IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 22, 2020, accepted January 20, 2021, date of publication January 25, 2021, date of current version February 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054176

Recognition of Students' Mental States in Discussion Based on Multimodal Data and Its Application to Educational Support

SHIMENG PENG AND KATASHI NAGAO
Department of Intelligent Systems, Graduate School of Informatics, Nagoya University, Nagoya 464-8601, Japan
Corresponding author: Katashi Nagao (nagao@i.nagoya-u.ac.jp)

This work was supported by the Microsoft Research Asia through the Grant of the 2020 MSRA Collaborative Research (CORE16).

ABSTRACT Students will experience a complex mixture of mental states during discussion, including concentration, confusion, frustration, and boredom, which have been widely acknowledged as crucial components for revealing a student's learning states. In this study, we propose using multimodal data to design an intelligent monitoring agent that can assist teachers in effectively monitoring the multiple mental states of students during discussion. We firstly developed an advanced multi-sensor-based system and applied it in a real university's research lab to collect a multimodal ''in-the-wild'' teacher-student conversation dataset. Then, we derived a set of proxy features from facial, heart rate, and acoustic modalities and used them to train several supervised learning classifiers with different multimodal fusion approaches single-channel-level, feature-level, and decision-level fusion to recognize students' multiple mental states in conversations. We explored how to design multimodal analytics to augment the ability to recognize different mental states and found that fusing heart rate and acoustic modalities yields better recognize the states of concentration (AUC = 0.842) and confusion (AUC = 0.695), while fusing three modalities yield the best performance in recognizing the states of frustration (AUC = 0.737) and boredom (AUC = 0.810). Our results also explored the possibility of leveraging the advantages of the replacement capabilities between different modalities to provide human teachers with solutions for addressing the challenges with monitoring students in different real-world education environments.

INDEX TERMS Educational support, data-driven application, multimodal learning analytics, multimodal sensing, students' mental states detection, supervised classification.

I. INTRODUCTION

Conversation-based discussion is one form of typical complex learning activities held in higher education today in which students are required to complete a series of complex learning tasks including answer questions, generate explanations, express opinions, and transfer acquired knowledge. A broad range of remarkable research has validated the idea that students' may consistently experience a mixture of multiple mental states, such as concentration/engagement, anxiety, delight, satisfaction, confusion, frustration, boredom etc., in complex cognitive learning [1]–[6]. Those mental states can be used as crucial components for inferring students' learning situations. Among them, negative emotional/mental states such as irritation, frustration, and anger are often aroused when students make mistakes, struggle at troublesome impasses, or face failure. Alternatively, a series of positive mental states such as delight, excitement, and satisfied are often aroused when they complete tasks, conquer challenges/difficulties, or gain insight [7]–[9]. Furthermore, D'Mello and Graesser [9] explored the dynamic changes in a student's learning-centered mental states, concentration, confusion, frustration, and boredom, when they complete complex learning activities such as conversation-based discussion. They suggest that a student commonly enters learning activities with a state of engaged concentration, and this state will remain until they reach a difficult impasse, which may result in their state transitioning to confusion. At this point, two transition paths are described that students may go through. One is that they go back to being engaged if the impasse has been resolved, which can be due to positive accomplishments brought about by solving problems

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

or achieving goals. Alternatively, if the impasse cannot be resolved, the student may get stuck, and their state may then transition to frustration, at which point, the student is unlikely to transition back to confusion or concentration, and if the state of frustration persists, it may be more likely to transition to boredom, and the student will finally abandon the pursuit of their learning goals. In his work, the definitions of those four mental states have also been clarified; (1) engaged concentration is the state of interest in being involved in activities, (2) confusion arises from a clear lack of understanding of the current content, (3) frustration is a state of dissatisfaction or annoyance with the content, and (4) boredom is a state of becoming weary due to no longer being interested in the content. An ideal teacher should be sufficiently sensitive to monitor the students' mental states during learning, especially for those negative ones such as confusion and frustration, and infer the need for latent assistance in order to provide personalized and adaptive coaching support. For the case of discussion activity, when a student is found to be confused with the current discussion opinion, the teacher should give further explanation, or when the student is found to be frustrated with the discussion content, the teacher can change how discussion topic is directed to help students regain their motivation to participant the discussion, thereby maximizing students' learning outcomes.

For one-to-one coaching activities, observing students' external responses, such as their facial expressions or speech statements, is a common way for teachers to monitor students' learning situations and to determine what kind of support to provide and at what times [10]. However, for the case of offline multi-participant discussion activities, teachers have difficulty capturing changes in the mental states of each participant, especially for those students who engage in few interactions in a discussion and who, most of the time, prefer to participate as if they were an audience member. In addition, the outbreak of COVID-19 around the World from early 2020 has changed people's daily lives, such as by requiring them to wear a mask to carry out daily communication. This has undoubtedly brought greater difficulties for teachers in observing the complex mental states of students in discussion activities, since facial expressions are not sufficiently available anymore. On the other hand, carrying out remote lectures or discussion activities has gradually become a popular form of modern coaching; however, because some students tend not to use cameras or tend to mute microphones during remote educational activities, this will lead to facial and auditory cues being completely unavailable at certain times, which undoubtedly brings another challenge for teachers in capturing the mental state of students.

There has been increasing attention on automatically detecting students' complex learning-centered mental states during learning, and research has benefited from online environments that make it possible to generate and accumulate massive amounts of high-frequency learning data. Most previous work has focused on detecting students' single mental state such as in terms of engagement, when they are interacting with an online tutor system or completing learning tasks in a computer environment, such as problem solving, essay writing, programming testing, and game design [4], [11]–[15]. Some of these works used uni-variate modality signals, that is, video [11], audio [4], [12], and physiological measures [13]. Most recently, with the emergence of modern sensors, opportunities to support novel methodological approaches to measure a student's mental state from various perspectives have been explored to improve recognition accuracy. The authors of [16] used facial cues and heart rate cues to predict students' engagement as they work on writing tasks in a computer environment; [14] integrated facial and EEG signals to describe a group of middle school students' engagement level while they interacted with an online learning tutor system.

However, it is still an open question on how to effectively monitor students' multiple mental states including concentration, confusion, frustration, and boredom while they interact with a human teacher in real-world learning activities. We attempt to explore these questions by leveraging multimodal data to design an intelligent monitoring agent that can effectively sense the multiple mental states of students in real-world learning activities. To achieve this goal, we developed an advance multi-sensor-based data collection system and applied it on an environment of a discussion held in a university's research lab to record the visual, physiological, and audio data of students while they interacting with a teacher. Then, we derived a series of proxy features from multimodal cues and used them to generate a set of machine learning models to predict the multiple mental states of students, as shown in Fig.1. In this study, we would like to explore how to fuse different modalities to provide the best combination in identifying different mental states. Meanwhile, we also investigate the possibility of using the supplementation and replacement capabilities of modalities to provide solutions for navigating situations in which certain modalities are unavailable in real-world educational settings.

A. NOVELTY AND CONTRIBUTIONS

There are several novel contributions in our work that are preliminarily different from relevant studies; (1) Instead of learning interactions between students and computer tutors or pre-designed script-based learning activities in both HCI or HHI environments, we are interested in paying attention to an ''un-plugged'' scenario in which students and their advisor teacher have a coaching-driven conversation in real discussion-based learning activities. Therefore, our study aims to analyze a series of ''true feelings'' exposed during these real conversations, increasing the applicability and practicality of our results for real-world coaching activities. (2) Since ''in-the-wild'' contexts with real operational environments and real teacher-student conversations pose unique challenges in terms of collecting, validating, and interpreting data, we developed a multi-sensor-based data-collection system for supporting the generation and accumulation of massive amounts of multimodal data in ''in-the-wild''

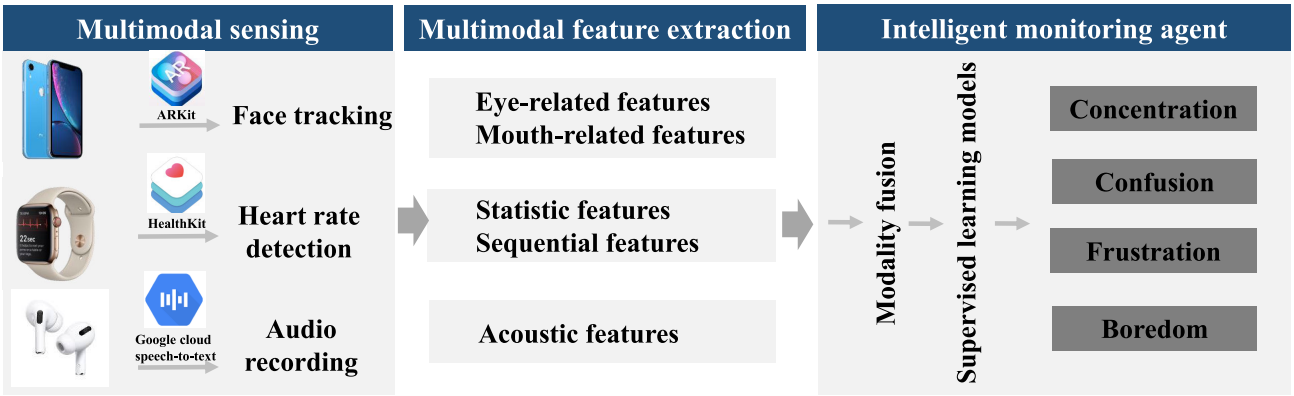


FIGURE 1. The framework for identifying students' multiple mental states based on multimodal data.

educational settings. This additionally aims to provide an enormous amount of rich high-frequency data resources to support other multi-angle analysis work in real-world educational activities. (3) With a few exceptions, most existing work has focused on using a uni-variate modality to analyze students' single mental state, such as engagement, or basic emotional states such as joy and sadness. In comparison, this study attempts to integrate multiple modalities including facial, heart rate, and acoustic cues to generate an intelligent monitoring agent that effectively senses multiple learning-centered mental states of students–concentration, confusion, frustration, and boredom. Our results provide evidence of the potential practical value of taking advantage of the supplemental and replacement capabilities between the different multimodal data to explore students' ''in-the-wild'' mental states in different real-world educational environments.

