Inferring Students' Self-Assessed Concentration Levels in Daily Life Using Biosignal Data From Wearables

Södergård, Caj; Laakko, Timo

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2023.3260061

Published: 01/01/2023

Document Version
Publisher's final version

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 February 2023, accepted 28 February 2023, date of publication 21 March 2023, date of current version 29 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260061

RESEARCH ARTICLE

Inferring Students' Self-Assessed Concentration Levels in Daily Life Using Biosignal Data From Wearables

CAJ SÖDERGÅRD[1] AND TIMO LAAKKO[2]
[1]NextAI, 02320 Espoo, Finland
[2]VTT Technical Research Centre of Finland Ltd., 02044 Espoo, Finland

Corresponding author: Caj Södergård (caj.sodergard@nextai.fi)

This work was supported in part by the Project "Adult Education and Corporate Staff Training (ADECO)" within the Digile Research Program in Digital Services mainly funded by the Finnish Funding Agency for Innovation TEKES (currently Business Finland), and in part by the Finnish Center for Artificial Intelligence.

This work involved human subjects or animals in its research. All ethical and experimental procedures and protocols were performed in line with the Finnish Personal Data Act. (1999/523).

ABSTRACT The ability to concentrate well is an important determinant of students' learning outcomes but remains poorly understood. In this work we investigated whether there exists a mapping between students' biosignals and perceived concentration levels. If we succeed in this mapping, a wearable can function as a Concentration Tracker, a novel feature that is missing from current wearables. For this, a wearable wristband was used to record students' heart rate, heart rate variability, skin temperature, skin conductivity and acceleration from body changes. Additionally, students self-assessed their concentration levels using a smartphone application. We improved the accuracy by utilizing a big amount of unlabelled biodata from outside the study sessions. Our best boosted regression tree model predicted students' concentration level with only 1.7% NMAE error. The predictions for a user not in the training set were much weaker; the best model, a convolutional neural network, achieved a prediction NMAE error of 30.7%. This implies that the users generated biosignals highly individually. Thus, models are not well transferable from one user to another without rooting them in user-specific data. Contrary to stress research, our results showed that skin conductivity had mostly a negative correlation with students' concentration levels. Also diverging from stress reactions, skin temperature had mainly a positive correlation. Conductivity and temperature were the two dominant predictors. Further, the results suggest that an element of deep, effortless concentration was present in the learning experience of the subjects. Altogether, our work demonstrates that a concentration tracking wearable for improving learning is technically achievable.

INDEX TERMS Affective computing, affective learning, artificial intelligence, biosignals, convolutional neural networks, educational technology, machine learning, mental concentration, semi-supervised learning, wearable sensors.

I. INTRODUCTION

Mental concentration is defined by the American Psychology Association as "the act of bringing together or focusing, as, for example, bringing one's thought processes to bear on a central problem or subject" [1]. A similar definition is in [2]. The ability to concentrate during studies has a well-documented positive impact on learning outcomes [3] and plays a central role in helping students develop other important life-skills such as critical thinking [4]. Students themselves recognize the value of being able to focus, yet many feel that they lack in this quality. For instance, a recent study showed that 80% - 90% of students procrastinate consistently and 50% problematically; 95% of them wish to

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

reduce it [5]. Despite the importance of the topic, methods to predict how concentrated a person feels him or herself in various situations, are to a large extent lacking. This paper aims to fill this gap. The focus of this paper is to develop a model to estimate how well a student is concentrating based on his or her biosignals. This model could be used in various applications to help students improve their concentration levels. For example, the model could enable a software to track students' biosignals and learn to give them personalised advice on study habits, such as when and where to study for optimal concentration.

So far, researchers have investigated the relationship between physiological signals and concentration mainly within strictly controlled experimental settings [6], and often with equipment that are impractical to use in daily life, such as the electrocardiograph [7]. Our aim, in turn, was to create a framework that could be used to track students' concentration levels in realistic settings. Therefore, we decided to use a multi-sensor wristband to collect the biosignal data even though other equipment may have been more accurate. It follows that our central research question is whether biosignal data collected from a wristband during students' daily life has sufficient predictive power for inferring concentration levels.

In order to investigate the relationship between biosignals and concentration levels, we had to define what exactly is meant by good concentration. Since the aim was to investigate this in a realistic daily life setting, it was not possible to focus on any precise theoretical definition of concentration. Instead, it was left up to the students themselves to define what a good concentration meant to them; that is, we asked them to provide self-assessed ratings of their concentration levels whenever they were studying. Another reason why we decided to investigate self-assessed concentration levels is due to its positive correlation with students' motivation [8] and thus learning outcomes. Furthermore, self-assessed levels of attention have been shown to correlate with objective assessments of attention gathered from the d2 test under pressure [9]. The correlation factor in that study was $r = 0.31$. As attention and concentration are closely related, it can be assumed that subjective and objective concentration levels also correlate. Further, self-assessment of intelligence has in multiple studies been found to correlate moderately with measured intelligence with r varying between 0.25 and 0.46 [10]. This is supporting evidence, as intelligence has been found to moderately correlate with attention [11].

Our data set therefore consisted of students' self-assessed concentration ratings, and the following biosignals measured with a wristband: heart rate, heart rate variability, skin temperature, skin conductivity, and acceleration data. Both Boosted Regression Tree and Convolutional Neural Network based machine learning models were trained on the biosignal data to predict students' concentration during the study sessions. Feature importance analysis and partial dependence plots were used to explore the relationship between the biosignals and students' concentration levels in more detail. 10-fold and user-based cross-validation on fully unseen new user's data were applied to evaluate the performance and generalization ability of the model. The main contributions of this work are:

- We have shown that it is possible to infer students' concentration levels from their biosignals in a realistic daily life setting in which we did not control for when, what and where they studied.
- We found evidence that students' self-assessed concentration levels in real-life settings cannot be fully explained by cognitive stress [12] as they had different biosignal patterns. Good concentration likely reflects a deep effortless state of mind maintained over the study time.
- By examining the relative importance of the different biosignals, we found that students' concentration levels were best predicted by their skin conductivity and skin temperature, and, to a small degree, by heart-rate variability. These results will help to design future studies and technical implementations.
- All the results were achieved using a single multi-sensor wristband to collect the biosignal data rather than a combination of more accurate and expensive medical equipment. This data-gathering set-up can thus hopefully inspire future studies and applications to use similar methodology and is scalable to large user groups.

Remainder of this paper is organized as follows. After summing up related research in Section II, we describe our method in Section III. The performance of the Boosted Regression Tree models, CNN models and other results are covered together with a critical analysis and a validating comparision to current literature in Section IV. Conclusion and ideas for future research are presented in Section V.

II. RELATED RESEARCH AND RESEARCH GAP

Commercial smartwatches and rings now have stress trackers as standard features, see [13]. They estimate stress mostly from heart rate variation, but recently increasingly using electrodermal (EDA) signals. Several wearables have skin temperature sensors, but none of these wearables claim to use temperature for stress evaluation. Empatica [14], which we use in this paper, has all of these sensors, but is not meant for the ordinary consumer.

Building on these sensors, there are numerous applications for relieving stress and anxiety [15] as well as depression [16] through various methods like relaxation, light therapy, and exercise. However, to the best of our knowledge, no wearable yet features mental concentration tracking. There is a clear reason for this; mental concentration is a very complex phenomenon and there are no clear metrics yet developed as we will notice in our review below on concentration in educational settings. Such studies demand tracking concentration in highly varying ''in-the-wild'' conditions, which requires dealing with large variances in input signals, and multiple sources of noise as well as challenging ground truth acquisition.

Our research lies at the intersection of two related areas: affective learning and affective computing. Affective learning focuses on the notion that emotions play a vital role in the learning process [8], [17], [18]. The aim of affective computing, on the other hand, is to utilize computational power to detect and analyse different emotional states in humans. It is imperative to note that in this context our usage of the word emotion differs from its everyday use; it does not only refer to classic emotions such as happiness or sadness but rather is used to capture the entire mental state of a person, also known as affect. In fact, research in affective computing and learning has identified that more abstract emotions such as curiosity, focus, flow, anxiety, and boredom are much more relevant to learning than the six basic emotions of Ekman (anger, fear, disgust, joy, surprise and sadness) [19].

Despite recent developments in affective computing, detecting affect during learning remains a tricky task. One major reason is that the construct of different emotions remains unclear. For the most part, affect detection has relied on classic emotional theories from psychology [20]. The two most important ones in the context of learning have been to detect emotions through physical expressions and behaviours (e.g., facial expressions), and through embodiments [20] such as changes in physiology (e.g., biosignals such as heart rate). Based on these two approaches, numerous systems have been developed to detect when a learner is, for instance, concentrated, bored, or confused.

A. THEORIES AND TESTS OF CONCENTRATION IN EDUCATION

In most previous concentration studies, the subjects have been studied in attention tests, where the subject solve specific tasks in laboratory settings. These tests are common in sports sciences, e.g. golf [21] and basketball [22]. The subjects are typically monitored both regarding test performance and biophysical reactions, like heart rate and brain waves. The tests are well documented, e.g., the Attentional Capacity Test (ACT) [23], Continuous Performance Tests (CPTs) [24], Conjunction search [25], and the Letter Cancellation Task [26]. Also, simple tasks, like typing speed [27] or reading and writing certain texts, have been used in testing [28]. However, these task-related tests are only applicable to a defined learning environment and not to the many varieties of everyday learning, which are studied in this paper. In addition, they resemble IQ tests in that they mainly measure focused effortful attention, i.e. cognitive load, which however is only one form of concentration.

