DEGREE PROJECT

Adaptive Learning

Per Grundtman

Computer Science and Engineering, masters level
2017

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering

Abstract

The purpose of this project is to develop a novel proof-of-concept system in attempt to measure affective states during learning-tasks and investigate whether machine-learning models trained with this data has the potential to enhance the learning experience for an individual.

By considering biometric signals from a user during a learning session, the affective states *anxiety, engagement* and *boredom* will be classified using different signal transformation methods and finally using machine-learning models from the *Weka Java* API.

Data is collected using an Empatica E4 Wristband which gathers skin- and heart related biometric data which is streamed to an Android application via Bluetooth for processing.

Several machine-learning algorithms and features were evaluated for best performance.

Preface

This master thesis was performed between April 2016 and November 2016 under the Department of Computer Science, Electrical and Space Engineering at Luleå University of Technology.

I would like to thank professor Peter Parnes for the opportunity, support and provision of resources, enabling me to work with this project. I also want to thank Niklas Karvonen for the rewarding discussions and Agneta Hedenström along with the teachers at Antnässkolan, Luleå, for the feedback and opportunity to test my system in a real-world scenario.

Table of Contents

Table of figures

Table of tables

Definitions and Abbreviations

A plethora of abbreviations re-occur throughout this theses of which are denoted with different names in the literature. To summarize and clarify these abbreviations, glossary and corresponding meanings are listed below. In-depth explanations can be found throughout the thesis in relevant subsections.

- ***Biometric sensors:*** This term is sometimes referred to as "psycho-physiological sensors" in the literature. Biometrics refer to the distinctive and measurable metrics of human characteristics which can be used to describe individuals.
- ***ITS:*** Intelligent Tutoring System
- ***MOOC:*** Massive Open Online Courses
- ***Accuracy:*** The word accuracy in this report is simply referring to the number of correctly classified instances among the total number of instances.
- ***Precision:*** The fraction of correctly classified positives (true-positives) among the total set of instances that were actually classified as being positive (true positives + false positives).
- ***Recall:*** The fraction of correctly classified positives among all relevant elements (true positives + false negatives). In other words: the fraction of correctly classified positives among the true positives and the elements that got classified as being negatives, though they should be classified as being positives.
- ***ROC-area:*** Also known as *ROC AUC* (Receiver Operating Characteristic Area Under the Curve) in the machine learning community. A measure of performance for a binary classifier which shows the trade-off between *Precision* and *Recall.* This value is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. A ROC-value of 0.5 implies that a classifier outputs predictions that are random. A ROC-value of 1 is optimal and implies a classifier which makes perfect predictions.
- ***F-measure:*** Also a measure of classifier performance. The F-measure is defined as $F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

  The F-measure can be interpreted as being the weighted average of the precision and recall, where a value of 1 is considered best and 0 worst.

Section 1: Introduction

As we proceed into the educational system of the 21th century with new challenges and many improvements to be made regarding efficiency and availability of educational platforms, the digitalization is important in letting us take the next step. There are a few reasons for this. Firstly, as unique individuals, we simply learn differently. While one student might struggle with basic algebraic concepts, a second student might easily get bored as the pace of the teaching is way too slow to stimulate this student and keep her motivated to advance in her path of learning. This is where Artificial Intelligence in Education (AIEd) comes into place.

There is much research going on regarding how we can measure affective states and use this meta-cognitive information to enhance learning. The purpose of this project is to investigate this area thoroughly and create a "*proof-of-concept*" system where psychophysiological (from now on referred to as *biometric*) sensors are used to discover the affective state during a learning procedure.

To enable measurement of affective states, an Empatica E4 Wristband (https://www.empatica.com/) was used to capture biometric data that has been linked to affective states in previous research. The Empatica E4 can collect data that are skin-related (Electrodermal Activity and body temperature) and heart-related (Heart-Rate, Heart-Rate Variability and Blood Volume Pulse).

1.1 Background

Can we view learning from another perspective and potentially get helpful insights and knowledge about particular measureable flaws in our education system? These might be small recurrent fragments in particular tasks which learners struggles with. By investigating where these obstacles arise, preventing measures can be taken at an early stage, making education more motivating and thus effective. This can be compared with having one teacher per student who gives personal feedback by not only recognizing results, but also *how* the student is experiencing the learning. This is just an example of how AI can be used in education and why this is a big area of research.

As massive open online courses (MOOCs) and intelligent tutoring systems (ITSs) become more widespread and intelligent learning environments gets deployed in public school classrooms but also steps outside the classrooms and into our homes, creating

probabilistic learner-models have become more and more important as a step into the next generation of learning-systems [1].

We need concrete options on how AIEd can be used to assist and support the teaching profession [2] and how we approach teaching overall today. A hot topic especially the last decade has been the investigation of how emotions and affective states are linked to learning and a learner's progress. Efforts to try to detect and measure these affective states through advanced technology is thus interesting for the whole field of technology-enhanced learning [1].

With this knowledge in mind, we want to further investigate how AI can be used to enhance the learning experience, by considering the mental state of the learner during a learning procedure.

1.2 Motivation

Research shows that learning is enhanced when empathy or support is present [3–5]. We also know that learning is connected to mood, meta-cognition, self-efficacy and other affective components. The ability to detect mood which can be directly linked to how well students' learn is thus important.

There is research suggesting that affective states such as boredom can be detected without physical sensors by analyzing log files from intelligent tutoring systems (e.g. [6]). Other research suggests that the use of these sensors can significantly improve the learning procedure because the algorithms calibrating e.g. the trails of exercises can be remarkably shorter than in an ITS which does not use sensors [5, 7].

A common denominator is however that some affective states such as boredom, frustration and confusion is particularly troublesome to detect without the use of biometric sensor technology. Referring to section 1.3, it is also important to keep the obtrusive level low in a learning environment or that will in turn backfire to our disaffection.

Keeping distractions away by using a minimal set of physical sensors is a challenge but also possibility to discover ways that can potentially improve student motivation. If the students affective state can be detected during learning, it can be used as augmentation for e.g. an ITS to calibrate the difficulty-level of assignments. Furthermore, this might help to maintain the motivational peaks we occasionally feel during learning, which is important when solving problems and learning new material.

1.3 Problem definition

This section describes the challenges and questions which is ought to be thoroughly investigated. The main objective of this project is to investigate whether how the learning experience can be improved by considering the affective state of the learner and if so, attempt to build a model that can distinguish between these states using machine-learning techniques.

In the context of this main objective, I shall also examine the points below and treat them throughout this report.

1.3.1 Theoretical problems

(P1) Investigate what psychological parameters affects our learning potential and how these affect the learning experience. More specifically: how can affective state recognition eventually improve our learning experience?

(P2) The second problem is attempting to define affective states and distinguish them from affective states with similar physical features, because it may e.g. result in an unfavorable type of feedback. Feelings/mood/affective states are not strictly defined, and all individuals react differently during these states. So, is it possible with a simplified model to map these affective state to quantifiable data-patterns that can be used to train a model using machine-learning techniques?

(P3) Investigate what has been done in related work regarding affective state recognition, and how biometric-data mining can be done with relatively simple technically means that are not too intriguing for the user, but also viable in terms of cost, should a similar system be deployed in public schools or in homes.

(P4) Would it make sense to use the data about affective states to provide external user feedback (feedback outside the screen)? Or alternatively be used in cooperation with an Affective Agent (AA) to calibrate the algorithms used to individualize the learning process even more effectively and accurately? The simple fact to when, what and how to quantify the input and turn it to helpful feedback (e.g. regulating difficulty level of the ITS).

1.3.2 Technical problems

The technical problems needed to be solved in this project are listed below.

(P5) Evaluate and decide on methods on how to extract data from the chosen sensors.

(P6) Decide on an optimal subset of features that can represent the chosen affective states along with which baseline values to extract, in order to calibrate a model for each participant using the system. With these decisions, evaluate different machine-learning algorithms that best fits the purposes of this classification problem. They might differ in suitability for classifying offline and online, regarding performance cost etc.

1.4 Delimitations

This is just a "proof-of-concept" project. Systems of this sort can be widely augmented in several ways and some of the concepts that are important to consider are the following.

- Attempting to measure affective states and emotions using biometric sensors is a big research field. This project is limited in trying to measure affective states as an interim target with relatively small resources, since the final result is not intended to become a working product for commercial purposes, but rather a study on available technology and methods which can be used e.g. to improve the learning experience for a student using an intelligent tutoring system. While some research suggests that a larger set of biometric sensors can improve the accuracy of these kind of systems, this project will attempt to deliver good accuracy with only a subset of biometric sensors.
- While the subject of improving self-efficacy (to make the student aware of his/her own ability to perform in a given situation [8]) is deeply connected to the topic of this project, no feedback system will be developed. At the beginning of the project, the idea was to give motivational feedback with the purpose to lessening unfavorable moods, e.g. leading to an affective state of 'stuck', frustration or boredom but also to keep up motivation when the learner is doing well or being under stimulated (inspiration from [5, 7, and 9]). Because of lack of time, the project will merely treat the part of affective-state recognition.
- Our mental attitude highly influences how we learn. Therefore, a large set of affective states would be needed to consider when calibrating the diversity and difficulty-level of the assignments in an ITS to be more accurate. In this project,

the set of affective states which will be attempted to measure will be inescapable simplified into three groups of emotions: *anxiety, engagement* and *boredom*. To strive for a larger set of affective states would not necessarily mean improving the system. The issues of mapping human behavioral patterns can be read more of in Section 2.

- The tests will be done using a real-life learning scenario with middle-school students, and not in conjunction with an ITS.

1.5 Thesis structure

Section 2 discusses related work to get an overview of what has been done in the field. Section 3 starts by explaining how the project was initialized and the final structure was determined. The section continues with project methodology, including experimental procedure and how the features were extracted from the sensor-data. Section 4 describes implementation while Section 5 describes the evaluation of the implemented system. Section 6 discusses improvements and further thoughts and the report ends with results in Section 7 and future works in Section 8.

Section 2: Related work

The areas of related work can roughly be categorized to '*general studies of biometric sensors and emotions'* and '*biometric sensors to augment ITS's'.*

2.1 General studies of biometric sensors and emotions

Similar experiments where physiological signals are analyzed to determine how the user is feeling have been conducted [5, 10–21]. In [16], physiological data was measured during sessions of the game Tetris in order to adjust the game depending on the current emotional state of the player. Data was analyzed based on complete 5-minute sessions on of the game. The researchers stated in the Future Work session that analyses should be conducted more frequently and not only during the period of a complete game, since feelings such as anxiety can arise at any time during a session.

2.2 Biometric sensors to augment ITSs'

We are only in the beginning of understanding how affect contributes to a learner's progress.

A lot of research regarding ITSs' has been conducted and how to infer a student's affective state, but mainly through methods using software (e.g. Machine Learning). However, in the last couple of years, research has also been conducted using physical sensors to augment existing ITS's.

In [7], it's described how an ITS called AutoTutor was augmented by classifying emotions by recognizing facial expressions, body movements and conversational cues.

In [5], a combination of physical sensors, including face-recognition, conductance-bracelets, eye-tracking, pressure sensitive mice and seats equipped with accelerometer were used to gather data which have the potential to provide information on a student's physiological responses which have evidently been linked to affective states such as frustration, boredom etc. Many interesting notes were taken from this paper and contributes to what is being done in this project. For instance that students self-reports of emotions can actually be inferred automatically from physiological data that is streamed to a tutoring software for students in a real public school setting. The future work from the same paper was to predict emotions reasonably well with only a subset of sensors available in the classroom.

Section 3: Theory and Methodology

This section describes the steps, leading up to the methods which finally was chosen for this project. It includes a description of the difficulties with measuring affective states and how this was solved by choosing the Empatica device.

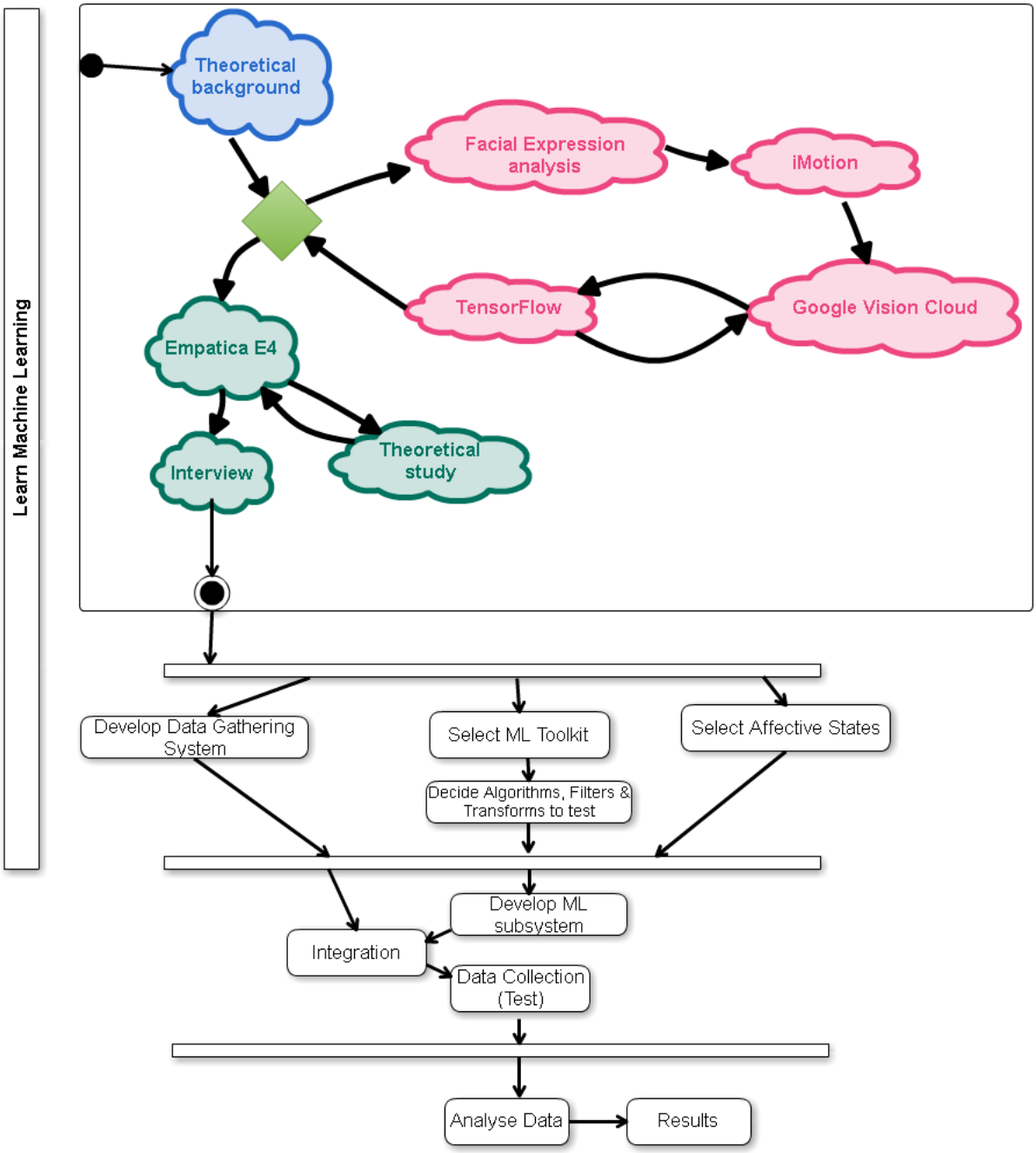


Figure 1: Methodology diagram

Figure 1 shows the overall structure of how this project was conducted and can be broken down in the following parts.

Machine Learning. With no previous experience in machine-learning, familiarization with this topic was carried out throughout the whole project.