II. RELATED WORK

The research regarding the detection of students' mental states in learning activities range from early studies that used uni-variate modality to characterize students' mental states in learning activities to the recent development of multimodal learning analytics (MMLA) to measure students' multiple mental states in learning activities from high-frequency multivariate modalities. Most previous work has focused on detecting students' mental states when they complete pre-designed learning tasks in a computer environment, and only a few exceptions focused on analyzing the mental states of students when they are interacting with a human teacher in the classroom or in an offline educational environment.

A. FACIAL-MODALITY-BASED DETECTION

With the development of computer vision technologies, there has been a rich body of research work that uses facial features extracted from video streams for the task of detecting human mental states. Hoque *et al.* [17] derived a set of mouth-related features from video to characterize smiling movements and explored the possibility of a ''smile'' being used to identify frustration or delight. De *et al.* and Gomes *et al.* [18], [19] employed eye-related features from facial signals like blinking and gaze to analyze students' concentration states during learning activities. Grafsgaard *et al.* [20] used a series of video-based cues to characterize facial expressions and predicted students' engagement, frustration, and learning gain. bosch *et al.* [21] used several facial related features extracted from video to describe the movement in the brow, eye, and lip areas, and they trained several supervised learning models including logistic regression and Bayes net on the basis of those visual features to predict multiple mental states when students are playing a physics game in a computer environment. This work validated the predictive ability of facial features by achieving AUC scores of 0.610 for boredom, 0.649 for confusion, 0.867 for delight, 0.679 for engagement, and 0.631 for frustration.

B. PHYSIOLOGICAL-MODALITY-BASED DETECTION

More recent work in this space has been able to accurately predict students' learning-centered mental states or simply basic emotional states when they are engaged in learning activities. Mental states are generally considered to be related to thoughts and feelings controlled by the autonomic nervous system (ANS), and their changes can be observed through physiological signals such as the heart rate (HR) and brain waves [22]–[24]. This theoretical fact makes the heart rate (HR), heart rate variability (HRV) or EEG signal the most widely used clues in the work of emotional/mental state detection. Hellhammer assessed HR changes before, during, and after cognitive tasks to measure students' stress level [25]. Pereira *et al.* used HR and HRV to predict students' stress state [26]. Muthukrishnan validated the predictive ability of HRV features in predicting students' learning performance [27]. In our previous work [28], we took advantage of the use of heart rate signals to predict the appropriateness of students' answers, and we suggested that their mental confidence toward correctly giving answers could be indicted by their HR and HRV features. Several pieces of

work [29]–[32] employed EEG signals in a prediction task regarding students' mental states such as engagement in a student-computer interactive learning environment.

C. ACOUSTIC-MODALITY-BASED DETECTION

It is widely believed that emotional/mental state information may be transmitted from speech signals and can be explicated from linguistic and audio channels. Students' learning in computer environments makes it easy to collect a large amount of text-related log records, and several pieces of research derived a number of linguistic signals on the basis of text content to predict learning-centered mental states. Reilly *et al.* [33] used a set of linguistic features to predict how students reached consensus and their level of understanding of problems. Kovanovic *et al.* [34] took advantage of text mining technologies to extract a number of text-based features from online discussion transcripts to predict students' cognitive presence. However, for face-to-face conversation-based learning activities, a secretary is often required to manually record the content of the discussion, which will bring high costs and low accuracy of the recorded text due to the deviation between the original intention of the speaker and the understanding of the recorder. In addition, emotion recognition in conversations (ERC) has become one of the hottest topics in the NLP field and is gaining increasing attention from the community. Castellano *et al.* [35] proposed a method for extracting speech features including MFCC, pitch contour, etc. with other modality cues to classify eight basic emotions: anger, despair, interest, irritation, joy, pleasure, pride, and sadness.

D. MULTIMODAL-LEARNING-ANALYTICS BASED DETECTION

There are several inspired pieces of related literature that use MMLA to detect the mental states of students during learning. The authors of [16] used facial cues and heart rate cues to predict student engagement as they work on writing tasks in a computer environment. They generated a set of supervised learning models based on logistic regression and Bayes net, achieving AUC scores of 0.660 for classifiers based on facial modality and AUC scores of 0.730 for classifiers based on a combination of facial and heart rate modalities. These results suggest that physiological data can extend beyond what can be easily perceived by humans (e.g., facial expressions). Chen *et al.* [36] analyzed a series of video records of one child solving math problems with his mom to extract a series of features from multiple modalities such as facial, acoustic, and other interactive cues in order to characterize the child's multiple mental states including confusion, frustration, joy, and engagement demonstrated during learning activities. Peng *et al.* [14] integrated multiple modalities of facial and EEG signals from a group of middle school students to describe their engagement level when they interacted with an online learning tutor system. Wampfler *et al.* [37] adopted a bio-sensor and stylus to record several physiological signals such as heart rate and skin temperature and writing-related features such as writing speed when students worked on math tasks in a computer environment to predict their mental states

E. CURRENT STUDY

The literature review shows that there has been a lot of work exploring the automatic measurement of the mental states of students when they engage in a series of pre-designed learning tasks in a computer environment. However, there are many questions that remain unanswered with regard to how to measure the mental states of students when they completing face-to-face ''conversation tasks'' with a human teacher, and how we could effectively design detectors to address the challenges with monitoring students in ''in-the-wild'' education environments.

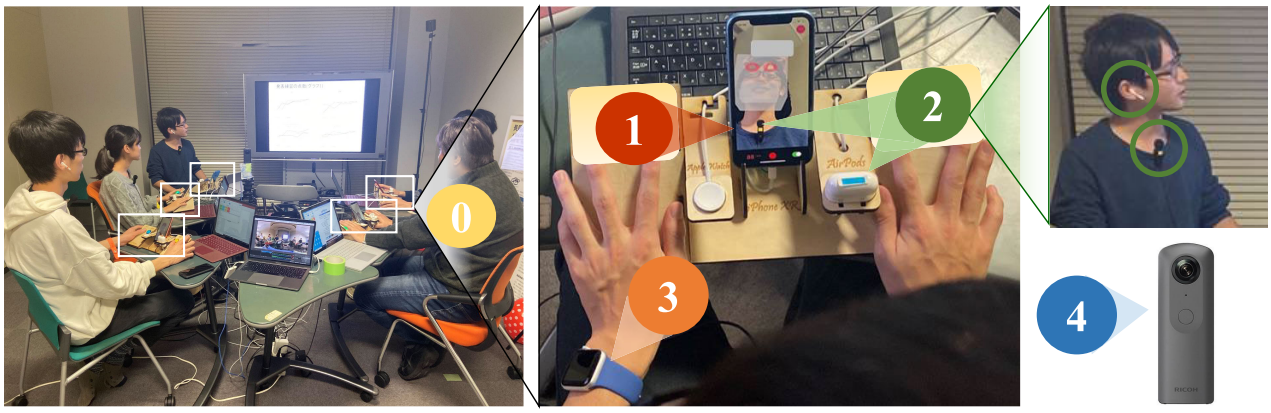
Therefore, in this study, we would like to challenge the research question of recognizing students' multiple learning-centered mental states, concentration, confusion, frustration and boredom, when they are having coaching-led conversation-based discussion with their teacher in the wild. We took advantage of modern sensor technologies to collect a multimodal dataset–we used the Apple Watch for real-time heart-rate data detection, integrated the ARKit frame work and iPhone front-camera to track and collect facial motion signals, and used AirPods together with pin microphones to record the audio of discussions. A video-audio-based retrospective annotation tool was developed to collect ground-truth measurements of the multiple mental states of students. A set of multimodal features were extracted and used to train a series of supervised classifiers with various multimodal fusion methods. We validated the performance of automatic detectors at the student level to ensure generalization to new students.

III. DATA COLLECTION METHODOLOGY

A. PARTICIPANTS AND EXPERIMENT SCENARIO

Data for our multimodal dataset was collected on the basis of participants including four graduate students (one female student and three male students) and their advisor professor. The students ranged in age from 21 to 24 years. The professor has been guiding these students for 2 years by holding regular small-group progress report meetings every week.

As shown in area (a) of Fig. 2, we selected a scenario held in a real-world university's research lab, which is the main way for the professors to check the research progress of students and provide appropriate guidance. This kind of research-progress-report meeting is held once a week, and the meetings go as follows: (1) students report their latest research progress in order; (2) the professor may ask questions about the details of the experiments based on the content of the current student's report, ask the student to explain in detail if some point is not clear, or conduct further discussion with student around a certain research problem. One regular meeting generally took around 3 hours in total, with an average length of around 50 minutes for each student's report chunk, including a 10-min presentation and



(a) Experimental scenario (b) Multimodal data collection

FIGURE 2. (a) Small group face-to-face conversation-based coaching discussion. (0) Data collection system was placed on a desk in front of each of participant. (b) Multi-sensor-based data-collection system. (1) ARKit running on iPhone for face tracking. (2) AirPods for recording audio. (3) Apple Watch for detecting heart rate. (4) Ricoh THETA for recording panorama video of the experiment in 360 degrees.

a 30–40-min. conversation discussion chunk. We observed that each student's conversation chunk were carried out only between the current presenter-student and the advisor professor. We selected this real-world discussion activity as our experimental scenario, and we attempted to explore how to detect the mental states of students when they have those conversations with their advisors. In addition, according to students' report, they were not subject to external distractions such as wearing devices to collect data, since they were completing a series of real conversations with their professor in a real regular learning activity.

