Retraction Notice

Title of retracted article: Recognizing Student's Learning-Centered Affective States in Conversation with Intelligent Multimodal Analytics

Author(s): Shimeng Peng, Shigeki Ohira, Katashi Nagao

* Corresponding author. Email:hou@nagao.nuie.nagoya-u.ac.jp

Journal: Creative Education
Year: 2020
Volume: 11
Number: 9
Pages (from - to): 1697-1719
DOI (to PDF): http://dx.doi.org/ 10.4236/ce.2020.119124
Paper ID at SCIRP: 102945
Article page: http://www.scirp.org/Journal/PaperInformation.aspx?PaperID=102945

Retraction date: 2020-11-19

Retraction initiative (multiple responses allowed; mark with **X**):
☐ All authors
X Some of the authors:
☐ Editor with hints from
- ○ Journal owner (publisher)
- ○ Institution:
- ○ Reader:
- ○ Other:

Retraction type (multiple responses allowed):
☐ Unreliable findings
- ○ Lab error
- ○ Inconsistent data
- ○ Analytical error
- ○ Biased interpretation
- ○ Other:

☐ Irreproducible results
☐ Failure to disclose a major competing interest likely to influence interpretations or recommendations
☐ Unethical research

☐ Fraud
- ○ Data fabrication
- ○ Fake publication
- ○ Other:

☐ Plagiarism
☐ Self plagiarism
☐ Overlap
☐ Redundant publication *
☐ Copyright infringement
☐ Other legal concern:

☐ Editorial reasons
- ○ Handling error
- ○ Unreliable review(s)
- ○ Decision error
- ○ Other:

X Other:

Results of publication (only one response allowed):
X are still valid.
☐ were found to be overall invalid.

Author's conduct (only one response allowed):
☐ honest error
☐ academic misconduct
X none (not applicable in this case – e.g. in case of editorial reasons)

* Also called duplicate or repetitive publication. Definition: "Publishing or attempting to publish substantially the same work more than once."

History
Expression of Concern:
☐ yes, date: yyyy-mm-dd
X no

Correction:
☐ yes, date: yyyy-mm-dd
X no

Comment:

The paper is withdrawn from "Creative Education" due to its indexing databases that don't meet the author's PhD examination.

This article has been retracted to straighten the academic record. In making this decision the Editorial Board follows COPE's Retraction Guidelines. The aim is to promote the circulation of scientific research by offering an ideal research publication platform with due consideration of internationally accepted standards on publication ethics. The Editorial Board would like to extend its sincere apologies for any inconvenience this retraction may have caused.

Editor guiding this retraction: Anita LIU (Editorial Assistant of CE)

Creative Education, 2020, 11, 1697-1719
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Recognizing Student's Learning-Centered Affective States in Conversation with Intelligent Multimodal Analytics

Shimeng Peng[1], Shigeki Ohira[2], Katashi Nagao[1]

[1]Department of Intelligent Systems, Graduate School of Informatics, Nagoya University, Nagoya, Japan
[2]Information Technology Center, Nagoya University, Nagoya, Japan
Email: {hou, nagao}@nagao.nuie.nagoya-u.ac.jp, ohira@nagoya-u.jp

How to cite this paper: Peng, S. M., Ohira, S., & Nagao, K. (2020). Recognizing Student's Learning-Centered Affective States in Conversation with Intelligent Multimodal Analytics. *Creative Education, 11,* 1697-1719.
https://doi.org/10.4236/ce.2020.119124

Received: August 20, 2020
Accepted: September 15, 2020
Published: September 18, 2020

Abstract

Coach-led conversations are the most common way for teachers to monitor students' learning progress and provide just-in-time interventions. Students will experience a complex mixture of learning-centered affective states during the conversations, including concentration, confusion, frustration, and boredom, which have been widely acknowledged as crucial components for inferring a student's learning states. Effectively recognizing students' learning-centered affective states, especially negative ones such as confusion and frustration, could indicate that a student has a need for assistance, and help a human teacher improve their perceptual and real-time decision-making capability in providing personalized and adaptive support in coaching activities. Many lines of research have explored the automatic measurement of students' single emotional states in pre-designed student vs. computer-teacher tasks. It still remains a challenge to detect the learning-centered affective states of students in face-to-face teacher-student coach-led conversations. Meanwhile, "in-the-wild" contexts with real operational environments and real teacher-student conversations pose unique challenges in collecting, validating, and interpreting data. In this study, we attempted to first develop an advanced multi-sensor-based system and applied it in small-scale meetings to collect multi-modal teacher-student conversation data. Then, we demonstrate a multimodal analysis framework to characterize students' learning-centered affective states from multiple perspectives: facial, audio, and physiological cues. A series of interpretable proxy features were derived from these modalities and used to train a set of supervised learning classifiers with various multimodal fusion approaches, signal-channel-level, feature-fusion-level, and decision-fusion-level, to recognize students' learning-centered affective states. We achieved a mean AUC of 0.76 for the facial and audio feature-level fusion classifier. Our results

provide evidence of the potential practical value of fusing multi-modal data to explore students' "in-the-wild" learning-centered affective states in teacher-student coach-led conversations.

Keywords

Learning Analytics, Multimodal Analytics, Machine Learning, Heart Rate, Facial Expression, Audio Features

1. Introduction

Conversation-based coaching discussion is one form of typical complex learning that requires students to answer casual questions, generate explanations, solve problems, and demonstrate and transfer acquired knowledge (Graesser, Ozuru, & Sullins, 2010). A broad array of learning-centered affective states are always aroused, accompanied with complex learning processes, such as concentration/engagement, anxiety, delight, confusion, frustration, and boredom (D'Mello, Craig, Fike, & Graesser, 2009; Forbes-Riley & Litman, 2011; Robison, McQuiggan, & Lester, 2009; Rodrigo & Baker, 2011a; Calvo & D'Mello, 2011; Rodrigo & Baker, 2011b). In the past few years, much research has validated the correlation of students' learning-centered affective states with measures of their short-term or long-term learning achievements (Pardos et al., 2014; Rodrigo et al., 2012; Calvo & D'Mello, 2010).

To more precisely identify the learning-centered affective states that students may experience during complex learning processes and to gain more insight in the future into the interplays and influence between those states with learning outcomes, a number of pieces of research are beginning to emerge on defining and describing emotions in the education field. Among the research, the dynamic affective states model proposed by D'Mello and Graesser in 2012 (D'Mello & Graesser, 2012) has often been used as a classical theoretically grounded model for intuitively understanding the dynamic changes in a student's learning-centered affective states when they complete complex learning activities such as conversation-based discussion. In this theory, they suggest that a student commonly enters complex learning activities with a state of engaged concentration to avoid failure when they are anxious, and this state will remain until they make mistakes or 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 concentrated if the impasse has been resolved, which can be due to positive accomplishments brought about by solving problems 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 may be more likely to transition to boredom if the state of frustration

persists and finally abandon the pursuit of their learning goals. It has been validated that students experiencing the flow of confusion->concentration are positively related with their learning outcomes, and vice versa, the confusion->frustration->boredom flow has a negative relationship with their learning outcomes in complex learning activities (Craig, Graesser, Sullins, & Gholson, 2004; Graesser et al., 2007).