Psychologists have developed several theories of how concentration arises. For example, Salomon [29] introduced the amount of invested mental effort (AIME) that reflects a voluntary allocation of effort, whereby people will invest greater effort in processing complex stimuli that cannot easily be accounted for by their existing mental schema. This concept is closely related to cognitive engagement, which depicts the motivation to acquire new ideas and skills [7]. A high degree of concentration is also reminiscent of the concept of flow, a state of positive and full immersion in an activity [30], [31], [32] that results from an appropriate balance between the challenge of a task and the skills of the practitioner. This deep effortless concentration on the activity one is engaged in is essential in everyday learning. Opposite to effortful attention, which typically is measured through objective performance tests described above, flow is best estimated using subjective self-reporting. However, the subjective and objective measurements are related [32]. Achieving a flow state is desirable as it correlates positively with performance measures in various pursuits such as writing and sports [33]. The physiology of flow has been investigated to some degree. Manzano et al. [34] studied piano players and assessed flow in three dimensions: challenge-skill balance, concentration, and autotelic. The authors found a positive correlation between the flow factors and the following variables: increased hr, reduced hrv, decreased respiratory depth and increased activity of the facial muscles. Keller et al. [35] arrived at similar results by observing test subjects in answering quiz questions on a computer as well as playing the game of Tetris. The difficulty of the tasks ranged between 'boredom', 'appropriately challenging' and 'overload'. Their results indicate that flow experiences combine subjectively positive experiences resulting from an appropriate skill-demand balance of the task, as well as physiological elements that reflect the tension and mental load (higher cortisol levels, lower hrv). On the other hand, Zheng and Spires [36] found that game-based learning introduced a flow experience, but it did not predict learning gains. Mansfield et al. [37] researched the relationship between flow and a similar concept called coherence. Coherence is defined physiologically by a smooth sine-wave hrv pattern and qualitatively as a harmonious state between emotional, cognitive and physical systems, which has been linked to improved performance [37]. In their study, the authors induced flow and coherence through different types of questionnaires and video games. However, their finding suggests that flow and coherence are independent of each other. These results underline the complex nature of flow; more research is needed before we can fully understand this concept and how it could be measured and applied to maximize performance and learning gains.

Lee et al. [38] and Son et al. [27] use video input (i.e., webcams) in the concentration evaluation. Video analysing can estimate, e.g., the effects of body movement on concentration. The benefits of pure video-based concentration estimation include that it is easy to use in the online evaluation. However, the evaluation ignores several significant factors that can be detected by biosignals, e.g. skin temperature and electrodermal signals.

A few studies have combined self-reporting concentration with measuring biosignals. One is Lokare and Netak's study [39], which uses four self-scored concentration levels from various predefined tasks (e.g., reading, listening, performing calculations, browsing, relaxing) and measures corresponding EEG signals. The prediction accuracy from

the best Machine Learning model is, however, rather modest (71%). In addition, the use of EEG is not possible in current consumer wearables like smartwatches and rings. A further limitation is that the data in the study comes from only one subject. Another example is Basterrech and Krömer's study [28], where the subjects distinguished between the mental states of high concentration and relaxation. However, there were only four subjects and the activities were specified to reading and writing a certain text in a laboratory setting.

B. DETECTING AFFECT FROM BEHAVIOURS DURING LEARNING

In recent years, log files from interactive learning environments (ILE) and tutorial systems (ITS) have provided researchers with useful data on how students behave during the learning process. Mavrikis [40], for instance, used machine-learning algorithms to predict how well students answer questions in an ILE, based on characteristics such as effort and confidence. Similarly, Baker et al. [41] used several machine-learning algorithms, such as decision trees, to detect students' affect, including concentration, confusion, frustration and boredom, based on how they interacted with an ITS. ITSs have also proven useful for detecting personal learning styles in self-regulated study environments [42], [43].

Most studies based on ILEs and ITSs characterize the learning process with variables such as the time it took to read a text and its difficulty. For instance, Mills and D'Mello [44] used this type of variables, along with students' answers to the question of whether their mind was wondering, as inputs into a supervised machine learning model. This model learnt to detect lapses in students' concentration while they were reading with 20 percentage points above chance. Hershkovitz and Nachmias [45] on the other hand used a hierarchical clustering algorithm to track the motivation of online learners and found that engagement was best predicted by the amount of time a student spent on a task and the average session length.

A weakness of the ITS and ILE based studies discussed above is that they try to infer student emotion from a discrete and pre-defined set of variables and are only applicable in the defined learning environment. To overcome this, many studies have focused on more holistic, observable, behavioural patterns such as students' facial expressions. Grafsgaard et al. [46], for instance, showed how computer software can analyse students' level of engagement, frustration, and how well they learned, from their facial movements. Similarly, Lewis et al. [47] showed that students' self-reported engagement and frustration after tutorial sessions could be predicted from their facial reactions in response to different types of questions asked by tutors during the sessions. In addition to facial expressions, affective information can also be extracted from students' postural changes [8]. As an example, D'Mello and Graesser [18] collected data from a pressure-sensing chair and fed this into a machine learning algorithm. The final model was able to identify boredom, engagement, flow, confusion, frustration and delight from students' body movements with accuracy ranging between 70% and 80%.

A limitation of the above works is that they require special equipment. Their applicability to measuring biosignals throughout students' daily life is hence limited. Typically, investigations that attempt to capture students' affect in an authentic study environment use the experience sampling method [30], in which data gathering occurs during the respondents' daily activities, for instance, regulated by a beeper alert [48]. However, this setup only allows the tracking of students' self-reported emotions but provides no way to track any related behaviours, such as physiological signals. To improve on this limitation, we conjecture that recent advancements in smartphones and wearables, and particularly their ability to collect biosignals, make them an attractive alternative in collecting data on students in realistic environments.

C. BRIDGING THE RESEARCH GAP

A concentration tracking wearable suited for students requires an understanding of concentration sensing that is currently lacking. As noted in the literature review above, most current measurements of performance and cognitive load from biosignals do not cover the totality of daily studies, which include many learning modes – attending lectures, doing group work, reading, writing, doing calculations etc. – in varying environments like at school, in the library and at home. In addition, these learning sessions last significantly longer - 0.5–2 hours – than in the typical attention tests. Therefore, in daily life studies, we need to measure sustained concentration, which is long enough to complete a task. In addition, effortless concentration - flow - is an essential part of studying as discussed above. As said, flow is better caught through subjective reporting than through classical attention tests. Another shortcoming in current approaches, as well noted above, is that they use special laboratory equipment, like EEG sensors or electrographs [7] for gathering biosignals. This is impractical in everyday life and not suiting for our aim of designing a wearable Concentration Tracker. To bridge this research gap, we selected to investigate through probing, how the students themselves perceive their concentration in various conditions. Their biosignals were at the same time measured with a multisensor wristband. Contrary to controlled laboratory conditions, the subjects were allowed to perform whatever form of study they wanted. In addition, we had a more representative set of subjects with 16 students of both genders than in many studies in the literature. Our method, based on signals from a wrist-worn wearable, allowed us to determine from the model what are the most important biophysical constituents in concentration as perceived by the subjects as well as how these constituents co-variate with levels of concentration. To our knowledge, this has not been done before. The results have an important bearing on enhancing education because they

enable a real-time concentration sensor to show under which conditions a student is expected to experience optimal concentration. In addition, e-learning systems can be made more adaptable, by presenting learning content according to the concentration level of the student.

D. RESEARCH AIMS

Based on bridging the research gap above, we formulate the following research aims:

1) To accurately predict how the students themselves perceive their concentration in various daily life learning conditions using biosignals from commonly available wearables. This aims to show that it is feasible to integrate concentration tracking to future wearables.
2) Determine which biosignals are most influencing the prediction accuracy. This result can guide the prioritization of signals to include in the implementation of the concentration tracker products.
3) Show that prediction accuracy increases significantly when non-labelled biosignals outside of study sessions are utilized. These are gathered in enormous amounts in the wearables.
4) To provide insights, whether deep, effortless concentration is present in the learning experience of the subjects.

III. METHOD

16 students (10 male, 6 female) from Haaga-Helia University of Applied Sciences in Helsinki took part in our experiment over a two-week period. The median age was 24 years ranging between 21 and 34 years. All participants signed an agreement whereby they gave their consent for us to collect personal data during the experiment in accordance with the Finnish Personal Data Act (1999/523). Over a two-week period, we collected data on their biosignals as well as their self-assessed concentration levels from study sessions.

Fig. 1 outlines the overall framework developed to evaluate the relationship between students' biosignals and their self-assessed concentration levels and to online predict concentration based on the input biosignal measurements. It also provides references to the respective sections of this paper.

A. MEASURING STUDY CONCENTRATION

The students used Android smartphones to manually record their self-assessed concentration levels during study sessions. These study sessions took place throughout the students' normal daily lives. We did not impose any constraints on when, how or what material the students studied. The only requirement was that the students were studying towards an upcoming exam in their typical study environment. In order to help them do this effectively, we developed the Concentration Sampler Application, which allows one to rate their concentration level on a scale from 1, very low, to 5, very high.

The user interface of the Concentration Sampler is shown in Fig. 2. The student launches the application on the

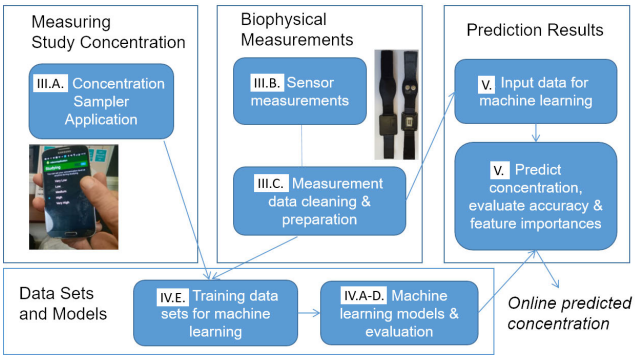


FIGURE 1. The layout of the technical work performed in our study. The blue boxes refer to Sections in this paper. The arrows depict the data flow from the measurements to training data for machine learning, models and evaluations ending up in concentration predictions. The training data originates both from the student's self-assessed concentration levels (Measuring Study Concentration) and from biosignals from multi-sensor wristbands (Biophysical Measurements). Trained machine learning models are used to online predict (Prediction Results) a student's concentration (based on the input measurement data) or to evaluate the accuracy or biosignal/feature importances of the models.