B. MULTI-SENSOR BASED MULTIMODAL DATA-COLLECTION SYSTEM

We developed a multi-sensor-based data collection system shown in area (b) of Fig. 2 to collect multimodal data involving facial, heart rate, and audio signals as students held discussions with the professor. Before the meeting, all participants were asked to initiate the data collection mechanism, which was placed on a desk in front of each of them, by choosing their name and meeting date and then pressing the record button on the iPhone; then, the multi-modal data-collection functions for each device were synchronized and started. All of the sensors were working with the same timestamp.

- Facial data collection: We used the ARKit framework to integrate the front-facing camera of the iPhone to detect the face and track the positions of the face with six degrees of freedom, and we then generated a virtual mesh overlaid over the face to simulate facial expressions in real-time, an example was shown as Fig. 3.
- Heart rate data collection: We employed HealthKit framework running on a paired Apple Watch to detect students' changes in heart rate for the entire discussion, The students were asked to wear the Apple Watch on their wrist, and it was started at the same time as the iPhone.
- Audio data collection: We used AirPods to synchronously record the participants' audio data. Participants wore the AirPods in each ear for the entire discussion. In the case that the audio data could not be recorded due to equipment failure or lack of power, we also required students to wear pin microphones to record their audio data. We also used Google Cloud Speech-to-Text to convert speech content into text content, and asked the speakers to manually modify the text that was translated incorrectly while listening to a recording of their speech as well as to add a period after each complete sentence so that entire speech statements could be divided into sentence units. These text data in sentence units were stored in csv format, and we used them as subtitles for the video in the annotation work that we will explain in the next section.
- Panorama video recording: We used a Ricoh THETA set in the middle of where the participants were seated to record panorama video of the experiment in 360 degrees; this was for providing audio-video reference for annotating participants' mental states.

FIGURE 3. An example of facial movements detected from a student.

C. OBSERVER RETROSPECTIVE ANNOTATION OF MENTAL STATES

Generally, there are two common ways of annotating mental states: self-report and observation by a third party.

Considering that self-reporting by the participants themselves may lack a certain degree of objectivity, our ultimate goal was to develop a monitoring agent that augments the perceptual ability of teachers (third party) in observing students' mental states, so we adopted the second method, annotation by third-party observation, to collect ground truth data of the students' mental states. We employed two independent annotators to do the annotation work. They included one professor and one PhD course student who both came from the same research lab as the participants but did not attend the experiment meetings. The annotators have rich experiences in annotating such kind of data through observing students' facial expressions, body movements and verbal speech. Such as they were asked to observe students' facial expressions, body language (hand gestures, eye contact) and speech cues (speed of speak, manner of speak, tempo of speak) to judge their certainty level of giving appropriate answers in Q&A session in discussion meeting and their overall confidence levels. Therefore, in this study, we employed these two annotators as external observers to annotate mental states of speaker-students by observing their facial expressions, upper body movements (since the participants were sitting, most of the body movements occurred on the upper body, such as raised hand and body leaned forward), speech cues (speed of speak, manner of speak, tempo of speak, and text content of speech).

To better implement this annotation work, a video-audio-based retrospective annotation tool was developed as shown in Fig. 4. All of the video segments (with a mean length of 10 secs for each video segment) of each student having conversations with their professor were extracted from the panorama video (10-second windows was inspired by a number of previous studies [38], [39]). Before annotation, we made clear the definitions of these four states to all of the annotators. Then, the annotators needed to watch the video segments and comprehensively observe the student's facial expressions, acoustic cues (speed of speak, manner of speak, tempo of speak), text content of speech (subtitle information of the current speech sentence displayed at the bottom of the screen), and upper body movements (hand movements, body positions). Finally, the annotators needed to annotate the student into one of the four mental states that they thought the student most clearly showed by selecting the corresponding buttons at the bottom of the screen. The annotators can repeat to watch the video as many times as they want, and if there was no clear mental state, they did not need to choose any buttons. We adopted Cohen's Kappa [40] to measure the inter-rater agreement of these two different annotators. According to the explanation of the kappa value, if it varies from 0.41 to 0.60, the agreement level is considered to be moderate, and if it falls within the range of 0.60–0.80, it is considered to indicate substantive agreement between different subjective opinions. If the kappa value is in the range of 0.81–0.99, the two annotators were considered to have almost reached perfect agreement.

FIGURE 4. Tool for annotating mental states.

D. MULTIMODAL DATASET

We recorded a total of 10 meetings, accumulating 1967 minutes worth of video-audio and physiological data with a mean length of 491 minutes for each student. There were 9507 video clips that needed to be annotated with a mean length of 10 secs for each clip. We computed the Cohen Kappa value of the judgement for each mental state between these two annotators (which we treated as a binary labeling task). We got a Cohen Kappa score of 0.64 and 0.71 for the inter-agreement level on the judgment of concentration and frustration, which suggests that these two annotators were in substantive agreement on their judgment of these states. Furthermore, we achieved a Cohen Kappa value of 0.44 for confusion and 0.50 for boredom, which indicates a moderate agreement level between the two annotators in their judgment of confusion and boredom. Finally, we obtained 1772 successful observations of mental states which received consistent judgment from the two annotators, and used that data as the ground-truth of the mental states of students in this study. Looking at the details of the data, concentration was the most common mental state observed by annotators (75.2%), followed by frustration (10.4%), confusion (9.6%), and boredom (4.8%).

IV. METHODOLOGY FOR RECOGNIZING MULTIPLE MENTAL STATES

In this section, we present how we designed multimodal analytics to develop an intelligent agent that identifies students' mental states including concentration, confusion, frustration, and boredom in conversation-based discussion activities. There are two main parts; in the first part, three types of proxy features are derived separately from three modality streams: the facial modality, heart rate modality, and acoustic modality. In the second part, we explain how we built several supervised learning classifiers based on those selected features using different modality fusion methods, signal-channel-level, feature-level, and decision-level fusion, to recognize learning-centered mental states. To reduce the dimension of the feature space and select the important features of each

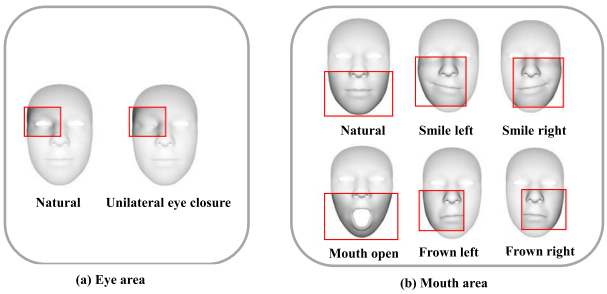


FIGURE 5. (a) Example of measurement of the movement of eye-closure and eye-opening, where the natural state is the state when the eyelid is opening with a relative coefficient of 0.00, and the maximum movement of the eyelid is the state when it is closed with a coefficient of 1.00. (b) Example of measurements of the actions of speaking and smile by calculating the movement of the lips along the vertical direction and the movement of both of the mouth corners among the four quadrants, in which the natural state of the mouth is the natural closed state without movements of the lips or mouth corners, with a coefficient of 0.00, while the respective maximum movements with a coefficient of 1.00.

modality, we apply a feature selection process (only on the training dataset) for the three kinds of extracted features separately. Leave-one-student-out cross-validation was performed to validate the recognition performance of each classifier.

A. EXTRACTION OF MULTIMODAL FEATURE SETS

1) EXTRACTING FACIAL FEATURES

As we introduced in the last section, with the aid of the ARKit library, which was integrated with the depth camera of the iPhone, various types of information regarding the students' face were detected, such as the face position and orientation, along with a series of blend shape coefficients to describe the facial expression of a recognized face in terms of the movements of specific facial features. The blend shape coefficient was a floating point number indicating the current position of the respective feature relative to its neutral configuration, ranging from 0.00 (neutral) to 1.00 (maximum movement).

Considering that previous lines of work have explored the effectiveness of using facial related features in the eye and mouth areas to analyze students' mental states in learning activities, we decided to focus on the facial cues of these areas as well. As shown in Fig.5, the related facial mesh that we adopted characterized the dynamic facial features in the eye and mouth areas. Area (a) of Fig.5, shows the measurement of the movement of closing the eyelids along the vertical direction, where the natural state is the state when the eyelid is opening (relative coefficient of 0.00), and the maximum movement of the eyelid is the state when it is closed (coefficient of 1.00). Area (b) shows measurements of the movement of the lips along the vertical direction and the movement of both of the mouth corners among the four quadrants, where the natural state of the mouth is the natural closed state without movements of the lips or mouth corners, which is detected as shown with the facial mesh in the upper left corner of (b). The facial mesh in the bottom left corner shows the mouth state in the open state when the lips move in the vertical direction. At the same time, the correlation coefficient depicting the mouth opening movement moves in the positive direction, and the maximum value is 1.00. In addition, the four facial meshes on the right side of panel (b) present measurements of the movement of the mouth corners in four quadrants, where we define the movement of the corners of the mouth in the first and fourth quadrants to represent the "smiling" state and its movement in the second and third quadrants to represent the "frowning" state. Then, based on those blend shape coefficients extracted from raw videos at an average frequency of 30.0 Hz, a sequence of dynamic facial features were derived to characterize the movement patterns of the eye and mouth. We used the first 300 frames (10 seconds) from each entire meeting video as a baseline in computing the features.

