Reading the Room – Automated, Momentary Assessment of Student Engagement in the Classroom: Are We There Yet?

BETSY DISALVO, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
DHEERAJ BANDARU, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
QIAOSI WANG, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
HONG LI, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA
THOMAS PLÖTZ, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA

When in front of a classroom, a skilled teacher can *read the room*, identifying when students are engaged, frustrated, distracted, etc. In recent years we have seen significant changes in the traditional classroom, with virtual classes becoming a normal learning environment. Reasons for this change are the increased popularity of Massive Open Online Courses (MOOCs) and the disruptions imposed by the ongoing COVID-19 pandemic. However, it is difficult for teachers to *read the room* in these virtual classrooms, and researchers have begun to look at using sensors to provide feedback to help inform teaching practices. The study presented here sought to ground classroom sensor data in the form of electrodermal activities (EDA) captured using a wrist-worn sensing platform (Empatica E4), with observations about students' emotional engagement in the class. We collected a dataset from eleven students over eight lectures in college-level computer science classes. We trained human annotators who provided ground truth information about student engagement based on in-class observations. Inspired by related work in the field, we implemented an automated data analysis framework, which we used to explore momentary assessments of student engagement in classrooms. Our findings surprised us because we found *no significant correlation* between the sensor data and our trained observers' data. In this paper, we present our study and framework for automated engagement assessment, and report on our findings that indicate some of the challenges in deploying current technology for real-world, automated momentary assessment of student engagement in the classroom. We offer reflections on our findings and discuss ways forward toward an automated *reading the room* approach.

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; • **Applied computing** → **Education**.

Additional Key Words and Phrases: Engagements; Students; Wearable; Electrodermal Activity; Momentary Assessments;

ACM Reference Format:
Betsy DiSalvo, Dheeraj Bandaru, Qiaosi Wang, Hong Li, and Thomas Plötz. 2022. Reading the Room – Automated, Momentary Assessment of Student Engagement in the Classroom: Are We There Yet?. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 112 (September 2022), 26 pages. https://doi.org/10.1145/3550328

1 INTRODUCTION

> "As emotional practitioners, teachers can make classrooms exciting or dull and leaders can turn colleagues into risk-takers or cynics. Teaching, learning and leading may not be solely

Authors' addresses: Betsy DiSalvo, bdisalvo@cc.gatech.edu, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA; Dheeraj Bandaru, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA; Qiaosi Wang, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA; Hong Li, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA; Thomas Plötz, School of Interactive Computing, Georgia Institute of Technology, Atlanta, USA.

2474-9567/2022/9-ART112
https://doi.org/10.1145/3550328

> emotional practices, but they are always irretrievably emotional in character, in a good way or a bad way, by design or default." [39]

For human teachers, detecting students' emotions is an important part of mentoring, motivating, and managing the classroom [47, 52, 60]. A skilled teacher can read the affect levels in classrooms and student engagement with learning; interpreting their students' emotions, and adjusting their approach to keep students engaged and motivated. As the traditional collocated models of education have changed with the explosion of Massive Open Online Courses (MOOCs) and the need for worldwide remote learning because of the COVID-19 pandemic, we are seeing emerging research on replicating the emotional work of teachers in physical classrooms to virtual platforms [1, 3, 5, 45].

However, it is difficult to automate affect detection. Much of the previous research that has tried to automate affect detection in learning environments has taken place in controlled learning environments, such as Intelligent Tutoring Systems (ITS) or educational games, rather than classrooms [22, 23, 68]. In these settings, highly trained human observer data has been mapped to video data of facial expressions, hand gestures, body posture, and computer interactions. While there has been some work tying data from video media (expressions, gestures, and postures) to classroom experiences, namely the work of Ashwin and Guddetti [3–5], in general, the social nature of the classroom has been treated differently than these individual learning environments. This is likely the result of discomfort with the use of videos in classrooms and the privacy/security risk they pose [18] and difficulties with controlling video quality.

Recently, several studies have sought to use less intrusive wearable technology to measure learners' affect in real-world classroom settings [20, 30, 34, 45, 53]. While this research is encouraging, we are critical of some components: existing studies centered on understanding the overall affect levels of a whole lecture session, which does not correspond to teachers' needs during class. We aim to explore if an automated assessment of affect can realistically support teachers' classroom practices by measuring it in a moment-by-moment fashion. To that end, we employed human observations to validate sensor data.

Previous studies have used self-report to measure students' emotional engagement during class. Self-reporting emotions or engagement at the end of an activity or class poses a number of issues regarding validity [45, 49, 53]. First, the self-reported data is subjective to each subject's understanding of the emotional measures asked. Second, reporting at the end of a class or activity only captures a momentary appraisal of the learners' emotional state. Third, the act of asking a student about their emotional state distracts them from learning – and thus changes their engagement with learning. Fourth, students find the task disruptive and time-consuming, so they are less likely to report consistently.

Using human observations of student engagement, we sought to address some of the above validity issues and provide just-in-time data that helps teachers *read the room*. Our goal was to ground less-intrusive wearable sensor data with the rich observational data that has been used in ITS studies. Although many affect states can be observed in classrooms, we focused on correlating biometric sensor data with trained observers in classrooms who recorded if students' affect is on or off class topic. Our goal was to identify patterns in the sensor data that showed engagement with the learning environment. We hoped the observation data would support the use of wearable sensors in the classroom. However, we failed to find correlations between human observations of momentary student engagement and sensor data in this study. As such, this paper calls for intensified research to develop automated affect detection technology that is ready for deployments in classrooms (real or virtual).

The contributions of this paper can be summarized as follows:

- We designed and conducted a study in which we explored to what extent wearable EDA (electrodermal activity) sensors and data analysis techniques can be used for momentary assessments of student engagement in college classroom scenarios.

- We trained human annotators to provided ground truth for student engagement based on live observations in the classroom.We collected a dataset on eleven participants over eight lectures of college-level computer science classes.
- We implemented and explored a range of data analysis methods that have been introduced in previous related work. In contrast to what previous work suggests, we were *not* able to recognize changes in student engagement from EDA sensor data.
- We reflect upon the implications of our findings on the prospects of automated, momentary assessment of student engagement in the classroom and offer perspectives on what would be required to realize the automated equivalent to a skilled teacher *reading the room*.

2 BACKGROUND AND RELATED WORK

2.1 Student Engagement

In this paper, we use the term "student engagement" to express when observed student behavior indicated students were outwardly paying attention, and their affect indicated emotional and cognitive engagement with the class. Our efforts to detect student engagement was not a perfect measure of the emotional and cognitive state of students because it was based on behaviors we could observe. We used observable behavior as the measure to simulate what a teacher might observe in a classroom setting. Our use of this term builds upon prior work in learning sciences, education, and computing literature.

2.2 Defining Affect for Learning Environments

Much of the literature we reviewed for this study use the terms "affect" or "emotions." Both terms are often used to discuss students' feelings in learning contexts. For the purpose of this paper, we will refer to affect as the marker of experienced emotion. In other words, the goal is to understand emotion, and *affect* is the indicator of emotions. In 2012, Calvo and D'Mello predicted five directions for affect-aware learning technologies that would characterize the next several years of research and development, including an increased emphasis on interventions in real classrooms instead of controlled labs [13]. However, in the ensuing years, most such technology has focused on communicating affect to inform intelligent tutor systems [35, 62] for use in asynchronous or out-of-school learning environments [14, 25, 37]. More recently, researchers have explored potential applications to synchronous online classroom environments.

2.3 Defining Engagement for Learning

The construct of "engagement" in learning is difficult to define because it is contextual, multidimensional, and–from a learning perspective–dependent upon learning goals and objectives [2]. How engagement is defined impacts the choice of technology used to measure the construct and how ground truth is assessed [46]. In much of the literature that uses multi-modal approaches (specifically biometric, behavioral, emotional, or cognitive data) engagement is typically defined as relating to emotions and interests. As a result, researchers often measure positive or negative valence and high or low arousal as proxies for engagement. However, valence may not be as helpful in learning settings as less pleasant emotions, such as confusion or frustration–at least for a short duration–can be linked to positive learning outcomes [46].

Di Lascio et al. monitored students' in-real-time engagement during lectures [20]. They adopted Fredricks and McColskey's definition of engagement as a meta-construct consisting of behavioral, emotional, and cognitive engagement [26]. For the study, Di Lascoio et al. focused on emotional engagement, which they argued "is linked to students' affective state and is connected to emotional reactions to teachers" [20]. In our classroom observations, we made a similar distinction regarding the affect detection of student engagement. We sought to considered students' affective states and observed students' behaviors as *on-task* or *off-task*. For example,

we identified emotionally and cognitively on-task behavior when students nodded along with the professor, frowned, or laughed in agreement with statements from the professor or classmates. We considered a behavior off-task when a student looked at their phone or closed their eyes. The goal was to track short-term behavior and emotions, to replicate how a teacher might *read the room.*

2.4 Wrist-Worn Sensing

For the context of our study, we examine the accuracy of wearable sensor data–specifically, data from students wearing the Empatica E4 [29]–in measuring moment-by-moment emotional engagement with the class material. Various modalities have been used to detect affect in learning environments, including speech [16, 58], facial expressions [19, 24, 71], and physiological (wearable) sensor data [20, 21, 31, 40]. This paper is concerned with the accuracy of affect detection using wrist-worn sensors because they show potential for use in online classroom settings.

While speech and cameras for affect detection in learning settings have promising findings, they raise some privacy and security risks. Students are generally wary of the potential abuse of such data in education [73], and considerable attention has been paid to the ethical implications of instructors in an age of "mechanized learning" [74]. For instance, the use of headbands to monitor children's brainwaves in schools in China came under significant scrutiny in 2019 and was subsequently suspended. These controversies highlight the delicate nature of the use of affect-sensitive technology in learning environments. In order to justify these risks, the benefits must be significant.

Wearable technology, especially wrist-worn platforms, incorporates a range of physiological sensors, including sensors of movement (accelerometers, gyroscopes, magnetometers), heart rate, temperature, EDA or galvanic skin response, and many more. These sensors allow us to capture and model the physiological parameters of affect, cognitive load, and stress accurately, continuously, inexpensively, and with little intrusion [12, 20]. For example, Hassib et al. applied electroencephalography (EEG) for audience engagement during presentations [40]. Di Lascio et al. investigated the potential opportunities in monitoring students' engagement in class [20] and audience engagement in conference presentation scenarios [31] using wearable devices that collect EDA data. As sensor technology has become more seamless, comfortable, and ubiquitous, detecting affect based upon a wearable device like a watch has become a direction for affect detection in learning research.

