British Journal of Educational Technology *Vol 50 No 6 2019* 3004–30
doi:10.1111/bjet.12854

Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach

Kshitij Sharma, Zacharoula Papamitsiou and Michail Giannakos

Kshitij Sharma is a senior researcher at the Computer Science department in the Norwegian University of Science and Technology. He received his Ph.D. in Computer Science from the Ecole Polytechnique Federale de Lausanne (EPFL, Switzerland). He has also worked as a Postdoctoral researcher in EPFL and in the University of Lausanne, Switzerland for 2 years before moving to Norway. His research interests include eye-tracking, MOOCs, collaborative learning, applied machine learning, multimodal learning and statistics. Zacharoula Papamitsiou is a senior researcher at the Computer Science Department of the Norwegian University of Science and Technology. She holds a Ph.D. from the University of Macedonia. Her domain of expertise is on adapting and personalizing learning services and on supporting learners' decision making using Learning Analytics. Michail Giannakos is an associate professor of interaction design and learning technologies at the Department of Computer Science of NTNU, and Research Director of the Center for Excellent IT Education (Excited). Giannakos has coauthored more than 100 manuscripts published in peer-reviewed journals and conferences (Computers & Education, BJET, ACM TOCE, CSCL, ICALT to mention a few). He has worked at several research projects funded by diverse sources like EC, Microsoft Research, The Research Council of Norway (RCN), US-NSF, German agency for international academic cooperation (DAAD) and Cheng Endowment. Address for correspondence: Kshitij Sharma, Department of Computer Science, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway. Email: kshitij.sharma@ntnu.no

Abstract

Students' on-task engagement during adaptive learning activities has a significant effect on their performance, and at the same time, how these activities influence students' behavior is reflected in their effort exertion. Capturing and explaining effortful (or effortless) behavior and aligning it with learning performance within contemporary adaptive learning environments, holds the promise to timely provide proactive and actionable feedback to students. Using sophisticated machine learning (ML) algorithms and rich learner data, facilitates inference-making about several behavioral aspects (including effortful behavior) and about predicting learning performance, in any learning context. Researchers have been using ML methods in a "black-box" approach, ie, as a tool where the input data is the learner data and the output is a given class from the chosen construct. This work proposes a methodological shift from the "black-box" approach to a "grey-box" approach that bridges the hypothesis/literature-driven (feature extraction) "white-box" approach with the computation/data-driven (feature fusion) "black-box" approach. This will allow us to utilize data features that are educationally and contextually meaningful. This paper aims to extend current methodological paradigms, and puts into practice the proposed approach in an adaptive self-assessment case study taking advantage of new, cutting-edge, interdisciplinary work on building pipelines for educational data, using innovative tools and techniques.

Introduction

Learning performance is strongly associated with learners' perception of task difficulty and on-task mental effort (Papamitsiou & Economides, 2015; Yen, Chen, Lai, Su, & Chuang, 2015) among others (eg, affective states, self-regulation behavior, perceived self-efficacy and expertise).

Practitioner Notes

What is already known about this topic

- Capturing and measuring learners' engagement and behavior using physiological data has been explored during the last years and exhibits great potential.
- Effortless behavioral patterns commonly exhibited by learners, such as "cheating," "guessing" or "gaming the system" counterfeit the learning outcome.
- Multimodal data can accurately predict learning engagement, performance and processes.

What this paper adds

- Generalizes a methodology for building machine learning pipelines for multimodal educational data, using a modularized approach, namely the "grey-box" approach.
- Showcases that fusion of eye-tracking, facial expressions and arousal data provide the best prediction of effort and performance in adaptive learning settings.
- Highlights the importance of fusing data from different channels to obtain the most suited combinations from the different multimodal data streams, to predict and explain effort and performance in terms of pervasiveness, mobility and ubiquity.

Implications for practice and/or policy

- Learning analytics researchers shall be able to use an innovative methodological approach, namely the "grey-box," to build machine learning pipelines from multimodal data, taking advantage of artificial intelligence capabilities in any educational context.
- Learning design professionals shall have the opportunity to fuse specific features of the multimodal data to drive the interpretation of learning outcomes in terms of physiological learner states.
- The constraints from the educational contexts (eg, ubiquity, low-cost) shall be catered using the modularized gray-box approach, which can also be used with standalone data sources.

Students' on-task mental effort is an important factor of their educational outcomes, such as their persistence in learning (eg, Jung & Lee, 2018) and their academic achievement (eg, Chen, 2017; Pardo, Han, & Ellis, 2017). According to Humphreys and Revelle (1984), effort is "the motivational state commonly understood to mean trying hard or being involved in a task. Effort is increased when the subject tries harder, when there are incentives to perform well, or when the task is important or difficult." In this study, the terms "effortful behaviour" and "engagement" are used interchangeably, and they refer to learners' conscious, intrinsically motivated and active involvement with the learning tasks.

Although engagement and active involvement with the learning activities lead to better educational outcomes, "true performance" is often overshadowed by effortless behavior commonly exhibited by students: "cheating," "guessing" or "gaming the system" behavioral patterns counterfeit these outcomes (eg, Baker, Corbett, Koedinger, & Wagner, 2004; Wise & Kong, 2005). In traditional classrooms, before instructors provide a set of tasks to their students, they need to be aware of the students' comprehensions, their ability level, as well as an estimation of effort needed to successfully accomplish those tasks, to prevent students from engaging in effortless behavior. Similarly, contemporary intelligent tutoring and adaptive learning systems automatically identify students' ability level (eg, estimate students' knowledge mastery from performance

indices) and select the most appropriate tasks to deliver to students accordingly, considering the required effort for those tasks (eg, the probabilities to guess or slip the solution, Baker, Corbett, Aleven, 2008; Gowda, Rowe, Baker, & Chi, 2011; Pelánek, 2016).

Motivation of the research and research question

Students' on-task effort exertion during adaptive learning activities is an important factor that affects their performance. Therefore, deeper understanding, explaining and predicting effortful behavior is expected to shed light to more complex learning mechanisms in adaptive learning settings.

Existing methods for the prediction and explanation of effortful behavior are usually based on learners' response time patterns and data coming from traditional computer activity logging (Papamitsiou & Economides, 2015, 2016; Wise & Kong, 2005). For example, van Gog, Kirschner, Kester, and Paas (2012) found that repeatedly measuring mental effort (using subjective rating scales and associating the measurements with response times) after performing individual tasks in a series, was favored for tasks that take longer than usual to complete.

Other computational methods operationalize effortful behavior as probabilities (Gowda *et al.*, 2011). In these cases, a guessing parameter is incorporated in learner models to describe the possibility of the learner to respond correctly in a generally random fashion (effortless) instead of actively seeking to determine the correct answers (effortful). For example, Backer *et al.* (2008) make contextual estimations of the probability for a student to have guessed or slipped.

Furthermore, more sophisticated measurements have also been employed for coding effortful interactions, focusing mostly on learners' engagement: the idea to employ effort-related multimodal physiological measures in the operationalization of student engagement is not new (D'Mello, Craig, & Graesser, 2009; Gilzenrat, Cohen, Rajkowski, & Aston-Jones, 2003; Mulder, 1986). Multimodal data provide educational technology researchers with an unprecedented opportunity to gain insights into and deeper understand learners' actions in diverse learning contexts (eg, D'Mello *et al.*, 2009; Furuichi & Worsley, 2018). For example, pupil dilation was found to be highly correlated with engagement, heart rate (HR) variability and cognitive load have been acknowledged to reflect self-regulatory capacity, whereas facial features have been extensively used for emotion recognition that are related to deeper learning (D'Mello *et al.*, 2009; Gilzenrat *et al.*, 2003; Mulder, 1986).

Towards explaining and understanding students' effortful behavior and learning performance, we propose building machine learning (ML) pipelines on multimodal physiological data collected during an adaptive learning activity. The physiological data sources include eye-tracking, electroencephalography, facial features and arousal data (HR, blood volume pressure (BVP), electrodermal activity (EDA) and skin temperature). Various combinations of such data sources have been used in the past to explain (Raca & Dillenbourg, 2014) and/or predict (Beardsley, Hernández-Leo, & Ramirez-Melendez, 2018) learning behaviors (Furuichi & Worsley, 2018) and/or performance (Junokas, Lindgren, Kang, & Morphew, 2018).

In the present study, we propose a shift in methodological paradigm for developing ML pipelines for educational data, and through a case study, we put the proposed approach into practice to predict learning performance and explain how the students achieve high performance by exhibiting effortful behavior. This understanding would enable us to identify appropriate moments, during the learning process, for giving actionable feedback to the students. As such, the research question that guided this work is: ***"What combinations of students' physiological data explain their effortful engagement and learning performance in adaptive learning conditions?"***

To address the research question, this study was contextualized and operationalized in adaptive self-assessment conditions. Self-assessment leads students to a greater self-awareness, by facilitating self-regulation of motivation and actions (McMillan & Hearn, 2008), and inherently promotes students' effortful behavior, because the result of the assessment is primarily important to the student herself (Papamitsiou & Economides, 2019). Furthermore, adaptation has shown a positive impact on learners' engagement with the activities (Normadhi *et al.*, 2019), which is also reflected on learners' performance (Barla *et al.*, 2010; Liu, McKelroy, Corliss, & Carrigan, 2017). In addition, previous work with multimodal data yielded results in other learning settings. However, although each of the individual data streams have their own challenges, their fusion is even more technically difficult and demanding (Ochoa & Worsley, 2016). As such, lack of previous work in adaptive learning contexts along with the challenges in fusing multimodal data, as well as extending or contradicting previous finding from other contexts motivated this study.

Contribution

A core contribution of the presented work derives directly from the study itself, and concerns the fusion of multimodal data and ML methods to predict learners' effortful behavior and performance in adaptive assessment tasks. Most of the recent approaches on measuring effort rely solely on response time patterns and guessing behavior patterns (eg, Chang, Plake, Kramer, & Lien., 2011; Wise, Kuhfeld, & Soland, 2019). Therefore, this study is the first one—to the best of our knowledge—that goes a step ahead from the commonly used clickstream data for effort estimation/predictions, by exploiting non-invasive high-frequency multimodal data.

Furthermore, due to the inherent particularities of multimodal data, the common approach that researchers have been using to analyze them and address the educational objectives they set, is to employ ML methods (eg, Di Mitri *et al.*, 2017; Mattingly *et al.*, 2019). However, although the authors describe the method they use, ML is presented as "black-box," ie, as a tool where the input data are the learner data and the output is a class/value from the chosen construct, without actually inspecting how/why the considered algorithms are accomplishing what they are accomplishing. For example, in a previously proposed framework for modeling learners' behavior and actions in an Intelligent Tutoring System, Conati and Kardan (2013) suggested an ML pipeline consisting of two phases. The pipeline first detects learners' behavioral clusters and then classifies a new learner to one of the predefined clusters, based on their logged actions. In Conati's and Kardan's (2013) framework, the idea was to first relate the clusters' features to the learning outcomes and then isolate in each cluster those behaviors that are responsible for the learning effects. Next, as new users interact with the system, they are classified in real time into one of the clusters generated by the behavior discovery phase. Still, this framework is not using multimodal data, and it is processing the learners' logs in an automated (ie, "black-box") manner.

This paper proposes a shift from the "black-box" approach to a "grey-box" approach, where the input features can be informed from the context and the theory/relevant research, the data fusion is driven by the limitations of the resources and contexts (eg, ubiquitous, low-cost, high precision, different experimental settings), and the ML method is chosen in an informed manner, rather than just as a way to obtain the optimal prediction/classification accuracy. In other words, this contribution aims to invite researchers to *shift from the optimal ends (outputs) to the optimal means (paths)*.

