Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus

Rafael Wampfler
Dept. of Computer Science
ETH Zurich, Switzerland
wrafael@inf.ethz.ch

Severin Klingler
Dept. of Computer Science
ETH Zurich, Switzerland
kseverin@inf.ethz.ch

Barbara Solenthaler
Dept. of Computer Science
ETH Zurich, Switzerland
solenthaler@inf.ethz.ch

Victor R. Schinazi
Dept. of Humanities, Social and Political Sciences
ETH Zurich, Switzerland
scvictor@ethz.ch

Markus Gross
Dept. of Computer Science
ETH Zurich, Switzerland
grossm@inf.ethz.ch

ABSTRACT

The role of affective states in learning has recently attracted considerable attention in education research. The accurate prediction of affective states can help increase the learning gain by incorporating targeted interventions that are capable of adjusting to changes in the individual affective states of students. Until recently, most work on the prediction of affective states has relied on expensive and stationary lab devices that are not well suited for classrooms and everyday use. Here, we present an automated pipeline capable of accurately predicting (AUC up to 0.86) the affective states of participants solving tablet-based math tasks using signals from low-cost mobile bio-sensors. In addition, we show that we can achieve a similar classification performance (AUC up to 0.84) by only using handwriting data recorded from a stylus while students solved the math tasks. Given the emerging digitization of classrooms and increased reliance on tablets as teaching tools, stylus data may be a viable alternative to bio-sensors for the prediction of affective states.

Keywords

Classification, Affective Computing, Stylus, Biometric Sensors

1. INTRODUCTION

Affective states are psycho-physiological constructs used to characterize the emotions (short-lived) and moods (long-lived) that arise and are experienced while individuals are engaged with a stimulus. Affective states play an important role in the educational context and can directly influence a student's learning gain [9, 26, 10]. For example, learning outcomes have been found to decrease if frustration is persistent during problem solving, whereas overcoming a state of frustration can have a positive effect on learning [10].

Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor Schinazi and Markus Gross "Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 198 - 207

Previous research has investigated the relationship between affective states and learning performance by attempting to detect the diverse emotions that occur during learning. The logic behind this approach is that, depending on the emotion of a student, appropriate actions can be taken in order to assist students during learning (e.g., adapting task elements in the case of intelligent tutoring systems (ITS) and self-regulation by providing affective feedback).

Previously, a wide range of data sources have been used to measure and predict affective states in the learning context including audio and video [32], interaction data [21, 14] and bio-sensors [8, 4]. Systems that rely on the analysis of audio (e.g., speech) and video (e.g., facial expression) data [32] cannot guarantee full anonymity and are subject to privacy issues. Given these limitations, researchers have attempted to derive the affective state of individuals based on interaction data which contain log data of the user's interaction with the learning system, such as input and error behavior, timing and help calls [21, 14]. Although large and powerful interaction data sets can be easily collected especially in online environments, the features are typically dependent on the learning domain and on the specific learning system. Attempts towards a cross-domain or cross-system engagement model have been presented (e.g., for learning spelling and math [18]), but these generalized methods typically have a lower accuracy as domain-specific features. Data from bio-sensors (e.g., measuring muscle activity [8] and heart rate [4]) have also been used to predict emotions. However, most of these devices are typically restricted to lab settings, expensive and difficult to operate, and somewhat intrusive. Recently, a variety of portable and low-cost bio-sensor devices have become available (e.g., Shimmer GSR+ and Polar H10). These devices have the potential to transform education research because they can be used to monitor a learner's physiological state at home or in a classroom.

In this paper, we explore a low-cost mobile setup to detect the affective state of students. Our goal is a system to detect affective states that is cheap and easy to operate, can be used outside a lab setting, is non-intrusive, and minimizes potential issues related to privacy. We consider bio-sensor data from skin conductance, heart measures, and skin temperature. In addition, and in contrast to previous work in

the field of learning systems, we also evaluate handwriting data recorded by a stylus to predict the affective state of students. Here, we use the fact that tablets bundled with a stylus are becoming increasingly available in households and classrooms and are inherently non-intrusive and mobile.

We propose a generic pipeline in which we process the data from the bio-sensors and stylus in order to extract a set of features for each of the sensors. We then use a classification model to predict the current affective region in the valence-arousal space of emotions [29]. Valence describes how much an emotion is perceived as positive or negative and arousal represents the intensity of the emotion. Our method allows researchers to define arbitrary areas of interest in the valence-arousal space, and can be applied to a wide range of applications and questions of interest. We evaluated our method by applying it to a math problem solving scenario in which participants provided answers in unstructured handwriting on a tablet device. Best performance was reached when data from all sensors was used for prediction (0.88 AUC). Interestingly, we reached a comparable performance using only the data acquired by the stylus (0.84 AUC). These results suggest that a simple tablet with a stylus can be sufficient to reliably predict a student's emotional state. Finally, we also explored whether the affective state model could be generalized over domains. For this purpose, we applied the trained model to a passive setting with picture stimuli leading to a performance of 0.68 AUC.

2. RELATED WORK

Affective States and Interaction Data. Due to their influence on learning gain, affective states play an important role in education in general and in particular during math learning [27, 21]. Boredom was shown to negatively influence the learning gain [9, 26], while engaged concentration can improve the learning outcome [9]. Interestingly, frustration and confusion can positively affect learning in case the student is able to resolve these states [10]. One line of research tries to predict these affective states based on logged user interactions only. Frustration, boredom, engaged concentration and confusion have been successfully predicted using interaction data for math tutoring systems [21, 14]. On the other hand, valence and arousal have been predicted using mouse and keyboard interaction data from writing compositions in free text [31]. Moreover, generalized models have been proposed, such as an engagement model for two different learning domains and tutors (spelling and math) [18]. Based on such automatically predicted affective states, different intervention strategies have been explored. An automatic student-centered affect-aware feedback loop was shown to increase the learning gain [14] while other work explored how teachers can provide better interventions based on real-time information about the evolution of student's affective states [11].

Biometric Sensors. Biometric sensors provide an objective measure of the physiological reactivity of users engaging with a learning environment while minimizing interference with the actual task [19, 4, 17, 30]. Indeed, education research has investigated the effectiveness of a variety of physiological signals used to infer affective states. Electrodermal activity, skin temperature, and heart rate were generally found to be good predictors of emotions [19, 17, 30] and mind wandering [4] across different tasks including math learning [17, 30], scientific text reading [4] and audio, visual and cognitive stimuli in general [19]. However, these previous works mainly focused on expensive, high quality sensors to provide medical grade accuracy for the measurement of physiological signals. In contrast, we focus on an affective tutor that can be used in learning systems, hence we gather such data in a non-intrusive and easy to use way.