2.5 Automated Assessments of Student Engagement in the Classroom through Body-Worn Sensors

Prior attempts at using machine learning and data analysis to trigger automatic detection of student engagement levels based on various sensor inputs have proven successful to varying degrees. However, so far, these prior studies have not sought to demonstrate the ability to successfully determine momentary cognitive engagement without multiple data inputs that can be intrusive to the environment and the study subject's overall learning environment.

One of the earliest methods to tackle this issue was template matching using EDA data to determine when a participant is startled. Identification of a startled reaction based on the calibration state heavily relied on the template matching methodology used during data analysis to identify startling episodes. This template matching technique is a variation of pattern matching algorithms that normalizes the raw data and then creates a threshold specific to each individual in the study [41].

Recent work has moved from using pattern matching to identify arousal in participants to using raw sensor data to identify general arousal, physiological synchrony, and momentary engagement to validate engagement. One of the first papers to use electrodermal sensors to identify momentary arousal based on surges in skin conductivity involved creating similar thresholds to those used by the template matching method [41]. However, these thresholds are based on statistical features of the data itself, such as standard deviation and peaks, rather than initial calibration engagement. The threshold created by Cain was based on the standard deviation of an

engineered feature named the "RCSC" which stands for Relative Change in Skin Conductivity[12]. In order to identify peaks as moments of engagement, we used a threshold based on the standard deviation of the sample and this threshold helped determine which peaks were of statistical relevance[12]. Another approach used to accurately identify moments of engagement within EDA data is through the creation of engagement levels within the data. Five engagement levels were created based on the statistical features of the data. Moments of engagement were identified in this method by using jumps from one level of engagement to a different level of engagement [20].

The most recent successful approach to tackling this problem of identifying engagement collected accelerometer, EDA, photoplethysmogram, video, and audio data and environmental data like noise and carbon dioxide data [28]. The EDA features used were created through EDA decomposition into tonic and phasic parts to better identify the physiological synchrony between the student and the teacher and the arousing and unarousing moments noted by previous studies. To identify the physiological synchrony between the two subjects, both Pearson Correlation Coefficient and Dynamic Time Wrapping Distance were utilized in two separate features. The prediction pipeline implemented by the study determined its ground truth using a survey based on the 5-point Likert scale to assign an engagement score and then adopted the LightGBM Regressor to predict the engagement score. They note that emotional engagement is the strength of these predictions. However, cognitive engagement was lower than the random baseline, and overall engagement was the most predicted of all single-dimensional engagements.

2.6 Student Engagement Measures

In this section, we consider student engagement measures used by researchers that might be used in validating student engagement with wearable sensor data. In Fredricks and McColskey's review of student engagement measures, they outline five methods [26]. First, and most commonly used, is student self-reporting. Second, and related to self-reporting, is experience sampling, where students are reminded (usually using technology) to report their current location, activities, and cognitive and affective states. Third is teachers' rating of students with a reflective checklist that can provide teachers with an aggregated class assessment of student engagement. Fourth is interviews conducted at the end of a class or course that often uses a stimulated recall process[61]. Fifth is observations developed to measure students' on- and off-task behavior as a gauge of student engagement.

Researchers have also used qualitative methods to establish a ground truth for their sensor data in educational settings. Similar to psychology and education research, the method most commonly used in wearable studies on student engagement is self-reporting [20]. Meta reviews of tutoring systems that use wearable sensors outline two methods for data training and validation of student engagement: self-report of emotional state and measurements under experimental methods that induce specific emotions [22, 23, 26].

There are exceptions. One is the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP). This is a time-sampling method for quantitative field observations of affect, on-task / off-task behavior in the classroom, educational gaming, and intelligent tutor contexts [56]. BROMP has been used as ground truth in validating sensor data from multiple sources such as motion sensor and video data, and EDA data with synthetic noise injection [68]. While some researchers have successfully found correlations between the BROMP labels and these types of sensors, others have found that they did not yield effective models of affect recognition with models performing "only marginally better than chance" [68]. Similar to BROMP, we conducted a time-sampling method for affect, valance, engagement, and disengagement in classroom content. However, we focused on physiological sensor data from wearables rather than the wide range of methods that could be used in a lab setting.

Our choice of observational methods was primarily motivated by our desire to understand if wearable sensors could identify the moment-by-moment changes of students' engagement during class time. The other methods used to measure student engagement; self reporting, experience sampling, teacher ratings and interviews, all provide reflective assessments that would not be useful to a teacher in the moment. While observation did

provide us with this real time data there are concerns that the subjective nature of observations involve inferring when individuals are engaged by their actions or discourse [70]. While the validity of observational data has been questioned because of students ability to hide their engagement levels during class [27, 61] other research indicates concurrent validity of observed and self-reported student engagement measures [72], and other have identified similar subjective issues with self-reported data [26, 70]. But as Sinatra suggests, the selection of student engagement measurement tools should align with the granularity of the data that researchers seek to address and the questions they seek to measure.

While numerous studies have explored affect detection related to learning„ these studies tend to take place in contexts related to ITS and learning games rather than in traditional classroom settings [22, 64, 75] (e.g., laboratories, online experimental settings, and school computer laboratories). As a result, these studies do not reflect the "blooming, buzzing confusion" of classroom settings [10]. Therefore it is unclear if the classroom context, in person or online, with all of the possible distractions and deviations from the learning purpose can be comparable. We are not suggesting that the affect detection is not accurate, rather that in a 60 – 90-minute online or physical classroom filled with students, technology breakdowns, and other distractions, the affect detected may be reporting emotions unrelated to learning.

3 CASE STUDY: DATA COLLECTION

Our work aims to extend previous work towards automated assessment of student affect and engagement by measuring moment-by-moment changes in affect and engagement – similar to how teachers *read the room*. Although self-reporting has been the prevalent method to identify affect in classroom settings for those seeking to validate wearable sensor data [20, 30, 34, 45, 49, 53], it does not reflect just-in-time data that would help teachers *read the room*. We decided, thus, to use third-party observations similar to the Baker-Rodrigo Observation Method Protocol (BROMP) [56]. This qualitative method can provide a detailed account of engagement levels changing across time within particular learning contexts [56].

We collected our data during a period of four weeks in 2019. In what follows, we provide details of the research context, participants, and data collected.

3.1 Research Context and Participants

The research team was made of experts in pattern recognition, applied machine learning, psychology, learning sciences, and qualitative methods. This expertise allowed us to apply a strong understanding of both quantitative and qualitative approaches to the study and background in the classroom context.

We recruited eleven students 18 - 30 years of age (four females and seven males) across two large, lecture-based computer science classes in a public U.S. institute (five students were at a Bachelor's level and six students were at a graduate level). Both courses had an average of 275 students. We chose these courses for three reasons. First, they were large enough to simulate many aspects of online learning environment, with teachers having difficulty tracking student engagement, which we deemed relevant given the goal of exploring the applicability of sensors within the complexity of online class scenarios. Second, the courses were popular and oversubscribed, meaning that not every student could be admitted and all the students who were admitted were eager to learn. Third, instructors were willing to have researchers observe classes, and the course schedule aligned with observers' schedules. However, we acknowledge that the there are many ways that a in person classroom with an observer present is not the same environment as an online course, which limits the transferability between context

A recruitment message was distributed through the class learning management system and announced at the beginning of several lectures. Once students reached out to the research team voluntarily, we introduced them to the study procedure and wearable devices, presented them with an informed consent form approved by our IRB,

Table 1. Distribution of observers to class sessions and participants across the study

Observer	Class	Week	Session	Participants
O1	B1	1, 2	1-4	P4,P5,P6
		3, 4	5-8	P7, P11
O2	B2	1, 2	1-4	P1,P2,P3
		3, 4	5-8	P8, P9, P10

gathered information about their demographic and their experience with lectures, using on-body sensors, and sharing physiological data.

We conducted observations with each participant during four lecture times for a total of 44 in-class observations. These observations lasted for the whole 90-minute class sessions and took place across two weeks. The Observers noted engagement data for each student. After practicing the observation in classrooms, we determined that given the attention demand on the observers' side, each observer captured data for a maximum of three students per class session. This is similar to other observation protocols [56]. Table 1 describes the distribution of observers, sessions, and participants across the study. While observations were taking place, each participant wore an Empatica E4 (wrist-worn) device that captured physiological signals.

3.2 Collected Data

3.2.1 Physiological Data. We used the Empatica E4 to collect students' physiological data. This device uses four sensors to measure blood volume pulse, acceleration, peripheral skin temperature, and electrodermal activity. Although the wearable nature of the E4 provides various advantages of continuously gathering sensor data, it can lead to unreliable data when used in conditions where free movements interfere with the sensors' functioning.

3.2.2 Observer-Annotated Data. Observers underwent extensive training to ensure that they could capture students' behaviors in relation to their class engagement and could attain an agreement on how to conduct observations. First, the research team adapted the work of [9] defining pleasure/activation states to a valence/arousal assessment 5-point Likert scale (see Table 2). Observers first used the instrument to assess the engagement of students recorded in Youtube class videos. Once they revised their notes and confirmed a mutual understanding of how to use the scale, they proceeded to calibrate their scoring process by attending six 90-minute class sessions where each annotated valence and arousal values of three random students in a round-robin fashion, moving to the next student after a 10-seconds span. After each session, we calculated observers' degree of agreement using different interrater annotation exchange methods (e.g., Conger Kappa, Krippendorff Alpha, Gwet's AC1). We then repeated the training until observers reached almost perfect agreement on recording participants' Valence values (0.8 Kappa coefficient) and moderate agreement on noting participants' arousal scales on average (0.76 Kappa coefficient) [44].

From there, observers were assigned session classes and participants to follow for collecting students' specific behaviors, valence and arousal, and on- or off-task status. Based on our experience during training sessions, we asked participants to sit in front of the classroom and located observers in the front corner of the classroom (see Figure 1). At the beginning of each session, the observer left the devices on the reserved seats to wait for students to arrive. Such an arrangement enabled observers to clearly notice students' facial expressions and behavior activities. Further, it minimized the students' self-conscious effect of knowing they were being observed. However, sitting in the front of the class and knowing the observers were present students may have altered their behavior. Observers collected data using a customized web application that offers predefined scales for valence and arousal and an expanding list of possible behaviors to note. There, they recorded students' data every time students' behaviors changed (e.g., bodily movements took place).

Table 2. An explanation for the valence and arousal scales we used for observation notes.