Besides proposing the aforementioned "grey box" approach, the present study puts into practice this approach, in a case study that as an authentic example showcases the whole process of building a ML pipeline. Specifically, this contribution, explicitly streamlines the process from gathering the multimodal data within a learning context—based on the educational constraints

(eg, ubiquitous, low-cost, high precision, different experimental settings)—to fusing and analyzing them, and to predicting a learning construct (ie, performance and effort, in this study), as a generic, "step-by-step" methodology.

Finally, it is important to make it explicitly clear that this contribution in no way claims that the methods used in this paper have not been previously used, or that the paper presents novel ways of handling the physiological data, or that the educational data are interpreted in a different way; the core scope of this contribution is to frame and standardize a commonly used research practice (ie, ML), and to point out the "grey-box" approach for improving upon the ways ML methods can be used with multimodal data in education.

Related work: Utilizing multimodal data to predict learning constructs

Physiological data from multimodal channels have been acknowledged regarding their potential to provide insights to educational technology researchers about learners' states and behaviors (Lane & D'Mello, 2019). In a recent selective review, Lane and D'Mello (2019) summarized how different physiological data (eg, gaze, facial features, fMRI, fNIRS, EMG, EEG) have been used in state-of-the-art approaches to measure learners' attention, focus, cognitive load and various affective states and learning strategies, and what is their capacity to inform and guide the design of cognitive, affective and metacognitive scaffolds.

In the above-mentioned approaches, the physiological data have been typically processed using ML, and were interpreted by considering contextual information, as well. The reason is that physiological measurements lack objective ground truth (D'Mello, Dieterle, & Duckworth, 2017), resulting in weak interpretation when compiled alone into constructs like affect and engagement. This limitation is also highlighted in Di Mitri, Schneider, Specht, and Drachsler (2018). In their study, the authors demonstrated how the ground truth for various "learner labels" was obtained, and they concluded that none of the listed approaches provides an objectively measurable ground truth (Di Mitri *et al.*, 2018). This "missing" contextual information, though, is common practice in educational research to be grounded in and obtained from the knowledge from relevant literature (eg, in formulating hypotheses).

Furthermore, it has been argued that configuring ML methods for the multimodal data that measure specific characteristics of the learner, is an adequate and recommended means (Giannakos, Sharma, Pappas, Kostakos, & Velloso, 2019). Nowadays, both the physiological data collection devices and the ML methods are rapidly being developed into cost-effective, consumer-off-the-shelf products (D'Mello *et al.*, 2017). This could explain the increasing adoption of such approaches in complex and open-ended learning settings (eg, programming, robotics, complex problem solving, Blikstein & Worsley, 2016), and not only in the case of online, in-front-of-the-screen or cognitive tutors.

One of the most prominent uses of multimodal data in combination with ML is affect prediction (Bosch, D'mello, Ocumpaugh, Baker, & Shute, 2016; D'Mello, Bosch, & Chen, 2018; Mattingly *et al.*, 2019). For example, Mattingly *et al.* (2019) predicted affective states (performance, intelligence, personality, mood, anxiety, health measures, exercise, sleep and stress) using the data collected from the physical activity and phone logs. D'Mello *et al.* (2018) also focused on predicting the affective states of their participants, using audio, facial expressions, HR, EDA, temperature and ECG. Furthermore, Bosch *et al.* (2016) predicted students' affective states while using an online physics tutor using facial features and movement data.

Taking advantage of technological advancements in "big data" capturing and processing, many studies carried out in educational settings have used multimodal data and focused either on

measuring learner engagement in individual/collaborative conditions (eg, Andrade, Delandshere, & Danish, 2016; Worsley & Blikstein, 2014), or on predicting individual learner performance in more diverse set-ups (eg, Spikol, Ruffaldi, Dabisias, & Cukurova, 2018).

Apparently, using physiological data to capture and measure learners' active on-task involvement and effortful behavior is not new (eg, D'Mello *et al.*, 2009; Kalsbeek & Ettema, 1963). For instance, pupillary response has a long association with the measurement of mental effort in response to cognitive variables. Marshall (2002) attempted to quantify small discontinuities in pupil size that are related to cognitive activity. Gaze has long been studied as an approach for understanding users' behaviors and cognitive states (Lai *et al.*, 2013). Moreover, Fairclough, Moores, Ewing, and Roberts (2009) found that electroencephalography (EEG) variables were sensitive to disengagement due to cognitive load. Furthermore, effort-related cardiovascular responses can be mapped to success importance until a maximum effort has been achieved (Wright & Kirby, 2001).

In particular, regarding engagement in individual learning, Andrade *et al.* (2016) used Multimodal Learning Analytics (MMLA) to automatically detect the moments when students' expectations are likely to influence their engagement with the knowledge ("epistemological frames"). The authors used speech, posture and gaze to model such moments, in order to understand the depth of students' engagement with the content, but they could not verify a direct relationship between the behavioral patterns in the multimodal data and "epistemological frames." However, in another study, Worsley and Blikstein (2014) verified this relationship. Specifically, the students collaborated in pairs to complete an engineering design task, and the authors used hand/wrist movement, electro-dermal activation, and voice activity detection, for modeling how students engage with the task, in terms of the reasoning strategies the used. Furthermore, in a face to face classroom setting, Pijeira-Díaz, Drachsler, Kirschner, and Järvelä (2018) utilized the EDA, Galvanic Skin Conductance, temperature and the accelerometer data, to measure simultaneous arousal levels among the students with respect to the students' mood, motivation, affect and collaborative engagement. Results shown that low arousal was the predominant state, whereas all students were never in high arousal states in the classroom, at the same moment. In the same context, Raca and Dillenbourg (2014) used the synchronization of students' gaze direction and body postures for predicting their self-reported attention. Attention has been found to be a strong construct of engagement (Kinnealey *et al.*, 2012; Mundy, Acra, Marshall, & Fox, 2006). The results showed that students with lower levels of attention were slower in reacting to the teacher than the focused students.

Furthermore, considerable amount of research has also been conducted to predict learning performance in diverse learning tasks, using multimodal data. Specifically, researchers have used EEG and behavioral data (eg, reaction time from clickstreams, Beardsley *et al.*, 2018) to predict students' recall, or gestures, postures and body movements to predict students' performance in repeating, recalling and association tasks (Junokas *et al.*, 2018). In a project-based learning case, Spikol *et al.* (2018) used objects created by the students in their respective projects, in combination with students' positions, hand gestures, facial expressions, audio, video and interaction patterns with the physical computing platform, aiming to predict the quality and correctness of the solution. Other researchers aimed to model learners' performance using either audio, video and the log data from a chemistry educational tutor (Liu *et al.*, 2019), or HR, BVP and other physiological data sources in self-regulated learning activities (Di Mitri *et al.*, 2017). While learners were solving mathematical problems, Smith, King, and Gonzalez (2016) recorded Kinect sensor data and dialogues and used these data sources to explain students' performance in terms of interaction patterns. In all studies, the prediction of performance achieved was highly accurate.

Except from predicting affective states and performance and explaining engagement, other studies employed multimodal data for other research objectives, as well, including modeling dialogue

acts (Ezen-Can, Grafsgaard, Lester, & Boyer, 2015; Worsley, 2018), idea creation (Furuichi & Worsley, 2018) or motivational intentions (Yu *et al.*, 2018), assessing presentation skills (Chen *et al.*, 2016; Ochoa *et al.*, 2018) and predicting collaborative coordination/synchrony between the collaborating peers (eg, Grafsgaard, Duran, Randall, Tao, & D'Mello, 2018; Schneider & Blikstein, 2015; Stewart, Keirn, & D'Mello, 2018; Worsley, 2014). Furthermore, substantial work has been done in the area of providing feedback using data in one or more modalities (Pardo, Poquet, Martínez-Maldonado, & Dawson, 2017). For example, EEG data and responses/clickstream data have been exploited for understanding affectional reactions to feedback (eg, Cabestrero *et al.*, 2018; Luft, Nolte, & Bhattacharya, 2013). In a slightly different context, Andrade (2017) used multimodal data (motion and gaze) to show how students' explanations of feedback loops differ while controlling an embodied simulation. Finally, Moridis and Economides (2012) provided effective feedback using Embodied Conversational Agents based on emotional facial expression and speech. The exhaustive list of multimodal data applications are beyond the scope of the work presented in this paper.

The manual analysis of learners' interactions (with a computer or peers) is a complicated and sometimes tedious process. Among the primary motivations of these studies was to reduce this human workload, and to select appropriate multimodal data for capturing learners' behaviors (Andrade *et al.*, 2016; Ochoa *et al.*, 2018). In line with these works, the long-term motivation of the present study is to fuse (or utilize combinations of) multimodal data sources for explaining learners' behavior in adaptive learning settings and build ML pipelines for facilitating the efficient processing of those data.

In a nutshell (See Table S1), the previous research has employed multimodal data in different classroom/experimental settings, with individual/collaborative tasks, to understand learners' behavior associated with high levels of performance, to explain learners' on-task engagement patterns or to support learners/instructors in an automated manner. In this contribution, we explore the combinations of multimodal data to predict learning performance and effortful behavior in an adaptive self-assessment procedure and explain how to build efficient ML pipelines for the educational multimodal data.

Pipeline from data collection to educationally meaningful interpretations

ML in multimodal educational data

As seen from the review of relevant literature, several channels of learners' physiological data can be recorded during students' participation in learning activities, aiming to help us understand and explain educational constructs (eg, Pijeira-Díaz *et al.*, 2018; Spikol *et al.*, 2018). Those multimodal data are usually collected with specialized equipment and include eye-tracking, EEG, facial expressions, heart-rates, gestures, postures, audio, video, body motions and many more. The information captured in those data is rich enough to describe different aspects and dimensions of the learners' states at any moment during the learning process. Furthermore, these data shall next be fed into an analysis cycle to (a) extract from them the educationally meaningful knowledge associated with human learning mechanisms and (b) answer to educationally critical questions about human learning processes.

However, typical analysis techniques (eg, analysis of variance, correlations, linear regressions) usually considered in educational settings, are employed mostly for hypothesis testing, where hypothesizing is guided from previous work or/and theories. When dealing with multimodal data, for analyzing the high volumes of those data (ie, predictions/classifications), other, more sophisticated techniques need to be configured. To predict the different aspects of learners' states at any moment from multimodal data during the learning process, ML algorithms should suit

the nature of the data: not all algorithms are appropriate for treating all kinds of data, and the purpose of the analysis also needs to be considered.

The studies reviewed in the related work section use the ML methods in combination with the multimodal data sources to predict various learning constructs in a given learning context. However, the description of the ML methods as it is used in these papers, serves only as a part of the research methods and not as an explication of the ML pipeline. To make this statement clear, let us consider the following hypothetical study, conducted in a video-based learning context. Let the learning task be as simple as watching a video (at learner's own pace) and next answer a test of multiple-choice questions, based on the video content. Let also the research objective be the prediction of the test score (performance), based on data collected during the video-watching task. For the needs of the (hypothetical) study, the researchers have recorded clickstream, eye-tracking and EEG data and facial videos. This example is a simplification of a typical study, similar to the ones mentioned in the related work section. The data analysis methods in a typical study like this one, would include features extraction and application of one or more prediction/classification algorithms. The reporting of results would be based on the algorithm yielding the most accurate prediction/classification results.

