Sensor-Based Recognition of Engagement During Work and Learning Activities

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Elena Di Lascio

under the supervision of
Prof. Silvia Santini

November 2021

Dissertation Committee

Prof. Gabriele Bavota	Università della Svizzera italiana, Lugano, Switzerland.
Prof. Fabio Crestani	Università della Svizzera italiana, Lugano, Switzerland.
Prof. Jakob E. Bardram	Technical University of Denmark, Copenhagen, Denmark.
Prof. Fahim Kawsar	Bell Labs Cambridge, UK and TU Delft, Netherlands.
Prof. Akane Sano	Rice University, Houston, Texas.

Dissertation accepted on 15 November 2021

Research Advisor	PhD Program Director
Prof. Silvia Santini	**Prof. Walter Binder**

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

ElenaDiLascio

Elena Di Lascio
Lugano, 15 November 2021

Abstract

Personal computing systems like e.g., laptop, smartphone, and smartwatches are nowadays ubiquitous in people's everyday life. People use such systems not only for communicating or searching for information, but also as digital companions, able to track and support their daily activities such as sleep, food intake, physical exercise and even work. Sensors embedded in personal computing systems enable the continuous collection of heterogeneous data about their users. Location, heart rate and more, can nowadays be measured reliably using such sensors. Processed with machine learning and data analytics techniques, sensors' data can be used to infer users' information such as the type of activity, the behaviour and even the affective states.

In this thesis, we investigate the feasibility of using data derived from personal devices to automatically recognize the affective state of *engagement* as it occurs during daily activities. We focus on the specific use cases of inferring students' engagement during learning activities and knowledge workers' engagement during work activities.

Engagement, generally considered in terms of the emotional and attentional involvement into an activity, is a well-known predictor of learning outcomes and job performance. Consequently, *engagement-aware systems* able to sense, recognize and promote engagement have a huge potential for improving the learning and work experience.

Measuring engagement has been for years central focus of research in psychology. Traditional methods, such as self-reports and observations, requiring significant manual effort from researchers and study participants, have been used for years to derive knowledge about engagement. Bulky devices measuring physiological parameters e.g., electrodermal activity and heart rate variability, have been also used to study engagement from a physiological perspective mostly in laboratory settings or during pre-defined activities. Today, taking advantage of the availability of personal devices and the sensors they are equipped with, computer science researchers are investigating methods for automatically measuring engagement in everyday activities, with little or no effort from users.

Despite the knowledge gained from years of research on engagement, its automatic assessment using sensor data is a challenging goal. Indeed, there is no a pre-defined mapping between sensor data and engagement, and it is not clear what transformation and combination of data can provide a reliable engagement assessment.

Furthermore, engagement definitions and its expressions are context-dependent, thus a system aiming to infer engagement should be able to retrieve and use information about the user's context. However, in the work environment, context information such as the type of activity, are difficult to infer. People use several tools to perform their tasks, work in different locations, alone and with others, making the activity inference challenging.

In this thesis, we target two main problems: (1) the engagement recognition problem; and (2) the activity recognition problem.

To evaluate our approaches we designed and ran three user studies and collected data in laboratory settings and in the *in-the-wild* e.g., during lectures in the classroom and during actual work days. Further, we performed an extensive data analysis.

Specifically, we first address the sensor data transformation and combination problem for engagement recognition. To this end, in the first study presented in this thesis, we leveraged electrodermal activity data and proposed a method for translating findings from educational research into sensor data representation, i.e., features. We then used the features in input to machine learning algorithms with the aim of recognizing students engagement during lectures. In the second study, we proposed a novel method to recognize a behavioural expression, i.e., laughter, that can be used for recognizing engagement. We leveraged typical physiological and body movement reactions of laughter, and quantify them using sensor data gathered from wristbands. In the third study, we investigated sensor-fusion strategies based on traditional machine learning and deep learning, and combined physiological data i.e., electrodermal activity and cardiac activity, together with context information to recognize engagement during work activities.

Second, we address the problem of recognizing activities in the workplace. To this end, we proposed a method to combine behavioral expressions such as physiological activation, physical movement, laptop and phone usage. We performed a thorough analysis and investigated which type of device and sensor data bring relevant information, especially for distinguishing between work and break activities.

We believe that the insights and technical contributions of this thesis aim to enable the design and development of engagement-aware systems able to support people during their daily activities.

Contents

Figures

Tables

Chapter 1

Introduction

Computing systems have dramatically changed over the past decades. Technological advances have enabled a tremendous increase of computing power and battery capacity, as well as a miniaturization of the devices going from "room size" to "palm size". A major consequence of such technological advances are the development and widespread adoption of mobile and wearable devices such as laptops, tablets, smartphones and smartwatches. Thanks to the proliferation of these easily accessible devices, used anywhere and at any time, computing systems have become ubiquitous.

Mobile and wearable devices allow a continuous and unobtrusive collection of large amount of *heterogeneous data* in every day life situations. Data gathered from sensors embedded in personal mobile and wearable devices, can be processed and provide information about, e.g., people behaviour, activity and even affective states [5].

For instance, inertial sensors can be used to determine the activity users are doing [6]; location and communication sensors can be leveraged to understand their mobility patterns [7; 8] and location preferences [9]; physiological data such as electrodermal activity and heart rate variability can be exploited to infer stress [10; 11] or engagement [12].

Emerging computing systems, such as *personal informatics system* [13], utilize information about individuals' behavior, activities, and affective states to support them in their daily lives. For instance, such systems could help individuals maintaining a healthy life style by promoting self-reflection or even inducing behavioral changes with *just-in-time adaptive interventions (JITAIs)* [14].

These technologies are already in widespread use to track, e.g., food intake, fitness, sleep quality and more [15; 16]. An example of component of personal informatics system is the commercial Fitbit wristband[1], this device enables users to track and improve their physical activity by setting goals (e.g., 10,000 steps per day) and through competition with friends [15].

Recently, there has been a growing interest from researchers and practitioners, in enhancing computing systems with *emotional intelligence* [17]. Affect and emotions play a fundamental role in several aspects of daily life such as decision making, communication, memory, physical health, and overall psychological well-being [18]. As a consequence, *emotion-aware systems* able to sense, interpret, adapt and even express emotions and, more generally affective states, have the potential for communicating more effectively with users, and for understanding, and supporting them [19; 20; 17]. Prof. Rosalind Picard first clearly outlined the potential and challenges of emotion-aware systems, and promoted the new field of research called *Affective Computing* [20].

Most of the initial efforts in the Affective Computing field were conducted as controlled studies in laboratory settings [21] in which participants were asked to wear bulky sensors and perform pre-defined tasks. The affective *features* were derived from video, audio and physiological signals [21] and used to create affective *models*. Today, the advances of mobile and wearable technologies as well as their widespread use in daily life, open plenty of new possibilities for Affective Computing, enabling the creation of systems able to infer and respond to individuals' affective states in real settings and during daily activities [17].

In this thesis, we investigate how a specific affective state, i.e., *engagement*, can be automatically recognized using data gathered from personal devices, and how information about engagement can be integrated in personal informatics systems to support people in their daily activities.

Several definitions of engagement exist and depend on the context [22; 23]. However, it is widely accepted that "*the way users engage in an activity is an essential component of their experience with the activity*" [24]. When engaged in a task, individuals generally have an emotional and attentional involvement with the task [24].

In this thesis, we focus on the problem of inferring engagement during activities. In particular, we focus on *students* and *knowledge workers* and on inferring their engagement during *learning* and *work activities*. The ultimate goal of the thesis is to advance the understanding of how engagement-aware systems can be designed and utilized, and to provide a set of methods and tools for their design and development.

[1] `https://www.fitbit.com/au/home`

1.1 Motivation

Engagement is a crucial component of the overall quality of the work and learning experience [22; 23]. Indeed, engagement has been linked to better learning [25] and increased performance [26].

Knowledge about engagement can be used by a computing system to understand if and when is an appropriate moment to send an intervention to promote engagement (e.g., when prolonged disengagement is assessed), to measure the effectiveness of an intervention (e.g., whether the intervention strategy increased or decreased the level of engagement) or to display information to the users in order to enable self-reflection (e.g., in which activity, moment of the day they felt more or less engaged).

Devising methods for automatically recognizing engagement is a key enabler to the mentioned intervention strategies and it is the core goal of this thesis.

The automatic recognition of engagement is a challenging task. Indeed, there is not an immediate and pre-defined mapping between sensor data and engagement and it is not known which sensor representation and combination of information could provide reliable engagement assessment.

In this thesis, we adopt a data-driven approach to discover quantitative relations between sensors data and engagement. We leverage data about electrodermal activity, cardiac activity, and physical movement, all gathered from physiological and inertial sensors embedded in non-intrusive wearable devices (i.e., wristbands). We derive and propose a set of features to quantify the physiological activation and a multi-modal expression (i.e., laughter) that can be used for detecting students and workers' engagement during their activities.

As already mentioned, effective affect-aware systems should consider the context of the user [27; 19]. A relevant context information to consider when creating interactive systems is the type of activity performed by the user [27]. The type of activity can provide not only relevant information for assessing engagement but can be exploited also as additional information to provide to the user [28] or as a trigger for interventions [3].

However, given the complexity of the work environment, recognizing what the user is doing is a challenging problem. Knowledge workers use different tools, work on several tasks, in different locations, alone and with others. The choice and combination of the appropriate devices and sensors to capture the complexity of the work environment is yet unexplored. A trade-off between invasiveness and completeness needs also to be considered.

Motivated by the importance of retrieving information about the users' activity, in this thesis we also investigated the role of data gathered from personal devices (i.e., laptop, smartphones and wristbands) to automatically recognize activities in the workplace. Further, we explored the role of context information in the assessment of workers' engagement.

The remainder of this chapter is organized as follows. Section 1.2 describes the engagement recognition and activity recognition problems and the associated challenges. Section 1.3 summarizes the research questions addressed in this thesis. An overview of the research methods we used and the contributions we made are presented in Section 1.4 and Section 1.5 respectively. Section 1.6 presents the list of publications resulted from the work described in this thesis. We conclude the chapter with the outline of the thesis.

1.2 Problem statement and open challenges

1.2.1 Engagement recognition problem

The engagement recognition problem is an example of the broad affect recognition problem. *Affect recognition (AR)* is a fundamental building block of the Affective Computing (AC) research field [29]. AR refers to the study of methods for automatically recognizing affective states based on *observables*, i.e., sensor recordings, and it can be conceptualized as a pattern recognition problem [29]. Recognizing the affective state of the user allows to answer the question: *"How is the user feeling?"*

Researchers showed that engagement manifests itself through several non-verbal *behavioural expressions* such as facial expressions, gestures, postures and movement, physiological activation and multi-modal expressions. Using specific sensors, e.g., cameras, inertial sensors, physiological sensors, it is possible to capture different *modalities* – e.g., face, body – and quantify peculiar behavioural expressions or *cues* of engagement – e.g., facial expressions, body movement, physiological activation. Sensor data is processed and transformed into *features*, which describe the behavioural cues of engagement (e.g., in terms of presence, absence, intensity, or descriptions). The features are paired with a target output i.e., *engagement labels* and used as input to supervised *machine learning (ML)* models [30]. The ML models are trained to learn the mathematical functions that map the input with engagement labels. The learned models are then applied to inputs extracted from new data to determine the engagement level of the individual.

One of the factors that makes engagement recognition a challenging task is the uncertainty related to sensor data *representation*, i.e., how meaningful information from sensor data can be extracted to quantify engagement's expressions. In particular, it is not evident what *transformation* needs to be applied to sensor data and which *combination* of sensor data would produce a reliable engagement assessment.

In machine learning, the features extraction or *features engineering* – i.e., the process of applying transformations to sensors' data – is critical to ensure the success of the assessment, especially for models with a shallow structure [31].

A common approach to the design of features is to embed *expert knowledge* in their design. Being AC an interdisciplinary field grounded on research from psychology and neuroscience, we argue that for creating representative features, inspiration from theories derived from these research fields should be taken. In this thesis we rely on engagement definition and theories from educational research and psychology such as the "Flow Theory" conceptualized by the psychologist Csikszentmihalyi [32]. Further details about engagement definitions and theories used in this thesis are described in Chapter 2.

Most of the previous approaches in sensor-based engagement recognition, do not consider existing theoretical knowledge for the design of features from physiological signal such as electrodermal activity (EDA). For instance, studies targeting the learning setting [33; 34], rather infer engagement using statistical description of the signals in the temporal or frequency domain or are based only on specific signals' characteristics such as EDA peaks i.e., *skin conductance responses (SCRs)* [35].

Despite these features provide complementary information, they do not take into account the relation between the expression of engagement and the theoretical explanation underneath, preventing a proper interpretation of the information derived from the sensors. Understanding how to map objective sensor data into explainable representation is still an open challenge.

When interacting, people naturally use different modalities to convey their emotional reactions. Each of the modalities bring unique fingerprints of the affective state. For this reason, in the Affective Computing field, it has been widely exploited the complementary information provided by different sensors through their combination, we refer to this approach as *multi-modal affect recognition* [29; 36]. Several strategies exist for combining sensor data at different levels, from very early stages in terms of raw data, to fusion of independent models [36].

Despite the great advancements in data fusion techniques presented in literature, fusing data from different sources is still an open challenge. For instance, the choice of the type and number of sensors and devices to use is not known a priori. The choice of the sensors should be guided by the trade-off between intrusiveness and quality of the information retrieved. The quality of the information can be considered in terms of relevance of the information for the problem addressed and in terms of noise in the signals generated by the sensors (e.g., due to physical movement or sensor malfunctioning).

Despite the good progress in data fusion in AC, several approaches treat multimodal expressions such as laughter, using a single modality [37], using data collected with privacy-invasive sensors as cameras and microphones [38; 39], or using multiple invasive sensors attached in different parts of the individuals' face and body [40].

A major challenge in multi-modal affect recognition consists in defining which data should be fused at which level [36]. Multi-modal fusion strategies used for the inferring workers' engagement using physiological sensors [41; 28] rely on the use of hand-crafted features and machine learning classifiers with a shallow structure [41; 28]. Despite the promising results obtained, these approaches fail to exploit the additional information derived by the complex, non-linear, low-level interaction that can occur simultaneously across different modalities [42; 43]. Deep learning presents a possible solution to this problem, allowing an automatic hierarchical construction of features within and across modalities [42] that can be merged at different stages [44].

In addition to the behavioural expressions of an individual, the current situation or *context* has an impact on the engagement recognition problem. Indeed, affective states in general and engagement in particular are context-dependent [19]. Information about context is particularly critical when aiming to build affective models using data collected "in-the-wild" where people act freely and are continuously exposed to different situations and interactions.

Particular behavioural expressions of engagement can be expected in specific situations but not in others. For instance, a worker could laugh when interacting with colleagues during a meeting or when watching a video during a break, while a laughter episode could be less likely to happen when the worker is focused on a work activity such as *coding*.

Individuals might be more engaged when dealing with particular activities or during specific parts of the day. Further, the physiological reactions might vary depending on the type of activity performed, and a change in the perceived engagement might not always result in a change of the physiological signals. For instance, during more "arousing" or "physiological demanding" activities (e.g., coding or attending meetings) the physiological reactions might be more pronounced compared to when workers are dealing with a less demanding activity, while the perceived level of engagement might be the same.

Despite its importance, retrieving and embedding context information into the assessment of affective states such as engagement, are significant open challenges [19]. For instance, what type of context information can be meaningful to consider and how to fuse it with physiological data is still unclear.

Most of existing approaches in the automatic recognition of workers' engagement are either conducted in laboratory settings (where the context is fixed) [45; 41], do not consider the role of context [46; 47] or indirectly retrieve it from digital activities only [28].

1.2.2 Activity recognition in the workplace problem

The problem of activity recognition belongs to the *Human Activity Recognition (HAR)* research field. HAR "aims to recognize activities from a series of observations on the actions of subjects and the environmental conditions" [48]. As for engagement recognition, also HAR can be conceptualized as a pattern recognition problem. Recognizing the activity the user is performing allows to answer the question: *"What is the user doing?"*

HAR is involved in the development of several applications, such as video surveillance, home monitoring, healthcare and human-computer interaction [48]. As discussed in [27], the activity the user is performing represents one of the main *primary* types of context information to consider when designing interactive computer systems. Systems able to sense and adapt to user's context are known as *context-aware* systems.

In this thesis, we focus on recognizing workers' activities in the workplace, and in particular we aim to distinguish between *work* and *break* activities.

One reason why activity recognition in the workplace is challenging is the complexity of the work environment. Workers use different tools (physical and digital), perform multiple tasks, work alone or with others, and in different locations (e.g., in their or others' offices or at home). As a consequence, understanding which sensor(s) or device(s) to use to gather data as well as which technique to use for inferring the workers' activity is not straightforward.

While the use of several sensors for a precise inference of workers' activities might seem a promising solution, it is important to consider that the use of more data sources do not necessarily imply more information. Indeed, data from similar sources might be redundant or not significant for the problem at hand. Further, a trade-off between completeness of information and users' burden needs to be considered. Indeed, users might feel overwhelmed by the use of multiple instruments and by the perception of being tracked all the time, and consequently stop using the system.

Several existing approaches that aim to quantify the activities in the workplace, focus on *digital activities* only [49; 3; 50], and determine the type of work and break activities using only data derived from computer usage (e.g., type of application used or website visited, interaction with keyboard or mouse). These approaches use pre-defined assumptions and heuristics for inferring the workers' activity, e.g., a worker takes a break when she is not interacting with the computer [49; 3; 51].

However, relying on heuristics prevents the recognition for being flexible to intra-activity variations. For instance, a worker can take a break using the laptop and watching a video, or can take a walk outside. Using heuristics based on laptop usage only lack of distinguishing between these situations. Further, the validity of the proposed heuristics in real settings has not been validated, by for instance letting workers confirm whether the system correctly assessed when they were taking a break nor is the correctness of the assumptions made verified otherwise.

Recently, more flexible, machine learning based approaches have been used to determine digital activities [52; 53]. However, considering only digital activities limits the possibility of properly capturing the multi-facet complexity of the work environment leading to possible miss-classification especially of activities that do not strongly require the use of the laptop as meetings, reading or taking a break.

Other approaches have used sensors as cameras [54], microphones [54], RF-radar [55] or Bluetooth beacons [56], placed in the workers' room, on the furniture [55] or in the building [56] to determine workers' work or break activities [56; 55; 54].

However, approaches relying on cameras and microphones, may be considered invasive in the workplace setting especially when other people risk being unintentionally monitored. Approaches relying on sensors in the environment require additional infrastructures to be placed and are not flexible in terms of work activities locations. For instance using sensors in the room might miss the difference between ones not being in the office because of being in meeting in another room or because taking a break outside.

1.3 Research questions

Following the considerations reported in the previous section, we now formulate the research questions addressed in this thesis. The broad question this thesis aims to address is:

"Is it possible to recognize students and knowledge workers' engagement using data gathered from mobile and wearable devices?"

To address this broader topic, and building upon the considerations summarized in the previous section, we formulate three specific research questions:

- **RQ1** How can features representing behavioral expressions of engagement be derived from physiological and movement data?
- **RQ2** How can information about context a) be fused with physiological data and b) impact the recognition of workers' engagement?
- **RQ3** How can activities in the workplace be accurately recognized?

1.4 Research method

To address the research questions formulated above, we propose a data-driven approach and conducted three user studies. In this section we outline the methodological approach we used for the studies described in the thesis. We motivate the decisions we took for collecting and analysing the data. For specific information about each of the studies we refer to the articles and the chapters of the thesis.

1.4.1 Data collection methods

Due to the lack of appropriate publicly available data sets for answering our researchers questions, we collected three data sets: the *Students Engagement Using EDA (SEED)* data set, the *USI_Laughs*, and the *WorkplaceDataSet*, a detailed description of the data sets is provided respectively in Chapter 3, Chapter 4 and Chapter 5.

To collect the above mentioned data sets we designed and ran three data collection campaigns. For all three studies we followed the ethics review policy in place at our institutions. The studies were approved by the research ethics delegate of the Faculty of Informatics at USI.

When designing a user study several choices needs to be made, including: the setting of the study; the duration; the population; the type of tools to use; the data collection protocol; and the strategies to adopt to protect the privacy of the study participants and of the collected data.

Figure 1.1. Data collection during lectures: students were asked to wear the Empatica E4 wristband while attending the lectures.

Setting of the study. The first decision to take is whether the study should be carried in laboratory settings or in the *in-the-wild*, i.e., in the natural setting of the subjects. Laboratory settings are suitable for testing novel approaches, for example to understand the feasibility of using new sensors for detecting particular reactions to stimuli. Natural settings add significant challenges in terms of continuous monitoring of data quality and quantity but increases the reliability and ecological validity of the results by taking into account the variability and uncertainty of real settings. In this thesis we conducted studies both in laboratory settings (study presented in Chapter 4 for laughter recognition) and natural settings, i.e., during actual lectures in classroom (study presented in Chapter 3) and during actual work activities in the workplace (study presented in Chapter 5).

Study duration. Given our goal of recognizing engagement during daily activities, we designed the studies to capture multiple occurrences of the activities. Consequently our user studies lasted one or multiple weeks.

Population. We investigated two populations, *students* – Bachelor and Master students – and *knowledge workers* from Academia (or *academics*) – PhD, senior researchers and professors. We selected the first category because fostering students' engagement is considered as one of the most effective countermeasures for resolving serious issues such as high drop out rates, disaffection, students' alienation and low academic performance [57]. We selected academics since they deal with multiple and diverse tasks, they tend to have irregular working hours and periodic stressful periods, thus having systems that can support their work are challenging to design but particularly useful.

Type of tools and data collection protocol. Refers to the choice of data type – in terms of sensor data and ground-truth –, and the data collection strategy. One of our main goals is to monitor participants in real settings when also other people might be present (e.g., in the classrooms or in the office). For this reason we use data collection instruments that are not invasive, do not limit users' movements, and are socially acceptable by the users. Specifically, we collected data using unobtrusive personal devices such as wristband, smartphone, and laptop that people can use anywhere and at anytime and limit the user's discomfort.

In particular, in all the studies we used the *Empatica E4*[2] wristband [58] to collect physiological and movement data. The E4 is a lightweight device equipped with four high-quality sensors which measures: the electrodermal activity (EDA), the blood volume pulse (BVP), the skin temperature (ST) and the 3-axis accelerometer data (ACC) [58]. Figure 1.1 shows an example of the setup we used for collecting students' EDA data during lectures using the E4 wristband. We collected information about laptop usage with *RescueTime*[3], a commercial monitoring tool that people can use for tracking digital activities. We designed and developed a custom Android application called *MEMOTION* for collecting data about phone usage. Participants installed MEMOTION on their smartphone. Data from RescueTime and MEMOTION are used for the study presented in Chapter 5.