Pre-study. The diversity of the project resulted in a comprehensive pre-study in the areas of artificial intelligence in education (AIEd), intelligent tutoring systems, psychosociological studies of emotions (affective states & metacognition), biometric sensors and related research in affective state recognition. This study continued throughout the whole project as I learned more and the angle of my topic changed.

Pre-implementation: Emotion detection using camera. During several weeks I investigated alternatives for how to use emotion detection through a normal web-camera. Several top-of-the-line technologies were found but they all came with a high subscription cost. Next step was to find free solutions, which led me to Google's Cloud Vision API [22] and the Microsoft Oxford Project [23]. These API's both had evaluation periods for free so I decided to put time into making a simple prototype web-application. Pictures was taken through the webcam, using WebRTC and uploaded to the Cloud Vision service which returned probabilities that a person in the image expressed different emotional expressions like joy, sadness etc. After some minor tests, I decided the accuracy, complexity and reliability to not be in favor of my project. I found that the accuracy of the emotion detection was not nearly as accurate as I would have needed in order for it to be useful for my proof-of-concept application.

I also learned about Google Tensor Flow [24], which is the powerful API which powers most of Google's AI services (Cloud Vision included). My hopes was that there would be a relatively easy way to train a model for facial expressions using Tensor Flow but came to the conclusion this would be a project itself which would consume too much time.

This led to a new theoretical study and the good opportunity to use the Empatica E4 wristband at LTU. The Empatica E4 contains three useful sensors for biometric readings which I had investigated already during the first pre-study, which made me settling for this technology.

Interview with teachers. To gather feedback for my project, a visit to Antnässkolan in Luleå was done (see Section 10 for a compilation of this meeting). During this little workshop, I also concluded to delimit the project to not implement any external (outside-the-screen) feedback which was the original plan.

Implementation phase: Data from the Empatica E4 can be saved in the internal memory of the device and can be retrieved using the Empatica E4 software for post-analysis. However to perform processing of the raw-data, development of an Android application was started to retrieve data for processing in real-time. Android was chosen because of compatibility with the Empatica Developer SDK. During this time, I settled for the *Weka* API since it's well-documented and widely used for machine-learning projects. Detailed information about system-architecture can be found in Section 4.2.

After software and hardware had been chosen, a thoroughly investigation of what features to extract from the sensors was performed, along with signal-processing methods needed before feature extraction. Information of feature-extraction can be found in Sections 3.5-3.6.

Testing and evaluation: Tests were performed at Antnässkolan in Luleå, where third grader were to attend math-exams while wearing the Empatica wristband. Section 3.4 describes the experiment in detail with test-evaluation described in Section 5.

3.1 Determining how detecting affective states can help to improve the learning experience

Finding the right type of feedback to "confront" boredom has turned out to be trickier than imagined. According to [25] the antithesis of a certain view of boredom is to eliminate coercion, to support the students own agendas and motivate the freedom of own choice. This would in a real-school environment mean to induce an "informational" classroom environment where the student is encouraged to participate in his/her own assessment and planning of learning rather than the usual "controlling" environment where the student is dependent on the teacher (or intelligent tutor).

In conclusion of what Belton & Priyadharshin writes in [25], boredom plays a significant role in the learning process, even though it's' an ambiguous concept which cannot (and perhaps not should) be easily countered, the concept of autonomy and control in the learning environment should be given more focus in the educational systems. This could be a potential improvement for future ITSs' to consider by e.g. enabling the student to manipulate his/her own trail of exercises in some way, when boredom is detected by the system.

For this project, I will use the answers from the questionnaire (Section 10) along with a somewhat simplified model of recurrent behavior of learners, to map the sensor inputs to an appropriate label of affective state, which in future work could be used to trigger proper interventions in the learning system.

3.1.1 Gathering data about affective state

In Ekman's "The nature of emotion" (As cited in [11]), it's stated that emotions typically last for seconds or at the longest up to minutes. For this reason, when reporting the current emotion to the application, this label of the currently experienced emotion was set as the label for the calculated data instances roughly half a minute back in time as well as the current instance of calculated data. This makes sense as the feeling might have arisen before the participant was asked for a state report.

3.2 Definitions of affective states

The level of difficulty must be high enough to maintain engagement, whereas it otherwise can lead to boredom should the task be too easy. If the level of the task is too high relative to the skill of the user, it might lead to anxiety and frustration, see Figure 2.

This report suggests to model the sensor inputs according to the "flow-theory" of Mihaly Csikszentmihalyi which says the quality of experience depend on the challenge experienced and skill required in specific situations (as cited in [26]). Thus this project merely focus on the three affective states *anxiety*, *engagement* and *boredom*. This is a popular model in psychological- and affective computing research, e.g. [10, 23].

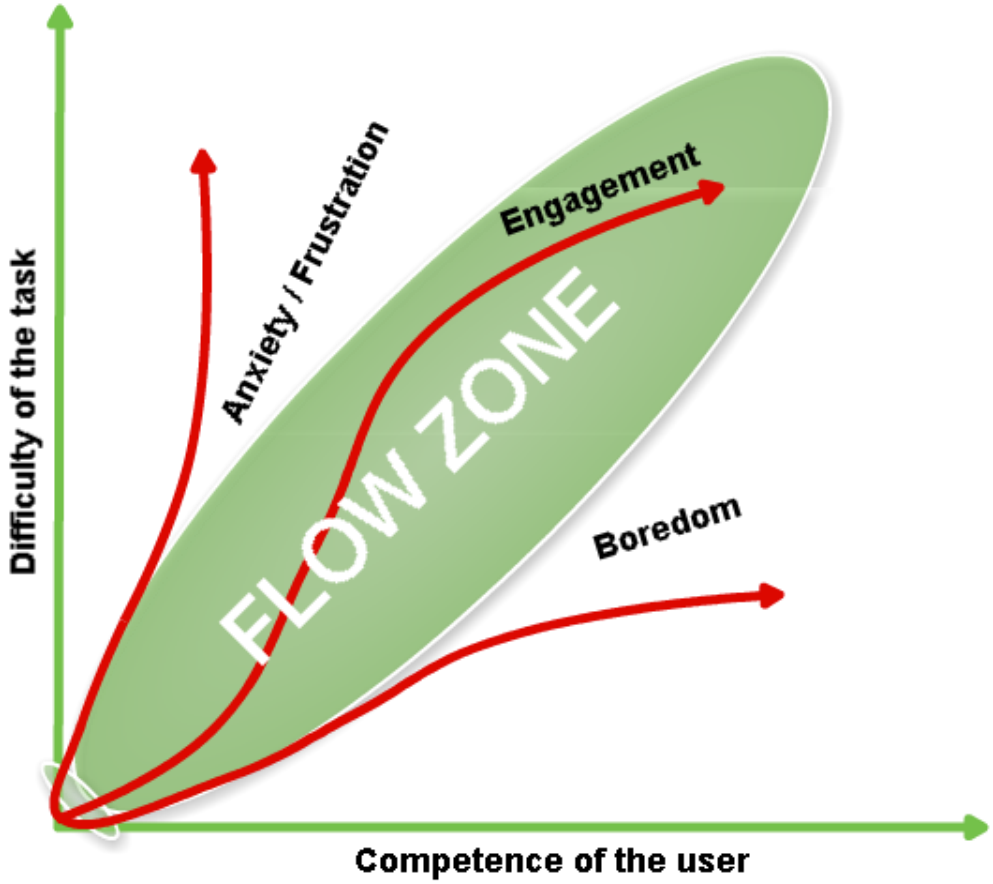


Figure 2: Flow zone factors

In this model the notions of what is being considered "good" vs "bad" feelings, are drastically simplify for the purpose of feasibility of this project.

Positive emotions:

Affective states that have been proven to contribute to work progress and the psychological state of being in *flow.*

- Arousal
- Engagement (positive excited)
- Interest

Negative emotions:

This category treats affective states being related to feeling a lack of motivation or being *stuck.*

- Boredom (under-stimulation, negatively calm, too easy task)
- Frustration (negative excitement, too difficult task)
- Anxiety (can't comprehend with the task)

3.3 Biometric sensors and their settings

As mentioned in Section 3, the measurement equipment of choice was an Empatica E4 device. The motivation is that it is equipped with several biometric sensors proven to be useful in detecting affective states, while also being less troublesome and intrusive for the user (referring to **P3** in the problem definition). Table 3 is a summary of affective states related to this project, which previously has been linked to relevant biometric readings (acquirable from the Empatica) in the literature. The sensors in the Empatica are explained in below. The accelerometer sensor was not used in this project. An overview of the system is depicted in Figure 3.

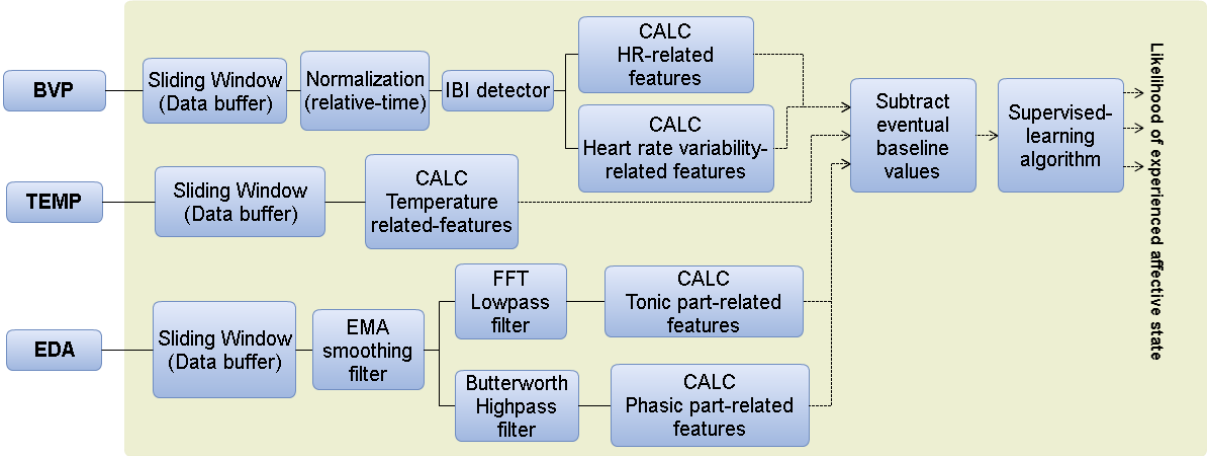


Figure 3: System diagram overview

Photoplethysmography (PPG) Sensor. Measures blood volume pulse (BVP). From the BVP signal, heartrate and heartrate variability related features can be extracted. This sensor operates at a sampling frequency of 64 Hz and uses no relative unit for measurement but a relative scale.

Galvanic Skin Response (GSR) Sensor. This sensor is used to measure sympathetic nervous system arousal and to derive features related to stress, engagement, and excitement.

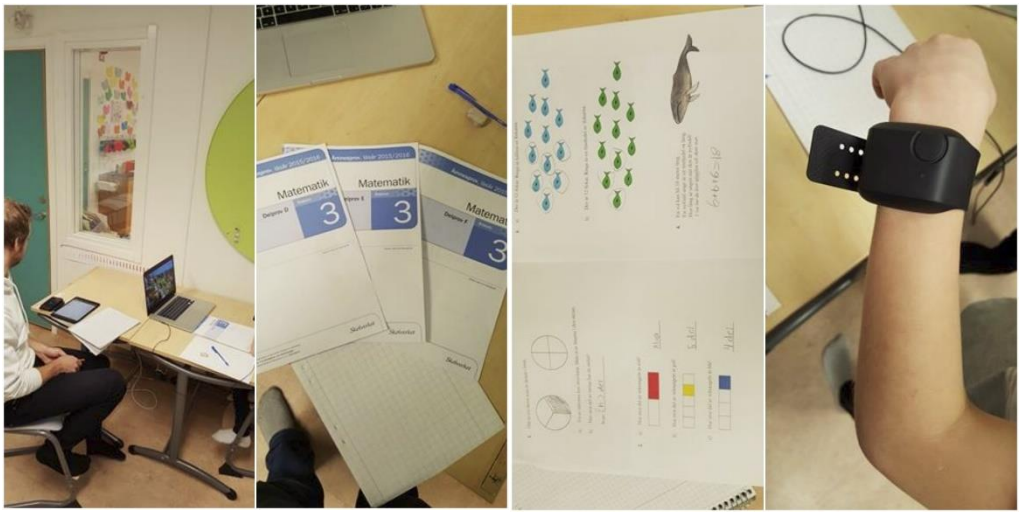
The GSR sensor operates at a sampling frequency of 4 Hz, using the unit microsiemens (μS). It measures electrodermal activity (EDA) in the skin and can be measured as either the skin's electrical resistance or as its electrical conductance. The Empatica device measures the electrical conductance, which is the mode of measurement which makes most sense, because the number of sweat glands and their secretion is linearly related to skin conductance, as cited in [28].

Infrared thermopile. Reads peripheral skin temperature with a sampling rate of 4 Hz with the Celsius degree unit.

3.4 Experimental procedure

This subsection describes how training data was gathered to test the affect-recognition system.

A test-scenario was made for third graders at Antnässkolan, Luleå. Two students at the age of 9 were to conduct a national exam in mathematics during a period of 30 minutes. The time limit of the test was decided because the focus-span of students in this age would probably falter too much with a longer test. The test sessions each consisted of three parts. In between each part, it was a short time slot of about 0.5-1 minute, where the participant was asked to self-report the perceived progress and was also asked how she felt at the end of the test part.



The Empatica E4 wristband was attached to the non-dominant hand [29] (non-mouse/writing hand) of the participant because the dominant hand might induce too many movement artifacts during the test which might prevent good data readings.

However it's worth mentioning that research have been conducted where comparisons of measuring EDA and BVP between the dominant and the non-dominant were done [28, 30, 31]. According to the results in e.g. [28, 32], interesting information might be lost if only measuring on one side. In our case, both participants were right-handed.

To get as good data-readings as possible, the EDA electrodes was aligned with the middle of the ring- and middle fingers.

3.4.1 Pre-test relaxation and calculation of baseline values

Before each session, the user was asked to watch a video clip of relaxing pictures and music used for meditational/concentration purposes, to calm their minds and base their biometric values at a neutral level, and also reduce the risk of the sensor data being affected by external stimuli.

To calibrate the data prior to analysis, these baseline values were calculated during this calming period of 2 minutes. Baseline calculations helps improving the accuracy during feature extraction and is recommended in the literature [11, 16, 19, 33]. The purpose of each individual baseline value is described in more detail under the subsections of Section 3.5.

Baseline values were calculated for each session for the EDA- and BVP data, to avoid circumstantial variations in the data (day-by-day fluctuations), as suggested in [29].

Since the system needs to be fed with affective-state labels as the training goes along, the participant was asked about how they felt with an interval of approximately 3-5 minutes, depending on if the participant was stuck on an assignment or not. By asking too frequently, these question might disturb the participant too much. At the same time, by not questioning frequently enough, changes in state of the participant might be missed and the system might risk getting trained incorrectly. These problems are described in more detail in Section 3.2.1.

3.5. Signal processing

Since biometric data is extremely noisy, it contains a lot of faulty data.

The Empatica wristband samples all data at relatively slow rates. Therefore I decided to be cautious with smoothing of the raw data. Smoothing was only performed on the signals which showed obvious signs of noise and hence I deemed it to be neccesary.

All data have been filtered before running any analysis to obtain more accurate results, with the exception of the infrared thermopile signal. In this case, features were extracted directly from the raw data. The type of filtering depends on the signal. Data smoothing was cautiously used, since some oscillations in the signals are of interest for finding biometric patterns used for feature extraction. Also, a sliding window technique was used with data overlap before any processing. The windowing technique is explained in greater detail in subsection 3.5.3 further down.

Table 1 shows the mathematical annotations used in the following subsections. Explanation of the data cleaning and methods for noise reduction is also explained in the subsections below.