It is still an open question as to how to measure students' learning-centered affective states, including concentration, confusion, frustration, and boredom, when they are having a coaching-driven conversation with their teacher. Most research has been focused on detecting students' affective states 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 (Peng, Chen, Gao, & Tong, 2020a; Grafsgaard et al., 2013; Bosch et al., 2016; Zaletelj & Košir, 2017). The unique challenges that students face in completing face-to-face "conversation tasks" involves their fast and accurately giving answers and explanations or searching for the knowledge required to handle a teacher's unpredictable questions. Having coach-led conversations with students has been the preferred way for teachers to examine the learning status of their students, and, at the same time, it is also vital for teachers to be sensitive enough to capture students affective states or infer the latent need for assistance in order to make more precise real-time decisions on what kind of support to provide and at what times.

While the work mentioned above focuses on detecting a student's single affective state such as in terms of engagement, a substantial amount of prior work measured students' affective states using univariate modality signals such as video (Grafsgaard et al., 2013), audio (Forbes-Riley & Litman, 2011), and physiological measures (Hussain et al., 2011). Modern sensors have rendered opportunities to support novel methodological approaches to measure students' multiple affective states from various perspectives and have been explored to improve recognition accuracy.

In this study, we attempt to adopt a multimodal analytic approach to recognize multiple learning-centered affective states in students including concentration, confusion, frustration, and boredom when they are having coach-based conversations with their teacher. To achieve this goal, we first develop a multi-sensor-based data collection system to record students' video-audio and physiological data as they have conversations with their teacher. Then, a multimodal analysis framework is demonstrated, in which a series of interpretable features is extracted from those modalities to characterize the students' multiple affective states from several aspects. These features are then used to generate a set of supervised learning models using different modality-fusion methods, that is, a single-channel level, feature-fusion level, and decision-fusion level. AUC scores are used to evaluate the accuracy of each classifier, using leave-one-student-out cross validation with the purpose of validating the predictive performance of our

extracted multimodal features and each modality fusion method.

Novelty and Contributions

There are several novel contributions we make and from aspects that are preliminarily different from relevant studies; simply put, 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 "unplugged" scenario in which students and their advisor teacher have a coaching-driven conversation on real learning activities. We tracked and recorded real-world conversations between students and their teacher in a weekly face-to-face meeting held by a university lab for up to 3 months. These conversations included the complete process by which the teacher checked the students' latest learning progress by asking questions, seeking explanations for uncertain contexts, and confirming the details of their work and made decisions in terms of where to direct coaching after listening to the students' answers. Therefore, our study innovatively aims to analyze a series of "true feelings" exposed during these real conversations, guaranteeing the applicability and practicality of our results for real-world coaching activities. 2) A multi-sensor-based data collection system is developed and applied in a small group meeting held in a university's lab, in which we recorded multimodal conversation data by a) using the iPhone to track the video-audio information of each meeting participant and b) using the Apple Watch to detect their heart rate (HR) signals. The audio information of the entire conversation was then transcribed into statements by Google Cloud Speech-to-Text. Our multimodal data collection system could support a 2 - 3-hour-long group meeting composed of conversation activities held amongst multiple participants. Multiple types of conversation data from participants including video-audio, textual, and physiological data were synchronized, captured, and stored structurally, and this shows the potential utility of our system in collecting multimodal datasets over a long period of time as well as in supporting the analysis of real-world teacher-student conversations. 3) With few exceptions, most existing work has focused on using a univariate modality to analyze a single learning-centered affective state, engagement, or basic emotional states such as joy and sadness. In comparison, this study attempts to integrate multiple modalities, that is, facial, audio, and physiological (heart rate) cues, to predict multiple learning-centered affective states, that is, concentration, confusion, frustration, and boredom, when students are having conversations with their teacher. In addition to that, lines of interpretable features were extracted from those multiple modalities to gain insight into each affective state. We trained several prediction models on these features by using different fusion methods and evaluated their prediction performance. Our results would provide evidence of the potential utility of the presented multimodal features and modality fusion methods in scaling up the analysis and recognition of students' multiple learning-centered affective states and guide teachers towards optimal instruction in real time.

2. Related Work

Most previous research on recognizing learning-centered affective states has focused on engaged concentration prediction tasks. From early studies in which researchers used univariate modalities (Bohus & Horvitz, 2009; Bosch et al., 2015), to most recently, with the emergence of modern sensors that provide support for recording multimodal data (MMD) in daily life, multimodal learning analytics (MMLA) has been frequently applied to analyzing emotions in education, such as (Monkaresi et al., 2016) using facial and heart rate cues to predict students' engagement when they completed writing tasks in a computer environment. (Chen et al., 2016) 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 conversational cues in order to characterize the child's affective states including confusion, frustration, joy, and engagement demonstrated during learning activities. (Peng, Chen, Gao, & Tong, 2020a) also integrated multiple modalities of facial and EEG signals from a group of middle school students to describe their engaged attention when they were interacting with an online learning tutor system.

2.1. Facial-Signal-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 affective states. (Kapoor, Burleson, & Picard, 2007) adopted several modalities including facial features to predict students' frustration with automated learning system. (Hoque et al., 2012) also computed a set of facial and other features from videos to predict if smiles indicated frustration or delight. (De Koning et al., 2010; Gomes et al., 2013) employed eye-related features from facial signals like blinking and gaze to analyze students' concentration states during learning activities. (Devillers & Vidrascu, 2007) characterized human smiles and laughter by monitoring mouth-noise related features. (Grafsgaard et al., 2013) used mouth features to predict overall levels of concentration, frustration, and learning gain. In our previous work, we also demonstrated how we used a number of facial-related features extracted from the eye and mouth areas to predict students' learning-centered affective states (Peng, Ohira, & Nagao, 2020b).

2.2. Physiological-Signal-Based Detection

More recent work in this space has been able to accurately predict student's learning-centered or simply basic affective states when they are engaged in learning activities. Affective states are generally considered to be related to thoughts and feelings controlled by the autonomic nervous system, and their changes can be observed through physiological signals such as heart rate (HR) and brain waves. This theoretical fact makes the ECG (heart rate) or EEG signal the most widely used clue in the work of affective state detection. (Stevens et al., 2007) employed heart rate signals in a prediction task regarding students' en-

gagement in a student-computer interactive learning environment. In our previous work (Peng, Ohira, & Nagao, 2018; Peng, Ohira, & Nagao, 2019), 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 heart rate features. Several pieces of work (Stevens et al., 2007; Cowley et al., 2013; Luft et al., 2013; Burt & Obradović, 2013; Peng et al., 2020a) analyzed brainwave EEG signals to understand the affective states of students during the learning process.

2.3. Audio-Signal-Based Detection

It is widely believed that affective-state information may be transmitted from speech signals and can be explicated from linguistic and audio channels. Emotion recognition in conversations (ERC) has become one of the hottest topics in the NLP field and is gaining increasing attention from the community. (Sikka et al., 2013) proposed a method that combines audio features with other visual descriptors to automatically detect seven emotion categories from video clips: anger, disgust, fear, happiness, neutral, sadness, and surprise. (Castellano, Kessous, & Caridakis, 2008) 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. (Yoon et al., 2019) used a deep learning method to exploit the textual and acoustic data of an utterance for emotion classification tasks in speech.

2.4. Present Approach

We attempted to adopt a multimodal analytics approach to recognizing students' multiple learning-centered affective states when they are having conversations with their teachers from the aspect of visual, audio, and physiological (heart rate) cues. To achieve this goal, different from the previous work, we employed a set of portable, commercial wearable sensors for collecting multiple-modality data; we used ARKit running on the iPhone to track the visual movement records of students, and AirPods were used as an audio data recording tool for the entire conversations, using Apple Watch to simultaneously measure the changes in the students' heart rate data throughout the conversations.