To collect ground-truth in the *in-the-wild* studies, we used the *experience sampling method (ESM)* [59] and asked participants to report their engagement by providing answers to validated questionnaires. The ESM has been widely used as instrument to collect ground-truth in the workplace [60; 61; 50] and in the classroom setting [34; 12]. With this strategy we collected *subjective* data about the perceived engagement. We used a retrospective strategy and prompted the questionnaires at the end of participants' activities. In this way we did not interrupt the flow of users' activities and avoided the participants to leave the engagement state due to the interruption caused by the ESM.

For the study on laughter recognition, conducted in laboratory settings, we asked external annotators to report *laughter episodes*, as common practice in the literature [62].

To facilitate the collection of ground-truth data we used paper-based questionnaires and diaries, we designed and developed smartphone applications, laptop widget, a *situated self-report* device [63] called *Devo* and used the *ANVIL* video

[2] https://www.empatica.com/research/e4/

[3] https://www.rescuetime.com/

annotation tool [64]. A comprehensive overview of existing methods for collecting ground-truth data is presented in Section 2.4.

Privacy and storage. In order to protect participants' privacy, we anonymized participants' data during the data collection phase and we stored all the data in the academic cloud storage service SWITCHdrive[4]. We used only third parties services as Empatica and RescueTime which encrypt data in the transmission or not store participants' personal data.

1.4.2 Data analysis methods

To analyse the data we collected we used quantitative methods including statistical analysis, traditional machine learning and deep learning [65; 66].

For recognizing engagement and activities we built upon the typical machine learning pipeline for Affect Recognition [29] and Human Activity Recognition [6] and rely on concepts and techniques presented in previous work [67; 68; 66; 29; 6]. Specifically, to analyse the collected raw data we followed the steps of: data cleaning and pre-processing; segmentation; feature extraction; and the classification pipeline. Each of these steps might have a significant impact on the model performance. We experimented with different techniques and strategies depending on the specific contribution. Figure 1.2 shows an exemplary schematic representation of the steps needed to process the data. We discuss below details about the different steps.

Data cleaning and pre-processing. To minimize the noise and to derive useful information from the data cleaning and pre-processing steps such as missing data handling; sensor data filtering and decomposition; and normalization; are usually suggested.

When collecting data in natural settings, data loss might be experienced. For instance, participants might forget to answer to questionnaires, as a consequence the ground-truth for a specific data instance will be lost. We discarded instances without labels since reconstructing subjective data could be highly inferential. Sensor data might be missing too, due to e.g., sensor malfunctioning, problems with data transmissions or participants forgetting to wear or switch on the devices. Missing data could be either discarded or imputed. To handle missing data, we first investigated the possible reasons behind, indeed missing data could be an information on its own, then we either discarded or imputed data based on the quantity of data missing and possible errors introduced by the reconstruction.

[4] `https://www.switch.ch/drive/`

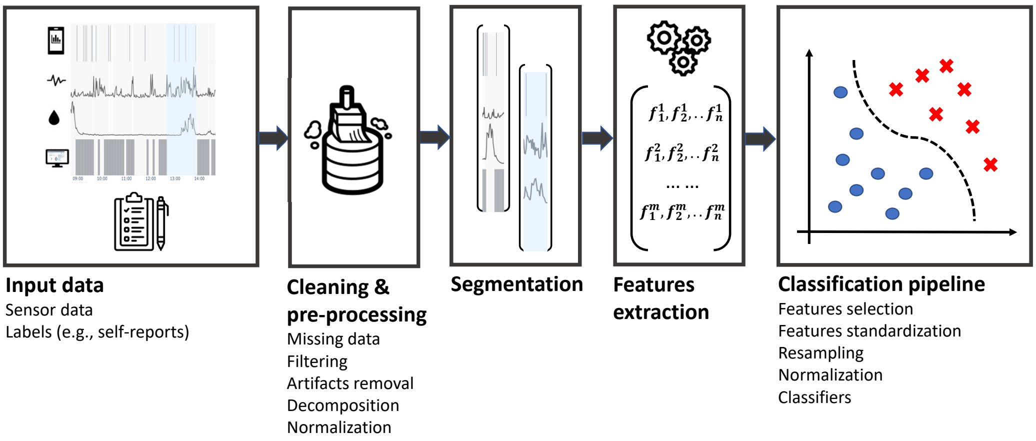


Figure 1.2. Exemplary schematic representation of the data analysis steps we used in this thesis.

The quality of physiological data collected in real-settings might be affected by the presence of noise, and *artifacts* due to e.g., physical movement, thermal regulation or device misplacement. To deal with this problem, we used manual techniques e.g., visual inspection, and automatic approaches, e.g., using filters such as the Butterworth low pass filter [69], and applying artifacts detection models with publicly available tools such as the *EDArtifacts*[5] developed and validated in our research group and presented in [70].

To extract additional information from sensor data, pre-processing steps such as *decomposition* or *aggregation* are often recommended [29]. In this thesis we decomposed the EDA signal in its *phasic* and *tonic* components [35] using tools such as *cvxEDA* available at[6] and presented in [71]. We also derived the ACC magnitude by combining the recordings from the three-axis.

Physiological signals of different individuals, even under the same experimental conditions, might present differences in the recordings (e.g., in the range of values) due to several factors e.g., different skin dryness, color, and thickness [67; 35]. This individuality might represent a problem especially when building models that leverage data of a group of people and apply them on new unseen subjects [67]. To account for differences among individuals and make the signals comparable, techniques as normalization (scaling) are usually applied. To deal with this problem, we used techniques such as min-max scaling or z-transformation [72].

[5] https://github.com/S.gashi/EDArtifact

[6] https://github.com/lciti/cvxEDA

Segmentation. Pre-processing steps are often followed by a segmentation procedure. Windows of fixed or variable sizes can be used to segment the signals. We often segmented the signal traces based on the activities performed by the user (e.g., lectures or workplace activities) and further use a *sliding-window* approach to augment the data set.

Features extraction. From the segmented traces, more often in a traditional machine learning pipeline, informative features are extracted and used as input to the classifiers. In this thesis we proposed features as well as extracted sensor-specific features used in the literature [67; 68; 73; 74; 75; 76; 77; 78; 79]. For extracting the features we either implemented algorithms or used existing tools such as *HeartPy* [80] for heart rate analysis and *EDAExplorer*[7] [68] for electrodermal activity analysis.

Classification pipeline. The engagement recognition and activity recognition problems can be conceptualized as *classification* tasks. Features are paired with the labels (engagement levels or activity type) and used as input to the classifiers. Additional steps such as features selection; features standardization; and resampling are often applied before providing data to the classifiers. The goal of these steps is to improve the model performance and speed up the learning process [72; 66].

The feature selection step allows to limit the *curse of dimensionality* problem [66] as well as to investigate the relevance of the specific features for the classification task. Depending on the studies we used features selection methods such as filter and wrapper methods. Specifically, we used the Sequential Forward Floating Selection (SFFS) algorithm [81] and filter methods based on the Kolmogorov-Smirnov (KS) non-parametric test as done in [82].

Our data sets often present an imbalanced distributions of the samples in the classes. To increase the importance of underrepresented classes and avoid algorithms to be skewed towards the majority classes only, we used resampling algorithms such as the Synthetic Minority Oversampling TEchnique (SMOTE) [83].

The SMOTE algorithm combines the over-sampling and under-sampling techniques, achieving better performance compared to only under-sampling [83].

In this thesis we experimented with several machine learning algorithms. For instance, we applied traditional machine learning pipeline using hand-crafted features as input to classifiers with a shallow structure e.g., Support Vector Machines (SVM) [66], Random Forest (RF) [66], and Gradient Boosting (GB) [66]. We also used end-to-end deep learning pipeline using raw sensor data as input

[7] `https://eda-explorer.media.mit.edu/`

to Convolutional Neural Networks (CNNs) [31].

Evaluation. To evaluate the performance of the model a validation procedure; performance metrics; and baselines to compare the models against, should be used. The goal of machine learning is to learn models from previously collected data that are able to generalize to unseen data [66].

To verify the generalization capabilities of the models, validation techniques are used. The data set is divided in *train* and *test* subsets. The model is built using the train set and then it is applied to the test set which contains the unseen data. Several validation strategies exist, and their choice should be guided by the goal of the analysis and the application scenario. In this thesis we used *user-independent* and *user-dependent* validation strategies [84]. The first is suitable to test the generalizability of the models to unseen users, the second to evaluate the performance of a new behaviour (or activity in our case) of known users [84]. We used user-independent validation strategies such as *leave-one-subject-out (LOSO)* and *leave-one-group-out (LOGO)* [29]. In LOSO, the model is trained with all users' data except one which is used for testing. In LOGO the model is trained with a group of users' data and tested with another group. Both methods test the ability of the models to generalize to unseen user(s). These approaches mimics the application scenario in which new user(s) use the system for the first time.

We also tested the ability of the model to generalize to data from unseen activities of seen users. In this case, data of the same participants might be in the train and test sets but data gathered in the same day is kept either in the train or test set to avoid similarity biases due to adjacent segments [84].

We considered several metrics to evaluate the models such as accuracy [65], balanced accuracy [85], precision [65], recall [65] and F-measures [65] to provide a complete picture of the performance. However, we set a metric of interest depending on the final goal of the system.

We compared the performance of the models with different baselines such as Random Guess (RG) and Biased Random Guess (BRG) classifiers. The RG classifier predicts the outcome uniformly at random. The BRG instead takes in consideration the distribution of the samples in the training set to take a biased decision and it is a suitable baseline for when dealing with imbalance data sets [86]. We also considered baselines such as heuristic-based classifiers and approaches presented in the literature, depending on the goal of the study.

1.5 Contributions

In this section we report a brief summary of the main contributions of this thesis.

Features and behavioural markers of engagement

We answer the first research question by proposing and analysing a set of features and behavioural markers that can be used as proxies for recognizing engagement. In particular, based on findings in educational research, we extracted a set of theoretically-motivated features from electrodermal activity to quantify the *momentary engagement* [87], the *reaction to the teacher* [88], and the *emotional arousal* [89], key components of students' emotional engagement. We tested the effectiveness of the derived markers in the automatic recognition of students' emotional engagement during lectures. We compared the performance obtained when using the proposed features in input to the classifier with the ones obtained when using features proposed in existing literature [90]. To evaluate the performance of the method we ran a data collection campaign *in-the-wild* and collect the *Student Engagement Using EDA (SEED)* data set. The *SEED*, after data cleaning, contains data from 24 students, 9 teachers, during 41 actual lectures. Results from this contribution have already been published as journal paper [A] in the PACM IMWUT (September 2018). Details about the study are presented in Chapter 3.

Further, we proposed a novel approach based on the combination of physiological and movement data, gathered from wristbands, to automatically quantify *laughter*, a key multi-modal expression of engagement. To evaluate our approach we conducted a study in laboratory settings and collected a data set, *USI_Laughs*, which contains data from 34 participants and is available to researchers. This study is described in Chapter 4. Results from this contribution have already been published as conference paper [B] at the PervasiveHealth conference (May 2019).

Fusion of sensor data and context information

We answer the second research question by proposing to combine physiological data (i.e., electrodermal activity and blood volume pulse) gathered from wristband together with context information (i.e., time of the day, day of the week and type of activity) derived from self-reports.

We investigated different fusion strategies based on traditional machine learning and deep learning. We tested the performance of the models using the context information alone and in combination with physiological data. To evaluate the proposed approach we ran a data collection *in-the-wild* with 13 knowledge workers performing their activities during actual work days resulting in the *WorkplaceDataSet*.

For this work we used a subset of the *WorkplaceDataSet* data set. An overview of this study is presented in Chapter 5. Results from this contribution have been accepted for publication and will appear in the proceedings of the ACII 2021 conference [C].

Activity recognition in the workplace

We answer the third research question by proposing a method to combine data derived from personal devices (i.e., laptop, smartphone and wristband) that people use and carry anywhere and anytime. We investigated the impact of combination of data from multiple sources.

We compared the performance of the proposed method with heuristic-based classifiers. To evaluate our approach we used a subset of the *WorkplaceDataSet*. More details about this contribution are presented in Chapter 5. Results from this contribution have already been published as journal paper [D] in the PACM IMWUT (September 2020).

1.6 Publications

The work presented in this thesis has been published or is under submission in peer-reviewed conferences and journals. The scientific articles derived from work described in this thesis are:

A. **E. Di Lascio**, S. Gashi, S. Santini: *Unobtrusive Assessment of Students' Engagement During Lectures Using Electrodermal Activity Sensors*. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 2, Issue 3, September 2018. 21 pages.

B. **E. Di Lascio**, S. Gashi and S. Santini: *Laughter Recognition Using Non-invasive Wearable Devices*. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), May 2019. 10 pages.

C. **E. Di Lascio**, S. Gashi, M. E. Debus and S. Santini. *Automatic Recognition of Flow During Work Activities Using Context and Physiological Signals.* Accepted for publication in: Proceedings of 9th International Conference on Affective Computing & Intelligent Interaction (ACII). 2021. 8 pages.

D. **E. Di Lascio**, S. Gashi, J. S. Hidalgo, B. Nale, M. Debus, S. Santini: *A Multi-sensor Approach to Automatically Recognize Breaks and Work Activities of Academic*

Knowledge Workers. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 3, Issue 3, September 2020. 21 pages.

During my Ph.D. I have also co-authored additional publications. The complete list of publications is reported in Appendix A.

1.7 Thesis outline

This thesis is structured in six chapters.

In Chapter 1 we introduced the motivation behind the work done in this thesis, we defined the research questions, presented the research methods and a brief summary of the main contributions and resulting publications.

In Chapter 2 we provide definitions, theories and concepts which build the foundation of this thesis as well as we discuss existing literature in engagement and activity recognition.

Chapter 3 presents our first user study that investigated the potential of EDA sensors embedded in wearable devices to recognize students engagement during lecture, we discuss our method for translating concepts from educational research into features extracted from EDA signals.

In Chapter 4 we describe our second user study and present our novel method for recognizing the behavioural expression of laughter using a combination of physiological and movement data.

Chapter 5 describes our last user study in monitoring knowledge workers' daily activities. Specifically, we explored the fusion of physiological signals and context information to derive workers' flow during work activities and further proposed to use a combination of personal devices to distinguish between work and break activities.

We summarize the thesis, along with conclusions and outlooks in Chapter 6

Chapter 2

Background and Related Work

In this thesis we rely on several definitions, concepts and techniques. In this chapter we provide an overview of these aspects.

In Section 2.1 we cover the definitions and theories of *engagement* that place the theoretical foundation of this thesis. We particularly focus on students and workers' engagement from a psychological and physiological perspective.

We describe existing methods for measuring engagement in Section 2.2. Section 2.3 provides an overview of investigated behavioural cues of engagement and the sensors used to describe them. We review existing methods for labeling and modeling engagement respectively in Section 2.4 and Section 2.5.

We cover definitions of *context* and *activity* in Section 2.6. We conclude the chapter with Section 2.7, in which we review existing literature in activity recognition in the workplace.

2.1 Engagement definitions and theories

In this section we introduce definitions of engagement that provide the theoretical foundation of this thesis. Specifically, Section 2.1.1 describes students' engagement theories derived from educational research and Section 2.1.2 provides an overview of emotional engagement from a psychophysiology perspective. In Section 2.1.3 we summarize theories of work engagement and, in Section 2.1.4 we provide a description of the psychophysiology of flow.

Engagement is related to a person's level of involvement and absorption into an activity [91] and it represents a fundamental component of the persons' experience with that activity [24]. Even though several context-specific definitions of engagement exist, researchers agree that when people are engaged in an activity, they are more likely to produce better outcome, enjoy and learn more [18]. Given the importance of engagement, several theories in psychology have emerged for defining and quantifying engagement [92; 57].

In the years these theories have been exploited in Computer Science (CS) research, specifically in Human-Computer Interaction (HCI) and Affecting Computing (AC) fields for addressing the problem of measuring *engagement* using sensing technology.

In this thesis we differentiate between studies aiming to quantify *user*'s engagement *with* technology with the goal of improving the user experience with technology, from studies concerning measuring *individual*'s engagement during their activities *using* technology with the goal of supporting people in their daily activities. Even though the difference seems subtle, we believe that it is important to consider that the final goal of the engagement-aware system in the two cases is different.

The goal of the first type of studies is to determine which are the characteristics of a technology artifact (e.g., user interface, web page, smartphone application) that increase the likelihood of user's engagement and design technology in a way that engage people [93]. In this direction, O'Brien rooted user's engagement *with* technology in Play, Flow and Aesthetic theories [94]. From the analysis of these theories authors of [94] defined user's engagement as the "quality of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control" [94].

The goal of studies in the second group is to investigate which are the characteristics of individuals when they are engaged in their activities, quantify them using sensing technology and build systems that can help people discover what conditions engage them (e.g., using personal analytics), and how these conditions can be promoted (e.g., using personal feedback).

Work done in this thesis belongs to the second category. In particular, we focus on quantifying the state of engagement as experienced by an individual (i.e., a student or worker) during an activity, and how it can be assessed using sensing technology. In this thesis we focus on definitions and theories of engagement that have been conceptualized for workers and students.

2.1.1 Students engagement definitions

There is a long history of research in educational literature aiming to define, quantify and measure students' engagement [23]. Researchers and educators have focused on students' engagement as key for addressing serious social issues such as high drop out rates, low academic achievement, students' alienation and boredom [95]. Students' engagement is a malleable state and can be influenced and shaped by teachers, peers and even technology [18; 96].

Students' engagement is linked to *motivation*. Motivation refers to reasons behind a given behaviour while engagement is considered in terms of action or manifestation of motivation [95]. In other words, in order to move from motivation to action, the individual (in this case the student) should be engaged in the activity. Furthermore, differently from motivation, engagement "reflects an individual's interaction with the context", so the engagement of the individual is directed to something in particular (e.g., a task, an activity) and can not be separated from it.

Scholars have focused on different populations of students (e.g., elementary, middle school, high school, university) used several terms (e.g., academic engagement, school engagement, student engagement in class), definitions, components (e.g., emotional, behavioural, academic, psychological) for students' engagement, and measured engagement at different levels (e.g., with prosocial institutions, with school, in the classroom, with learning activities) and granularity (e.g., individual students in the learning activity, group of students in a class). For a comprehensive review of students' engagement we refer to [23; 97].

In this thesis we focus on *university students*, we consider the well accepted and widely used definition of students engagement as "a meta-construct that includes behavioural, emotional and cognitive engagement" proposed by Fredricks et al., [57], that we describe below. We investigate the *emotional* component of engagement of individual students in the learning activity of attending the lecture in classroom.

Behavioural, cognitive and emotional engagement.

According to the definition of Fredricks et al. [57], students' engagement can be divided in three components: *behavioural* engagement (e.g., time on task), *emotional* engagement (e.g., interest and value) and *cognitive* engagement (e.g.., self-regulation and learning strategies) [95].

Behavioural engagement is based on the idea of *participation* in social, in extracurricular or in academic activities. It is generally assessed based on a set of observable behaviours as class attendance, time on homework, effort (e.g., time on task) and adherence to classroom rules. Behavioural engagement can be also considered in terms of *academic* engagement, which refers to the presence or absence of a set of students behaviours in classroom such as participating in classroom tasks, asking questions, writing, or reading aloud [95].

Cognitive engagement is broadly defined as “students’ level of investment in learning” and incorporates elements such as motivation, effort, learning strategy and self-regulated learning. The term cognitive engagement is used for describing aspects as the cognitive strategies used by the student (i.e., whether the student used sophisticated or superficial learning strategies), self-regulatory or meta-cognitive strategies (i.e., how students manage the learning process in terms of acquiring and planning information). In the classroom, cognitive engagement has been also expressed as the amount of effort put by the student in understanding a topic [98].

Emotional engagement considers the emotional involvement of the student in the process of learning. It focuses on the presence of positive and supportive emotions (e.g., interest) and absence of negative or withdrawal emotions (e.g., boredom, anxiety)[95; 57].

Several theories exist for defining emotions and affective states [29]. In this thesis we consider the *circumplex model of affect* proposed by Russel at the end of 1970s [99] and widely adopted in Affective Computing research [29].

The circumplex model of affect is a two-dimensional model in which affective states are represented as discrete points described in terms of the two axis of *valence* and *arousal* [99]. The valence axis defines the polarity of the affective states, e.g., how positive or negative it is. The arousal component refers to the intensity or activation, and it describes how energized or deactivated an individual feels [99]. Students’ emotional arousal is a relevant component to consider in learning since it can influence performance and memory [100; 89].

An important component of emotional engagement in classroom is *interest*. The complex relation between interest and engagement can be represented using two metaphors, a “hook” and a “switch" as reported in [87].

The “hook” refers to the simplest interest, when an activity or a features of an activity in class, elicits interest, the student “readily engages with that activity” this case is also referred as *situational engagement* or *momentary engagement*. In addition to situational engagement there is also students’ personal interests.

The “switch” metaphor in this case connects the personal interests with the opportunity to express them in the class activity, this takes into account not only the momentary engagement but also the past students’ experience. Whether interest derives from specific activities or personal interests, it is fundamental for ensuring engagement in classroom [87].

Students’ emotional engagement is also linked to positive and negative reactions to the teacher [95; 57]. Studies from educational research show that as the teacher’s emotional response in classroom increases, so does their students' emotional response [88; 101]. Researchers, have studied the interaction between students and teachers in classroom using the theory of *emotional contagion* using traditional methods (i.e., surveys, rating-scales) [88; 101].

The theory of emotional contagion suggests that people communicating to each other, "automatically mimic and synchronize expressions, vocalizations, postures and movements and consequently converge emotionally as a result of the activation and/or feedback from such mimicry" [102]. Moreover it suggests that all participants in an interaction are susceptible to each other's emotions and the contagion effect [102].

In this thesis we focus on students' emotional engagement and consider the components of emotional arousal, interest, interaction with the teacher as basis for extracting and proposing a set of features from electrodermal activity that could be used as proxies for describing these components. More details about the procedure we followed are explained in Chapter 3. In the following section we describe how emotional engagement have been considered from a psychophysiology perspective.

2.1.2 Psychophysiology of emotional engagement

The relation between human affective states and the physiological reactions of the body has been central focus of psychophysiology[1] research for over a century [104]. Several theories have been proposed in years connecting the emotional and the physiological reactions, however the direction/causality that links the two is still an open question [104; 29].