Annotation	**Meaning**
X	Complete data-set of the current window
x_i	Value of the data-point at index $i \in X$
N	Size of the data-set of the current window
n	Index of the data-point in the data-set X

Table 1: Mathematical annotations

3.5.1 Acquiring data from the EDA signal

EDA signals consists of two parts: One part is a slowly changing, low-frequency signal called the *tonic part* or skin conductance level (SCL). The tonic signal usually changes over a period of a few minutes and is commonly used as a measure of arousal [33]. The other part is called *phasic*. The phasic part of the signal reflects reactions based on

external stimuli like spontaneous/sudden events of high arousal and is recognized by high frequency peaks which arises very suddenly (1-2 seconds), and drops off at about double the time it took for the signal to rise. These peaks are usually referred to as skin conductance responses (SCR). Example figures of a *tonic* and a *phasic* signal can be seen in Section 10.5.

First, an exponential moving average filter (1) with the fairly high alpha factor of α=0.64 was used to remove noise. The reason is that the sampling frequency of the signal is only 4 Hz and by having a lower alpha factor, I risk over-smoothing and the interesting fluctuations of the signal might not be visible.

$$f(x_n) = x_{n-1} + \alpha(x_n - x_{n-1}) \quad (1)$$

To extract the phasic part of the signal, I used a Butterworth high pass filter (2) which is a biquad digital filter of the 2nd degree.

$$f(n) = a1 \cdot x_n + a2 \cdot x_{n-1} + a3 \cdot x_{n-2} - b1 \cdot f(x_{n-1}) - b2 \cdot f(x_{n-2}) \quad (2)$$

$$c = \tan(\pi \cdot \frac{f}{sampRate})$$

$$a1 = \frac{1}{(1 + r \cdot c + c^2)}$$

$$a2 = -2 \cdot a1$$

$$a3 = a1$$

$$b1 = 2 \cdot (c^2 - 1) \cdot a1$$

$$b2 = (1 - r \cdot c + c^2) \cdot a1$$

The resonance value of the filter was set to 8.0 and the cut-off frequency to 0.05 Hz. The reason for the 0.05 Hz cut off is because the longest assumed duration of a SCR is 8 seconds. With a 0.05 Hz cut off I'm therefore assured to see the whole part of the signal. The resonance parameter was manually tweaked to avoid over-shot of the signal and keep the filter stable while also not erasing important fluctuations in the signal that are of importance for the feature extraction.

SCR's rise approximately in 1-2 seconds and decays over a period of 2-6 seconds and the maximum frequency of a phasic response is about 0.33 Hz according to the literature. Therefore I needed to extract the phasic part of the signal at 0.66 Hz according to the Nyquist Theorem to make sure the whole phasic response is recorded. This was of course not a problem since the sampling frequency was well over the 0.66 Hz minimum. The filter drastically eliminates the signal under 0.05 Hz and the SCR's could now be detected with double thresholding inspired from [21]; by observing two zero-crossings on the x-axis: from negative to positive and from positive to negative.

The amplitude of SCRs can differ quite noticeably. Inspired from [21], we looked at the maximum amplitudes between each double zero-crossing in each window. First, a max peak between such zero-crossings needed to exceed a lower threshold at 0.02 μS of the filtered signal to qualify as a SCR. Next, the highest amplitude found between all zero-crossings whose amplitude were of at least 20% of the largest amplitude of this data segment were considered to be SCRs.

The tonic signal has a tendency to continuously rise and fall approximately linearly when a person is experiencing arousal when engaged in a task for instance. Therefore, only considering the amplitude of the tonic signal can be misleading, and be destructive for classifying purposes. For this reason, features related to this part of the signal were extracted by looking at the trend of the signal rather than the amplitude explicitly.

To prepare the data for statistical analysis, I first normalized the raw signal with the running mean of each sliding window. Thereafter the normalized smoothed signal was filtered using a low-pass filter based on Fast Fourier Transform (3) with a cut-off frequency of 0.05 Hz. This was to neglect eventual SRCs.

$$X_n = \sum_{k=0}^{N-1} x_n \cdot e^{-\frac{i2\pi kn}{N}} \tag{3}$$

Two examples of how this filter removes high frequency spikes in the signal is shown in Figure 12 and Figure 13 in Section 10.5.

To consider the trend of the tonic part of the signal I implemented a simple method inspired of [34] called *EDA Positive Change* (EPC) as an alternative way to measure arousal out of the tonic EDA signal. The idea of this method is to simply calculate the sum of all discrete positive changes in the EDA signal in the window.

Instead of taking the median values between local minima of the raw signal, and measure EPC's from this smoothed signal, the number of EPC's were calculated from the output of the low-pass FFT-filter.

Another interesting feature from the tonic signal is to look at the slope of the tonic signal. Instead of looking at the mean of the raw signal, simple linear regression was chosen to extract the first derivative of each window (see Figure 14 & Figure 15 in Section 10.5).

3.5.2 Acquiring data from BVP/PPG Signal:

The main purpose of the BVP sensor is to calculate the heart-rate of the user. The heart rate can be derived by computing the time-length between adjacent local peaks in the BVP signal. These time-lengths are called Inter-Beat-Intervals (IBI) and can be used to calculate the instantaneous heart-rate as well as the average heart-rate (pulse) over a period of time.

Many of the features used in this project are based on the IBI value. By applying simple transforms of the IBI, interesting biometric patterns can be found which has been shown to correlate with affective states in the literature.

The Empatica developer SDK included a sample application which calculates IBI values from the wristband. Their method was accurate when the wrist was held completely still. However, this was not very reliable since even the smallest of movement artifacts resulted in getting no IBI data at all. I solved this with a simple method:

First, all timestamps for the BVP raw-data points in the window were normalized with the first timestamp-value in the window, as depicted below.

$$timestamp(x_i) = timestamp(x_i) \quad - timestamp(x_0), \qquad \mathrm{i} = 0, 1, 2, \dots, \mathrm{N} \tag{4}$$

After normalization, the raw BVP signal was smoothed using formula (1) with α=0.12. Then the local peaks were calculated by comparing neighboring data points with one another of the smoothed signal until a peak was found. From this, the relative time differences between adjacent peaks were calculated. This method is sensitive to noise artifacts so a threshold of minimum time allowed between IBI's of 0.4 sec was

used, which corresponds to a max-heart rate allowed of 150 bpm. The heart-rate was calculated as the inverse of the IBI.

$$HR = \frac{1}{IBI} \cdot 60 \tag{5}$$

Looking at the method in Figure 4, we can see a BVP signal which is noisy. An IBI wavelength include diastolic points [30] that can be seen in the same figure. To extract credible time-intervals between these peak amplitudes, one of the points in each pair of diastolic points was removed with the filter (see the blue signal) and the IBI can be calculated. The minimum threshold mentioned above takes care of possible local maximas being too close to each other (probable noise artifacts).

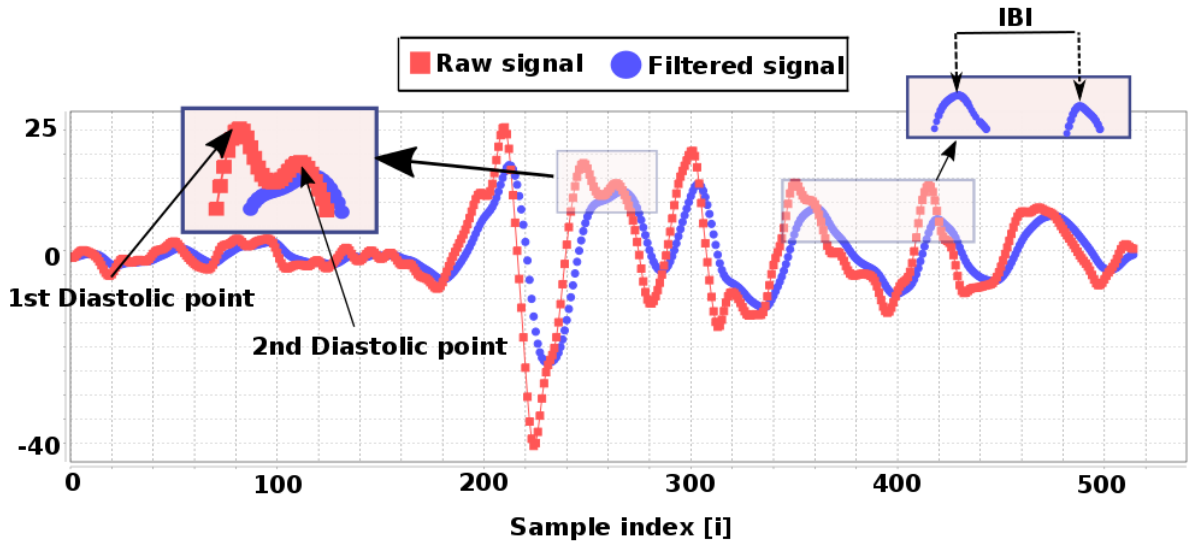


Figure 4: IBI extraction

To investigate the credibility of this method, IBI values were compared from our method with the values retrieved from the sample application in real-time. The results turned out to be good with the chosen window size (see Section 3.5.3) and hence useful since I retrieved reasonably accurate IBI result even if a participant was moving her arm.

The IBI-values between the two methods differed with a maximum of 10 %. From these calculations, heart-rate and heart-rate variability-related features could be extracted.

3.5.3 Sliding windows

For my experiment, I used 64 second windows with a 32 second overlap (50% overlap) as shown in *table* 2. This relatively large window-size made sure not to miss any interesting patterns in the biometric signals. The reason for the window-size being a power of two is because the FFT implementation required it to work properly.

Sensor	Sampling frequency (Hz)	Samples	Overlap
Galvanic Skin Response (GSR)	4	256	128 (50%)
Infrared Thermopile	4	256	128 (50%)
Photoplethysm--ography (PPG)	64	4096	2048 (50%)

Table 2: Windowsize, sampling frequency and overlap for each respective sensor

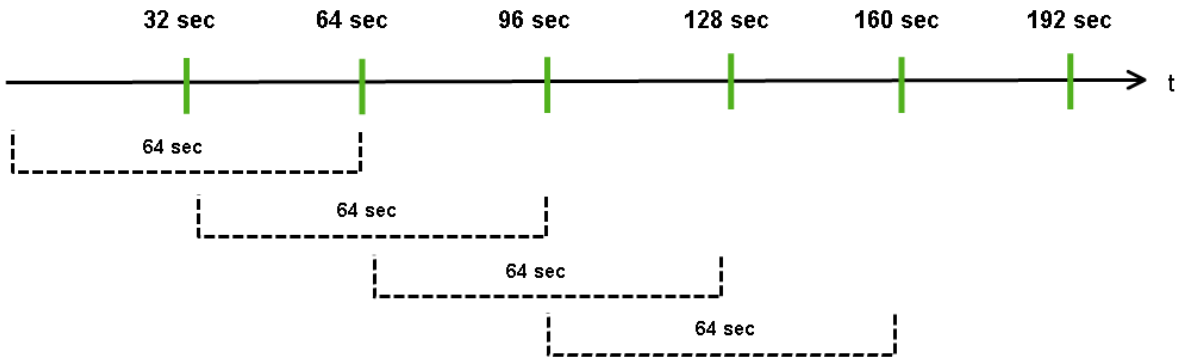


Figure 5: Windowsize

3.6. Feature Extraction

Features are an important term in this report. In machine learning and pattern recognition, a *feature* refers to a measurable pattern observed in some data-set to e.g. explain a phenomenon (an explanatory variable). *Features* are crucial for a machine-

learning algorithm in order to classify a state, or to discriminate between classes of a phenomenon.

The choices of features are based on research listed in *table 3* where certain biometric readings have been linked to affective states. The set of features chosen for viability test are listed in the subsection below. A large number of features are intentionally chosen for test and will be pairwise compared during post-test analysis, to eliminate possible redundant attributes for the classifiers.

Type of signal	**Previously linked to**
Skin-related	
EDA	Valence and arousal [20–22], Engagement [19], Frustration [20]
TEMP	Boredom, engagement and anxiety [9, 19, 24]
Heart-related	
Heart-rate (HR)	Frustration [9, 19]
Heart-rate variability (HRV)	Anxiety [19]
BVP	Frustration [19] Valence and Arousal [19]

Table 3: Biometric measurements previously paired with affective states

3.6.1 Proposed Feature set

All features included in the tests are shown in table and explained in more detailed below.

Sensor	**Feature**
EDA	MaxToStDev
	MinToStDev
	SCLSlope
	EPC
	AUCPhasic
	MaxPeakAmplitude
	MinPeakAmplitude
	MeanSCRAmplitude
	NumOfSCRsPerWindow

	SumOfSCRAmplitudesPerWindow
TEMP	MeanTemp
	LocalStandardDeviation
	MeanTempPeakAmpl
	MaxTempPeakAmpl
	MeanTempDifference
BVP	HRMeanDifference
	HRMean
	HRVarianceDifference
	BVPMean
	BVPRange
	HRLocalStandardDeviation
	SDNN
	RMSSD

Table 4: Features

This is a group of features inspired from the literature [11, 19, 21, 28, 30, 33, 35–37]. Some standard features that are used in the literature, re-occur for the different sensors, like standard deviation (6), variance (7) and running mean (8)

$$StDev(X) = \sqrt{Var(X)}, \tag{6}$$

$$Var(X) = \frac{1}{N-1} \cdot \sum_{i=0}^{N-1} (x_i - Mean(X))^2 \tag{7}$$

$$Mean(X) = \frac{1}{N-1} \cdot \sum_{i=0}^{N-1} x_i \tag{8}$$

3.6.2 Features extracted from EDA signal

A SCR is considered to be a legitimate skin conductance response when it fulfils the requirements in the algorithm described in Section 3.5.1.

- **MaxToStDev -** The max amplitude of the low-pass filtered signal subtracted with the standard deviation of this window
- **MinToStDev -** The min amplitude of the low-pass filtered signal subtracted with the standarddeviation of this window

- **SCLSlope -** The first derivative of the linear regression prediction of the SCL after filtered with the FFT low pass filter.
- **EPC -** Number of positive EDA changes per time-window.
- **AUCPhasic** - The area under the curve of the smoothed high-pass filtered phasic signal.
- **MaxPeakAmplitude** - The maximum peak amplitude of the phasic signal, normalized per time-window.
- **MinPeakAmplitude** - The minimum peak amplitude of the phasic signal, normalized per minute.
- **MeanSCRAmplitude** - The mean SCR amplitude of the window.
- **NumOfSCRsPerWindow** - The number of peaks in the phasic signal, normalized per time-window.
- **SumOfSCRAmplitudesPerWindow** - The sum of phasic peak amplitudes of the phasic signal, normalized per time-window.

3.6.3 Features extracted from the temperature signal

- **MeanTemp** - The running mean temperature
- **LocalStandardDeviation** - the running standard deviation of a window. Describes changes in the signal during this time window
- **MeanTempPeakAmpl** - The mean peak amplitude in the temperature signal
- **MaxTempPeakAmpl -** max peak amplitude
- **MeanTempDifference** - The difference between the mean temperature during the baseline and during the task

3.6.4 Features extracted from the PPG (BVP) Sensor

Some features are similar to that of the EDA signal with the exception that the signal was not normalized for any of the following features.

An IBI is considered legitimate when it fulfils the requirements of the algorithm explained in Section 3.5.2.

- **HRMeanDifference** - The difference between the mean heart rate amplitude during the baseline and during the task.
- **HRMean** - The mean heart rate.

- **HRVarianceDifference** - The difference between the variance in the heart rate during the baseline and during the task *n*.
- **BVPMean -** The average value of the BVP signal.
- **BVPRange -** The range of amplitudes in the window.
- **HRLocalStandardDeviation -** Corresponds to the standard deviation of the heart rate.
- **SDNN** - The standard deviation of beat-to-beat (NN) intervals, normalized by a baseline.
- **RMSSD** - The square root of the mean of the square of the successive differences between adjacent NN intervals [38], normalized by a baseline

$$RMSSD = \sqrt{\left(\frac{1}{N-1}\sum_{k=0}^{N-1}(IBI_{i+1} - IBI_i)^2\right)} \quad (9)$$

Section 4: Implementation

This section describes how the affective-state recognition system was implemented and how the feature- and algorithm problem (P6) in Section 1.3.2 was solved.

