND WINDMEIJER, F. A. 1997. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics 77,* 2, 329–342.

CAO, Z., SIMON, T., WEI, S.-E., AND SHEIKH, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. 7291–7299.

CELEPKOLU, M. AND BOYER, K. E. 2018a. The importance of producing shared code through pair programming. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. SIGCSE '18. Association for Computing Machinery, New York, NY, USA, 765–770.

CELEPKOLU, M. AND BOYER, K. E. 2018b. Predicting student performance based on eye gaze during collaborative problem solving. In *Proceedings of the Group Interaction Frontiers in Technology (GIFT 2018)*. GIFT'18. Association for Computing Machinery, New York, NY, USA.

CELEPKOLU, M. AND BOYER, K. E. 2018c. Thematic Analysis of Students' Reflections on Pair Programming in CS1. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE 2018)*. ACM, 771–776.

CELEPKOLU, M., FUSSELL, D. A., GALDO, A. C., BOYER, K. E., WIEBE, E. N., MOTT, B. W., AND LESTER, J. C. 2020. Exploring middle school students' reflections on the infusion of cs into science classrooms. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE 2020)*. SIGCSE '20. Association for Computing Machinery, New York, NY, USA, 671–677.

CHAE, S. W. 2016. Perceived proximity and trust network on creative performance in virtual collaboration environment. *Procedia Computer Science 91*, 807–812.

CHAN, C.-L., JIANG, J. J., AND KLEIN, G. 2008. Team task skills as a facilitator for application and development skills. *IEEE Transactions on Engineering Management 55,* 3, 434–441.

CHAN, L. H. AND CHEN, C.-H. 2010. Conflict from teamwork in project-based collaborative learning. *Performance Improvement 49,* 2, 23–28.

CHEN, Y. 2018. Perceptions of EFL college students toward collaborative learning. *English Language Teaching 11,* 2, 1–4.

CHOWDHURY, S. A., STEPANOV, E. A., AND RICCARDI, G. 2016. Predicting user satisfaction from turn-taking in spoken conversations. In *INTERSPEECH 2016*. 2910–2914.

CIMATTI, B. 2016. Definition, development, assessment of soft skills and their role for the quality of organizations and enterprises. *International Journal for Quality Research 10,* 1, 97–130.

CLARKE, A. D. AND TATLER, B. W. 2014. Deriving an appropriate baseline for describing fixation behaviour. *Vision Research 102*, 41–51.

CUKUROVA, M., LUCKIN, R., MILLÁN, E., AND MAVRIKIS, M. 2018. The nispi framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education 116*, 93–109.

CUKUROVA, M., ZHOU, Q., SPIKOL, D., AND LANDOLFI, L. 2020. Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough? In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK 2020)*. 270–275.

DAOUDI, I., TRANVOUEZ, E., CHEBIL, R., ESPINASSE, B., AND CHAARI, W. 2020. An edm-based multimodal method for assessing learners' affective states in collaborative crisis management serious games. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 596–600.

DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 4171–4186.

DEWIYANTI, S., BRAND-GRUWEL, S., JOCHEMS, W., AND BROERS, N. J. 2007. Students' experiences with collaborative learning in asynchronous computer-supported collaborative learning environments. *Computers in Human Behavior 23,* 1, 496–514.

DI MITRI, D., SCHEFFEL, M., DRACHSLER, H., BÖRNER, D., TERNIER, S., AND SPECHT, M. 2017. Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. In *Proceedings of the 7th International Learning Analytics & Knowledge Conference (LAK 2017)*. LAK '17. Association for Computing Machinery, New York, NY, USA, 188–197.

D'MELLO, S., STEWART, A. E., AMON, M. J., SUN, C., DURAN, N. D., AND SHUTE, V. 2019. Towards dynamic intelligent support for collaborative problem solving. In *Proceedings of the 20th Artificial Intelligence in Education Conference (AIED 2019)*. 59–65.

ECHEVERRIA, V., MARTINEZ-MALDONADO, R., AND BUCKINGHAM SHUM, S. 2019. Towards collaboration translucence: Giving meaning to multimodal group data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

ELOY, L., EB STEWART, A., JEAN AMON, M., REINHARDT, C., MICHAELS, A., SUN, C., SHUTE, V., DURAN, N. D., AND D'MELLO, S. 2019. Modeling team-level multimodal dynamics during multiparty collaboration. In *Proceedings of the 21st International Conference on Multimodal Interaction (ICMI 2019)*. 244–258.