Scale	Valence (negative ->positive emotion)	Arousal (sleep ->wide awake)
1	Extremely unhappy, angry, extremely frustrated	Extremely sleepy, already fall asleep during class
2	Seems a little upset, unhappy, negative	Tired, low energy
3	Neutral, neither negative or positive emotion exhibited	Neutral, neither energetic nor sleepy
4	Happy, smiling, enjoying the lecture, etc.	Actively engaged, interact with the class and instructors. e.g., raise their hands during class, answer instructors' questions by themselves or in front of the whole class, nodding or shaking their heads, smiling, etc.
5	Extremely happy, e.g., instructor cancelled class/quiz/exam/assignment, students would be cheering loudly, very excited	Actively engaged, interact with the class and instructors. e.g., raise their hands during class, answer instructors' questions by themselves or in front of whole class, nodding or shaking their heads, smiling, etc.

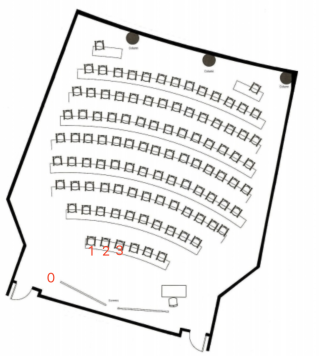


Fig. 1. Experimental setting for data collection in class B. The numbers 1 to 3 shown with red color mark the positions for the three participants in data collection. The position 0 in red color shows where researchers sat in the classroom to observe the participants.

3.3 Comparison to Previous Assessment Studies

Our work is motivated and inspired by previous work that demonstrated the effectiveness of GSR-based assessments of student engagement. As outlined in Sec. 2 the majority of previous studies operate in a similar yet not identical scenario. Our focus is on momentary assessments that would enable real-time interventions aiming at (re-)engaging students in case their attention slips. For the sake of transparency and better comparability, we

Table 3. Comparative summary of previous studies on GSR-based assessment of student engagement and ours.

Study	Ref.	Summary
Measuring Electrodermal Activity to Capture Engagement in an Afterschool Maker Program	[12]	Detects moments of engagement of a single session using the standard deviation of the given EDA data session
Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors	[20]	Classifies entire participant sessions as engaged or not-engaged based on various features including an innovative momentary engagement classification based on large deviations in the EDA data.
StartleCam: A Cybernetic Wearable Camera	[41]	Template-matching to identify moments of being startled
Our Paper	–	Attempts to predict moments of engagement based on Electrodermal Activity of indivdual students

contrast the most relevant previous studies that our work is based on to our study in Tab. 3. Note that more studies have been described in the literature, yet those listed in the table are representative of the state-of-the-art as most other systems and studies closely resemble those. Our study is the first that attempts to predict *moments* of student (dis-)engagement based on EDA data.

4 DATA ANALYSIS

Based on the data collected in our study (as described in the previous section) and inspired by previous studies as described in related work; we implemented an automated analysis system that aims at detecting correlations between ground truth observations of student engagement and sensor data collected through the wrist-worn Empatica E4 platform. For the sake of our exploration, we implemented and validated a range of analysis methods, which helped us understand the data analysis problem – and why it is so challenging.

Figure 2 provides an overview of our procedure. The automated analysis of the collected sensor data with regards to the student engagement observations requires preprocessing of both data sources. The observers' annotations of student engagement needed to be converted into a ground truth to explore possible correlations with the sensor-collected data (top-left in the figure; Sec. 4.1). The recorded sensor data also need to undergo a dedicated preprocessing step that targets data cleaning and normalization, for which we followed the state-of-the-art as described in related literature and summarized in Section 2. The former is necessary because electro-dermal activity data and physiological signals are often noisy for they tend to be easily affected by, for example, movements (top-right of Fig. 2; Sec. 4.2). Furthermore, we had to trim and synchronize the collected data to accommodate for the practicalities of our study (top-center part of Fig. 2; Sec. 4.3). In a deployment all input data are preprocessed in this manner. Subsequently, a sliding window procedure extracts analysis windows that comprise sequences of consecutive sensor readings – ten seconds in our case, motivated by the characteristics of EDA data that are known to change only relatively slowly over time [20]. Feature vectors are then extracted for individual frames and forwarded to a subsequent classification backend [11]. Alternatively, heuristics-based analysis on the raw, preprocessed sensor data are pursued allowing us to inspect the EDA data more closely to understand the challenges of automated, momentary engagement assessment.

The actual data analysis approach explores a range of techniques previously used in related work on engagement prediction using body-worn sensing platforms. In particular, we evaluate using classification models [15, 42, 43], and–for the more detailed exploration of EDA data–slope-based predictive [12], and discrete level jump predictive models [20]. Figure 2 illustrates each of these methods that all result in predictions of student (dis-)engagement

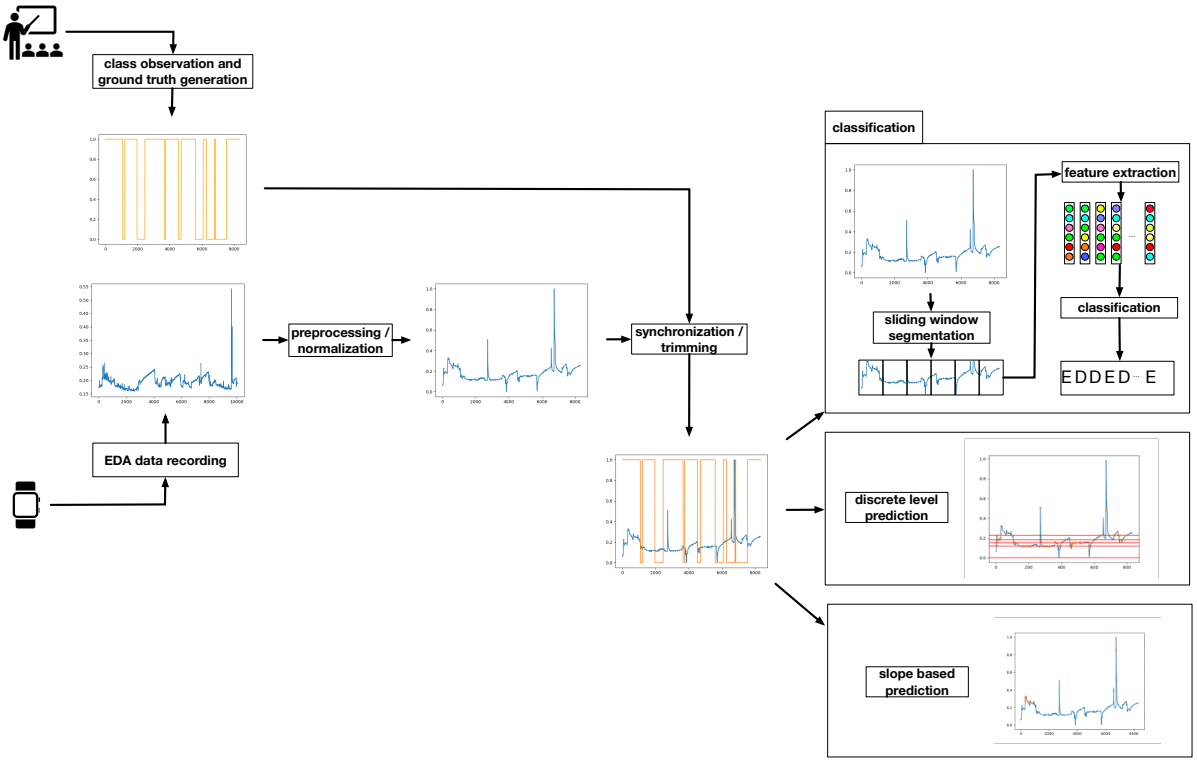


Fig. 2. Data analysis workflow from sensor input to result analysis.

that are superimposed to the ground truth on the observation/ recording timelines for individual sessions. In what follows, we provide detailed explanations of the components of our data analysis approach. Sec. 4.5 will also provide details of the model training and evaluation protocol that we employed for our study.

4.1 Converting Observation Data into Engagement Ground Truth

We needed to convert the observation data collected into ground truth that could determine when students were engaged or not. To work toward our goal of finding possible correlations between sensor readings and observation data, we explored the needed conversion from observation to ground truth from different approaches. Each provided a new perspective for future correlation explorations. A first approach consisted of mapping the valence and arousal scales to engagement scales. Levels 1-2 would reflect low engagement, level 3 a medium engagement, and levels 4-5, high engagement. Another approach used the behavior data noted by our observers. We cleaned overlapping annotations and typos in the nearly 40 specific behaviors noted by our observers into a codebook of behaviors. We then grouped these behaviors using two classifications that could help us describe engagement in different ways: *i)* engaged vs. non-engaged; and *ii)* engaged with/ without physical behavior vs. non-engaged with/ without physical behavior. Table 4 provides details for those different classifications of behavior data.

For the explorations reported in this paper, we chose binary engagement annotations that are derived from mapping the specific behaviors listed in Table 4 to engaged and not engaged. Our trained expert annotators

Table 4. Different classifications for students' behaviors that served as ground truth for future correlation explorations: engaged vs not engaged and engaged with/without physical activities vs not engaged with/without physical activities.

Engagement	**Physicality**	**Specific behaviors**
Engaged	With physical activities	'Asking instructor question','Class-related discussion', 'Laptop','Laughing','Laughing listening','Nodding', 'Shaking head', 'Nodding and shaking heads', 'Smiling','Touching face','Writing notes','Laughing taking notes', 'Answer questions','Respond to teacher','Writing and yawning', 'Raising hand'
	Without physical activities	'Listening/looking at presentation/instructors', 'Reading notes'
Not Engaged	With physical activities	'Checking phone','Taking picture of the slide', 'Drinking water','Eat mints','Falling asleep', 'Yawning','Looking around','Non-class-related chatting', 'Talking','Stretching','Sketching','Coughing','Tie hair', 'Playing with hair','Frowning','Clean nose','Making faces'
	Without physical activities	'Zoning out','Sleep','Closing eyes' 'Head on hands"

observed student behaviors according to the behavior specifications given in the right part of Table 4. Even though our eventual goal is to automatically assess student engagement at the level of granularity of the specific behaviors as listed in the table, for this study we concentrate on the first, most fundamental assessment – are students engaged or not, and when does their engagement change. Accordingly, we converted detailed behavior observations into a binary ground truth as per the operational definitions given in Table 4.