- Eye-related features: We used coefficients describing the changes in the closure of the eyelids over the left and right eyes to detect eye-blink events, which have often been used as a proxy in recognizing mental states. We took the average of the eye-lids' movement coefficient of both eyes when the Pearson's r score was equal or higher than 0.70. However, when head rotation outside this range was detected, as often happens in "in-the-wild" uncontrolled environments as in our study, we only used the movement coefficient of the visible eye. The raw eyelid-movement coefficient time series was further denoized using a Savitzky-Golay filter [41] with a window of 15 frames to remove artifacts introduced when the device occasionally lost track of faces, leading to incorrect measurements. We then applied peak detection [42] methods to detect the local maximum (peak, eye-shut) and local minimum (valley, eye-opening). Eye blinks were detected by identifying a complete cycle from open (low coefficient) to close (high coefficient) and then back to open. We filtered out fake blinks by setting a threshold of 0.50 as the minimum peak coefficient since it may indicate eye squinting and a minimum between-peak duration of 0.40s since an eye-blink cycle is around 0.40 to 0.60s. Eye-blink rate was calculated based on the identified eye-blink events as one of the eye-related features. In addition, we derived two other related features to describe the sustained duration of eye-closure and eye-opening. Presumably, when a student's concentration level is heightened, the duration for which their eyes remain open may increase, while eyes closed for a long period of time may indicate that a student is squinting or feels bored.
- Mouth-related features: Like the action of the eyes opening and closing, mouth movement dynamics may reveal students' underlying cognitive and mental processes manifested through prototypical patterns such as "smiling," which reflects a positive mental state of feeling accomplished or happy or "frowning," suggestive of a negative mental state such as confusion or frustration. We define the "smiling" state as when two corners of

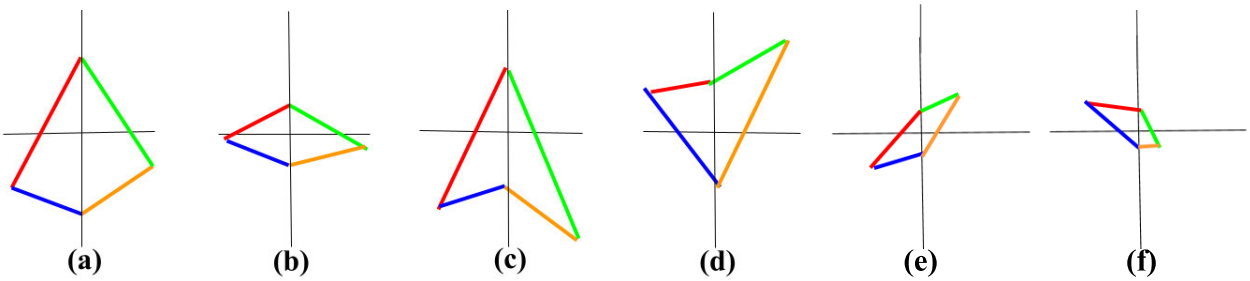


FIGURE 6. Example of mouth movement patterns observed from one student when he conversed with teacher: left and right corners of lips and middle points of upper and lower lips. (a) Mouth open, (b) mouth close, (c) frown, (d) smile, (e) the movement of the corners of the mouth is detected in the first and third quadrants, (f) the movement of the corners of the mouth is detected in the second and fourth quadrants.

the mouth respectively appear in the first and fourth quadrants with a minimum movement coefficient of 0.30, and when the movement of the corners of the mouth is detected in the second or third quadrant with a maximum movement coefficient of -0.30, we regard the state at this time as the "frowning" state. Besides these basic patterns of mouth movements, from the data itself, we also found a state of interest, that is, the corners of the mouth show a slanted line, which means that the movement of the corners of the mouth is detected in the first (or second) and third (or fourth) quadrants. Presumably, when a student is confused or frustrated with the discussion content, the movements in the mouth area would appear diagonal. To filter out fake "diagonal lines" that occur due to actions made when speaking, we set a threshold for an effective slanted line, that is, when the slope of the line between the corners of the mouth falls between -0.57 and 0.57 (The angle between the line of corners and the x-axis ranged from 30° and 150°). Then, the sustained duration of "frowning", "smiling" and "diagonal lines" were calculated by measuring the movement of both mouth corners among the four quadrants. Furthermore, it is generally believed that the visual cues that describe the actions of mouth during speaking could be used to reveal the mental sates of students in conversations. Considering that, we also measured the movement of lips along the vertical direction to capture the mouth open-close actions during speaking, and calculated the velocity and acceleration of mouth open-close actions as another two mouth-related features. In Fig. 6, we present those patterns of mouth movements observed from a student when he conversed with the teacher.

We measured eye- and mouth-related dynamic events for a given time window of 3 sec. and then computed several statistical features including mean, standard derivation (std.), max, min, range, and root mean square (RMS) over the entire video segments.

2) EXTRACTING HEART-RATE FEATURES

We detected students' heart rate (HR) from the sensor on the Apple Watch, and the data was a uni-variate continuous value within the range of 0–150 beats per minute reported at a frequency of approximately 1.0 Hz. Considering the individual differences of the participants, the first 5 minutes of HR data before each experiment was used as a baseline in computing the HR features. We first sampled the HR values to the same frequency as the facial data and then experimented with two different methods of extracting features from those values.

- Aggregated heart rate features: One of the methods was deriving a series of simple statistic features including the mean, standard deviation (std.), root mean square successive difference (RMSSD), max, min, variance, slope, mean gradient, and spectral entropy for the entire segments.
- Sequential pattern heart rate features: In the second method, we explored rich feature representations that can describe the moment-by-moment dynamic changes in the HR value using symbolic aggregate approximation (SAX) [43], [44], which was done in two steps. First, the piecewise aggregate approximation (PAA) [45] algorithm was applied to the standardized raw sampled heart-rate time series $T = \{t_1 \dots t_n\}$, with zero mean and unit variance, where T is the time of each speech video segment. We then divided the time series of length T seconds into w ($w = 5$) equal-length segments and represented the w-dimensional space with a real vector $\overline{T} = \{\bar{t}_1 \dots \bar{t}_w\}$, where the ith element of $\overline{T}$ was computed with the following (1):

$$\bar{t}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(j-1)+1}^{\frac{n}{w}i} t_j \tag{1}$$

Second, we mapped the PAA sequences of values into a finite list of symbols. The discretion threshold was chosen so that the distribution of symbols was approximately uniform. We chose an alphabet of size 3 $\{a, b, c\}$ to represent the PAA sequences to reflect the underlying dynamics of heart rate transition among three levels, i.e., low, medium, and high. In Fig. 7, we give an example of the SAX representation "cbaaa" generated from a raw heart rate time series as a way of characterizing temporal dynamic patterns. Then, we used a featurization method,

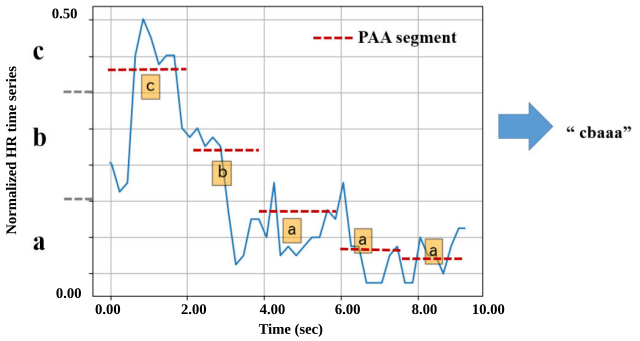


FIGURE 7. Example of sequences generated from HR time series using SAX representation.

"Bag-of-Words," where each word is a SAX pattern such as "cbaaa." Altogether, we had 243 "words" of HR SAX patterns.

3) EXTRACTING ACOUSTIC FEATURES

We used openSMILE [46] to extract audio features. OpenSMILE is often used for automatically extracting the features of audio signals and also for classifying speech and music signals. Since openSMILE is used by the OpenEAR project for emotion recognition [47], various standard feature sets for emotion recognition are available on openSMILE. We used The INTERSPEECH 2009 Emotion Challenge feature set, which contains 384 standard audio features that have been validated in terms of prediction ability regarding the task of recognizing emotional/mental states. These features are based on 16 base contours (MFCC 1–12, RMS energy, F0, zero crossing rate, and HNR) and their first derivatives (with 10-ms time windows). Features for a whole chunk were obtained by applying 12 functions [mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE)]. The entire audio data of each discussion was recorded by AirPods and stored in mp3 format. We converted it to wav format and segment them based the same start and end timestamp of each video segment we introduced before. We then used the openSMILE API to extract 384 features for the acoustic channel.

B. SUPERVISED CLASSIFICATION OF STUDENTS' MULTIPLE MENTAL STATES

We built a line of supervised learning models with three different multimodal fusion methods as shown in Fig. 8. The feature vectors for each modality are denoted as f_{hr}, f_a, and f_f, which respectively represent a set of modalities of $M = \{heart\ rate, acoustic, facial\}$. First, three baseline prediction models were separately built on the basis of individual channels: f_{hr}, f_a, and f_f. Second, we built four feature-level fusion prediction models in which we combined the three modalities together and trained a multi-label classifier called the "Combo. classifier," along with three other classifiers based on two modalities each time ($f_{hr}+f_a; f_{hr}+f_f; f_a+f_f$). Considering that if the numbers of features that we adopt from different modalities are extremely unbalanced, the modality for which more features are used will dominate the final prediction results. we performed feature selection to separately selected features for each modality and ranked them according to feature importance for prediction, which we explain in detail in the following section. We chose a similar number of features from each modality to use to build each feature-level classifier. Finally, we also built decision-fusion level classification models, in which we used three single-channel-level classifiers as base classifiers to make classifications on the same test instances separately. We then voted on the prediction results (the probability of belonging to each category) which were denoted as O_{hr}, O_a, and O_f, respectively, and the result of the base classifier with the highest decision probability was selected as the final decision of each instance. The advantage of building decision-level fusion learning models is that, even in the case that some of the modality information was corrupted due to signal noise, was missing, or could not be captured due to occlusion or sensor artifacts, etc., which often occurs in the data collecting process in "in-the-wild" environments, we could still obtain final prediction results by training the available base classifiers on the instances.