Stylus. Predicting affective states based on stylus data is still a relatively new research topic. Likforman-Sulem et al. [24] predicted anxiety, depression and stress based on figure drawings and writing given words. Fairhurst et al. [12] conducted an experiment for predicting stress and happiness by letting participants writing down a given list of words and describing a visual scene in own words. Instead of predicting a fixed set of affective states, our approach can capture different affective regions which can be defined according to the researchers need. Our approach is not restricted to copying predefined sentences and figures but works with arbitrary handwriting and drawing. To our knowledge, this is the first work to leverage stylus data in order to predict the affective state of a student during math solving.

3. METHOD

We present a classification pipeline that automatically predicts affective states based on low-cost and mobile bio-sensor and stylus devices. Our pipeline assumes that we have access to reports on affective states of users based on the circumplex model of affects [29]. The circumplex model is a two-dimensional model representing affective states in terms of valence and arousal. The classification task then amounts to classifying regions within this space using a combination of signals from bio-sensor and stylus devices. For this purpose, we build a generic affective predictor (Figure 1). Recorded stylus and bio-sensor data are preprocessed and the relevant features are extracted to train a classification model for the specific affective regions. We design our predictor to work unobtrusively in the background of any ITS.

3.1 Input Signals

During the task solving process bio-sensor and stylus data are recorded.

Electrodermal activity (EDA). EDA is an indicator of the emotional state of a person reflected by the variation in the electrical characteristics of the skin as a result of sweating [2]. EDA is quantified by measuring the amount of current flowing between electrodes attached to the skin. Changes in affective states can lead to subtle variations in the level of sweat that can be detected as the changes in the current. Typically, the EDA signal is decomposed into tonic (low frequency) and phasic (high frequency) components.

Interbeat Intervals (IBIs). IBIs are the time intervals between consecutive heartbeats in normal heart function. This natural variation is also known as heart rate variability (HRV). The heart rate (HR) can be easily computed as the inverse of the IBI averaged over a certain time window.

Skin Temperature (ST). ST measures the thermal response of human skin. Vasoconstriction (e.g., provoked by

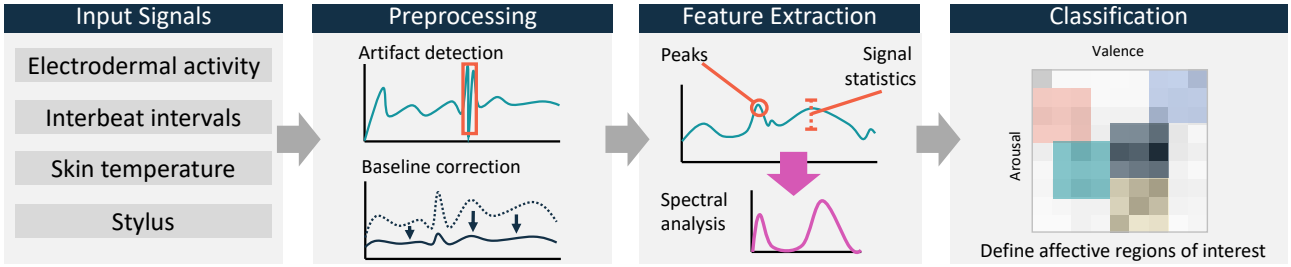

Figure 1: The classification pipeline. Stylus and bio-sensor data are gathered during task solving processes. After preprocessing the signals, features are extracted and used to classify the affective regions of interest.

an affective state) can result in an increase in blood flow and a consequent increase in ST [19].

Stylus. Tablet devices often come equipped with stylus pens as accessories that can provide precise and pressure-sensitive input. Stylus data consists of the applied pressure during writing and the pixel positions of the written text. From these measurements, handwriting characteristics related to time and ductus can be calculated.

3.2 Preprocessing of Signals

During preprocessing, the raw input signals are filtered in order to detect artifacts from movement and muscle contraction. The signals are also corrected for differences between individuals using baseline recordings for each individual.

Artifact detection. We follow the procedure outlined by Greco and colleagues [15] to decompose the EDA into tonic, phasic and an additive white Gaussian noise component with a convex optimization approach that accounts for signal filtering and detrending. For IBIs, detrending is not necessary in the preprocessing [36], and we use the criterion beat difference for artifact detection [16].

Baseline correction. Similar to previous work [30, 17], we collect baseline data for all sensors in order to account for individual differences in stylus and bio-sensor signals related to writing habit, ambient temperature and dryness of the skin. Baseline data is collected while individuals remain in a relaxed state (e.g., watching a nature video). We search for the minimum value of each bio-sensor signal during the relaxation phase over a 10 seconds window using a sliding window approach to be robust against outliers. Due to possible signal lags, we search the minimum for each signal separately. We then normalize the bio-sensor data by subtracting the feature values calculated over the corresponding 10 seconds interval of the baseline from the actual feature values computed during task solving. Stylus data is normalized by subtracting a baseline for all features computed over handwriting of an English sentence.

3.3 Feature Extraction

In the proposed pipeline, we extract several different feature types from the stylus and bio-sensor signals. Where appropriate, we compute basic statistics for these features types including the mean, standard deviation (SD), minimum and maximum and the linear trend (slope of a fitted linear regression line). A summary of all extracted features is presented in Table 1.

EDA. For EDA, we decompose the signal into phasic and tonic components and calculate standard statistics (i.e., mean, SD, min, max, slope). For the phasic component, we also calculate the area under the curve (AUC) [3] and the number of peaks using zero-crossings of the smoothed gradients of the signal [19]. Based on the extracted peaks, we further compute amplitude statistics (i.e., mean, min, max) [38].

IBI. From the IBI recordings, we extract temporal and frequency features. In the temporal domain, we calculate the percentage of successive IBIs that differ by more than 50 milliseconds (pNN50) and 20 milliseconds (pNN20) as well as the SD and root mean square of successive differences between adjacent IBIs (SDSD and RMSSD) [34, 25]. For the frequency domain, it is well known that the distribution of spectral power gives an indication of physiological activation [3]. Therefore, we extract a feature related to the high frequency (HF) band of 0.15-0.40 Hz by a Fast Fourier transform of the cubic spline interpolated signal [34, 25]. Based on the IBIs, we compute the heart rate for which we extract several standard statistics (i.e., mean, SD, min, max, slope).

ST. We extract several statistics (i.e., mean, SD, min, max, slope) from the temperature signal [38, 35].