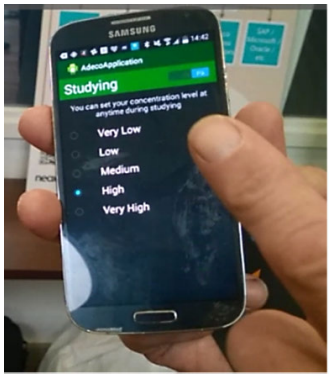


FIGURE 2. The student rates with a smartphone application the felt concentration on a Likert Scale (1...5), when studying.

smartphone at the beginning of the study session, and selects from the touch screen one of the 5 concentration levels on a Likert scale once or several times during the study session. The session id as well as user- and time-labelled ratings are then automatically uploaded from the application to our cloud database as shown in Fig. 5.

Students were asked to use the application to update their concentration rating whenever they felt it changed during a study session. Each study session could therefore potentially consist of several sub-sessions of different concentration levels. Any continuous period with a constant concentration rating was considered a single observation in our data set. We did however require all the study sessions to be at least two minutes long as this eliminated instances in which a student had accidentally launched the concentration rating application on their phone for a short amount of time. An average study session, with a continuous concentration rating, lasted 44 minutes. In total, we ended up with concentration ratings for 130 study sessions. The concentration levels (1 to 5) given by the users in the sessions were distributed as in Fig 3. The total number of study sessions and biodata samples are given

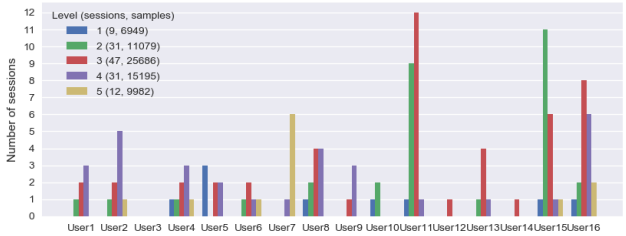


FIGURE 3. The rating of concentration of the 130 sessions of 16 students. The total number of study sessions and biodata samples are given for each rating level. Notice that we have biodata for all users including User 3.

TABLE 1. Biosignal measurement with Empatica E3; *acc* stands for acceleration, *hr* for heart rate, *eda* for electrodermal activity and *st* for skin temperature.

Variable	Measurement details
acc	Three-dimensional accelerometer with sampling frequency 32 Hz.
hr	Photoplethysmography sensor measures light reflected from oxyhemoglobin. Beat detection algorithm by E3 calculates HR at 64 Hz.
eda	Two electrodes at wrist measure electrical conductance, measured on scale 0 µS-100 µS, resolution of 900 pS Frequency of 4Hz.
st	Recorded with optical infrared thermometer with accuracy of 0.2 °C.

FIGURE 4. Empatica E3 wristband and biosignals. These photos depict the multiple sensors of the E3 wristband and a screenshot of raw biosignals for one student. EDA stands for Electrodermal Activity, HR for Heart Rate and ST for Skin Temperature. The graph was created by uploading the raw data to Empatica web portal.

for each rating level. The diagram shows that all the 5 rating levels are fairly well represented both in the sessions and the biodata, however with a clear emphasis on levels 2-4. We can see that the rating data is scarce for several users, even missing completely for user 3. This scarcity of ground truth data underlines the need to analyze all users as a group and avoid user-specific prediction models.

B. MEASURING BIOSIGNALS

We collected four raw biosignals using an Empatica E3 [14] wristband: 3-axis-acceleration, heart rate (*hr*), electrodermal activity (*eda*), and skin temperature (*st*) (see Table 1 for more detail and [14] for full description). The students wore the wristband on the less dominant hand on the ventral area of the wrist (Fig. 4). The biosignals for each person were recorded over the entire experimental period, rather than only during study sessions. This was done so that all the signals could later be standardized and modelled with a maximum of each individual user's own biodata. The longest continuous measurement session for one user lasted 36 h 45 minutes, which fits into the 38 h battery time of E3 [14].

In Fig. 6, the raw median skin conductivity and temperature biosignals of the 16 students are shown. For *eda*, most students show significant signal variations above the median, whereas the temperature has smaller variations, however with clear errors towards lower temperatures. Altogether, we collected 76 MB of biodata from the 16 users.

Fig. 5 shows the experimental setup of the entire data-gathering system. The user has a smartphone for concentration ratings and a wristband for biosignals measurements.

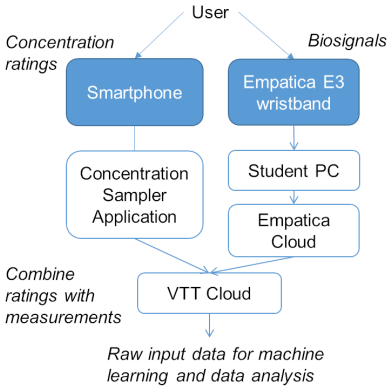


FIGURE 5. We collected students' self-assessed concentration ratings and biosignals and transferred the combined raw data to our internal database for machine learning and data analysis.

Both the ratings and measurement data form the input for further processing, combining, and storing in the VTT cloud (which provides raw input data for machine learning). A few times during the experiment, students attached the wristbands to their PCs to charge them and upload the collected data to the Empatica Connect cloud server. From there it was transferred to our research institution, VTT's, cloud in a time-stamped format. We built our data gathering system around the Open Shift cloud platform from Redhat and the open-source database Postgres. After data gathering, we formed 5-second samples of all biosignals by averaging the measurements within this time window. This resulted in

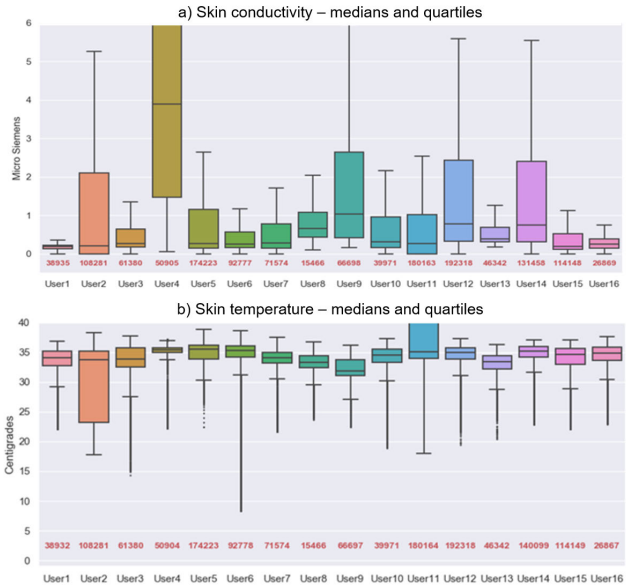


FIGURE 6. Skin conductivity (*eda*) (a) and skin temperature (*st*) (b) measurements for the 16 students. Medians, quartiles, and whiskers are shown and in addition outliers for low temperatures. The numbers of the 5-second samples are in red. The average measurement time for *eda* is 122,5 h and for *st* 120 h. For skin temperature, User 11 is clipped, because of clear misreadings. These are raw data before data cleaning and standardization.

an array, where each row expressed the biosignal values and ratings at that instant.

C. DATA CLEANING AND PREPARATION

Even if the measurement data contained plenty of outliers, as can be seen from Fig. 6, we only filtered away observations, where the skin temperature was measured to over 40 degrees Celsius and where the user clearly is moving, not lose information. This clipping of high-temperature artifacts reduced the amount of data by 4.3%. We sought to eliminate situations, where the student was moving and therefore most probably not studying, by inspecting the variance *acc_var* of the resultant three-axis acceleration vector during and outside of study sessions. We concluded that the test person is lying down or sitting when the *acc_var* is under the threshold of 1.5 ms^2 (see Fig. 7). This reduced the amount of data by 1.5%.

From the measurement data, we computed a set of additional features. The raw *hr* biosignals were used to derive heart rate variability. It was calculated as a time series of the Root Mean Square of Successive Normal to Normal Interval Differences (*rmssd*) over 30 s sliding windows of *hr*, as proposed by Nussinovitch et al. [49]:

$$rmssd_t = \left(\frac{1}{N-2} \sum_{n=2}^{N} [I(n) - I(n-1)]^2 \right)^{\frac{1}{2}} \quad (1)$$

where $I(n)$ is the n-th normal-to-normal interval, and N is the total number of normal-to-normal intervals in the current 30 s window at time t. To avoid problems caused by missing *hr* data, we required that each 30 seconds interval contained

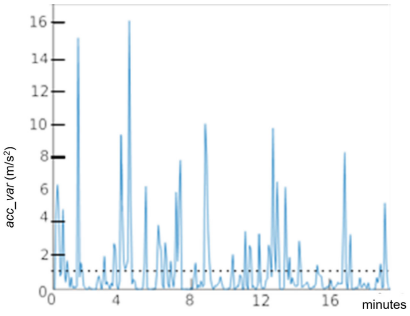


FIGURE 7. Example of acceleration variance (*acc_var*) data. This image shows a snapshot of acceleration variance for one test subject. The horizontal dotted line depicts the threshold above which the person was moving too much to be studying. The snapshot used here purposefully depicts a period during which the person was moving a lot at times, and stayed still at other times.

at least 10 seconds worth of *hr* samples, otherwise we skipped the interval. We calculated the *sdnn* (Standard Deviation of Normal-to-Normal Intervals) similarly. We also derived another feature - Body movement frequency (*bm*) - as the centroid position on the frequency axis. A higher *bm* number corresponds to higher frequency body movements, or more precisely, hand movements.

$$\begin{aligned} F_t(u) &= \mathrm{DCT}(acc\_var(t)) \;\; ; u = 0, 1, \ldots N-1 \\ &= \left(\frac{2}{N}\right)^{\frac{1}{2}} \sum_{i=0}^{N-1} [\Lambda(t) \cos(\frac{\pi u}{2N}(2i+1))] acc\_var(t-i) \\ \Lambda(i) &= \begin{cases} \frac{1}{\sqrt{2}} & \text{for } i = 0 \\ 1 & \text{otherwise} \end{cases} \\ bm &= k/N \text{ for } k \text{ such that } \sum_{u=0}^{k-1} F_t(u) = 0.5 \sum_{u=0}^{N-1} F_t(u) \end{aligned} \quad (2)$$

where $acc\_var(t)$ is the variance of the three-dimensional xyz-acceleration vector calculated at time t, $F_t(u)$ is frequency component u at time t and DCT stands for discrete cosine transformation. We computed $F_t(u)$ and *bm* over a sliding window of $N = 32$ samples of 5 seconds each = 160 s. In addition, we computed a frequency measure *eda_freq* from *eda* similarly as we calculated *bm* from *acc_var*.