We used a combination of dynamic visual facial, audio, and heart rate features extracted from video-audio and physiological records. We then trained several predictive machine learning models using different multimodal fusion methods and evaluated the predictive performance of each classifier on a personal level. We discuss the feature importance for each modality and compare the distinct recognition abilities for each multimodal fusion method.

3. Methodology for Collecting Multimodal Conversation Dataset

3.1. Participants and Study Protocol

The participants were 4 graduate students (one female and three male) and their

advisor professor, and 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. These participants have become accustomed to this form of coach-led conversations so that they would not be disturbed by wearing contact devices during the data collection process, therefore ensuring the reliability of our results. Data for our multimodal dataset was collected on the basis of these students when they had a conversation with the professor, in which they reported their weekly research progress separately and the professor then initiated coaching-driven conversations with them.

3.2. Data Collection System and Procedure

A regular research-progress-report meeting was held in the university's research lab, in which students reported their latest research progress separately while displaying content related in the form of a slide presentation. There would be a continuous conversation between the current presenter-student and advisor professor, in which the professor started such as by asking questions, requiring detailed explanations regarding the content being presented, and the student would give corresponding responses. One regular meeting always took around 3 hours in total, with an average length of around 50 minutes for each student report, which included a 10-min presentation and a 30 - 40-min. conversation chunk. Each student's conversation chunk would be carried out only between the current presenter-student and the advisor professor.

Before the meeting, as shown in **Figure 1**, all participants were asked to initiate a face tracking function that ran on the iPhone XR, which was placed on a desk in front of each of them, by choosing their name and pressing the record button; the function was developed using ARKit. A paired Apple Watch, which they wore on their wrist, was started at the same time to detect changes in their heart rate. AirPod earphones were also worn in order to collect the audio data generated during the entire conversation between the students and the advisor professor. After the meeting, the video-audio and heart-rate data of each student were then stored as csv and mp3 files listed with the username and date of the experiment on their iPhone and then were transferred to a server for analysis with the permission of each participant.

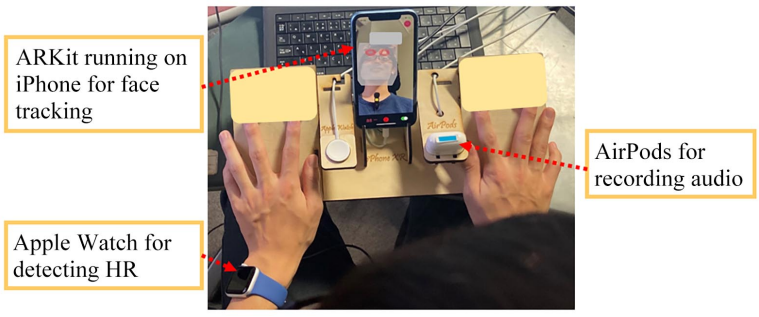

Figure 1. Multiple sensors for collecting conversation data.

3.3. Observer Annotation of Affective States

We recorded panorama video in 360 degrees using a Ricoh THETA, set between students during the whole meeting. As shown in Figure 2, a video-audio-based annotation tool was developed, in which the panorama video was played with corresponding conversation **subtitles** presented on the bottom of the screen, and subtitles appeared for each sentence. Before the annotation work, the annotators were asked to choose a video to be annotated along with its corresponding subtitles file in csv format, as shown in the top-left area in Figure 2. When a video is played, the screen automatically switches to the current speaker. The annotator needs to comprehensively observe the speaker's facial expressions and his or her speech information, including audio cues and the text of the current speech displayed at the bottom of the screen in order to make a suitable judgment regarding the speaker's affective state shown while they speak. Then, the annotator chooses one of the buttons representing four affective states at the bottom of the screen, and if there is no clear affective state, they do not need to choose any buttons. We employed two independent annotators to complete the annotation work. They were one professor and one PhD course student who both came from the same research lab as the participants but did not attend the meetings. We adopted Cohen's Kappa (Landis & Koch, 1977) to measure the inter-rater agreement of these two different annotators. If the kappa varied from 0.41 to 0.60, the agreement level was considered to be moderate, and if it fell within the range of 0.60 - 0.80, it was considered to indicate substantive agreement between the different subjective opinions. If the kappa was in the range 0.81 - 0.99, the two annotators were considered to have almost reached perfect agreement.

3.4. Multimodal Dataset

We recorded a total of 10 meetings, accumulating around 2000 minutes worth of video-audio and physiological data with a mean length of approximately 500 minutes for each student. There were 9507 video segments that needed to be annotated with a mean length of 10 secs for each clip. Table 1 shows the Cohen Kappa value for each affective state (which we treated as a binary labeling task), along with the number of video segments that received consistent judgment from the two annotators. 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 the concentration and frustration affective states. Looking at the details of the data, there were 1360 video segments that were annotated as showing concentration and 186 video segments that were annotated as showing frustration from both of the annotators. 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. In addition, there were 171 video segments and 55 video segments considered to indicate confusion and boredom by both annotators. Therefore,

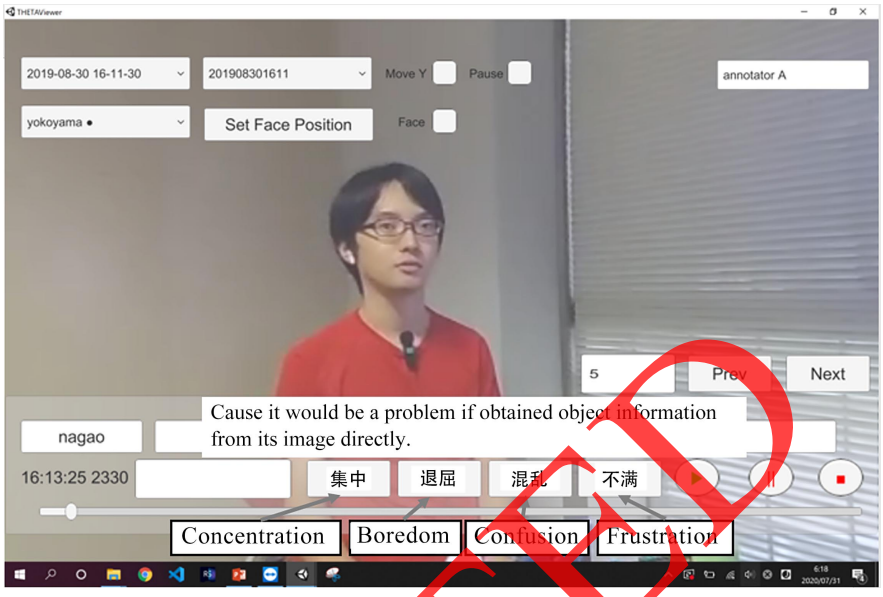

Figure 2. Video-based tool for annotating learning-centered affective states.

Table 1. Cohen Kappa value and consistent judgment numbers for each affect.

Affective states	Cohen Kappa	Consistent segment numbers
Concentration	0.64	1360
Confusion	0.44	171
Frustration	0.71	186
Boredom	0.50	55

the multimodal dataset we used in this study included 1772 labeled segments, along with the speaker's audio and heart rate data recorded synchronously.