1) FEATURE SELECTION

Considering that using all features for each modality we extracted may decrease the performance of the learning prediction models, we applied RELIEF-F [48] to select features to reduce the dimensionality of raw features and extract the important features of each modality regarding the prediction tasks. We did so on training data only. RELIEF-F algorithm can deal with multi-class problems and is more robust with incomplete and noisy data. It randomly selects an instance R and then searches for k nearest instances from the same class called "nearest hits instances" as well as k nearest instances from each different class called "nearest misses." Then it updates the weight of all attributes depending on R, nearest hits, and nearest misses. A feature importance list will be returned in which features are ranked by weight. To decide the subset of features of each modality to be used, we selected several proportions used to extract a feature subset from each modality and validated the predictive performance by using each proportion of features. Due to there being 252 HR related features, 384 acoustic related features, and only 48 facial related features, we separately tested 3 different proportions of facial related features with (0.30, 0.50, 0.70), as well as 4 different proportions both of HR and acoustic features with (0.05, 0.08, 0.10, 0.15). We will report the proportions of each modality that provided the best predictive performance in the results section.

2) SUPERVISED CLASSIFIERS AND VALIDATION

Due to the class distribution of mental states being highly skewed, which is a common situation in the work regarding detecting emotional/mental states of human in the wild. We adopt Synthetic Minority Oversampling Technique

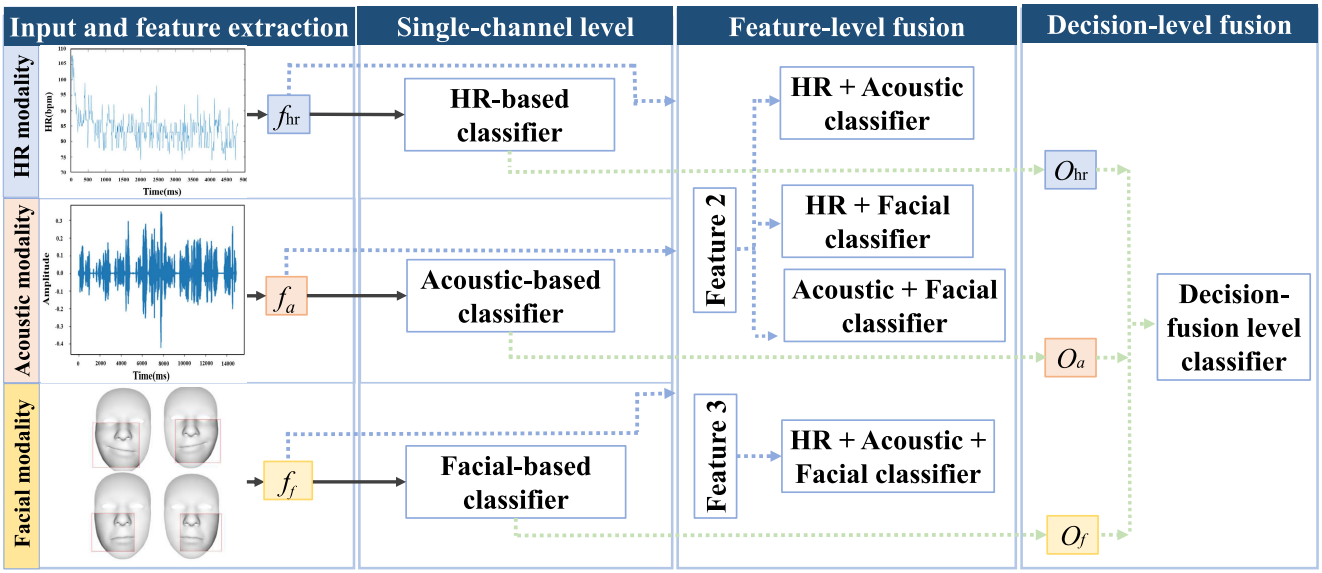


FIGURE 8. Supervised learning mental-state-prediction classifiers with different multimodal fusion approaches.

(SMOTE [49]) in the training data (we do not use it in testing data) to cope with data imbalance in order to improve model fitting. SMOTE creates synthetic instances in minority class by projecting new data points in the feature space between an instance and randomly chosen nearest within-class neighbors. We built a set of multi-class classifiers based on three kinds of supervised-learning machine learning models including support vector machine (SVM), random forest (RF), and multi-layer perceptron (MLP). We then performed leave-one-student-out cross-validation to evaluate the prediction performance of each classifier. Due to the receiver operating characteristic (ROC) curve being insensitive to changes in class distribution [50], the area under curve (AUC) of the ROC curve (chance level = 0.5) scores was used as our primary evaluation metric. We will report the aggregate AUC (Since AUC is defined only for binary classes, we calculated AUC for each class with others and average the results [50]) to measure the overall performance of each classifier in identifying all mental state classes as well as the performance of each modality fusion method in recognizing each mental state class separately. The confusion matrix for each mental state class will also be reported.

V. RESULTS

A. OVERALL PERFORMANCE

Table. 1 presents the aggregate AUC scores of each multi-class classifier in recognizing all mental states classes and the models that achieved the best overall performance (in bold) along with the feature numbers we used from each modality when the best performance was yielded.

For the overall performance of single-channel classifiers. The RF classifiers (using 300 trees, along with balanced class weights and hyperparameter optimization using randomized search with 100 iterations) did a better job both in using the single HR modality and in using the single acoustic modality to recognize students' multiple mental states by demonstrating a better identification capability. Among them, the RF classifier based on the HR modality achieved a mean AUC of 0.704, which was slighter better than the MLP classifier with a mean AUC of 0.690. In addition, there were 10 top-ranked HR related features that we used to generate the HR modality based classifiers and that achieved the best performance. Meanwhile, the RF classifier also showed outstanding performance in using acoustic cues to recognize all of the mental state classes, which yielded a mean AUC of 0.728, stronger than the SVM classifier with a mean AUC of 0.694 and better than the MLP classifier with a mean AUC score of 0.721; the first 20 top-ranked acoustic features we used to achieve the best performance. However, a difference was noticed for the facial-based single-channel classifiers, that is, the MLP model (7 layers; with active function of relu along with using cross-entropy as loss function) could learn the facial features better than the other two supervised learning models, with a mean AUC score of 0.718, a small advantage

TABLE 1. Mean AUC scores of each classifier with different modality fusion approach.

Fusion approaches	SVM	RF	MLP	No.Features
HR	0.673	**0.704**	0.690	10
Facial	0.651	0.716	**0.718**	15
Acoustic	0.694	**0.728**	0.721	20
HR+Acoustic	0.683	**0.759**	0.737	30
HR+Facial	0.680	**0.725**	0.724	25
Acoustic+Facial	0.685	**0.739**	0.731	35
Combo.	0.701	**0.763**	0.741	45
Decision-voting	0.687	0.733	**0.734**	⋆

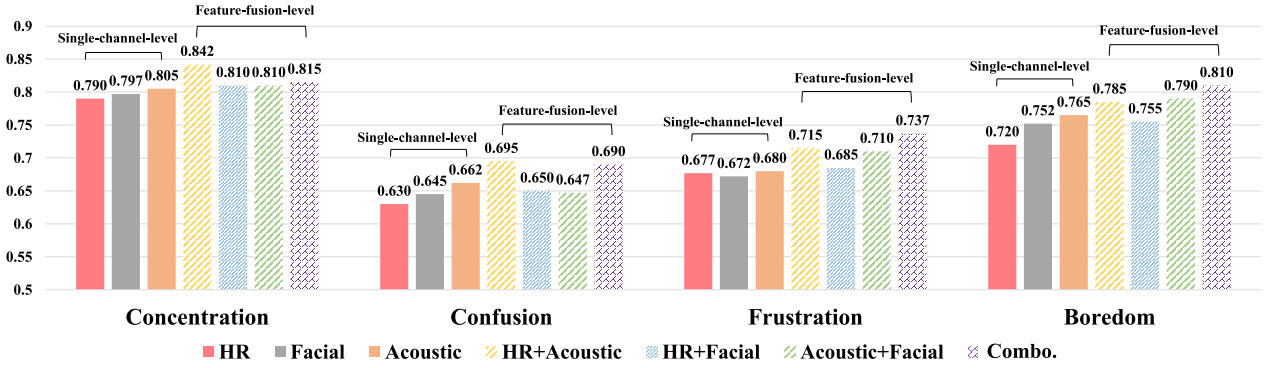


FIGURE 9. Recognition performance for each mental state class using different multimodal fusion approaches.

over the RF model, which had a mean AUC of 0.716, but stronger than the SVM model, which had a mean AUC of 0.651. For both multi-class classifiers that were built using feature-level fusion and decision-level fusion approaches, we adopted the most predictive features from each channel we introduced above to build classifiers.