EYBEN, F., SCHERER, K. R., SCHULLER, B. W., SUNDBERG, J., ANDRÉ, E., BUSSO, C., DEVILLERS, L. Y., EPPS, J., LAUKKA, P., NARAYANAN, S. S., ET AL. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing 7,* 2, 190–202.

EYBEN, F., WÖLLMER, M., AND SCHULLER, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 2010 International Conference on Multimedia*. 1459–1462.

FORSYTH, C., ANDREWS-TODD, J., AND STEINBERG, J. 2020. Are you really a team player? profiling of collaborative problem solvers in an online environment. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. ERIC, 403–408.

GADZICKI, K., KHAMSEHASHARI, R., AND ZETZSCHE, C. 2020. Early vs late fusion in multimodal convolutional neural networks. In *Proceedings of the 23rd International Conference on Information Fusion (FUSION 2020)*. IEEE, 1–6.

GAMBÄCK, B. AND SIKDAR, U. K. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the 1st Workshop on Abusive Language Online*. 85–90.

GOGATE, M., ADEEL, A., AND HUSSAIN, A. 2017. Deep learning driven multimodal fusion for automated deception detection. In *Proccedings of 2017 IEEE Symposium Series on Computational Intelligence (SSCI 2017)*. IEEE, 1–6.

GOUD, T. T., SMRITHIREKHA, V., AND SANGEETHA, G. 2017. Factors influencing group member satisfaction in the software industry. In *Proceedings of the 2nd Conference on Data Engineering and Communication Technology (ICDECT 2016)*. Springer, 223–230.

GRAFSGAARD, J. F., WIGGINS, J. B., BOYER, K. E., WIEBE, E. N., AND LESTER, J. C. 2013. Automatically recognizing facial indicators of frustration: a learning-centric analysis. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 159–165.

GRIFFITH, A. E., KATUKA, G. A., WIGGINS, J. B., BOYER, K. E., FREEMAN, J., MAGERKO, B., AND MCKLIN, T. 2022. Investigating the relationship between dialogue states and partner satisfaction during co-creative learning tasks. *International Journal of Artificial Intelligence in Education*, 1–40.

HAO, J., LIU, L., VON DAVIER, A., KYLLONEN, P. C., AND KITCHEN, C. 2016. Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining. In *The 9th International Conference on Educational Data Mining (EDM 2016)*. 382–387.

HASLER-WATERS, L. AND NAPIER, W. 2002. Building and supporting student team collaboration in the virtual classroom. *Quarterly Review of Distance Education 3,* 3, 345–52.

HAYASHI, Y. 2019. Detecting collaborative learning through emotions: An investigation using facial expression recognition. In *Proceedings of the 15th International Conference on Intelligent Tutoring Systems (ITS 2019)*. Springer, 89–98.

HOUSSAMI, N., MACASKILL, P., MARINOVICH, M. L., DIXON, J. M., IRWIG, L., BRENNAN, M. E., AND SOLIN, L. J. 2010. Meta-analysis of the impact of surgical margins on local recurrence in women with early-stage invasive breast cancer treated with breast-conserving therapy. *European Journal of Cancer 46,* 18, 3219–3232.

HUANG, K., BRYANT, T., AND SCHNEIDER, B. 2019. Identifying collaborative learning states using unsupervised machine learning on eye-tracking, physiological and motion sensor data. *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, 318–323.

KAPP, E. 2009. Improving student teamwork in a collaborative project-based course. *College Teaching 57,* 3, 139–143.

KATUKA, G. A., BEX, R. T., CELEPKOLU, M., BOYER, K. E., WIEBE, E., MOTT, B., AND LESTER, J. 2021. My partner was a good partner: Investigating the relationship between dialogue acts and satisfaction among middle school computer science learners. In *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning (CSCL 2021)*. International Society of the Learning Sciences, 51–58.

KATUKA, G. A., WEBBER, A. R., WIGGINS, J. B., BOYER, K. E., MAGERKO, B., MCKLIN, T., AND FREEMAN, J. 2022. The relationship between co-creative dialogue and high school learners' satisfaction with their collaborator in computational music remixing. *Proceedings of the ACM on Human-Computer Interaction 6,* CSCW1, 1–24.

KHALEGHI, B., KHAMIS, A., KARRAY, F. O., AND RAZAVI, S. N. 2013. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion 14,* 1, 28–44.

KIM, J., KWON, Y., AND CHO, D. 2011. Investigating factors that influence social presence and learning outcomes in distance higher education. *Computers & Education 57,* 2, 1512–1520.

KU, H.-Y., TSENG, H. W., AND AKARASRIWORN, C. 2013. Collaboration factors, teamwork satisfaction, and student attitudes toward online collaborative learning. *Computers in Human Behavior 29,* 3, 922–929.