However, there are a few problems with such studies: (a) the ML pipeline is not explicitly mentioned in the study, limiting the generalizability of the method; (b) the approach is not modular which limits the investigated feature space; (c) there is no investigation based on the feature selection and the prediction/classification algorithms; (d) the ML methods are used in a black-box manner, ie, there is lack of step-by-step understanding of the method; and (e) the methodologies don't systematically allow researchers to introduce contextual, theoretical and background information to the analysis, eg, utilizing data features that are educationally and contextually meaningful.

The "Grey-box" of ML and how to use it

This paper demonstrates a generalized and modular ML pipeline. The present method explicates how a literature/hypothesis-driven "white-box" approach for feature extraction (ie, ground truth contextual knowledge) can be combined with a computation-driven "black-box" approach (ie, ML) for feature-fusion into a "grey-box" approach. In the "*grey-box approach*" the features are extracted based on the theory or relevant literature and the need of the analysis, and then ML does the necessary computational analysis, that follows the boundaries set from the "white-box" part of the approach, avoiding in this way the results that can't be interpreted to inform an instructional decision. In the "grey-box" approach, there is a clearer understanding (than the "black-box") about the combinations of feature selection and prediction/classification algorithms. Moreover, this pipeline can be fine-tuned for the research needs, eg, one can use only facial videos and wristband data in a setting that demands high levels of pervasiveness, while in a controlled lab setting, eye-tracking and EEG can be added to the data sources. Thus, the proposed technique allows us to utilize ML advantages, but at the same time embrace, the important for learning sciences, information about the context, as well as appropriate theoretical grounds and related works.

The primary goal of ML algorithms, in educational contexts, is for the machine (ie, the algorithm) to "learn" from the educational data and use this "lesson" to predict/classify learning events, with sufficient efficiency, accuracy and scalability. ML algorithms have been found to achieve the most accurate predictions, but in most of the cases utilize a "black-box" approach, that makes it difficult to explain relationships or produces educationally meaninglessness results. In this section, we explain this process in simple terms, targeting to showcase that enhancing educational processes with artificial intelligence mechanisms can facilitate human learning and produce educationally meaningful implications.

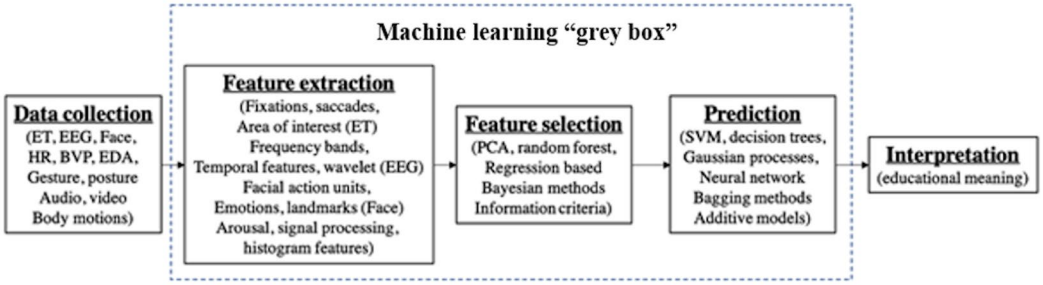

Figure 1: The different modules of a generalized machine learning pipeline [Colour figure can be viewed at wileyonlinelibrary.com]

Specifically, we generalize the process of building ML pipelines, using multiple sources of learners' physiological data. The whole process is synopsized in Figure 1. We use the term "module" to describe every separate phase or step of the pipeline. The modularity of the pipeline reflects how fine grained can be the different phases of the pipeline building process.

As seen in the Figure 1, the ML pipeline consists of three different modules, namely: feature extraction, feature selection and prediction. The first step in this process—after data collection—is to provide the machine with the possible features of the multimodal data (ie, feature extraction step). The feature extraction step is highly dependent on the types and the specifications of the data collections and analyses employed. In other words, the available data and related educational theory guide and determine feature extraction.

However, not all features are equally meaningful to the machine: the feature extraction procedure results in a multitude of features, which might or might not satisfactorily explain an outcome of interest (if the features extraction is inspired by the literature, this might come from the difference in the learning settings). In this case, just like with the human learning, the machine needs to "make-sense" from the data, and therefore, only those features from the input data that shall be useful to the machine, have to be selected. The selection of the features is implemented in the second module of the pipeline, named feature selection. Feature selection utilizes different techniques, such as principal component analysis (PCA), Random Forests (RF), Bayesian methods and information criteria (the list is not exhaustive). The last step of the pipeline is utilizing the selected features to "train" the machine to make predictions or classifications, this happens using an appropriate prediction/classification algorithm (eg, Support Vector Machines (SVM), Neural Networks, decision trees). The result of this process shall be interpreted in an educationally meaningful manner by the human end-user, in order to guide decision making and instructional interventions, accordingly.

Methods—Case study

Participants

Thirty-two undergraduate students (15 females [46.9%] and 17 males [53.1%], aged 18–21 years-old [M = 19.24, SD = 0.831]) at a European University enrolled in an online adaptive self-assessment procedure for the Web Technologies course (related to front-end development). The participants undertook the self-assessment task individually, at a University lab, especially equipped and organized for the needs of the experimental process, for approx. 45 minutes each student, on October 2018.

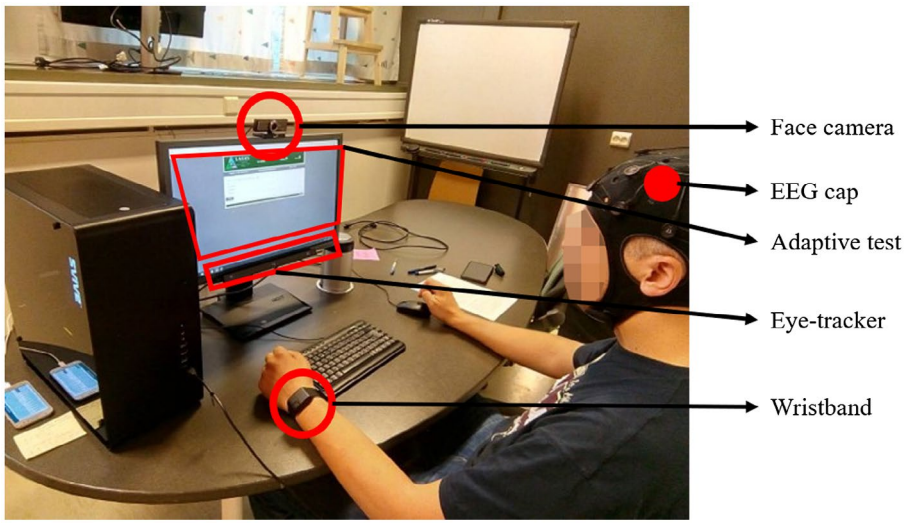

Figure 2: The experimental setup—The participant is connected to all data collection devices and is ready to take the self-assessment test
[Colour figure can be viewed at wileyonlinelibrary.com]

Study design and experimental procedure

Prior to their participation, all students signed an informed consent form that explained to them the procedure and was giving the right to researchers to use the data collected for research purposes. After granting consent, the participants had to wear a wristband and an EEG cap and be connected to all the data collection devices (ie, eye-tracker, wristband, EEG, cameras). The experimental setup is illustrated in Figure 2. Then, the actual adaptive self-assessment test started and the students had to answer to the test items.

Each item had two to four possible answers, but only one was the correct. Every time the student submitted an answer to an item, her mastery class was revised and the next item was delivered to her, according to the correctness of the answer and the distinguishability of the items. Specifically, the selection of the next item was based on entropy, a maximum information gain strategy from Information Theory. The goal was to select the item that has the greatest expected reduction in entropy, ie, that better fits the learner's mastery class, based on the answers she provided on the previous items. For adapting the self-assessment, the Measurement Decision Theory (MDT) (Rudner, 2003) was utilized (for the full description of the adaptation mechanism, and the preparation of the item bank, please see the supplementary material-Appendix E).

Finally, the test score was made available to the students, along with their full-test results, including all the items they had answered to, their responses, the correctness of the responses, and the option to check the correct answer to the items that they had submitted wrong answers, to rethink and self-reflect.

The participation to the procedure was optional. The adaptive self-assessment tests were offered to facilitate the students' self-preparation before the final exams, to help them track their progress, align with their learning goals and self-reflect. The scores on these tests had no participation to the final grade (ie, no rewards as external motivation).

It should be clarified that the decision to conduct the study in a self-assessment testing context was grounded on previous research that demonstrates that students who take practice tests often

outperform students in non-testing learning conditions such as restudying, practice, or filler activities. A recent meta-analysis examined the effects of practice tests versus non-testing learning conditions, and the results revealed that practice tests are more beneficial for learning than restudying and all other comparison conditions (Adesope, Trevisan, & Sundararajan, 2017). Furthermore, it has been argued that self-assessment leads students to a greater awareness, by training them to self-regulate their motivation and behavior, as well as by promoting metacognition and fostering reflection on their own progress in knowledge or skills, and finally, to understanding themselves as learners (Nicol & MacFarlane-Dick, 2006).

Data collection

We collected sensor data from four different sources: eye-tracking, EEG, facial video and arousal data from wristband (HR, BVP, EDA and skin temperature (TEMP)). The details of the apparatus and the setup of each device can be found in Appendix F.

Furthermore, during the study we also computed participants' effort and performance for the whole assessment session (ie, the dependent variables-outcomes of interest).

Effort

As stated in the Introduction, effortful behavior and on-task engagement are considered synonyms in this study. In other words, effort is an indicator ofhow much engaged the learners are in completing the tasks. In this study, for the effort calculation, the response time effort (RTE) measurement was employed (Wise & Kong, 2005). RTE measures the proportion of items thatthe students try to solve (solution behavior) instead of guessing the answers (guessing behavior). Details can be found in Appendix E.

Performance

The measurement used for students' performance in this study was the score the students achieved on the self-assessment test. For the score computation, only the correct answers were considered, without penalizing the incorrect answers (ie, without negative scores), due to the adaptive nature of the test. Specifically, the selection of the next item to deliver to students was guided by the correctness of the previous answer, and as such, the incorrect answers participated in formulating the "degree of difficulty" of the test. Furthermore, due to the adaptive nature of the test, the students had to respond to and solve different number of items. Overall, a minimum of 10 and a maximum of 20 items were used to classify the students based on their diagnosed mastery level. To overcome these issues concerning the score computation, each student's j learning performance (LP) was calculated as: $\mathrm{LP}_j = \frac{\sum_{i=1}^{k} d_i z_i}{k}$, where k is the number of items and according to the correctness of the student's answer on each item i, with $z_i \in \{0,1\}$ and the difficulty of the item, d_i. Each item had been previously weighted based on its difficulty level (see supplementary material) and contributed differently to the overall self-assessment score, ranging from 0.5 points (easy) to 1 point (medium) to 1.5 points (hard). The final score was on a [0–10] scale.

Building the pipeline

Feature extraction

After collecting the data, we proceeded to the feature extraction step. Specifically, we defined the eye-tracking features based on events (ie, fixations (Reichle, Warren, & McConnell, 2009); saccades, (Russo *et al.*, 2003), pupil diameter (Prieto, Sharma, Kidzinski, Rodríguez-Triana, & Dillenbourg, 2018)), and we computed the mean, variance, maximum and median and other statistics of those events (eg, number of fixations and saccades and the ratio of fixations and saccades).

For the EEG data stream, we defined band specific features. We calculated band powers of alpha, lower beta and theta bands (Worden, Foxe, Wang, & Simpson, 2000) from all the 17 channels. Band power is calculated as the root mean square of a signal over a period. The bands are frequency ranges and are strongly correlated to cognitive states (Hassib, Khamis, Friedl, Schneegass, & Alt, 2017). For example, the alpha band power has been associated with attention (Huang, Jung, & Makeig, 2007), the lower-beta band is related tomemory and theta is related to cognitive load (Kumar & Bhuvaneswari, 2012).