For instance, according to the *James-Lange* theory [105], proposed by the psychologists William James and the physiologist Carl Lange in the 19th century, emotions occur as physiological reactions to events. The theory suggests that the physiological reaction is a consequence of an external stimulus, while the emotional reaction depends on how the individual interpret the physical reactions.

In contrast to the *James-Lange* theory, the *Cannon-Bard* theory [106], proposed by Walter Cannon in 1920 and expanded by physiologist Philip Bard during the 1930s, the emotions and the physiological reactions are experienced *simultaneously* [106].

For an overview of existing theories of the psychophysiology of emotions we refer to [104].

[1] Psychophysiology is a particular branch of psychology that regards the study of the bodily functions and physiological bases of "psychological phenomena –– the experience and behavior of organisms in the physical and social environment" [103]

Besides the different theories, the physiological patterns connected to emotional responses are often linked to the the activity of the *Autonomic Nervous System (ANS)*. The *ANS*, which is a part of the *Peripheral Nervous System (PNS)*, is responsible for the regulation of involuntary body function as blood flow, heartbeat, digestion, and breathing.

The *ANS* "plays a key role in directing physiological responses to external (e.g., events) or internal stimuli (e.g., thoughts)" [29] and its activation is thus suitable for understanding affective states [29].

The ANS is composed by two complementary branches the *Sympathetic Nervous System (SNS)* and the *Parasympathetic Nervous System (PNS)*. The activity of each of these systems is prevalent under specific conditions. The activation of the *SNS*, also known as *physiological arousal*, is dominant during "fight-or-flight" reactions and exercises.

An increment of the activation of the SNS corresponds to high arousal states. The main function of the SNS is to prepare the body and provide energy for responding to a stimulus, it does so by increasing a set of physiological parameters (e.g., accelerates the heart rate, increase the breathing rate, activate the sweat secretion).

In contrast, the activation of PNS is dominant during quiet and resting conditions, regulating the "rest and digest" functions. The goal of the PNS is to conserve energy [107] by for instance slowing the heart rate and the breathing rate.

Typical measures of the activity of the ANS, as studied in psychophysiology literature, are the *electrodermal activity* – regulated by the SNS only, it thus represents a direct measure of the physiological arousal – and *cardiac activity* – regulated by both branches.

Based on findings from psychophysiology literature, researchers have investigated the emotional component of the students' engagement in terms of the students' body reactions [108; 33; 24; 34].

In this thesis we focus on the use of electrodermal activity as a proxy for students' emotional engagement.

Considering the lecture as the external stimulus to which students are exposed, physiological parameters derived from the electrodermal activity can be used to determine the physiological and thus the emotional arousal component of students' emotional engagement [33]. For instance McNeal et al. [33] monitored the engagement of 17 students during a course on environmental geology. Authors tested different pedagogical approaches (frontal lecture, discussion, movie watching) on subsets of the students and reported statistical differences in terms of mean of the EDA traces, that they consider as proxy of engagement. Authors

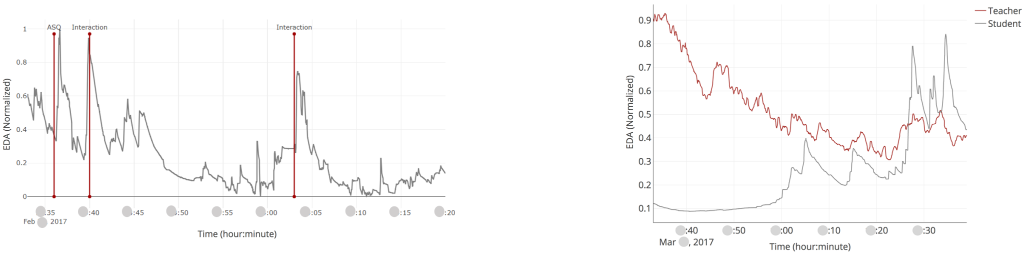


Figure 2.2. Example of teacher's (red or dark color) and student's (grey or light color) EDA signal alignment over time.

noticed higher engagement, during discussion and movie watching compared to lectures [33].

In a recent work Cain et al. [90] proposed to use EDA as proxy for momentary engagement. Specifically, the authors proposed that evident increments of physiological arousal could be used as indicators of momentary engagement[90]. The authors used EDA peaks to identify children's momentary engagement during learning activities [90].

The emotional contagion phenomenon, described before, has been also investigated from a physiological perspective [109]. In particular the *physiological synchrony* has been underlined as one of the possible mechanism under emotional contagion [109].

The *physiological synchrony* refers to the association between the physiological activity of two or more people [110]. Previous works have successfully connected the physiological synchrony derived from the EDA to individual's engagement, during conversations [111] or in children-adult interactions [67]. For a complete review on physiological synchrony definitions and metrics we refer to [110].

In this thesis we build upon previous work and derive a set of features that could represent the emotional arousal of the student, the momentary engagement and the reaction to the teacher.

An anecdotal example of the connection between EDA and momentary engagement, as we considered in this thesis, is presented in Figure 2.1. It is possible to observe that the EDA of a student during a lecture presents high peaks when the teacher interacts directly with the student, or when an in-class exercise is prompted, these "arousing moments" could be interpret as proxies for the *momentary engagement*.

Figure 2.2 shows an anecdotal example of physiological synchrony between a student and a teacher. It is possible to observe variations of the alignment in the responses (no alignment at the beginning of the lecture, more alignment towards the end). More details about the method we used for quantifying students' engagement during lectures is presented in Chapter 3.

2.1.3 Knowledge workers engagement definitions

Knowledge workers – often referred to also as *information* workers – typically perform "*non-routine, cognitive, or creative work*" [112] and they have "*significant responsibility for structuring and managing [their] work*" [113].

Examples of knowledge workers are programmers, engineers, scientists, design thinkers and academics. In the study presented in Chapter 5 we investigate a particular category of knowledge workers: the *academics*.

Researchers agree in considering engagement at work as a "positive, fulfilling, work-related state of mind" [114].

According to Sonnentag et al. [114], work engagement is not only referred as broad concept that compromise all the work activities and the *whole work* in general. Rather, work engagement "emerges in the process of working, that is, when dealing with a specific tasks" [114].

Engagement at work can fluctuate from day to day but also from task to task [114], however this fluctuations are not random and events or experiences are predictors of engagement. For instance, the type of task the worker is dealing with can have a substantial contribution to the experience of work engagement [114].

Several theories exist in the definition of engagement at work [115; 116; 117; 26]. Among them, prevalent theories have either consider work engagement in relation with burnout (either as antipodes or as independent) [116], or in terms of the optimal experience of *flow* corresponding to high levels of engagement [32; 26].

Studies in the first category, conceptualizing engagement as a persistent state rather than a momentary and specific one, as opposite to the *flow* theory which considers engagement as a peak experience emerging while performing an activity.

A summary of the two theories is described below. However, in this thesis we rely on the flow theory which has been also studied from a psychophysiology perspective [1; 118] supporting the possibility of using physiological parameters such as electrodermal activity and cardiac activity for measuring engagement, which is the aim of this thesis.

Work engagement and burnout

Initial studies defined work engagement as opposite to *burnout* [115]. Burnout defined as "prolonged response to chronic emotional and interpersonal stressors on the job" is described by three dimensions: *exhaustion*, *cynicism* and *inefficacy* [115; 116].

Exhaustion, or low energy, is one of the most reported aspects of burnout and refers to psychological fatigue experienced by overtired employee.

Cynicism, or low identification, corresponds to a pessimistic or indifferent attitude towards work. Lastly, inefficacy is defined in terms of reduced personal accomplishment.

In contrast to burnout, Maslach et al. [116] defined work engagement as "a persistent, positive affective-motivational state of fulfillment in employees" and described by three dimensions: *vigor*, *dedication*, and *absorption* [116].

Vigor, considered the opposite of exhaustion, is characterized by high level of energy and willingness in investing effort in the work. Dedication, in contrast to cynicism, refers to the enthusiasm of workers about their job that let individuals experience a sense of significance. Absorption, is characterized by total immersion in one's work to the point that while working, employees do not notice time passing.

Schaufeli et al. [117; 119] after considering the results of a factor analysis to verify this theory, decided to move away from the definition of engagement as antipodes of burnout. Indeed authors observed that while the first two dimensions of engagement (i.e., vigor and dedication) resulted to be the opposite of the first two dimension of burnout (i.e., exhaustion and cynicism), this was not verified for the third component (i.e., absorption against inefficacy).

Thus authors [117] proposed to maintain the three components of work engagement but consider it independently from burnout and developed the self-report measure *Utrecht Worker Engagement Scale (UWES)* [120].

Other studies, suggest that the core components of work engagement were actually vigor and dedication, while absorption seems more related to the concept of flow [119].

Flow at work

The concept of flow was firstly introduced by the psychologist Mihalyi Csikszentmihalyi in the 1970s [32], as an *optimal* experience characterized by "intense focus and concentration" [32].

When individuals are in this state, they are totally involved in the activity and nothing else seems to matter. Flow occurrence and characterization have been studied in several domains, such as sports, gaming, and leisure activities [121; 92].

Based on findings of the previous studies, Bakker [26] applied the concept of *flow* to the work context. Bakker defined flow at work as "short-term peak experience at work that is characterized by absorption, work enjoyment and intrinsic work motivation" [26].

Absorption refers to the sense of total involvement in the activity. Intrinsic motivation refers to the need of performing the work activity because of the pleasure and satisfaction it elicits. Enjoyment, refers to the perceived enjoyment during the pleasurable work activity [26].

A central element of the flow theory regards the balance between the task difficulty and the ability of the individual to address it. The optimal state of flow during activities is likely to occur when the challenges induced by the task at hand are in optimal balance with the individual's skills, opposite to when activities are too easy (leading to boredom) or too difficult (leading to anxiety or stress) [122; 121]. Furthermore it has been observed that the experience of flow is more likely to occur when both, challenge and skills, are above the individual's average experienced daily activities [121].

Interestingly, researchers noted that flow is more likely to occur in work activities rather than in leisure activities [123], maybe due to difficulty for individuals to create challenging situations that require matching skills in the free time [123]. For a comprehensive description of the advances in flow research with refer to [92].

In the following section we describe how the state of flow has been characterized from a physiological point of view.

2.1.4 Psychophysiology of flow

Initial studies measuring flow used retrospective methods such as self-reports. More recently, researchers in psychophysiology started investigating potential physiological indicators of flow [1; 118]. A comprehensive overview of existing literature in psychophysioloy of flow is presented in [1].

The exact connection between physiological parameters and flow during daily activities is still debated. However, based on the challenge-skill balance prerequisite for the flow experience, from a physiological perspective, flow has been

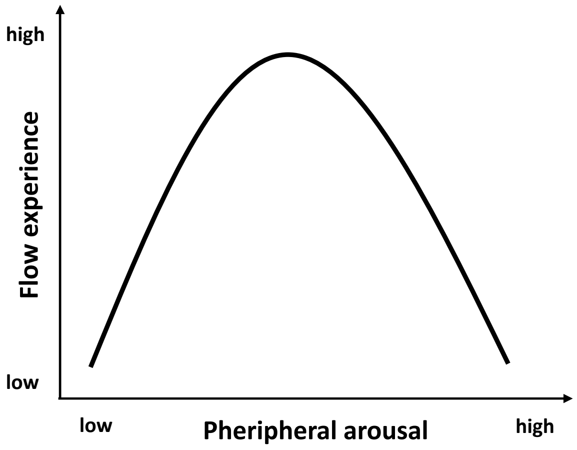


Figure 2.3. Diagram describing the inverted-U connection between the flow experience and the peripheral arousal as proposed in [1].

often linked to a moderate level of arousal, neither too high (as stress or anxiety) or too low (as boredom or relaxation) [118].

Based on existing literature, Peifer describes the experience of flow as characterized by "optimized physiological activation" [1]. Despite the different physiological demands that different activities might request, Peifer identifies a common "ground" for the phsyiological activation happening during flow: "the full concentration of all body functions to the given activity and the down regulation of all functions that are irrelevant for task fulfillment" [1].

According to Peifer, the "optimized physiological activation" [92] corresponds to a "moderate peripheral arousal following a U-shaped function of activation" as schematically represented in Figure 2.3.

Electrodermal activity and cardiovascular activity are in general the most used physiological parameters originated from the peripheral nervous system considered in psychophysiology studies of flow [124; 118; 125].

Very low or very high cardiovascular activation corresponds respectively to relaxation or stress states. Thus, according to the model proposed by Peifer, the optimal state of flow should generate cardiovascular reactions that lie between the two [1]. On the same line, the mental effort generated during flow should be lower to the one during stressful conditions, thanks to ability of the individual to deal with the challenges introduce by the task. This corresponds to lower physiological arousal (and consequently reduced EDA responses) during flow compared to stressful situations, however the challenges introduced by the task will nonetheless activate the sympathetic nervous system and thus generating higher EDA responses compared to relaxation [1].

While the model proposed by Peifer [1] is grounded on existing literature, the exact relation between physiological parameters and flow is still not certain. For instance, while studies have confirmed the Peifer's observations [125], others have instead found a linear relation between flow and peripheral arousal [126].

Motivated by previous research, in this thesis we investigate the role of physiological parameters as electrodermal activity and cardiovascular activity, to automatically recognize flow at work. However, we take a data-driven "bottom-up" approach based on machine learning that could allow discovering new unknowns and complex relations between physiology and flow as also discussed in [47].

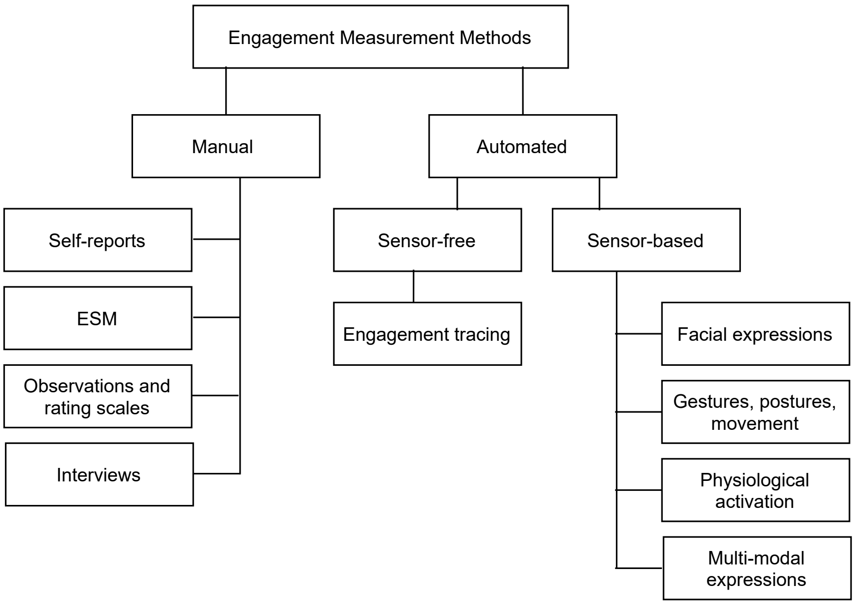


Figure 2.4. Taxonomy of engagement measurements methods.

2.2 Engagement measurements methods

Measuring engagement refers to the process needed for deriving information about the individual's engagement during an activity.

We divide the methods for measuring engagement in two main categories: *manual* and *automated*. Methods belonging to the first category, often referred to as as *traditional* methods, require manual effort from participants and/or researchers. Manual methods are principally used in organizational psychology and educational research [95; 92].

Methods in the second category refer to strategies mostly adopted in HCI research, that could be integrated in computer systems [127; 24; 12; 28] and require little or no input from the individual.

Given the focus of this thesis on students and knowledge workers, we mostly report methods for measuring the engagement of these populations.

The main manual methods for measuring engagement are: self-reports; experience sampling; observations and rating scales; and interviews. We divide the automated measurements methods in: sensor-free, and sensor-based methods. For conducting the studies presented in this thesis, we adopted automated, sensor-based methods. Figure 2.4 shows a schematic representation of the categories of existing measurements methods for engagement as considered in this thesis.

2.2.1 Self-reports

Self-reports survey measures are one of the most used traditional methods for obtaining subjective information.

This method has been largely used to measure students and workers' engagement [95; 92; 128]. It has been widely adopted to assess students' engagement in classroom because it is easy and practical to administer in a classroom setting at relatively low cost [95]. Students are generally asked to select the response to items that reflects different aspects of engagement.

Self-reports enable the collection of students' subjective impressions and perceptions, differently from objective information derived from behavioral indicators (e.g., lecture attendance, homework completion) [95]. In the educational context self-reports are particularly indicated for assessing emotional and cognitive engagement since these aspects are not directly observable, and using other methods that require an external observer (e.g., a teacher) are considered highly inferential [129].

One limitation of self-reports is that students might not be honest in their answers, especially under specific circumstance (i.e., if the teacher administer them and in case the anonymity of the answers is not granted) [95; 127] hampering the possibility of getting valid measurements [95; 129]. Further, self-reports often contains items that do not reflect the particular context (e.g., the lecture) but rather describe broader circumstances (e.g., "I work hard at school" [57]), so their usage is limited when the goal is to measure engagement in relation of contextual factors. Lastly, completing a self-report is often time consuming and participants might not have time or willingness to answer to many questions.

2.2.2 Experience sampling

The *Experience Sampling Method (ESM)*, often also referred as *Ecological Momentary Assessment (EMA)*, is a widely used technique to measure human behaviour in real-settings [130].

In ESM studies, participants are requested to provide answers to self-reports multiple times per day reporting about their activities, emotions or other aspects of their daily life [130].

The self-reports are usually provided in terms of answers to short, identical questionnaires that are prompted on a device accompanied by a notification. Initial ESM studies were conducted using electronic pagers or alarm watches. For example, participants of the study presented in [131], carried an electronic pager together with a paper questionnaire, and completed a paper form when receiving the pager signal [131].

Recently, advances in mobile technology and their widespread in daily life have enabled new possibilities for ESM studies. Personal devices as smartphones and smartwatches are now often used as promising and sophisticated platforms for delivering self-reports [59]. Utilizing the sensors embedded in these devices is also possible to adjust the content and the sampling strategy based on the sensed participant's availability and context [59].

The ESM strategy is widely adopted in *flow* research [92]. Csikszentmihalyi [32] was one of the first recognizing the value of this method for measuring flow during daily activities.

The ESM has been identified also as a valuable strategy to measure students' engagement in the classroom [95] and workers' flow during work activities [132; 128]. The ESM allows researchers to collect measures of engagement in the moment, reducing the recall failure typical of self-reports which are retrospective. Further, ESM studies enable researcher to investigate patterns and variations of engagement across time and situations [95]. However, the ESM strategy requires time and effort from participants, thus the success and quality of the method highly depends on the compliance of the respondents.

Furthermore, it is important to consider a proper design of the ESM studies when measuring engagement and flow, indeed with the ESM there is the risk of interrupting the user during the activity causing the individual to leave the flow state [1]. A possible solution to this problem consists in assessing flow right after the activity [1].

2.2.3 Observations and rating scales

Observational methods require external observers (either researchers or the teachers in case of students' monitoring) to "judge" how engaged individuals *appear*. Observers are generally asked to complete checklists or rating scales to report the presence or absence of specific behavioral indicators of engagement.

This method have been used to measure students' engagement either at the classroom or individual level [95].

Typical academic behaviors indicative of engagement during lectures are writing, answering questions, participating in classroom tasks [95].

One of the main advantages of the observational method is that it can provide a detailed description of the contextual factors which occur when students' are or not engaged. However, this method is time consuming, and the reliability of the measures depend on the proper training of the observer [95].

Furthermore, even if this method can provide useful information about the behavioral indicators of engagement, it is limited for the assessment of students' cognitive and emotional involvement. The last two indicators are not directly observable and students might have learned how to mask their emotions [95].

For instance, researchers, in initial studies on students' engagement [133], observed that many students who appeared "off-task", in subsequent interviews reported to be highly engaged, while students assessed as engaged where actually not thinking about the material presented during the lecture [133].

2.2.4 Interviews

The interviews are a less common method to measure students and workers' engagement [95].

Interviews are typically structured or semi-structured with questions that are pre-designed and participants are requested to provide answers in an open-ended and not structured way [95]. This method allows to collect detailed insights and information about the reasons behind possible variability of the level of engagement as perceived by the participants. However, the good quality of the data derived from the interviews, is impacted by the skills, bias and knowledge of the interviewer. Further, the tendency of social desirability of the participants could hamper the reliability of the answers.

All the traditional methods for measuring engagement have advantages and disadvantages and their applicability depends on the particular context.

In general, traditional methods are time-consuming, involve large manual effort and the quality of the data often depends on the researchers and the participants. Further, a key characteristic of engagement is that it is experienced during the activity. Thus, developing methods for continuously and automatically measuring engagement could enable a better understanding of the engagement state as well as create systems and interfaces that can optimally respond to the user's engagement.

Recent advances in mobile and wearable technologies have enabled initial progresses in the automatic measurement of engagement. In the followings we describe the main automated methods for measuring engagement.

2.2.5 Sensor-free methods

The *sensor-free* method does not require the use of sensors for measuring engagement.

This method consists in monitoring individual's interaction with the computer system to derive actions that are stored in log files. Information from log-files are processed to derive information such as *reaction time*, performance and errors [134] which are then used to infer engagement. This method was mainly used in initial studies on computer-based learning and *intelligent tutoring systems (ITS)* [135]. For example, the *engagement tracing* is a sensor-free method defined in [135] are widely used in ITS [127; 134].

The *engagement tracing* is often considered as a *semi-automated* method given the indirect involvement of the student and the teacher in the process [134]. The method for estimating learner's engagement consists in evaluating the accuracy and timing of learners' responses to problems and practice questions [135]. For instance, very short response times or random answers to easy questions can be used as indicators of student's disangagement using probabilistic inference [135].

Although this method is widely adopted, it requires the students to use instruments (e.g., the laptop) that enable the system to automatically calculate the time and accuracy of the responses. Sensor-free methods are more suitable for computer-based learning. Indeed, students attending lectures in classrooms might not always use their computers making the log file information not always available. Further, especially the engagement tracing method, requires the instructor to provide questions and exercises during the lecture which might not always be the case during lectures.

2.2.6 Sensor-based methods

Sensor-based automated measurements methods consist in using data derived from sensors (e.g., cameras, inertial sensors, physiological sensors) to measure different modalities (e.g., face, body) to quantify behavioural expressions or cues of engagement (e.g., facial expressions, body movement, physiological activation). Features extracted from sensors are used to describe the behavioural expressions of engagement. Features together with engament labels are used as input the machine learning algorithms that allows to automatically determine, the level of engagement.