LAHAT, D., ADALI, T., AND JUTTEN, C. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE 103,* 9, 1449–1477.

LAN, Z.-Z., BAO, L., YU, S.-I., LIU, W., AND HAUPTMANN, A. G. 2014. Multimedia classification and event detection using double fusion. *Multimedia Tools and Applications 71,* 1, 333–347.

LAW, Q. P., SO, H. C., AND CHUNG, J. W. 2017. Effect of collaborative learning on enhancement of students' self-efficacy, social skills and knowledge towards mobile apps development. *American Journal of Educational Research 5,* 1, 25–29.

LIU, R., DAVENPORT, J., AND STAMPER, J. 2016. Beyond log files: Using multi-modal data streams towards data-driven kc model improvement. *Proceedings of the 9th International Conference on Educational Data Mining (EDM 2016)*, 436–441.

LOES, C. N. AND PASCARELLA, E. T. 2017. Collaborative learning and critical thinking: Testing the link. *The Journal of Higher Education 88,* 5, 726–753.

MA, Y., CELEPKOLU, M., AND BOYER, K. E. 2022. Detecting impasse during collaborative problem solving with multimodal learning analytics. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK 2022)*. 45–55.

MADAIO, M., LASKO, R., OGAN, A., AND CASSELL, J. 2017. Using temporal association rule mining to predict dyadic rapport in peer tutoring. *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, 318–323.

MAGERKO, B., FREEMAN, J., MCKLIN, T., REILLY, M., LIVINGSTON, E., MCCOID, S., AND CREWS-BROWN, A. 2016. Earsketch: A steam-based approach for underrepresented populations in high school computer science education. *ACM Transactions on Computing Education (TOCE) 16,* 4, 1–25.

MAGNISALIS, I., DEMETRIADIS, S., AND KARAKOSTAS, A. 2011. Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE Transactions on Learning Technologies 4,* 1, 5–20.

MALMBERG, J., JÄRVELÄ, S., HOLAPPA, J., HAATAJA, E., HUANG, X., AND SIIPO, A. 2019. Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning? *Computers in Human Behavior 96*, 235–245.

MANGAROSKA, K., SHARMA, K., GAŠEVIĆ, D., AND GIANNAKOS, M. 2022. Exploring students' cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity. *Journal of Computer Assisted Learning 38,* 1, 40–59.

MANGAROSKA, K., SHARMA, K., GIANNAKOS, M., TRÆTTEBERG, H., AND DILLENBOURG, P. 2018. Gaze insights into debugging behavior using learner-centred analysis. In *Proceedings of the 8th International Conference on Learning Analytics & Knowledge (LAK 2018)*. 350–359.

MAO, Y. 2019. One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

MARTINEZ-MALDONADO, R., ECHEVERRIA, V., SANTOS, O. C., SANTOS, A. D. P. D., AND YACEF, K. 2018. Physical learning analytics: A multimodal perspective. In *Proceedings of the 8th International Conference on Learning Analytics & Knowledge (LAK 2018)*. 375–379.

MATSUDA, Y., FEDOTOV, D., TAKAHASHI, Y., ARAKAWA, Y., YASUMOTO, K., AND MINKER, W. 2019. Estimating user satisfaction impact in cities using physical reaction sensing and multimodal dialogue system. In *Proceedings of the 9th International Workshop on Spoken Dialogue System Technology*. Springer, 177–183.

MEE, R. W. AND CHUA, T. C. 1991. Regression toward the mean and the paired sample t test. *The American Statistician 45,* 1, 39–42.

MESSINGER, D. S., MATTSON, W. I., MAHOOR, M. H., AND COHN, J. F. 2012. The eyes have it: making positive expressions more positive and negative expressions more negative. *Emotion 12,* 3, 430.

MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

MOLINILLO, S., AGUILAR-ILLESCAS, R., ANAYA-SÁNCHEZ, R., AND VALLESPÍN-ARÁN, M. 2018. Exploring the impacts of interactions, social presence and emotional engagement on active collaborative learning in a social web-based environment. *Computers & Education 123*, 41–52.

MURRAY, G. AND OERTEL, C. 2018. Predicting group performance in task-based interaction. In *Proceedings of the 20th International Conference on Multimodal Interaction (ICMI 2018)*. 14–20.

NAKANO, Y. I., NIHONYANAGI, S., TAKASE, Y., HAYASHI, Y., AND OKADA, S. 2015. Predicting participation styles using co-occurrence patterns of nonverbal behaviors in collaborative learning. In *Proceedings of the 17th International Conference on Multimodal Interaction (ICMI 2015)*. 91–98.