Furthermore, using face videos we defined expressions and features from different face regions (eyes, nose, mouth, jawline). Following a best practice of the literature, we extracted the facial Action Units (AUs, Cohn, Ambadar, & Ekman, 2007) using the OpenFace library (Amos, Ludwiczuk, & Satyanarayanan, 2016). Figure 3 shows the AUs detected in this study.

Finally, we defined features from the arousal data using the distributions of the data coming from the four different sensors (ie, HR, EDA, TEMP, BVP) (Kikhia *et al.*, 2016). From the Empatica E4 wristband we extracted the following statistical features: mean, median, variance, skewness, maximum for all the different data streams.

An overview of the extracted features from all data sources can be found in Appendix B.

Feature selection

As mentioned in the previous section, several feature selection techniques can be employed. In this study, we compare two commonly used feature selection techniques, ie, PCA and RF, briefly described in Appendix F.

Prediction algorithms

SVM, decision trees and Gaussian process regression (GPR) are used to predict the student performance and effort, using the selected features. Brief descriptions of the prediction methods can be found in Appendix F.

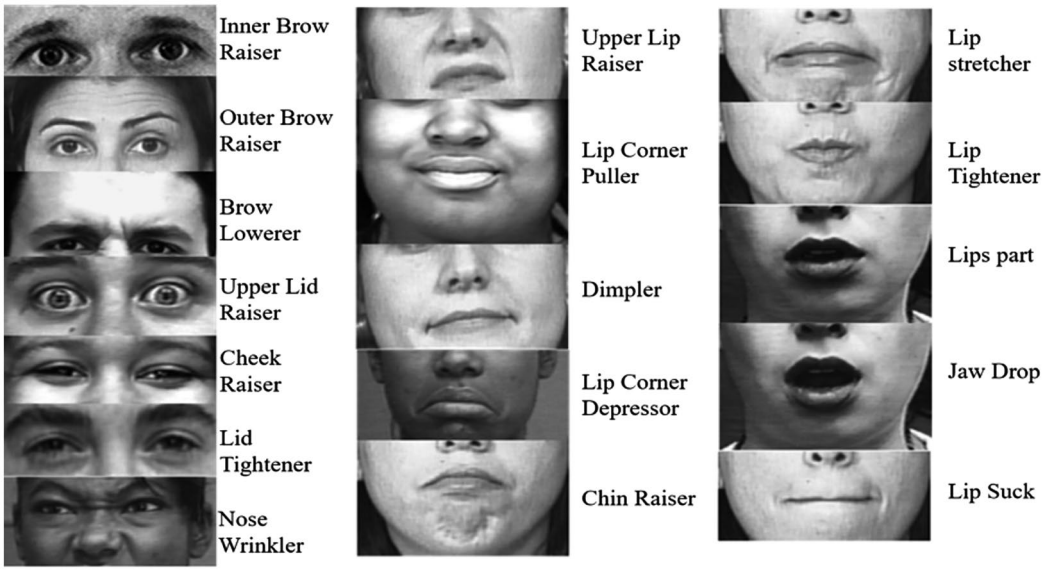

Figure 3: Action units extracted using the OpenFace Library. Action unit 45 (not shown in the figure) is "Blink"

Figure 4: The different stages of the specific machine learning pipeline built in this study [Colour figure can be viewed at wileyonlinelibrary.com]

The overall ML pipeline built in this study is illustrated in Figure 4:

Data analysis

To identify how the combinations of the multimodal data can predict performance and effort in the adaptive assessment, we divided the whole dataset into training and testing subsets, retaining data from the quarter (ie, eight) of the participants for testing. We repeated these four times, to use every quarter for training, this is a common ML technique. Further, we performed a four-fold cross-validation to remove the sampling bias from the training set. To evaluate and compare the different predictive models (solutions), we used the Normalized Root Mean Squared Error (NRMSE). NRMSE is the proposed metric for student models (Pelánek, 2015), and is used in most of the articles in learning technology (Moreno-Marcos, Alario-Hoyos, Muñoz-Merino, & Kloos, 2018) for measuring the accuracy of learning prediction.

Results

In this section, we present the prediction results for effort and performance and the feature selection results for the scenarios with the lowest NRMSE.

Prediction results

All solutions provided prediction accuracy (all in NRMSE hereinafter) ranging from 12.08% to 39.44% for effort, and from 6.21% to 25.49% for performance. For simplicity reasons, we present only the top results in the following subsections (ie, less than 15% for effort and less than 10% for performance). The comprehensive list of selected features for all possible combinations of modalities, feature selection algorithms and prediction methods, for both dependent variables (ie, total 300 combinations) can be found in Appendix D. The basic idea is to have the NRMSE value as low as possible, ie, as close to 0 as possible. However, the interpretation of "*how good a given value of NRMSE is*" can be based on the range of the predicted variable. For example, in our case, the two predictables were performance (scale 0–10) and effort (scale 0–20). We achieve 6.21% and 12.08% NRMSE for performance and effort, respectively. One can interpret the performance error as an error of 0.6 marks on the scale of 0–10. Similarly, the interpretation of error in effort prediction could be translated to about two questions difference in the actual number of guessed answers.

Prediction of effort

Table 1 illustrates the results for the prediction of students' on-task effort, in the self-assessment activity. Both feature selection methods (ie, PCA and RF) combined with the SVM radial prediction algorithm, provide the optimal prediction, with the lowest error rate being 12.08% in both solutions. Both optimal solutions combine features from the same sources (ie, eye-tracking, faces and wristband data). Figure 5 shows the results of predicting the performance (least NRMSE) using SVM Radial with the different values of learning rate and the radius size. This predictor

Table 1: Prediction results for students' effort in adaptive self-assessment

Feature selection	*Prediction algorithm*	*Modalities*	*NRMSE (%)*
RF	Svm radial	**ET-Face-WB**	**12.08**
	Svm radial	Face-WB	13.09
PCA	SVM Linear	EEG	14.77
		EEG-Face-WB	13.56
		EEG-WB	13.53
		ET	12.81
		ET-EEG-Face	13.82
		ET-Face	14.17
	SVM radial	All	14.30
		EEG	13.77
		EEG-Face-WB	13.98
		EEG-WB	13.97
		ET	14.20
		ET-EEG	12.36
		ET-EEG-Face	14.30
		ET-EEG-WB	13.23
		ET-Face	13.34
		ET-Face-WB	**12.08**
		ET-WB	12.19
		Face	12.27
		Face-WB	14.68
		WB	13.68

ET = eye-tracking; **EEG** = electro encephalograph; **WB** = wristband data. The values in bold depict the best NRMSE results.

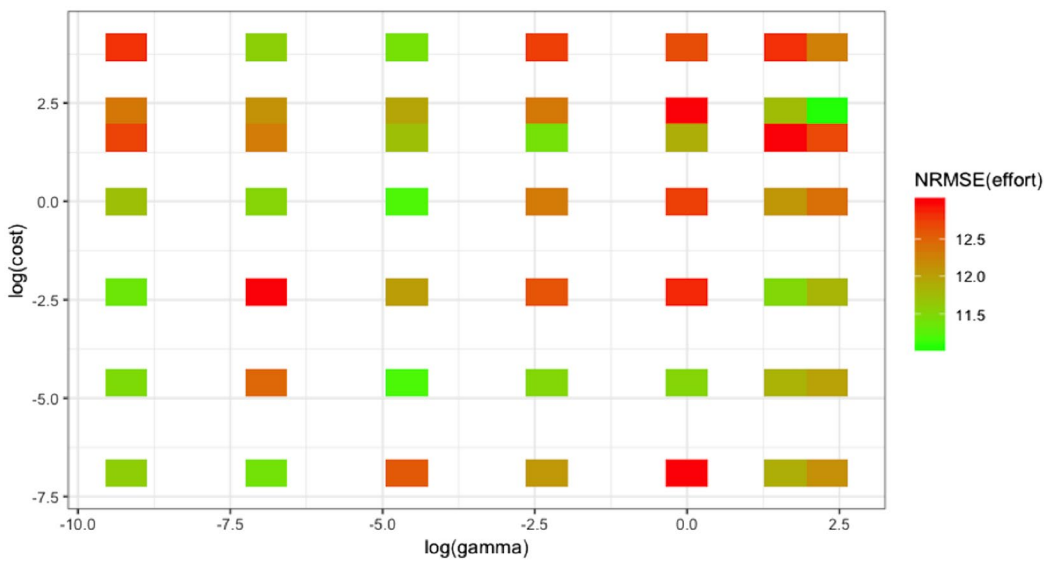

Figure 5: Prediction results for effort using SVM radial with PCA features using the combination of eye-tracking, face and wristband data
[Colour figure can be viewed at wileyonlinelibrary.com]

uses the features from eye-tracking, facial expressions and wristband data. On the other hand, the red curve in the Figure 7 shows the results of predicting the performance (highest NRMSE)

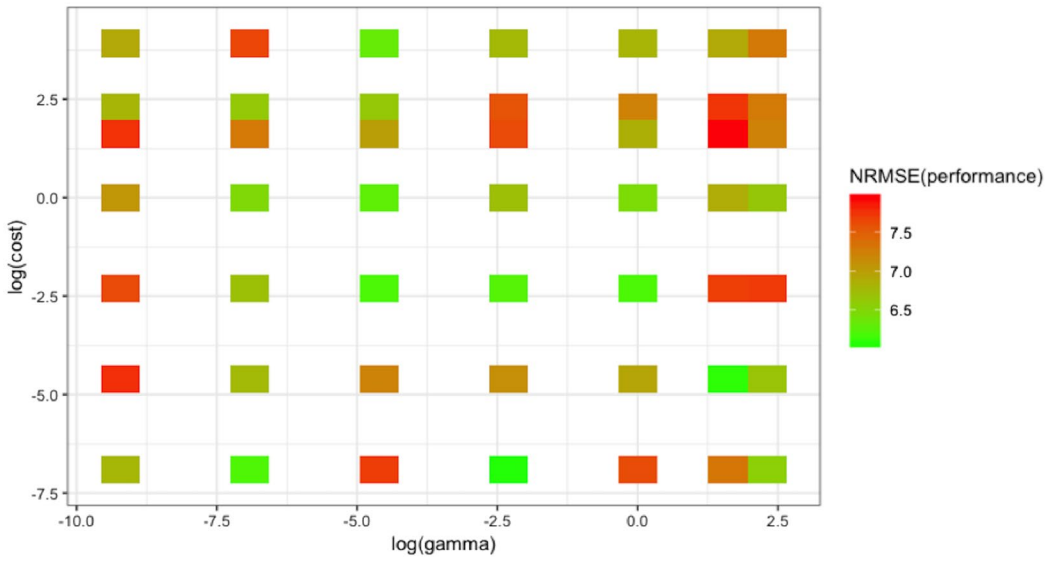

Figure 6: Prediction results for performance using SVM radial with PCA features using the combination of eye-tracking, face and wristband data
[Colour figure can be viewed at wileyonlinelibrary.com]