In the following sections we provide a brief overview of the main behavioural cues, sensors, labeling procedures and models adopted in the literature for the automatic recognition of engagement.

2.3 Behavioral cues of engagement and sensors

Most of the existing approaches in estimating engagement [136; 137; 138; 67] and particularly students [127; 12; 24; 34; 30] and workers' engagement [28; 45; 47; 41] rely on nonverbal behavioural cues of engagement such as: facial expressions [139; 24]; gestures, postures and movement [140; 141]; physiological activation [142; 47]; and multimodal expressions [143]. These expression are either considered individually [127; 47] or combined [24; 28; 41].

In this thesis we mainly rely on physiological activation cue and on a particular multi-modal expression i.e., laughter.

Engagement theorists have identified two types of data that can be used to measure engagement: *internal* to the individual and *external* observable factors [25; 24]. The first is more suitable when targeting affective and cognitive components of engagement, the second for the behavioral one. Among the behavioural cues that can be quantified using sensors, the physiological activation represents a good proxy for the internal factor, while the others are key components of the external factor. In the followings we summarize the main behavioural cues used in the literature and how they have been quantified using sensors.

2.3.1 Facial expressions

Researchers have hypothesized that facial expressions represent one of the main cues for humans to judge the perceived engagement of others [127; 134]. Based on this assumption several approaches for recognizing students' engagement rely on this modality [134; 127; 24]. For instance, McDuff et al. consider, as one of the possible cues, facial expressions for workers' engagement recognition [28].

Cameras embedded in personal computers, mobile phones or positioned in the classroom or in the office, can be used to record face images. These images can be then processed with Computer Vision techniques to automatically derive facial information that are then translated into expressions of engagement [134]. A comprehensive review of Computer Vision techniques for students' engagement recognition in online learning is presented in [134].

Two main methods exist for the analysis of facial expressions: *part-based* and *appearance-based* [134].

The part-based method refers to the range of techniques that focus on parts of a face (e.g., eyes, mouth, nose, forehead, chin and so on) [134]. A widely used method for describing the parts of the face is the *Facial Action Coding System (FACS)* proposed by Ekman et al. [144; 145]. The FACS framework describes facial expressions in terms of *Action Units (AUs)* which measure the action of individual or groups of facial muscles [144; 145].

While measuring the *AUs* represents a descriptive analysis of behaviour, deriving facial expressions (e.g., frustration, happiness) is an inferential process [144; 145]. For instance, happiness is often described as a combination of *AU12* and *AU6*.

Over the years, psychologists and neuroscientists have widely used the FACS for facial expression analysis. Currently, several computer vision-based systems, such as the *Computer Expression Recognition Toolbox (CERT)* [146], can recognize accurately several *AUs* and provide their intensity.

For instance, Grafsgaard et al. [139] used the *CERT* system to track facial expressions during computer-based tutoring. Authors observed that upper face movements, including eyebrow raising (inner and outer), brow lowering, eyelid tightening, and mouth dimpling, were predictors of engagement, frustration and learning.

The CERT system was used also in [127] for the automatic assessment of students' engagement in online learning as perceived by observers. Authors found that among the most discrimitative features, the *AU10* (upper lip raiser), was positively correlated to perceived engagement while the *AU1* (inner brow raiser) and the *AU45* (eye closure) – indicative of the student has tuned out or looking down away – were negatively correlated with engagement.

Techniques in the appearance-based category, extract features from the whole face region [134] and analyze the changes in the face's surface in static and dynamic space [24]. Monkaresi et al. used a appearance-based method called *Local Binary Pattern (LBP)* for texture description, part-based features and heart rate measurements to determine when students were or not engaged in online learning activities [24]. Authors obtained the best engagement recognition performance when fusing the channels [24].

Alternative ways for measuring facial expressions could involve the use of Electromyographic (EMG) signals which capture the electrical activity of muscles through electrodes placed on the face [147], however this method is often considered invasive. Alternatively, recently there is a growing interest in using inertial sensors embedded in earables such as the *eSense*[2], to measure facial muscle movements and consequently facial expressions [148].

2.3.2 Gestures, postures and movement

Researchers postulate that the cognitive processes are constrained by the environment and by the coupling of action and perception. Thus, cognitive and affective states are expected to be manifested in the body language [149].

Observable expressions such as head or hand gestures, body postures and movement, provide important hints of non-verbal communication and mental states and can be used for engagement detection [134; 150; 28; 151].

Depending of the type of gestures and parts of the body involved, gestures and postures can be quantified using sensors as cameras [134], depth cameras as Microsoft Kinect [28], inertial sensors – accelerometer (ACC), gyroscope (GYRO) – embedded in wrist-worn wearable devices or earables, or using textile pressure sensors embedded in the chair to measure body weight and weight changes [150; 149].

Body postures in case of students and workers have been mostly monitored when subjects were in a seated position [28; 140], due to these individuals spending most of the time of their activities seated on a chair. When highly engaged in a task, a worker or student could be expected to have a upright posture, in contrast to a slouched posture typical of a disengaged person [28].

In this direction, D'Mello et al. equipped a chair with a *Body Pressure Measurement System (BPMS)* and investigated how body postures impacted the engage-

[2] `https://www.esense.io/`

ment recognition while participants interacted with the *AutoTutor* system [140]. Authors found that the flow state was associated with minimal movement on the chair and heightened pressure on the seat. This indicates that when learners are mentally focused on a task, they do not spend large amount of cognitive processes in trivial body motions. Further, when in *flow*, individuals tend to get closer to the source of stimulation (i.e., shorter distance between the nose and the screen). Authors also found that boredom was manifested by a higher pressure on the back of the chair, suggesting learners leaning back and detaching from the learning environment. Rapid changes in pressure were also found during boredom, indicating learners fidget when they are mentally disengaged from the tutor [140]. According to [140], an advantage in monitoring body postures for determining engagement is that, in comparison with facial expressions and gestures, these cues are unconscious and unintentional thus less affected by social editing.

Head and hand gestures are also relevant proxies for recognizing engagement or disangagement [134]. For instance, hand over-face gestures have been observed to be highly prevalent in online learning settings and they seem to represent a relevant proxy for understanding learner's affective states such as focus or boredom [152; 141]. For instance, Grafsgaard et al. [152] observed one-hand-to-face and two-hands-to-face gestures were respectively associated with reduced frustration or focus.

Methods for recognizing gestures indicative of active participation, and consequently engagement, such as hand raised, have been also investigated in [153; 154].

Researchers have also explored the role of movements (e.g., movement of the head [151], eye [134], and body [155]) in the detection of workers [155] and students' [151] engagement. For instance Ara et al. [155] used continuous sensing of motion rhythms derived from a 3-axis acceleromenter embedded in work badges, to quantify flow experience at work. Authors found that the motion rhythm around 2-3Hz was moderately correlated with flow experienced at work [155].

2.3.3 Physiological activation

Researchers have frequently connected engagement to an increased level of arousal or alertness and used central and peripheral physiological responses to measure it [24; 67; 156; 137]. Physiological responses are very hard if not impossible to control by humans and provide an *internal* indicator of engagement, thus they are suitable for measuring emotional and cognitive components of engagement.

Existing approaches relying on physiological parameters to recognize engagement often use wearable physiological sensors to quantify the brain [157; 158], cardiac [24; 12] and electrodermal activity [159; 12].

Brain activity

Measurements of the *brain activity* obtained using *electroencephalogram (EEG)* have been used for measuring cognitive engagement [137; 156; 157]. Using electrodes placed on the scalp, the EEG measures the ionic current of neurons of the brain. An "engagement index" derived from EEG has been proposed in the literature [160].

The engagement index is a combination of neural oscillations at different frequencies that reflects "visual processing and sustained attention" [157]. For more details about the engagement index, we refer to [160]. The index has been used to measure audience [158] and learners' engagement [156] and to log engagement during work tasks [157]. Common devices used for measuring the brain activity are brain headsets such as the *Emotiv Insights*[3].

Cardiac activity

The activity of the cardiovascular system is regulated by interplay of the two branches of the Autonomic Nervous System (ANS), i.e., the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS) [103; 161].

The *cardiac activity* can be quantified using parameters such as the *heart rate (HR)* and the *heart rate variability (HRV)* i.e., the variation between beat-to-beat intervals.

For instance, during stressful or arousing conditions the SNS increases the HR that is then brought back to normal by the PNS in the resting condition. The HRV measurements are often used to quantify the interaction between the two branches of the ANS, for instance, an increased/decreased HRV is indicative of an increased/decreased activity of the PNS and SNS respectively [103; 161].

The main physiological measurements for cardiac activity are the *electrocardiogram (ECG)* and the *photoplethysmogram (PPG)*. The first captures the electrical activity of the heart by placing electrodes on the skin. The PPG data, also known as *Blood Volume Pulse (BVP)* is obtained using an optical method.

Devices often used for recording cardiac activity for students and workers' engagement are chestbands [47], wristbands [12], and ear clip [41]. Cameras

[3] `https://www.emotiv.com/insight/`

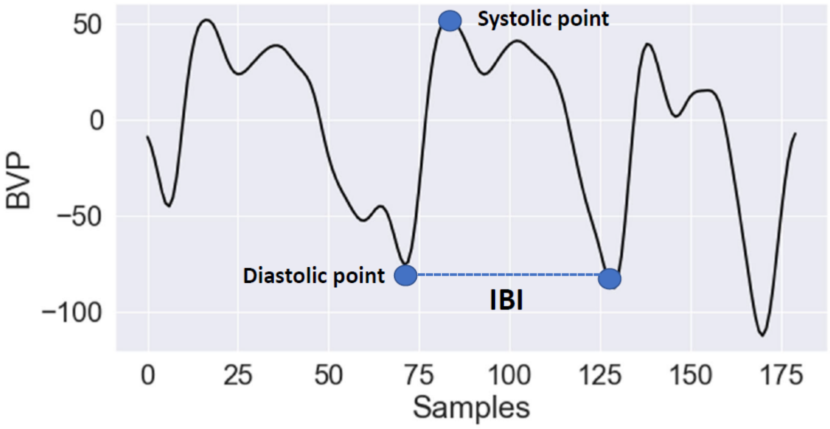


Figure 2.5. Example of the BVP signal we collected using the Empatica E4 wristband.

have been also used to derive heart rate measurements using computer vision techniques [24].

In the studies presented in this thesis we collected the BVP signal, with a sampling rate of 64Hz, using the E4 wristband equipped with a PPG sensor.

The BVP indicates changes of the blood volume in the peripheral blood vessels [162]. The raw BVP can be used to monitor changes in the blood volume and consequently characterize the cardiac activity. Since the size of the blood vessels is controlled by the SNS, the BVP has been widely used as measure of sympathetic arousal [163; 164; 165]. The BVP signal can be used also to derive measures of heart rate and heart rate variability.

Indeed, the HR and HRV can be calculated from the time intervals between peaks (or valleys) in the BVP referred to as *inter-beat-interval (IBI)* [162]. The IBI is often referred to as RR interval (the R letter is used in relation to the R-peaks of traditional ECG signal) or NN interval (the letter N is often used for indicating “normal” RR intervals, i.e, intervals that are free of artifacts and represent normal cardiac timing).

Significant points of the BVP signal, often used to derive the IBI, are the *diastolic*, i.e., local minima, and the *systolic*, i.e., local maxima, points. The difference between the diastolic and systolic points, corresponding to the difference in terms of the PPG reflected and absorbed light between the most and least oxygenated blood conditions, can be used for estimating the vasoconstriction of the individual. Figure 2.5 shows an example of the BVP signal we collected.

Several features have been proposed in the literature to quantify the cardiac

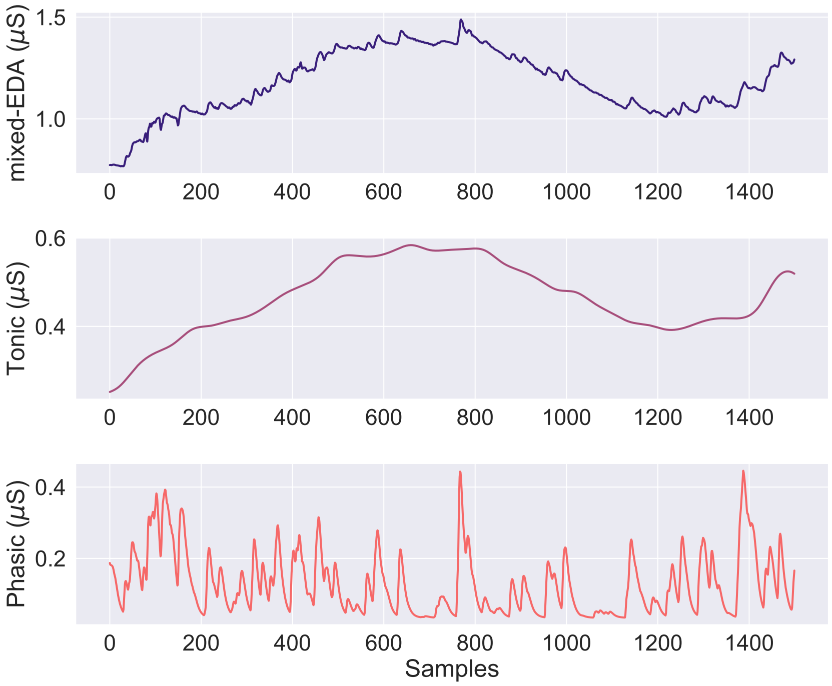


Figure 2.6. Example of the EDA signal (top) and its tonic (center) and phasic components (bottom).

activity [76; 77; 78; 79].

Examples of features describing the HRV in the time domain are: the mean of the RR intervals and its standard deviation (SDNN), the standard deviation of differences between adjacent RR (SDSD), the square root of the mean of the sum of the squares of differences between RR (RMSSD). The IBI is often also processed in the frequency domain using methods such as the Fast Fourier Transform (FFT), from which the Power Spectral Density (PSD) is estimated. Typical frequency-domain features are: total spectral power of all RR samples in the 0.05-0.15Hz – low frequency (LF) – , between 0.15-0.5Hz – high frequency (HF) – and the ration between the two (HF/LF).

Features representing the HRV in the temporal and frequency domain have been used for recognizing students [24; 12] and workers' [45; 47] engagement. For instance Rissler et al., found that the HF/LF ratio, representative of the autonomic balance [29], was the most discriminative feature for distinguishing low from high levels of flow experienced by workers [47].

Electrodermal activity

The *electrodermal activity (EDA)* also known as *galvanic skin response (GSR)* broadly refers to changes of the permeability of the skin due to the eccrine sweat glands activity [35]. The EDA is frequently measured in terms of *skin conductance (SC)* [71].

The activity of the eccrine sweat glands is regulated by the Sympathetic Nervous System (SNS), thus EDA signals are considered as a direct measure of the SNS and thus of the physiological arousal [35].

Given its connection with alertness and arousal, the EDA signal is a promising proxy for measuring students [12; 34] and workers' [28] engagement.

The EDA signal can be measured using small electrodes placed on the skin. Ad hoc sensors or wrist-worn devices as the Empatica E4 [58] have been used for collecting EDA data with the goal of measuring students [12] and workers' engagement [28].

In the studies presented in this thesis we used the E4 wristband to collect EDA data. The E4 measures the electrical conductance of the skin thanks to a small current passing between two dry electrodes positioned on the skin. The E4 uses a sampling rate of 4Hz for EDA measured in microSiemens (μS).

The EDA is characterized by *Skin Conductance Responses (SCRs)*, whic are signal's peaks occurring in response to stimulus and are indicative of the SNS activity [35]. The SCRs have a peculiar shape, consisting in a steep increment to a peak and a consequent exponential decay. The SCRs typically lasts 1-5 seconds and have an amplitude of at least $0.01(\mu S)$ [35; 68].

It is common practice in the literature to decompose the EDA signal, to which we often refer to as *mixed-EDA*, into its *tonic* and *phasic* components [35; 67; 166]. These two components differ for timescale and relationship with the stimuli [71] and they can thus be used to infer different information from the EDA signal. More precisely, the tonic component corresponds to a slowly varying signal and it is characterized by one-minute scale fluctuations [167]. The phasic component is superimposed to the tonic one and it is characterized by rapid changes and spike-like features [166]. The phasic component is related to stimulus responses connected to novelty, significance and generally attention. The tonic component is instead connected to general alertness and physiological arousal [168]. Figure 2.6 shows an example of an EDA signal we collected using the E4 wristband, as well as the tonic and phasic components.

Statistical features, and features describing the SCRs are often extracted from EDA and its components to describe the physiological and thus emotional arousal [67; 68; 73; 74; 75]. Examples of statistical features often used are: minimum, maximum, mean, standard deviation, dynamic range (i.e., difference between maximum and minimum). Features as the mean and standard deviation of the first and second derivatives of the EDA signal are often computed to determine the direction and amplitude of signal's changes. Typical SCRs features used are: area under the curve, rise time, decay time, number of peaks, peaks' width and amplitude.

In this thesis we investigated the role of typical EDA features together with a set of theoretically-motivated features that we proposed, in the recognition of students' emotional engagement during lectures. More details are presented in Chapter 3.

Further, we investigated fusion strategies based on EDA and BVP signals, together with context, to recognize workers' flow state during work activities. More details are presented in Chapter 5.

2.3.4 Multi-modal expressions

These type of expressions involve the concurrent interaction of multiple modalities such as physiology, face, body, voice.

Examples of multi-modal expressions connected to engagement or the lack thereof and investigated in the classroom or during work activities (e.g., meetings) are laughs [169; 170] and yawns [143; 171]. To detect these expressions multiple sensors (e.g., cameras, inertial sensors, physiological sensors) are often involved. In this thesis we focus on the automatic recognition of laughter.

Laughter

Laughter is a multi-modal behavioral expression characterized by the combination of several body reactions. Humans express and perceive laughter in different ways. However, laughter is naturally and universally recognizable from all individuals [172].

Specific behavioural patterns make laughter easily identifiable, indeed a laughter episode (i.e., the whole multi-modal event related to laughter [173]) is characterized by an *onset*, when the face suddenly change into a smiling-expression, an *apex*, when the rhythmical exhalation and vocalization happen, and an *offset*, the moment when the vocalization ends and returns to a smile [172]. In addition to the vocal and facial expression generated during laughter episodes, physical and physiological reactions are expressed as well. However, their investigation is under explored compared to the vocal and facial cues.

Several body movements happen during laughter episodes. As reported by Ruch and Ekman in [174], the movements related to the forced expiration during laughter episodes, generate vibrations of the trunk and shaking of the shoulders which could be captured using accelerometers either placed on the body or on the extremities [174]. We refer to these movements as *respiration-movements*. In addition to those, other idiosyncratic body movements not related to the respiration, defined as *full-body movement* in [175], might be observed , e.g. “rocking violently sideways” or “hands throwing” [174]. These movements are more expected during social interactions among group of people to express the intensity of the emotional arousal [175; 173].

Very little is known about the physiological reactions generated during laughter. Ruch in [176] describes laughter as a behavioural indication of higher exhilaration and provides a physiological characterization of laughter episodes [176].

In particular, the author [176] underlines that “characteristic cardiovascular changes and fluctuations in electrodermal activity (EDA) can also be observed during laughter” [176]. The author reports also acceleration of the heart rate, diastolic and systolic blood pressure increments as well as changes in the peripheral blood pressure between laugh and smile [176]. He noticed that significant changes in the EDA happens during laughter, however whether they come from the respiration changes or not remains still an open question [176]. In general, it seems that a humorous stimulus activate the Sympathetic Nervous System and it is then measurable with electrodermal and cardiac activity [177].

Based on what stated in the literature, we hypothesize that laughter can be recognizable using the combination of physiological and respiration-movement data. In this thesis we propose a novel method to recognize laughter based on the combination of physiological sensors (EDA and BVP) and inertial sensors (ACC) gathered from a wrist-worn device, more details are presented in Chapter 4.

Even though researchers have widely demonstrated the validity of the above-mentioned behavioural expressions for detecting engagement, it is important to notice that the choice of the expressions to investigate and the sensors to use depends on several factors.

Particular behavioral expressions might be expected in specific contexts and not in others. For instance, in this thesis we consider laughter episodes as multimodal behavioural expression of engagement. However, this expression might occur in situations in which social interactions take place [178] (e.g., during meetings, lectures) or during breaks, while it might be less likely that individuals laugh while focused in a work activity. Thus, it is important to have information about the context in which engagement is monitored in order for the measurement to be effective.

The choice of the behavioural cues depends on the component of engagement to measure. In this thesis we focus on the affective component of engagement and measure students' emotional engagement and workers' flow. Thus, we chose to mainly focus on *internal* indicators of engagement and in particular on the physiological activation cue.

Further, the choice of the sensors should be also guided by constraints of the workplace and classroom in terms of ease of access, comfort level, unobtrusiveness and social acceptance [41]. For instance, given the interaction of the subjects with computers, computer's frontal cameras could represent a valid sensor to use in online learning for detecting individual students' facial expressions or even physiological parameters. However, having an always-on cameras in contexts such as classrooms or offices can be considered privacy invasive as discussed in [41; 158]. For instance, many institutions do not allow cameras to be used in classrooms and students who are not part of the study, or are not willing to be monitored are difficult to exclude from the recordings [179].

EEG sensors or chest bands are often considered not socially acceptable or uncomfortable, thus requiring subjects to use these sensors for prolonged time might cause individuals to drop the study or stop using the system.

In this thesis we use physiological signals (EDA and BVP) and inertial sensors (ACC) gathered from an off-the-shelf wrist-worn device. This device, is not invasive and being a watch-like device could be more acceptable by e.g., students [180]. Further, with this device, we do not connect engagement recognition to specific environment or task.

2.4 Engagement labeling methods

The main methods for gathering labels about students and workers' engagement during activities consists in using *ESM* and *external raters*. Methods for gathering labels (or *ground-truth*) are grounded on the traditional methods for measuring engagement presented before. Cognitive signals could be also used for deriving engagement labels. In the followings we provide a brief overview of engagement labelling methods used in the literature.

2.4.1 Engagement labeling using ESM

The ESM is generally used when the goal is to assess the engagement as perceived by the individual. The ESM is based on the assumption, that being engagement a subjective experience, the best way of getting ground-truth about individual's engagement is by directly asking the subject itself. This method is also more suitable when the goal is to determine the affective or cognitive components of engagement since they are not directly observable [95].