4. Methodology for Recognizing Affective States

4.1. Multimodal Feature Sets for Characterizing Affect

4.1.1. Extracting Facial Features

As mentioned, we employed ARKit packages running on the iPhone to track students' faces. Depth sensor data was used to generate a single facial mesh over a user's face, and various types of information regarding the user's face was detected, including its position, orientation, and a series of blend shape coefficients to represent corresponding values of specific facial features recognized by ARKit. The blend shape coefficient was a floating point number indicating the current position of the respective feature relative to its neutral configuration, ranging from 0.0 (neutral) to 1.0 (maximum movement). Figure 3 shows a set of examples of the facial mesh that we adopted to measure the dynamic facial features in the eye and mouth areas.

We extracted a series of dynamic facial features describing movement patterns of the eye and mouth at an average frequency of 30 HZ. The first 300 frames (10 seconds) from each entire meeting video were used as a baseline in computing the features.

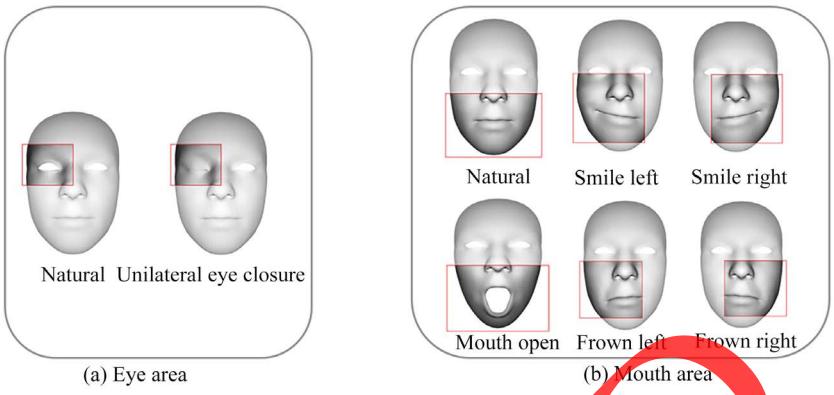

Figure 3. (a) Example of characterizing blinking of right eye by measuring closure of eyelid, in which natural expression had coefficient of 0.0, and maximum movement of right eye blinking had coefficient of 1.0. (b) Example of describing mouth and smile movements by measuring degree of openness of lips in vertical directions and positions of mouth corners in four quadrants, in which natural state of mouth had coefficient of 0.0, while maximum movement of lips and mouth corners had coefficient of 1.0.

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 affective states. We took the average of the eyelids' movement coefficient of both eyes when the Pearson r score was 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 denoised using a Savitzky-Golay filter with a window of 15 frames to remove artefacts introduced when the device occasionally lost track of faces, leading to incorrect measurement. We then applied peak detection (Du et al., 2006) methods to detect the local maximum (peak, eye-opening) and local minimum (valley, eye-shut). Eye blinks were detected by identifying a complete cycle from open (high coefficient) to close (low coefficient) and then back to open. We filtered out fake blinks by setting a threshold of 0.20 as the maximum valley coefficient and a minimum between-peak duration of 0.40 since an eye-blink cycle is around 0.40 to 0.60 s. We estimated the eye-blink frequency on the basis of the detected 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 affective processes manifested through prototypical patterns such as "smiling," which reflects a positive affect of feeling accomplished or happy or "frowning," suggestive of a negative affect such as confusion or frustration. We computed the sustained duration of "smiling" and "frowning" by measuring the position of

both mouth corners in 2D space.

In addition, in conversations, the visual features that characterize mouth actions during speaking could be signatures in inferring a student's affective states. Therefore, we measured the velocity and acceleration of mouth open-close movements along the vertical direction by computing lip movement coefficients. In **Figure 4**, we present four such patterns of mouth movements. As shown, subplots (a) and (b) are examples of a mouth that is open and closed. Subplot (c) is an example of a "frown," with the left and right corners of the mouth pulled downward, and subplot (d) is the pattern of "smiling," where the two corners of the mouth are pulled up

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.

4.1.2. Extracting Heart Rate Feature

We detected students' heart rate (HR) from the sensor on the Apple Watch, and the data was a univariate 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. 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. 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) (Lin et al., 2003; Lin et al., 2007), which was done in two steps. First, the piecewise aggregate approximation (PAA) (Matthews et al., 2002) algorithm was applied to the standardized raw sampled heart-rate time series $T = \{t_1, t_2 \ldots t_n\}$ 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 $\tilde{T} = \{\tilde{t}_1, \tilde{t}_2 \cdots \tilde{t}_w\}$, where the ith element was computed with the following Equation (1).

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

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 **Figure 5**, 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.

4.1.3. Extracting Audio Features

We employed openSMILE (Eyben, Wöllmer, & Schuller, 2010) 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 (Schuller, Steidl, & Batliner, 2009), 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 affective 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 functional (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)). Figure 6 shows an example of (a) a raw audio segment and example audio features, that is, (b) zero cross point, (c) F0, and (d) MFCC-12, we extracted using openSMILE.

4.2. Multimodal Analytics Framework for Tasks of Recognizing Affective States

In this section, we will demonstrate how we implement multi-modal analytics based on the multiple modality features we extracted in the previous section to generate a series of supervised learning models to predict a student's learning-centered affective states.

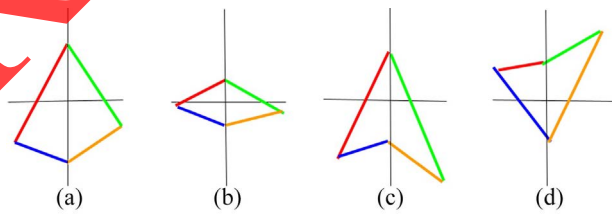

Figure 4. Example of mouth movement patterns defined with four key points of mouth: left and right corners of lips and middle points of upper and lower lips. (a) Mouth open, (b) mouth close, (c) frown, (d) smile.

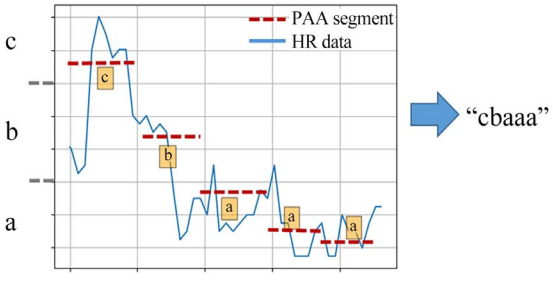

Figure 5. Example of sequences generated from HR time series using SAX representation.