Previous works have shown that standardized biosignals offer higher predictive power than non-standardized ones [7]. This makes sense as the relative range and behavior of biosignals is unique to everyone. Each biosignal time series $X_{i,j}^{(t)}$ was therefore standardized, separately for every student, by subtracting the median and dividing by the median absolute deviation:

$$Z_{i,j}^{(t)} = \frac{X_{i,j}^{(t)} - \mathrm{median}(X_{i,j})}{\mathrm{MAD}(X_{i,j})} \quad (3)$$

(3) is calculated for all biosignals i of each student j. This is a common robust standardization method and is appropriate here as we had some very large outliers in the raw data due to occasional problems with the measuring instrument.

The process above resulted in a time-series dataset of within-subject standardized biosignal variables. In particular, the above standardization was calculated using all the recorded data, rather than just data from study sessions. This was done to make the standardization more robust and also was intuitively appealing as we could then interpret the standardized biosignals as deviations from the students' typical baselines. The final step in our pre-processing was to match concentration ratings to the biosignals; for each study session, we had a time series of standardized biosignals.

We consider a *session* as a continuous measurement period of biosignals of a user. We sample the biosignals over a 5-second period by using the average value. These data make up the *Baseline set* with over 1,4 million rows (see Table 2). Note that all biosignals nor ratings need not be always available or recorded during the session. During a session, the user could rate their concentration several times.

Because of the requirement of a 5-second continuous measurement and misreadings by the wristband, not all rows in the dataset contain complete measurements of a specific time.

IV. MACHINE LEARNING ALGORITHMS AND DATA SETS USED

In this study, we used Boosted Regression Trees (BRT) and Convolutional Neural Network (CNN) methods to classify students' study-session concentration based on their biosignals.

A. BOOSTED REGRESSION TREES

BRT is a popular machine learning model [50], [51], [52], [53]. The algorithm falls under the category of ensemble models since the final model is a combination of several individual regression trees. Each individual regression tree tries to classify the dependent variable using binary splits based on chosen independent variables. The algorithm determines the split points and variables so that the difference between real and predicted outcomes is minimized [54]. Using just a single decision tree is, however, very unstable; a small change in data could result in a completely new tree. Hence, rather than just using a single tree, BRT joins several trees together using boosting. This is a stage-wise process in which at each step a new tree is added so that the model's error is reduced. The very first tree is thus the one that best describes the entire data set, the second tree is the one that best estimates the residual that remains after fitting the first tree, and so on [52].

We chose to use BRT as it has several advantages over traditional regression techniques. First, subsequent splits in the trees implicitly take account of variable interactions, and as we are fitting many such trees, the model can handle complex nonlinearities [54]. Second, since only split points matter, the model is insensitive to outliers. Third, compared to other machine learning classifiers such as artificial neural networks, BRT results are easier to interpret. In the implementation, we used the XGBoost library [55].

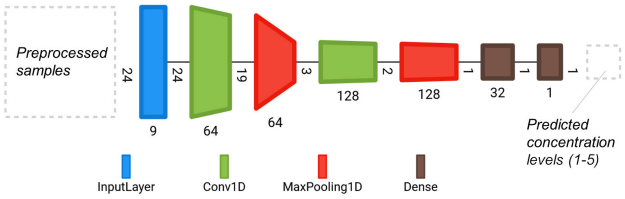


FIGURE 8. CNN network architecture (visualized with the help of [64]) including model layers, and their input and output dimensions.

B. CONVOLUTIONAL NEURAL NETWORKS

A significant property of deep neural networks is the learning of high-level features in the hidden layers. This reduces the demand of feature engineering and input data handling. A convolutional neural network (CNN) [56] is a deep neural network with one or more convolutional layers. CNNs have shown very successful results in different kinds of applications [57], [58], [59], [60], [61], [62]; most commonly, CNNs are used in image processing where input image data is given as a 2D grid of pixels.

In our case, the input data is given as time-series data, which can be represented as a 1D grid of samples at regular time intervals [58]. For training the CNN model, the data were reshaped into samples, each of which consisted of 24 consecutive 5-second time-steps (i.e., two minutes) of measurements within user sessions.

We used the following architecture for the CNN model (see Fig 8). First, in the network, there is an input layer, which takes samples with the dimension 24×9 (24 time-steps and 9 features) and provides input to the first 1D convolution layer (having kernel size 6) followed by a 1D pooling layer (pooling size 6). After this, there is the second 1D convolution (kernel size 2) and pooling (size 2) layers. Finally, we have two dense layers, of which the last layer outputs the prediction. The activation function ReLU (Rectified Linear Unit) was used with the convolution layers and the first dense layer. To compile the network, we used an RMSprop (Root Mean Square Propagation) optimizer, which maintains a moving average of the square of gradients and uses momentum. The prediction output is then a floating point number (i.e., concentration value). In the implementation, we used Tensorflow with Keras [63].

C. EVALUATION OF THE MODELS

We used 10-fold cross-validation (10-CV) for evaluation. The data samples were first picked randomly from the various users in order to evaluate all users equally. The dataset was then split into 10 equal folds, of which 9 were used to train the model and the one remaining fold was used as a test fold to evaluate how well the model was able to predict concentration level. Each model's performance was judged as its average over the 10 test folds.

As a check of how well the models can be generalized to a totally unseen user, we also applied a measure that we called

User-based cross-validation (U-CV). Here, we evaluated the models for each user separately. The models are trained purely with other users data, i.e. cross-validation with fully unseen data. For each user, we take the current user's data for testing and the data from other users for training. Finally, the average is computed from the obtained metrics values. Our method is similar to the leave-one-subject-out (LOSO) method [65], [66], where each subject in turn is left out for testing. However, we further restrict the testing set to be fully complete data (the *Ref Set* in our case, cf. Table 2) and require that the testing data in the evaluation is not used for filling in incomplete training data. These restrictions ensure that data of the test and training sets are not derived from each other. This is because we applied several methods to complete the missing data values.

D. METRICS USED IN THE EVALUATIONS

We used the following metrics to compare and evaluate the results obtained with different models and data sets:

- RMSE (Root Mean Squared Error).
- MAE (Mean Absolute Error).
- NMAE% (Normalized Mean Absolute Error %).
- Accuracy.

In computing the above metrics, the self-assessed ratings (=targets) and predicted concentration values are floating-point numbers in the range 1-5. NMAE% gives an error estimate complementing the MAE metric; it is defined as MAE divided by the average target value, in percentage. If our task is cast not as a regression problem, but as a classification problem, accuracy is the ratio of the correct and the total number of predictions. Since the predicted values are floating-point numbers, we interpret a predicted value to be correct if it differs from the target less than 0.5. Figure 2 shows that a ''naive'' classification policy of always predicting concentration level 3 gives the baseline accuracy of 0.37 if based on the number of samples with ratings. The random choice accuracy for the baseline set is 0.24. However, NMAE is a more informative measure than accuracy, because the concentration prediction of, say 3.2, tells more than its rounding to class 3.

E. DATA SETS

The preprocessed original, so-called, baseline data set (hereafter *Baseline set*) contained the columns *timestamp*, *user_id*, *session_id*, *acc_var*, *hr*, *rmssd*, *sdnn*, *st*, *eda*, *eda_freq*, *bm*, *location*, and the target ground truth column *concentration*.

There were several issues to be considered with the *Baseline set*:

- Plenty of biosignal measurement samples (15 - 197 Ksamples per user), but massively missing bio-signal data values in the samples; only 0.27% of the samples were complete with all biodata and labels (=concentration ratings) present.
- Only 4.8% of the complete biosignal samples had user concentration ratings (=labelled data).
- Some users had much more biosignal measurement data than others, and distributions of rating values (1-5) were heavily unbalanced as seen in Fig. 3.
- Data originated from quite a low number of test users (16 users).

In order to cope with the above issues, we created several data sets. First, we constructed a reference data set (''*Ref set*'') from the *Baseline set* by filtering out the rows containing empty values. In addition, we built a larger *Labelled set* holding all rows that have concentration ratings ie. labels, but that might have missing biosignal values.

Then, we created additional data sets by filling and augmenting incomplete data samples and ratings within the *Labelled* and *Baseline sets*. We also applied resampling to create more balanced data sets. All these sets were used in 10-CV evaluations, and in training the models for U-CV validation. In U-CV validation *Ref set* was used as a basis for testing data. Note that all sets can be automatically generated from the *Baseline set*.

1) FILLING MISSING VALUES

We used two alternative methods to fill the missing data values in the *Labelled set*: a) interpolation and b) machine-learning based augmentations.

a) By using linear interpolation the missing (= NaN valued) biosignals in the data set are filled column by column within user session time series data. Further, NaN values from session start and end are assigned with the nearest non-NaN value. The resulting data set: ''*Filled set 1*''.

b) By using the trained BRT model to fill missing values in two steps. First, we filled missing *hr* values as a function of *acc_var*, *st*, and *eda*. The model is trained by the rows including complete *acc_var*, *st*, *eda* and *hr* values. Then, missing *hr* values are predicted by the model. Second, we filled the missing *rmssd*, *sdnn* and *bm* values each separately as a function of *acc_var*, *st*, *eda* and *hr*. Each model hyper-parameters were optimized using grid search and 10-CV. The resulting data set: ''*Filled set 2*''.

In addition, the missing *location* values were marked as ''other'' (the location value was one of the following: ''at home'', ''in school'', ''other'').

2) PSEUDO-LABELING DATA WITH SEMI-SUPERVISED LEARNING

Semi-supervised learning [67], [68], [69], [70] utilizes the large amount of unlabeled data to improve the model training that otherwise is limited by the small amount of labelled data. Since our *Baseline set* contained only 4.8% labelled data (of which only 0.27% were complete rows), we used semi-supervised learning to complement concentration values.