Stylus. From the stylus data, we derive features related to the pressure applied by the pen as well as timing and location information. Previous research has successfully employed these features to predict affective states [24, 12]. From the pressure data, we compute standard statistics (mean, SD, max, min) per stroke and average these over an entire task. Additionally, over each task we compute the slope of a linear regression fit to the pressure values and the statistical skewness of the pressure distribution. We also compute standard statistics (i.e., mean, SD, max, min, slope) of the speed and acceleration of the strokes. For the handwriting data, we discriminate between the actual writing process and the think time while completing the task [24]. During writing there are always small time gaps between strokes which cannot be attributed to thinking but belong to the writing process itself. Because writing patterns are different for every user, we infer an individual threshold for each user to distinguish if the time between two strokes belongs to thinking or to the actual writing process. We chose this threshold as the 80 % cut-off value of the distribution of the time between the strokes over the stylus baseline (cropping the right tail of the distribution). Based on this threshold, we derive a feature measuring the percentage of writing (i.e., the time spent in the writing process). Additionally, we compute the

statistics (i.e., mean, SD, max, min) on the speed between consecutive strokes having time differences below threshold (writing process) and on the distance between strokes having time differences above threshold (thinking).

Table 1: Extracted bio-sensor and stylus features. For each signal, the features are sorted according to their importance (based on our experiments). The 10 most predictive features are highlighted in bold. SD refers to the standard deviation.

Signals	Features
EDA	**Phasic AUC**, **Phasic Mean**, **Tonic SD**, Tonic Max, Tonic Mean, Tonic Min, Phasic SD, # Phasic Peaks, Tonic Slope, Max Phasic Peak Amplitude, Min Phasic Peak Amplitude, Phasic Slope, Mean Phasic Peak Amplitude
Heart	**IBI SDSD**, **IBI RMSSD**, **IBI SD**, IBI pNN20, HR Mean, IBI High Frequency, IBI pNN50, IBI Mean, HR Min, HR Max, HR SD, HR Slope
Temperature	Max, Mean, Min, Slope, SD
Stylus	**#Strokes/Mean Speed**, **Mean Distance between Strokes**, **Max Distance between Strokes**, **SD Distance between Strokes**, Mean Pressure, Max Pressure, Mean Stroke Acceleration, Max Stroke Acceleration, Max Stroke Speed, Max Speed between Strokes, Mean Speed between Strokes, SD Speed between Strokes, SD Stroke Speed, SD Stroke Acceleration Excluded[1]: %Writing, {SD, Slope, Skewness} Pressure, {Mean, Min, Slope} Stroke Speed, {Min, Slope} Stroke Acceleration, Min Speed between Strokes, Min Distance between Strokes, #Strokes/Minute

[1] Excluded due to our experimental setup (see Section 5.1)

3.4 Classification

To train our classification algorithms ground truth is built by defining arbitrary non-overlapping regions of interest in the two-dimensional valence and arousal space based on the affective labels which can be gathered, for example, through self-reports or expert labelers. We then use a classification model to predict the affective region an individual is likely to be in during task solving based on the recorded bio-sensor and stylus data. Before applying the classification algorithm, we standardize all features to have zero mean and unit variance. We propose the usage of four different classifiers (i.e., Random Forest, Support Vector Machine, k-Nearest Neighbors and Gaussian Naive Bayes). We select these classifiers because they are among the most widely used in machine learning and have shown to provide good results on bio-sensor and stylus data [37, 24, 13]. All models are evaluated using leave-one-user-out cross-validation which ensures that data from the same user is not in the testing and training set at the same time. Hyperparameter optimization is performed using nested cross-validation and randomized search.

4. EXPERIMENT

We conducted a controlled lab experiment with 88 participants in order to test our pipeline. In the experiment, we recorded bio-sensor and stylus data while participants solved approximately 40 math tasks chosen to trigger different affective states. The math tasks were chosen because they are an integral part of the educational curriculum. However, instead of relying on a math based ITS, we have designed specific math tasks to increase the probability of evoking a wider range of affective states.

4.1 Experimental Setup

Participants. We recruited 88 participants (45 female) between ages of 18 and 29 (mean = 22.1, SD = 2.0) from 10 different engineering and natural science departments of the second and third year of the Bachelor program of an university. We excluded participants suffering from cardiovascular pathologies, smokers, and participants suffering from evident mental pathologies (score $>$ 4 in the Patient Health Questionnaire [22]). In order to control for external factors, we kept the humidity and room temperature at an average of 21.7 °C (SD = 0.59 °C) and 32.6 % (SD = 5.3 %), respectively. Figure 2 presents the experimental setup.

Sensors. We measured EDA and wrist acceleration using a Shimmer GSR+ device. To test the accuracy of the device, we compared its measurements with a state of the art ADInstruments PowerLab 8/35 device (connected through the ADInstruments FE116 GSRAmp signal amplifier) over a 23 minute recording of an user watching a nature video and picture stimuli. Results revealed a strong and significant cross-correlation value of 0.96 (p-value $< 10^{-100}$) between the two signals. These results suggest that the smaller, mobile and more affordable Shimmer GSR+ device may be sufficient to detect changes in affective states. During the experiment, the Shimmer GSR+ device was worn on the non-dominant hand with the electrodes placed at the proximal phalanx of the index and middle finger [7]. Data was recorded at a sampling rate of 100 Hz. As part of the Shimmer GSR+ setup, we also attached an optical pulse sensor providing a photoplethysmogram signal on the ring finger. However, photoplethysmogram data was of poor quality and consequently discarded from analysis. Prior to electrode attachment, we asked participants to wash their hands with lukewarm water [5]. Heart activity was measured using a Polar H10 chest belt. The Polar H10 belt provides IBIs and post-processed heart rate data by monitoring electrical changes on the surface of the skin. A predecessor of this device (Polar H7) was shown to provide accurate data when compared to an expensive lab device (Cosmed Quark T12x system) [28]. We recorded the skin temperature using the infrared thermopile sensor of the Empatica E4 device (sampling rate = 4 Hz; resolution = 0.02 °C). Since the sensor was attached to the dominant hand (used for writing during the tasks), other signals that the wristband can provide (EDA and blood volume pulse) were heavily affected by motion artifacts and discarded from the analyses.

During the experiment, participants interacted with a Huawei MediaPad M2 10.0 running Android 5.1 to solve the different math tasks. All interactions with the tablet were conducted with a Wacom Bamboo Ink stylus at an average sampling rate of 250 Hz (SD = 25 Hz) and with 2048 levels of pressure sensitivity. The signals from the bio-sensor devices were streamed to the tablet using the Bluetooth Low Energy protocol. We also recorded the behaviour of participants with the front camera of the tablet and a GoPro HERO3 camera.