For the feature-level fusion classifiers, first of all, the RF classifiers displayed an over-all outstanding classification ability for all modality fusion methods. Second, we could order these four feature-fusion methods as Combo. > HR + Acoustic > Acoustic + Facial > HR + Facial. We also noticed that fusing channels provided a significantly notable overall recognition performance improvements over each individual channel. Among them, fusing the HR and acoustic channels helped with improving the overall recognition performance by increasing the mean AUC scores by 5.5% over HR individual channel and by 3.1% over acoustic individual channel. In particular, combining the three modalities (AUC = 0.763) showed the best recognition ability over using only any single modality and any of the other combination methods in identifying the mental states of students.

For the decision-level fusion classifiers, we used a voting method on the output of each base single-channel classifier and made the final decision for the classification results. Building these classifiers will be of benefit for some inevitable situations in the data collecting process in "in-the-wild" educational environments; that is, in cases where some modality cannot be detected due to obstacle occlusion or the target students move outside the detectable area, we can still obtain final prediction results by voting on the output of each available single-channel classifier. From the results shown in the last row, we got a mean AUC score of 0.734 for the MLP classifiers and 0.733 for the RF classifier, which guarantees that our models could still work well in real-world educational settings.

B. RECOGNITION PERFORMANCE FOR EACH MENTAL STATE CLASS

One of our ultimate goals is to provide human teachers with solutions to help them design multimodal analytics to meet the requirements of identifying different mental states in different educational environments. Therefore, we examined the performance of the classifiers based on each individual channel and the classifiers based on the feature-fusion level in discriminating each mental state class as shown in Fig. 9. In addition, both of classifiers are based on the RF learning models that achieved the best performance as we presented above.

From the perspective of helping teachers augment their just-in-time decision-making capabilities, we are more interested in how to effectively recognize students' states of confusion and frustration as much as possible. As shown in the second bar group of Fig. 9, fusing the HR and acoustic modality (yellow bar, AUC = 0.695) was more effective than any of the other fusion approaches; in addition, fusing those two modalities helped with improving the prediction ability over only using the single HR modality and single acoustic modality, increasing the AUC score by 6.5% and 3.3%. Contrary to expectations, only fusing the external cues, the acoustic and facial modalities (green bar, AUC=0.647), reduced the capability of the individual acoustic channel used to recognize the state of confusion with decreasing the AUC score by 1.5%. However, adding the physiological cue of the heart rate to the combination of acoustic and facial channels reversed the decrease in AUC scores from combining only the external cues, becoming another outstanding model that contributed the second best performance in accurately identifying the state of confusion. These interesting results may suggest that, for the task of recognizing students' confusion in such interactive conversation activities, the observational information that external clues can provide is limited and confused to some extent. Instead, psychological cues can make up for this deficiency by revealing the confusion state of the students. Furthermore, combining the three modalities yielded the second best performance in recognizing the state of confusion with an AUC of 0.690, which is slightly lower than the performance of only combining HR and acoustic modality. This result provides support to our argument that we can take advantage of the replacement capabilities of different combinations between modalities to address the monitoring

challenges, such as situations in which students are wearing a mask in an offline conversation activity or have turned off their camera in an online learning environment. Our proposed monitoring agent can still recognize that students are in a state of confusion by using the combination of HR and acoustic modalities. Meanwhile, for identifying the state of frustration, the combination of three modalities achieved the best recognition performance with an AUC of 0.737 as shown in the third bar group of Fig. 9. These results indicate that multimodal data can provide complementary information to each other, which results in augmenting the overall recognition ability in identifying the mental states of students.

Furthermore, from the perspective of providing teachers with solutions regarding what kind of discussion topic students are more interested in or are prone to lose interest in order to help the teachers with adjusting and arranging their coaching strategies, we are more concerned about how to design multimodal analytics to effectively detect concentration or boredom in students during the discussion process. As shown in the first bar groups of Fig. 9, the combination of the HR and acoustic modalities (yellow bar, AUC = 0.842) showed an overall outstanding recognition ability than any of the other modality-fusion methods in recognizing the state of concentration. What is more, the fusion of the HR and acoustic modalities enabled the modalities to provide each other with additional information, enhancing the ability to recognize the state of concentration better than the classifiers using only the single HR or single acoustic modality. For recognizing the state of boredom, the best pair was the combination of the three modalities (purple bar, AUC = 0.810).

We also reported the confusion metric for each mental state class using different multimodal fusion methods, as shown in the tables. 2,3, 4,5, 6,7, 8. In each confusion metric, the value (in bold) of each row shows the percentage of inputs belonging to that class that were correctly classified.

TABLE 2. Confusion matrix for recognizing mental states using HR single-channel based classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.700**	0.230	0.060	0.010
Confusion	0.162	**0.634**	0.110	0.094
Frustration	0.040	0.130	**0.650**	0.180
Boredom	0.020	0.100	0.160	**0.720**

Similar with the results we reported previously, the HR + acoustic modalities showed a good ability in accurately identifying the states of concentration and confusion, This indicates that the combination of physiological and acoustic cues can adequately meet the requirement for identifying concentration and confusion in students in learning environments such as when most of the students' facial information cannot be detected due to the camera being turned off in a remote course or the student wearing a mask. Additionally, the combination of three modalities did a better job in accurately recognizing the states of frustration and boredom. Furthermore, we noticed that the classifier based on the fusion of three modalities, the classifier based on the fusion of acoustic and facial modalities, and the classifier based on the fusion of HR and acoustic modalities all showed quite similar capabilities in accurately identifying the state of boredom. These results provide more evidence to verify the possibility of using the replacement capabilities between different combinations of modalities to provide solutions to address the monitor challenges such as the certain modalities are unavailable in real-world educational settings.

TABLE 3. Confusion matrix for recognizing mental states using facial single-channel based classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.707**	0.132	0.112	0.049
Confusion	0.120	**0.640**	0.120	0.120
Frustration	0.101	0.210	**0.642**	0.047
Boredom	0.060	0.098	0.100	**0.742**

TABLE 4. Confusion matrix for recognizing mental states using acoustic single-channel based classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.743**	0.109	0.098	0.050
Confusion	0.110	**0.655**	0.213	0.022
Frustration	0.120	0.190	**0.659**	0.030
Boredom	0.070	0.090	0.080	**0.760**

TABLE 5. Confusion matrix for recognizing mental states using HR and acoustic modality fusion based classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.811**	0.122	0.037	0.030
Confusion	0.130	**0.690**	0.130	0.050
Frustration	0.070	0.170	**0.700**	0.060
Boredom	0.010	0.090	0.130	**0.770**

TABLE 6. Confusion matrix for recognizing mental states using HR and facial modality fusion based classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.775**	0.164	0.030	0.030
Confusion	0.100	**0.649**	0.150	0.100
Frustration	0.070	0.148	**0.682**	0.100
Boredom	0.060	0.070	0.130	**0.740**

TABLE 7. Confusion matrix for recognizing mental states using acoustic and facial modality fusion based classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.791**	0.120	0.049	0.040
Confusion	0.120	**0.600**	0.240	0.040
Frustration	0.103	0.190	**0.707**	0.000
Boredom	0.050	0.079	0.080	**0.790**

VI. APPLICATION TO EDUCATIONAL SUPPORT

In this section, We propose how to use our proposed intelligent monitoring agent from the aspects of real-world data circulation to implement educational support applications as shown in Fig. 10, as well as elaborate the novelty and contributions of this research and highlight its social value. In