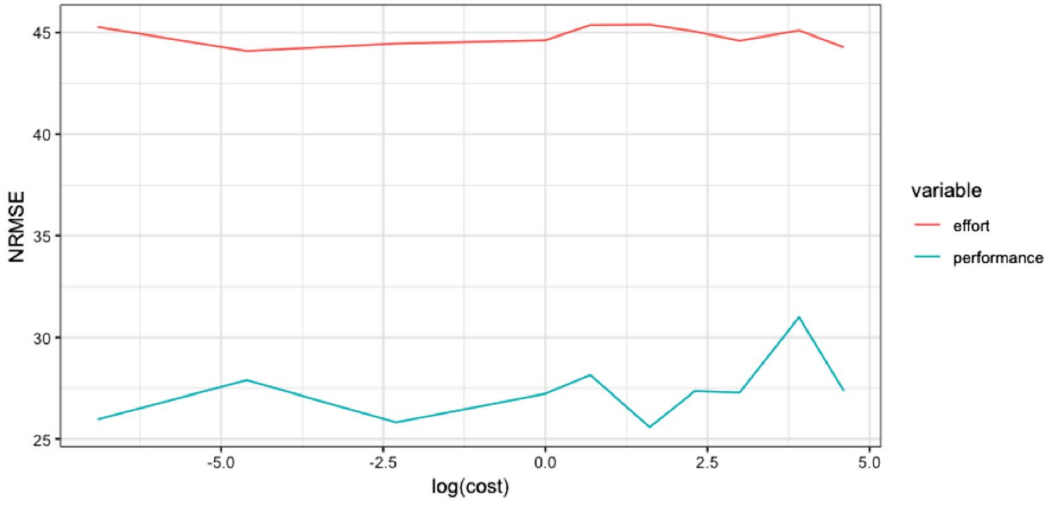

Figure 7: Prediction results for performance using SVM radial with PCA features using the facial data (blue line). Prediction results for effort using SVM radial with PCA features using the combination of eye-tracking, face and EEG data (red line)
[Colour figure can be viewed at wileyonlinelibrary.com]

using SVM Linear with the different values of learning rate. This predictor uses the features from facial expressions.

Prediction of performance

Table 2 demonstrates the results for the prediction of performance. As seen in this table, PCA outperforms RF (both optimal predictions utilize SVM radial); the lowest error rate using PCA

Table 2: Prediction results for students' performance

	RF feature selection				PCA feature selection			
Modalities	*Gaussian poly*	*Model trees*	*SVM poly*	*SVM radial*	*Gaussian poly*	*SVM linear*	*SVM poly*	*SVM radial*
All	9.67				8.21	7.21	8.40	7.65
EEG				9.34	8.14	8.04	8.36	8.23
EEG-Face					8.18	7.57	8.40	7.76
EEG-Face-WB				9.76	8.12	7.59	8.37	7.71
EEG-WB				9.50	8.13	7.79	8.36	7.47
ET				9.41	8.15	7.18	8.38	7.46
ET-EEG	9.48			9.77	8.15	7.35	8.39	6.66
ET-EEG-Face				9.82	8.15	7.15	8.38	7.49
ET-EEG-WB				9.51	8.16	8.99	8.36	8.05
ET-Face					8.15	8.70	8.37	8.07
ET-Face-WB	8.10		9.55	**7.99**	8.14	7.55	8.39	**6.21**
ET-WB	9.36				8.14	7.27	8.40	7.40
Face	9.00			9.06	8.14	7.76	8.39	6.49
Face-WB	7.41	7.70	8.55	8.07	7.99	8.49	8.38	7.78
WB	9.36		9.98	9.01	8.20	7.50	8.40	7.75

Gaussian poly = Gaussian process models with polynomial kernel; SVM Linear = SVM with linear kernel; SVM Poly = SVM with polynomial kernel; SVM Radial = SVM with radial kernel. ET = eye-tracking; WB = wristband. The missing values in the table are all more than 10% NRMSE. The values in bold depict the best NRMSE results.

is 6.21%, while the lowest error rate using RF is 7.99%. In terms of modalities, we see that the same combination of modalities (ie, eye-tracking, facial action units and wristband data) results the optimal prediction, in both cases. Figure 6 shows the results of predicting the performance (least NRMSE) using SVM Radial with the different values of learning rate and the radius size. This predictor uses the features from eye-tracking, facial expressions and wristband data. On the other hand, the blue curve in the Figure 7 shows the results of predicting the performance (highest NRMSE) using SVM Linear with the different values of learning rate. This predictor uses the features from eye-tracking, facial expressions and EEG data.

Feature selection

In this section, we present feature selection results using PCA and RF. Those features were used to predict students' effort and performance, with the lowest NRMSE (ie, SVM with radial kernel). The complete results from the feature selection module can be seen in Appendices B and C.

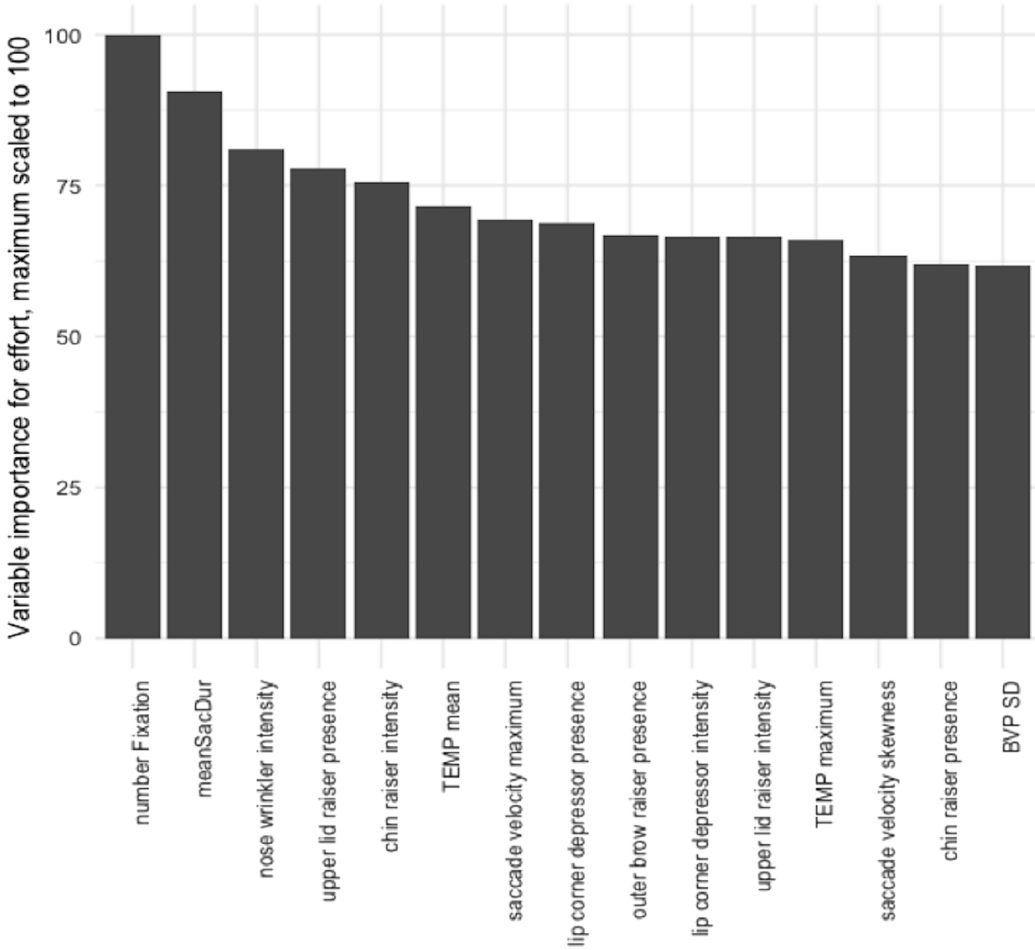

Figure 8: Variable importance with RF for effort prediction (the maximum importance is normalized at 100) [Colour figure can be viewed at wileyonlinelibrary.com]

1. **Feature selection with RF for effort prediction:** the lowest NRMSE is obtained by combining the eye-tracking, facial and wristband features (Table 1). Figure 8 depicts the top 15 features used to predict students' effort. The most important feature (ie, important predictors) is the number of fixations (eye-tracking) during the test, followed by average saccade duration (eye-tracking) and intensity of nose wrinkler (face). Finally, the most important feature from wristband is the average temperature.
2. **PCA feature selection for effort prediction:** as explained in Feature Selection sub-section, a threshold on the number of components to use in the prediction module was set at 90% of the variance explained in the data. This resulted in the top 17 components (explain 91.06% of variance). Figure 9 illustrates the correlations between the top 34 features (coming from the 17 components) and effort. The most correlated feature is the maximum BVP (wristband), followed by lip corner puller presence (face) and inner brow raiser presence (face) (in absolute values). Finally, the most correlated feature from eye-tracking is the number of saccades.
3. **RF feature selection for performance prediction:** again, the lowest NRMSE is obtained by combining the eye-tracking, facial and wristband features (Table 2). Figure 10 shows the top 15 features used to predict students' performance. The most important feature (top predictor) is the number of fixations (eye-tracking) during the test, followed by blink presence (face) and kutrosis of temperature (wristband).
4. **PCA feature selection for performance prediction:** the 90% threshold on the variance explained resulted in 19 components, in this case (explain 92.27% of variance). Figure 11 illustrates the correlations between top 38 features (from the 19 components) and effort. The most important features are the inner brow raiser presence (face) and its intensity (face)

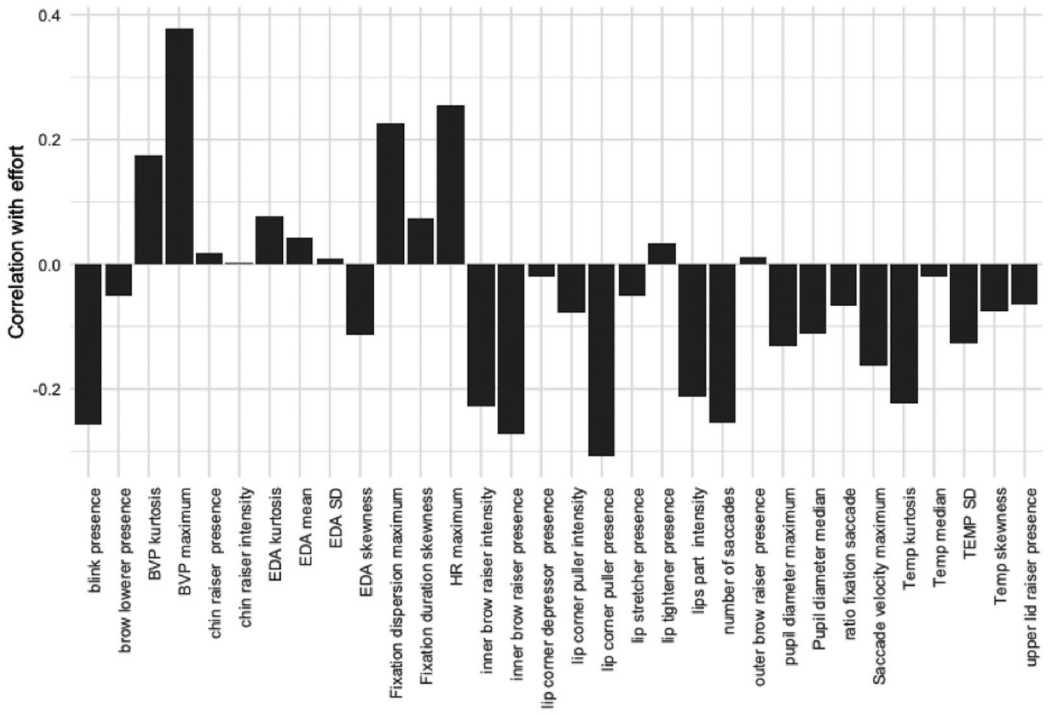

Figure 9: Pearson correlation between the effort and the top 34 features corresponding to each of the top 17 PCA components