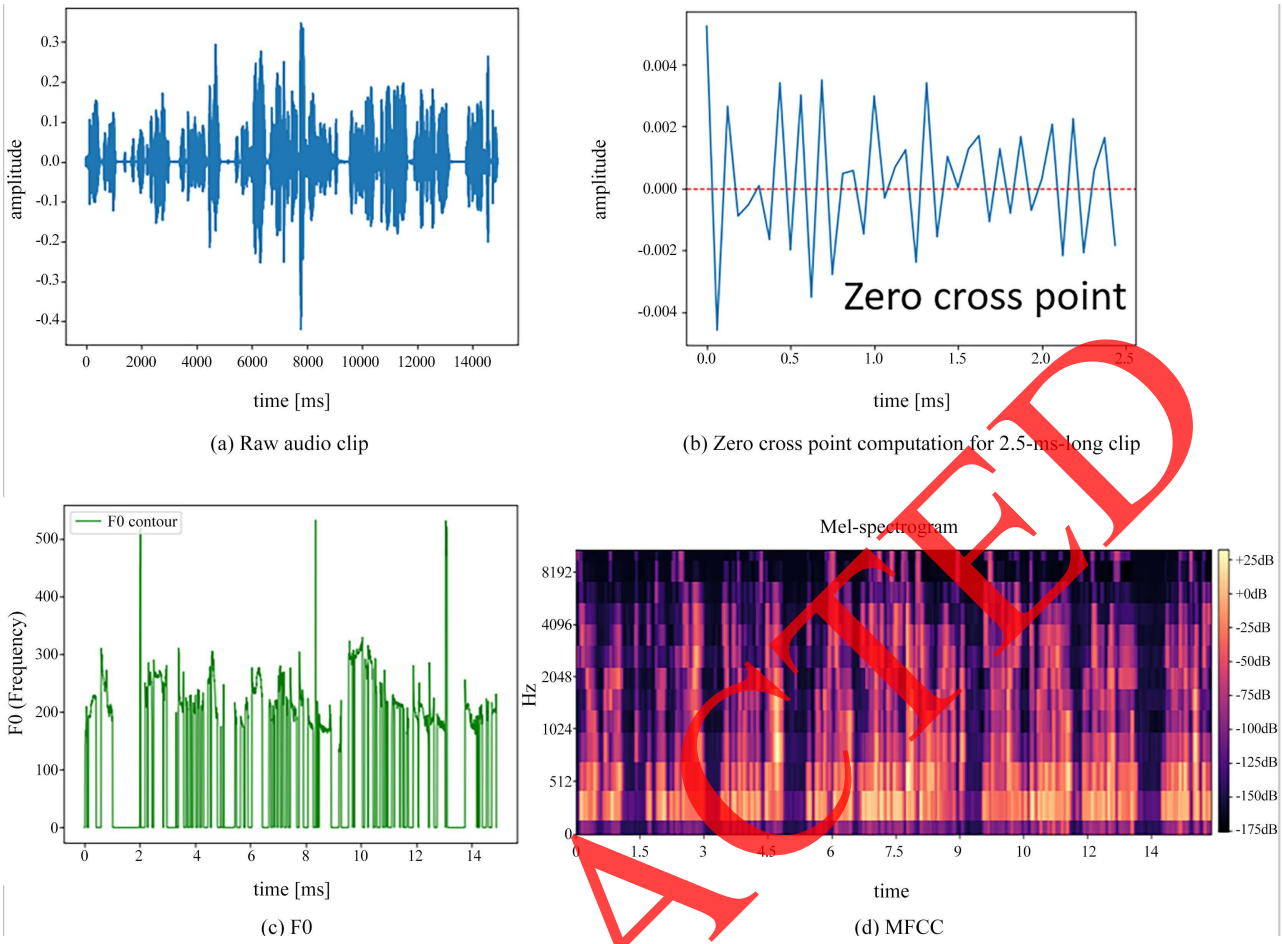

Figure 6. (a) is example raw audio segment that presents amplitude along time. (b) We cut 2.5-ms-long sub audio clip from raw audio segment as example to show how we compute zero cross point, which is amplitude value that crosses x-axis, and we then compute zero crossing rate for given time windows. (c) Example for F0 feature. (d) Example for MFCC-12 features.

We built a line of supervised learning models with three different multi-modal fusion methods as shown in Table 2. First, three baseline prediction models were separately built on the basis of individual channel: HR features, facial features, and audio features. 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 "Combination 3 classifier", along with three other classifiers based on two modalities each time (HR + Facial; HR + Audio; Facial + Audio). For feature-level fusion predictive models, considering the different numbers of each modality's features, we 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 model. Since we only had 10 features for the HR modality but 384 features for the audio modality, 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. Finally, we also built decision-fusion level classification models, in which we used three single-channel classifiers as base classifiers to make

Table 2. Supervised learning affective-state-prediction classifiers with different multimodal fusion approaches.

Multimodal fusion approaches	Classifiers	Modality
Single-channel	HR-based classifier	HR
	Facial-based classifier	Facial
	Audio-based classifier	Audio
Feature-level	Combination3 classifier	HR + Facial + Audio
	HR_ Facial-based classifier	HR + Facial
	HR_ Audio-based classifier	HR + Audio
	Facial_ Audio-based classifier	Facial + Audio
Decision-level	Decision-level Voting classifier	*

classifications on the same test instances separately. We then voted on the prediction results (the probability of belonging to each category), 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 "in-the-wild" environments, we could still can train the predictive learning models on the instances using a decision-level fusion approach even though some modalities are not available.

4.2.1. 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 (Kononenko, 1994; Urbanowicz et al., 2018) 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 is an extension of the RELIEF algorithm that can deal with multi-class problems and is more robust with incomplete and noisy data. It is similar to the RELIEF algorithm, which randomly selects one instance R but, in addition for k nearest instances from the same class called "nearest hits instances" and also searches for k nearest instances from each different class called "nearest misses." It then 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 of the top-ranked features from each modality and validated the predictive performance. Due to there being 10 HR features, 42 facial features, and 384 audio features, we separately tested 2 different proportions of HR with (0.50, 1), 3 different proportions of facial features with (0.30, 0.50, 0.70), as well as 4 different proportions of audio 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.

4.2.2. Classifiers and Validation

We built a set of multiclass classifiers based on three kinds of supervised-learning machine learning models including support vector machine (SVM), random forest (RF), and multilayer perceptron neural network (MLP). We then performed leave-one-student-out cross-validation to evaluate the prediction performance of each classifier. The Area under the ROC curve (AUC) scores were used as our primary evaluation metric, and we will report the AUC scores of each classifier.

5. Classification Results and Discussion

In this section, we report the mean AUC score of each classifier regarding the affective-state multiclass classification tasks built using different multimodal fusion approaches. We will also explain the proportions we used from the feature importance ranking list of each modality for making the feature subsets for the different classifiers. Furthermore, we will report the mean AUC for each affective-state classification task of the best fusion models.

Table 3 is a summary of the mean AUC for each classifier using different modality fusion approaches. As is shown, for single-channel classifiers, the audio channel-based classifiers achieved an overall better classification performance than the other single-channel based classifiers regarding the tasks of recognizing multiple learning-centered affective states. Among the audio channel-based classifiers, the RF classifier performance had a better prediction ability with a mean AUC of 0.76 than the SVM classifier with a mean AUC of 0.68 and was moderately stronger than the MLP classifier with a mean AUC score of 0.75; the same was also observed for single-channel based classifiers. However, for the facial single-channel based classifiers, the MLP classifier achieved the best mean AUC score with 0.73, which was higher than the other two classifiers, which may indicate that MLP can learn interpretable facial features better than other traditional supervised learning models. In addition, we got the best mean AUC scores for the facial single-channel-based classifiers when we chose the 15 top-ranked facial features, and when we chose the first 20 top-ranked audio features, we achieved a better prediction performance for the audio single-channel based classifiers; actually, for the HR single-channel-based classifiers, we got extremely similar classification results when we set the proportions of the HR features to 0.50 and 1.0. In consideration of balancing the numbers of features used from each modality, we finally chose to set the proportion to 1.0, which means using all of the HR features for building classifiers. For both the feature-level fusion and decision-level fusion approaches, we adopted the most predictive features from each channel to build classifiers.