4.2 Sensor Data Preprocessing

The Empatica E4 sensor readings cover a range of signals, including galvanic skin response (GSR, also referred to as electrodermal activities, EDA), heart rate and heart rate variability (inter heart-)beat interval, IBI), skin temperature, and accelerometry. Motivated by related work on affective computing [65] and practical considerations such as ease of use and objectivity of sensor readings, we utilize EDA as the main data source for our study.

Sensor data preprocessing aims at cleaning up the recorded data with regards to noise or any other detrimental artifacts that the sensors may pick up [32] in a recording session in the classroom. We first identify excessive arm motions by analyzing signal energy from the tri-axial accelerometer stream and eliminate portions that breach a filtering threshold, i.e., we eliminate portions of the sensor data from further processing for which the measured acceleration magnitude is more than two standard deviations away from the session mean. It is well known that clean EDA readings require tight skin contact with the sensor and that motion artifacts are often detrimental to the quality of the sensor data [20, 43, 67]. Consequently, our thresholding approach effectively eliminates such motion artifacts. Furthermore, the EDA signals are smoothed through a kernel-based median filter with the kernel size set to E4 sampling rate plus one the filter length itself set to five samples [7]. This filter helps further eliminating signal artifacts that are not of relevance for our analysis but preserves the typical shape of EDA signals.

Considering the value of EDA signals recorded from different students may differ significantly [51], the signal is then range-normalized using min-max normalization, which transforms the minimum value of the signal to 0 and maximum value of the signal to 1 and transforms all the values in between accordingly.

The GSR data can contain important information associated with changes in emotional engagement. To extract such features, we decomposed the signal into *tonic* and *phasic* components [8, 20, 43, 67]. The background *tonic*

component reflects the skin conductance level (SCL), which indicate general changes in autonomic arousal. The rapid *phasic* components describes the skin conductance response (SCR), which suggests changes associated with a stimulus. In this work, we applied one of the most widely used cvxEDA method by Grecoet *et al.* [36], which uses convex optimization to decompose the signal. Inspired by the feature extraction in [20], we calculated the general arousal features from the original GSR signal, SCL, and SCR data.

4.3 Data Trimming and Synchronizing for Data Collection

In order to set up our data collection study with minimal interruptions to the normal class routing, all wrist-worn sensing platforms (Empatica E4) were configured such that they started recording before the students arrive in class, and continue to do so until after the students had left. We handed over the devices before students came to class and collected them after the class had finished.

This protocol, while practical, results in portions of the data not being usable because they are covering time periods that are not related to the study. Furthermore, we recognize that students' level of excitement is often high when first entering the class and that they often start moving more towards the end than during the class. Consequently, these portions of the data may also not be of highest relevance for our study. As such, we removed the first and the last five minutes from our analysis.

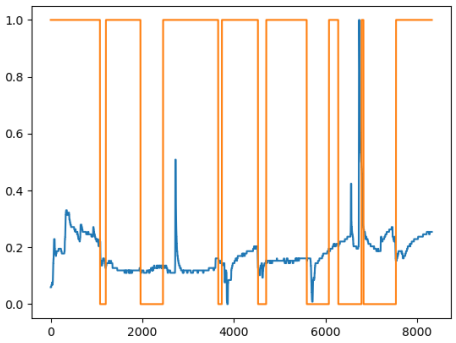
We synchronized the start-times for both sensor and observation data to the first annotation time that observers recorded as valid.

4.4 Predicting Student Engagement from Sensor Data

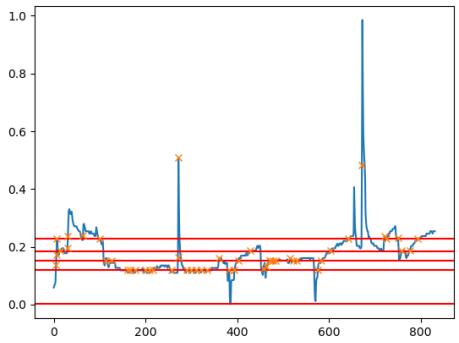
The overarching goal of our study was to explore to what extent EDA sensor data collected through a wrist-worn device can be used to automatically recognize changes in student engagement. We particularly sought to automatically *read the room* such that teachers and human-computer interfaces in online classrooms could adapt to student engagement in the moment. This endeavor is appealing, especially for online learning where students have poor video quality, will not turn on their cameras, or the video conferencing system will not allow teachers to see all of the students.

After preprocessing, trimming, and synchronization, as described in the previous sections, the recorded data is now fed into our analysis back-end that aims at replicating the ground truth observations of student engagement with technical means. We explore the utility of three assessment methods that have been described in related work: *i)* Classification Methods; *ii)* Slope-based Predictive Models [12]; and *iii)* Discrete Level Jump Predictive Model [20]. Figure 3 illustrates the three different methods using an exemplary recording session. Figure 3a shows processed EDA data (blue) for an exemplary recording session, along with processed, binary ground truth annotation regarding engaged vs not engaged overlaid in orange.

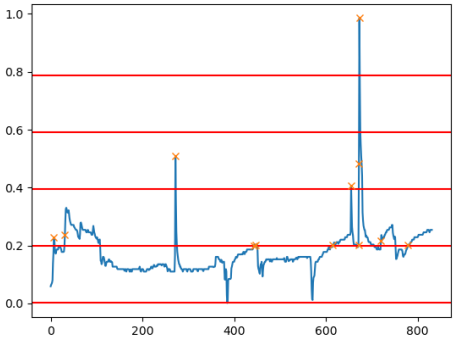
4.4.1 Classification Models. Based on the preprocessed EDA data, which are subsequently segmented into ten-second, non-overlapping analysis windows, can extract a feature representation that converts our sensor data (from individual analysis frames) into a format that can be used for analysis using machine learning based classification backends. Our features are adopted from previous work in the field that validated signs of general arousal [15, 42, 43]. This feature set is computed on a ten-second moving window based on the EDA signal, the SCL data, and the SCR data. The SCL and SCR data are the decomposition of the EDA signal into the tonic and phasic components of the EDA signal respectively, (as described in Sec. 4.2). The features extracted for each analysis window are as follows: *i)* mean and standard deviation of EDA signal; *ii)* mean and standard deviation of SCL signal; *iii)* mean and standard deviation of SCR signal; *iv)* number of peaks in EDA signal; *v)* number of peaks in SCL signal; *vi)* number of peaks in SCR signal; *vii)* average peak amplitude of EDA signal; *ix)* average peak amplitude of SCL signal; *x)* average peak amplitude of SCR signal; *xi)* area under the curve for EDA signal; *xii)* area under the curve for SCL signal; and *xiii)* area under the curve for SCR signal. Utilizing this feature set



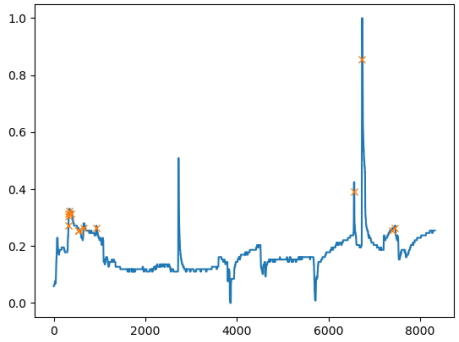
(a) Processed EDA data for an exemplary session (blue) with binary ground truth annotation overlaid (orange).



(b) Quintiles defined by equal numbers of observations(red lines) defined levels of engagement.



(c) Quintiles defined by equal length (red lines) defined levels of engagement.



(d) Slope-based EDA analysis of student engagement.

Fig. 3. Illustration of data analysis methods explored for predicting student engagement from sensor data.

as the input for the classification models, we evaluate the four main categories of statistical classifiers, namely: *i)* instance based learning – *k* Nearest Neighbors; *ii)* descriptive modeling – Naive Bayes; *iii)* discriminative learning – Decision tree and its ensemble variant, i.e., Random Forests; and *iv)* kernel based learning – Support Vector Machines (SVM). Note that the limited size of our dataset renders the use of contemporary modeling techniques such as deep neural networks infeasible for our exploration study. The binary classification models take feature representations of (portions of) sensor data as input and classify these portions as either engaged or not engaged. Each ten-second window of a recording session, which typically runs for about 65 minutes, i.e., the standard lecture duration, is associated with binary ground truth annotation regarding student engagement as defined in Sec. 4.1.

4.4.2 Slope-based Predictive Model. Ideally, we would develop a fully automated system for which we have implemented the data analysis pipeline as described above. In order to better understand the challenges of the analysis task, we have also implemented two alternatives to the ML-based classification backends, in which we

utilize the raw, yet preprocessed, sensor data for direct, heuristics-based analysis. Again, our analysis methods are inspired by related work in the field that suggests that our overarching goal of automatically "reading the room" may be achievable.

For the first variant, EDA data is analyzed based on the slope of the signal where positive slopes indicate arousal and negative slopes indicate unarousal. To ensure that not all positive slope segments are defined as arousal a threshold is set for each session that is analyzed. Any data point that is more than one standard deviation away from the mean in either direction is considered significant and thus indicating a relevant slope (positive, or negative). All positive slopes, as defined above, are then considered moments of arousal and all other data points are considered unaroused moments, which is in line with previous work in the field [12].

4.4.3 Discrete Level Jump Predictive Model. The second variant of non-classifier based data analysis is based on the definition of levels of engagement as proposed by Di Lascio [20] The preprocessed EDA signal is used to determine a total of five discrete levels of engagement. These five levels are used by the predictive model to find moments at which the signal changes level and identifies these moments as moments of arousal or unarousal determined by whether the signal goes from a lower level to a higher level or vice versa respectively.

Using piecewise aggregate approximation [50], the dimensionality of the EDA data is reduced before further analysis. Piecewise aggregate approximation takes an input signal and takes a specified segment window of the signal and calculates the mean of that segment. This mean value is then used as the new value for that portion of the signal, which results in effective dimensionality reduction. In our case, we reduce the dimensionality of the signal by a factor of ten. Once the EDA data from a session is converted, we identify five discrete levels for each session as per related work [20]: *1)* very low; *2)* low; *3)* normal *4)* high; and *5)* very high. Using these levels we predict momentary cognitive engagement as moments when the participant moves from one engagement level into a higher one and moments of disengagement as moments where participants move from a one engagement level into a lower one. In order to determine the five discrete levels for each of the sessions we explore two methods: *i)* levels based on number of observations; and *ii)* levels based on signal value. Based on these prediction we can then perform the comparison to the binary ground truth annotation (Sec. 4.1) in our case study. Figure 3b and 3c illustrates the use of our discrete level jump predictive model for engagement analysis from EDA data.