We adapted pseudo-labelling [62], [68], [71] considering one user's data at a time. First, we trained a model using *Ref set*. In this phase, we used the hyper-parameters of the BRT models [72] for 10-CV. Then, we used the created model to label the unlabelled, but otherwise complete measurement

TABLE 2. Data sets.

Data set	Sample rows	Rows per user (min,median,max)	Concentration (mean, sd)
Baseline	1442450	15395, 69095, 197542	3.15 ± 1.16
Labelled	68891	0, 4033, 8661	3.15 ± 1.16
Ref	3748	0, 101, 1052	2.80 ± 1.22
Resampled 1	5760	360, 360, 360	3.13 ± 0.97
Resampled 2	5760	360, 360, 360	3.26 ± 1.04
Pseudo-Labelled	270021	400, 8858, 53984	3.03 ± 0.71
Filled 1	35172	0, 2238, 5933	3.46 ± 1.19
Filled 2	26267	0, 1544, 5020	3.22 ±1.16

data samples of the *Baseline set*. In addition, we generated labelled training sets for each user for U-CV evaluation to ensure that labelling was done purely with the other users' data. The resulting data: "*Pseudo-Labelled set*".

3) BALANCING DATA SETS BUILDING ON THE LABELED DATA

When balancing data sets [73], we took the biosignal data and concentration ratings of each user equally into account. There are several alternative ways to handle unbalanced datasets with re-sampling. You can make a dataset balanced by either removing samples from over-presented classes (under-sampling) or adding more samples to under-represented classes (over-sampling).

We applied over-sampling to take each user's data equally into account in training the model by using *Ref set* as the basis and the *Pseudo-Labelled* and *Filled sets* as auxiliary sets in combining the result into the balanced sets. We used 30 minutes of measurement data of each user. The balancing proceeded as follows: The aim is to keep user sessions data together and, first, pick only complete data samples from *Ref set*. If not available, then use a) *Pseudo-Labelled* or b) *Filled set 1* to provide additional samples to the resulting set. The resulting data sets: a) "*Resampled set 1*" b) "*Resampled set 2*", respectively.

In addition, we applied sample weighting on the data sets: pseudo-labelled sample rows and filled sample rows were indicated with "labelled" and "filled" flags. This information was used in the training phase to give less weight to the labelled and filled samples than to the original ground truth data (*Ref set*). When training the BRT or CNN models, we used the sample weight 0.5 for weighting the loss function.

Table 2 summarizes the data sets size, row counts per user, the number of user sessions and concentration rating mean and standard deviation values.

V. RESULTS

We trained the models and predicted the concentrations by combining input data over 2 minutes (i.e. 24 time-steps) sample periods. The prediction is fast; it takes only 1 ms for BRT and 37 ms for CNN on a standard quad-core laptop with a CPU clock rate of 1.90 GHz for the sample.

Table 3 reports the validation results for the best-performing BRT and CNN models by using the 10-CV evaluations. As noted above in Section IV-D, the baseline accuracy to compare with is the random accuracy 0.24 and "naive" accuracy 0.37.

The models were optimized separately for 10-CV and U-CV with different hyper-parameters using grid search. For the BRT model, the parameters "number of iterations" (i.e., boosting stages) and "maximum depth" were set to values 1000 and 6 in 10-CV, and 100 and 1 in U-CV evaluations, respectively, while the learning rate was set to 0.1. For CNN, we used the number of epochs 100 and the batch size 64 (i.e., the number of samples per iteration) in the evaluations. In all optimizations, the *Ref set* was used.

Table 3 shows very good accuracies using the standard 10-fold Cross Validation (10-CV) measure. BRT achieves an almost perfect score (NMAE=1.7 ± 0.014%) for the *Pseudo-Labelled set*. This corresponds to the following confusion matrix giving the accuracy of 0.999:

$$\begin{pmatrix} 6796 & 32 & 0 & 0 & 0 \\ 9 & 38180 & 15 & 0 & 0 \\ 0 & 67 & 182771 & 4 & 0 \\ 0 & 0 & 16 & 27416 & 0 \\ 0 & 0 & 0 & 4 & 14711 \end{pmatrix}$$

Thus, semi-supervised learning decreased the estimation error from 3.7 to 1.7%. CNN performs here clearly worse with an NMAE=9.9%, (accuracy 0.83) using *Pseudo-Labelled set*. As a comparison, we tried BRT on a dataset, where the concentration ratings were binarised into only two classes (Good/No Good). However, the 10-CV accuracy was then clearly lower: 0.8, with a baseline "naive" accuracy of 0.66.

User-based cross-validation (U-CV) in Table 4 shows, as expected, that the models are highly user-dependent and therefore not adaptable on users of which there are no training data. It shows further, that CNN outperforms BRT for this totally unseen dataset. There, CNN has an NMAE 30.7% (with standard deviation ± 38.5%) for *Resampled set 1*, and corresponds to the following confusion matrix and an accuracy of 0.47:

$$\begin{pmatrix} 0 & 88 & 76 & 1 & 0 \\ 3 & 24 & 28 & 0 & 0 \\ 17 & 83 & 211 & 24 & 0 \\ 0 & 5 & 22 & 14 & 4 \\ 1 & 0 & 9 & 3 & 72 \end{pmatrix}$$

BRT has NMAE of 35,7% for the *Ref set* and corresponding accuracy 0.26 (±0.28). Even if these accuracies are low with a huge variation, it must be noted, that they refer to the case with a new user that the system has not seen before. An approach to handle the high user dependency of the models is to cluster the training data into several clusters based on similarity [74], [75]. We tested clustering so that we clustered the 16 users data into three clusters, and aimed to find the optimal combination of users data by U-CV evaluation. Then, we evaluated new users data with the model trained with best-fitted cluster. In this test, the clustering was shown to decrease the NMAE error using the BRT model about 0.3%, whereas the CNN results were not improved. One reason for the small improvement is likely to be, that BRT by itself efficiently

TABLE 3. Evaluation of BRT and CNN models and various datasets with the 10-CV method. *Ref set* is used in hyperparameter optimizations.

Data set	Method	NMAE%	RMSE	MAE	Accuracy
Ref set	BRT	3.71 ±0.326	0.046 ±0.0126	0.104 ±0.0098	0.960 ±0.0112
	CNN	26.7 ±7,854	1.332 ±1.121	0.743 ±0,261	0.511 ±0.090
Resampled set 1	BRT	2.77 ±0.137	0.025 ±0.0054	0.087 ±0.0041	0.981 ±0.0069
	CNN	19.1 ±6,041	0.804 ±0,491	0.592 ±0,178	0.597 ±0.156
Resampled set 2	BRT	2.24 ±0,234	0.023 ±0,0060	0.073 ±0,0073	0.982 ±0,0083
	CNN	19.7 ±5,709	0.850 ±0,505	0.638 ±0,170	0.525 ±0,107
Pseudo-Labelled set	BRT	1.72 ±0.016	0.006 ±0.0001	0.052 ±0.0005	0.999 ±0.0001
	CNN	9.90 ±0.516	0.189 ±0.234	0.300 ±0.015	0.822 ±0.015
Filled set 1	BRT	2.21 ±0.079	0.021 ±0.0018	0.074 ±0.0024	0.982 ±0.0026
	CNN	17.9 ±2.341	0.672 ±0.163	0.600 ±0.077	0.557 ±0.062
Filled set 2	BRT	6.84 ±0.163	0.126 ±0.0079	0.220 ±0.0046	0.876 ±0.0050
	CNN	19.0 ±2.374	0.694 ±0.170	0.614 ±0.075	0.539 ±0.064

TABLE 4. U-CV validation of BRT and CNN models. *Ref set* is used both in hyperparameter optimization and - opposite to as in 10 CV - also as a test set.

Data set	Method	NMAE%	RMSE	MAE	Accuracy
Ref set	BRT	35.7 ±41.9	1.234 ±1.131	0.937 ±0.477	0.256 ±0.278
	CNN	32.0 ±34.2	0.959 ±0.959	0.769 ±0.446	0.396 ±0.350
Resampled set 1	BRT	37.6 ±46.4	1.462 ±1.750	0.984 ±0.623	0.302 ±0.320
	CNN	30.7 ±38.5	0.798 ±0.704	0.692 ±0.401	0.469 ±0.352
Resampled set 2	BRT	36.7 ±46.5	1.236 ±1.456	0.932 ±0.544	0.236 ±0.286
	CNN	31.1 ±38.1	0.809 ±0.658	0.709 ±0.378	0.443 ±0.320
Pseudo-Labelled set	BRT	36.6 ±46.8	1.408 ±1.855	0.945 ±0.641	0.283 ±0.313
	CNN	36.5 ±46.0	1.089 ±0.892	0.853 ±0.433	0.321 ±0.345
Filled set 1	BRT	39.1 ±48.9	1.565 ±2.06	1.009 ±0.660	0.246 ±0.268
	CNN	37.1 ±41.1	1.235 ±0.918	0.893 ±0.450	0.357 ±0.330
Filled set 2	BRT	36.8 ±44.6	1.520 ±2.22	0.971 ±0.679	0.248 ±0.258
	CNN	37.3 ±38.7	1.283 ±1.06	0.922 ±0.455	0.357 ±0.288

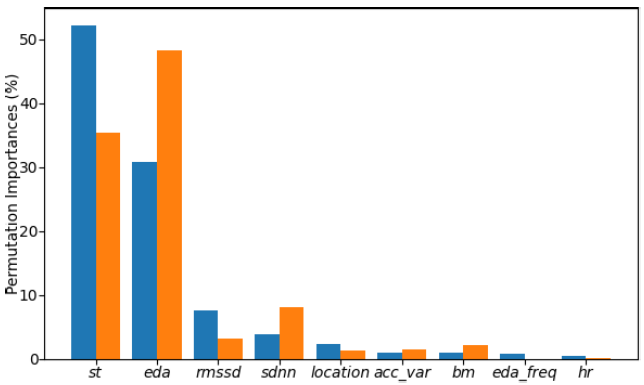


FIGURE 9. Relative feature importance of the different biosignals in the best performing BRT models. The model optimized for 10-CV is shown in blue, and for U-CV evaluations in orange bars.

separates the users from each other and, thus, implicitly builds user-specific sub-models. Therefore, we decided not to do further tests with clustering.