The ESM questionnaires can be triggered during the activity (*concurrent* reports) or at the end of the activity (*retrospective* reports) [181]. For instance, Monkaresi et al. [24] sent auditory probes every two minutes during a writing activity, for asking students to verbally report about their level of engagement [24]. The proposed method allows to get a fine-grained labels and could be beneficial for building precise real-time engagement models. However, this method is highly invasive and might distract the individual causing her to leave the engagement or flow state [92; 41]. Sending the questionnaire at the end of the activity can overcome this problem, however being retrospective, this approach might suffer of "recall bias" [182] i.e., participants could report their engagement inaccurately. The ESM questionnaires for collecting labels are often prompted on a device such as smartphone or laptop, or using a paper form.

Recently, *Situated Self-Reporting (SSR)* devices have been proposed as alternative tools for collecting self-reports in situ [63]. According to the definition provided by Paruthi et al. [63] a SSR device is a "situated device intended to be placed in a location to optimize user's self-reporting efficacy". In contrast to smartphones or laptop, SSR are often single-purpose devices, and their design should guarantee a minimized access time and fit with the study's context and physical environment [63].

In general, the *ESM* strategy is often used when user studies are conducted in real-settings as during actual lectures in classroom as in [34; 12] or in the workplace as in [47], while *external raters* are mostly involved when targeting laboratory settings especially during simulated learning activities as in [127].

2.4.2 Engagement labeling using external raters

The *external raters* method is based on the traditional *observations and rating scales* method. The external raters method consists in asking human observers to watch recorded videos of subjects performing a task and assign an *engagement score* based on their estimation of the subject's engagement or attention level [127; 171].

Guidelines are often provided to the raters for determining the level of engagement, for instance Whitehill et al. [127] provided the following guidelines to the observers considering: *Not engaged at all* – e.g., looking away from computer and obviously not thinking about task, eyes completely closed. *Nominally engaged* – e.g., eyes barely open, clearly not "into" the task. *Engaged in task* – student requires no admonition to "stay on task". *Very engaged* – student could be "commended" for his/her level of engagement in task. *X*: The clip/frame was very unclear, or contains no person at all [127].

The external raters method is more suitable when the goal is to develop a system able to quantify the engagement as perceived by others (e.g., similar to a teacher that "judges" the students from how engaged they appear) [127]. However, methods involving observations are highly inferential [95], lack of the "learner's perspective" [181], are expensive and impractical in large settings (e.g., when classrooms with multiple students are monitored).

2.4.3 Engagement labeling using cognitive signals

Cognitive measurements derived from physiological signals as EEG could be also used as alternative methods for generating labels about engagement as done in [137] for modeling user's engagement with the smartphone [137].

The EEG-based engagement index introduced before and proposed in [160] could be for instance used for this purpose. The main advantage of using cognitive signals stands in the possibility of gathering a continuous assessment of engagement without interrupting the flow of the task in which the user is involved in. However, in the context of lectures the usage of EEG headsets could be considered invasive by students, given the uncommon usage of these devices in everyday life, students could have perceived them as unnatural. Similarly, in the workplace context, the usage of the EEG headset might be inconvenient for prolonged data collection.

However, the use of cognitive signals could be beneficial for benchmarking the engagement during short and specific work activities (e.g., short coding sessions), in which the user could feel more comfortable in wearing the device.

In this thesis we use the retrospective ESM strategy. Specifically, we use validated questionnaires sent at the end of the activity to get labels about engagement. We chose this strategy due to the focus on the affective component of engagement as perceived by the individual (i.e., the student and the worker), to not interrupt the individual during the activity, and since we conducted user studies in real-settings where multiple people were monitored concurrently (i.e., multiple students in classrooms).

To collect labels in our "in-the-wild" studies we used paper forms, we designed and developed a smartphone application, laptop widget and a SSR device called *Devo*. Details about the methods we used are described in details in Chapter 5.

2.5 Engagement modeling

Refers to the creation of models that map features into engagement labels. Machine learning (ML) algorithms are used for this purpose.

2.5.1 Machine learning algorithms for modeling engagement

One categorization of ML algorithms refers to the amount of *supervision* needed by the algorithm, this divides the strategies in: *supervised*, *unsupervised*, *semi-supervised* and *reinforcement learning* [72].

In this thesis we focus on the *supervised* approach. In *supervised learning*, the data set consists in a set of labeled instances $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where the element $\mathbf{x}_i$ is a *feature vector*.

Each dimension $j = 1, ..., D$ of the feature vector contains a value called *feature* ($x^{(j)}$). For instance, if each example in the data set describes the physiological activation of a student during a lecture, the first feature $x^{(1)}$ could correspond to the number of *EDA*'s *SCRs* generated by the student during the lecture. If the label y_i belongs to a finite set of *classes* $\{1, 2, ..C\}$, e.g., $\{engaged, non\_engaged\}$ used in [24], the learning problem is called *classification*. While, if the label is a real number, e.g., an *engagement score* used in [12], the learning problem is called *regression* [72].

To the best of our knowledge, supervised learning is the prevalent method used in the literature for targeting the students' and workers' engagement recognition problem [24; 47; 41; 24; 28; 171; 12; 127; 183].

Most of the existing studies target the problem as a classification task [24; 47; 41; 24; 28; 171], either binary classification [24; 47; 28] – as *yes-or-no* phenomenon (e.g., engaged vs non-engaged) [24] or *low-high* levels [47; 45; 28] – or multi-class classification – as multiple levels (e.g., *low-medium-high*) of engagement [171] or engagement against other affective states as stress or boredom [41] –. Few studies used a regression strategy [12; 127; 183].

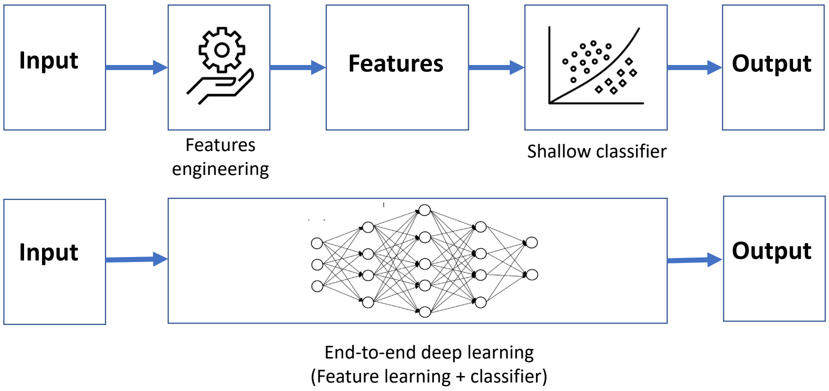


Figure 2.7. Traditional machine learning (top) and end-to-end deep learning (bottom). The figure is adapted from [2].

In this thesis we target the engagement recognition as a classification task.

We consider another categorization of ML algorithms: *traditional machine learning* and *end-to-end deep learning* as in [2] and summarized in Figure 2.7.

In traditional machine learning, features are explicitly engineered by the human expert and the algorithms usually have a *shallow* structure, i.e., the parameters are learnt directly from the training samples' features [72]. In particular, in *traditional machine learning* the feature extraction and model construction are performed in a separate manner.

Examples of shallow learning algorithms adopted in engagement recognition are *Support Vector Machine (SVM)* [72] used in [45], and *Random Forest (RF)* [72] used in [47].

One advantage of traditional ML methods, consists in the interpretability of the feature space. Indeed, traditional ML methods allow, using techniques such as features selection or feature importance, to determine which are the most discriminative features that lead to a particular output (e.g., engagement levels). Knowing which are the most relevant features for solving a specific problem, could allow to build the trust of the user with the system. For instance, a worker can better understand what are the important information, and in turns behaviours, that often lead her to be disengaged. Researchers could better understand and describe the phenomenon of interest.

However, the process of transforming sensor data into useful features for engagement recognition, requires several steps (i.e., pre-processing, features extraction and features selection) which are often time consuming and demanding.

Instead of relying on hand-crafted features, *end-to-end deep learning* uses data representation learning, and transforms data into abstract representation enabling features to be learnt directly from raw sensor data [2]. In deep learning the emphasis is put on "learning successive layers of increasingly meaningful representations" [31] holding the potential of discovering new patterns in data previously unknown.

Deep learning models enabling these layered representation are called *neural networks (NNs)* [31]. A layer can be considered as a "data-processing module" [31].

Different types of layers and networks exist and their choice and success depends on the task to solve. Examples of NN, which we use in this thesis, are the *Convolutional Neural Networks (CNNs)* [31]. CNNs use *convolutional layers* specialized in extracting spatial structures in one or more dimensions.

CNNs are widely used features extractors for various data sources (e.g., images, text, time-series) [184], CNNs are particularly suitable for capturing local dependencies and translation invariant features in the data [43]. CNNs with multiple layers allow representation modularity, initial layers extract low-level and simpler features directly from the raw data, going deeper in the network the extraction of more high-level features is obtained [43]. Typically used for images, CNNs have been also successfully used for processing sensor data (1D structure) in the Human Activity Recognition [185; 186] and Affective Computing [36] fields.

Temporal dynamics in sequential data (e.g., video, audio) can be exploited using layers as *Long Short Terms Memory (LSTM)* [31] and *Recurrent Neural Networks (RNNs)* [31]. These methods are also widely used for affect and activity recognition [185; 36].

In the context of students' and workers' engagement recognition, deep learning methods [187; 188; 189] are less investigated compared to the traditional ML ones [28; 45; 47; 41; 24; 127; 12]. Few recent approaches [187; 188; 189] applied deep learning methods to image and video data to recognize students' engagement in e-learning environment. These approaches use the *Dataset for Affective States in E-Environments (DAiSEE)* presented in [190] which became also part of the *Emotion in the Wild (Emotiw)* challenge [191]. In other contexts such as engagement recognition during gaming [192] or in children interaction with robots [30] physiological signals in input to deep learning models have been used.

2.5.2 Multi-modal engagement recognition

As discussed before, engagement is a multi-dimensional construct and can be expressed through different behavioural cues. This suggests that using a combination of different modalities, measured using sensors, to represent engagement could improve the engagement recognition performance [30].

In the Affective Computing field, it has been widely demonstrated that exploiting the complementary information provided by different sensors through their combination can improve the affect recognition performance [29; 36].

Several solution for combining data have been proposed in the literature [36]. Existing strategies allow to combine data at different levels, from raw data to independent models [36].

Fusion strategies can be broadly divided in *feature fusion* and *classifier ensemble*. The first refers to the combination of information extracted from the different modalities to create a feature vector used in input to the classifier. In the second case, the decisions took from independent classifiers, trained using the information from a single modality, are merged in a unique decision [193].

In the classifier ensemble approach, it is important to train independent and uncorrelated classifiers (i.e., using different algorithms or different features [72]) increasing the possibility of making different type of errors and thus increasing the ensemble's accuracy [66]. The predictions of different classifiers can be combined using a *hard-voting* or *soft-voting* approach [66]. In the first case, the predicted class is the one that gets the majority of the votes, in the second case, the class probabilities estimated by each classifier are combined with a statistic (e.g., mean, median) and the class with the highest probability is predicted [66].

The majority of fusion strategies adopted for students' and workers' engagement rely on the use of hand-crafted features and shallow models [28; 41; 12].

The *feature fusion* strategy using shallow classifiers, often referred to as *features concatenation (FC)* [186], is obtained by concatenating in a single feature vector a set of hand-crafted features derived independently from each modality using sensor-specific processing techniques [186]. The FC is a widely used approach, however it does not allow to exploit the complex relation that might exist across different data sources at different levels [42; 186; 43; 43].

Deep learning presents a possible solution to this problem, allowing an automatic hierarchical construction of features within and across-modalities [186; 42]. Deep neural networks allow additional flexibility in the development of feature fusion strategies, allowing features automatically extracted from sensor modalities to be merged at different stages [193; 44].

In this thesis we use traditional ML for students' engagement recognition, we use a feature-concatenation strategy for laughter detection, and explore different fusion strategies based on traditional ML and end-to-end deep learning for workers' engagement recognition. More details are presented in Chapter 3, Chapter 4 and in Chapter 5.

2.6 Context definitions

The role of *context* in the design of responsive computer systems have been widely investigated in the fields of Human-Computer Interaction (HCI) and Ubiquitous Computing (UbiComp). Context-aware sytems, are a category of computer systems that *adapt* to context by changing their behaviour according to the "sensed" context.

The central premise for investigating context is to mimic the interactions among humans. When communicating, humans are indeed capable to convey information and appropriately react. This capability is mainly due to the richness of their language and the "implicit understanding of everyday situation", or *context*. Thus, giving systems the ability to recognize and respond to context could improve their interactions with humans [27].

Several definitions of the term context exist, in this thesis we rely on the widely used definition provided by Dey et at al. [27]:

"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."

In other words, according to the authors, any information that can characterize the situation of the *entity* is considered as context. In this thesis we consider as *entity* the users of the engagement-aware system, i.e., students and workers.

The authors argue that *primary* context types important to consider for describing the situation are the *location*, *time*, *identity* and *activity*. [27].

In this thesis we focus on *activity* and *time* context types referring to the workplace. The activity describes *what* is occurring in the situation and in particular what the user is doing, the time refers to *when* it is happening. In the followings we provide definitions of workplace activities as well as briefly describe existing methods for automatically recognize them.

2.6.1 Workplace activities definitions

Human activities involve the movements of one or several parts of the body and have an inherent hierarchical structure [48].

According to Zhang et al. [48] human activities can be considered in terms of a three-level categorization. At the bottom the authors place the *atomic actions* at the limb level performed by a specific part of the body (e.g., arm, hands, upper body). An example of a atomic action is "rising the left hand".

In the second level, the authors consider the *actions*. The term action refers to whole body movements composed of several atomic actions in sequential order by a single individual with no objects involved. Examples of actions are "walking", "running".

In the upper level, *interactions*, are human activities that involve either two or more people, or objects. An example of interaction is the "cooking" activity which involves a single (or multiple) person(s) interacting with object(s) (e.g., pots and pans).

In this thesis we consider human activities at the upper level and use the term *interaction* and *activity* interchangeably. We focus on recognizing workers' activities in the workplace, and in particular we aim to distinguish between *work* and *break* activities. During these activities people interact with different object(s) (e.g., laptop, books, pens) and with different people (e.g., colleagues).

For the definition of work activity or *task* we rely on the one proposed by Meyer et al. [52], which describe a work task as a: *"well-defined work assignment with a specific goal that people divide their work in"* [52]. Examples of work activities, according to the authors, are fixing a code bug, or preparing for a meeting.

Knowledge workers are often responsible for managing and self-organizing their work, thus it is important to consider a work task in terms of how workers divide the work day. Further, as also pointed in [53], knowledge workers use different digital and physical tools for performing their task and do not think about their activities in terms of the type of application they use. Indeed the same application can be used for different purpose. Thus not making assumptions on what tools or devices define work activities is an important aspect to consider.

In this thesis we further investigate the role of the type of work activities in the assessment of workers' flow. Different categorization of type of work activities at different granularities have been investigated in the literature [194; 61; 53; 52], in this thesis we refer to a set of activities identified in [194] and in [61] which are also common to the ones identified in [53; 52] such as *meeting, learning, read/write, planning, email, coding, research project, other*.

Despite the broad consensus on the existence and benefits of breaks, there is no a single definition for breaks [195].

In contrast to external and internal interruptions [196], which are not triggered or desired by the user, work breaks can be considered as “moments in which workers voluntarily pause their work” [195]. The term *work break* comprises of a large amount of different activities [195] which vary across individuals [195] and do not necessarily imply the user to leave her workstation [197].

Recently, work breaks for knowledge workers have been divided in *digital breaks* (e.g., watching videos, visiting social media website, consulting news) and *physical breaks* (e.g., activities which involve physical activities such as walking, going to take a coffee, doing physical exercises) [3; 51]. The benefits of both types of work breaks have been widely proved. Physical breaks allow workers to interrupt sedentary work and consequently improve their physical health, while digital breaks can increase productivity and reduce boredom [195; 3].

In the following section we provide an overview of existing automated methods, proposed in the literature, for the automatic recognition of activities in the workplace.

2.7 Activity recognition in the workplace

Existing approaches on activity recognition in the workplace rely on automated methods. We divide existing works in: *sensor-free* and *sensor-based* methods.

Sensor-free

Most of existing work aiming to characterize knowledge workers’ activities rely on *computer interaction data* [61; 52; 50; 53]. These approaches quantify *digital* or *online* activities using a combination of computer interaction data and ESM [61; 52; 50; 53].

Studies in this category relied on computer interaction data to quantify the multi-task, task-switch behaviour [52], or to connect the activity type to the perceived engagement [50].

Data about the interactions of the workers with the computer and/or enabling devices (e.g., keyboard, mouse), are often collected using custom built-in software as in [50; 61; 198; 52] or using third party services as *uLog*[4] such as in [53].

Typical information retrieved are mouse and keyboard activity, documents opened, websites visited, application used [53; 50; 61; 28]. Examples of features

[4] https://www.noldus.com/observer-xt/modules

extracted, in a time window of few seconds [61], five [53] or 10 minutes [50], are time spent on a given application (or category of application) as Facebook (or social media category) [50], most used application (or category), number of mouse clicks, number of spaces and backspaces from the keyboard [53; 50; 61].

Features are used to analyze, using descriptive or inferential statistics, patterns of workers' behaviour with the aim of describing the work dynamics and work fragmentation [61] or the connection with attention levels [50].

Approaches relying on computer interaction data, often use heuristics (e.g., time spent in a state higher than a pre-defined threshold) and characterize breaks as moments in which computer *inactivity* or "idle state" (i.e., lack of interaction of the user with the computer or devices) is observed as opposed to work activities distinguished by lack of inactive moments [61; 3]. Other approaches, considered breaks as times spent interacting with applications related to social media, news or shopping, hint for *digital breaks* [51; 199; 200].

Type of work activities have been derived using the name or title of the application or website in a segment and then mapped to high-levels categories using a pre-defined mapping generated with a semi-automated open-coding process [61; 201].

Few approaches, used computer interaction features and self-reported activities in input to machine learning models to automatically recognize the type of work activity as defined by the user [53; 52].

Information about the workers' current activity have been also retrieved by personal calendars as in [28] in particular for meetings.

Computer interaction data represent a valid proxy for quantifying online or digital activities of workers. However, workers perform several activities that do not strictly require the use of computers (e.g., take breaks, participate in meetings, read books). Complementing digital information with data derived from sensors could represent a promising solution to overcome this limitation.

We summarize below existing sensor-based approaches for recognizing activities in the workplace.

Sensor-based

Sensor-based methods for the recognition of activities in the workplace rely on data gathered from sensors such as cameras [54; 51], microphones [54], RF-radar [55], accelerometer [56; 202] and bluetooth beacons [56].

Features derived from sensor data are then used to represent cues such as desk presence, conversations, gestures and movements. These features are then processed using pre-defined heuristics [56; 51], statistical models [54] or machine learning models [55] for the automatic assessment of activities in the workplace.

Several approaches aim to recognize *static* desk-work activities such as *reading*, *hand-writing*, *keyboard typing* [202; 55]. Others focus on *dynamic* work activities such as *walking*, *climbing stairs* and *sitting-standing* [202]. Few existing approaches considered activities involving other people such as *face-to-face conversations* [54].

Most of the existing studies presented in the literature were conducted in laboratory or controlled settings, in which researchers asked participants to perform activities using a pre-defined protocol [202; 55].

For instance, Avrahami et al. in [55] explore the use of a RF-radar sensor placed under the worker's desk for the recognition of deskwork activities in an experimental setting. The RF-radar sensor, equipped with several antennas, measures the strength of the signal reflected by the area above and generate a 3D representation that authors mapped in a 2D image representation.

According to the authors [55], the RF signal representation generated should serve as hint for recognizing the gestures and the activities happening on the desk. The authors used the image pixels intensity in input to shallow classifiers and manage to recognize activities in the office such as *reading*, *writing* and *eating* with accuracy in 83-95% range [55].

Olivier et al. in [54] used data derived from microphones, cameras, keyboard and mouse as input to Layered Hidden Markov Models to infer the type of activity the user was performing. In particular, the authors used a two-layered representation approach. In the first layers the authors used audio data to derive particular office's sounds (e.g., conversations, music, keyboard typing), and video data as *person detector*, to infer the presence of one or multiple people in the office. In the second layer of the model, the output of the first layer together with features that characterized keyboard and mouse activity are used to determine the higher-level activity the worker is performing. With their proposed method, the authors distinguished among work activities such as *phone conversation*, *face to face conversation*, *ongoing presentation*, *distant conversation*, *nobody in the office*, and *user present* [54].

The main sensor-based cues used in previous work for determining whether a worker is taking a break, or not, are: absence from the desk [56; 51] (sensed using cameras [51] or Bluetooth beacons [56]), physical movement [56] (assessed by counting the number of steps).

For instance, Cambo et al. [56] designed and evaluated a break recommendation system called *BreakSense*. To detect when workers were taking a break, the authors used a combination of physical movement and distance from the desk. In particular, the authors used the *receiver signal strength indicator* (RSSI) of Bluetooth beacons placed in the office to calculate the distance of the worker from their room, and gathered physical activity information from Microsoft Band 2 wristband. Based on pre-defined heuristics, the *BreakSense* system identifies as breaks moments in which the worker is *far* from the room and the physical state changed to *walking* [56].

Kaur et al. [51] investigated the use of data derived from different sources (e.g.., audio, video, computer interaction) to predict opportune moments for recommending breaks or transitions to another work task. In their implementation, authors considered as breaks moments in which workers visited *distracting* websites, hint for *digital breaks* and in which absence of data (video and computer interaction data) was recorded, hint for *physical breaks* [51].

Building upon findings presented in previous work, in this thesis we used data gathered from personal devices (i.e., wristbands, smartphone and laptop) to capture different cues (i.e., physiological activation, physical movement, phone and laptop usage). We extracted and concatenated features from these data sources and used them in input to machine learning algorithms for automatically distinguishing between work and break activities. More details about the proposed approach and the results we obtained are presented in Chapter 5.

Chapter 3

Recognition of Students' Emotional Engagement From Electrodermal Activity Data

Students' engagement is known to be a predictor of students' learning progress and academic achievements [25]. Fostering students' engagement is considered one of the most effective countermeasures to prevent students' drop out, disaffection, alienation and low academic performance [57].

Technology can play a fundamental role in understanding and reacting to students' engagement or lack thereof during lectures. An engagement-aware system can provide students with information about their own level of engagement with the goal of driving self-reflection and possibly behavioral changes. If information about students' engagement during lectures can be made available to teachers, they can too self-reflect about their teaching performance, design and evaluate methods to (re-)engage students.