Ground truth annotation for the non-classifier based approaches is provided based on engagement *changes*, i.e., moments (in form of the previously mentioned ten second signal windows) where changed from engaged to disengaged, or the other way around, are observed (again, as per the operational definitions for behavior observations and the procedure for obtaining the ground truth annotation as described in Sec. 4.1).

4.5 Evaluation Protocol

The ultimate application case for our work is to provide students with commodity devices such as the Empatica E4 for recording EDA data, and to conduct real-time sensor data analysis that directly informs either a human teacher or feeds into a human-computer interface with the goal of adapting content delivery "on-the-fly" and as required by individual learners. This is how teachers in classical instruction scenarios *read the room* and respond to momentary demands. As such, the obvious evaluation protocol for an automated analysis procedure would be based on a student-independent model that is deployed "as is". Such a model would be derived from large amount of annotated sample data, collected from many different students.

Consequently, our initial evaluations of the machine-learning based recognition models are based on a *leave-one-participant-out (LOPO)* protocol, in which we hold out the data for one participant to test on, and use the data from the remaining ten participants in our dataset for model training. We report results for all participants, i.e., each participant's data is held out once, along with total averages.

Arguably, the LOPO evaluation protocol is the most challenging one because it requires a recognition model to generalize to unseen participants [66]. In addition to the non-personalized LOPO evaluation protocol, we also

explore to what extent some amount of personalization affects the machine-learning based models. To do so, we used a *leave-one-session-out (LOSO)* protocol in which we tested the models on data from individual class sessions for a participant while we trained the models with the data from all other participants (all sessions) plus all sessions of the target participant minus the chosen test session. We iterate through all possibilities to leave out individual sessions and average the results on a per participant basis.

Our dataset is imbalanced with regard to the distribution of engaged vs. non-engaged portions (see Tab. 5 for details). It is well known that such class imbalance may have a detrimental impact on training effective machine learning based recognition systems [33, 66]. To counter possible negative effects, we conducted additional experiments in which we applied a random undersampling technique to counter the class imbalance [33] (training data only). Results for the machine-learning based methods are reported separately for both LOPO and LOSO evaluations – and contrasted to the results from using classifiers that were trained directly on the imbalanced datasets.

The two non-machine leaning based methods do not require a training step because the slope and jump estimations are conducted based on a global heuristic (as informed by previous work in the field [12, 20]). As such, evaluation results are reported for entire sessions where the objective is to correctly identify those ten second windows in which student engagement changes as per the ground truth definition.

We evaluated the prediction results through both accuracy and F1 scores. The former is used to be consistent with previous work in the field. However, as mentioned above our dataset exhibits imbalanced class distribution for which F1 scores better reflect the true recognition capabilities. The presentation of our results also includes breakdowns of the class distributions, which provides more insights into the underlying data and helps us understand the challenges of the analysis task.

5 RESULTS

Sec. 4.5 described the data collection and annotation protocol. Our analysis study is based on observations of eight class sessions for eleven students that highly trained human observers annotated. Ground truth annotations were generated as described in Sec. 4.1, and the sensor data was analyzed through the methods described in the previous section. For classifier training, optimization, and validation we employed the protocol described in Sec. 4.5. For contextualization we also report classification results achieved using two baseline classifiers: *i)* a "Random" classifier that assigns 'engaged' or 'disengaged' according to a uniform random distribution; and *ii)* a "Biased Random" classifier, which assigns the majority class (from training) to all test windows.

5.1 Leave-One-Participant-Out (LOPO) Evaluation of ML-Based Analysis

Tables 6 and 7 show the results for the LOPO-based evaluation of our ML-based classifiers without and with class balancing during model training, respectively. We report results for each of the five explored analysis models in terms of accuracy and (macro) F1 score, along with averages over all participants and standard deviations.

It can be seen that, on average, none of the classification approaches lead to satisfying analysis results. Across the sessions, we see F1 scores of less than 45%, which indicates that–much to our surprise and in contrast to what related work suggests–the automated, momentary assessment of student engagement through wrist-worn EDA sensors need further evaluation. To contextualize this finding, we emphasize again the rigor that went into our data collection and especially the annotation procedure. We employed state-of-the-art processing pipelines (as suggested by related work) that are deemed suitable for the required data analysis. It is widely known that the activity recognition chain (ARC [11]), a variant of which our approach essentially resembles, represents the de-facto standard for such analysis scenarios. More sophisticated, recent analysis techniques such as Deep Learning based methods [38, 57] are not directly applicable in our envisioned scenario due to training data limitations and the highly idiosyncratic behaviors and EDA responses of individual students.

Table 5. Class distribution (engaged / non-engaged) in our dataset (listed per session and averaged per participant).

Participant / Session	#Engaged Windows	#Dis-Engaged Windows
P1-1	158	87
P1-2	236	132
P1-3	238	50
P1-4	194	70
P1 total	826	339
P2-1	190	61
P2-2	316	44
P2-3	264	109
P2-4	259	49
P2 total	1029	263
P4-1	143	64
P4-2	158	19
P4-3	205	9
P4-4	208	15
P4 total	714	107
P5-1	163	46
P5-2	156	21
P5-3	174	23
P5-4	184	36
P5 total	677	126
P7-1	200	27
P7-2	191	32
P7-3	163	38
P7-4	198	11
P7 total	752	108
P8-1	207	136
P8-2	257	92
P8-3	228	98
P8 total	692	326
P9-1	309	30
P9-2	300	52
P9-3	275	65
P9 total	884	147
P11-1	200	24
P11-2	189	36
P11-3	158	45
P11-4	173	36
P11 total	720	141
Avg	209.8	51.9
Std	48.35	34.02

Even the artificial balancing of the training data does not lead to noteworthy improvements: The results in Tab. 7 are more or less the same as when not balanced (Tab. 6).

Table 6. Results for discrimination between engaged and non-engaged episodes in our student dataset for machine-learning based classifiers trained *without* class balancing and following a leave-one-participant-out (LOPO) evaluation protocol.

Test	**Bayes**		**K-NN**		**DT**		**RF**		**SVM**		**Random**		**Biased Random**	
Participant	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
P1	0.32	0.3	0.67	0.46	0.64	0.51	0.68	0.44	0.71	0.41	0.49	0.47	0.71	0.41
P2	0.8	0.45	0.73	0.47	0.68	0.48	0.74	0.46	0.8	0.44	0.48	0.44	0.8	0.44
P4	0.76	0.46	0.79	0.51	0.71	0.5	0.78	0.48	0.87	0.47	0.53	0.44	0.87	0.47
P5	0.82	0.49	0.78	0.52	0.67	0.48	0.71	0.46	0.84	0.46	0.5	0.43	0.84	0.46
P7	0.87	0.46	0.8	0.53	0.7	0.47	0.83	0.5	0.88	0.47	0.52	0.44	0.88	0.47
P8	0.68	0.42	0.67	0.46	0.63	0.48	0.67	0.44	0.68	0.4	0.47	0.46	0.68	0.4
P9	0.74	0.5	0.75	0.49	0.69	0.5	0.77	0.5	0.86	0.46	0.5	0.43	0.86	0.46
P11	0.84	0.46	0.78	0.5	0.7	0.49	0.79	0.47	0.84	0.46	0.48	0.43	0.84	0.46
Avg	0.73	**0.44**	0.75	**0.49**	0.68	**0.49**	0.75	**0.47**	0.81	**0.45**	0.50	**0.44**	0.81	**0.45**
Std	0.18	0.06	0.05	0.03	0.03	0.01	0.06	0.02	0.08	0.03	0.02	0.01	0.07	0.03

Table 7. Results for discrimination between engaged and non-engaged episodes in our student dataset for machine-learning based classifiers trained *with* class balancing and following a leave-one-participant-out (LOPO) evaluation protocol.

Test	**Bayes**		**K-NN**		**DT**		**RF**		**SVM**		**Random**		**Biased Random**	
Participant	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
P1	0.29	0.25	0.48	0.46	0.5	0.47	0.49	0.46	0.39	0.39	0.51	0.49	0.29	0.23
P2	0.25	0.24	0.51	0.45	0.49	0.43	0.51	0.46	0.5	0.47	0.49	0.45	0.2	0.17
P4	0.19	0.19	0.51	0.42	0.49	0.4	0.48	0.41	0.79	0.58	0.49	0.42	0.13	0.11
P5	0.21	0.2	0.53	0.46	0.5	0.43	0.41	0.37	0.42	0.37	0.52	0.46	0.16	0.14
P7	0.19	0.19	0.51	0.43	0.55	0.44	0.51	0.43	0.49	0.41	0.49	0.42	0.12	0.11
P8	0.6	0.5	0.52	0.49	0.56	0.52	0.52	0.5	0.62	0.49	0.5	0.48	0.32	0.24
P9	0.18	0.17	0.49	0.42	0.49	0.4	0.46	0.41	0.38	0.34	0.48	0.4	0.14	0.13
P11	0.55	0.45	0.5	0.42	0.47	0.4	0.46	0.4	0.52	0.43	0.49	0.42	0.16	0.14
Avg	0.31	**0.27**	0.51	**0.44**	0.51	**0.44**	0.48	**0.43**	0.51	**0.44**	0.50	**0.44**	0.19	**0.16**
Std	0.17	0.13	0.02	0.03	0.03	0.04	0.04	0.04	0.14	0.08	0.01	0.03	0.07	0.05

5.2 Leave-One-Session-Out (LOSO) Evaluation of ML-Based Analysis

Tables 8 and 9 show the results for the LOSO-based evaluation of our ML-based classifiers without and with class balancing during model training, respectively. Again, we report results for each of the five explored analysis models in terms of accuracy and (macro) F1 score and average over the results.

We have to conclude that even personalization does not lead to improvements in the classification performance. Again, balancing the class distributions had no noticeable effect on the capabilities of our machine learning based classifiers.

5.3 Evaluation of Non-ML-Based Analysis

In an effort to understand the reasons for the mediocre classification performance, we conducted a second set of experiments in which we relaxed the constraints on the ground truth annotation. Instead of aiming for accurate classification of each ten-second sensor data window into engaged or disengaged, we now look at detecting when students change from engaged to disengaged and vice versa. We argue that this scenario also resembles a variant of in-class practices where instructors at some point notice changes to engagement and would respond with adaptation to their teaching accordingly. The automated analysis is now focused on detecting those changes. We employ the non-classifier based approaches described in detail in Sections 4.4.2 and 4.4.3. These methods analyze

Table 8. Results for discrimination between engaged and non-engaged episodes in our student dataset for machine-learning based classifiers trained *without* class balancing and following a leave-one-session-out (LOSO) evaluation protocol.