4.1 System Components

Biometric data was gathered using an Empatica E4 device, which streamed the data to an Android device, where all data-processing was done in real-time. When a data-collecting session was completed, all processed data was saved as an .arff-file in a user-specific folder on the Android device, along with a tailor-made model for that specific user. The .arff-format is the default format for data processing in the *Weka API* which was used for the machine-learning part of the system.

The system consists of the following components:

- **Empatica E4 (device):** This device was mounted on the wrist of the test subject in order to collect skin- and heart-related biometric data for emotion-recognition.
- **Android application:** All data processing coming from the Empatica device was done on this application. The developed application included a simple GUI with buttons for *creating/loading* user profiles, *start/stopping* the data stream, *start/stopping* feature calculations, *labeling the data* with the correct class label, *activate real-time classification*, and a button to *generate an .arff file* containing all calculated features from this session.

The overall deployment structure of the system is shown in Figure 6. The biometric data from the Empatica E4 was collected via Bluetooth and processed in an Android application for feature extraction, model training and the possibility for real-time recognition of affective states.

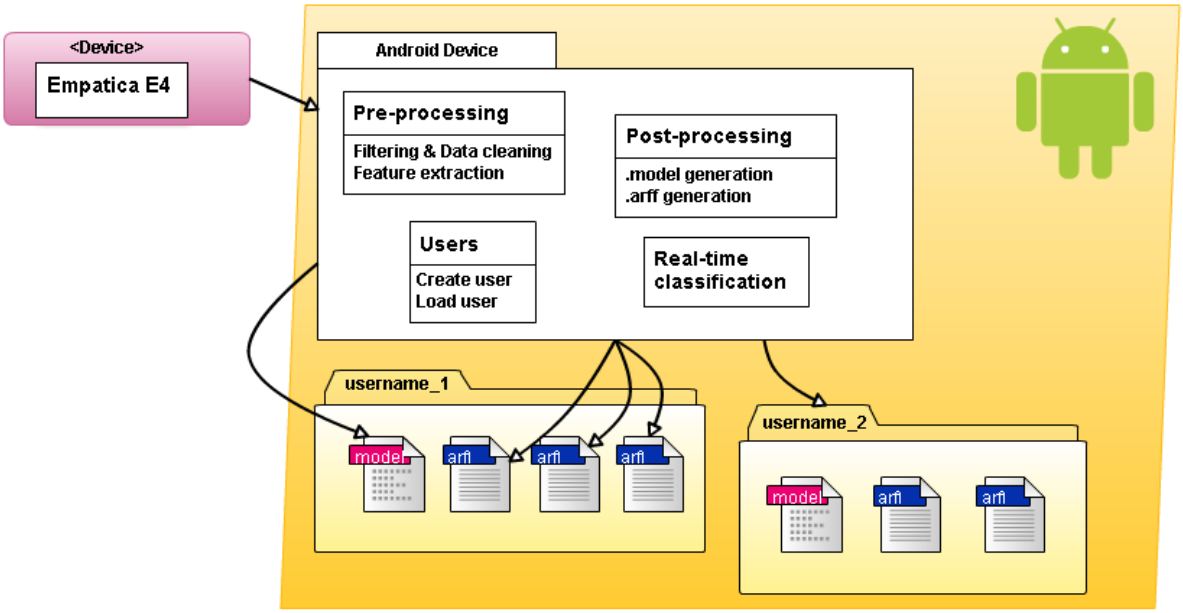


Figure 6: Deployment diagram

4.2 System architecture

This subsection describes the software architecture for the pre- and post-processing of features. The application was built as an extension of a sample application included in the Empatica Mac SDK. The sample application included methods needed to communicate with the Empatica E4 wristband.

4.2.1 Application state diagram

Figure 7: Application state diagram is an overview of system states, describing the system flow during interaction with a user of the application. The basic rule of the system is that a user profile must be loaded (or created) prior to usage.

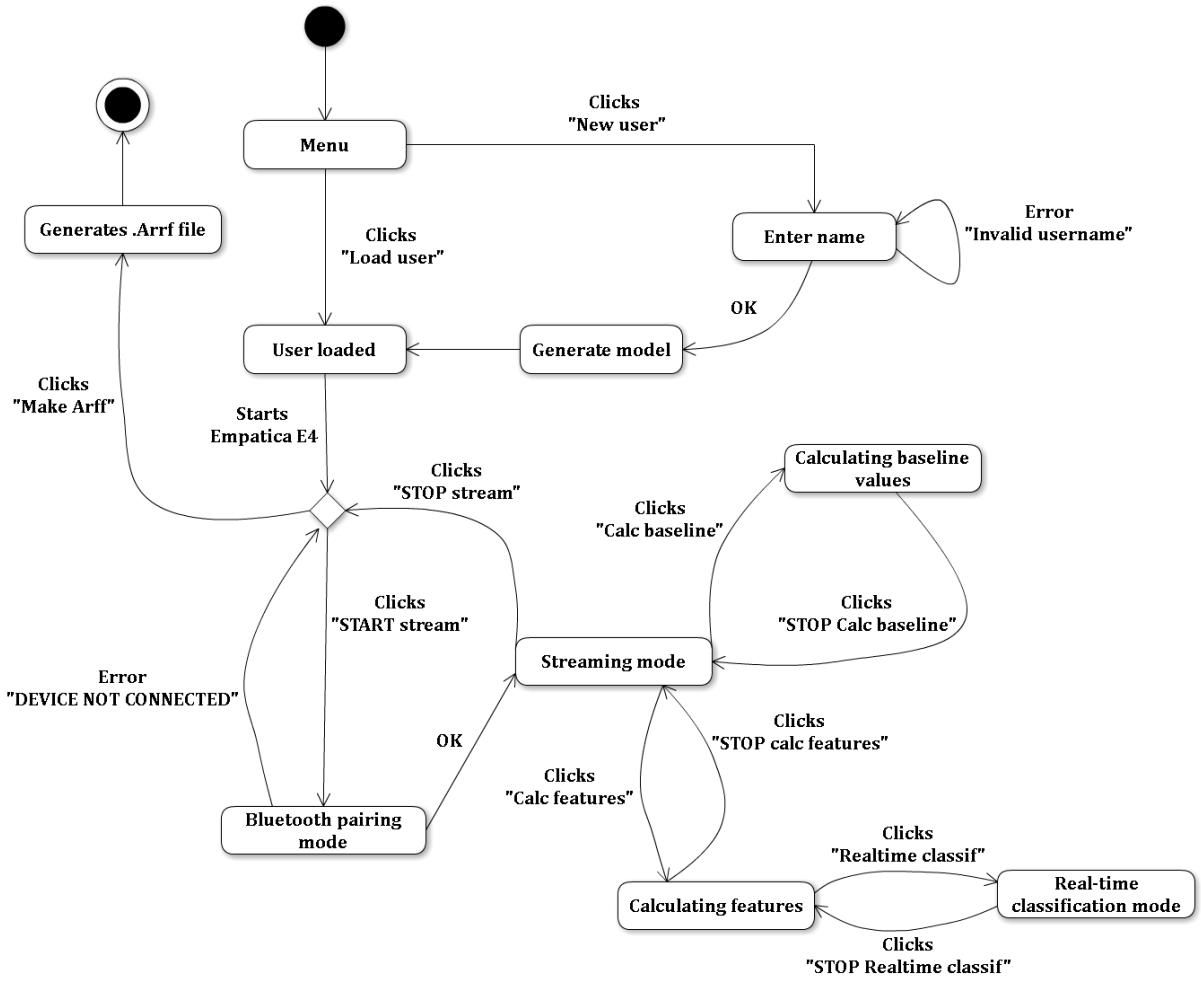


Figure 7: Application state diagram

4.2.2 Pre-processing

To understand the processing flow, we can look at Figure 8: Pre-processing state diagram where the process of gathering raw-data and then extract features is visualized in more detail.

Raw-data is not saved in any files for post-processing. Instead, raw-data is streamed in real-time to the application and when a window is filled for one type of sensor-data, filtering is done on the raw-data before performing feature calculations. When all features have been extracted for this window, a flag is set to verify the calculations are completed and ready to be stored. When this step is done for all three sensor data

buffers, all extracted features are merged into one '*Instances*' object along with the chosen class-label for this data instance (*anxiety*, *engagement* or *boredom*) and written to another buffer where all instances are saved. All functionality needed in order to understand this procedure, is depicted in the figure. The *TEMP* data is not depicted in much detail, but follows the same general steps as the *EDA* and *BVP* data, except that the *TEMP* data is never filtered. This procedure is iterated until the training-session is completed.

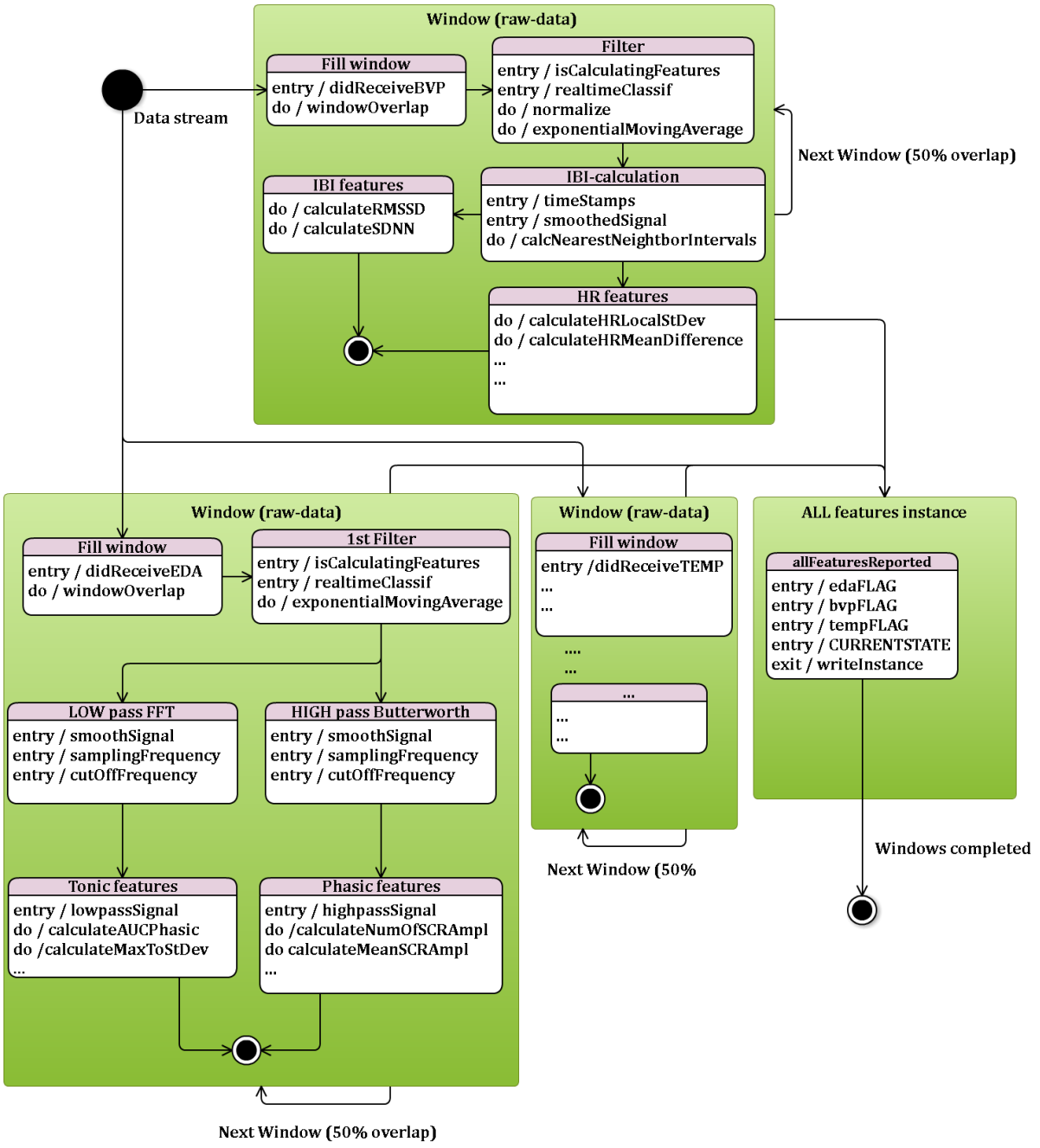


Figure 8: Pre-processing state diagram

4.2.3 Post-processing

After all data has been gathered and hence the pre-processing step is done, all data from this training session is saved into an *.arff* file. There is no need to specify saving destination, as the file will be automatically stored in the user-specific folder for the profile that is currently loaded. As shown at the '*Generate model*' step in Figure 7: Application state diagram. A machine-learning model is also created and saved as a *.model* file in the same folder upon profile creation. The file structure is shown in the figure below.

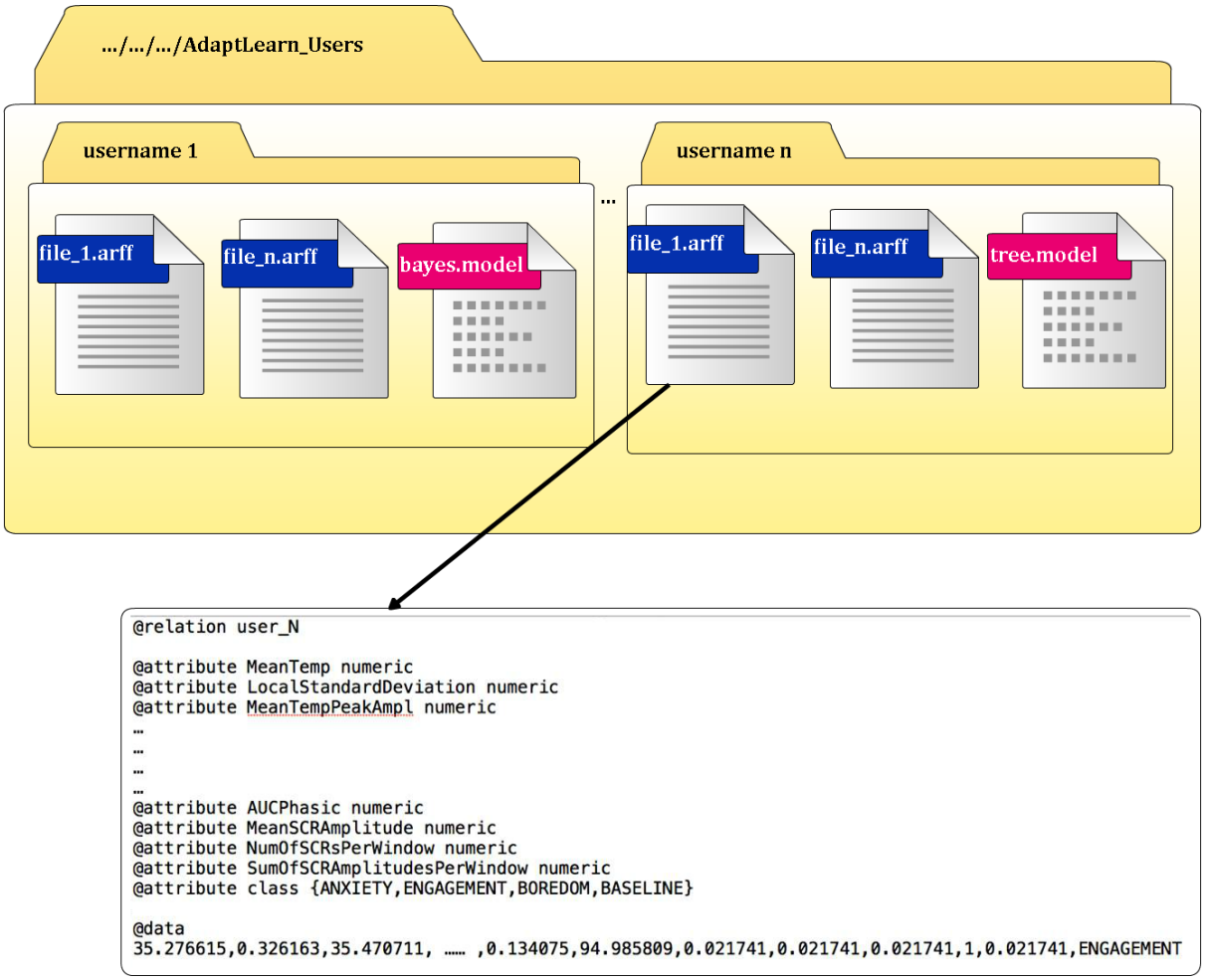


Figure 9: File structure

4.2.4 Real-time classification

When a user profile has been loaded successfully (see '*User Loaded*'-state in Figure 7: Application state diagram) the user-specific model will be automatically trained on all available .arff files for this user. This is done by using a functionality for incremental training available in the *Weka* API. This makes real-time classification possible, by making affective-state predictions based on this newly trained model. This mode is

activated if the *Realtime classif* button is clicked. It is still possible to save the data as an .arff file after this session is completed.