For the feature-level fusion classifiers, first of all, the RF classifiers displayed an overall outstanding classification ability for all modality fusion methods. Second, we could order these four feature-fusion methods as Facial + Audio = Combination 3 > HR + Audio > HR + Facial. Combining the three modalities did not help with improving the prediction ability over only combining the facial

Table 3. Mean AUC scores of each classifier with different modality fusion approach.

Multimodal fusion channels	SVM	RF	MLP	Number of features	Best fusion approach
HR	0.68	**0.69**	0.68	10	
Facial	0.66	0.72	**0.73**	15	
Audio	0.68	**0.76**	0.75	20	✓
Combination 3	0.70	**0.76**	0.72	45	
HR + Audio	0.69	**0.75**	0.71	30	
HR + Facial	0.65	**0.74**	0.67	25	
Facial + Audio	0.69	**0.76**	0.75	35	✓
Decision-voting	0.68	0.74	**0.75**	*	

and audio channels, which indicates that the HR channel did not provide additional information for prediction tasks, but it did not introduce any other noise that decreased the prediction abilities as well. In addition, fusing the HR and facial channels yielded more accurate prediction results over fusing the facial single channel with AUC scores was increased by 0.02, which suggests that the HR channel can provide external information on visual facial cues, that is, physiological cues can reveal internal hidden information that can not been found through observing a human's external visual facial cues.

For the decision-level fusion classifiers, in which we used a voting method on the output of each single-channel classifier and made the final decision for the classification results, we got a mean AUC score of 0.75 for the MLP classifiers and 0.74 for the RF classifier, which guarantees that our models could still work well even though some modalities were not available.

Furthermore, we would like to give a comprehensive demonstration of the performance of the experiments on recognizing affective states. We chose the RF classifiers that demonstrated a better recognition performance with all three multimodal fusion approaches and report the mean AUC scores of the classifiers for both single-channel fusion and feature-level fusion approaches on classifying each affective state. Figure 7 presents the mean AUC of each affective state class of the single-channel fusion RF classifier. As is shown, the audio modality showed an overall better ability in recognizing all affective state classes than the other two modalities, reaching an AUC of 0.8 accuracy in identifying the concentration class. From an educational aspect, we hope that our classifier could identify negative states as much as possible, especially the ability to recognize the state of frustration in order to help teachers in direct-coaching intervention in real time. All of the single-channel level RF classifiers could precisely recognize the state of frustration with AUC scores of over 0.70, which indicates the effectiveness of the predictive abilities of the features we proposed from each modality.

Figure 8 shows the mean AUC scores of each affect class of the feature-level fusion RF classifiers. We can see that even though the Facial + Audio feature-level fusion RF classifier showed much better classification abilities than the

other fusion set, however, the Heart rate + Audio feature-level fusion RF classifier did a better job at identifying the states of concentration and frustration with increased AUC scores by 0.01 and 0.03 for each, while the Facial + Audio feature-level fusion classifier was better at recognizing the affective states of confusion and boredom. These remarkable results suggest that we can take good advantage of the identification abilities of a set fusing different modalities in recognizing different affective states. Furthermore, the feature-level fusion RF classifiers also showed more powerful identification abilities in recognizing the state of frustration than single-channel classifiers, which validated our proposal that multimodal analytics could increase the recognition performance of learning-centered affective state classification than single modal analytic approaches.

Finally, we looked into the classification performance for each affective state at the student level. All of the classification experiments were evaluated by performing level-one-student-out cross-validation in order to further examine the proposed method's ability to recognize the learning-centered affective states of each student. The facial-audio feature-level fusion RF classifiers were chosen in this part because they had the best classification abilities in the recognition tasks. Figure 9 presents the ROC and AUC scores of each affect class for each student. The proposed Facial + Audio feature-level fusion model showed an excellent affect recognition ability for each test student. In particular, for some test students (C and D), our proposed feature-level fusion classifier yielded AUC scores close to 0.80 in recognizing the state of frustration. Furthermore, for all test students,

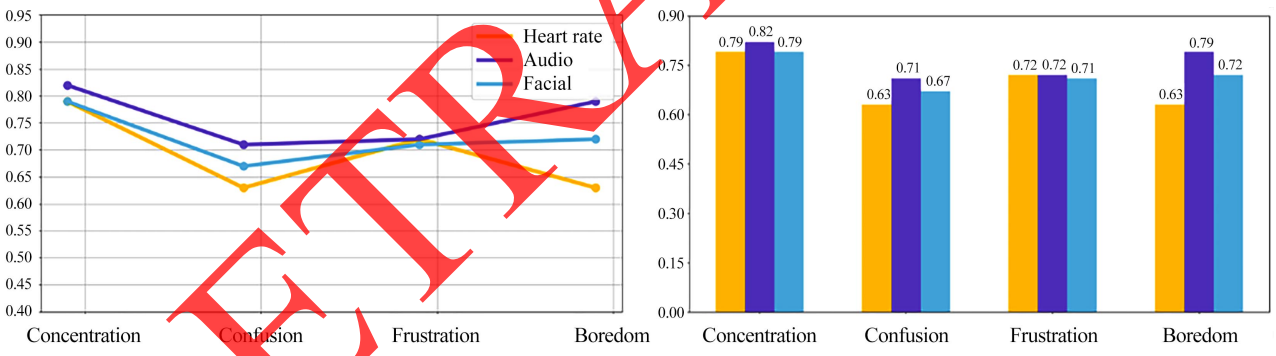

Figure 7. Mean AUC scores of each affect class of single-channel fusion RF classifiers.

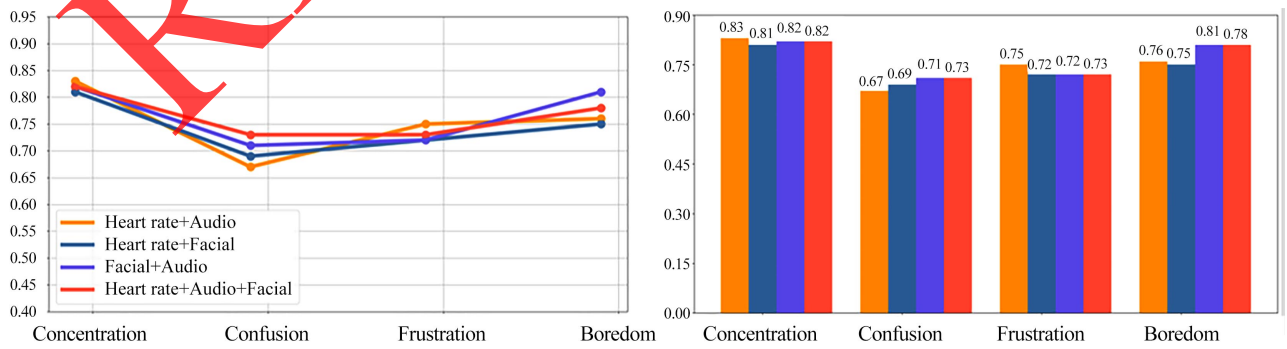

Figure 8. Mean AUC scores of each affect class of feature-level fusion RF classifier.