Test	**Bayes**		**K-NN**		**DT**		**RF**		**SVM**		**Random**		**Biased Random**	
Participant	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
P1-1	0.62	0.38	0.61	0.43	0.58	0.45	0.6	0.39	0.64	0.39	0.53	0.53	0.64	0.39
P1-2	0.58	0.38	0.62	0.48	0.57	0.48	0.61	0.45	0.64	0.39	0.55	0.54	0.64	0.39
P1-3	0.18	0.18	0.75	0.43	0.73	0.48	0.81	0.52	0.83	0.45	0.47	0.41	0.83	0.45
P1-4	0.73	0.42	0.71	0.53	0.66	0.52	0.67	0.46	0.73	0.42	0.51	0.47	0.73	0.42
P2-1	0.76	0.43	0.71	0.47	0.65	0.5	0.72	0.47	0.76	0.43	0.47	0.43	0.76	0.43
P2-2	0.47	0.43	0.82	0.48	0.76	0.5	0.82	0.46	0.88	0.47	0.49	0.39	0.88	0.47
P2-3	0.68	0.41	0.67	0.47	0.64	0.5	0.7	0.49	0.71	0.41	0.47	0.45	0.71	0.41
P2-4	0.82	0.5	0.72	0.45	0.7	0.55	0.73	0.47	0.84	0.46	0.47	0.4	0.84	0.46
P4-1	0.69	0.41	0.63	0.47	0.58	0.44	0.65	0.49	0.69	0.41	0.48	0.46	0.69	0.41
P4-2	0.88	0.47	0.84	0.59	0.71	0.49	0.8	0.45	0.89	0.47	0.46	0.38	0.89	0.47
P4-3	0.86	0.49	0.86	0.46	0.73	0.46	0.9	0.52	0.96	0.49	0.52	0.39	0.96	0.49
P4-4	0.74	0.44	0.86	0.56	0.77	0.45	0.82	0.49	0.93	0.48	0.55	0.43	0.93	0.48
P5-1	0.78	0.46	0.69	0.47	0.66	0.49	0.74	0.47	0.78	0.44	0.46	0.41	0.78	0.44
P5-2	0.88	0.47	0.78	0.5	0.65	0.42	0.72	0.42	0.88	0.47	0.53	0.41	0.88	0.47
P5-3	0.87	0.47	0.85	0.49	0.69	0.48	0.66	0.44	0.88	0.47	0.53	0.42	0.88	0.47
P5-4	0.8	0.44	0.77	0.49	0.7	0.45	0.81	0.53	0.84	0.46	0.54	0.47	0.84	0.46
P7-1	0.87	0.55	0.81	0.56	0.76	0.55	0.83	0.5	0.88	0.47	0.45	0.38	0.88	0.47
P7-2	0.86	0.46	0.79	0.44	0.67	0.47	0.82	0.49	0.86	0.46	0.5	0.42	0.86	0.46
P7-3	0.78	0.44	0.75	0.51	0.67	0.48	0.78	0.5	0.81	0.45	0.49	0.45	0.81	0.45
P7-4	0.92	0.48	0.89	0.55	0.79	0.46	0.9	0.52	0.95	0.49	0.48	0.34	0.95	0.49
P8-1	0.6	0.38	0.59	0.45	0.58	0.48	0.59	0.43	0.6	0.38	0.47	0.47	0.6	0.38
P8-2	0.7	0.43	0.64	0.47	0.6	0.45	0.63	0.43	0.74	0.42	0.51	0.47	0.74	0.42
P8-3	0.7	0.41	0.65	0.43	0.63	0.5	0.68	0.47	0.7	0.41	0.5	0.48	0.7	0.41
P9-1	0.53	0.44	0.76	0.48	0.73	0.53	0.75	0.48	0.91	0.48	0.49	0.39	0.91	0.48
P9-2	0.85	0.46	0.75	0.48	0.65	0.45	0.74	0.44	0.85	0.46	0.48	0.42	0.85	0.46
P9-3	0.8	0.45	0.75	0.53	0.69	0.54	0.76	0.46	0.81	0.45	0.49	0.44	0.81	0.45
P11-1	0.86	0.52	0.79	0.48	0.74	0.48	0.85	0.51	0.89	0.47	0.47	0.38	0.89	0.47
P11-2	0.84	0.46	0.75	0.46	0.66	0.44	0.77	0.49	0.84	0.46	0.5	0.42	0.84	0.46
P11-3	0.77	0.43	0.69	0.46	0.67	0.51	0.74	0.48	0.78	0.44	0.47	0.42	0.78	0.44
P11-4	0.82	0.48	0.77	0.47	0.67	0.48	0.78	0.51	0.83	0.45	0.53	0.46	0.83	0.45
Avg	0.74	**0.44**	0.74	**0.48**	0.68	**0.48**	0.75	**0.47**	0.81	**0.45**	0.49	**0.43**	0.81	**0.45**
Std	0.15	0.06	0.08	0.04	0.06	0.03	0.08	0.03	0.09	0.03	0.03	0.04	0.09	0.03

the raw, preprocessed EDA data for significant changes across different levels of discretization. Table 10 lists the detection results for this second set of experiments, again for each session and averaged over the entire dataset. Analysis results are given as accuracy values (for consistency with previous work)[20]and as F1 scores, which is the more realistic measure for the severely imbalanced distribution of engagement changes (typically between 10 and 30 instances in standard 65-minute sessions, i.e., some 390 ten-second data windows).

6 DISCUSSION

As education has moved to online classroom environments, including MOOCs and remote learning due to COVID, there is a desire to help teachers *read the room*, to understand students' emotional engagement levels, and adjust their teaching accordingly.. Previous work in the broader field of affective computing in general, and mobile and

Table 9. Results for discrimination between engaged and non-engaged episodes in our student dataset for machine-learning based classifiers trained *with* class balancing and following a leave-one-session-out (LOSO) evaluation protocol.

Test	**Bayes**		**K-NN**		**DT**		**RF**		**SVM**		**Random**		**Biased Random**	
Participant	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
P1-1	0.34	0.32	0.45	0.4	0.53	0.48	0.41	0.36	0.4	0.33	0.48	0.47	0.36	0.26
P1-2	0.35	0.28	0.48	0.48	0.51	0.48	0.44	0.43	0.37	0.34	0.48	0.47	0.36	0.26
P1-3	0.16	0.15	0.5	0.45	0.55	0.46	0.53	0.47	0.24	0.24	0.51	0.46	0.17	0.15
P1-4	0.29	0.25	0.52	0.49	0.58	0.54	0.55	0.53	0.32	0.29	0.45	0.42	0.27	0.21
P2-1	0.31	0.3	0.57	0.53	0.49	0.47	0.45	0.43	0.55	0.54	0.5	0.47	0.24	0.2
P2-2	0.16	0.15	0.52	0.38	0.54	0.41	0.55	0.41	0.51	0.39	0.5	0.42	0.12	0.11
P2-3	0.34	0.3	0.46	0.44	0.55	0.53	0.49	0.48	0.58	0.52	0.54	0.52	0.29	0.23
P2-4	0.19	0.17	0.48	0.41	0.48	0.44	0.43	0.38	0.2	0.19	0.49	0.41	0.16	0.14
P4-1	0.31	0.27	0.55	0.51	0.58	0.57	0.48	0.47	0.3	0.23	0.44	0.42	0.31	0.24
P4-2	0.2	0.2	0.51	0.45	0.55	0.45	0.49	0.41	0.36	0.33	0.47	0.38	0.11	0.1
P4-3	0.09	0.09	0.53	0.38	0.48	0.35	0.44	0.33	0.84	0.53	0.51	0.39	0.04	0.04
P4-4	0.26	0.23	0.47	0.37	0.48	0.38	0.43	0.32	0.13	0.13	0.47	0.38	0.07	0.06
P5-1	0.21	0.17	0.44	0.41	0.49	0.44	0.46	0.42	0.41	0.4	0.43	0.4	0.22	0.18
P5-2	0.59	0.41	0.46	0.39	0.44	0.38	0.39	0.34	0.36	0.32	0.54	0.46	0.12	0.11
P5-3	0.16	0.15	0.51	0.38	0.4	0.35	0.42	0.34	0.21	0.21	0.51	0.42	0.12	0.1
P5-4	0.73	0.42	0.5	0.42	0.5	0.42	0.48	0.42	0.73	0.42	0.51	0.42	0.16	0.14
P7-1	0.79	0.46	0.52	0.44	0.56	0.47	0.48	0.42	0.64	0.52	0.56	0.46	0.12	0.11
P7-2	0.25	0.25	0.52	0.44	0.51	0.42	0.45	0.39	0.51	0.43	0.48	0.4	0.14	0.13
P7-3	0.28	0.28	0.52	0.46	0.55	0.47	0.46	0.43	0.4	0.36	0.54	0.49	0.19	0.16
P7-4	0.42	0.33	0.57	0.39	0.56	0.41	0.5	0.38	0.55	0.39	0.51	0.4	0.05	0.05
P8-1	0.56	0.4	0.54	0.53	0.51	0.5	0.5	0.5	0.44	0.39	0.48	0.48	0.4	0.28
P8-2	0.26	0.24	0.46	0.44	0.41	0.39	0.42	0.41	0.6	0.56	0.42	0.4	0.26	0.21
P8-3	0.36	0.34	0.52	0.5	0.5	0.48	0.5	0.49	0.59	0.49	0.48	0.45	0.3	0.23
P9-1	0.11	0.11	0.5	0.4	0.53	0.4	0.38	0.33	0.24	0.24	0.48	0.38	0.09	0.08
P9-2	0.77	0.47	0.44	0.39	0.52	0.43	0.41	0.38	0.46	0.36	0.49	0.42	0.15	0.13
P9-3	0.24	0.23	0.49	0.44	0.59	0.52	0.47	0.43	0.46	0.43	0.48	0.42	0.19	0.16
P11-1	0.58	0.43	0.49	0.39	0.46	0.35	0.43	0.35	0.45	0.33	0.5	0.41	0.11	0.1
P11-2	0.7	0.5	0.52	0.43	0.53	0.44	0.54	0.47	0.56	0.44	0.52	0.47	0.16	0.14
P11-3	0.61	0.49	0.52	0.48	0.51	0.46	0.48	0.44	0.56	0.5	0.47	0.45	0.22	0.18
P11-4	0.46	0.44	0.55	0.47	0.5	0.41	0.52	0.48	0.71	0.52	0.47	0.43	0.17	0.15
Avg	0.37	**0.29**	0.50	**0.44**	0.51	**0.44**	0.47	**0.41**	0.46	**0.38**	0.49	**0.43**	0.19	**0.15**
Std	0.20	0.12	0.04	0.05	0.05	0.06	0.05	0.06	0.17	0.12	0.03	0.04	0.09	0.06

ubiquitous computing in particular, established that it is possible to automatically recognize affect and various forms of engagement from EDA data recorded through wrist-worn sensing platforms.[12, 20, 28] However, most of this previous work has focused on whole class period assessments rather than the momentary analysis. Our envisioned automated *reading the room* scenario requires us to measure moment-by-moment emotional engagement.