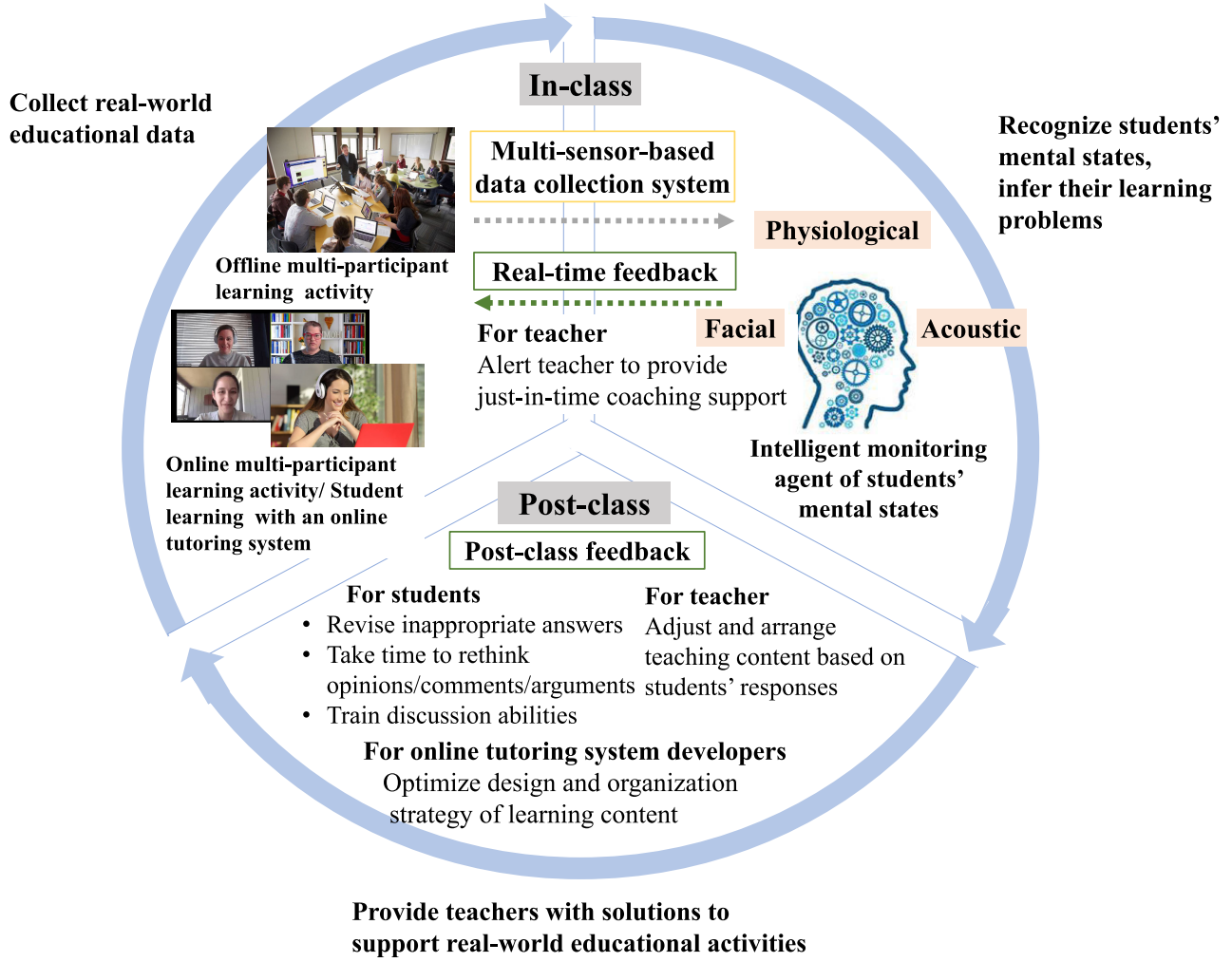


FIGURE 10. Overview of educational support system.

TABLE 8. Confusion matrix for recognizing mental states using combo. classifier.

	Concentration	Confusion	Frustration	Boredom
Concentration	**0.801**	0.122	0.067	0.010
Confusion	0.141	**0.672**	0.147	0.040
Frustration	0.070	0.130	**0.742**	0.058
Boredom	0.030	0.079	0.090	**0.800**

this study, we aim to provide solutions for teachers to address the challenges with monitoring the multiple mental states of students in real-world coaching activities involving multiple students. At present, the limitations of teaching environments and tools have brought about various obstacles for teachers in effectively coaching students. For example, for both offline or online teaching activities, capturing the mental states of all students in time to infer their learning problems and traditionally evaluating students' learning statuses (such as by observing their expressions and answers) is difficult due to the unavailability of observable information. In order to help teachers find alternative solutions to effectively identify these statuses and to solve these serious difficulties facing the educational community, our proposed intelligent monitoring agent leverages machine intelligence to effectively assist "busy" teachers as well as educators in remote coaching environments to augment their perceptual abilities so that they can more effectively monitor multiple learning-centered mental states. In particular, the agent may identify those students who are in need of assistance by recognizing that they are having difficulty learning (by detecting mental state) so that the teachers can provide appropriate alternative solutions that are adapted to the unique limitations in real-world educational settings.

For the target applications, we suggest applying the agent both for offline and online discussion learning environments involving multiple participants, as well as for the activity of students learning through an online tutoring system. Students learning with online tutor systems has been a commonly used self-learning method; however, for this case the monitoring of learning situations is often not sufficient. The lack of external intervention may lead to high drop-off rates or

other aspects of sub-optimal learning outcomes. As shown in Fig.10, We firstly suggest using the proposed multi-sensor-based data collecting system to collect multimodal data including heart rate, facial, and audio data from students who taking part in those learning activities. Then, we input this information into the proposed mental-state recognition models to detect the mental states of students. For the cases of multi-participant coaching activities, our proposed mental state detector can not only assist ''busy'' teachers to effectively monitor multiple students' mental states, but also take advantage of the replacement capabilities between different combinations of modalities. This may provide human teachers with alternate solutions for addressing the challenges with monitoring students when in online learning activities, the visual or audio modalities may not be available. Similarly, in offline learning activities, the facial modality is sometimes not available when students wear masks. Additionally, the proposed system could also act as a virtual teacher with high practicality to support the student who learn through an online tutoring system.

Furthermore, we also suggest several feedback functions for helping cultivate learning outcomes. We designed a real-time feedback mechanism to alert teachers of the negative states of students during learning, such as when students are confused with opinions or when they start feeling hopeless or frustrated with the learning content, helping the teacher make just-in-time decisions. This may help to avoid negative mental states, which, if left to persist through a lack of external intervention, could result in students losing interest and motivation in learning, hindering learning progress and decreasing learning outcomes.

In addition, we do not recommend always giving feedback to students in real time because this will interfere with their thinking process. In those cases, the unresolved contents can be returned as post-class materials. For the purpose of training, we hope to summarize and give feedback on students' unresolved content after a discussion and encourage them to spend much time reconsidering the appropriate answers and thinking about the opinions/comments that they did not understand during a discussion. Another type of post-class feedback we would like to provide to teachers is to help them discover which discussion content the students showed a high level of concentration for and where the students became bored and sleepy. This will assist teachers in effectively designing and arranging the lecture items or discussion strategies to increase the students' interest in learning as much as possible. Beyond providing post-class feedback to students and teachers, we also suggest designing post-class feedback functions for online learning tutoring system developers regarding what kind of learning content students show clear concentration for with a high level of interest and for what kind of learning content students lose motivation. We would like to feed back such information to system developers to provide them with solutions for optimizing the learning components.

VII. CONCLUSION

In this study, we aimed to leverage machine intelligence to generate an intelligent monitoring agent to provide teachers with solutions to challenges with monitoring students regarding the recognition of multiple mental states including concentration, confusion, frustration, and boredom in different educational environments. We proposed that multimodal data can be used to augment the practicality of the monitoring agent. To validate our arguments, we explored how to effectively design multimodal analytics to improve the ability to identify students' specific mental states, as well as how to take advantage of the supplementation and replacement capabilities between modalities to address challenges when modalities cannot be collected, in order to improve the practicality of the monitoring agent.

To achieve these goals, we first took advantage of modern sensor technologies to accumulate and archive a massive multi-modal dataset, for which we used the Apple Watch for real-time heart-rate data detection, integrated the ARKit framework and iPhone front-camera to track and collect facial motion signals, and AirPods with pin microphones to record the audio of discussions. We used our data collection system to record and accumulate an ''in-the-wild'' conversation-based discussion dataset generated between a teacher and students in a real university research lab. We then derived a series of interpretable proxy features from visual, physiological, and audio modalities separately to characterize multiple mental states. For visual, we extracted lines of facial related features to describe dynamic patterns of eye blinking and mouth movements (speaking and smiling). For the physiological modality, in addition to the use of statistic features, we also attempted to capture moment-by-moment temporary patterns from heart-rate time-series data by extracting SAX HR sequences. For audio, we used the openSMILE tool to compute numbers of features for capturing students' mental states from acoustic cues. Then, we trained a set of supervised learning SVM, RF, and MLP classifiers separately using different multimodal fusion approaches including single-channel-level, feature-level, and decision-level fusion for recognizing students' multiple mental states.

From the results of this study, we suggest taking advantage of the combination of HR + acoustic to better recognize the states of concentration and confusion. In addition, using this combination to develop a monitoring agent can also overcome such challenges when students' facial modality cannot be detected due to masks being worn or cameras turned off in online learning environments. Furthermore, fusing three modalities can better recognize the states of frustration and boredom. What is more, the results also indicate the possibility of recognizing boredom in students by using only the acoustic and facial fusion based classifier or the HR and acoustic fusion based classifier. These results provided experimental evidence in support of our arguments that using effectively designed multimodal analytics by taking advantage of the supplementation and replacement capabilities between modalities makes it possible

to adapt to different challenges in real-world education environments.

A. LIMITATIONS AND FUTURE WORK

We would like to point out the limitation of this study and propose the future work. The main limitation is collecting enough reliable ground truth data samples on the mental states of students when they are in real-world discussion activity. In this work, we selected those data samples that matched two experienced annotators as the ground truth data. However, we noticed the relative low inter-rater agreement among the annotators on some specific mental state of students, such as the state of boredom. We think that when students are completing discussion with their professor in a real-world learning activity, some of the states, especially the state of boredom, is hard to identify, which makes it difficult for annotators to form a unified judgment standard for judging the bored state. This is an inevitable limitations brought from ''in-the-wild'' data. To address this limitation, we would like to make such attempts in the future work, (1) adopt the affect detection protocol such as Baker-Rodrigo Observation Method Protocol (BROMP) to pre-train the annotators before encoding the data in order to improve the low inter-rater agreement; (2) use the self-report method along with the third party observations since the students will be invited to watch their discussion video and report their mental states using our developed annotation tool.

ACKNOWLEDGMENT

The authors would like to thank Asst. Prof. Shigeki Ohira for his help in collecting the ground truth measures of the experiments, and the members from Nagao Lab for participating in the data collection process. They would also like to thank Andy Gao for his help in proofreading the contents of this article.

REFERENCES

[1] S. D'Mello, S. Craig, K. Fike, and A. Graesser, ''Responding to learners' cognitive-affective states with supportive and shakeup dialogues,'' in *Proc. Int. Conf. Hum.-Comput. Interact.* Berlin, Germany: Springer, 2009, pp. 595–604.

[2] M. Rodrigo, R. Baker, S. Mello, ''Comparing learners' affect while using an intelligent tutoring system and a simulation problem solving game,'' in *Proc. Int. Conf. Intell. Tutoring Syst.* Berlin, Germany: Springer, 2008, pp. 40–49.

[3] J. Robison, S. McQuiggan, and J. Lester, ''Evaluating the consequences of affective feedback in intelligent tutoring systems,'' in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–6.