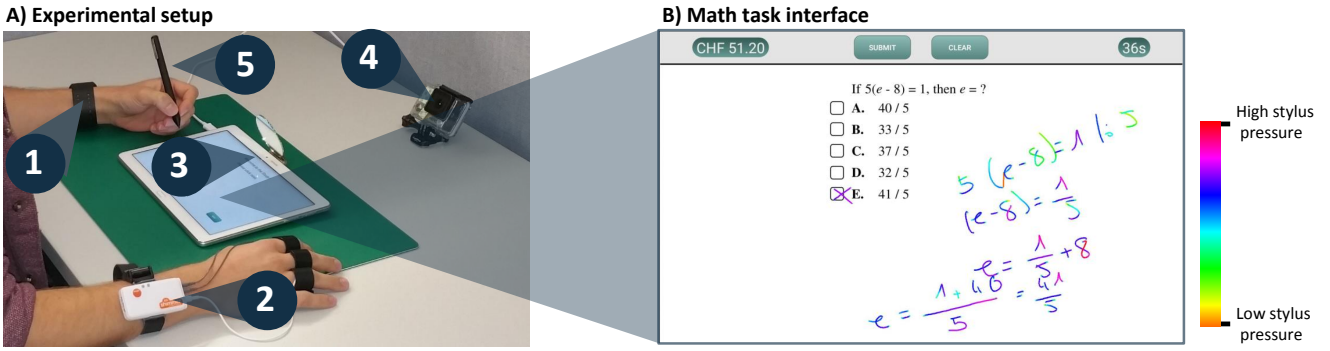

Figure 2: A participant completing the math tasks. A) During each session data is recorded from different devices. (1) An Empatica E4 recording skin temperature on the dominant hand. (2) A Shimmer GSR+ measuring skin conductance and wrist acceleration on the non-dominant hand. Participant behaviour was recorded by (3) the tablet front cam and (4) a GoPro HERO3. All interactions with the tablet were conducted with a stylus (5). Participants also wore a Polar H10 chest belt (not visible in the image) for recording heart activity. B) The task interface allows participants to write solution paths directly onto the screen (the stylus pressure is color-coded for visualization purposes only).

4.2 Experimental Procedure

For measuring the affective states of the participants we have used the self-assessment manikin (SAM) [6]. The SAM provides valence and arousal labels on a scale from 1 (most negative, lowest arousal) to 9 (most positive, highest arousal). For triggering the affective states we have used math tasks and pictures from the International Affective Picture System (IAPS) [23]. The IAPS is a database of 1182 pictures typically used in emotion research and has been standardized in terms of valence and arousal based on SAM ratings. We used the IAPS to investigate whether the affective model for the math tasks generalized to passive tasks, such as watching pictures (Section 5.6). As such, the set of IAPS pictures presented to the participants was sampled to cover similar affective regions as those expected to be evoked by the different math tasks.

An overview of the study procedure is presented in Figure 3A. The experiment lasted an average of 90 minutes for each participant. Upon arriving at the lab, participants completed a demographics questionnaire and were given an oral overview of the procedure. This included an explanation of the SAM questionnaire based on 4 example pictures from the IAPS presented on paper. Next, participants started working independently on the tablet by first watching a 7 minute nature video (bio-sensor baseline), followed by the stylus baseline that consisted of writing an English sentence with the stylus. Participants were then presented with 40 pictures from the IAPS in random order. Each picture was shown for 10 seconds and was directly followed by the SAM rating (valence and arousal) and a 10 second fixation cross. In total, we collected 3400 ratings from all participants. After rating the IAPS pictures, participants were asked to watch the nature video one more time before completing the math tasks. Before finishing the experiment, participants completed a paper questionnaire about their overall mood, comfort level while wearing the sensors, nervousness and sweating level.

4.3 Experimental Tasks

To trigger different affective states, we have created three different math task conditions by varying the difficulty level, available time for completion and monetary reward of the task. These types of manipulations were shown to be effective at eliciting different affective states in reading comprehension [4] and math tasks [32].

Task design. The math tasks were taken from an ACT data set [1] that provided difficulty ratings from 0.12 (most difficult) to 0.96 (simplest). We conducted a pilot study (exact same conditions, 11 participants) to get an indication of the time needed to solve the different tasks. Based on this timing information and the tasks from the ACT data set we generated the following three conditions.

1) *Repetitive condition.* For the *repetitive condition* we created random variants (by substituting the numerical values in the task) of two easy tasks from the ACT data set (difficulty = 0.76 and 0.83). The time available to solve each task was set between 60 and 75 seconds at random. This provided participants with more than sufficient time to come up with a solution for each task. Correctly solving a task in the *repetitive condition* granted only a minor monetary reward (+CHF 0.2) and a minor penalty (-CHF 0.2) for incorrect solutions. The *repetitive condition* was designed to trigger emotions such as boredom and fatigue.

2) *Challenge condition.* For the *challenge condition* we selected math tasks from the ACT data set with medium difficulty (difficulty $\in [0.58, 0.69]$) and provided participants with a larger monetary reward (+CHF 2) for correct solutions and the same small penalty as the *repetitive condition* (-CHF 0.2) for incorrect solutions. Participants were provided with sufficient time to solve the tasks based on data from the pilot study (min = 53 seconds, max = 93 seconds). The *challenge condition* was designed to provide diversified tasks for a more engaging and interesting experience, while the larger monetary reward provided a bigger incentive (higher-stakes) for participants to perform well with relatively small penalty in case of mistakes.

3) *Overchallenge condition.* For the *overchallenge condition*, we selected the math tasks with high difficulty in the

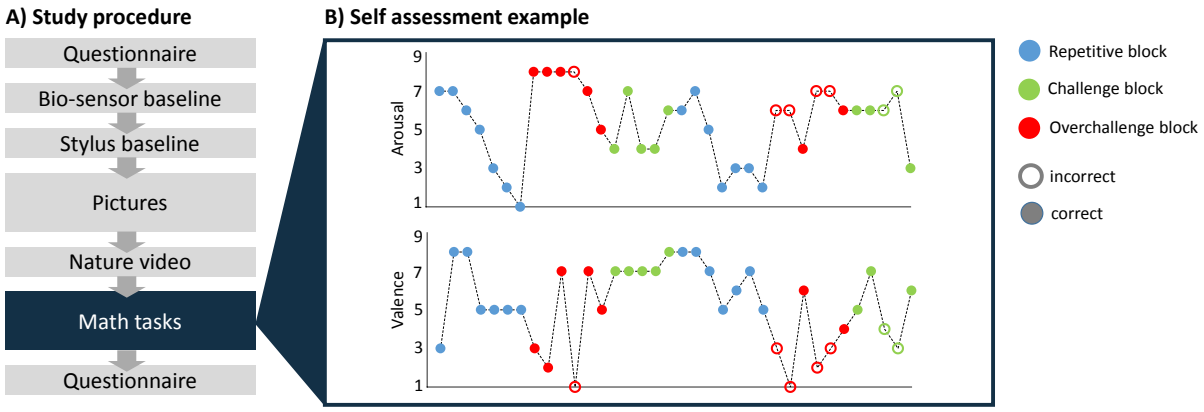

Figure 3: Overview over the different parts of the study. A) Overall experimental procedure. B) Changes in valence and arousal for one participant in relation to task type and answer.