While previous work suggests and encouraged us that such an endeavor seems possible, our study unveiled that the problem is more complex than one would expect. We have carefully designed and conducted a case study in which we followed best practices as documented in related work regarding sensor data recording and analysis. Our trained observers diligently annotated student engagement in class based on detailed operational definitions.

Table 10. Results of automated analysis of student engagement using our heuristics based approach. Sessions were manually annotated by our trained expert observers according to the operational definitions specified in Table 4.4.1. Behavior annotations were then subsequently converted into engagement changes indications, either from disengaged to engaged or vice versa. Accuracy and F1 scores listed for each session (# engagement changes given in parentheses) and for each of the three analysis methods as described in Sections 4.4.2 and 4.4.3.

Session	**Slope**		**Jump**		**Jump 2**	
(# engagement changes)	Acc	F1	Acc	F1	Acc	F1
P1-S1 (7)	0.94	0.32	0.96	0.33	0.98	0.33
P1-S2 (12)	0.94	0.32	0.91	0.32	0.96	0.33
P1-S3 (9)	0.94	0.32	0.94	0.32	0.96	0.33
P1-S4 (14)	0.9	0.32	0.93	0.32	0.95	0.35
P2-S1 (13)	0.89	0.32	0.89	0.33	0.92	0.33
P2-S2 (9)	0.95	0.32	0.94	0.32	0.96	0.33
P2-S3 (21)	0.86	0.33	0.91	0.32	0.93	0.32
P2-S4 (15)	0.91	0.32	0.91	0.34	0.95	0.32
P4-S1 (15)	0.91	0.32	0.87	0.34	0.95	0.33
P4-S2 (11)	0.88	0.31	0.87	0.33	0.9	0.34
P4-S3 (11)	0.92	0.32	0.89	0.31	0.94	0.32
P4-S4 (11)	0.94	0.32	0.91	0.35	0.95	0.35
P5-S1 (18)	0.86	0.34	0.95	0.33	0.96	0.33
P5-S2 (12)	0.91	0.32	0.9	0.34	0.94	0.32
P5-S3 (13)	0.92	0.32	0.94	0.32	0.96	0.33
P5-S4 (27)	0.84	0.34	0.9	0.33	0.93	0.32
P7-S5 (21)	0.86	0.38	0.89	0.31	0.92	0.32
P7-S6 (17)	0.88	0.36	0.83	0.31	0.87	0.31
P7-S7 (21)	0.85	0.31	0.9	0.33	0.92	0.34
P7-S8 (9)	0.89	0.31	0.85	0.31	0.89	0.31
P8-S5 (28)	0.87	0.36	0.96	0.33	0.97	0.33
P8-S6 (11)	0.94	0.32	0.94	0.32	0.96	0.33
P8-S7 (20)	0.91	0.36	0.86	0.32	0.93	0.32
P9-S5 (11)	0.94	0.32	0.92	0.32	0.95	0.33
P9-S6 (11)	0.89	0.32	0.92	0.32	0.93	0.33
P9-S7 (17)	0.89	0.31	0.92	0.34	0.94	0.32
P11-S5 (17)	0.85	0.31	0.85	0.32	0.87	0.33
P11-S6 (17)	0.84	0.31	0.85	0.33	0.88	0.34
P11-S7 (29)	0.81	0.33	0.83	0.31	0.87	0.33
P11-S8 (25)	0.84	0.31	0.7	0.3	0.79	0.3
Avg	.89	**.33**	.90	**.32**	.93	**.33**
Std	.0027	.0025	.0235	.0057	.0173	.0045

Our results show that current approaches may not be able to accurately assessing student engagement in a momentary manner.

In what follows, we discuss and contextualize our findings and offer suggestions on what should be future directions of research and development that may get us, as a community, closer to the goal of automated, momentary assessment of student engagement in the classroom.

6.1 From Prediction to Exploration

The emerging research to measure emotional engagement through wearable technology has grounded the assessment of these systems through self-reporting periodically during class or at the end of class. These approaches to assessing tend to miss the whole story of student engagement. They focus on predicting if students perceived engagement happened after the class has taken place. This is useful for a reflective assessment of teachers' effectiveness, but will not provide just-in-time data to teachers in online classrooms. We recognize that the physical classrooms used in this study provides a different context than online classes, and computer science classes are a specific context. Nevertheless, the lack of correlation found may be transferable to other settings and should be investigated.

Leveraging research from the learning sciences [56], we sought to emulate how teachers measure if students are on-task or off-task, and emotionally engaged. The goal was to explore how sensor data could be used in the moment to improve educational outcomes. Our study's exploratory approach highlights two aspects to consider in conducting studies evaluating sensor data's capabilities that respond to real classroom contexts. First, the definition, collection, and use of ground truth data should be tied to the goal of measurements. Second, the involvement of educational experts in devising, monitoring, and analyzing the collected data is necessary for identifying the objectives of the research and providing methodological expertise beyond analysis of sensor data.

6.2 Reading the Room: Are We There Yet?

Our analysis demonstrates no correlation between the moment-by-moment observations that a teacher might make in the classroom and the wearable sensor data collected. This calls into question the effectiveness of current wearable sensors in detecting student engagement in classroom settings that might be useful for teachers. This limitation might be tied to the context of the classroom. First, the wearable sensor data was able to measure affect, however, it was not necessarily tied to on-task or off-task affect in the classroom. For example, while a student might be emotionally engaged during class, the sensors could not tell if that engagement came from the professor's performance or unrelated tasks, such as reading an email. Second, it is unclear if the physiological signs can measure the small changes in student affect during a course. For example, if a student is mildly interested then becomes a little less interested, it is not clear that any physical changes would occur that would be measurable with current sensors. We must also consider that the observations, not the sensors, were not accurate information about student engagement; students, who knew they were being observed and sitting in the front of the classroom, could have been putting on a performance to act engaged or not.

6.3 Towards Future Evaluation Studies

This work calls into question if the current state of wearable sensors accurately track student engagement in a moment-by-moment manner. This suggests that we need to seek other methods of validation for student engagement and discuss if there are other alternatives for remote sensing that will be helpful to teachers during class time.

We could pursue video-based approaches that track facial expressions, hand gestures, and postures as other researchers have demonstrated better success in measuring student engagement with these tools in controlled settings. Turning on cameras in classes also may increase engagement because students who are monitored are more likely to stay on task [69]. However, videos in online classrooms have significant barriers. Technically the quality of the video dramatically changes its usability. Many of the studies on student engagement detection using video rely on tightly controlled environments where the quality of the camera and lighting are consistent [63]. From our experience with teaching online, we observed students are in public spaces, family living rooms, or their car (hopefully parked) where the video quality was poor or frequently interrupted. In online video

conferencing systems, we often observe something as simple (and uncontrollable) as cloud coverage dramatically decreasing the quality of videos.

Students' concerns with privacy will also likely limit the functionality of such video-based sensors. During remote learning due to COVID-19, many students turned off their cameras; the online video format made them uncomfortable about how they looked or what their personal space might communicate about them [55]. This is particularly true for students from low socio-economic backgrounds. In addition, the work that has sought to use videos in online classes has been used primarily as a predictive measure of student performance or teacher engagement, not as a measure to improve engagement in the moment [54].

There may be ways to augment the information received from wearable sensor data with multi-modal inputs from behavior data, such as clickstream data [59]. Previous work has focused on clickstream data for prediction [17] and helping students with self-regulation [48]. This work is promising in assisting MOOC students in managing their learning. However, this work has not focused on just-in-time feedback to assist teachers in live lectures and is often noisy and difficult for teachers to interpret and use to improve their courses [6]. However, by combining clickstream data with wearable sensor data we may better be able to determine on-task and off-task behavior and emotional engagement.

Finally, we need to question if the emotional sensing that teachers do is even related to the physiological data that wearable devices can provide. Are there other measures that can be used that might yield more effective data for teachers? Should we conduct interviews with teachers and educational experts on classroom behavior to concretely identify the methods that current teachers use?

6.4 Limitations

There are a number of limitations to this work. The use of observational methods, while accepted in the field of intelligent tutors, have been called into question in psychology research because of the ability of students to fake engagement or perhaps look uninterested even if they are engaged. For the granularity of our study, seeking moment-by-moment measurements of engagement, we do not have a better tool to use. However, the work would be more robust with a corresponding retrospective self-assessment or recall interview to triangulate the findings.

We also recognize that the physical classroom and the context of computer science likely impact the way that people experience, behave, and express their emotions. It may also be that the presence of observers or sitting at the front of the class changed the way that students acted. In light of these limitations, we hope this work will be taken as a first step to investigating ways to help online teachers *read the room.*

7 CONCLUSION

When schools went online during the COVID-19 pandemic, we saw teachers struggle with online classroom engagement, and researchers sought to help them address these issues with technology. This study demonstrated that using rigorous qualitative observations in classroom settings and quantitative analysis of wearable sensor data did not correlate with emotional engagement in the class. However, there are promising future explorations and directions to help us understand if wearable data can be used to measure student affect to improve online education.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive criticisms and suggestions that helped us improving the quality of this paper. Furthermore, we thank Yi He, Shan Jing, David Joyner, and Lauren Wilcox for their help in conceptualizing and conducting this study. This material is based upon work supported by the National Science Foundation under Grant No. 1842693.

REFERENCES

[1] Sasirekha Anbusegaran. 2021. Unobtrusive Assessment Of Student Engagement Levels In Online Classroom Environment Using Emotion Analysis. (2021).

[2] James J Appleton, Sandra L Christenson, and Michael J Furlong. 2008. Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools* 45, 5 (2008), 369–386.

[3] TS Ashwin and Ram Mohana Reddy Guddeti. 2020. Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems* 108 (2020), 334–348.

[4] TS Ashwin and Ram Mohana Reddy Guddeti. 2020. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies* 25, 2 (2020), 1387–1415.

[5] TS Ashwin and Ram Mohana Reddy Guddeti. 2020. Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. *User Modeling and User-Adapted Interaction* 30, 5 (2020), 759–801.