To enable the creation of feedback systems to (re-)engage students, it is however necessary to first devise effective methods to recognize students' engagement.

In the study presented in this chapter we propose to recognize students' engagement using Electrodermal activity (EDA) data. EDA is a measure of individuals' physiological arousal and, as extensively discussed in Section 2.1 and Section 2.3.3, it is considered as a valid proxy for measuring engagement [67; 34; 33].

While there exist several different definitions of engagement, in this thesis, we consider the well-accepted characterization of students' engagement as the combination of behavioral, emotional and cognitive engagement [95]. We then focus in particular on the emotional engagement – to which, for simplicity, we also refer to as *engagement* in the rest of the chapter. Emotional engagement is linked to students' affective state and is connected to emotional reactions to teachers [57].

To address the engagement recognition problem, we built upon literature from educational research and identified three components of emotional engagement: *momentary engagement* [87], *emotional arousal* [89] and *reaction to the teacher* [88]. We then derived and proposed a set of features from EDA to represent these components.

Further definitions of emotional engagement and the considered components are also provided in Section 2.1.

In contrast to existing work – which considered the recognition of students' engagement in e-learning contexts or in laboratory settings [24; 127] – we focus on the learning activity of attending lectures and monitored actual lectures in classroom setting.

Few studies examined student engagement during lectures in the classroom [33; 34; 179]. Fujii et al. [179] used a camera to record students during lectures and assess their engagement based on their head pose. However, as also reported by the authors, this system has significant privacy issues, especially for students that are not willing to be monitored. To overcome this problem, we proposed to collect EDA data using unobtrusive wristbands.

In contrast to cameras as used in [179], collecting data using wristbands ensures data to be gathered only from students willing to be monitored. Further, being lightweight and unobtrusive, wristbands do not limit students' movements allowing them to perform natural learning activities as taking notes. The authors of [34] and [33] used wearable EDA sensors to measure students' engagement. However, these approaches used simple sensor data representation (i.e., statistical features), and performed statistical analysis only. With respect to these works [33; 34], we collected a larger and more heterogeneous data set, derive a set of theoretically-motivated features and design a machine learning pipeline to automatically recognize students' engagement.

In the study presented in this chapter, we distinguish between *engaged* and *non-engaged* students, we ran an extensive analysis to evaluate the impact of the different features on the assessment of engagement. Our results show that non-engaged students can be identified with high reliability. Using a Support Vector Machine, for instance, we achieve a recall of 81% – which is a 25 percentage points improvement with respect to a Biased Random classifier. Features related to the momentary engagement – computed using a method we proposed – resulted the most discriminative ones.

To evaluate our approach we collected the *Student Engagement Using EDA (SEED)* data set, which – after data cleaning – contains data from 24 students, 9 teachers, and 41 lectures. At the time it was collected (2017), this data set was the largest and most diverse data set collected in-situ during lectures using unobtrusive physiological sensors.

The study presented in this chapter targets the first research question of this thesis, (*RQ1: How can features representing behavioral expressions of engagement be derived from physiological and movement data?*). The results of the presented work have been published as journal paper [A] in the PACM IMWUT (September 2018). Part of the text written in this chapter is reported from the paper [A].

The reminding of this chapter is structured as follows. We summarize existing literature in engagement recognition in Section 3.1. In Section 3.2 we describe the data collection procedure of the *SEED* data set. Section 3.3 and Section 3.4 provide an overview of the data analysis and a discussion of the results. In Section 3.5 and in Section 3.6 we discuss the implications and limitations of our approach. We conclude with a brief summary of the chapter.

3.1 Related work on sensor-based engagement recognition

In this section we review existing literature on engagement recognition using mobile and wearable devices with particular focus on students. Additional discussion about existing literature in this topic is presented in Chapter 2.

Several authors investigated the use of mobile and wearable devices for the continuous and longitudinal monitoring of students' stress [203; 204], happiness [205] and overall well-being [159; 206].

Differently from these studies, we aim using data from wearable devices (i.e., wristbands) to monitor students during a particular learning activity, i.e., attending a lecture in the classroom, and focus on the affective state of *engagement*.

Several automated methods to quantify engagement in different activities have been proposed in the literature [156; 138; 67; 136]. Examples include the use of electroencephalography (EEG) for audience engagement during presentations [156], or during the interaction with the smartphone [137]. Facial expressions have been used for measuring the engagement of television viewers [138].

In contrast to these approaches we rely on the use of unobtrusive devices. EEG headsets are bulky and given their uncommon usage in everyday life, students might consider them as invasive and unnatural. Cameras, besides being cumbersome and expensive to install, make it difficult to monitor only those students that want to be monitored as discussed in [179].

In our work, we focus on monitoring students' engagement using electrodermal activity data (EDA) collected unobtrusive wristbands.

Several attempts have been made at recognizing engagement and general affective states using EDA signals [67; 136; 34; 142; 166; 90; 111]. Some studies explore the use of EDA for detecting audience engagement [136; 142] and for predicting children engagement in child-adult interactions [67].

A core contribution of these and similar approaches – and of our work too – consists in designing features of EDA signals and use them to infer self-reported or observed engagement [67; 136; 142].

Other authors, however, focus on quantifying engagement in different scenarios than the one considered in this study. Since engagement is known to be context-dependent and since there is no unique definition for it [67], considering the specific characteristics of engagement in each scenario is crucial.

While most of the existing literature considers only one dimension of the emotional engagement [34; 33], we explored the role of three different components in the characterization of students' emotional engagement: emotional arousal [89], reaction to the teacher [88] – that we consider in terms of physiological synchrony [110] –, and momentary engagement [87].

Cain et al. [90] considered the interaction of two children during learning activities and propose to use the peaks of the EDA signal to detect the momentary engagement of the children.

We build upon their work and propose several new features that capture the presence (or absence) of arousing moments during lectures. Our results show that two of the new features we proposed contribute the most to the discrimination of engaged and non-engaged student. When used as input to a SVM classifier, these two features allow to achieve the best recognition performance.

Other authors investigated the role of *physiological alignment* – also referred to as *physiological synchrony* – between two individuals to infer engagement in social interactions [111; 67]. Similarly, we investigate if considering the physiological synchrony between teachers and students helps inferring students' emotional engagement. Several attempts at recognizing students' engagement and general affective states have also been made. However, as also reported in [24], a direct comparison of different techniques is difficult. The data sets used in other studies usually differ and the labeling methods used are often different. Specifically, while some authors prefer observations or interviews to gather ground-truth data [127], others rely on experience sampling method (ESM) and asked students to report engagement using self-reports [34].

We too rely on self-reports for labeling the data set. This method has been shown to be one of the most appropriate for explicitly measuring students' emotional engagement [95].

Researchers from the Intelligent Tutoring Systems (ITS) community have put significant effort in the development and evaluation of intelligent systems to enhance learners' experience [207; 208; 209]. These authors, however, focus mainly on fostering engagement in an e-learning environment, not in a classroom settings.

A well-explored technique to assess students engagement during learning activities consists in the automated detection of facial expressions [127; 24; 210]. Results from the work presented in [127] show a positive correlation between automatic engagement recognition from facial expressions and human observations of engagement. This approach was however tested in a laboratory setting only and involved the use of a camera, which is an invasive sensor.

Only very few studies measured students engagement during real lectures [33; 34; 179]. Both Wang et al. [34] and McNeal et al. [33] relied on wearable EDA sensors to detect students' engagement.

Wang et al. [34] measured the engagement of 17 students in a classroom and in a remote location during one lecture using EDA sensors. They used ANOVA for reporting affective states differences in the two locations based on questionnaires. Authors did not find significant difference in terms of EDA in the two groups. Data was gathered during one lecture only – which limits the generalizability of the obtained results.

We instead based our analysis on data from 41 different lectures given by 9 different teachers and extensively analyze the collected data.

Mcneal et al. [33] monitored the engagement of 17 students during a term. They tested different pedagogical approaches on subsets of the students and report statistical differences in terms of mean of the EDA traces. In contrast, we relied on a far more heterogeneous data set in terms of courses and students, and ran extensive data analysis.

Lastly, researchers in the artificial intelligence for education (AIED) and the learning analytics and knowledge (LAK) communities focused their attention on detecting students' affective states to foster their performance and learning [211; 212].

Arroyo et al. [211], for instance, created a system that automatically collects physiological data and students' self-reports about their emotional state.

However, they used first generation EDA sensors – which provided much less accurate data than the sensors we used – and did not propose a specific technique to infer engagement from EDA data as we do.

3.2 Data collection of the SEED data set

At the time of writing, data sets available for recognizing learners' engagement are four and contain video data only. A summary of the existing data sets suitable for learner's engagement recognition is presented in [134]. Our goal is to recognize engagement during lecture using EDA data. The lack of an appropriate existing data set for answering our research question, motivated us to collect the *SEED* data set. We provide below details about the participants, the equipment used to collect the data and the data collection procedure.

Participants

We recruited the instructors of five different courses, two at the Bachelor's level (B1, B2) and three at the Master's level (M1, M2, M3). Since some of the lectures we monitored were taught by teaching assistants, we collected data from a total of 11 instructors.

We further recruited 27 students (6 females and 21 males) of age between 21 and 44 (M = 25.64, STD = 4.70) who attended one or more of the monitored courses. The set of participants that attended B1 is the same that attended B2 (only one student did not attend B2). One student attended both M1 and M2.

To recruit student participants, we presented the study during the first lecture of each course. We thereby informed students about the purpose and duration of the study as well as about the devices used and the data collection procedure.

Students who volunteered to participate signed an informed consent form and received a bag of chocolates as a courtesy.

Collected data

Figure 3.2 shows, for each of the monitored courses, the number of teachers and students that participated in the study as well as the number of lectures during which data was collected. A lecture is a single teaching unit. We collected data during 62 different lectures, most of which lasted around 45 minutes (Mean=47.76, STD=17.35).

Sensor data. We recorded physiological data, i.e., electrodermal activity (EDA), of students and teachers using the Empatica E4 wristband [58].

Course	Stud	Teach	Lect
B1	12	2	16
B2	11	4	18
M1	4	1	8
M2	5	2	14
M3	7	2	6
TOTAL:	27	11	62

Figure 3.2. Key data about our study. “Stud” and “Teach” indicate the number of students and teachers that registered for the study for each course. “Lect” indicates the total number of lectures during which we collected data.

Ground-truth data. Refers to the subjective measurement of students’ emotional engagement during lectures. In this study we used the experience sampling method (ESM) and asked participants to fill-in a questionnaire at the end of each lecture. As reported in the educational research literature, questionnaires are a good explicit instrument for measuring students’ emotional engagement [95] because they enable students to report their internal state [24]. Methods relying on external observers or teacher rating scales have instead been reported to be highly inferential [95], as also discussed in Section 2.1 and Section 2.2.

The questionnaire we used is structured in two parts: The first part consists in a screenshot of the Photographic Affect Meter (PAM) [213] while the second part is an emotional engagement questionnaire.

For the analysis reported in this chapter, we considered only the emotional engagement questionnaire. The questionnaire consists of five items related to the emotional engagement dimension of the validated “University Student Engagement Inventory” (USEI) questionnaire [214]. Specifically, the questions are: (1) I didn’t feel very accomplished in this lecture; (2) I felt excited by the work in this lecture; (3) I liked being at this lecture; (4) I am interested in the work done in this lecture; (5) My classroom is an interesting place to be. We thereby slightly adapted the questions to the lecture context. For instance, instead of *I feel excited about the school work* we used *I felt excited by the work in this lecture*.

We used a 5-point Likert-scale (from *strongly disagree* to *strongly agree*). We selected the USEI questionnaire because it is one of the few questionnaires developed for university students. Indeed most of the existing surveys developed for assessing students emotional engagement target younger students [214].

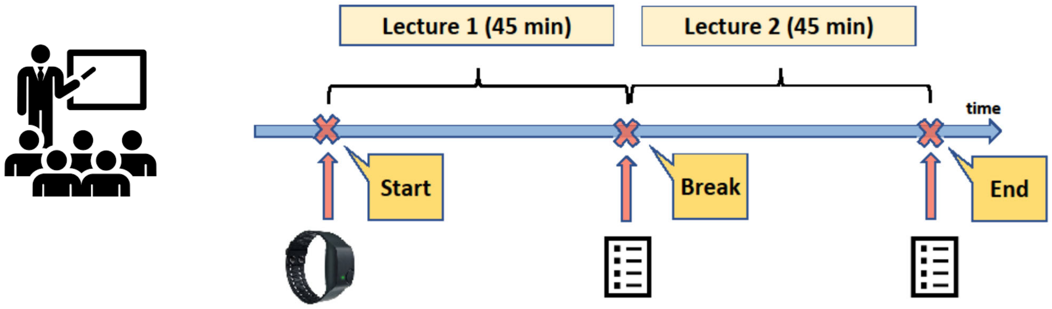


Figure 3.3. Schematic representation of the data collection procedure we followed for collecting the SEED data set.

Data collection procedure

One day before the start of each lecture we charged the devices and synchronized them to a laptop. Shortly before the lecture started, a GoPro camera was placed in an appropriate position for recording the teacher.

Figure 3.1 depicts the setup in one of the monitored classrooms. Once a student or a teacher arrived in the classroom we tied an E4 to the wrist of their dominant hand and handed in the questionnaires. Students were requested to fill-in the questionnaire after each teaching unit (i.e., during the break and at the end of the lecture). Figure 3.3 provides a schematic representation of the data collection protocol we used. At the end of each lecture we recollected the questionnaires as well as the devices. The study procedures were approved prior to the start of the study by the faculty of Informatics ethics delegate.

3.3 Data analysis

The primary goal of our work is to devise a method to discriminate between engaged and non-engaged students using EDA signals. We first cleaned the data set and pre-processed the EDA signals. Then, we extracted a set of representative features and used them as input for the engagement recognition pipeline. We provide below details about the procedure we followed.

Data cleaning

While nearly all lectures of courses B1, B2, M2, and M3 lasted roughly 45 minutes, those of course M1 turned out to be of variable (and usually much longer) length. This hampers the possibility to fairly compare this course with the others and we thus excluded it from our analysis. This left us with 314 traces of EDA data in 44 different lectures. We thereby refer to a *trace* as to the EDA signal collected from a student during a single lecture.

Then, as common procedure in the literature [67], we discarded the signals that presented a large amount of evident artifacts like, e.g., flat responses, abrupt drops and device quantization errors. We did this through visual inspection as in [67]. Figure 3.4 shows representative examples of discarded signals. This cleaning step brought to the elimination of 93 signals, leaving us with 221 data traces. Lastly, we further eliminated 24 signals for which we had incomplete or empty questionnaires.

The number of data traces left after this last cleaning step is 197. These final traces are gathered from 24 students during 41 lectures given by 9 teachers. Their average duration is of 42.87 minutes (STD = 8.86 minutes).

The cleaning of the data set caused the elimination of 117 data traces – equivalent to 37% of the total – out of the 314 raw data traces we had collected. Although this is a considerable reduction, the size both of the final data set and of the reduction itself are comparable to the figures reported by other authors for similar studies.

Hernandez et al. [67], for instance, used 51 data traces – after a data set reduction of 28% – for assessing the engagement of children during playful interactions with adults. Silveira et al. [166] collected EDA signals from 34 persons to classify their ratings of three movies. Martella et al. [215] collected accelerometer data from 35 audience's participants to recognize their enjoyment during a performance.

Visual inspection of EDA signals is a cumbersome and time-consuming procedure. This motivated us, in a subsequent work [70], to develop an automatic approach to recognize EDA artifacts from data collected in ambulatory settings, as well as to release an open-source dashboard, the *EDArtifact*[1], that facilitates the visual inspection process. We used the method presented in [70] to process EDA signals in the study presented in Chapter 5 aiming to recognize workers' engagement. More details about the approach we used are described in Section 5.4.

Pre-processing of EDA data

The pre-processing of EDA data is a crucial step to improve the quality of the collected signals. We followed the same pre-processing procedure proposed in [67],

[1] `https://github.com/shkurtagashi/EDArtifact`

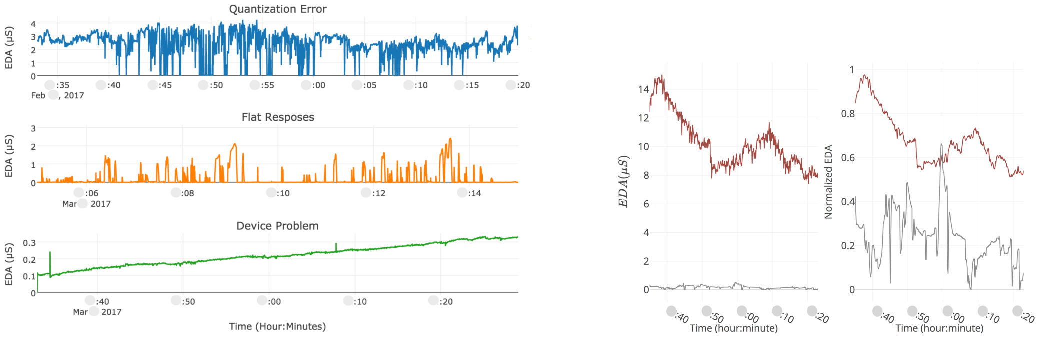


Figure 3.5. Normalizing EDA sensor because of difference in range of EDA values between teacher (red, or dark color) and student (grey, or light color).

which is inspired by [35] and includes the following steps: artifacts removal; normalization; and decomposition.

Artifacts removal. EDA signals collected in uncontrolled environments using wearable devices can be affected by artifacts that can significantly hamper the quality of the analysis [216]. Motion artifacts in particular can cause peaks in the signal, which in turn might be confounded with EDA responses [216]. To attenuate the influence of motion artifacts (MAs) we used a low-pass filter, as it is standard procedure in the literature [217; 67; 11; 203; 218]. More specifically, we used a median filter, as in [217], with a 5-seconds window. This filter reduces the artifacts but preserves the typical EDA edges [217].

Normalization. The amplitude of EDA signals recorded from different persons may present large differences [103]. Since these differences hamper the possibility to directly compare signals [35], we normalized each of the collected data traces using the same method used in [219]. We used the formula in Equation 3.1:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

Where x' is the normalized value, x is the original one, and $min()$ and $max()$ indicate the minimum and maximum functions. Thereby, we normalized the entire signals instead that only the tonic component as in [219]. After normalization, all EDA traces have values in the [0,1] range. Figure 3.5 shows the same EDA signals of a teacher and a student without (left side) and with (right side) normalization. The plot illustrates how normalization makes it possible to compare different EDA signals. Normalization was applied on each individual trace, of a given participant during a specific lecture, independently.

Decomposition. Several mathematical approaches have been proposed in the literature to decompose an EDA signal into its *phasic* and *tonic* components [220; 166]. In this work, we utilized the recently proposed *cvxEDA* method [71] by Greco et al., which uses convex optimization to decompose the signal[2].

Features

Building upon findings in educational research, we considered three components of students' emotional engagement, and characterized them using specific features gathered from students' physiological responses: emotional arousal [89]; reaction to the teacher [88], that we map using physiological synchrony; and momentary engagement [87].

We accordingly define three set of features – indicated as F_{emo}, F_{syn}, and F_{mom} – that can be extracted from the EDA signals and that quantify these three characteristics.

Table 3.1 provides the list and description of all considered features. While we derive most of the features from the existing literature, we also propose a new subset of features to capture momentary engagement (indicated with (*) in Table 3.1).

Emotional arousal features. Emotional arousal is particularly important in the learning context since it can influence memory and performance [89]. To derive information about the emotional arousal and thus the physiological arousal, we used features already proposed in the literature [67; 221; 168], including the mean, standard deviation, area under the curve, average peak amplitude, and number of peaks of the phasic, tonic and raw signals. We computed the number of peaks using a minimum amplitude of 0.01 (normalized) and distance of 1 second as in [67].

Physiological synchrony features. As a further potential indicator of a students' engagement during lectures we considered their emotional reaction to the teacher [57]. Studies from educational research show that as the teacher's emotional response in classroom increases, so does their students' emotional response [88; 101]. This is a phenomenon similar to the *physiological synchrony* – association between the physiological activity of two or more people [110] –, which has been observed connected to emotional engagement in other contexts [67; 111; 222]. More details are reported in Section 2.1.1.

To capture the physiological synchrony between teachers and students quantitatively we used three existing synchrony measures: Dynamic Time Warping

[2] An implementation of this method is publicly available at: `https://github.com/lciti/cvxEDA`

Feature	Description
EMOTIONAL AROUSAL (F_{emo})	
avg_eda/ avg_phasic / avg_tonic	Arithmetic mean of the raw EDA signal/ phasic/ tonic
std_eda/std_phasic/std_tonic	Standard deviation of the raw EDA signal / phasic / tonic
n_p_eda/ n_p_phasic/ n_p_tonic	Number of peaks of the raw EDA signal/ phasic/ tonic
p_a_eda/ p_a_phasic/ p_a_tonic	Average peak amplitude of the raw EDA signal/ phasic/ tonic
auc_eda/ auc_phasic/auc_tonic	Area under the curve of the raw EDA signal/ phasic/ tonic
PHYSIOLOGICAL SYNCHRONY (F_{syn})	
dtw	Dynamic time warping distance
pc	Pearson correlation coefficient
ssi	Single session index
MOMENTARY ENGAGEMENT (F_{mom})	
arousing_normalized*	Ratio between the number of arousing moments and the length of $\bar{S_b}$
arousing_ratio*	Ratio between arousing and unarousing moments
arousing_num *	Number of arousing moments during the lecture
unarousing_num*	Number of unarousing moments during the lecture
unarousing_num_Cain	Number of unarousing moments during the lecture [90]
arousing_num_Cain	Number of arousing moments during the lecture [90]
arousing_ratio_Cain	Ratio between unarousing and arousing moments [90]
level_i*	Ratio between the number of Level_i and the length of $\bar{S_l}$ ($i = 1,2,3,4,5$)

Table 3.1. Summary of the 30 features used in this work (level_i describe generally five features: level_1, level_2,level_3, level_4 and level_5) . The features we proposed have been indicated with an asterisk (*).

(DTW) [67], Pearson Correlation (PC) [67] and Single Session Index (SSI) [222]. DTW and PC have been explored in [67] to analyze the physiological synchrony between children and adults during playful interactions. The SSI is a single index that represents the physiological synchrony between two persons during their interaction. We computed the SSI following the procedure proposed in [222]. Although we use existing indicators to compute the physiological synchrony, to the best of our knowledge we are the first to investigate quantitatively this phenomenon between teachers and students.