Section 5: Evaluation

This section explains the evaluation of how well the affective-state recognition system performed based on the math test on third graders (see Section 3.4). All statistical analysis was done using the *Weka Toolkit*.

The selection of good features are critical for modeling a problem in a meaningful way. With many features that are redundant, the likelihood that predictions will be based on disturbing artifacts increases. Redundant features in this case mean they basically provide the same information as other features so they can be removed without any loss of information. The same goes with irrelevant features (features which does not affect the result in a significant manner). Some machine learning algorithms, especially the instance-based ones like the k-nearest neighbor which determine classification and regression predictions by using small neighborhoods in the attribute space, might suffer greatly. Thus by having redundant attributes, the prediction can fail badly or end up in skewed results.

For slimming down the number of possible statistical results that would be of interest for this report, a broad statistical test scenario (referred to as '*Test A'* from now on) was performed where a group of classifiers were trained using different feature-selection methods explained below.

5.1 Feature selection

It is overly optimistic to assume the set of suggested features will give viable predictions in its entirety. Since an increasing number of predictors lessens the experimental power, I want to omit features that either correlates too much with other features, or have no significant influence of the predicted outcomes. This classic problem in making prediction models is called overfitting, when the set of features is large compared to a relatively small available set of data [15].

For this reason several feature selection methods have been used. *Method 1* is a *Wrapper*-method (returns the sub-set of features with most predictive power among all possible combinations). *Method 2* and *Method 3* are so called *Filter*-methods which means they rank each attribute based on its predictive power. For the tests (see Section 10.2), I choose to keep the top six attributes with most predictive power from *Method 2*

and top five attributes from *Method 3*. These numbers seemed to generate the best results.

5.1.1 Method 1: CfsSubsetEval

First a wrapper feature selection method called *CfsSubsetEval* was used which is based on a best-first-search. This method considers the individual predictive ability of each feature along with the degree of redundancy between them. This is done to evaluate the worth of a specific subset of features.

5.1.2 Method 2: InfoGainAttributeEval

This method measures the information gain (entropy) of an attribute with respect to the class. It is thus a method which uses a ranker search method. The ranker search means no specific sub-set of features will be calculated as in the case of *Method 1* above. Instead it just calculates the predictive power for each attribute.

5.1.3 Method 3: CorrelationAttributeEval

This is a so called Filter method which also ranks the attribute, based on its correlation with the class value. It calculates something called the 'Pearson product-moment correlation coefficient' which is a value between +1 to -1, where 1 is total positive linear correlation, 0 is no linear correlation and -1 is total negative linear correlation.

5.1.3 Sensor specific evaluation

For the sake of research interest in Test A, statistical analysis was also performed by evaluating classifier performance on each sensor-specific feature subset individually (e.g. TEMP-related features alone), as well as all combinations of the three sensors as well, e.g. EDA/BVP, EDA/TEMP, TEMP etc.

5.2 Pre-test evaluation after video-game session

To evaluate the system prior to the real tests, I trained the system during a 60-minute session of the intense online multiplayer game *Dota 2*. The hypothesis was that training an affective-state recognition system during a game generates strong emotional impulses where the distinction between states, e.g. *frustration* and *engagement* is often

very clear, as opposed to the likelihood that biometric readings from a learning scenario will not be as easy to differentiate between.

During this session, engagement (positive excitement) was reported as the current class label until a negative event happened which obviously triggered frustration-related stimuli e.g. sudden increases in pulse and body temperature.

After the session, a group of prediction algorithms was used for 10-fold cross validation. The accuracy and ROC-area values was beyond expectation and returned almost a perfect result. Reasons for this might be that the sensor output is strong because games generally generates discernible adrenaline rushes and other clear signs of changes in excitement (positive to negative and vice versa). A second possible reason is that self-reporting was done, so each affective state can be observed in matter of seconds (instead of having to answer periodically occurring questions, or fill out forms during/after test [11, 19, 39]).

These results are obviously not included in the results because of biased data. However, this test proved to be useful to tweak feature parameters and also to get a first glimpse of classification algorithm evaluation as a guideline for further testing.

5.3 Test A: Analyzing data from Antnässkolan

The following results comes from the analysis of Test A which can be viewed in its entirety in Section 10.2.2.

5.3.1 10-fold Cross-validation for one test subject

The test results from 10-fold cross validation including the full set of features generated ROC-area, F-measure values and accuracy rates over 97%. These results are overly optimistic because of the probable chance that the models have been over-fitted (as expected). Therefore these results have been discarded.

The combination of all features from the EDA/BVP sensors did not generate extraordinary results (best 91% accuracy), but have also been discarded due to risk of overfitting, and the fact that more realistic set of features generated about the same results anyway.

Here are the six best results in terms of best feature subsets (features that correlate the most and thus have the best predictive power), from 10-fold cross validation on S1

and S2 respectively with corresponding values of validity (ROC-area, Precision, Recall & F-measure):

BEST Feature selection methods	BEST Classifier	Accuracy	Validity (weighted average between the classes)			
			ROC-Area (mean)	Precision	Recall	F-measure
EDA & TEMP	k-NN	85.45%	0.854	0.852	0.855	0.853
Method 3	k-NN	78.18%	0.705	0.762	0.782	0.768
EDA	k-NN	76.36%	0.767	0.738	0.764	0.750
Method 1	k-NN	74.54%	0.707	0.748	0.745	0.743
Method 2	Random Forest	72.73%	0.737	0.703	0.727	0.715
BVP	Random Forest	69.09%	0.588	0.637	0.691	0.656

Table 5: Best Algorithm/Feature-set combo for S1

BEST Feature selection methods	BEST Classifier	Accuracy	Validity (weighted average between the classes)			
			ROC-Area (mean)	Precision	Recall	F-measure
EDA & TEMP	SMO	97.22%	0.974	0.974	0.972	0.972
Method 1	k-NN	91.67%	0.918	0.918	0.917	0.918
Method 2	Random Forest	88.89%	0.975	0.889	0.889	0.889
EDA	Random Forest	88.89%	0.969	0.889	0.889	0.899
Method 3	Random Forest	88.89%	0.937	0.889	0.889	0.889
BVP	k-NN	72.22%	0.726	0.730	0.722	0.717

Table 6: Best Algorithm/Feature-set combo for S2

For the full statistical analysis which Table 5: Best Algorithm/Feature-set combo for S1 and Table 6: Best Algorithm/Feature-set combo for S2 are based on, see 10.2.

Section 6: Discussion

This section discusses technical and theoretical problems along with considerations faced during the course of the project. First the problems raised in Section 1.3 are treated, following by further discussion.

Looking at the results from previous section, we can also see that the suggested system have the possibility to distinguish between affective states. A deeper analysis of the results can be found in Section 7.

6.1 Theoretical problems

(P1) Investigation of what psychological parameters affects our learning potential was done from the beginning of this project. As most of the defined problems needed to be solved, getting a basic understanding of how humans learn and what psychological factors affects our ability to process data, was crucial before getting into the technicalities on how to perform this project. Section 3 treats this question but to claim the question to be solved would be bold, since the subject itself is an on-going research area. However findings described in this section made it clear that by considering the mood and psychological factors of the individual, learning can be made much more effective and enjoyable.

(P2) The problem of defining affective states and map them to labels was solved by combining Ekman's and Priyadharshin's models for mood, feelings and state of flow into the three categories Anxiety, Engagement and Boredom. Reasons for this simplified labeling are available in Section
3.1 Determining how detecting affective states can help to improve the learning experience.

(P3) The majority of this report is rather theoretical and summarizes both related research as an overview of affective state recognition works (see Section 2), but also attempts to sum these technologies in solutions that are non-obtrusive and could be deployed easily in learning environments (at school or at home).

An important concern for this project was to identify suitable sensors, which desirably would reflect the activity of the nervous system in the sensor output. This task is all but trivial since all individuals are expected to react very differently according to the literature.

Even if these reflections would be accurately read by the sensors, nothing says with certainty they were triggered from the expected event but could be due to other factors such as personal background as conscious/subconscious thoughts or even external factors from the environmental setting during the data collection. Unlike speech recognition or facial expression recognition where correct class labels of a given data point is self-evident, the acquisition of a high-quality physiological signal database with confidence is an intricate task, as cited in [21].

As a remark, this question was solved by choosing the Empatica E4 wristband with the grounds that this device utilizes sensors, well-known in the literature for gathering high-quality biometric data, while it also can be worn for a long period of time, without being bothering. This cannot be said about e.g. facial electromyograms or electro-encephalograms where sensors needs to be attached to the head or face.

One final remark regarding distribute availability: As this report is written, the Empatica wristband pricing is very high for commercial uses, but similar technologies do exist for much more affordable prices.

(P4) It is important to work with the potential risk of inappropriate interactions in an intelligent learning environment. For instance, while interacting with an affective agent, it might be awkward for the learner if the agent is overly excited about the learner's progress so it lessens the motivation rather than encourages the student to continue the task [13]. A lot of meta-cognitive related aspects of ITS may be very useful to help the learning procedure. By making the learner aware of her own thinking patterns and progress, actions can be taken to regulate the thinking of the student. This is one of the core discussion topics in the [2] which is a motivational argument for AIEd. So, while metacognitive feedback to increase a learner's 'self-awareness' can be highly positive, caution must also be considered. By e.g. increasing the self-awareness when a learner is not "stuck" in a task, the learner may actually enter the "stuck"-mode rather than experiencing the positive aspects of frustration which is part of the learning procedure [13].

The short answer to this question is that external feedback may be rather annoying and take focus away from the task, while feedback affecting the system itself (internal feedback) is easier to find directly useful.

Furthermore, this question is discussed in Section 6.7 and the discussion continues in a small scale in Section 8.

6.2 Technical problems

(P5) The sensors used in this project are frequently used in affective-state recognition research. However, there is an on-going debate and continuous investigation for ultimate filtering methods and classification algorithms that can be used.

The methods used are described in great detail in Section 3.5. Many improvements can be made regarding the choices of filters, their parameters and algorithms for extracting biometric patterns from the chosen device (Empatica E4). Suggestions for improvements are discussed in Section 8.

(P6) The last problem to solve was to decide upon an ultimate set of features in combination with machine-learning algorithms to distinguish between the chosen affective state labels. This was done through an in-depth study of related work and lastly by testing and perform statistical analysis on the results from the test, using various feature-selection methods.

6.3. Threats to validity

Some thoughts, especially from the test-scenario in Antnässkolan (Test A) and flaws that affects the validity of these tests are discussed below:

6.3.1 Number of participants:

The number of participants surely makes evaluation of data harder and caution needs to be taken before drawing conclusions. Further testing where the same participant attends several tests, maybe in different test scenarios, would be interesting to generalize the trained model for each individual. In a future scenario, having these on-going tests for several days might improve the accuracy for a general model even further, since day-by-day fluctuations in the biometric data would thus be taken into consideration.

6.3.2 Participant bias

Even though external bias was considered in the experiments, as described in Section 3.4.1, one cannot be assured that some stimuli are not triggered due to e.g. thoughts, family history or events during the day and not from the test-scenario itself. The age of

the participant also affects the validity of the results, since system is trained based on observations and interviews during the test, in this case.

6.3.3 Class labels

During these tests, both participants mainly expressed two of the three affective states chosen for this project. In this case, having only two labels might have sufficed. This is yet another reason for further testing to be of interest.

6.3.4 External validity

As opposed to the scenario described in Section 5.2 where self-report was done directly to the application, one cannot underestimate the potential of the interviewer during the test. This problem is discussed in Section 3.1.1.

6.4 Practical problems related to the data collection

6.4.1 Sensors and noise

There are several scenarios that adds noise to the mean of the green channel of the PPG sensor which undoubtedly resulted in some inaccuracy of the IBI prediction which in turn affected the HR and HRV measures.

The biggest affecting factor for the PPG sensor seemed to be moving artifacts, as it was hard to get any signal at all when the wrist of the user was not still. Other noise can also be due to the factors treated in Section 3.4.

If the sensor is strapped too tightly, the pressure against the wrist can cut off the circulation and dampen or flatten the pulse wave completely. The BVP sensor has to be tight enough to hold firmly but not so tight as to cut off the circulation.

If the participant is very nervous, their stress reaction can cause the circulation in the hands to be drastically reduced. This can be seen as a very weak pulse or no pulse at all (as cited in [40]).

6.4.2 Alternative way to extract SCR's from the EDA

The sampling rate for the EDA signal was eventually too slow so that SCR’s were difficult to extract. There are numerous methods which can be found in the literature on how to extract skin conductance responses from the EDA signal. E.g. using a method based on an amplitude threshold measured per time-unit.

Ideal would be to have a much higher sampling frequency than the one provided by the Empatica E4. Modern algorithms that are delineating and separating phasic components from each other require sample rates at about a minimum of 200-400 Hz [33].

6.5 Ethical considerations

Biometric readings from a participant are considered to be highly personal data. This must in a real-world setting be considered with high caution. Disclosure of such data has been shown to be at potential risk for misuse in illegal activities in the literature.

The data gathered in this project was only used for statistical analysis and was not distributed outside this lab-setting. The participants also remain anonymous according to good ethical practice.

6.6 Emotion recognition for other purposes

Even though the results from the gaming-test described Section 5.2 were not included in this report, this test infer that the system might work in any general setting, independent of what state-recognition purpose it would serve. E.g. for gamification purposes or an eventual contribution to the E-health community by e.g. discovering panic attacks or anxiety in a patient.

6.7 External feedback from ITS's

Due to lack of time, the system in this project was never tested in conjunction with an ITS. However, some ideas gathered from the literature are discussed here.

If the sensors does not pick up signs up frustration even though a task has not been completed for a long time, we might assume the learner is interested, explorative, and curious. Or we don't care about that data at all.
Otherwise, the system should intervene appropriately by e.g. adjusting the difficulty-level of the assignment. This subject is treated consistently in the literature [e.g. 4, 36–40].

6.7.1 Music

In Giles 'A little background music, please' (as cited in [37]), studies suggests that music in the classroom may be beneficial to pupils' behavior and performance and also that

most pupils function very well with music in the background and that the right music at the right time can make them less stressed, more relaxed, happier and more productive. This is one of many cases when research suggests music to be a helpful factor in regulating affective states among students to something beneficial for the learning process.