[6] Rachel Baker, Di Xu, Jihyun Park, Renzhe Yu, Qiujie Li, Bianca Cung, Christian Fischer, Fernando Rodriguez, Mark Warschauer, and Padhraic Smyth. 2020. The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *International Journal of Educational Technology in Higher Education* 17, 1 (2020), 1–24.

[7] Jorn Bakker, Mykola Pechenizkiy, and Natalia Sidorova. 2011. What's your current stress level? Detection of stress patterns from GSR sensor data. In *2011 IEEE 11th international conference on data mining workshops*. IEEE, 573–580.

[8] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media.

[9] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[10] Ann L Brown. 1992. Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The journal of the learning sciences* 2, 2 (1992), 141–178.

[11] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 1–33.

[12] Ryan Cain and Victor R Lee. 2016. Measuring electrodermal activity to capture engagement in an afterschool maker program. In *Proceedings of the 6th Annual Conference on Creativity and Fabrication in Education*. 78–81.

[13] Rafael A Calvo and Sidney D'Mello. 2012. Frontiers of affect-aware learning technologies. *IEEE Intelligent Systems* 27, 6 (2012), 86–89.

[14] Jingjing Chen, Bin Zhu, Olle Balter, Jianliang Xu, Weiwen Zou, Anders Hedman, Rongchao Chen, and Mengdie Sang. 2017. FishBuddy: promoting student engagement in self-paced learning through wearable sensing. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 1–9.

[15] Adrian Colomer Granero, Félix Fuentes-Hurtado, Valery Naranjo Ornedo, Jaime Guixeres Provinciale, Jose M Ausín, and Mariano Alcañiz Raya. 2016. A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents. *Frontiers in computational neuroscience* 10 (2016), 74.

[16] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 1 (2001), 32–80.

[17] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. 2016. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the sixth international conference on learning analytics & knowledge*. 6–14.

[18] João Roberto de Toledo Quadros, Fabio Paschoal, and Laercio Brito Gonçalves. 2017. Facial recognition system for automatic presence control in a classroom. In *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 1–6.

[19] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 653–656.

[20] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement During Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21.

[21] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2019. Laughter Recognition Using Non-invasive Wearable Devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 262–271.

[22] Sidney D'Mello. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology* 105, 4 (2013), 1082.

[23] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 1–36.

[24] Prakash Duraisamy, James Van Haneghan, William Blackwell, Steve Jackson, G Murugesan, and KS Tamilselvan. 2019. Classroom engagement evaluation using computer vision techniques. In *Pattern Recognition and Tracking XXX*, Vol. 10995. International Society for Optics and Photonics, 109950R.

[25] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. Investigating Visitor Engagement in Interactive Science Museum Exhibits with Multimodal Bayesian Hierarchical Models. In *International Conference*

on Artificial Intelligence in Education. Springer, 165–176.
[26] Jennifer A Fredricks and Wendy McColskey. 2012. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement*. Springer, 763–782.
[27] Kathryn A Fuller, Nilushi S Karunaratne, Som Naidu, Betty Exintaris, Jennifer L Short, Michael D Wolcott, Scott Singleton, and Paul J White. 2018. Development of a self-report instrument for measuring in-class student engagement reveals that pretending to engage is a significant unrecognized problem. *PloS one* 13, 10 (2018), e0205828.
[28] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. 2020. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–26.
[29] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. 2014. Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 39–42.
[30] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2018. Using students' physiological synchrony to quantify the classroom emotional climate. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 698–701.
[31] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–19.
[32] Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. 2020. Detection of artifacts in ambulatory electrodermal activity data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–31.
[33] Aurélien Géron. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".
[34] Michail N Giannakos, Kshitij Sharma, Sofia Papavlasopoulou, Ilias O Pappas, and Vassilis Kostakos. 2020. Fitbit for learning: Towards capturing the learning experience using wearable sensing. *International Journal of Human-Computer Studies* 136 (2020), 102384.
[35] Benjamin Goldberg, Keith W Brawner, and Heather K Holden. 2012. Efficacy of measuring engagement during computer-based training with low-cost electroencephalogram (EEG) sensor outputs. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 56. SAGE Publications Sage CA: Los Angeles, CA, 198–202.
[36] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2015. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2015), 797–804.
[37] Reza Hadi Mogavi, Xiaojuan Ma, and Pan Hui. 2021. Characterizing Student Engagement Moods for Dropout Prediction in Question Pool Websites. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.
[38] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proc. Joint Int. Conf. Artificial Intelligence (IJCAI)*.
[39] Andy Hargreaves. 2000. Mixed emotions: Teachers' perceptions of their interactions with students. *Teaching and teacher education* 16, 8 (2000), 811–826.
[40] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. 2017. EngageMeter: A system for implicit audience engagement sensing using electroencephalography. In *Proceedings of the 2017 Chi conference on human factors in computing systems*. 5114–5119.
[41] Jennifer Healey and Rosalind W Picard. 1998. Startlecam: A cybernetic wearable camera. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*. IEEE, 42–49.
[42] Rui Henriques, Ana Paiva, and Claudia Antunes. 2013. Accessing emotion patterns from affective interactions using electrodermal activity. In *2013 humaine association conference on affective computing and intelligent interaction*. IEEE, 43–48.
[43] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. 2014. Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 307–317.
[44] J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.
[45] Charlotte Larmuseau, Pieter Vanneste, Jan Cornelis, Piet Desmet, and Fien Depaepe. 2019. Combining physiological data and subjective measurements to investigate cognitive load during complex learning. *Frontline Learning Research* 7, 2 (2019), 57–74.
[46] Celine Latulipe, Erin A Carroll, and Danielle Lottridge. 2011. Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1845–1854.
[47] Mark R Lepper, Maria Woolverton, Donna L Mumme, and J Gurtner. 1993. Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. *Computers as cognitive tools* 1993 (1993), 75–105.

[48] Qiujie Li, Rachel Baker, and Mark Warschauer. 2020. Using clickstream data to measure, understand, and support self-regulated learning in online courses. *The Internet and Higher Education* 45 (2020), 100727.
[49] Qing Li, Yuan Ren, Tianyu Wei, Chengcheng Wang, Zhi Liu, and Jieyu Yue. 2020. A Learning Attention Monitoring System via Photoplethysmogram Using Wearable Wrist Devices. *Artificial Intelligence Supported Educational Technologies* (2020), 133.
[50] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and knowledge discovery* 15, 2 (2007), 107–144.
[51] David T Lykken and Peter H Venables. 1971. Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology* 8, 5 (1971), 656–672.
[52] Debra K Meyer and Julianne C Turner. 2006. Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review* 18, 4 (2006), 377–390.
[53] Ritayan Mitra and Pankaj Chavan. 2019. DEBE feedback for large lecture classroom analytics. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. 426–430.
[54] Ahmed Ali Mubarak, Han Cao, and Salah AM Ahmed. 2021. Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies* 26, 1 (2021), 371–392.
[55] Lorenz S Neuwirth, Svetlana Jović, and B Runi Mukherji. 2020. Reimagining higher education during and post-COVID-19: Challenges and opportunities. *Journal of Adult and Continuing Education* (2020), 1477971420947738.
[56] Jaclyn Ocumpaugh. 2015. *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual.* Technical Report. Technical Report. New York, NY: Teachers College, Columbia University ….
[57] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.
[58] Maja Pantic and Leon JM Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* 91, 9 (2003), 1370–1390.
[59] Jihyun Park, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, and Mark Warschauer. 2017. Detecting changes in student behavior from clickstream data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 21–30.
[60] Reinhard Pekrun, Anne C Frenzel, Thomas Goetz, and Raymond P Perry. 2007. The control-value theory of achievement emotions: An integrative approach to emotions in education. In *Emotion in education*. Elsevier, 13–36.
[61] Penelope L Peterson, Susan R Swing, Kevin D Stark, and Gregory A Waas. 1984. Students' cognitions and time on task during mathematics instruction. *American Educational Research Journal* 21, 3 (1984), 487–515.
[62] Sintija Petrovica. 2013. Adaptation of tutoring to students' emotions in emotionally intelligent tutoring systems. In *2013 second international conference on e-learning and E-technologies in education (ICEEE)*. IEEE, 131–136.
[63] Phuong Pham and Jingtao Wang. 2015. AttentiveLearner: improving mobile MOOC learning via implicit heart rate tracking. In *International conference on artificial intelligence in education*. Springer, 367–376.
[64] Phuong Pham and Jingtao Wang. 2018. Predicting learners' emotions in mobile MOOC learning via a multimodal intelligent tutor. In *International Conference on Intelligent Tutoring Systems*. Springer, 150–159.
[65] Rosalind W Picard. 2000. *Affective computing*. MIT press.
[66] Thomas PlÖtz. 2021. Applying machine learning for sensor data analysis in interactive systems: Common pitfalls of pragmatic use and ways to avoid them. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–25.
[67] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. 2010. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE transactions on Biomedical engineering* 57, 5 (2010), 1243–1252.
[68] Jonathan P Rowe, Bradford W Mott, and James C Lester. 2015. Opportunities and Challenges in Generalizable Sensor-Based Affect Recognition for Learning.. In *AIED Workshops*.
[69] Chanel M Schwenck and Jessica D Pryor. 2021. Student perspectives on camera usage to engage and connect in foundational education classes: It's time to turn your cameras on. *International Journal of Educational Research Open* 2 (2021), 100079.
[70] Gale M Sinatra, Benjamin C Heddy, and Doug Lombardi. 2015. The challenges of defining and measuring student engagement in science. , 13 pages.
[71] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*. 33–40.
[72] B Troy Frensley, Marc J Stern, and Robert B Powell. 2020. Does student enthusiasm equal learning? The mismatch between observed and self-reported student engagement and environmental literacy outcomes in a residential setting. *The Journal of Environmental Education* 51, 6 (2020), 449–461.
[73] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz, and Betsy Disalvo. 2020. Sensing Affect to Empower Students: Learner Perspectives on Affect-Sensitive Technology in Large Educational Contexts. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 63–76.
[74] James E Willis III and Viktoria Alane Strunk. 2015. Ethical responsibilities of preserving academicians in an age of mechanized learning: Balancing the demands of educating at capacity and preserving human interactivity. In *Rethinking machine ethics in the age of ubiquitous*

technology. IGI Global, 166–195.
[75] Timothy J Xeriland. 2018. *Intelligent Tutor Systems Addressing Student Disengagement: Adding Formative Reappraisal to Enhance Engagement and Learning*. Michigan State University.