NEUBAUER, C., WOOLLEY, J., KHOOSHABEH, P., AND SCHERER, S. 2016. Getting to know you: A multimodal investigation of team behavior and resilience to stress. In *Proceedings of the 18th International Conference on Multimodal Interaction (ICMI 2016)*. 193–200.

OCHOA, X., CHILUIZA, K., MÉNDEZ, G., LUZARDO, G., GUAMÁN, B., AND CASTELLS, J. 2013. Expertise estimation based on simple multimodal features. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI 2013)*. 583–590.

OLSEN, J. K., SHARMA, K., RUMMEL, N., AND ALEVEN, V. 2020. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology 51,* 5, 1527–1547.

PRAHARAJ, S., SCHEFFEL, M., SCHMITZ, M., SPECHT, M., AND DRACHSLER, H. 2021. Towards automatic collaboration analytics for group speech data using learning analytics. *Sensors 21,* 9, 3156.

PUGH, S. L., SUBBURAJ, S. K., RAO, A. R., STEWART, A. E., ANDREWS-TODD, J., AND D'MELLO, S. K. 2021. Say what? automatic modeling of collaborative problem solving skills from student speech in the wild. *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)*, 55–67.

RADU, I., TU, E., AND SCHNEIDER, B. 2020. Relationships between body postures and collaborative learning states in an augmented reality study. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*. Springer, 257–262.

RAJENDRAN, R., KUMAR, A., CARTER, K. E., LEVIN, D. T., AND BISWAS, G. 2018. Predicting learning by analyzing eye-gaze data of reading behavior. *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*, 455–461.

RAMACHANDRAM, D. AND TAYLOR, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine 34,* 6, 96–108.

REILLY, J. M., RAVENELL, M., AND SCHNEIDER, B. 2018. Exploring collaboration using motion sensors and multi-modal learning analytics. *International Educational Data Mining Society*.

REILLY, J. M. AND SCHNEIDER, B. 2019. Predicting the quality of collaborative problem solving through linguistic analysis of discourse. *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, 149–157.

RODRÍGUEZ, F. J., PRICE, K. M., AND BOYER, K. E. 2017. Exploring the pair programming process: Characteristics of effective collaboration. In *Proceedings of the 48th ACM Technical Symposium on Computer Science Education (SIGCSE 2017)*. 507–512.

SAGHAFIAN, M. AND O'NEILL, D. K. 2018. A phenomenological study of teamwork in online and face-to-face student teams. *Higher Education 75,* 1, 57–73.

SCHERER, S., WEIBEL, N., MORENCY, L.-P., AND OVIATT, S. 2012. Multimodal prediction of expertise and leadership in learning groups. In *Proceedings of the 1st International Workshop on Multimodal Learning Analytics*. 1–8.

SCHNEIDER, B. 2019. Unpacking collaborative learning processes during hands-on activities using mobile eye-trackers.

SCHNEIDER, B., SHARMA, K., CUENDET, S., ZUFFEREY, G., DILLENBOURG, P., AND PEA, R. 2018. Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups. *International Journal of Computer-Supported Collaborative Learning 13,* 3, 241–261.

SCHULTZ, J. L., WILSON, J. R., AND HESS, K. C. 2010. Team-based classroom pedagogy reframed: The student perspective. *American Journal of Business Education 3,* 7, 17–24.

SHARMA, K., PAPAMITSIOU, Z., OLSEN, J. K., AND GIANNAKOS, M. 2020. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK 2020)*. 480–489.

SINCLAIR, A. J. AND SCHNEIDER, B. 2021. Linguistic and gestural coordination: Do learners converge in collaborative dialogue?. *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)*, 431–438.

SNOEK, C. G., WORRING, M., AND SMEULDERS, A. W. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*. 399–402.

SO, H.-J. AND BRUSH, T. A. 2008. Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. *Computers & Education 51,* 1, 318–336.

SPIKOL, D., RUFFALDI, E., LANDOLFI, L., AND CUKUROVA, M. 2017. Estimation of success in collaborative learning based on multimodal learning analytics features. In *Proceedings of the 17th International Conference on Advanced Learning Technologies (ICALT 2017)*. 269–273.

SRIVASTAVA, N., NAWAZ, S., NEWN, J., LODGE, J., VELLOSO, E., M. ERFANI, S., GASEVIC, D., AND BAILEY, J. 2021. Are you with me? measurement of learners' video-watching attention with eye tracking. In *Proceedings of the 11th International Learning Analytics & Knowledge Conference (LAK 2021)*. 88–98.