ACT data set (difficulty $\in [0.25, 0.53]$). Participants received small monetary rewards for correct solutions (+CHF 0.2) and a large penalty (-CHF 2) for incorrect solutions. The time to solve each task was set to be insufficient for most participants based on data from the pilot study (min = 25 seconds, max = 51 seconds). The *overchallenge condition* was designed to provide a frustrating and annoying experience to participants.

The math tasks were presented in six blocks (2 in each condition) each containing a different number of tasks (*repetitive condition* 13 tasks, *challenge condition* 5 tasks, *overchallenge condition* 6 tasks). A similar block design for math tasks was already applied in previous work [32]. Moreover, we believe that a sequence of tasks is necessary to trigger an affective state. The first 3 blocks presented were randomly sampled. However, the succeeding 3 blocks were fixed to the same order as the first 3 blocks (but contained different tasks). In addition, the maximum time for each block was limited to 5 minutes to ensure that the math part of the experiment does not go over 30 minutes. After each block, a fixation cross was shown for 30 seconds to reduce potential carry-over effects of affective states. At the end of each math task, participants were asked to fill in the 9-point SAM scale to report their current valence and arousal level (in total, we have collected 3026 ratings from the participants). Figure 3B depicts the changes in the valence and arousal ratings for one participant in relation to the block type and task answer (correct vs. incorrect). We see that for the repetitive tasks, valence and arousal are decreasing over time leading to a shift towards boredom. Additionally, for incorrectly solved tasks, valence drops and arousal tends to increase. After the repetitive blocks we see a decrease in valence and an immediate steep increase in arousal that may be attributed to the increase in difficulty from the repetitive block to the overchallenge block. On average participants finished with CHF 44.3 (min = CHF 22.2, max = CHF 62.8). At the end of the experiment, each participant was compensated with a minimum of CHF 40.

Math task interface. Participants were asked to provide a solution path for every task anywhere on the screen and then to select their answers from 5 multiple-choice alternatives (see Figure 2B). Participants received immediate feedback on whether their answer was correct. A timer located on the top right corner of the interface informed participants about the time left to respond and started to blink when less than 10 seconds remained. When the time was up and the participant did not submit a solution, the answer was considered wrong. The cumulative amount of money earned was displayed on the top left of the interface.

5. RESULTS

We compared different versions of our classification pipeline using only a subset of the sensors with a focus on the difference between stylus and bio-sensors. All results are based on Random Forest (using 500 trees, balanced class weights and hyperparameter optimization using randomized search with 100 iterations) given that this was the best performing classifier. In order to measure the performance of our classifiers, we have used accuracy (chance level = 1/# classes) and micro-averaged area under curve (AUC) of the receiver operating characteristic (ROC) curve (chance level = 0.5), which aggregates the contributions of all classes to compute the average metric. Because both metrics are affected by class imbalance, we have also considered the macro-averaged AUC (chance level = 0.5) which is the average of the class-wise AUCs giving each class the same weight. To derive the SD for each metric, we employed an additional 10-fold cross-validation.

5.1 Study Validation

Our study was designed to trigger affective states across the entire valence-arousal space. As a first step, we investigated if our study design worked by examining if the different parameters acted as intended. In our task design we varied task difficulty, monetary reward and the available time for task completion. We have performed a per task Kendall's tau correlation analysis between these 3 parameters and the arousal and valence ratings of the participants. For the task difficulty and the percentage of remaining time, we have found high correlations for both valence (-0.2; p-value $< 10^{-59}$ and 0.22; p-value $< 10^{-80}$) and arousal (0.27; p-value $< 10^{-102}$ and -0.27; p-value $< 10^{-117}$). Participants shifted towards frustration (decreasing valence and increasing arousal) with increasing task difficulty or with a reduction in the time remaining to complete the task. In-

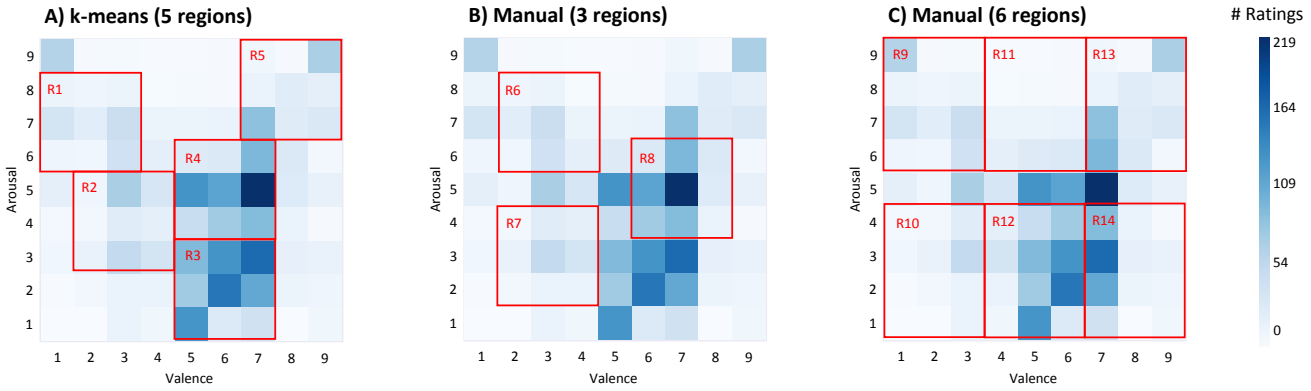

Figure 4: Heat maps showing the distribution of the participants' ratings on the math tasks. The red rectangles represent the different regions. A) 5 regions automatically chosen using k-means clustering. B) Three regions manually selected. C) Six regions manually selected.

terestingly, the effect size on valence and arousal is almost identical. In contrast, monetary reward appears to have a much larger effect on valence (0.47; p-value $< 10^{-295}$) than on arousal (-0.06; p-value $< 10^{-4}$). Altogether, it appears that our tasks worked as intended. Accounting for potential superficial correlations (e.g., task duration) is an important part of our study design. We found a significant Kendall's tau correlation between the task duration and the user ratings of 0.17 (p-value $< 10^{-48}$) and -0.11 (p-value $< 10^{-22}$) for arousal and valence, respectively. Because we have extracted the stylus features over the whole tasks, we have excluded all features having a significant Spearman correlation to the task duration (features greyed out in Table 1).