We also assessed the relative importance of the different biosignal variables in the final models. For this, we used the permutation "feature importance" [76] in the Scikit-learn library [77]. Feature importance is defined as the decrease in a model score when values of a single feature are randomly shuffled [76], [77]. Hence, it breaks the relationship between the feature and the target; the score decrease indicates how much the model depends on the feature. Fig. 9 shows feature importance of *st*, *eda*, *rmssd*, *sdnn*, *location*, *acc_var*, *bm*, *eda_freq* and *hr* using two BRT models, the hyper-parameters of which were optimized for 10-CV and U-CV, and the models were trained with *Ref set*. For these models, the relative importance *st*, *eda*, *rmssd* and *sdnn* were (51.6%, 35.3%), (31.2%, 48.3%), (7.7%, 3.2%), and (3.9%, 8.1%), respectively. *location*, *acc_var*, *bm*, *eda_freq* and *hr* made up the small remainder. When considering a practical implementation of a concentration measuring system, the importance of features derived in this work helps to decide if it is worth including certain sensors in the measuring device. It might be motivated to leave out features having only a minor effect on the prediction accuracy. Likewise, you could train alternative models based on what measurement data are available, but at least include the most important features.

In order to investigate the exact shape of the relationship between the different biosignals and the concentration variable, we calculated variable partial dependence plots (PDP) [51]. A PDP depicts how changes in the value of a biosignal contribute to the target concentration value, whilst holding all the other biosignals constant at their mean values. PDPs for the different biosignals are displayed in Fig. 10 using the best performing BRT model, according to the 10-CV evaluation using *Ref set*. Starting from the most important features, skin temperature *st* clearly rises with concentration indicating that higher concentration, contrary to stress, is associated with a more relaxed state. Skin conductivity *eda* behaves much in the same way as *st* with higher concentration going together with lower *eda* (=more relaxation), even if there is a dip in the concentration at 50% *eda*. Interestingly,

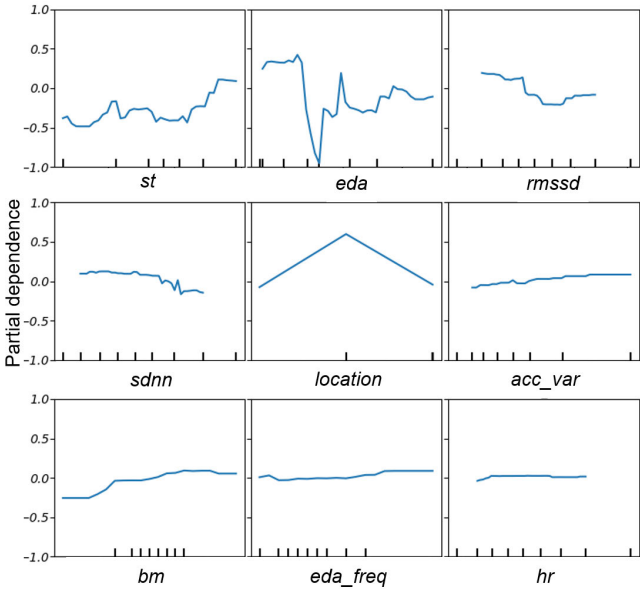


FIGURE 10. Partial dependence plots of the different biosignals using BRT model. The ticks on the X-axis represent data deciles (data shown between percentiles [0.2,0.8]). The Y-axis shows the expected relative contribution on the predicted concentration value.

rmssd and slightly *sdnn* react contrary to *st* and *eda* showing lower values (=less relaxation) at higher concentration. Thus, it seems that heart rate variation measures different aspects of concentration than *st* and *eda*, but as it is contributing much less to the prediction, this does not change the overall pattern of higher concentration implying higher relaxation. For the least important features, *location* values show that at home (middle of the x-axis) the concentration is higher than at school (x-axis left) or other places (x-axis right). *bm* rises clearly with concentration, which can be interpreted so that as studying often means writing on a keyboard, higher concentration means more intense typing. *acc_var*, *eda_freq* and *hr* do not show any clear positive or negative correlation with the concentration.

A. CRITICAL ANALYSIS AND DISCUSSION

In comparison to other research, the most relevant stickyards are the results in [39] and [28], which are reviewed in Section II of this paper. Reference [39] achieved an accuracy of 71% in predicting the concentration with their ML algorithm using four self-scored concentration levels (low, medium-low, medium-high, and high), whereas [28] reached between 71%-83% accuracy in determining if the four subjects are concentrated or not. Both based their predictions of concentration on EEG signals, which we did not use in our work. As mentioned in Section II, using a pressure-sensing chair allowed to identify mental states with accuracies also between 70% and 80% [18]. Tervonen et al. [66] reviewed 14 publications in detecting mental state, stress and emotions and the accuracies varied between 51% and 97%, the average being 80%. They reached themselves 67% accuracy. Thus, our best models (cf. Table 3) achieved better accuracy than any of these. However, the accuracies are not directly comparable in the different studies above since the training data, classification (concentration levels), and evaluations vary a lot.

In addition to high prediction accuracies, our results show which biosignals are most influential in predicting concentration during learning sessions. This allows the system to be further adapted on a case-by-case basis to utilize the most significant biosignal measurements available. We used both 10-CV and U-CV optimized models to predict the biosignal feature importances. Importantly, both methods show that *eda* and skin temperature *st* are the most important biosignals, and in general the feature importances are rather similar in both models (cf. Fig. 9). Furthermore, since the accuracy of the 10-CV optimized model is high, the feature importance prediction based on this model is also highly reliable.

A third advantage is that the trained models can be used for fast online prediction of the user's concentration. The length of the sample period can be adjusted according to the actual study session. With a minimum of 2 minutes sample period of recorded biosignal data, the prediction completes on a standard laptop in 1-37 ms in our example case, cf. Section V.

As noted above (cf. Section II-C "Bridging the research gap"), most previous research does not cover daily studies, which have many learning modes and significantly longer learning sessions than in the typical attention tests. Another shortcoming in current approaches to gathering biosignals, as well noted above, is that they use special laboratory equipment (like EEG sensors or electrographs) that are impractical in everyday life. We approached these shortcomings by investigating the relationship between the students' biosignals and self-assessed concentration levels within complete study sessions. Using 10-CV evaluation, our best BRT model predicted students' concentration level with only 1.7% NMAE error. The results show that the method can accurately predict how the students themselves perceive their concentration in various daily life learning conditions using commonly available wearables. That is the major *strength* of this paper and distinguishes it from comparable work in the literature as elaborated below.

There are several *limitations*. Although we had a large baseline data set, with plenty of biosignal measurements (over 1 million sample rows), there were lots of incomplete data with missing biosignal values, non-labelled samples and unbalanced distributions ("Data sets" in Section IV). Other limitations included a fairly small number of trial users (16). However, we managed to enhance the results significantly by semi-supervised learning and balancing and preprocessing the data sets for machine learning. Using U-CV evaluation, the prediction results were much weaker and the best CNN model achieved a prediction NMAE error of 30.7%. The results imply that the users generated biosignals highly individually. Thus, models are not well transferable from one user to another without rooting them in user-specific data. It remains a future work to elaborate if 10-CV and U-CV

results would converge if more training data from more users are available. When scaling up this kind of system to mass use, there are currently evident limitations that mostly relate to the use of wearables in general:

1) Wearable devices still have challenges in the convenience of use 24 hours a day, especially for non-technical people. One such bottleneck is still a relatively short battery duration. However, the fast technical development and the current wide and growing adoption of smartwatches and sports trackers continuously improve usability.
2) Wearables are heterogeneous and do not always measure the biosignals accurately enough. E.g., too loose contact between the sensors and the skin may introduce noise or even lead to missing data, as seen in this study. Movement and other physical conditions of the user may impact the accuracy. Partially, this can be overcome by signal filtering as exemplified in this study. Online user alerts and guidance are also needed.
3) There is a constant need for expanding the set of biosignals measured to improve the descriptive power. For measuring concentration, EEG is expected to improve prediction accuracy, but there is currently no solution available that is suitable for mass-adoption.
4) There is a risk that the externally stored data from the wearables are leaked. Therefore, privacy has to be cared for. Even if we do not address this limitation in this paper, it has to be handled professionally in a commercial system. The same goes for the rights of the user to his/her own biodata, where the protocols proposed e.g. by the MyData movement [78] are relevant. E.g. the school authorities should typically have access only to the anonymized student data.

The *impact* of this work in real-life scenarios are significant. A Concentration Tracker wearable based on the methods and model laid out in this paper helps the student to develop the ability to concentrate and thereby improve the learning outcomes. As one example, the student is able to more consciously select the places and times when he/she is most concentrated based on feedback from the Concentration Tracker. The organisations providing education can likewise benefit, by getting information about which curriculum, teaching methods, environments, time schedules and other factors have a positive impact on the concentration of the students. Naturally, the privacy of the student data must be cared for. Finally, education research will gain as they get valuable insight into the mechanisms affecting study concentration.

VI. CONCLUSION

We examined through probing, how the students themselves perceive their concentration in various conditions. Their biosignals were at the same time measured with a multisensor wristband. Contrary to controlled laboratory conditions in literature, the subjects were allowed to perform whatever form of study they wanted. In addition, we had a more representative set of subjects with 16 students of both genders than in many studies in the literature. Our method predicted perceived concentration very accurately based on signals from a wrist-worn wearable. It allowed us to determine from the model what are the most important biophysical constituents in concentration as perceived by the subjects as well as how these constituents co-variate with levels of concentration. To our knowledge, this has not been done before. The results have an important bearing on enhancing education, because they enable a novel real-time concentration sensor showing under which conditions a student is expected to experience optimal concentration.