STEWART, A. E., KEIRN, Z., AND D'MELLO, S. K. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction 31,* 4, 713–751.

STEWART, A. E., KEIRN, Z. A., AND D'MELLO, S. K. 2018. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI 2028)*. 21–30.

TEIXEIRA, J. P., OLIVEIRA, C., AND LOPES, C. 2013. Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology 9*, 1112–1122.

THOMAS, C. AND JAYAGOPI, D. B. 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.

TIAM-LEE, T. J. Z. AND SUMI, K. 2019. Emotional experience of students interacting with a system for learning programming. In *The AAAI-21 Workshop On Affective Content Analysis*.

TSAN, J., VANDENBERG, J., ZAKARIA, Z., BOULDEN, D. C., LYNCH, C., WIEBE, E., AND BOYER, K. E. 2021. Collaborative dialogue and types of conflict: An analysis of pair programming interactions between upper elementary students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE 2021)*. 1184–1190.

TSENG, H., WANG, C., KU, H., AND SUN, L. 2009. Key factors in online collaboration and their relationship to teamwork satisfaction. *Quarterly Review of Distance Education 10,* 2, 195–206.

VANDENBERG, J., RACHMATULLAH, A., LYNCH, C., BOYER, K. E., AND WIEBE, E. 2021. The relationship of cs attitudes, perceptions of collaboration, and pair programming strategies on upper elementary students' cs learning. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education*. 46–52.

VIVIAN, R., FALKNER, K., FALKNER, N., AND TARMAZDI, H. 2016. A method to analyze computer science students' teamwork in online collaborative learning environments. *ACM Transactions on Computing Education (TOCE) 16,* 2, 1–28.

VRZAKOVA, H., AMON, M. J., STEWART, A., DURAN, N. D., AND D'MELLO, S. K. 2020. Focused or stuck together: multimodal patterns reveal triads' performance in collaborative

problem solving. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK 2020)*. 295–304.

WALKER, E., RUMMEL, N., AND KOEDINGER, K. R. 2014. Adaptive intelligent support to improve peer tutoring in algebra. *International Journal of Artificial Intelligence in Education 24,* 1, 33–61.

WEI, W., LI, S., OKADA, S., AND KOMATANI, K. 2021. Multimodal user satisfaction recognition for non-task oriented dialogue systems. In *Proceedings of the 23rd International Conference on Multimodal Interaction (ICMI 2021)*. 586–594.

WORSLEY, M. 2012. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI 2012)*. 353–356.

WORSLEY, M. 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.

WORSLEY, M. AND MARTINEZ-MALDONADO, R. 2018. Multimodal learning analytics' past, present, and potential futures. In *CrossMMLA@ LAK*.

YAN, Z., PEI, M., AND SU, Y. 2017. Children's empathy and their perception and evaluation of facial pain expression: An eye tracking study. *Frontiers in Psychology 8*, 2284.

YANG, S., YU, X., AND ZHOU, Y. 2020. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *Proceedings of 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*. IEEE, 98–101.

YE, G., LIU, D., JHUO, I.-H., AND CHANG, S.-F. 2012. Robust late fusion with rank minimization. In *Proceedings of the 25th International Conference on Computer Vision and Pattern Recognition (CVPR 2012)*. IEEE, 3021–3028.

ZEITUN, R. M., ABDULQADER, K. S., AND ALSHARE, K. A. 2013. Team satisfaction and student group performance: A cross-cultural study. *Journal of Education for Business 88,* 5, 286–293.

ZHANG, X., MENG, Y., DE PABLOS, P. O., AND SUN, Y. 2019. Learning analytics in collaborative learning supported by slack: From the perspective of engagement. *Computers in Human Behavior 92*, 625–633.

ZHENG, L. AND HUANG, R. 2016. The effects of sentiments and co-regulation on group performance in computer supported collaborative learning. *The Internet and Higher Education 28*, 59–67.

ZHONG, B., WANG, Q., AND CHEN, J. 2016. The impact of social factors on pair programming in a primary school. *Computers in Human Behavior 64*, 423–431.

ZHU, B., LAN, X., GUO, X., BARNER, K. E., AND BONCELET, C. 2020. Multi-rate attention based gru model for engagement prediction. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI 2020)*. 841–848.

ZHU, C. 2012. Student satisfaction, performance, and knowledge construction in online collaborative learning. *Journal of Educational Technology & Society 15,* 1, 127–136.

ZHU, J., LI, H., LIU, T., ZHOU, Y., ZHANG, J., AND ZONG, C. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. 4154–4164.