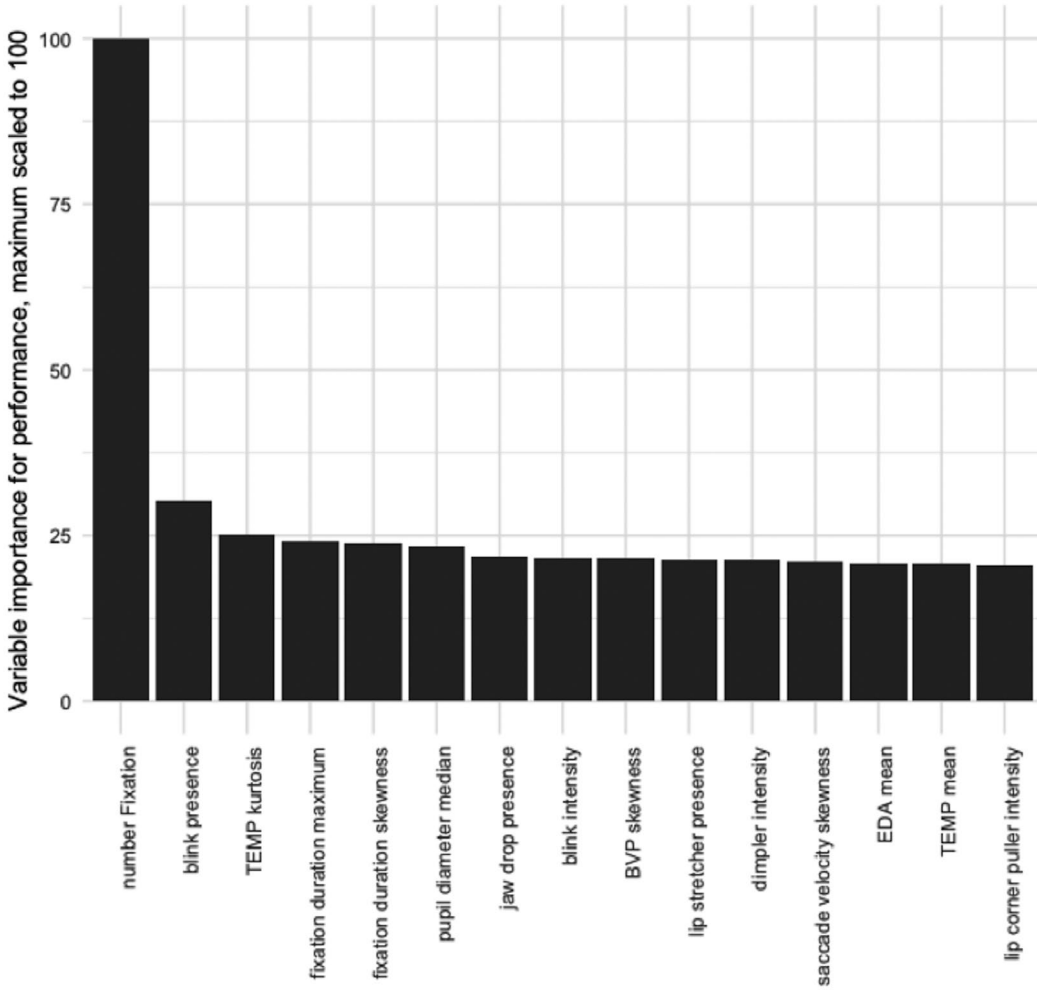

Figure 10: Variable importance with RF for performance prediction, (the maximum importance is normalized at 100)

followed by chin raiser presence (face). The most important feature from eye-tracking and wristband data are number of saccades and maximum BVP respectively.

Discussion

Previous studies revealed significant findings in terms of what physiological data are appropriate for explaining students' behavior (Lane & D'Mello, 2019) and modeling and predicting their emotions, engagement with the tasks and performance, in diverse learning settings (D'Mello *et al.*, 2009; Di Mitri *et al.*, 2018; Fairclough *et al.*, 2009; Marshall, 2002; Spikol *et al.*, 2018). In these settings, the exploitation of ML techniques was proposed to reduce human workload during the analysis of learners' interactions, and to select appropriate multimodal data for capturing learners' behaviors, (Andrade *et al.*, 2016; Ochoa *et al.*, 2018). However, the lack of previous results in

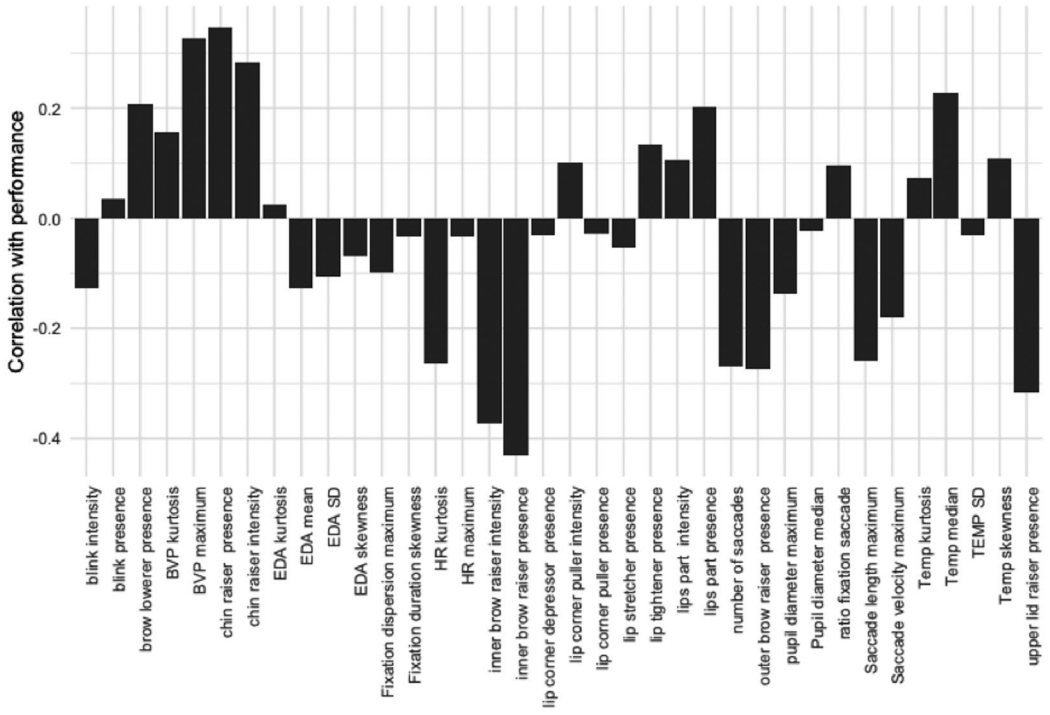

Figure 11: Pearson correlation between the performance and the top 38 features corresponding to each of the 19 PCA components

adaptive learning conditions, a gap in the methodology to systematically employ multimodal data fusion and analysis for prediction purposes, and the need to explicate and generalize the "step-by-step" building of a ML pipeline motivated the present study.

The main aim of the study was to identify the most important constructs from multiple modalities and their combinations that predict students' effortful behavior and performance (or lack thereof) in adaptive learning procedures. The majority of previous research lack objective ground truth (Di Mitri *et al.*, 2018) and do not explain the educational constraints (eg, ubiquity, low-cost, high precision, different experimental settings) based on which the features in use were selected. To address the research question, this work suggests a "grey-box" approach as a generic methodology for building ML pipelines for multimodal educational data and exemplifies its usage in a study that employs data from four physiological data sources.

Specifically, we collected EEG, eye-tracking, facial expressions and wristband data from 32 students, while they were answering an adaptive self-assessment test. Next, we extracted features from the data sources (eg, number of fixations, blink presence, BVP) that have been commonly used in literature (D'Mello *et al.*, 2018; Huang *et al.*, 2007; Kikhia *et al.*, 2016; Reichle *et al.*, 2009), to add ground truth contextual knowledge (Di Mitri *et al.*, 2018), that would be necessary for the interpretation of the findings later on. After the feature extraction step, we employed two feature selection algorithms (PCA, RF) to find the set of a few important variables contributing to the learning outcomes, ie, that strongly correlate with effort and performance. Using the selected features, we predicted the outcomes of interest, by configuring three commonly used ML algorithms (SVM, Model Trees, GPR).

Findings and interpretations

Extending previous results (eg, Fairclough *et al.*, 2009; Marshall, 2002), and coinciding with more recent findings (eg, Di Mitri *et al.*, 2017; Junokas *et al.*, 2018; Spikol *et al.*, 2018), our findings suggest that although individual modalities can be a good proxy for performance and effort, fusing features from different modalities has the potential to further increase prediction accuracy. In other words, and in line with Giannakos *et al.* (2019), it is confirmed that data fusion produces more consistent and accurate predictions than those from individual data sources.

It is important to mention, though, that the features to be fused were not randomly selected (ie, not in a "black-box" approach, just because the ML algorithms perform better with those ones), but they were chosen in a literature-driven "white-box" approach, taking advantage of the most appropriate background work.

Specifically, one interesting finding is that both feature selection methods (PCA, RF) returned fixation duration, saccade duration and saccade velocity histogram-based features (ie, maximum, skewness, mean) from eye-tracking data. Fixation duration has been found to be correlated with learners' attention (Abernethy & Russell, 1987; Reichle *et al.*, 2009), whereas saccade duration often indicates task difficulty (Bestelmeyer *et al.*, 2006; Vuori, Olkkonen, Pölönen, Siren, & Häkkinen, 2004). Furthermore, skewness of saccade velocity histogram often reveals students' anticipation patterns (Liao *et al.*, 2005).

Furthermore, facial action units are known to be related to students' emotions (Cohn *et al.*, 2007; Lewinski, den Uyl, & Butler, 2014). For example, upper lid raiser indicates happiness; dimpler and lip corner puller are constituents of contempt; inner brow raiser and lip corner depressor contribute to sadness; and finally, lip and lid tightener are the main components of anger. Both PCA and RF yielded them among the most important ones. This result ties emotions to performance and effortful engagement in a more direct manner than previously reported (Linnenbrink & Pintrich, 2002, 2003). Moreover, using only facial features (ff) results in NRMSE that is comparable to the best model for performance ($NRMSE_{ff}$ = 7.76%; $NRMSE_{best}$ = 6.21%) and effort ($NRMSE_{ff}$ = 12.27%; $NRMSE_{best}$ = 12.08%). This finding is interesting in that it supports that using only facial data is a satisfactory, yet low-cost and ubiquitous solution.

Moreover, regarding wristband data, histogram-based features were selected for the best predicting model, as well. Those data streams have recently been found to be related to learners' perceived performance and satisfaction (Sharma, Pappas,Papavlasopoulou, & Giannakos, 2019). In the past, they had found to be good predictors of engagement (Worsley & Blikstein, 2014). Also, using features from only wristband (WB) data results in NRMSE that is comparable to the best model for performance ($NRMSE_{wb}$ = 7.50%; $NRMSE_{best}$ = 6.21%) and effort ($NRMSE_{wb}$ = 13.68%; $NRMSE_{best}$ = 12.08%). This result is useful in terms that using only wristband data can provide accurate mobile and ubiquitous solutions.