We used the collected data to train BRT and CNN models to predict how well a student had concentrated. We applied semi-supervised learning to complement the largely missing concentration and biosignal values. The best model reached a 10-CV NMAE prediction error of only $1.7 \pm 0.016\%$, where semi-supervised learning more than halved the error. We also evaluated the models with our U-CV method, where the testing data comes from a user contributing with no training data. This depicts the case, where the system handles a new user. The prediction error for the best model was then much bigger - NMAE was $31 \pm 38\%$, which shows the high user dependency of our predictions. When the test set contains even a minor part of data from a user that has contributed to the training set, our best models were able to utilize the user-dependent data almost perfectly, but for totally unseen user data, the models failed to show practical value. Future studies could clarify, how this fruitful mix of general and user-specific data will change with a much higher number of users than in this study.

The results for known users are promising, considering that the study was conducted in an entirely realistic daily life setting, without any constraints on where, how, and what the students studied. The results were achieved using a state-of-the-art wristband rather than more accurate laboratory equipment that are impractical to use in daily life. As wearable technology improves further, it will be possible to reach similar results with consumer-grade devices. Once the prediction models are trained on the data from the various users, the execution of the prediction is fast and can be performed in real-time either in the device or on the server/cloud side. One application is to utilize a wristband to regulate the type and difficulty of material given to a student. It would also be intriguing to examine how employees' concentration levels correlate with biosignals in an office setting.

Our results showed that for the examined subjects mental concentration behaves predominantly contrary to cognitive stress, i.e., good concentration goes together with a relaxed state of mind. Skin conductivity went down when students' concentration levels raised, while skin temperature went up together with the concentration levels. Even if heart rate variability dropped with better concentration, i.e., behaved similarly as in stressful situations, this did not change the overall pattern of higher concentration implying higher

relaxation. This is because skin temperature and conductivity affect the concentration predictions much more than heart rate variability.

Hence, students' concentration levels in this study are not fully described by cognitive stress, which has been the focus of most previous research. This likely reflects that good concentration can also arise during non-stressful tasks such as reading. Furthermore, it is possible that students gave themselves better concentration ratings whenever they felt happy after a study session. If this were to be the case, it could mean that factors such as the difficulty of the material studied played a major role. Better self-assessed concentration might hence be much closer to the affective concept of flow than to cognitive stress. In future work, we need to improve our understanding of how students' self-assessed concentration links to their school performance.

Naturally, the students' biosignals were also affected by factors unrelated to studying, such as what was happening in their surroundings, which made it harder to extract clean patterns.

A. FUTURE DIRECTIONS

One major future direction is to implement the method into a wearable that works online in real-time or near-to-realtime using local and cloud back-end computation. This *Concentration Tracker* would display to the user the current concentration value preferably in a graphical form in the same way that current smartwatches display your stress level. In addition, you could get the variation over the day, week a.s.o. on the wearable or mobile phone or other terminal connected to the cloud. To be viable, the application software should be downloadable on commercial wearables using their de-facto platforms. It would also be interesting to investigate other applications than learning, for example, alertness while driving cars or operating other technical systems.

Semisupervised learning has already shown promising results in this work. We aim to further research semisupervised learning methods with more training data. It would also be essential to evaluate the effects of *missing biosignal data* on the accuracy of the overall results, based on partial data, where only some of the required biosignals are present. Another methodological development is to use *transfer learning* leveraging pre-trained psycho-physical models and datasets of emotions like stress. This is motivated because emotions have many similarities with mental concentration.

EEG registration would increase the accuracy of concentration estimation. As pointed out several times in this work, EEG cannot be measured with current commercial wearables. However, recent research has shown that it is possible to construct EEG wearables, e.g. attached to the ear [79]. When these devices are available, they would complete the sensor set used in this work.

Finally, this technology could be developed to offer insights to the student and others about the importance of *lifestyle* for achieving good concentration. In particular, the impact of sleep quality, exercise and nutrition is central. As sleep quality and amount of exercise are straightforward to measure from the wearable, the influence of healthy choices on good concentrations can clearly be illustrated from the gathered data.

APPENDIX A SUPPLEMENTARY DATA

The Baseline dataset can be accessed from IEEE Dataport doi: https://dx.doi.org/10.21227/as25-6r07.

ACKNOWLEDGMENT

The authors would like to thank Hermanni Hälvä, Ville Antila, and Jouni Soitinaho for their help with the initial stages of investigation. They are also thankful to Alain Boyer, Kristian Södergård, and Staffan Södergård for their help with the video accompanying this submission.

REFERENCES

[1] *APA Dictionary of Psychology*. Accessed: Nov. 12, 2022. [Online]. Available: https://dictionary.apa.org/concentration

[2] L. M. Brand. (2010). *The Effect of Technology on Attention and Concentration Within the Classroom Context*. [Online]. Available: http://hdl.handle.net/10500/3452

[3] J. T. Guthrie, S. L. Klauda, and A. N. Ho, "Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents," *Reading Res. Quart.*, vol. 48, no. 1, pp. 9–26, Jan. 2013.

[4] R. M. Carini, G. D. Kuh, and S. P. Klein, "Student engagement and student learning: Testing the Linkages," *Res. Higher Educ.*, vol. 47, no. 1, pp. 1–32, Feb. 2006.

[5] S. Piers, "The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure," *Psychol. Bull.*, vol. 133, no. 1, pp. 65–94. 2007.

[6] D. M. Pendleton, M. L. Sakalik, M. L. Moore, and P. D. Tomporowski, "Mental engagement during cognitive and psychomotor tasks: Effects of task type, processing demands, and practice," *Int. J. Psychophysiol.*, vol. 109, pp. 124–131, Nov. 2016, doi: 10.1016/j.ijpsycho.2016.08.012.

[7] J. Pärkkä, M. Ermes, and M. van Gils, "Automatic feature selection and classification of physical and mental load using data from wearable sensors," in *Proc. 10th IEEE Int. Conf. Inf. Technol. Appl. Biomed.*, Nov. 2010, pp. 1–5.

[8] R. W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker, "Affective learning—A manifesto," *BT Technol. J.*, vol. 22, no. 4, pp. 253–269, 2004.

[9] C. Mengelkamp and R. S. Jager, "Self-estimates of attention performance," *Psychol. Sci.*, vol. 49, no. 3, p. 223, 2007.

[10] H. Holling and F. Preckel, "Self-estimates of intelligence—Methodological approaches and gender differences," *Personality Individual Differences*, vol. 38, no. 3, pp. 503–517, Feb. 2005.

[11] L. Schmidt-Atzert, M. Bühner, and P. Enders, "Messen konzentrationstests konzentration?" *Diagnostica*, vol. 52, no. 1, pp. 33–44, Jan. 2006.

[12] Q. Liu, Y. Liu, X. Leng, J. Han, F. Xia, and H. Chen, "Impact of chronic stress on attention control: Evidence from behavioral and event-related potential analyses," *Neurosci. Bull.*, vol. 36, no. 11, pp. 1395–1410, Nov. 2020.

[13] M. Sawh. (2022). *Stress Wearables: Best Devices That Monitor Stress and How They Work*. Accessed: Nov. 12, 2022. [Online]. Available: https://www.wareable.com/health-and-wellbeing/stress-monitoring-wearables-explained-7969

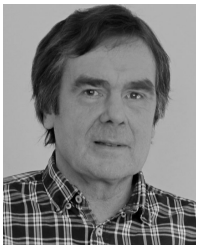
[14] M. Garbarino, M. Lai, S. Tognetti, R. Picard, and D. Bender, "Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *Proc. 4th Int. Conf. Wireless Mobile Commun. Healthcare Transforming Healthcare Through Innov. Mobile Wireless Technol.*, Nov. 2014, pp. 39–42.

[15] X. Li, M. C. Rozendaal, K. Jansen, C. Jonker, and E. Vermetten, "Things that help out: Designing smart wearables as partners in stress management," *AI Soc.*, vol. 36, no. 1, pp. 251–261, Mar. 2021.