Momentary engagement features. A further important component of emotional engagement is *interest* [87]. When an activity in class elicits interest and the student "*readily engages with that activity*" [87], this is referred to as *momentary engagement* [87].

In a recent work Cain et al. [90] suggest that evident increments of physiological arousal identify situations of momentary engagement. Sudden peaks in the EDA signal may however be caused by several stimuli or correspond to situations of acute stress – due to, e.g., the teacher asking a question to the student – rather than to emotional engagement. Even in this case, however, these peaks represent

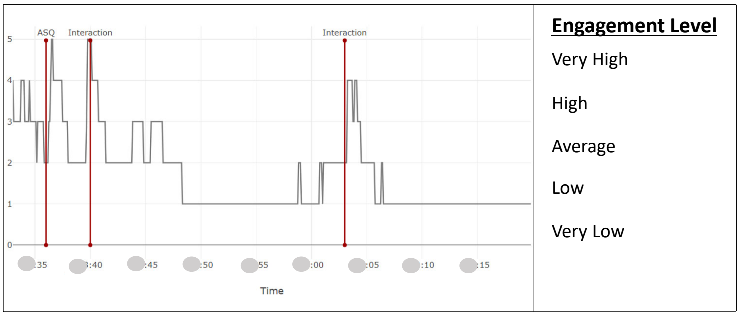


Figure 3.6. Discretized EDA signal and corresponding engagement levels.

valuable information because they contrast to periods without any peaks, which are likely to be indicative of lack of interest and boredom [217].

To capture students' momentary engagement we considered a set of twelve features, listed in Table 3.1. Three of them – `unarousing_num_Cain`, `arousing_num_Cain`, and `arousing_ratio_Cain` – indicate the number of unarousing and arousing moments during a lecture and their ratio, computed using the procedure described in [90].

Thereby, arousing moments, or *highlights*, are defined as those time intervals during which the EDA signal shows a significant increment [90]. All other moments are unarousing moments.

To compute the number of unarousing and arousing moments Cain et al. first subsample the EDA signal by computing the mean of 10-seconds long, not-overlapping windows of the signal [90]. They then compute the difference between subsequent samples, obtaining what they call the *relative change in skin conductivity* (RCSC) [90]. The standard deviation of the RCSC is used as a threshold to discriminate unarousing from arousing moments. More specifically, samples between ±0.25 standard deviation correspond to unarousing moments whereas samples outside this range identify arousing moments.

The method used by Cain et al. for identifying unarousing moments is simple and effective yet sensitive to the presence of artifacts or small random variations in the EDA signal, which could be erroneously identified as arousing moments.

To cope with this problem we proposed an alternative method to identify arousing moments. As in [90], we first reduce the dimensionality of the signal using piecewise aggregate approximation (PAA) [223]. PAA divides the signal into blocks of M signal values[3] and computes the mean of the values in each block. PAA thus transforms the signal into a sequence $\bar{S}$ of the means of the sub-

[3] We use $M = 20$. Since the sampling rate of the E4 wristband is set to 4 Hz, a block corresponds to snippet of the signal of 5 seconds.

sequent blocks [223]. In contrast to Cain et al. we then discretize the signal $\bar{S}$, by mapping its values to five levels of equal width, obtaining a new signal $\bar{S_l}$. We considered five levels to have a correspondence between values of the EDA signals and engagement levels indicated on the questionnaires used to gather ground truth data (Level 1 = Very Disengaged to Level 5 = Very Engaged). Figure 3.6 shows an example of discretized signal and the corresponding levels.

From the discretized signal $\bar{S_l}$ we compute the relative change between two levels as $\Delta l = \bar{S_l}[i] - \bar{S_l}[i-1]$, where $\bar{S_l}[i]$ indicates the current level. We then divide the signal into time intervals of equal length (30 seconds in our current design) and define each time interval as an arousing moment if $\Delta l \geq 1$.

The fulfillment of this condition indicates that there has been an increment of the arousal level during the time interval. Time intervals that do not fulfill the condition are instead labelled as unarousing moments.

The output of this procedure is binary signal $\bar{S_b}$ which we used to compute a subset of the momentary features.

The main difference between our method and the method used by Cain et al. consists in the fact that we discretized the signal and then used the detection of "jumps" between levels to identify arousing moments. This makes our method less sensitive to noisy fluctuations of the signal, which could be misinterpreted as peaks.

Furthermore, the discretization of the signal allows us to consider the levels of EDA as an additional feature, which is significantly discriminative in the recognition of students' engagement. We use our alternative method to compute arousing and unarousing moments and compute the features *arousing_num*, *unarousing_num*, *arousing_ratio*, *arousing_normalized*, and *level_i* (with $i = 1, 2, 3, 4, 5$). These are thus novel features whose discriminative power is investigated for the first time in this work.

The feature *arousing_ratio* is the ratio between the number of arousing moments and the number of unarousing moments during a lecture. This feature models the impact of the number of moments in which a student shows a momentary interest towards the lecture compared to the number of moments in which a student was not showing any significant change in her physiological arousal.

We captured the role of the EDA amplitude in the momentary engagement features by considering separately the percentage of time a student has been in one of the engagement levels respect to the whole lecture, which we map into *level_i* feature (with $i = 1, 2, 3, 4, 5$).

Summary of the considered feature sets

As discussed above we considered in our study three main sets of features: features related to the emotional arousal of a student – indicated with F_{emo} – ; features related to the reaction to the teacher, quantified using physiological synchrony between students and teachers (F_{syn}); and features related to the momentary engagement of students (F_{mom}).

We further divided the latter set into two subsets: F_{mom_cain} and F_{mom_new}. The first contains the three features *unarousing_num_Cain*, *arousing_num_Cain*, and *arousing_ratio_Cain*, which rely on the computation of unarousing and arousing moments using the method by Cain et al. The second contains the remaining momentary engagement features, which are computed using our own method to identify unarousing and arousing moments.

Engagement recognition pipeline

In order to identify engaged and non-engaged students during the lecture we setup a standard binary classification pipeline, described below.

Labeling. We leveraged the answers to the USEI questionnaire to assign students for each lecture to either the *engaged* or the *non-engaged* class. To do so we reverse the first item of the questionnaire, as described in [214] and then compute a single *engagement score* from the five items of the questionnaire by computing (and rounding) the average of the answers as in [215]. We labelled as *engaged* students whose score is higher or equal to four (corresponding to "Agree") and as *non engaged* the others. This is because we assume that an engaged student would provide answers at the positive extremes of the scale, as also discussed in [215]. This labeling procedure leads to an imbalanced data set with 120 instances i.e., lectures, in the non-engaged class (which we considered as positive class) and 77 in the engaged class.

Classifiers. In order to perform the classification task we selected three well-known classifiers. In particular we used the Logistic Regression (LR) [65], Linear Discriminant Analysis (LDA) [65] and the Support Vector Machine (SVM) [224] with a linear kernel. We used the hyperparamenters of the classifiers set by default by the Python library we used for the implementation (i.e., scikit-learn[4]), except for the SVM for which we set C = 100 empirically. We did not perform an automatic optimization of the hyperparameters due to the limited amount of data at disposal for training the model. Indeed, the hyperparameter optimiza-

[4] https://scikit-learn.org/stable/

tion procedure would have required an additional validation set with consequent reduction of the number of training samples. To put the obtained classification results in context, we use Random Guess (RG) and Biased Random Guess (BRG) classifiers as baseline competitors. We compared the performance between the classifiers and the baselines using a paired t-test [225]. We tested the p-values against the threshold $\alpha < 0.05$ and report also the corrected threshold (obtained using the Bonferroni correction method) $\alpha_c = \alpha/n = 0.005$, where $n = 10$ in our case. Further, we report the Cohen's d effect size measure.

Metrics. While the ultimate goal of our work is to investigate whether it is possible to assess students' engagement using physiological data, the automatic engagement recognition is especially beneficial for non-engaged students. Those are indeed the students who need to be re-engaged and ameliorate their performance [24]. For this reason we considered the non-engaged class as the positive one. To evaluate the performance of the classifiers we use several performance metrics: *recall*, *precision*, *accuracy*, and *F1*, all defined according to [65]. We further consider the *F2* metric – which weights the recall higher than the precision – as defined in [226]. Lastly, to quantify the classification error we consider the *false discovery rate* (FDR) [65], which indicates how many of the students classified as non-engaged were actually engaged.

Feature Selection. We used the *Sequential Forward Floating Selection (SFFS)* algorithm [81] to identify the most relevant features. We use the recall as the target metric of the SFFS algorithm. This is because we aim at reducing the number of false negatives, i.e., the number of non-engaged students that are erroneously classified as engaged. A large number of false negatives would indeed make the system miss non-engaged students and thus prevent the understanding of the reasons of their lack of engagement as well as hamper the use of interventions to re-engage them.

Validation Procedure. We trained and tested the classifiers using a nested cross-validation approach [227]. This method iterates twice over the data, once for running the feature selection process (inner loop) and once for actually training and testing the classifiers (outer loop).

In the outer loop we created the folds so that the training and test sets in a single fold do not contain data of the same student. In particular, similarly to [67; 138], we divide the data into n groups (whereas n is the number of students in our data set, i.e., n=24), and each group contains the data of one student only. A number k of folds ($k \leq n$, k=10 in our case) is then created ensuring that the data of the same group does not appear in two different folds. The groups assigned to a fold are selected at random and the number of distinct groups in each fold is approximately the same, i.e., the folds are approximately balanced. This procedure – to which we refer to as the *leave one group out (LOGO)* – ensures that the classifiers are trained in a student-independent manner and, thus, that specific characteristics of individual students do not alter the validity of the final results and their generalizability to other student populations [228; 138; 67].

Once the training and test sets are defined in each fold, we scaled the features as recommended in the literature using the training data only (from different participants and lectures) [227]. Further, to ensure that the distribution of non-engaged and engaged students is balanced, we re-sample the training set using the *SMOTE* algorithm [83].

In the inner loop we use five-fold cross-validation to make the SFFS algorithm select the best features [81]. Since nested cross-validation is a computationally expensive procedure, we used five folds in the inner loop to speed up the execution of the SFFS algorithm. Once the inner loop ends, we used the selected best features to evaluate the performance of the classifiers in the outer loop as done in [67] and explained above.

To evaluate the impact of specific characteristics of the data of individual students on the overall classification error, we additionally performed experiments using also the *leave-one-student-out (LOSO)* validation approach [29]. At each iteration, the data of only one student is used for testing, while the data of all the other students is used for training.

For both LOGO and LOSO validation approaches, the final performance of each classifier is computed as the mean of the performance achieved by the classifier in each iteration.

3.4 Results and discussion

We report and discuss the results from the engagement recognition pipeline in Section 3.4.1. Then, to further reflect on the connection between EDA responses and the emotional engagement of students during lectures, we analyze the correlation between EDA features and the engagement score obtained from self-reports. Results from the correlation analysis are reported in Section 3.4.2.

3.4.1 Engagement recognition results

Table 3.2 shows the performance obtained using the considered classifiers trained using different set of features and applying the LOGO validation protocol de-

Feature set	Classifier	Recall	FDR	Precision	Accuracy	F1	F2
F_{all} + SFFS	RG	53	40	60	51	53	52
	BRG	56	34	66	51	54	53
	LDA	77	31	69	64*	65	70
	LR	75	32	68	63	64	68
	SVM	**81***,**	36	64	60	**65**	**72**
F_{best_7}	SVM	59	32	**67**	58	58	57
F_{best_4}		79*,**	41	59	59	63	70
F_{best_2}		**81***,**	35	65	60	**66**	**71**
F_{mom_cain}	SVM	35	52	48	42	36**	35**
F_{mom_new}		54	39	61	53	55	53
F_{mom}	SVM	53	41	59	58	50	50
F_{emo}		58	36	64	54	57	56
F_{syn}		52	36	64	51	53	51
F_{all}		59	41	59	50	53	55

Table 3.2. Summary of the performance obtained using different models (subset of features and classifiers). An asterisk (*) indicates significantly different performance (p<0.05) from the RG and Cohen's $|d| > 0.8$, two asterisks (**) indicate significantly different performance from the BRG and Cohen's $|d| > 0.8$. None of the carried tests showed a p-value $< \alpha_c$. In the F_{best_7} scenario we use all the 7 features extracted at least once from SFFS. In the F_{best_4} only the first 4 features and in F_{best_2} the best 2 features (level_5 and arousing_ratio). F_{mom} refers to the use of all the momentary features while F_{mom_cain} and F_{mom_new} refer respectively to the use of only the momentary features obtained by the Cain et al. procedure and the one obtained using the features we proposed. F_{emo} refers to the use of emotional arousal features while F_{syn} to only the use of synchrony features. F_{all} considers the use of all features together.

scribed in Section 3.3. We discuss below the main results obtained in the recognition of students' engagement.

Performance obtained using all features and SFFS for feature selection

We start by commenting on the performance obtained using all features (i.e., all features listed in Table 3.1) as input to the classifiers in combination with the SFFS algorithm for feature selection (F_{all}+SFFS).

The results in Table 3.2 show that all classifiers outperform both the RG and

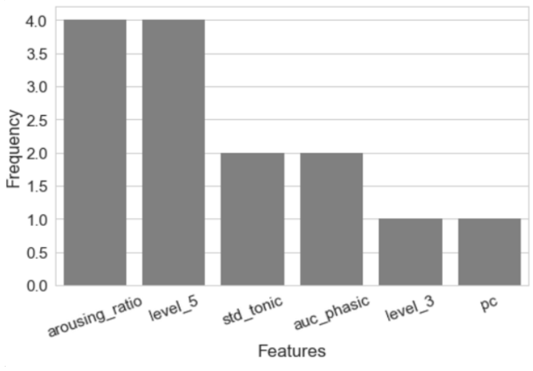


Figure 3.7. Frequency of occurrence of the features selected at least once during the feature selection process using the SFFS algorithm.

BRG baseline classifiers in terms of recall.

In particular, SVM achieves the best performance (tested with the paired t-test [225], p <0.05), with a recall of 81%, which is 25 and 28 percentage points higher than for RG and BRG. SVM achieves the best performance also in terms of F2 (72%), which is 20 and 19 percentage points higher than the corresponding performance of RG and BRG. The fact that other performance metrics (FDR, precision, accuracy, F1) present no or little improvement over the baseline classifiers might be due to the fact that we used the recall as target metric during the feature selection process. This makes the classifiers maximize the recall – and thus their ability to successfully identify non-engaged students.

From Table 3.2 we observe that SVM has a FDR of 36%. This means that if, on average, the classifier identifies 100 students as non-engaged, 36 of them were actually engaged. Even though this value of FDR may seem high, it still represents a reasonably good result for our particular scenario. Indeed, as also observed in [24], the cost paid in misclassifying engaged students as non-engaged is less relevant compared to missing the chance of identifying non-engaged students.

Since the SVM resulted to be the best performing classifier among those considered, we present below results obtained using SVM as reference classifier.

Dominant features selected during the feature selection procedure

Of the 30 features listed in Table 3.1 only seven are selected by the SVM at least once during the feature selection process. Figure 3.7 shows the frequency of occurrence of the selected features.

We can observe that the most frequently selected – i.e., the most discriminative – features are the *arousing_ratio* and the *level_5* (4 times) features. The other most frequently selected features are the *std_tonic* and the *auc_phasic* (2 times) as well as the *auc_eda*, the *level_3* and the *pc* (once) features.

The fact that the selected features belong to all the three main feature subsets we defined (F_{emo}, F_{syn}, and F_{mom}) suggests that all the components we considered to characterize students' emotional engagement – emotional arousal, physiological synchrony and momentary engagement – contribute to the detection of non-engaged students.

However, the momentary features display a stronger discriminative role in identifying non-engaged students. In particular, the two dominant ("best") features *arousing_ratio* and *level_5* capture information about the arousing moments of a student – which in turn we can consider as representing the number and intensity of the "highlights" experienced by the student during a lecture. This confirms educational theories that outline the importance of momentary engagement for ensuring the engagement of students with learning activities [87].

Most of the selected features contain information about the amplitude of the signal (e.g., *level_5*, *auc_phasic*, *auc_eda* and the *level_3*) rather than its temporal variation. The selection of the *pearson* feature, indicates that the physiological synchrony with the teacher contributes to the recognition of students' engagement, although not in an as determinant manner as the momentary engagement.

Performance obtained using only the best features

Building upon the observations reported above about dominant features we ran further experiments to verify the performance obtained by SVM if trained using: all the seven best features selected at least once by the SFFS algorithm (F_{best_7}); only the first four best features – *arousing_ratio*, *level_5*, *std_tonic* and *auc_phasic* – (F_{best_4}); and only the best two features – *arousing_ratio* and `level_5` – (F_{best_2}). We thereby use the LOGO protocol described in Section 3.3. The obtained results are reported in Table 3.2 and show that SVM achieves the highest recall (81%) if it is trained with only the two best features (*level_5* and *arousing_ratio*).

Instead, SVM achieves the highest precision if the seven best features are used as input, however at the cost of a significant loss in terms of recall (which decreases to 59%).

Further, we note that the use of features derived from the phasic and tonic components – in particular *auc_phasic* and *std_tonic* – does not lead to performance improvements in terms of recall.

Overall, these results show that using only two of the novel momentary features we proposed in combination with SVM allows us to achieve the best performance in terms of recall as well as a good balance between precision and recall.

Performance obtained using different subsets of momentary features

We further compare the performance obtained by SVM trained using only the new momentary features we proposed (F_{mom_new}) against the features extracted using the method proposed by Cain et al. [90] (F_{mom_cain}). Table 3.2 shows that using only the features in F_{mom_new} SVM achieves a recall that is 19 percentage points higher than the recall obtained using the features in F_{mom_cain}. Similarly, precision, F1 and F2 are significantly ($p < 0.05$) higher and FDR significantly lower when using F_{mom_new} instead of F_{mom_cain}.

We believe that the superiority of our features is due to the fact that our method for identifying arousing moments is more robust against noise and artifacts with respect to the algorithm proposed by Cain et al. [90].

Performance obtained using the individual feature subsets

To further investigate the discriminative power of different components of the emotional engagement, we ran experiments using the individual feature subsets F_{emo}, F_{syn}, and F_{mom} as input to the classifiers as well as all of them together (F_{all}) but without any feature selection process.

The corresponding results, reported in Table 3.2, show that the performance obtained using the single feature subsets are comparable. The combination of the features provides a little improvement (1 percentage point) in terms of recall.

A part from the recall, the emotional arousal features provide slightly better performance. We believe this is due to the fact that the phasic component of the EDA signal presents a significant correlation with engagement as we discuss in Section 3.4.2.

These results show that considering separate feature subsets – and thus separate components of the emotional engagement – separately leads to the same final performance. The combination of the subsets brings little improvement.

Furthermore, the lack of a feature selection process leads to low overall performance. Analysing the different combinations of feature subsets and the results of the feature selection process instead allows us to understand which features – and which components of the emotional engagement – are the most relevant and thus allows us to obtain significantly better overall performance.

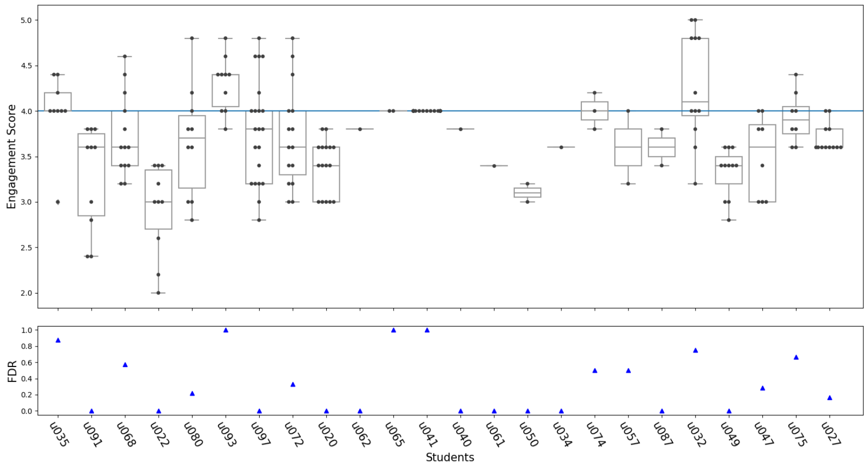


Figure 3.8. Students' engagement distribution during lectures and their correspondent FDR. The blue horizontal line represents the threshold used for dividing the students in the engaged and non-engaged classes. The students' username presented in this figure are randomly generated, to maintain the privacy of the participants.

Per-student classification error

Lastly, we analyzed the impact of the data from individual students on the classification error of the most successful model (i.e., SVM with F_{best_2}).

To this end, we trained and tested the model using the LOSO validation procedure described in Section 3.3. We obtained the following results: precision: 62 (STD=37), accuracy: 68 (STD=31), recall: 77 (STD=37), FDR: 33 (STD=37), F1: 68 (STD=36), F2: 71 (STD=36). While for the LOGO validation approach we obtained: precision: 65 (STD=26), accuracy: 60 (STD=21), recall: 81 (STD=25), FDR: 35 (STD=26), F1: 66 (STD=24), F2: 71 (STD=24).

The results we obtained using the LOSO validation procedure are similar to the results obtained using the same model with the group-students cross-validation approach. This similarity is not surprising since also the LOGO approach does not use the data of the same student for training and testing the model.

At the same time, the standard deviation of the performance obtained across different LOSO folds is significantly higher (12 percentage point higher in terms of recall) than the corresponding standard deviation obtained using the LOGO. We believe that this higher instability is due to the fact that with the LOSO validation procedure the size of the test sets varies significantly (i.e., between 1 to 21 data points) across folds.

Figure 3.8 shows the distribution of the engagement score of each student (upper plot) along with the corresponding FDR obtained with the LOSO procedure when the data of the same student is used in the test set (lower plot).

We observe that the model recognizes the engagement of 11 students without error (FDR = 0) and that, in general, it performs better for students who tend to be non-engaged.

The worst performance is obtained for students who reported several times an engagement score equal to four (e.g., u035, u065, u041), which is the threshold we use to discriminate engaged and non-engaged students.

We believe that students who often display an engagement score of four were not highly engaged and, thus, their physiological reaction is not appreciably different from that of the non-engaged students, leading them to be misclassified. However, more data from highly engaged students as well as possibly additional sources of information about students' engagement is needed to verify this hypothesis.