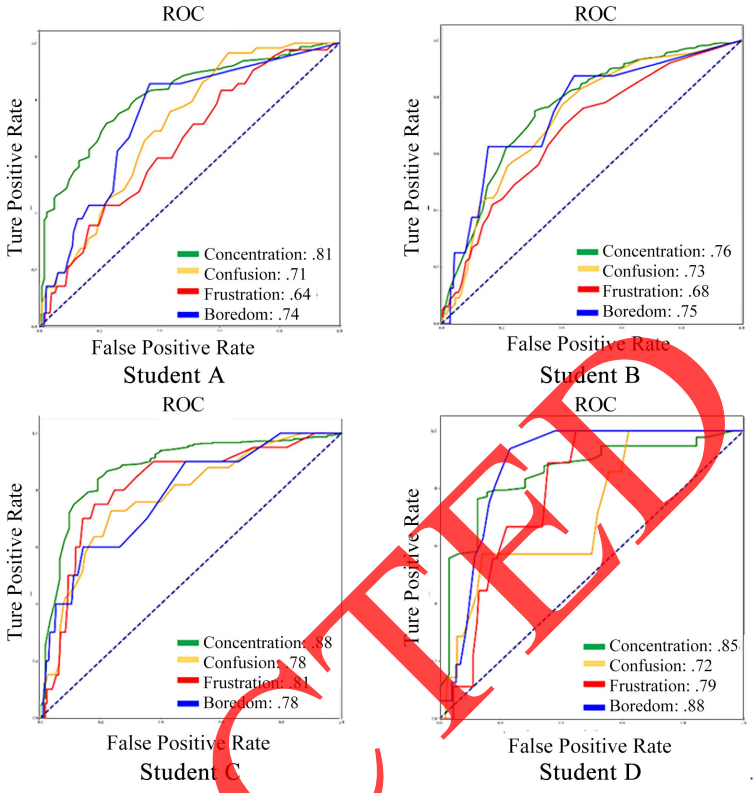

Figure 9. AUC scores of each affective state of facial + audio fusion RF classifiers at person level.

the classifier's recognition abilities in identifying the states of concentration, confusion, and boredom lead to AUC scores higher than 0.70. These exciting results have validated the outstanding predictive abilities of our proposed features and multimodal fusion approach on identifying students' learning-centered affective states "in-the-wild" and on conquering the influence of individual difference in the affective-state recognition problem.

6. Conclusion and Future Work

In this study, we attempted to use multimodal analytics to recognize students' learning-centered affective states including concentration, confusion, frustration, and boredom displayed during coach-led conversations with their professor. To achieve this goal, we first developed an advanced multi-sensor-based multimodal data collection system that can support the long-term recording of a set of multimodal conversational data including the speaker's audio-video information and physiological data (heart rate). A three-month-long multimodal "in-the-wild" conversational dataset was collected in a university research lab. This dataset recorded four students' video-audio and heart rate data as they had coach-led conversations with their professor. 500 minutes were accumulated for each student. We adopted a third party annotation approach in which we employed two independent annotators to annotate the video segments of conversations by using a video-audio-based annotation tool and created 1772 labeled

speech video segments with an average length of 10 sec.

We derived a series of interpretable proxy features from visual, audio, and physiological modalities separately to characterize the students' learning-centered affective states. For visual, we extracted lines of facial related features to describe dynamic patterns of eye blinking and mouth movements (speaking and smiling). For audio, we used the open SMILE tool to compute numbers of features for capturing students' affective states from acoustic cues. 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. 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 learning-centered affective states. We built three single-channel-level classifiers for individual modalities, facial, audio, and HR, separately. Then, four feature-level fusion classifiers were trained on combinations of three modalities: HR+ audio modalities, HR + facial modalities, and facial + audio modalities. Finally, a decision-level fusion classifier was generated by running a voting mechanism on the outputs of base single-channel level classifiers.

We performed leave-one-student-out cross-validation to evaluate the performance of these classifiers and reported the mean AUC scores for the aggregated-level, affective-state level, and person level. Evaluating the aggregated-level, feature-level fusion classifiers had an advantage over all single-channel classifiers, which indicates that fusing different modalities can provide addition information on individual modalities in order to improve the predictive abilities. Furthermore, the Facial + Audio feature-level fusion classifier yielded the best accuracy with an AUC of 0.76 in detecting learning-centered affective states compared with the other fusion models. The decision-level fusion approach also achieved AUC scores of 0.75 for the RF classifier and MLP classifier, which guarantees the practical ability of the proposed method, when faced with inevitable phenomena in "in-the-wild" data sets, such as some modalities not being available due to device problems. We also checked the recognition performance of the classifiers at the affective-state level. The Heart rate + Audio feature-level fusion RF classifier showed outstanding identification abilities in recognizing the states of concentration and frustration, while the Facial + Audio fusion classifiers were good at recognizing the states of concentration and boredom, which suggests that the flexible usage of different feature-level fusion sets can be beneficial in recognizing special affective states. A person-level evaluation was done last; our remarkable results provide evidence that our proposed multimodal analytics could overcome the influence of individual differences in students' affective states tasks.

For future work regarding our results, 1) we are looking forward to extending the size of our multimodal conversational dataset. In our previous work (Peng, Ohira, & Nagao, 2020b), in which we used small-scale facial and heart-rate modalities only for recognizing affective states, MLP classifiers performed far worse

than the other classifiers. However, in the present study, MLP started to show close to or even better identification abilities with feature-level fusion classifiers, which suggests a large-scale dataset can provide the opportunity of using more complex deep-learning models in the study of recognizing learning-centered affective states. 2) The professor's multimodal data were also collected, on which we are going to take a deep dive and analyze the potential utility of interaction behavior patterns in predicting students' learning-centered affective states. 3) In terms of application, we are going to launch our learning-centered affect states recognition model for use with our data collection system. We aim to alert teachers when students are facing an "assistance dilemma" shown through their confusion and frustration, in order to let teachers provide a timely and adaptive intervention to improve students' learning outcomes.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Bohus, D., & Horvitz, E. (2009). Models for Multiparty Engagement in Open-World Dialog. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 225-234). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1708376.1708409

Bosch, N., D'Mello, S. K., Baker, R. S., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2016). Detecting Student Emotions in Computer-Enabled Classrooms. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 4125-4129). Palo Alto, CA: AAAI Press.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 379-388). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2678025.2701397

Burt, K. B., & Obradović, J. (2013). The Construct of Psychophysiological Reactivity: Statistical and Psychometric Issues. *Developmental Review, 33,* 29-57. https://doi.org/10.1016/j.dr.2012.10.002

Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing, 1,* 18-37. https://doi.org/10.1109/T-AFFC.2010.1

Calvo, R., & D'Mello, S. K. (Eds.) (2011). *New Perspectives on Affect and Learning Technologies*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-9625-1

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In C. Peter, & R. Beale (Eds.), *Affect and Emotion in Human-Computer Interaction* (pp. 92-103). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-85099-1_8

Chen, L., Li, X., Xia, Z., Song, Z., Morency, L. P., & Dubrawski, A. (2016). Riding an Emotional Roller-Coaster: A Multimodal Study of Young Child's Math Problem Solving Activities. *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 29 June-2 July 2016, 38-45.