[16] S. Lee, H. Kim, M. J. Park, and H. J. Jeon, "Current advances in wearable devices and their sensors in patients with depression," *Frontiers Psychiatry*, vol. 12, Jun. 2021, Art. no. 672347, doi: 10.3389/fpsyt.2021.672347.
[17] A. Landowska, "Affective learning manifesto—10 years later," in *Proc. Eur. Conf. e-Learn. (ECEL)*, Jan. 2014, pp. 281–288.
[18] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Appl. Artif. Intell.*, vol. 23, no. 2, pp. 123–150, Feb. 2009, doi: 10.1080/08839510802631745.
[19] S. K. D'Mello and R. A. Calvo, "Significant accomplishments, new challenges, and new perspectives," in *New Perspectives on Affect and Learning Technologies*. New York, NY, USA: Springer, 2011, pp. 255–271, doi: 10.1007/978-1-4419-9625-1_19.
[20] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.
[21] Y. Sakai, T. Yagi, and W. Ishii, "EEG analysis of mental concentration in golf putting," in *Proc. 5th Biomed. Eng. Int. Conf.*, Dec. 2012, pp. 1–5.
[22] C. C. Dereceli, "An examination of concentration and mental toughness in professional basketball players," *J. Educ. Training Stud.*, vol. 7, no. 1, pp. 17–22, 2019.
[23] A. M. Weber and S. J. Segalowitz, "A measure of children's attentional capacity," *Develop. Neuropsychol.*, vol. 6, no. 1, pp. 13–23, Jan. 1990.
[24] C. A. Riccio, C. R. Reynolds, and P. A. Lowe, *Clinical Applications of Continuous Performance Tests: Measuring Attention and Impulsive Responding in Children And Adults*. Hoboken, NJ, USA: Wiley, 2001.
[25] F. Mori, F. A. Naghsh, and T. Tezuka, "The temporal change of attentional levels under different music environments," in *Proc. Int. Conf. Comput. Supported Educ.* Cham, Switzerland: Springer, 2014, pp. 52–67.
[26] S. Kumar and S. Telles, "Meditative states based on yoga texts and their effects on performance of a letter-cancellation task," *Perceptual Motor Skills*, vol. 109, no. 3, pp. 679–689, Dec. 2009.
[27] H. S. Nguyen, Y. Takahata, M. Goto, T. Tanaka, A. Ohsuga, and K. Matsumoto, "Estimating the concentration of students from time series images," in *Proc. 35th Int. Conf. Comput. Appl.*, in EPiC Series in Computing, vol. 69, G. Lee and Y. Jin, Eds. Manchester, U.K.: EasyChair, 2020, pp. 224–229. [Online]. Available: https://easychair.org/publications/paper/7tch
[28] S. Basterrech and P. Krömer, "A nature-inspired biomarker for mental concentration using a single-channel EEG," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 7941–7956, Jun. 2020.
[29] G. Salomon, "Introducing AIME: The assessment of children's mental involvement with television," *New Directions Child Adolescent Develop.*, vol. 1981, no. 13, pp. 89–102, Sep. 1981.
[30] M. Csikszentmihalyi, "Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi," *Flow Found. Positive Psychol., Collected Works Mihaly Csikszentmihalyi*, vol. 15, no. 3, pp. 1–298, 2014. [Online]. Available: http://link.springer.com/10.1007/978-94-017-9088-8
[31] M. Csikszentmihalyi and J. Nakamura, "Effortless attention in everyday life: A systematic phenomenology," in *Effortless Attention: A New Perspective in the Cognitive Science of Attention and Action*. 2010, pp. 179–189.
[32] J. Marty-Dugas, L. Howes, and D. Smilek, "Sustained attention and the experience of flow," *Psychol. Res.*, vol. 85, no. 7, pp. 2682–2696, Oct. 2021.
[33] W. D. Wilder, M. Csikszentmihalyi, and I. S. Csikszentmihalyi, "Optimal experience: Psychological studies of flow in consciousness," *Appl. Artif. Intell.*, vol. 24, no. 4, p. 690, 1989.
[34] Ö. De Manzano, T. Theorell, L. Harmat, and F. Ullén, "The psychophysiology of flow during piano playing," *Emotion*, vol. 10, no. 3, pp. 301–311, Jun. 2010.
[35] J. Keller, H. Bless, F. Blomann, and D. Kleinböhl, "Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol," *J. Exp. Social Psychol.*, vol. 47, no. 4, pp. 849–852, Jul. 2011, doi: 10.1016/j.jesp.2011.02.004.
[36] M. Zheng and H. A. Spires, "Fifth graders' flow experience in a digital game-based science learning environment," *Gamification, Concepts, Methodologies, Tools, Appl.*, vols. 3–4, pp. 1433–1450, Jun. 2015.
[37] B. E. Mansfield, B. E. Oddson, J. Turcotte, and R. T. Couture, "A possible physiological correlate for mental flow," *J. Positive Psychol.*, vol. 7, no. 4, pp. 327–333, Jul. 2012.
[38] W. Lee, J. Oh, and J. Shim, "A new approach to estimate concentration levels with filtered neural nets for online learning," *Complexity*, vol. 2022, pp. 1–8, Apr. 2022.
[39] V. T. Lokare and L. D. Netak, "Concentration level prediction system for the students based on physiological measures using the EEG device," in *Intelligent Human Computer Interaction*, M. Singh, D.-K. Kang, J.-H. Lee, U. S. Tiwary, D. Singh, and W.-Y. Chung, Eds. Cham, Switzerland: Springer, 2021, pp. 24–33.
[40] M. Mavrikis, "Data-driven modelling of students' interactions in an ILE," in *Proc. 1st Int. Conf. Educ. Data Mining*, 2008, pp. 87–96.
[41] R. S. 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," in *Proc. 5th Int. Conf. Educ. Data Mining (EDM)*, 2012, pp. 126–133, [Online]. Available: http://eric.ed.gov/?id=ED537205
[42] F. Bouchet, R. Azevedo, J. S. Kinnebrew, and G. Biswas, "Identifying students' characteristic learning behaviors in an intelligent tutoring system fostering self-regulated learning," in *Proc. 5th Int. Conf. Educ. Data Mining (EDM)*, 2012, pp. 65–72. [Online]. Available: http://files.eric.ed.gov/fulltext/ED537188.pdf
[43] J. L. Sabourin, B. W. Mott, and J. C. Lester, "Early prediction of student self-regulation strategies by combining multiple models," in *Proc. 5th Int. Conf. Educ. Data Mining (EDM)*, 2012, pp. 156–159.
[44] C. Mills and S. D'Mello, "Toward a real-time (day) dreamcatcher: Detecting mind wandering episodes during online reading," in *Proc. 8th Int. Conf. Educ. Data Mining*, 2015, pp. 69–76. [Online]. Available: http://eric.ed.gov/?id=ED560533
[45] A. Hershkovitz and R. Nachmias, "Developing a log-based motivation measuring tool," in *Proc. 1st Int. Conf. Educ. Data Mining*, 2008, pp. 226–233.
[46] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Proc. 6th Int. Conf. Educ. Data Mining (EDM)*, 2013, pp. 43–50.
[47] B. Lewis, I. Smith, M. Fowler, and J. Licato, "The robot mafia: A test environment for deceptive robots," in *Proc. 28th Modern Artif. Intell. Cogn. Sci. Conf. (MAICS)*, 2017, pp. 189–190.
[48] T. Litmanen, K. Lonka, M. Inkinen, L. Lipponen, and K. Hakkarainen, "Capturing teacher students' emotional experiences in context: Does inquiry-based learning make a difference?" *Instructional Sci.*, vol. 40, no. 6, pp. 1083–1101, Nov. 2012.
[49] U. Nussinovitch, K. P. Elishkevitz, K. Katz, M. Nussinovitch, S. Segev, B. Volovitz, and N. Nussinovitch, "Reliability of ultra-short ECG indices for heart rate variability," *Ann. Noninvasive Electrocardiol.*, vol. 16, no. 2, pp. 117–122, Apr. 2011.
[50] G. Ridgeway, "The state of boosting," *Comput. Sci. Statist.*, vol. 31, no. 31, pp. 172–181, 1999. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.276
[51] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[52] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Animal Ecol.*, vol. 77, no. 4, pp. 802–813, Jul. 2008.
[53] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.
[54] J. Leathwick, J. Elith, M. Francis, T. Hastie, and P. Taylor, "Variation in demersal fish species richness in the oceans surrounding New Zealand: An analysis using boosted regression trees," *Mar. Ecology Prog. Ser.*, vol. 321, pp. 267–281, Sep. 2006.
[55] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vols. 13–17, 2016, pp. 785–794.
[56] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[57] M. Maier, D. Elsner, C. Marouane, M. Zehnle, and C. Fuchs, "DeepFlow: Detecting optimal user experience from physiological data using deep neural networks," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1415–1421.
[58] N. Ganapathy, R. Swaminathan, and T. Deserno, "Deep learning on 1-D biosignals: A taxonomy-based survey," *Yearbook Med. Informat.*, vol. 27, no. 1, pp. 98–109, Aug. 2018.
[59] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, May 2013.

[60] V. Gliner, N. Keidar, V. Makarov, A. I. Avetisyan, A. Schuster, and Y. Yaniv, "Automatic classification of healthy and disease conditions from images or digital standard 12-lead electrocardiograms," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Oct. 2020.
[61] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Commun.*, vol. 11, no. 1, p. 1760, Apr. 2020.
[62] B. H. Kim and S. Jo, "Deep physiological affect network for the recognition of human emotions," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 230–243, Apr./Jun. 2020.
[63] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques TO Build Intelligent Systems*. Sebastopol, CA, USA: O'Reilly Media, 2019.
[64] A. Bäuerle, C. van Onzenoodt, and T. Ropinski, "Net2 vis—A visual grammar for automatically generating publication-tailored CNN architecture visualizations," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 2980–2991, Jun. 2021.
[65] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect recognition—A review," *Sensors*, vol. 19, no. 19, p. 4079, Sep. 2019.
[66] J. Tervonen, K. Pettersson, and J. Mäntyjärvi, "Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors," *Electronics*, vol. 10, no. 5, pp. 1–19, 2021.
[67] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3635–3641.
[68] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2891–2902, Nov. 2019.
[69] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 6256–6268.
[70] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5065–5074.
[71] M. González, C. Bergmeir, I. Triguero, Y. Rodríguez, and J. M. Benítez, "Self-labeling techniques for semi-supervised time series classification: An empirical study," *Knowl. Inf. Syst.*, vol. 55, no. 2, pp. 493–528, May 2018.
[72] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.
[73] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.
[74] W. Yang, "Personalized physiological-based emotion recognition and implementation on hardware," Ph.D. dissertation, School Inf. Technol., Telecommun., Electron., Sorbonne Univ., Paris, France, 2018. [Online]. Available: https://theses.hal.science/tel-02494690
[75] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.
[76] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.
[78] *MyData*. Accessed: Nov. 12, 2022. [Online]. Available: https://www.mydata.org/
[79] M. Kim, S. Yoo, and C. Kim, "Miniaturization for wearable EEG systems: recording hardware and data processing," *Biomed. Eng. Lett.*, vol. 12, pp. 239–250, 2022, doi: 10.1007/s13534-022-00232-0.

CAJ SÖDERGÅRD is a Professor of digital services and working in his own company NextAI. After working in the industry, he was in a variety of positions at VTT Technical Research Centre of Finland Ltd., including the Team Leader and the Center Leader. He has developed big data methods and applications for more than 40 years, starting from image processing, mainly for the media industry and more recently for applications within nutrition, environment, learning, and bio-economy. He has 270 publications and five patents. Recently, he was on the board and the Vice President of the AI, Data, and Robotics Association. He was a member of the EU High-Level Expert Group on the European Open Science Cloud.



TIMO LAAKKO received the D.Sc. (Tech.) degree in information processing science from the Helsinki University of Technology (now Aalto University), in 1994. From 1999 to 2004, he was a Research Professor of multimedia with VTT Technical Research Centre of Finland Ltd., Espoo, where he is currently a Senior Scientist. He has managed and participated in several research projects concentrated widely on data-driven services, mobile applications, and context awareness. His current research interests include machine learning and data-driven (including personal data) applications and services.