Our predictors (the features from physiological data) and dependent variables are continuous valued, making it possible for the students to have multiple different types of effortful behavior simultaneously (eg, higher attention, lower cognitive load and negative emotions such as anger), providing a finer-grained explanation of effort, and a more accurate estimation of performance, as well. This extends results from previous studies that used classes (eg, Worsley, 2018) and hence were not able to predict continuous variables or they predicted continuous variables, yet not with multimodal data (Sharma, Jermann, & Dillenbourg, 2015). The present study exemplified how to solve this problem by keeping the dependent variables continuous, claiming that low error rate is feasible by fusing multimodal data.

However, the feature space that can be computed includes features that are not always easy to use and are not always interpretable in educationally meaningful ways in different learning settings. In the demonstrated "grey-box" approach, the literature-driven feature selection process allows us to select those features from which we can make sense by explaining them in educational terms. This gives us an opportunity to continue exploring multimodal data to provide actionable feedback to both the students and teacher with little amount of training.

Conclusions

The inherent particularities of adaptive learning, ie, the fact that the tasks are tailored to the detected mastery level and abilities of the learners, as well as the estimation of learners' performance directly from the adaptive learning environment itself, make this context a very special learning setting for the study of learners' effortful engagement, which has been extensively studied in other learning conditions (Galla *et al.*, 2014; Hughes, Luo, Kwok, & Loyd, 2008).

This study demonstrated a consolidated analysis of fused multimodal educational data, collected during an adaptive self-assessment activity, using sophisticated ML methods for prediction purposes. The implications of the suggested approach are discussed in this section.

Implications for research and practice

First, this study adds to the educational technologies research by providing a generalized and modularized "grey-box" methodology for building ML pipelines for multimodal educational data, aiming to justify each step in the process, to predict effortful engagement and performance in adaptive learning settings, and bridge the existing gaps in relevant literature. This is the first study—to the best of our knowledge—that explicitly determines the steps of the pipeline building process, grounds the selection of multimodal features on relevant literature, fuses the diverse multimodal data and simplifies a series of sophisticated artificial intelligence techniques to shed light to the "black-box" of ML for educationally meaningful outcomes.

Most of the past works have used the modules we present; however, (a) they do not mention the other options available to the researchers (eg, Andrade *et al.*, 2016; Ochoa *et al.*, 2018); (b) they describe the algorithms only as part of the methods section without associating the feature selection process to the possible constraints of the educational context (ie, in a "white-box," hypothesis formulation manner); (c) they do not provide a literature-driven explanation for the selection of the features (they rely on with what features the ML algorithm performs better).

The thorough analysis showcased that multimodal data fusion and different ML algorithms can provide useful predictions about students' effort exertion and performance that are easy to interpret in terms of physiological learner states. The demonstrated "grey-box" approach is a methodological "tool" in the hands of educational technologies researchers and professionals, to support them identify those features within the physiological data that are grounded in previous contextual knowledge and can best explain the learning situation they are trying to understand.

Considering the restrictions (constraints) from the educational context (eg, ubiquity, low-cost, high precision, different experimental settings), the step-by-step modularized methodology can be utilized not only for the multimodal data but also for separate standalone data sources such as clickstreams, postures, gestures, gaze We shown 150 examples of different pipelines that can be built for each of our dependent variables: these pipelines consist of different modules, ie, data sources, feature selection and prediction algorithms. The options for the different modules are not limited to the ones presented in this paper. Moreover, due to the nature of the modules employed, such pipelines are transferable to other contexts beyond adaptive self-assessment.

Implications for adaptive learning

The modularized methodology presented in this study can help learning design professionals, as a tool, to identify and integrate specific features into the adaptive learning environments (based on physiological data), to prevent cognitive students' overload and effortless behavior (eg, guessing).

The most important features from the pipeline which are critical in terms of learner effortful behavior (eg, blinks, brow raiser presence, nose wrinkler), can be extracted and delivered back to the students, thus opening the learner models to them (Bull & Nghiem, 2002). These features can also be used in an aggregated fashion to display on the dashboards for teachers (Martinez-Maldonado, Echeverria, Santos, Santos, & Yacef, 2018; Prieto, Sharma, Dillenbourg, & Jesús, 2016). As we have mentioned before, the most important features (in terms of how they contribute to improving prediction accuracy for both performance and effort) can be explained in educational terms, extending previous studies that claim that the same features are important for learning as well. For example, attention is a key factor in achieving high performance (eg, Greenfield, DeWinstanley, Kilpatrick, & Kaye, 1994; Harris, Danoff Friedlander, Saddler, Frizzelle, & Graham, 2005; Sharma, Alavi, Jermann & Dillenbourg, 2016) and emotions also play a significant role in explaining learning processes (eg, Frenzel, Pekrun, & Goetz, 2007; Meyer & Turner, 2002; Pekrun, 2006). Actionable feedback is one of the most important issues to be dealt with in adaptive learning. By using physiological data from different channels, we showed in this paper, that with a few minutes of interaction to be able to provide this kind of feedback (The average duration of the self-assessment test was 8 minutes and 41.93 seconds, SD = 2 minutes and 0.56 seconds). This provides possible paths to implement actionable feedback systems for learners.

Acknowledgements

This work is supported from the Norwegian Research Council under the projects FUTURE LEARNING (number: 255129/H20) and Xdesign (290994/F20). This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

Statement on open data, ethics and conflict of interest

As it is possible to identify participants from the data, ethical requirements do not permit us to share participant data from this study.

Participation was voluntarily, and all the data collected anonymously. Appropriate permissions and ethical approval for the participation requested and approved.

There is no potential conflict of interest in this study.

References

Abernethy, B., & Russell, D. G. (1987). Expert-novice differences in an applied selective attention task. *Journal of Sport Psychology*, *9*(4), 326–345.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). *Openface: A general-purpose face recognition library with mobile applications*. Pittsburgh, PA: CMU School of Computer Science.

Andrade, A. (2017, March). Understanding student learning trajectories using multimodal learning analytics within an embodied-interaction learning environment. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 70–79). Vancouver, Canada: ACM.

Andrade, A., Delandshere, G., & Danish, J. A. (2016). Using multimodal learning analytics to model student behaviour: A systematic analysis of epistemological framing. *Journal of Learning Analytics*, *3*(2), 282–306.

Baker, R. S. J. D., Corbett, A. T., Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In B. P. Woolf, E. Aïmeur, R. Nkambou, & S. Lajoie. (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems, ITS 2008* (pp. 406–415). Berlin, Germany: Springer.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behaviour in the cognitive tutor classroom: When students "game the system". In *Proceedings of ACM CHI 2004: Computer-Human Interaction* (pp. 383–390). Vienna, Austria.

Barla, M., Bieliková, M., Ezzeddinne, A. B., Kramár, T., Šimko, M., & Vozár, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computers & Education*, 55(2), 846–857.

Beardsley, M., Hernández-Leo, D., & Ramirez-Melendez, R. (2018). Seeking reproducibility: Assessing a multimodal study of the testing effect. *Journal of Computer Assisted Learning*, *34*(4), 378–386.

Bestelmeyer, P. E., Tatler, B. W., Phillips, L. H., Fraser, G., Benson, P. J., & Clair, D. S. (2006). Global visual scanning abnormalities in schizophrenia and bipolar disorder. *Schizophrenia Research*, *87*(1–3), 212–222.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238.

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(2), 1–26.

Bull, S., & Nghiem, T. (2002, June). Helping learners to understand themselves with a learner model open to students, peers and instructors. In *Proceedings of workshop on individual and group modelling methods that help learners understand themselves, International Conference on Intelligent Tutoring Systems* (Vol. *2002*, pp. 5–13). Biarritz, France.

Cabestrero, R., Quirós, P., Santos, O. C., Salmeron-Majadas, S., Uria-Rivas, R., Boticario, J. G., … Ferri, F. J. (2018). Some insights into the impact of affective information when delivering feedback to students. *Behaviour & Information Technology*, *37*(12), 1252–1263.

Chang, S.-R., Plake, B. S., Kramer, G. A., & Lien, S.-M. (2011). Development and application of detection indices for measuring guessing behaviours and test-taking effort in computerized adaptive testing. *Educational and Psychological Measurement*, *71*(3), 437–459. https://doi.org/10.1177/0013164410385110

Chen, I. S. (2017). Computer self-efficacy, learning performance, and the mediating role of learning engagement. *Computers in Human Behaviour*, *72*, 362–370.

Chen, L., Feng, G., Leong, C. W., Joe, J., Kitchen, C., & Lee, C. M. (2016). Designing an automated assessment of public speaking skills using multimodal cues. *Journal of Learning Analytics*, *3*(2), 261–281.

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. In R. W. Levenson, J. A. Coan, & J. J. B. Allen (Eds.), *The handbook of emotion elicitation and assessment* (pp. 203–221). New York: Oxford University Press.

Conati, C., & Kardan, S. (2013). Student modeling: Supporting personalized instruction, from problem solving to exploratory open ended activities. *AI Magazine*, *34*(3), 13–26.

Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., & Specht, M. (2017, March). Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 188–197). Vancouver, Canada: ACM.

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*(4), 338–349.

D'Mello, S., Dieterle, E., & Duckworth, A. (2017). Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist*, *52*(2), 104–123.

D'Mello, S. K., Bosch, N., & Chen, H. (2018, October). Multimodal-multisensor affect detection. In *The handbook of multimodal-multisensor interfaces* (pp. 167–202). Association for Computing Machinery and Morgan & Claypool.

D'Mello, S. K., Craig, S. D., & Graesser, A. C. (2009). Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, *4*(3–4), 165–187.

Ezen-Can, A., Grafsgaard, J. F., Lester, J. C., & Boyer, K. E. (2015, March). Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 280–289). New York, NY: ACM.

Fairclough, S. H., Moores, L. J., Ewing, K. C., & Roberts, J. (2009, September). Measuring task engagement as an input to physiological computing. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009* (pp. 1–9). Amsterdam, the Netherlands: IEEE.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics—A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*(4), 497–514.

Furuichi, K., & Worsley, M. (2018, October). Using physiological responses to capture unique idea creation in team collaborations. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 369–372). Jersey City, NY: ACM.

Galla, B. M., Wood, J. J., Tsukayama, E., Har, K., Chiu, A. W., & Langer, D. A. (2014). A longitudinal multilevel model analysis of the within-person and between-person effect of effortful engagement and academic self-efficacy on academic performance. *Journal of School Psychology*, *52*(3), 295–308.

Giannakos, M. N., Sharma, K., Pappas, I. O., Kostakos, V., & Velloso, E. (2019). Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, *48*, 108–119.

Gilzenrat, M. S., Cohen, J. D., Rajkowski, J., & Aston-Jones, G. (2003, November). Pupil dynamics predict changes in task engagement mediated by locus coeruleus. *Society for Neuroscience Abstracts*, *515*, 19.

Gowda, S. M., Rowe, J. P., Baker, R. S. J. D., & Chi, M. (2011). Improving Models of Slipping, Guessing, and Moment-By-Moment Learning with Estimates of Skill Difficulty. *Educational Data Mining*, *2011*, 198–208.

Grafsgaard, J., Duran, N., Randall, A., Tao, C., & D'Mello, S. (2018, May). Generative multimodal models of nonverbal synchrony in close relationships. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 195–202). Xi'an, China: IEEE.

Greenfield, P. M., DeWinstanley, P., Kilpatrick, H., & Kaye, D. (1994). Action video games and informal education: Effects on strategies for dividing visual attention. *Journal of Applied Developmental Psychology*, *15*(1), 105–123.

Harris, K. R., Danoff Friedlander, B., Saddler, B., Frizzelle, R., & Graham, S. (2005). Self-monitoring of attention versus self-monitoring of academic performance: Effects among students with ADHD in the general education classroom. *The Journal of Special Education*, *39*(3), 145–157.