6.7.2 Lightning settings.

Research regarding how different lighting settings, where exposure of lightning with different illuminances and correlated color temperature (CCT) can affect the quality of sleep, mood, alertness and what's interesting for our project; the learners perceived self-efficacy of the subject studied. In [43], a dynamic lightning system was used on Dutch elementary school children. They used instruments to measure the pupils' concentration and thus try to validate earlier findings of how effects of lightning conditions on children's concentration in elementary schools.

The results were not easy to evaluate, since overall, the positive effects of different lighting settings (artificial and daylight) seemed to be dependent on both numerous variables such as age and season. The tests found no correlation between lightning setting/concentration and gender. The tests were also conducted on "normal" children without learning disabilities.

6.8 New technologies appear!

While this writing was done, the first ever tutoring-application was released which considers the ever changing emotional state of the user such as *boredom* and *frustration* and adjusts the difficulty of the task to optimize the learning-experience (see [44]).

Section 7: Conclusions and remarks

This section describes final conclusions and the analysis of the results of the affective-state recognition system.

There are several factors to take into consideration when analyzing the results from Test A. First off, let us look at the distribution of classes in the data sets from S1 and S2.

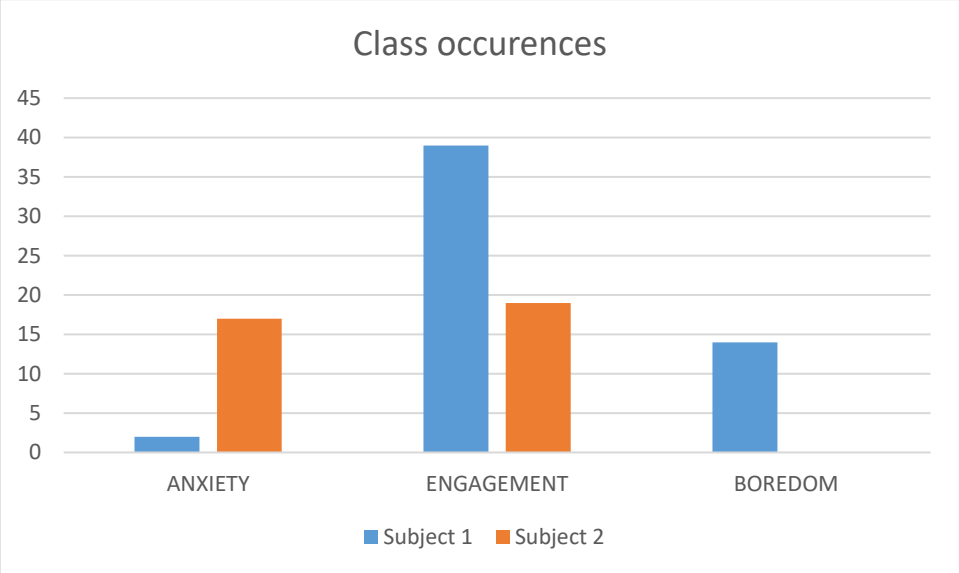


Table 7: Class occurences (Test A)

As depicted in the figure, S1 and S2 demonstrated widely different signs of emotional states during the tests. S1, who experienced the math exams to be relatively easy, experienced *engagement* throughout the test until *boredom* eventually took over. Only a short duration of *frustration* (*anxiety*) was shown at the start of one exercise when the participant did not understand the question properly.

The state of S2 fluctuated between *engagement* and *anxiety* as she struggled a fair amount with the exercises, asking for help to explain the tasks several times during the session.

As such, none of the participants showed clear signs of all three states. For this reason, the accuracy shown in Table 8: Best results (S1) and Table 9: Best results (S2) might be this high because there were never a case when the classifier had to distinguish between all three states, but only two. A conclusion is however that the classifier can distinguish between the two states of which each participant displayed, with a fair amount of accuracy.

Feature selection methods	Classifier	Accuracy	Validity (weighted average between the classes)			
			ROC-Area (mean)	Precision	Recall	F-measure
~~EDA & TEMP~~	~~k-NN~~	**~~85.45%~~**	**~~0.854~~**	**~~0.852~~**	**~~0.855~~**	**~~0.853~~**
Method 3	k-NN	**78.18%**	**0.705**	**0.762**	**0.782**	**0.768**

Table 8: Best results (S1)

The best results for S1 is shown above. The combination of *EDA & TEMP* features gave an accuracy of 85.45% with high validity values. Considering that this combination included 15 features, the accuracy does not differ too much with the second best result, where the k-NN algorithm was used with feature selection *Method 3* which only used 5 features.

Looking at the results from S2 below. The *EDA & TEMP* features in combination with the SMO algorithm generated an accuracy of 97.22%. As with the results from S1, this best numbers differ greatly with the second best results for each test. One might assume this may be because of overfitting since the set of *EDA & TEMP* features is big compared to the relative small training data-sets. The *EDA & TEMP* results are therefore decided to be neglected for both tests. If we look at the remaining result based on Method 3, an interesting observation can be done. As seen in Table 7: Class occurences (Test A), the distribution of classes for S1 is not evenly spread but the classes occurring belongs to *engagement* and *boredom*. The features chosen to be generate these results comes from all three sensors, which is not the case with the results from S1.

Going back to Table 9: Best results (S2) below, we see that 4 results are included after the *EDA & TEMP* result has been neglected. This is because they show about the same results, with mainly slight changes in validity scores. By scrutinizing what features were chosen for these results, it shows that Method 1-3 agreed upon almost exact same set of features, which all comes from the EDA signal. The class-occurrences from the S2 data-set were also very evenly distributed, which makes these results the most reliable.

These observation leads to the conclusion that *anxiety* and *engagement* have the potential of being classified with high precision, by essentially looking at patterns derived from the EDA signal, while distinguishing between *engagement* and *boredom* might be possible with good results by looking at a compilation of features from all the sensors in the Empatica E4.

BEST Feature selection methods	**BEST Classifier**	**Accuracy**	**Validity (weighted average between the classes)**			
			ROC-Area (mean)	Precision	Recall	F-measure
~~EDA & TEMP~~	~~SMO~~	~~**97.22%**~~	~~**0.974**~~	~~**0.974**~~	~~**0.972**~~	~~**0.972**~~
Method 1	k-NN	**91.67%**	**0.918**	**0.918**	**0.917**	**0.918**
Method 2	Random Forest	**88.89%**	**0.975**	**0.889**	**0.889**	**0.889**
EDA	Random Forest	**88.89%**	**0.969**	**0.889**	**0.889**	**0.899**
Method 3	Random Forest	**88.89%**	**0.937**	**0.889**	**0.889**	**0.889**

Table 9: Best results (S2)

Section 8: Future work

Future work can be done to augment the emotion-recognition system with e.g. an ITS but work can also be done to improve the performance of the system as well. This section describes a few suggestions.

Standardization or Normalization of sensor-data. To judge by the majority of results in relevant research literature, biometric sensor systems based on machine-learning produces the best results when used on the individual from which the classifier was trained. This minimizes a huge factor of randomness in the classifiers, should the classifier be general.

There are however ways to standardize the training procedure so the system can be "trained by some, and used by anyone". There are a lot of discussion on which methods are best for acquiring this.

In this system, some features based on amplitudes and magnitudes of raw-data (filtered or not), can be augmented to be proportion/percentage-based and this could possibly lead to higher accuracy in general use of the system (used on a random individual which had nothing to do with the training of the classifiers

Improve signal filter for the raw EDA signal. To improve the extraction process for SCRs, a Welsh-filter is suggested instead of the Butterworth, which might increase the accuracy of the system further. The Welch filter might may require more computational power than the solution for this thesis. However, for the relative small sets of data require, this might not a problem

Alternative training method after data collection. An idea could be to perform classifier evaluation using an entropy method by clocking the tests and mapping the timestamps of events during test to investigate biometric patterns for the participant. This means to consider the density of data points in relation to task instead of using supervised-learning. This could be valuable to verify the validity of the system.

Alternating the sizes of the sliding windows. For this thesis, assumptions were that the window-size would not have any significant impact on the outcome, based on results from the literature (see Section 3.5.3). However this is a blind assumption and should be tested properly.

Different window sizes are suggested in the literature. In some studies, the accuracy of a classifying algorithm is connected to the size of the windows (see [9]). The window

sizes in the literature ranges from 5 seconds to up to 10 minutes with varying results. In [19] where EDA, EEG and Gaze-tracking data were recorded for affective state prediction, window sizes from 5-60 were tested and the results indicated that the window size was of minor importance relative to the precision and recall values acquired for these kind of sensor data.

Combining more sensors. An example would be that a small increase in skin conductivity in combination with a smile (facial expression analysis) might imply satisfaction which could lead to some sort of celebratory feedback from the system [13].

Section 9: References

[1] S. D'Mello, "A selective meta-analysis on the relative incidence of discrete affective states during learning with technology.," *J. Educ. Psychol.*, vol. 105, no. 4, pp. 1082–1099, 2013.

[2] M. Griffiths and L. B. Forcier, *Intelligence Unleashed*. 2016.

[3] W. Chen and R. Persen, "A recommender system for collaborative knowledge," *Front. Artif. Intell. Appl.*, vol. 200, no. 1, pp. 309–316, 2009.

[4] M. F. Caro, D. P. Josyula, M. T. Cox, and J. A. Jiménez, "Design and validation of a metamodel for metacognition support in artificial intelligent systems," *Biol. Inspired Cogn. Archit.*, vol. 9, pp. 82–104, 2014.

[5] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school," *Front. Artif. Intell. Appl.*, vol. 200, no. 1, pp. 17–24, 2009.

[6] S. K. D. Mello, S. D. Craig, A. Witherspoon, and B. Mcdaniel, "Automatic detection of learner ' s affect from conversational cues," pp. 45–80, 2008.

[7] S. K. D. Mello, S. D. Craig, B. Gholson, S. Franklin, R. Picard, and A. C. Graesser, "Integrating Affect Sensors in an Intelligent Tutoring System," *Affect. Interact. Comput. Affect. Loop Work. 2005 Int. Conf. Intell. User Interfaces*, pp. 7–13, 2005.

[8] B. J. Zimmerman, "Self-efficacy: An essential motive to learn," *Contemp Educ Psychol*, vol. 25, no. 1, pp. 82–91, 2000.

[9] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, "Predicting affect from gaze data during interaction with an intelligent tutoring system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8474 LNCS, pp. 29–38, 2014.

[10] K. Muldner, W. Burleson, and K. VanLehn, "'Yes!': Using Tutor and Sensor Data to Predict Moments of Delight during Instructional Activities," *User Model. Adapt. Pers.*, pp. 159–170, 2010.

[11] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," *Proc. - Int. Conf. Softw. Eng.*, vol. 1, no. May, pp. 688–699, 2015.

[12] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence:

analysis of affective\nphysiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[13] R. W. Picard and W. Burleson, "Affective Agents: Sustaining Motivation to Learn Through Failure and a State of 'Stuck,'" in *Social and Emotional Intelligence in Learning Environments Workshop In conjunction with the 7th International Conference on Intelligent Tutoring Systems*, 2004.

[14] S. Schmidt and H. Walach, "Electrodermal activity (EDA) - State-of-the-art measurement and techniques for parapsychological purposes," *J. Parapsychol.*, vol. 64, pp. 139–163, 2000.

[15] D. Bondareva, C. Conati, R. Feyzi-Behnagh, J. M. Harley, R. Azevedo, and F. Bouchet, "Inferring learning from gaze data during interaction with an environment to support self-regulated learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7926 LNAI, pp. 229–238, 2013.

[16] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," *Proc. 12th Int. Conf. Entertain. media ubiquitous era - MindTrek '08*, p. 13, 2008.

[17] D. G. Cooper, I. Arroyo, B. P. Woolf, K. Muldner, W. Burleson, and R. Christopherson, "Sensors model student self concept in the classroom," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5535 LNCS, pp. 30–41, 2009.

[18] M. C. Desmarais and R. S. J. D. Baker, "A review of recent advances in learner and skill modeling in intelligent learning environments," *User Model. User-adapt. Interact.*, vol. 22, no. 1–2, pp. 9–38, 2012.

[19] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger, "Using Psycho-physiological Measures to Assess Task Difficulty in Software Development," *Proc. 36th Int. Conf. Softw. Eng.*, pp. 402–413, 2014.

[20] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *Int. J. Hum. Comput. Stud.*, vol. 65, no. 8, pp. 724–736, 2007.

[21] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. Journal Article, pp. 419–427, 2004.

[22] "Google Cloud Vision," 2016. [Online]. Available:

https://cloud.google.com/vision/.

[23] "Microsoft Oxford Project," 2016. [Online]. Available: https://www.microsoft.com/cognitive-services.

[24] "Google TensorFlow," 2016. [Online]. Available: https://www.tensorflow.org/.

[25] T. Belton and E. Priyadharshini, "Boredom and schooling: a cross-disciplinary exploration," *Cambridge J. Educ.*, vol. 37, no. 4, pp. 579–595, 2007.

[26] G. B. Moneta and M. Csikszentmihalyi, "The effect of perceived challenges and skills on the quality of subjective experience.," *J. Pers.*, vol. 64, no. 2, pp. 275–310, 1996.

[27] J. Chen, "Flow in games (and everything else)," *Commun. ACM*, vol. 50, no. 4, p. 31, 2007.

[28] A. Kushki, J. Fairley, S. Merja, G. King, and T. Chau, "Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites," *Physiol. Meas.*, vol. 32, no. 10, pp. 1529–1539, 2011.

[29] B. C. Manager, "E4 Wristband: User Manual," pp. 1–32, 2015.

[30] "Utilizing the PPG/BVP signal." [Online]. Available: https://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal.

[31] "What should I know to use EDA data in my experiment?" [Online]. Available: https://support.empatica.com/hc/en-us/articles/203621955-What-should-I-know-to-use-EDA-data-in-my-experiment-.

[32] R. Picard, S. Fedor, and Y. Ayzenberg, "Response to Commentaries on 'Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry,'" *Emot. Rev.*, vol. 8, no. 1, pp. 84–86, 2016.

[33] J. Braithwaite, D. Watson, J. Robert, and R. Mickey, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments," …, pp. 1–42, 2013.

[34] D. Leiner, A. Fahr, and H. Früh, "EDA Positive Change: A Simple Algorithm for Electrodermal Activity to Measure General Audience Arousal During Media Exposure," *Commun. Methods Meas.*, vol. 6, no. 4, pp. 237–250, 2012.

[35] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. Tai, "A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 387–426,

2014.

[36] M. J. Timms, “Letting Artificial Intelligence in Education out of the Box: Educational Cobots and Smart Classrooms,” *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 701–712, 2016.

[37] S. Hallam and J. Price, “Can the use of background music improve the behaviour and academic performance of children with emotional and behavioural difficulties?,” *Br. J. Spec. Educ.*, vol. 25, pp. 88–91, 1998.

[38] T. Force, H. rate variability European Society of Cardiology, physiological interpretation and clinical use Of measurement, T. F. of T. E. S. of C. American, T. North, and S. of P. and Electrophysiology, “Guidelines Heart rate variability,” *Eur. Heart J.*, vol. 17, pp. 354–381, 1996.

[39] J. a Healey, “Wearable and automotive systems for affect recognition from physiology,” p. 158, 2000.

[40] D. C. Combatalade, “Basics of HEART RATE VARIABILITY Applied to Psychophysiology,” *Signal Processing*, vol. 1, no. February 2010, p. 31, 2010.

[41] P. H. Winne and A. F. Hadwin, “International Handbook of Metacognition and Learning Technologies,” *Int. Handb. Metacognition Learn. Technol.*, vol. 28, no. January, pp. 197–211, 2013.

[42] J. Kay, P. Reimann, E. Diebold, and B. Kummerfeld, “MOOCs : So Many Learners , So Much Potential ...,” pp. 2–9, 2013.