5.2 Data Analysis

Input Signals. Given that we detected a very low amount of artifacts across participants (EDA = 0.015 % and IBI = 0.71 %), we refrained from removing them from the analysis. Visual inspection of the ST recordings revealed a slow linear increase of the temperature over the course of a participant's session. This change in temperature may be due to the skin warming up under the wristband and independent of the affective state of the participants. We removed this linear trend from all measurements by subtracting the result of a linear least-squares fit to the signal. We did not observe any other artifacts for ST. The bio-sensor features listed in Table 1 have been computed using a window of 10 seconds since the minimum task duration was 10 seconds. For the stylus features, we have used an implicit window over the entire task. In addition, we have excluded all data points having at least one missing value.

Clustering of Ratings. Figure 4 presents the distribution of the participants' ratings in the valence-arousal space (dark and light blue refers to a high and low number of data points, respectively). A v-shape is visible with most ratings being made at a valence and arousal level of 7 and 5, corresponding to a positive medium intense state (e.g., interest). Several ratings were made at the extremes (top left and top right) of the valence-arousal space corresponding to states of distress and excitement that are associated with very good and very poor performance. To uncover the underlying clusters in the data, we have applied k-means clustering in this two-dimensional valence and arousal space. Using the Bayesian information criterion, we found an optimal number of 5 clusters. We defined region boundaries (shown by the red rectangles in Figure 4A) as the arithmetically rounded value of the centroid of each cluster plus and minus the standard deviation of the participants' ratings in the corresponding cluster. We observed that the regions are all of equal size and cover the area of the v-shape. Based on Russell's [29] and Scherer's [33] categorization we identify the following regions, their sizes and corresponding affective states: Region R1 (213 data points; frustrated, annoyed), region R2 (284; bored, taken aback), region R3 (965; attentive, serious), region R4 (861; expectant, confident), region R5 (295; excited, triumphant). Together, it appears that the math task covered a broad range of affective states relevant for learning and that positive states (R3, R4, R5) dominate.

5.3 Classification Performance

Figure 5A and Table 2 present the predictive performance of the model based on the 5 defined regions. Using all sensors, the model achieved an accuracy of 65 % (chance level = 20 %). Here, the slightly lower value for the macro-averaged AUC (0.83) compared to the micro-averaged AUC (0.88) may be related to class imbalance. Figure 5C depicts the confusion matrix based on all sensors. The matrix shows that regions R1 and R2 are more difficult to predict than the other regions. This may be due to the lower number of data points collected for these regions. As expected, the larger the distance between the regions, the easier it is for the model to discriminate between them.

Feature Importance. Table 1 presents the 10 most important features (in bold). The features are sorted according to their relative importance which we computed using permutation feature importance (permuting each feature 100 times and measuring the mean decrease in micro-averaged AUC). We obtained the same relative feature importance ordering using the Gini importance measure. EDA and heart measures provided 3 out of the 10 most important features and stylus features contributed with 4 of the most important features. There were no ST features among the top ten features. Regarding the heart measures, the features related to IBIs were more important than HR features. An

interesting observation can be made for the stylus features. Features related to the distance between strokes appear to be more important than speed between stroke features indicating that the spread of writing attributed to thinking (i.e., how the writing space is covered) provides more information than the actual writing behaviour.

5.4 Sensor Comparison

Bio-Sensors. If we consider the individual sensors (Figure 5B), ST performs substantially worse (−0.11 AUC) compared to EDA (0.80 AUC) and heart rate measures (0.81 AUC). The combination of all the bio-sensors (Figure 5A) provides only marginal performance improvements (+0.05 AUC) compared to the individual sensors.

Stylus. Our most important finding is that the stylus performs equally well as the bio-sensors (Figure 5B), rendering the data from the bio-sensors redundant and unnecessary for the prediction of affective states. The performance of the stylus is only marginally inferior (−0.02 AUC) when compared to the combination of all bio-sensors. In contrast, the combination of the bio-sensors and the stylus achieves a slightly higher performance (+0.02 AUC) compared with the bio-sensors and stylus alone (Figure 5A). This might be an indication that they may contain complementary information, although the difference appears to be small.

5.5 Affective Region Analysis

In order to investigate the ability of our pipeline to predict different affective regions based on the recorded bio-sensor and stylus data we have defined two additional coverings of the valence and arousal space (Figure 4B and 4C). Based on Russell [29] and Scherer [33] we have manually defined specific regions associated with frustration (annoying; region R6, 185 data points), boredom (taken aback; region R7, 199) and interest (engaged concentration, flow; region R8, 720) as shown in Figure 4B. Being able to distinguish these 3 regions is important in education due to their impact on learning gain [9, 26, 10]. To cover the valence and arousal space evenly, we have manually defined the 6 regions shown in Figure 4C, dividing arousal in two and valence in 3 components (The number of data points from region R9 to R14 are 287, 154, 134, 852, 432 and 506). The results for both space partitionings are listed in Table 2 (note that chance level for the accuracy is 33 % for 3 regions and 16.66 % for 6 regions). The performance of the classification of 3 regions outperforms the one for 5 and 6 regions in terms of accuracy. On the other hand, when taking into account the AUC, there is no substantial difference in performance between the different coverings. This difference between accuracy and AUC stems from the fact that predicting only 3 regions is a much easier task than predicting 5 or 6 regions. This is in line with the finding that the accuracy for predicting 5 regions is slightly higher than for 6 regions. Nevertheless, we can conclude that we have seen that our approach is able to provide good results for 3 different coverings. Thus, we come to the conclusion that our pipeline is rather flexible being able to handle different regions in the valence-arousal space. Compared to previous work relying on fixed affective states, our approach has the advantage that the regions do not have to be pre-defined allowing for much more flexible use.

Table 2: Performance of Random Forest on the math data for different signals and regions. $\mathbf{AUC_{micro}}$ and $\mathbf{AUC_{macro}}$ represent micro-averaged and macro-averaged AUC, respectively. The chance level for accuracy is 1/# regions and for AUC it is 0.5. The standard deviations are given in brackets.