Cowley, B., Ravaja, N., & Heikura, T. (2013). Cardiovascular Physiology Predicts Learning Effects in a Serious Game Activity. *Computers & Education, 60,* 299-309. https://doi.org/10.1016/j.compedu.2012.07.014

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and Learning: An Exploratory Look into the Role of Affect in Learning with AutoTutor. *Journal of Educational Media, 29,* 241-250. https://doi.org/10.1080/1358165042000283101

D'Mello, S., & Graesser, A. (2012). Dynamics of Affective States during Complex Learning. *Learning and Instruction, 22,* 145-157. https://doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S., Craig, S., Fike, K., & Graesser, A. (2009). Responding to Learners' Cognitive-Affective States with Supportive and Shakeup Dialogues. In J. Jacko (Ed.), *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction* (pp. 595-604). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-02580-8_65

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2010). Attention Guidance in Learning from a Complex Animation: Seeing Is Understanding? *Learning and Instruction, 20,* 111-122. https://doi.org/10.1016/j.learninstruc.2009.02.010

Devillers, L., & Vidrascu, L. (2007). Real-Life Emotion Recognition in Speech. In C. Müller (Ed.), *Speaker Classification II* (pp. 34-42). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-74122-0_4

Du, P., Kibbe, W. A., & Lin, S. M. (2006). Improved Peak Detection in Mass Spectrum by Incorporating Continuous Wavelet Transform-Based Pattern Matching. *Bioinformatics, 22,* 2059-2065. https://doi.org/10.1093/bioinformatics/btl355

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1459-1462). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1873951.1874246

Forbes-Riley, K., & Litman, D. (2011). When Does Disengagement Correlate with Learning in Spoken Dialog Computer Tutoring? In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *International Conference on Artificial Intelligence in Education* (pp. 81-89). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-21869-9_13

Gomes, J., Yassine, M., Worsley, M., & Blikstein, P. (2013). Analysing Engineering Expertise of High School Students Using Eye Tracking and Multimodal Learning Analytics. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining.* International Educational Data Mining Society.

Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. (2007). Emotions and Learning with Auto Tutor. *Frontiers in Artificial Intelligence and Applications, 158,* 569.

Graesser, A., Ozuru, Y., & Sullins, J. (2010). What Is a Good Question? In M. McKeown, & G. Kucan (Eds.), *Bringing Reading Research to Life* (pp. 112-141). New York: Guilford.

Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. (2013). Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. *Educational Data Mining 2013.*

Hoque, M. E., McDuff, D. J., & Picard, R. W. (2012). Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. *IEEE Transactions on Affective Computing, 3,* 323-334. https://doi.org/10.1109/T-AFFC.2012.11

Hussain, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011). Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. In G. Biswas, S.

Bull, J. Kay, & A. Mitrovic (Eds.), *International Conference on Artificial Intelligence in Education* (pp. 131-138). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-21869-9_19

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic Prediction of Frustration. *International Journal of Human-Computer Studies, 65,* 724-736. https://doi.org/10.1016/j.ijhcs.2007.02.003

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of RELIEF. In F. Bergadano, & L. De Raedt (Eds.), *European Conference on Machine Learning* (pp. 171-182). Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-57868-4_57

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33,* 159-174. https://doi.org/10.2307/2529310

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (pp. 2-11). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/882082.882086

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: A Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery, 15,* 107-144. https://doi.org/10.1007/s10618-007-0064-z

Luft, C. D. B., Nolte, G., & Bhattacharya, J. (2013). High-Learners Present Larger mid-Frontal Theta Power and Connectivity in Response to Incorrect Performance Feedback. *Journal of Neuroscience, 33,* 2029-2038. https://doi.org/10.1523/JNEUROSCI.2565-12.2013

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., & Warm, J. S. (2002). Fundamental Dimensions of Subjective State in Performance Settings: Task Engagement, Distress, and Worry. *Emotion, 2,* 315-340. https://doi.org/10.1037/1528-3542.2.4.315

Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2016). Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing, 8,* 15-28. https://doi.org/10.1109/TAFFC.2016.2515084

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics, 1,* 107-128. https://doi.org/10.18608/jla.2014.11.6

Peng, S., Chen, L., Gao, C., & Tong, R. J. (2020a). Predicting Students' Attention Level with Interpretable Facial and Head Dynamic Features in an Online Tutoring System (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence, 34,* 13895-13896. https://doi.org/10.1609/aaai.v34i10.7220

Peng, S., Ohira, S., & Nagao, K. (2018). Automatic Evaluation of Students' Discussion Skill Based on Their Heart Rate. In B. McLaren, R. Reilly, S. Zvacek, & J. Uhomoibhi (Eds.), *International Conference on Computer Supported Education* (pp. 572-585). Berlin: Springer. https://doi.org/10.1007/978-3-030-21151-6_27

Peng, S., Ohira, S., & Nagao, K. (2019). Prediction of Students' Answer Relevance in Discussion Based on their Heart-Rate Data. *International Journal of Innovation and Research in Educational Sciences (IJIRES), 6,* 414-424.

Peng, S., Ohira, S., & Nagao, K. (2020b). Reading Students' Multiple Mental States in Conversation from Facial and Heart Rate Cues. *Proceedings of the 12th International Conference on Computer Supported Education, 1,* 68-76.

https://doi.org/10.5220/0009564000680076

Robison, J., McQuiggan, S., & Lester, J. (2009). Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems. In C. Muhl, D. Heylen, & A. Nijholt (Eds.), *Proceedings of International Conference on Affective Computing & Intelligent Interaction* (pp. 37-42). Los Alamitos, CA: IEEE Computer Society Press. https://doi.org/10.1109/ACII.2009.5349555

Rodrigo, M. M. T., & Baker, R. S. J. d. (2011a). Comparing the Incidence and Persistence of Learners' Affect during Interactions with Different Educational Software Packages. In R. Calvo, & S. D'Mello (Eds.), *New Perspective on Affect and Learning Technologies* (pp. 183-200). New York, NY: Springer. https://doi.org/10.1007/978-1-4419-9625-1_14

Rodrigo, M. M. T., & Baker, R. S. J. d. (2011b). Comparing Learners' Affect While Using an Intelligent Tutor and an Educational Game. *Research and Practice in Technology Enhanced Learning, 6,* 43-66.

Rodrigo, M. M. T., Baker, R. S., Agapito, J., Nabos, J., Repalam, M. C., Reyes, S. S., & San Pedro, M. O. C. (2012). The Effects of an Interactive Software Agent on Student Affective Dynamics While Using an Intelligent Tutoring System. *IEEE Transactions on Affective Computing, 3,* 224-236. https://doi.org/10.1109/T-AFFC.2011.41

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. *10th Annual Conference of the International Speech Communication Association*, Brighton UK, 6-10 September 2009, 312-315.

Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., & Bartlett, M. (2013). Multiple Kernel Learning for Emotion Recognition in the Wild. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (pp. 517-524). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2522848.2531741

Stevens, R. H., Galloway, T., & Berka, C. (2007). EEG-Related Changes in Cognitive Workload, Engagement and Distraction as Students Acquire Problem Solving Skills. *International Conference on User Modeling*, Springer, Berlin, Heidelberg, 187-196.

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2018). Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining. *Journal of Biomedical Informatics, 85,* 168-188. https://doi.org/10.1016/j.jbi.2018.07.015

Yoon, S., Byun, S., Dey, S., & Jung, K. (2019). Speech Emotion Recognition Using Multi-Hop Attention Mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2822-2826). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICASSP.2019.8683483

Zaletelj, J., & Košir, A. (2017). Predicting Students' Attention in the Classroom from Kinect Facial and Body Features. *EURASIP Journal on Image and Video Processing, 2017,* Article No. 80. https://doi.org/10.1186/s13640-017-0228-8