[43] P. J. C. Sleegers, N. M. Moolenaar, M. Galetzka, A. Pruyn, B. E. Sarroukh, and B. V. D. Zande, “Lighting and student concentration,” *Light. Res. Technol.*, 2012.

[44] “Little Dragon: Emotional Learning.” [Online]. Available: https://www.indiegogo.com/projects/little-dragon-emotional-learning-mobile-education#/. [Accessed: 04-Nov-2016].

[45] A. Heraz and C. Frasson, “Towards a Brain-Sensitive Intelligent Tutoring System: Detecting Emotions from Brainwaves,” *Adv. Artif. Intell.*, vol. 2011, pp. 1–13, 2011.

Section 10: Appendices

10.1 Interview with teachers at Antnässkolan, Luleå, June 14, 2016

I started off by presenting the project and inform about the area of ITS's and its probable role of importance in the future for the education of the 21st century. To clarify my intentions on what kind of feedback I sought for, I also mentioned how big of an impact the state of our mind has during educational exercises.

The interview was planned with precise questions for the sake of crucial content but intentionally expanded to a more open discussion (workshop style). Here I summarize the most important content of the discussion results. Worth mentioning is, that I explicitly kept the theme or subject of the discussion to 'mathematics in school' instead of general education, even though much of the feedback can of course be applied to any subject in an educational environment.

Do students get continuous feedback about their progress in school?

Response: Not nearly as much as the teachers would want to give. Feedback is given as often as possible at least once every semester. However they all agreed that more frequent feedback would be much better for the learner.

Do you think this kind of feedback would be good to increase students' motivation to learn?

Response: YES. They all agreed here

How to deal with a student who seems to be…

…frustrated/angry due to the learning content?

Response: Of course this depends on the individual but basically adjusting the level or nature of the task often works, or even to give the student a short break to get her mind more balanced again.

…under-stimulated?

Response: This is a problem. The teachers agreed that there are two types of students who need more resources: The ones struggling and the ones experiencing the educational material to not be challenging at all. And mostly the

ones struggling are the ones that get more attention from the teachers which might leave the under-stimulated students with too little support.

How fast should you as a teacher be to intervene when a student appears to be frustrated/bored when she studying?

Response: The general answer is to intervene quickly to maintain the spikes on a student's flow-curve. If intervening quickly by e.g. adjusting the difficulty-level of the task (preferably without the student knowing about it), the probability that the motivation will persist, increases.

Is it common for you to use computer programs to teach children math today?

Response: Not as much, it's mainly old-school teaching with books.

What kind of feedback do these systems (computer-programs) provide the learner?

Response: Just simple words like "Great job!" or smileys. Oh, how the kids love smileys! What also seems to work are reward systems, like progress bars and giving students stars (rating) which they gladly wants to improve, by re-doing the tasks over and over until the score improves.

I asked about suggestions for external feedback (outside the screen feedback) like having a little robot shouting out in joy when the student does well. Or changing the light and putting on music etc.

Response: To sum up: "KEEP IT SIMPLE". The teachers all agreed that external feedback would only cause distraction, except for the fact that the robot would get destroyed (their words). It was also a matter of cost but even so they seemed to agree upon the distractive nature of such devices. Feedback on-screen seemed however to be much welcomed. Like simple encouraging text-messages, star-systems and progress bars. This kind of feedback is however most often incorporated in ITS systems of today.

Any other proposals?

Response: A common denominator was that gamification of learning systems seems promising and works very well generally. This includes having reward systems and progress bars, both for the individual but also to see the progress of the group as a whole.

Another suggestion which often works is to reward the students with mini-games (Tetris-style) in the middle of a learning session. Or an artificial buddy which motivates

and maybe does expressions like “hey let’s proceed with the task so we can play that game!” Artificial buddies are however a research field in itself.

A suggestion was also to keep track on *when e.g.* frustration arises, e.g. always after lunch or during certain kinds of task? This could be another parameter for a machine learning classifier that earns about an individual’s learning progress.

10.2 Test A

10.2.1 Algorithm and feature evaluation results from Subject 1 and 2 respectively

Feature selection methods	**Classifier**	**Accuracy**	**Validity (weighted average between the classes)**			
			ROC-Area (mean)	Precision	Recall	F-measure
EDA	Naive Bayes	70.91%	0.627	0.663	0.709	0.680
	k-NN	**76.36%**	**0.767**	**0.738**	**0.764**	**0.750**
	Random Forest	74.55%	0.822	0.704	0.745	0.718
	SMO	70.90%	0.500	0.503	0.709	0.588
BVP	Naive Bayes	58.18%	0.587	0.743	0.582	0.581
	k-NN	65.56%	0.563	0.618	0.655	0.632
	Random Forest	**69.09%**	**0.588**	**0.637**	**0.691**	**0.656**
	SMO	70.91%	0.517	0.503	0.709	0.588
EDA & TEMP	Naive Bayes	67.27%	0.624	0.669	0.673	0.668
	k-NN	**85.45%**	**0.854**	**0.852**	**0.855**	**0.853**
	Random Forest	76.36%	0.867	0.725	0.764	0.733
	SMO	72.73%	0.531	0.767	0.727	0.629
Method 1 (2 feats)	Naive Bayes	63.64%	0.679	0.598	0.636	0.615
	k-NN	**74.54%**	**0.707**	**0.748**	**0.745**	**0.743**
	Random Forest	69.09%	0.755	0.685	0.691	0.688
	SMO	70.91%	0.500	0.503	0.709	0.588
Method 2 (5 feats)	Naive Bayes	0.49%	0.614	0.704	0.491	0.473
	k-NN	70.91%	0.705	0.709	0.709	0.709
	Random Forest	**72.73%**	**0.737**	**0.703**	**0.727**	**0.715**
	SMO	70.91%	0.522	0.503	0.709	0.588
Method 3 (5 feats)	Naive Bayes	61.82%	0.726	0.705	0.618	0.622
	k-NN	**78.18%**	**0.705**	**0.762**	**0.782**	**0.768**
	Random Forest	76.36%	0.785	0.746	0.764	0.739
	SMO	70.91%	0.508	0.503	0.709	0.588

Table 10: Algorithm and Feature results (Subject 1)

Feature selection methods	**Classifier**	**Accuracy**	**Validity (weighted average between the classes)**			
			ROC-Area (mean)	Precision	Recall	F-measure
EDA	Naive Bayes	83.33%	0.898	0.852	0.833	0.832
	k-NN	80.56%	0.759	0.807	0.806	0.806
	Random Forest	**88.89%**	**0.969**	**0.889**	**0.889**	**0.899**
	SMO	86.11%	0.868	0.893	0.861	0.859
BVP	Naive Bayes	52.78%	0.604	0.538	0.528	0.522
	k-NN	**72.22%**	**0.726**	**0.730**	**0.722**	**0.717**
	Random Forest	58.33%	0.618	0.582	0.583	0.582
	SMO	44.44%	0.427	0.390	0.444	0.387
EDA & TEMP	Naive Bayes	83.33%	0.898	0.838	0.833	0.833
	k-NN	91.67%	0.915	0.918	0.917	0.917
	Random Forest	94.44%	0.981	0.944	0.944	0.944
	SMO	**97.22%**	**0.974**	**0.974**	**0.972**	**0.972**
Method 1 (3 Feats)	Naive Bayes	86.11%	0.920	0.863	0.861	0.861
	k-NN	**91.67%**	**0.918**	**0.918**	**0.917**	**0.918**
	Random Forest	88.89%	0.978	0.889	0.889	0.889
	SMO	86.11%	0.865	0.873	0.861	0.861
Method 2 (6 feats)	Naive Bayes	86.11%	0.913	0.873	0.861	0.861
	k-NN	88.89%	0.873	0.889	0.889	0.889
	Random Forest	**88.89%**	**0.975**	**0.889**	**0.889**	**0.889**
	SMO	86.11%	0.865	0.873	0.861	0.861
Method 3 (5 feats)	Naive Bayes	86.11%	0.944	0.873	0.861	0.861
	k-NN	80.56%	0.732	0.806	0.806	0.805
	Random Forest	**88.89%**	**0.937**	**0.889**	**0.889**	**0.889**
	SMO	86.11%	0.865	0.873	0.861	0.861

Table 11: Algorithm and Feature results (Subject 2)

10.2.2 Full detailed statistical evaluation of data from Subject 2 (S2)

Feature selection methods	**Classifier**	**Accuracy**	**Validity (weighted average between the classes)**			
			ROC-Area (mean)	Precision	Recall	F-measure
EDA (10 feats)	Logistic Regression	75.00%	0.828	0.777	0.750	0.759
	Naive Bayes	83.33%	0.898	0.852	0.833	0.832
	k-NN	80.56%	0.759	0.807	0.806	0.806
	REPTree	88.89%	0.861	0.889	0.889	0.889

	Random Forest	**88.89%**	**0.969**	**0.889**	**0.889**	**0.899**
	J48 Tree	**91.67%**	**0.885**	**0.918**	**0.917**	**0.916**
	SMO	86.11%	0.868	0.893	0.861	0.859
BVP (8 feats)	Logistic Regression	63.89%	0.641	0.646	0.639	0.638
	Naive Bayes	52.78%	0.604	0.538	0.528	0.522
	k-NN	**72.22%**	**0.726**	**0.730**	**0.722**	**0.717**
	REPTree	41.67%	0.401	0.409	0.417	0.410
	Random Forest	58.33%	0.618	0.582	0.583	0.582
	J48 Tree	47.22%	0.446	0.470	0.472	0.471
	SMO	44.44%	0.427	0.390	0.444	0.387
TEMP (5 feats)	Logistic Regression	61.11%	0.628	0.610	0.611	0.609
	Naive Bayes	58.33%	0.619	0.588	0.583	0.562
	k-NN	66.67%	0.639	0.667	0.667	0.667
	REPTree	50.00%	0.421	0.461	0.500	0.424
	Random Forest	**69.44%**	**0.748**	**0.694**	**0.694**	**0.694**
	J48 Tree	44.44%	0.404	0.440	0.444	0.441
	SMO	58.33%	0.574	0.583	0.583	0.572
EDA & BVP (18 feats)	Logistic Regression	83.33%	0.932	0.912	0.833	0.866
	Naive Bayes	80.56%	0.844	0.834	0.806	0.803
	k-NN	80.56%	0.788	0.807	0.806	0.806
	REPTree	88.89%	0.861	0.889	0.889	0.889
	Random Forest	**91.67%**	**0.964**	**0.918**	**0.917**	**0.917**
	J48 Tree	91.67%	0.885	0.918	0.917	0.916
	SMO	80.56%	0.813	0.834	0.806	0.803
EDA & TEMP (14 feats)	Logistic Regression	88.89%	0.953	0.894	0.889	0.889
	Naive Bayes	83.33%	0.898	0.838	0.833	0.833
	k-NN	91.67%	0.915	0.918	0.917	0.917
	REPTree	88.89%	0.861	0.889	0.889	0.889
	Random Forest	94.44%	0.981	0.944	0.944	0.944
	J48 Tree	91.67%	0.885	0.918	0.917	0.916
	SMO	**97.22%**	**0.974**	**0.974**	**0.972**	**0.972**
BVP & TEMP (13 feats)	Logistic Regression	61.11%	0.641	0.615	0.611	0.611
	Naive Bayes	58.33%	0.619	0.585	0.583	0.584
	k-NN	**66.67%**	**0.689**	**0.681**	**0.667**	**0.654**
	REPTree	52.78%	0.498	0.516	0.528	0.442
	Random Forest	61.11%	0.686	0.612	0.611	0.604
	J48 Tree	50.00%	0.573	0.497	0.500	0.497
	SMO	52.78%	0.522	0.524	0.528	0.522
All features (23 feats)	Logistic Regression	80.56%	0.921	0.829	0.806	0.817
	Naive Bayes	77.78%	0.835	0.795	0.778	0.776
	k-NN	88.89%	0.870	0.894	0.889	0.889
	REPTree	88.89%	0.861	0.889	0.889	0.889
	Random Forest	**97.22%**	**0.981**	**0.974**	**0.972**	**0.972**
	J48 Tree	91.67%	0.885	0.918	0.917	0.916
	SMO	91.67%	0.918	0.918	0.917	0.917

Method 1 (3 feats)	Logistic Regression	86.11%	0.916	0.863	0.861	0.861
	Naive Bayes	86.11%	0.920	0.863	0.861	0.861
	k-NN	**91.67%**	**0.918**	**0.918**	**0.917**	**0.918**
	REPTree	86.89%	0.870	0.889	0.889	0.889
	Random Forest	**88.89%**	**0.978**	**0.889**	**0.889**	**0.889**
	J48 Tree	91.67%	0.867	0.918	0.917	0.916
	SMO	86.11%	0.865	0.873	0.861	0.861
Method 2 (6 feats)	Logistic Regression	86.11%	0.932	0.863	0.861	0.861
	Naive Bayes	86.11%	0.913	0.873	0.861	0.861
	k-NN	88.89%	0.873	0.889	0.889	0.889
	REPTree	88.89%	0.861	0.889	0.889	0.889
	Random Forest	**88.89%**	**0.975**	**0.889**	**0.889**	**0.889**
	J48 Tree	**91.67%**	**0.885**	**0.918**	**0.917**	**0.916**
	SMO	86.11%	0.865	0.873	0.861	0.861
Method 3 (5 feats)	Logistic Regression	86.11%	0.915	0.863	0.861	0.861
	Naive Bayes	86.11%	0.944	0.873	0.861	0.861
	k-NN	80.56%	0.732	0.806	0.806	0.805
	REPTree	88.89%	0.861	0.889	0.889	0.889
	Random Forest	**88.89%**	**0.937**	**0.889**	**0.889**	**0.889**
	J48 Tree	**91.67%**	**0.885**	**0.918**	**0.917**	**0.916**
	SMO	86.11%	0.865	0.873	0.861	0.861

10.3 Physiological tools for input

A variety of sensors with the possibility to detect affective states and mood of the user are mentioned in the literature and is suggested below. Top of the line versions of these sensors have been part of similar experiments by e.g. the Affective Computing Group (ACG) at MIT, but also in more cost-friendly versions by the research team in [17] where they also focused on reducing the user-intrusiveness of the sensors (making the sensors less annoying). Most of the sensors will also be uncared-for because of the cost, but also for the complexity of use for this project. In a project of bigger scale, a mix of all sensor types below would possibly result in more accurate results in terms of recognizing mood, assuming knowledge and resources to quantify all the data in a manner that makes sense is at disposal.

Galvanic skin response sensor (GSRS). By using two electrodes, an imperceptibly tiny voltage can be applied across the skin to measure the conductance of the skin.

This can be useful for determining a person's affective state at some extent and have been used for this reason in several research projects [ref, ref, and ref]. The baseline for an individual's skin conductance does vary for many reasons such as gender and situation. The definition of skin conductance is however considered to be a function of the sweat gland activity and the skin's pore size. The sweat gland is partially controlled by the sympathetic nervous system of our body. For instance when we experience anxiety, there will be a swift increase in skin conductance due to increased activity in the sweat glands. This change happens in a matter of seconds and can thus be useful to determine a change in affective state for an individual. The opposite reaction of our body regarding skin conductivity would be reabsorption. As it turns out, this can be mapped to when we get startled. Also the level of arousal we experience can be reflected to an increase in skin conductivity, as the skin's capacity to conduct the current passing through the skin then increases. Suggested best uses of GSR sensors to measure electrodermal activity can be read about thoroughly in [33]. GSR-sensors have been used frequently by research teams in related experiments [11, 12, 19, 28, 32].

Pressure sensitive mouse. This type of sensor could be used to discover symptoms of frustration by detecting tension in the grip of the user. Unfortunately, there are no commercially sold products but like the pressure sensitive chairs, they have to be custom-ordered which will also result in high production costs. Occurrences of tests using this technology are mentioned in the literature [e.g. 15]

Pressure sensitive chair. Postural movement have been shown to indicate to correlate to interest (as cited in [13]). Other uses of postural information could be used by a ‘smart buddy’ (on-screen or a humanoid-like robot on the table?) e.g. to mimic the posture and movements of the learner to increase sympathy between the two. This is however slightly outside the scope of this project.