Regions	Signals	$\mathbf{AUC_{micro}}$	$\mathbf{AUC_{macro}}$	Accuracy
k-means	EDA	0.80 (0.02)	0.75 (0.03)	50 % (4 %)
(5 Regions)	Heart	0.81 (0.01)	0.73 (0.01)	52 % (2 %)
	Temperature	0.69 (0.03)	0.59 (0.03)	37 % (4 %)
	Stylus	0.84 (0.01)	0.76 (0.02)	59 % (2 %)
	Bio-Sensors	0.86 (0.01)	0.81 (0.02)	60 % (2 %)
	Bio-Sensors & Stylus	0.88 (0.01)	0.83 (0.02)	64 % (2 %)
Manual	EDA	0.81 (0.02)	0.69 (0.04)	66 % (2 %)
(3 Regions)	Heart	0.79 (0.02)	0.66 (0.03)	62 % (3 %)
	Temperature	0.76 (0.01)	0.60 (0.04)	60 % (3 %)
	Stylus	0.83 (0.02)	0.72 (0.02)	67 % (3 %)
	Bio-Sensors	0.84 (0.01)	0.76 (0.03)	67 % (1 %)
	Bio-Sensors & Stylus	0.87 (0.01)	0.80 (0.02)	67 % (2 %)
Manual	EDA	0.80 (0.02)	0.72 (0.03)	46 % (3 %)
(6 Regions)	Heart	0.78 (0.01)	0.72 (0.02)	44 % (2 %)
	Temperature	0.70 (0.02)	0.61 (0.02)	35 % (3 %)
	Stylus	0.81 (0.01)	0.75 (0.02)	48 % (2 %)
	Bio-Sensors	0.85 (0.02)	0.80 (0.02)	57 % (4 %)
	Bio-Sensors & Stylus	0.87 (0.02)	0.83 (0.03)	61 % (3 %)

5.6 Model Transfer

In addition to the math tasks, we have also gathered bio-sensor data as well as valence and arousal ratings from the participants while they observed pictures from the IAPS. We have used this data to investigate our model's capacity to generalize to more passive tasks, such as looking at pictures. To predict the affective regions of interest, we applied our model trained on the bio-sensor data recorded during math task solving to data collected while participants viewed and rated the set of IAPS pictures. When we consider the 5 different regions (Figure 4A), the model's accuracy reaches 39 % (chance level = 20 %, $\mathrm{AUC_{micro}} = 0.68$, $\mathrm{AUC_{macro}} = 0.64$). When we train and evaluate a model directly on the picture data, we achieve a slightly better classification performance (accuracy = 42 %, $\mathrm{AUC_{micro}} = 0.74$, $\mathrm{AUC_{macro}} = 0.66$). There may be several reasons behind the suboptimal performance when predicting affective states during the picture task. These include sociocultural aspects when rating emotions based on pictures (e.g., rating how it is expected), old and low resolution pictures from the IAPS data set, media influence desensitizing participants to the content of the IAPS, and the fact that the math and picture domains are very different. Together these initial results indicate that building a general predictor of affective states might be possible, but further experiments are necessary.

6. CONCLUSION

In this paper we presented a generic pipeline for predicting affective regions of interest using bio-sensor and stylus data. We validated our pipeline for the case of math solving tasks and demonstrated that our pipeline can accurately predict various regions in the valence-arousal space (up to 0.88 AUC). In addition, we have compared different input signals with each other. The performance of the Shimmer GSR+ and Polar H10 have been on the same level (up to 0.81 AUC). Moreover, we found that the classification performance using only stylus data is comparable to the classification performance based on the bio-sensors. Taking into account the emerging digitization of education and the spread of tablets in schools and private households, these results

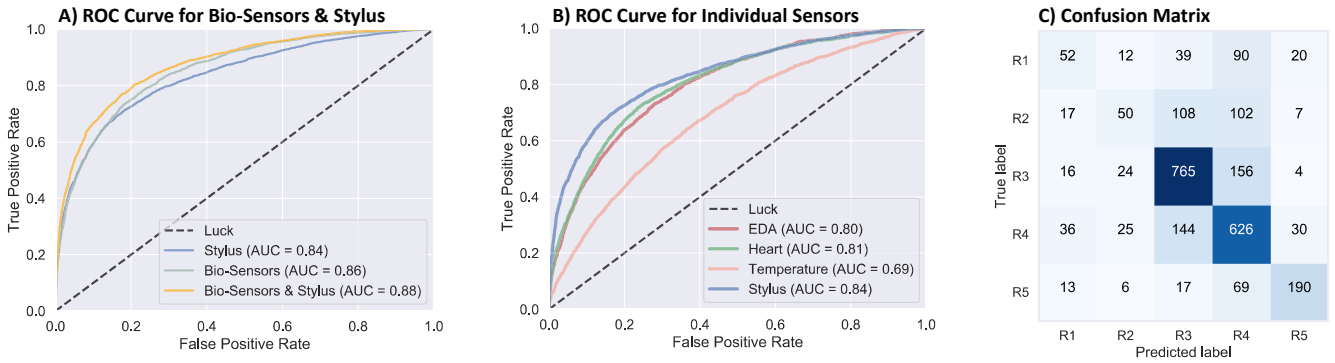

Figure 5: ROC curves and micro-averaged AUC scores for 5 regions chosen by k-means clustering for (A) the bio-sensors, stylus and the combination of bio sensors and stylus and (B) the individual bio-sensors & stylus. (C) The confusion matrix is computed by using the combination of bio-sensors and stylus.

make the stylus a preferred alternative to bio-sensors for measuring affective states in classrooms. Using bio-sensors in classroom settings can be cumbersome and costly as it requires the purchase and synchronization of several devices. In contrast, systems that depend on a stylus only are cheaper than systems relying on bio-sensor devices, and styluses often come bundled with mobile devices, such as tablets or smartphones. In addition to being cheaper and more ubiquitous, styluses are easier to setup (e.g., no attachment of electrodes, no motion artifacts) and less intrusive. Furthermore, stylus data is not only restricted to digital devices but can also be recorded using digital pens. Finally, we have demonstrated the possibility of a generalized model for predicting affective states by applying the model trained on the data from the math tasks (active part) to pictures from the IAPS (passive part) reaching a performance of 0.68 AUC.