[4] K. Forbes-Riley and D. Litman, ''When does disengagement correlate with learning in spoken dialog computer tutoring?'' in *Proc. Int. Conf. Artif. Intell. Educ.* Berlin, Germany: Springer, 2011, pp. 81–89.

[5] R. A. Calvo and S. K. D'Mello, *New Perspectives on Affect and Learning Technologies*, vol. 3. New York, NY, USA: Springer, 2011.

[6] M. M. T. Rodrigo and R. S. D. Baker, ''Comparing the incidence and persistence of learners' affect during interactions with different educational software packages,'' in *New Perspectives on Affect and Learning Technologies*. New York, NY, USA: Springer, 2011, pp. 183–200.

[7] S. Craig, A. Graesser, J. Sullins, and B. Gholson, ''Affect and learning: An exploratory look into the role of affect in learning with AutoTutor,'' *J. Educ. Media*, vol. 29, no. 3, pp. 241–250, Oct. 2004.

[8] A. Graesser, P. Chipman, B. King, B. McDaniel, and S. D'Mello, ''Emotions and learning with auto tutor,'' *Frontiers Artif. Intell. Appl.*, vol. 158, p. 569, Dec. 2007.

[9] S. D'Mello and A. Graesser, ''Dynamics of affective states during complex learning,'' *Learn. Instruct.*, vol. 22, no. 2, pp. 145–157, Apr. 2012.

[10] M. R. Lepper, M. Woolverton, D. L. Mumme, and J. Gurtner, ''Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors,'' *Comput. Cogn. Tools*, vol. 1993, pp. 75–105, 1993.

[11] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, ''Embodied affect in tutorial dialogue: Student gesture and posture,'' in *Proc. Int. Conf. Artif. Intell. Educ.* Berlin, Germany: Springer, 2013, pp. 1–10.

[12] D. Bohus and E. Horvitz, ''Models for multiparty engagement in open-world dialog,'' in *Proc. SIGDIAL Conf. 10th Annu. Meeting Special Interest Group Discourse Dialogue*, 2009, pp. 225–234.

[13] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. K. D'Mello, ''Affect detection from multichannel physiology during learning sessions with autotutor,'' in *Proc. Int. Conf. Artif. Intell. Educ.* Berlin, Germany: Springer, 2011, pp. 131–138.

[14] S. Peng, L. Chen, C. Gao, and R. J. Tong, ''Predicting students' attention level with interpretable facial and head dynamic features in an online tutoring system (student abstract),'' in *Proc. AAAI*, 2020, pp. 13895–13896.

[15] N. Bosch, S. K. D'Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, ''Detecting student emotions in computer-enabled classrooms,'' in *Proc. IJCAI*, 2016, pp. 4125–4129.

[16] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, ''Automated detection of engagement using video-based estimation of facial expressions and heart rate,'' *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, Jan. 2017.

[17] M. E. Hoque, D. J. McDuff, and R. W. Picard, ''Exploring temporal patterns in classifying frustrated and delighted smiles,'' *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 323–334, Jul. 2012.

[18] B. B. de Koning, H. K. Tabbers, R. M. J. P. Rikers, and F. Paas, ''Attention guidance in learning from a complex animation: Seeing is understanding?'' *Learn. Instruct.*, vol. 20, no. 2, pp. 111–122, Apr. 2010.

[19] J. Gomes, M. Yassine, M. Worsley, and P. Blikstein, ''Analysing engineering expertise of high school students using eye tracking and multimodal learning analytics,'' in *Proc. Conf. Educ. Data Mining*, 2013, pp. 1–5.

[20] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester, ''Automatically recognizing facial expression: Predicting engagement and frustration,'' in *Proc. Educ. Data Mining*, 2013, pp. 1–5.

[21] N. Bosch, S. K. D'Mello, J. Ocumpaugh, R. S. Baker, and V. Shute, ''Using video to automatically detect learner affect in computer-enabled classrooms,'' *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 2, pp. 1–26, Aug. 2016.

[22] M. N. Levy and P. J. Schwartz, *Vagal Control of the Heart: Experimental Basis and Clinical Implications*. Nordrhein-Westfalen, Germany: Futura Publishing Company, 1994.

[23] A. Camm, M. Malik, J. Bigger, G. Breithardt, and S. Cerutti, ''Heart rate variability: Standards of measurement, physiological interpretation and clinical use. Task force of the European society of cardiology and the North American society of pacing and electrophysiology,'' in *Circulation*, vol. 93. Dallas, TX, USA, 1996, pp. 1043–1065.

[24] U. R. Acharya, K. P. Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, ''Heart rate variability: A review,'' *Med. Biol. Eng. Comput.*, vol. 44, no. 12, pp. 1031–1051, Dec. 2006.

[25] J. Hellhammer and M. Schubert, ''The physiological response to trier social stress test relates to subjective measures of stress during but not before or after the test,'' *Psychoneuroendocrinology*, vol. 37, no. 1, pp. 119–124, Jan. 2012.

[26] T. Pereira, P. R. Almeida, J. P. S. Cunha, and A. Aguiar, ''Heart rate variability metrics for fine-grained stress level assessment,'' *Comput. Methods Programs Biomed.*, vol. 148, pp. 71–80, Sep. 2017.

[27] S. Mukherjee, R. Yadav, I. Yung, D. P. Zajdel, and B. S. Oken, ''Sensitivity to mental effort and test retest reliability of heart rate variability measures in healthy seniors,'' *Clin. Neurophysiol.*, vol. 7, pp. 2059–2066, Apr. 2011.

[28] S. Peng, S. Ohira, and K. Nagao, ''Prediction of students' answer relevance in discussion based on their heart-rate data,'' *Int. J. Innov. Res. Educ. Sci.*, vol. 6, no. 3, pp. 414–424, 2019.

[29] R. H. Stevens, T. Galloway, and C. Berka, ''Eeg-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills,'' in *Proc. Int. Conf. User Modeling*. Berlin, Germany: Springer, 2007, pp. 187–196.

[30] B. Cowley, N. Ravaja, and T. Heikura, "Cardiovascular physiology predicts learning effects in a serious game activity," *Comput. Edu.*, vol. 60, no. 1, pp. 299–309, Jan. 2013.
[31] C. D. B. Luft, G. Nolte, and J. Bhattacharya, "High-learners present larger mid-frontal theta power and connectivity in response to incorrect performance feedback," *J. Neurosci.*, vol. 33, no. 5, pp. 2029–2038, Jan. 2013.
[32] K. B. Burt and J. Obradovi, "The construct of psychophysiological reactivity: Statistical and psychometric issues," *Develop. Rev.*, vol. 33, no. 1, pp. 29–57, Mar. 2013.
[33] J. M. Reilly and B. Schneider, "Predicting the quality of collaborative problem solving through linguistic analysis of discourse," *Int. Educ. Data Mining Soc.*, 2019.
[34] K. Kitto, M. Hatala, and G. Siemens, "Towards automated content analysis of discussion transcripts: A cognitive presence case," in *Proc. 6th Int. Conf. Learn. Anal. Knowl.*, 2016, pp. 15–24.
[35] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect Emotion Human-Computer Interaction*. Berlin, Germany: Springer, 2008, pp. 92–103.
[36] L. Chen, X. Li, Z. Xia, Z. Song, L.-P. Morency, and A. Dubrawski, "Riding an emotional roller-coaster: A multimodal study of young child's math problem solving activities," *Proc. Int. Educ. Data Mining Soc.*, 2016, pp. 1–8.
[37] R. Wampfler, S. Klingler, B. Solenthaler, V. Schinazi, and M. Gross, "Affective state prediction in a mobile setting using wearable biometric sensors and stylus," in *Proc. The 12th Int. Conf. Educ. Data Mining (EDM 2019)*, 2019, pp. 198–207.
[38] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014.
[39] S. D'Mello, R. Dale, and A. Graesser, "Disequilibrium in the mind, disharmony in the body," *Cognition Emotion*, vol. 26, no. 2, pp. 362–374, Feb. 2012.
[40] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 4, pp. 159–174, Mar. 1977.
[41] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares Procedures.," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.
[42] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, Sep. 2006.
[43] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proc. 8th ACM SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery*, 2003, pp. 2–11.
[44] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining Knowl. Discovery*, vol. 15, no. 2, pp. 107–144, Aug. 2007.
[45] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM Trans. Database Syst.*, vol. 27, no. 2, pp. 188–228, 2002.
[46] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[47] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1–8.
[48] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1994, pp. 171–182.
[49] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[50] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

SHIMENG PENG received the M.S. degree in information systems from the Department of Intelligent Systems, Graduate School of Informatics, Nagoya University, in 2018, where she is currently pursuing the Ph.D. degree. Her research interests include multiple wearable sensor signals, including heart rate, facial, and acoustic, with machine learning methods to detect and understand students' complex mental states in education activities for the purpose of improving their learning outcome.

KATASHI NAGAO received the B.E., M.E., and Ph.D. degrees in computer science from the Tokyo Institute of Technology, in 1985, 1987, and 1994, respectively. Since 1987, he has been researching natural language processing and machine translation systems with the IBM Research, Tokyo Research Laboratory. In 1991, he began conducting research projects on natural language dialogue, multiagent systems, and human–computer interaction at Sony Computer Science Laboratories, Inc. From 1996 to 1997, he was a Visiting Scientist with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, USA. He rejoined IBM's Tokyo Research Laboratory and launched the Semantic Transcoding Project in 1999. He joined Nagoya University as an Associate Professor at the Graduate School of Engineering in 2001. Since 2002, he has also been researching artificial intelligence and computer-assisted education as a Professor with the Graduate School of Information Science, Nagoya University. The Graduate School of Information Science was reorganized into the Graduate School of Informatics in 2017.