Hassib, M., Khamis, M., Friedl, S., Schneegass, S., & Alt, F. (2017, November). Brainatwork: logging cognitive engagement and tasks in the workplace using electroencephalography. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia* (pp. 305–310). Stuttgart, Germany: ACM.

Huang, R. S., Jung, T. P., & Makeig, S. (2007, April). Multi-scale EEG brain dynamics during sustained attention tasks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007* (Vol. *4*, pp. IV-1173). Honolulu, HI: IEEE.

Hughes, J. N., Luo, W., Kwok, O. M., & Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology*, *100*(1), 1–14.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*, *91*, 153–184.

Jung, Y., & Lee, J. (2018). Learning engagement and persistence in massive open online courses (MOOCS). *Computers & Education*, *122*, 9–22.

Junokas, M. J., Lindgren, R., Kang, J., & Morphew, J. W. (2018). Enhancing multimodal learning through personalized gesture recognition. *Journal of Computer Assisted Learning*, *34*(4), 350–357.

Kalsbeek, J. W. H., & Ettema, J. H. (1963). Scored regularity of the heart rate pattern and the measurement of perceptual or mental load. *Ergonomics*, *6*(3), 306–307.

Kikhia, B., Stavropoulos, T. G., Andreadis, S., Karvonen, N., Kompatsiaris, I., Sävenstedt, S., … Melander, C. (2016). Utilizing a wristband sensor to measure the stress level for people with dementia. *Sensors*, *16*(12), 1989.

Kinnealey, M., Pfeiffer, B., Miller, J., Roan, C., Shoener, R., & Ellner, M. L. (2012). Effect of classroom modification on attention and engagement of students with autism or dyspraxia. *American Journal of Occupational Therapy*, *66*(5), 511–519.

Kumar, J. S., & Bhuvaneswari, P. (2012). Analysis of Electroencephalography (EEG) signals and its categorization–A study. *Procedia Engineering*, *38*, 2525–2536.

Lai, M. L., Tsai, M. J., Yang, F. Y., Hsu, C. Y., Liu, T. C., Lee, S. W. Y., … Tsai, C. C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, *10*, 90–115.

Lane, H. C., & D'Mello, S. K. (2019). Uses of physiological monitoring in intelligent learning environments: A review of research, evidence, and technologies. In T. Parsons, L. Lin, & D. Cockerham (Eds.), *Mind, brain and technology. Educational Communications and Technology: Issues and Innovations*. Cham, Switzerland: Springer.

Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, *7*(4), 227–236.

Liao, K., Kumar, A. N., Han, Y. H., Grammer, V. A., Gedeon, B. T., & Leigh, R. J. (2005). Comparison of velocity waveforms of eye and head saccades. *Annals of the New York Academy of Sciences*, *1039*(1), 477–479.

Linnenbrink, E. A., & Pintrich, P. R. (2002). The role of motivational beliefs in conceptual change. In *Reconsidering conceptual change: Issues in theory and practice* (pp. 115–135). Dordrecht, the Netherlands: Springer.

Linnenbrink, E. A., & Pintrich, P. R. (2003). Motivation, affect, and cognitive processing: What role does affect play. In *Annual Meeting of the American Educational Research Association*, Chicago, IL.

Liu, M., McKelroy, E., Corliss, S. B., & Carrigan, J. (2017). Investigating the effect of an adaptive learning intervention on students' learning. *Educational technology research and development*, *65*(6), 1605–1625.

Liu, R., Stamper, J., Davenport, J., Crossley, S., McNamara, D., Nzinga, K., & Sherin, B. (2019). Learning linkages: Integrating data streams of multiple modalities and timescales. *Journal of Computer Assisted Learning*, *35*(1), 99–109.

Luft, C. D. B., Nolte, G., & Bhattacharya, J. (2013). High-learners present larger mid-frontal theta power and connectivity in response to incorrect performance feedback. *Journal of Neuroscience*, *33*(5), 2029–2038.

Marshall, S. P. (2002). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants, 2002* (pp. 7–7). Scottsdale, AZ: IEEE.

Martinez-Maldonado, R., Echeverria, V., Santos, O. C., Santos, A. D. P. D., & Yacef, K. (2018, March). Physical learning analytics: A multimodal perspective. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 375–379). Sydney, Australia: ACM.

Mattingly, S. M., Gregg, J. M., Audia, P., Bayraktaroglu, A. E., Campbell, A. T., Chawla, N. V., … Gao, G. (2019, April). The Tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (p. p. CS11). ACM.

McMillan, J. H., & Hearn, J. (2008). Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons*, *87*(1), 40–49.

Meyer, D. K., & Turner, J. C. (2002). Discovering emotion in classroom motivation research. *Educational Psychologist*, *37*(2), 107–114.

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., & Kloos, C. D. (2018). Prediction in MOOCs: A review and future research directions. *IEEE Transactions on Learning Technologies*, 1–1.

Moridis, C. N., & Economides, A. A. (2012). Affective learning: Empathetic agents with emotional facial and tone of voice expressions. *IEEE Transactions on Affective Computing*, *3*(3), 260–272.

Mulder, G. (1986). The concept and measurement of mental effort. In *Energetics and human information processing* (pp. 175–198). Dordrecht: Springer.

Mundy, P., Acra, C. F., Marshall, P., & Fox, N. (2006). Joint attention, social engagement, and the development of social competence. In P. J. Marshall & N. A. Fox (Eds.), *The development of social engagement: Neurobiological perspectives* (pp. 81–117). Oxford Scholarship Online.

Nicol, D. J., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.

Normadhi, N. B. A., Shuib, L., Nasir, H. N. M., Bimba, A., Idris, N., & Balakrishnan, V. (2019). Identification of personal traits in adaptive learning environment: Systematic literature review. *Computers & Education*, *130*, 168–190.

Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018, March). The RAP system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 360–364). Sydney, Australia: ACM.

Ochoa, X., & Worsley, M. (2016). Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, *3*(2), 213–219.

Papamitsiou, Z., & Economides, A. A. (2015). A temporal estimation of students' on-task mental effort and its effect on students' performance during computer based testing. In *IEEE 18th International Conference on Interactive Collaborative Learning (ICL2015), Florence* (pp. 1136–1144). https://doi.org/10.1109/ICL.2015.7318194

Papamitsiou, Z., & Economides, A. A. (2016). Process mining of interactions during computer-based testing for detecting and modelling guessing behaviour. In *Third International Conference on Learning and Collaboration Technologies* (pp. 437–449). Toronto, Canada.

Papamitsiou, Z., & Economides, A. A. (2019). Exploring autonomous learning capacity from a self-regulated learning perspective using learning analytics. *British Journal of Educational Technology, 50*(6), 3138–3155. https://doi.org/10.1111/bjet.12747

Pardo, A., Han, F., & Ellis, R. A. (2017). Combining university student self-regulated learning indicators and engagement with online learning events to predict academic performance. *IEEE Transactions on Learning Technologies*, *10*(1), 82–92.

Pardo, A., Poquet, O., Martínez-Maldonado, R., & Dawson, S. (2017). Provision of data-driven student feedback in la & EDM. *Handbook of Learning Analytics*, 163–174.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341.

Pelánek, R. (2015). Metrics for evaluation of student models. *Journal of Educational Data Mining, 7*(2), 1–19.

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education., 98*(2016), 169–179.

Pijeira-Díaz, H. J., Drachsler, H., Kirschner, P. A., & Järvelä, S. (2018). Profiling sympathetic arousal in a physics course: How active are students? *Journal of Computer Assisted Learning*, *34*(4), 397–408.

Prieto, L. P., Sharma, K., Dillenbourg, P., & Jesús, M. (2016, April). Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 148–157). Edinburgh, Scotland: ACM.

Prieto, L. P., Sharma, K., Kidzinski, Ł., Rodríguez-Triana, M. J., & Dillenbourg, P. (2018). Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data. *Journal of Computer Assisted Learning*, *34*(2), 193–203.

Raca, M., & Dillenbourg, P. (2014). Classroom social signal analysis. *Journal of Learning Analytics*, *1*(3), 176–178.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*(1), 1–21.

Rudner, L. M. (2003, April). The classification accuracy of measurement decision theory. In *Annual Meeting of the National Council on Measurement in Education* (Vol. *2325*). Chicago.

Russo, M., Thomas, M., Thorne, D., Sing, H., Redmond, D., Rowland, L., … Balkin, T. (2003). Oculomotor impairment during chronic partial sleep deprivation. *Clinical Neurophysiology*, *114*(4), 723–736.

Schneider, B., & Blikstein, P. (2015). Unraveling students' interaction around a tangible interface using multimodal learning analytics. *Journal of Educational Data Mining*, *7*(3), 89–116.

Sharma, K., Alavi, H. S., Jermann, P., & Dillenbourg, P. (2016, April). A gaze-based learning analytics model: in-video visual feedback to improve learner's attention in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 417–421). Edinburgh, Scotland: ACM.

Sharma, K., Jermann, P., & Dillenbourg, P. (2015). *Identifying styles and paths toward success in MOOCs*. Madrid, Spain: International Educational Data Mining Society.

Sharma, K., Pappas, I., Papavlasopoulou, S., & Giannakos, M. (2019). Towards automatic and pervasive physiological sensing of collaborative learning. In *Proceedings of 13th International Conference on Computer Supported Collaborative Learning* (pp. 684–687). Lyon, France: International Society of the Learning Sciences (ISLS).

Smith, C., King, B., & Gonzalez, D. (2016). Using multimodal learning analytics to identify patterns of interactions in a body-based mathematics activity. *Journal of Interactive Learning Research*, *27*(4), 355–379.

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, *34*(4), 366–377.

Stewart, A. E., Keirn, Z. A., & D'Mello, S. K. (2018, October). Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In *Proceedings of the 2018 on International Conference on Multimodal Interaction* (pp. 21–30). Boulder, CO: ACM.

Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology*, *26*(6), 833–839.

Vuori, T., Olkkonen, M., Pölönen, M., Siren, A., & Häkkinen, J. (2004, October). Can eye movements be quantitatively applied to image quality studies? In *Proceedings of the Third Nordic Conference on Human–Computer Interaction* (pp. 335–338). Tampere, Finland: ACM.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer -based tests. *Applied Measurement in Education*, *16*, 163–183.

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The Effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183–192. https://doi.org/10.1080/08957347.2019.1577248

Worden, M. S., Foxe, J. J., Wang, N., & Simpson, G. V. (2000). Anticipatory biasing of visuospatial attention indexed by retinotopically specific-band electroencephalography increases over occipital cortex. *Journal of Neuroscience*, *20*(RC63), 1–6.

Worsley, M. (2014, November). Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviours. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 1–4). ACM.

Worsley, M. (2018, March). (Dis) engagement matters: identifying efficacious learning practices with multimodal learning analytics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 365–369). Sydney, Australia: ACM.

Worsley, M., & Blikstein, P. (2014, November). Deciphering the practices and affordances of different reasoning strategies through multimodal learning analytics. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 21–27). Istanbul, Turkey: ACM.

Wright, R. A., & Kirby, L. D. (2001). Effort determination of cardiovascular response: An integrative analysis with applications in social psychology. *Advances in Experimental Social Psychology*, *33*, 255–307.

Yen, C. H., Chen, I. C., Lai, Su C, & Chuang, Y. R. (2015). An analytics-based approach to managing cognitive load by using log data of learning management systems and footprints of social media. *Educational Technology & Society*, *18*(4), 141–158.

Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., Lai, K. R. (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning*, *34*(4), 358–365.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.