There are some potential limitations to the approach presented here. First of all, the setup is restricted to a lab environment and the population of Bachelor students may limit generalization to students at other levels. We are optimistic that our approach also works outside a controlled setting and for a broader population. Participants reported that the setup was comfortable and that they could act in a natural way. In addition, we assume that given a proper baseline correction the signals are also predictive for a heterogeneous group of people. Another limitation is the restriction to math tasks. Similar to bio-sensor data, we believe that handwriting data carries affective information independent of the task. Thus, we expect our approach to work also in other domains involving handwriting, such as solving exercises for different school subjects and writing essays.

Future research from our lab will test and refine our pipeline for multiple domains. Potential refinements include using non-linear IBI features and frequency features for skin temperature. Additionally, an in depth analysis of handwriting that takes into account the slant of the handwriting could further improve the classification performance. Another interesting direction would be to make use of large existing bio-sensor databases for semi-supervised learning by using auto-encoders to infer an efficient feature embedding [20].

7. REFERENCES

[1] ACT. The act technical manual, 2017.

[2] M. Benedek and C. Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010.

[3] A. Betella, R. Zucca, R. Cetnarski, A. Greco, A. Lanatà, D. Mazzei, A. Tognetti, X. D. Arsiwalla, P. Omedas, D. De Rossi, et al. Inference of human affective states from psychophysiological measurements extracted under ecologically valid conditions. *Frontiers in neuroscience*, 8:286, 2014.

[4] N. Blanchard, R. Bixler, T. Joyce, and S. D'Mello. Automated physiological-based detection of mind wandering during learning. In *Proc. ITS*, pages 55–60. Springer, 2014.

[5] W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. roth, M. E. Dawson, and D. L. Filion. Publication recommendations for electrodermal measurements. *Psychophysiology*, pages 1017–34, 2012.

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

[7] R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas. *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[8] C. Conati and H. Maclaren. Modeling user affect from causes and effects. In *Proc UMAP*, pages 4–15. Springer, 2009.

[9] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, NY, 2008.

[10] R. S. J. d Baker, S. M. Gowda, M. Wixon, J. Kalka, A. Z. Wagner, A. Salvi, V. Aleven, G. W. Kusbit, J. Ocumpaugh, and L. Rossi. Towards sensor-free affect detection in cognitive tutor algebra. *International Educational Data Mining Society*, 2012.

[11] M. Ez-Zaouia and E. Lavoué. Emoda: a tutor oriented multimodal and contextual emotional dashboard. In *Proc. LAK*, pages 429–438. ACM, 2017.

[12] M. Fairhurst, M. Erbilek, and C. Li. Study of automatic prediction of emotion from handwriting samples. *IET Biometrics*, pages 90–97, 2015.

[13] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*, pages 402–413. ACM, 2014.

[14] B. Grawemeyer, M. Mavrikis, W. Holmes, S. Gutierrez-Santos, M. Wiedmann, and N. Rummel. Affecting off-task behaviour: how affect-aware feedback can improve student learning. In *Proc. LAK*, pages 104–113. ACM, 2016.

[15] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Trans. on Biomedical Engineering*, 63(4):797–804, 2016.

[16] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar. cstress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proc. of the international joint conference on pervasive and ubiquitous computing*, pages 493–504. ACM, 2015.

[17] I. Jraidi, M. Chaouachi, and C. Frasson. A hierarchical probabilistic framework for recognizing learners' interaction experience trends and emotions. *Advances in Human-Computer Interaction*, 2014:6, 2014.

[18] T. Käser, G.-M. Baschera, A. G. Busetto, S. Klingler, B. Solenthaler, J. M. Buhmann, and M. Gross. Towards a framework for modelling engagement dynamics in multiple learning domains. *International journal of artificial intelligence in education*, 22(1-2):59–83, 2013.

[19] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.

[20] S. Klingler, R. Wampfler, T. Käser, B. Solenthaler, and M. Gross. Efficient feature embeddings for student classification with variational auto-encoders. In *Proceedings of EDM*, 2017.

[21] V. Kostyuk, M. V. Almeda, and R. S. Baker. Correlating affect and behavior in reasoning mind with state test achievement. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 26–30. ACM, 2018.

[22] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

[23] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 2008.

[24] L. Likforman-Sulem, A. Esposito, M. Faundez-Zanuy, S. Clémençon, and G. Cordasco. Emothaw: A novel database for emotional state recognition from handwriting and drawing. *IEEE Transactions on Human-Machine Systems*, 47(2):273–284, 2017.

[25] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European heart journal*, 17(3):354–381, 1996.

[26] M. Miserandino. Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of educational psychology*, 88(2):203, 1996.

[27] Z. A. Pardos, R. S. J. D. Baker, M. O. C. Z. San Pedro, S. M. Gowda, and S. M. Gowda. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proc. LAK*, pages 117–124. ACM, 2013.

[28] D. J. Plews, B. Scott, M. Altini, M. Wood, A. E. Kilding, and P. B. Laursen. Comparison of heart-rate-variability recording with smartphone photoplethysmography, polar h7 chest strap, and electrocardiography. *International journal of sports physiology and performance*, 12(10):1324–1328, 2017.

[29] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[30] S. Salmeron-Majadas, M. Arevalillo-Herráez, O. C. Santos, M. Saneiro, R. Cabestrero, P. Quirós, D. Arnau, and J. G. Boticario. Filtering of spontaneous and low intensity emotions in educational contexts. In *Proc. AIED*, pages 429–438. Springer, 2015.

[31] S. Salmeron-Majadas, R. S. Baker, O. C. Santos, and J. G. Boticario. A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios. *IEEE Access*, 6:39154–39179, 2018.

[32] M. Saneiro, O. C. Santos, S. Salmeron-Majadas, and J. G. Boticario. Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. *The Scientific World Journal*, 2014, 2014.

[33] K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[34] F. Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258, 2017.

[35] Y. Shi, M. H. Nguyen, P. Blitz, B. French, S. Fisk, F. De la Torre, A. Smailagic, D. P. Siewiorek, M. al'Absi, E. Ertin, et al. Personalized stress detection from physiological measurements. In *International symposium on quality of life technology*, pages 28–29, 2010.

[36] C.-S. Yoo and S.-H. Yi. Effects of detrending for analysis of heart rate variability and applications to the estimation of depth of anesthesia. *Journal of Korean Physical Society*, 44:561, 2004.

[37] J. Zhou, K. Hang, S. Oviatt, K. Yu, and F. Chen. Combining empirical and machine learning techniques to predict math expertise using pen signal features. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 29–36. ACM, 2014.

[38] M. Züger and T. Fritz. Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proc. of the Conference on Human Factors in Computing Systems*, pages 2981–2990. ACM, 2015.