Face-expression recognition camera (expression analysis). Possibility to detect eyebrow raise, head nod and shake, mouth smiles, fidgets, and blink rate. Probably the most scalable sensor because a standard web-camera can be used to detect at least some of these expressions with the right software. Much research have already been conducted using cameras to detect mood.

Electrooculography (EEG) headset. By measuring electrical activity in the brain, it has been shown that specific frequency bands can be connected to mental states and e.g. to measure cognitive workload [19, 45].

Speech to Emotion software. Using biological signals to perform real time analyses of the user's emotion to sense stress, pleasure, or disorder.

10.4 Github repository

https://github.com/PerGrundtman/AdaptiveLearning.git

10.5 Figures

An exemplary SCR in a signal with phasic activity I recorded is shown in Figure 10 for illustration. **A**=SCR Start, **B**=SCR Max Amplitude, **C**=SCR Approximate endpoint.

Figure 11 shows a signal with no significant phasic activity and the slowly changing trend of skin conductance can be observed.

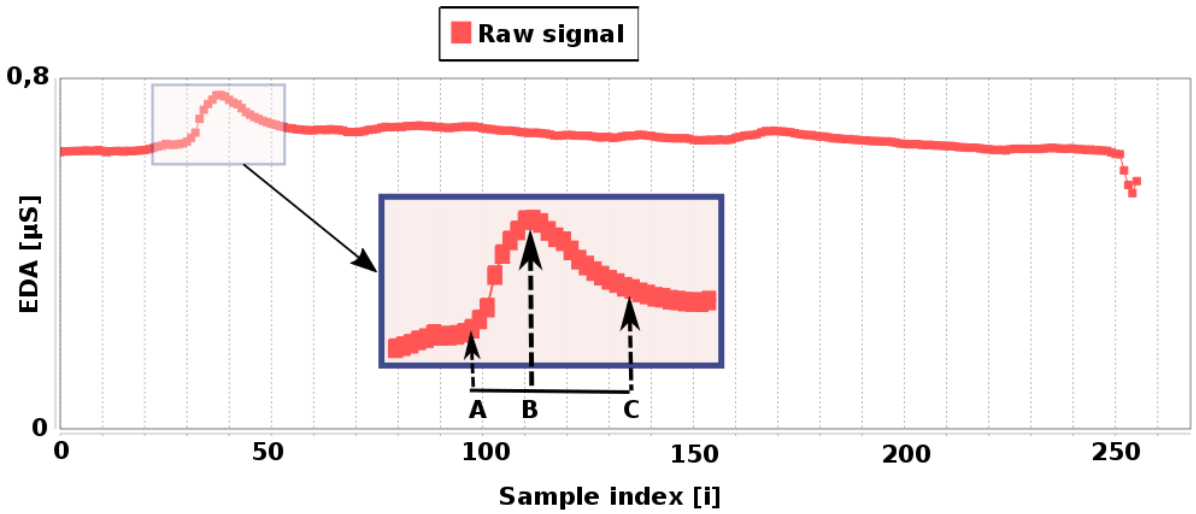


Figure 10: An exemplary SCR.

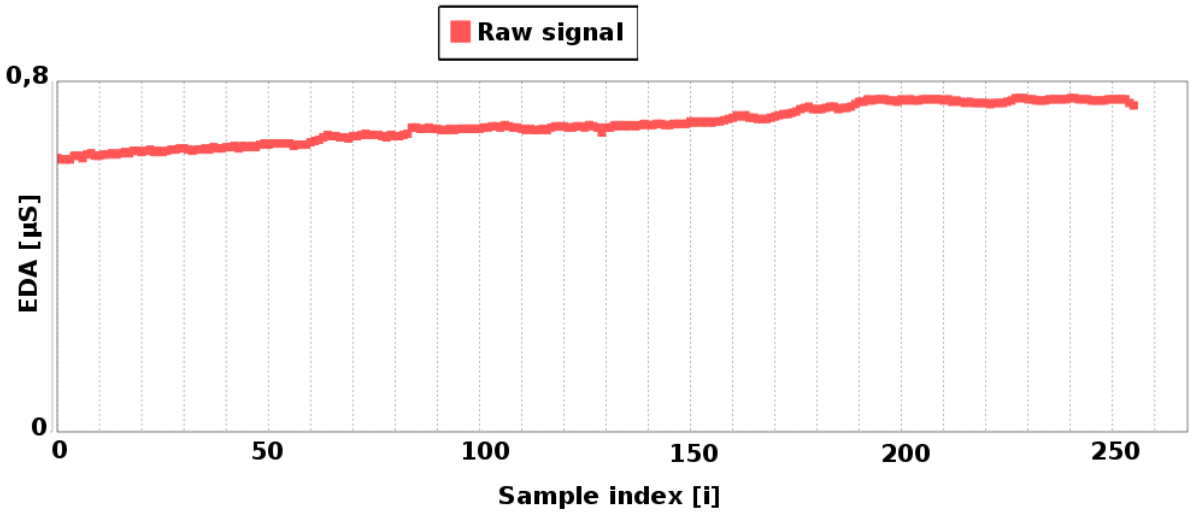


Figure 11: An exemplary part of an EDA signal which contains no SCR's but only the close to linear change in the signal.

Figure 12: An EDA signal containing much noise and Figure 13: Smooth EDA signal with an observed SCR shows examples of the EDA signal after they have been filtered with a FFT low pass filter.

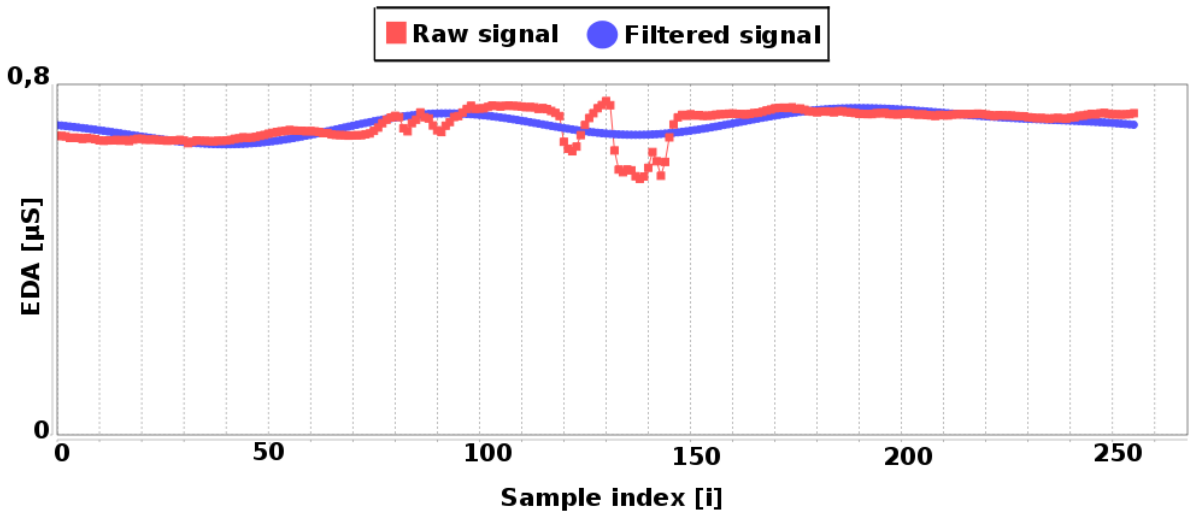


Figure 12: An EDA signal containing much noise

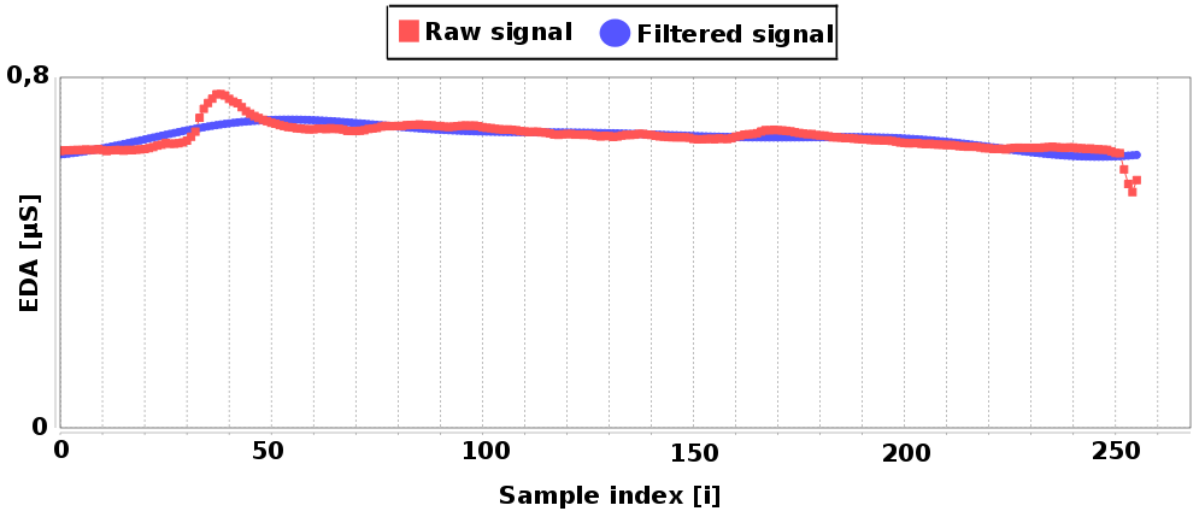


Figure 13: Smooth EDA signal with an observed SCR

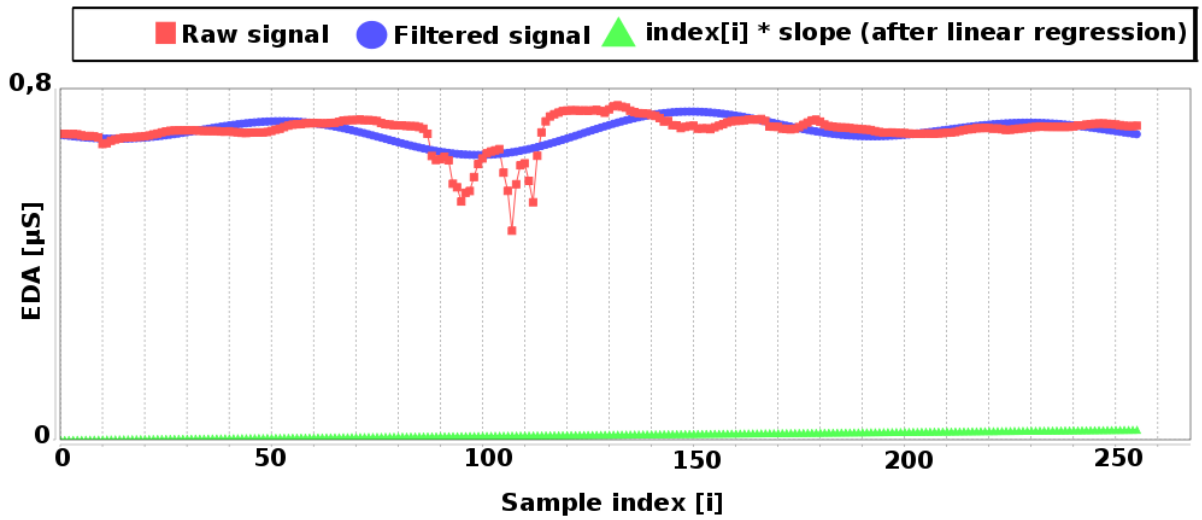


Figure 14: Approximation of mean slope of the window with a noisy signal

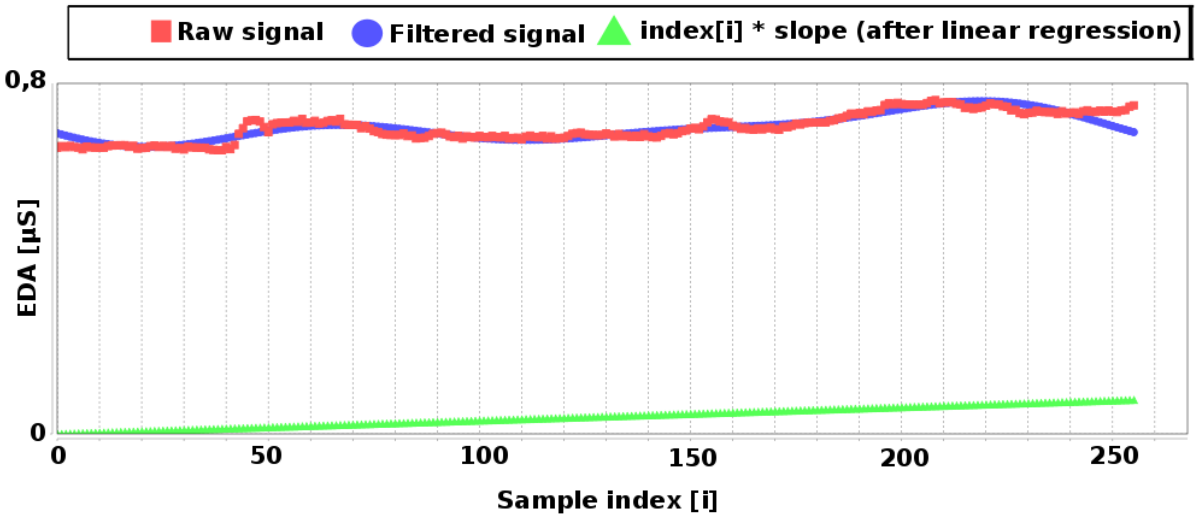


Figure 15: Approximation of mean slope of the window with a clean signal

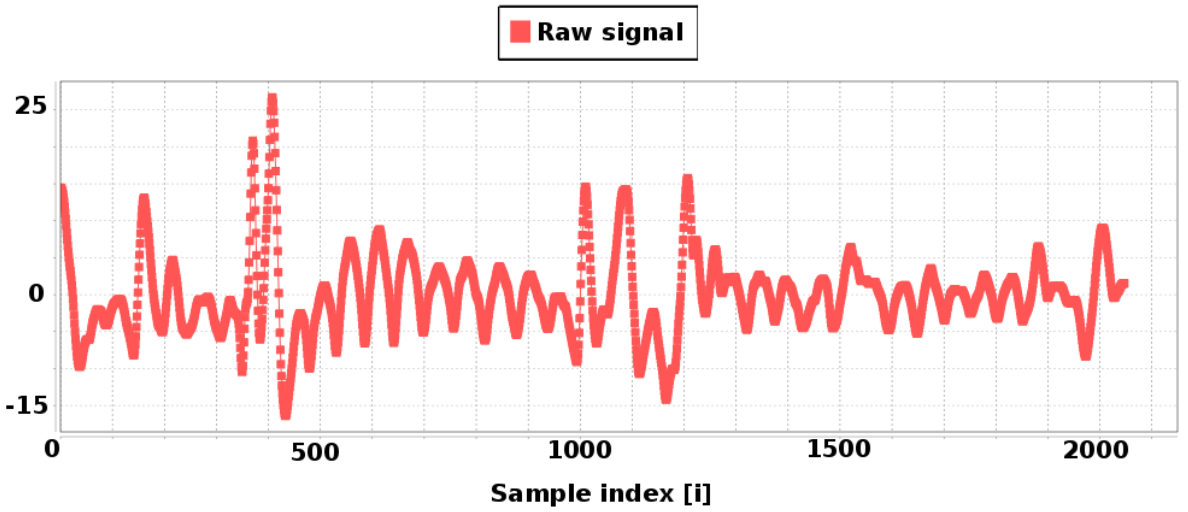


Figure 16: Example of a BVP signal