Lastly, Figure 3.8 also shows that while most of the students show a variable distribution of the engagement across lectures, many tend to be non-engaged. This tendency of students to be non-engaged during lectures is reported also by Wang et al. [34] and outlines the need for devising interventions to re-engage students.

3.4.2 Correlation of features with self-reported engagement

We computed the correlation between EDA features and the engagement score using a *per-student* approach similar to the one followed in [229]. Specifically, we first computed the mean – for each student and over all lectures – of both the engagement score and the value of each feature. We do not compute the correlation between engagement scores and features in each lecture because the data for the same student is likely to be correlated across different lectures [230]. The samples would thus not be independent, hampering the use of standard correlation measures. We then explored the correlation between the resulting values using the Spearman's rank correlation, given the non-normal distribution of the variables [231].

We observe that the percentage of time a student spent in the "very engaged" level (`level_5`,($r = 0.49$, $p < 0.05$)) and the ratio between arousing and unarousing moments (`arousing_ratio`,($r = 0.44$, $p < 0.05$) are positively correlated with the emotional engagement. This suggests that the physiological response of an engaged student tends to show more frequently high levels of electrodermal activity and that more engaged students tend to experience more arousing moments ("highlights") during lectures. This in turn confirms the relevant role of these momentary feature as a proxy of engagement.

Lastly, we observe that most of the features extracted from the phasic component of the EDA signal are also positively correlated with engagement: `auc_phasic` (r = 0.47, $p < 0.05$), `n_p_phasic` (r = 0.41, $p < 0.05$), `avg_phasic` (r = 0.55, p < 0.05). This underlines the fact that the phasic component models the response to external stimuli and consequently is connected to engagement [67].

3.4.3 Summary of the main findings

In the study presented in this chapter, we demonstrate that it is feasible to automatically recognize students' engagement during lectures using EDA data. Our main findings are:

- The SVM classifier achieved the highest performance in terms of recall (81%), which is 25 and 28 higher than the one achieved respectively by RG and BRG.

- Only seven features (F_{best_7}), among the 30 used, were selected at least once in the features selection procedure. The selected features belong to the three considered subsets (F_{emo},F_{syn},F_{mom}), indicating the contributions of all the components (i.e., emotional arousal, physiological synchrony and momentary engagement) in the engagement assessment.

- The momentary features resulted to be the most discriminative. Specifically, the best overall performance are obtained when using two of the momentary features we proposed i.e., *arousing_ratio* and *level_5*, as input to the SVM classifier. This result confirms the relevance of momentary engagement for the assessment of students' emotional engagement [87] as well as the strength of the new features we proposed.

- Correlation results indicate that the higher the number of "highlights" experienced by the students during the lecture, the higher is their perceived engagement during the lecture.

Overall, we believe that the results we obtained can be generalized to other populations of students. Our data set contains multiple data traces from multiple students with a variable data distribution, as observable from Figure 3.8. The plot shows that the number of data points across students varies (Min = 1, Max = 21, Mean = 8.21, STD = 5.81) and that the data distribution is quite spread.

3.5 Implications

One practical implication from our results is the possibility of using EDA sensor and wearable devices to monitor students' engagement during actual lectures. Therefore, researchers and practitioners aiming to design engagement-aware systems for monitoring students in classroom, could rely on our approach.

In particular, wristbands equipped with EDA sensors could be used as input device for the system. Momentary features could be extracted to quantify the physiological activation expression of engagement. These features could be used alone, or in combination with other source of information, as input to the engagement model. For instance, the occurrence of laughter episodes automatically recognized, for example using the method we present in Chapter 4, can be combined with the momentary features to enrich the engagement representation.

Information about the inferred engagement can be then used by the system to deliver appropriate interventions to the students or make the teachers aware of the students' perceived engagement.

We envision a system that provides insightful visualizations, accessible, for instance, through a dashboard at the end of each lecture. The inferred engagement could be paired together with lecture-specific information, such as teaching material, length of the lecture, questions asked. These additional information could be automatically gathered or manually input by the student or the teacher to allow them to reflect on the learning experience over one or multiple lectures and courses.

Since the emotional engagement is a predictor of learning outcome [57], reaching awareness on a recurring lack of engagement has the potential to motivate students to change their behavior – or to drop a specific course in favor of another. This is in line with the trend towards a *quantified-self* in education, which has been recently explored in the literature [212; 232; 233] and that we also discussed in [234].

On the other side, teachers can benefit from feedback about students' engagement provided after individual lectures. For instance, teachers can observe whether the same students are non-engaged across lectures and try to re-engage them. Similarly, teachers can test different teaching methods and evaluate their effect on students' engagement.

While we do not argue that this feedback alone can determine whether a specific method is more effective than others, we do believe in the benefit of providing feedback to teachers at the end of every lecture instead of just once or few times per term. The importance of providing more and more diverse feedback to teachers is also outlined by a recent, large-scale project aimed at providing teachers with "*information that they can trust from measures that are fair and reliable*" [235].

While students can profit from feedback gained analyzing their data only, teachers need to access students' physiological data to obtain useful feedback. The release of such data, even in anonymous form, can however have negative impacts.

For instance, teachers may – consciously or unconsciously – de-anonymize the data, which may in torn lead to a penalization of the non-engaged students. Adequate procedures to preserve the privacy of the students must thus be put in place. Moreover, teachers should be advised to rely on different feedback measures (e.g., in class-tests, questionnaires) to obtain a comprehensive picture of students' engagement.

3.6 Limitations

While our results show that it is feasible to use EDA signals to discriminate between engaged and non-engaged students during lectures, further research is needed to overcome some of the current limitations of our work.

First, a further improvement of the performance is needed in real settings. In particular, reducing the FDR from the current 36% to a lower value is desirable to prevent engagement-aware systems to send inappropriate interventions or to provide students with wrong information about their engagement during lectures. Using additional engagement cues, e.g., gestures, body moments, laughs, could help improving the engagement recognition performance.

Second, collecting ground-truth data at a higher granularity and using different methods may open up new possibilities for the engagement recognition analysis. In this study we collected ground-truth data only at the end of each lecture, where a lecture lasted on average 42 minutes. An interesting possibility for future work is to embed short questionnaires about engagement in live interaction platforms such as ASQ [236].

Irrespectively of the method used, though, the collection of responses from students *during* lectures is inherently critical, because it interferes with the normal flow of the lecture and it is a cause of distraction itself. This remains an open challenge that should be addressed in future work.

3.7 Summary of Chapter 3

In this chapter we presented our findings about the recognition of students' engagement during lectures using EDA data gathered from wearable devices. Specifically, we identified three components of students' emotional engagement in the educational research literature: emotional arousal, reaction to the teacher and momentary engagement. We then derived a set of EDA features that could represent these components.

To evaluate our approach we collected the *SEED* data set, which, at the time it was collected, was the largest and more heterogeneous data set containing physiological data of students during actual lectures.

We ran an extensive analysis and explored the role of different features and classifiers. We observed that the best overall performance are provided by the SVM classifier used in combination with two of the momentary features we proposed (*arousing_ratio* and *level_5*).

This confirms the importance of momentary engagement in the recognition of students' engagement [87] as well as the strength of the new features we propose. Results of this study have been published in [A].

Overall, our findings open up novel possibilities to design engagement-aware systems for supporting students during learning activities. Feedback strategies could be developed on top of the engagement recognition engine to, e.g., allow students to self-reflect on their engagement (or lack thereof) and teachers to test adequate strategies to re-engage non-engaged students.

Chapter 4

Recognition of Laughter Episodes From Physiological Signals and Body Movement Data

Laughter is a multi-modal behavioural expression characterized by the combination of several body reactions such as facial expressions, vocal tone, movement, respiration and physiological activation [173]. Laughter is an universal behavioral sign of positive emotions, it is known for reducing boredom and boosting engagement [237]. Laughter plays also a key role in communication giving positive feedback to the interlocutor and strengthening social bonds, as discussed in [62].

In activities that involve social interactions such as in team projects, lectures, meetings, presentations and breaks, laughter episodes often occur [238; 239; 178].

For instance, researchers have investigated the role of humor and laughter in the classroom [238; 239] and observed that they are effective "tools" for teachers to gain and maintain students' attention and to create a positive climate in the classroom [238; 239]. Humor and laughter have the potential of reducing tension and stress in the classroom, facilitating the learning experience and creating a positive relation among students and teachers [238; 239].

The role of laughter in business meetings has been also extensively investigated in [178] in terms of interaction phenomenon. For instance, authors of [178] observed that shared laughter is associated with task accomplishment, and when strategically invited by the managers, shared laughs enable to create a relaxed work environment [178].

Given the occurrence of laughs during work and learning activities, and their positive impact on students and workers' mood, we believe that an engagement-aware system aiming to support people in their daily activities should be able to detect laughs.

In the study presented in this chapter, we propose and evaluate a novel method to recognize laughter episodes. Specifically, we propose to use a combination of *physiological signals* – to quantify a person's electrodermal and cardiovascular arousal [176] – and *body movements* – such as vibrations of the shoulders or the trunk caused by laughter-induced exhalations [174] collected using wristbands.

Most of the existing work in laughter recognition, consider laughter as an audio-visual event and thus rely on the analysis of audio and video data only [62; 240]. The reliability of these methods is usually high, with typical accuracy values ranging between 70% and 90% [173]. However, using cameras and microphones to obtain the necessary video and audio data is often unfeasible, especially in noisy environments or under poor light conditions, or when multiple people interact with each other, e.g., during meetings or lectures [173; 241]. To overcome these limitations, we rely on physiological and body movement data collected using wristbands. Indeed, as also discussed in the study presented in Chapter 3, wristbands can be easily used for collecting data from multiple people while guaranteeing the privacy of who do not want to be monitored.

We distinguish between laughter and non-laughter episodes and test different combination of sensors and classifiers. Further, we test the robustness of our method against *confounding variables* such as cognitive load and clapping hands, which could generate similar physiological and movement responses to laughter and consequently generate miss-classification errors.

Our results show that using a combination of the features extracted from the physiological signals and the body movement data, as input to a Support Vector Machine, we can achieve an accuracy of 81%. Further, we demonstrate that the signatures left by laughter episodes on physiological and body-movement data differ significantly from those caused by slightly intense motions or cognitive load tasks.

To evaluate our approach we collected the *USI_Laughs* data set and made it available to the research community.

The study presented in this chapter targets the first research question of this thesis (*RQ1: How can features representing behavioral expressions of engagement can be derived from physiological and movement data?*). The results of the presented work have been published as conference paper [B] at the PervasiveHealth conference (May 2019). Part of the text written in this chapter is reported from the paper [B].

The reminder part of this chapter is structured as follows. We provide an overview of the existing literature in laughter recognition in Section 4.1. Section 4.2 describes the data collection of the *USI_Laughs* data set. Section 4.3 and Section 4.4 provide an overview of the data analysis method we used and a discussion of the obtained results. In Section 4.5 we discuss the implications of our approach. In Section 4.6 we discuss the limitations of the presented method. We conclude with a brief summary of the chapter.

4.1 Related work on laughter recognition

In this section we provide an overview of existing literature on laughter recognition. A description of laughter and its characterization is presented in Section 2.3.4.

Several authors use acoustic features to distinguish laughter from speech. Performance, in terms of correct classification rate, obtained using acoustic processing only, ranges from 70% to 90% [173]. Recently, Hagerer et al. [37] deployed a laughter detection system on smartphone and wearable using audio data only.

More recently, several authors explored the use of other sensory channels in addition to the audio source. The authors of [38; 39], for instance, combined acoustic signals with visual-facial expressions. Reuderink et al. [39] integrated the results of separate audio and visual models using a decision-level fusion approach to distinguish between laughter and non-laughter segments.

Petridis et al. [38] extensively analyzed the role of features derived from audio, video (from which they extracted features for characterizing facial expression and head pose) and their combination through features-fusion technique [242]. They also confirm that, on average, considering laughter as a multimodal behaviour rather than uni-modal, increases the performance of the model, especially for female subjects [38].

Our approach also exploits the multi-modal nature of laughter. However, we focus on the role of physiological and body movement characteristics, which are less investigated in comparison to the visual and vocal expressions.

Few approaches integrate body movements in the recognition of laughter [175; 40; 243]. Niewiadomski et al. [175] investigate the role of full-body movement in laughter recognition during social interactions. They use a motion capture system to identify and extract 13 features that could characterize typical body movements during laughter episodes, which they provide as input to several classifiers. Their method achieves a F1 score of 74%. Even though this performance may appear modest, the authors demonstrate that the use of body movement could be a valid alternative to audio and facial expression modalities, which are not always available in real-world scenarios [175].

Urbain et al. [40] investigated laugh-induced movements and their combination with respiration sensors and audio. The authors positioned polystyrene spheres on the participants' shoulders to track their movements and propose a Body Laughter Index (BLI) to capture those movements [40]. Despite the robustness of their method, as the authors also report, the presented system is too invasive to be used in real settings.

Even though the results presented in the above mentioned approaches are very promising, they all suffer from the limitations related to the sensors involved. In particular, performance derived from video and audio sources are highly dependent on the environmental conditions in which the system is deployed, for example in noisy or smoky places or in presence of poor light conditions [173; 241]. They may also raise significant privacy concerns, especially if used in public places.

To overcome these limitations, other methods have been proposed to detect laughter [244; 245]. For example, Consentino et al. [244] place a set of inertial measurement units (IMUs) and electromyography sensors (EMG) on the subjects' torso to measure the moments and the muscles activation during laughter.

Very few approaches consider the activation of the Sympathetic Nervous System (SNS), in terms of cardiovascular and electrodermal arousal in laughter episodes and most of them are tested in laboratory settings [173] using bulky and invasive devices.

Tatsumi et al. [246] collected physiological data from 10 participants using EDA, electrocardiogram (ECG), facial electromyogram (FEMG) and movement of the diaphragm placing electrodes on participants' body and face for detecting "hidden laughter", i.e., when participants almost laugh but without showing their expressions [246]. The method proposed by the authors achieves an accuracy of 85%.

Although the placement of several sensors could provide more precise measurements, it is impracticable in real-world settings, furthermore it is invasive and not ecological for the participants and can consequently alter their emotional state [173].

As stated in the literature review on quantitative laughter detection from Cosentino et al. [173], and we also demonstrated in this work, new advances in wearable technologies could solve the above mentioned issues and allow laughter recognition to be integrated in real systems.

4.2 Data collection of the USI_Laughs data set

Most of the data sets publicly available for laughter recognition do not contain physiological data. Only the *RECOLA* data set presented in [247], contains EDA data and ECG data collected with an invasive device (Biopac MP36[1]). However this data set does not contain movement data about body movement, an important modality to investigate for describing laughter episodes [173] as also confirmed by our results. A recent overview of the existing data sets suitable for laughter recognition is provided in [248]. The lack of an appropriate data set for investigating our research question, motivated us to collect a new data set. In this section we provide details about the participants, the equipment we used and the procedure we followed to collect the *USI_Laughs* data set.

Participants

We recruited 34 participants (28 males and 6 females) of age between 22 and 37 (Mean = 26.70, SD = 4.04). We allowed the participants to decide whether to take part to the experiment alone or with another participant. Given laughter being a contagious behaviour [172], we considered that the possibility of participating to the experiment with another person could increase the number of laughter episodes. This procedure allowed us to collect data from 16 pairs and two individuals.

Collected data

Sensor data. We gathered physiological data – blood volume pulse (BVP) and electrodermal activity (EDA) – and body movement data – accelerometer (ACC) – using the E4 wristband [58].

Ground-truth data. It is referred to the type of expressions and actions the participants performed during the experiment. We asked three external observers to annotate the video recordings of the participants, as common practice in the literature [62]. To collect the video recordings we used a GoPro camera equipped with a microphone.

Data collection procedure

We conducted the study in a laboratory setting. Despite the constraints of laboratory settings, this approach ensures a detailed analysis of the phenomenon

[1] `https://www.biopac.com/product/mp36r-system-54/`

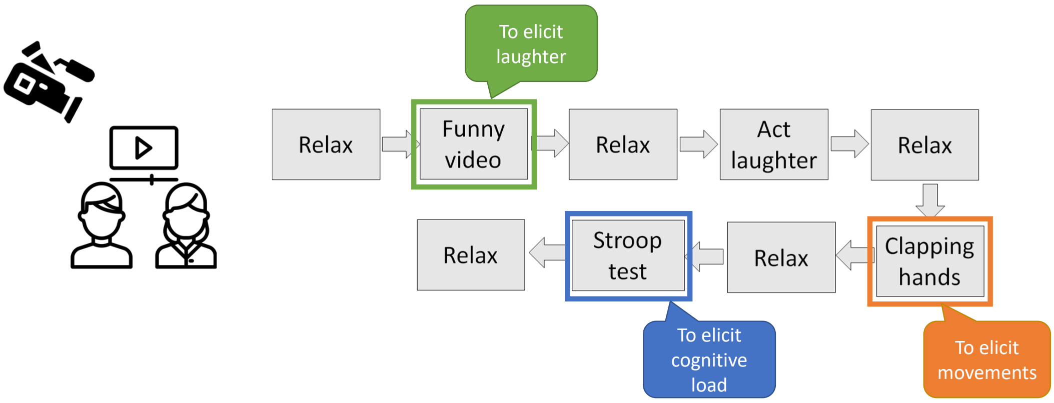


Figure 4.1. Experimental protocol we designed for collecting the USI_Laughs data set.

of interest, the replicability of the procedure [10] and it is suitable for initial investigation of new methods. Before starting the experiment, we asked the participants to wear an Empatica E4 to each wrist. We did this step to guarantee that even if one of the devices was malfunctioning, or the data was corrupted, we could still use the data from the other device. Before starting the experiment we provided a general description of the study and all the participants signed an informed consent agreement. The study procedures were approved prior to the start of the study by the faculty of Informatics ethics delegate. As suggested in the literature we did not disguise the actual purpose of the study because this should guarantee a spontaneous reaction of the participants [249].

The participants were asked to follow the instructions presented in a video and react as more spontaneously as possible. We put only two constraints: limit the movements of the hands – to reduce the presence of motion artifacts in the signals [216] – and, if they participate with a friend, to avoid to talk to each other – to exclude that the physiological response was due to a social interaction rather than to a laughter episode.

Once all the instructions were given, the participants were left alone in the room to not influence their behaviour. To elicit natural laughter we created a ten minutes video concatenating ten funny video (average duration = 69 seconds, Min = 7 seconds, Max = 151 seconds) from YouTube similarly to [249]. We alternate tasks with moments where the participants were asked to relax and breathe normally, we refer to them as *relax*. If the participants were two they performed the tasks one after the other. Figure 4.1 describes the protocol we followed. In particular, participants were asked to relax for 60 seconds, after that they watched the funny videos and then relax again. We asked them to perform an acted laughter – to analyze the difference between real and acted laughter, however this analysis is not part of the scope of this contribution – . After relaxing for other 30 seconds participants clapped their hands three times – to analyze the difference between the body movements generated by laughter episodes and the ones from slightly intense motions –. Additional 30 seconds of relax then they performed the Stroop test – to compare the activation of the SNS due to laughter episodes and the one due to cognitive load tasks –, designed and widely used for cognitive load inference [250].

The experiment ended with final 60 seconds of relax. The total duration of the experiment was 16 minutes.

We provided sweets as compensation for the participation in the study.

4.3 Data analysis

In this section we describe the procedure we followed for annotating and cleaning the data, pre-process sensors data, extract the features and classify the episodes.

Data annotation and cleaning

Three external annotators were asked to watch the video recordings of the participants and annotate laughter and non-laughter episodes as explained below.

Laughter episodes. Several definitions of laughter exist and no standard rules have been defined for the annotation, as also discussed in [62]. To annotate the laughter episodes we rely on the broad definition provided in [251] which states: *"Laughter is defined as being any perceptibly audible expression that an ordinary person would characterize as laughter if heard under everyday circumstances"* used also in [62].

One annotator labeled the laughter episodes of each participants using video and audio data and the ANVIL video annotation tool [64]. From this step we collected 612 data segments. These segments were then annotated by other two raters. We asked the raters to use the following labels: *laughter*; *smile*; *not-sure* and *none*. For annotating laughter episodes, raters were requested to consider the definition presented before and suggested in [251]. They should have reported *smile* if no sound or movement was identified, *not-sure* if the rater was not sure of the label and *none* if the segments could not be identified neither as laughter nor smile.

Then two sub-labels for characterizing laughter episodes were also provided: *sound* and *intensity*. In the sound category the raters were requested to identify whether the laughter was *voiced* or *unvoiced* based on the definition provided in [38]. We asked the annotators to rate the perceived intensity of the laughter as *low*, *medium* or *high* as suggested in [252]. There is no a precise definition of laughter's intensity given the fact that the intensity dimension is naturally used by people to describe laughter, as discussed in [253]. We obtained the final labels and sub-labels for each episode using majority voting among the three raters similarly to [62]. We discarded a segment if all the three raters provided different labels. Out of the 612 segments 27 were discarded from this procedure. In this study we focus on the medium and high intensity laughter episodes. Thus we consider 316 laughter episodes (out of 585) of average duration of 3.18 seconds (SD = 1.70 seconds) labeled as medium or high intensity.

Non-laughter episodes. In order to create a balanced data set we extracted 316 non-overlapping non-laughter segments using the following procedure. We considered the section of the experiment between the first and second relax periods, from which we removed the segmented laughter episodes and an additional segment of three seconds before and after the episode, to take into account the onset and offset of the laugh [39]. Moreover we discarded segments that one of the raters identified as either movements or talking episodes, to reduce the presence of artifacts in the signals. For each laughter episode we selected a random non-laughter segment of the same duration of the laughter segment.

Subset of the data set. Before proceeding with further analysis we considered episodes of duration of at least three seconds. Indeed, we considered that an EDA response's rise time lasts 1-3 seconds [35], so we considered three seconds as minimum duration to be able to capture meaningful physiological information. We also observe that to a lower intensity corresponds a shorter duration of the episode, confirming what stated in [252]. This suggests that laughter recognition using physiological signals is more suitable for higher intensity emotional reactions, the short duration of the low intensity laughs (Mean = 1.52 sec) could not allow a detectable change in the physiological parameters. The final data set we use for this analysis consists of 189 laughter and 189 non-laughter episodes from 31 subjects, the average number of laughter episodes per subject is 6 (Min = 1, Max = 18) of average duration of 4.19 sec (Min = 3 sec, Max = 13 sec).

Pre-processing of sensor data

We used sensor-specific pre-processing techniques suggested in the literature [67; 254; 165; 255] to remove noise and decompose the signals. We used the data from the non-dominant hand of the subject and pre-process the signals collected

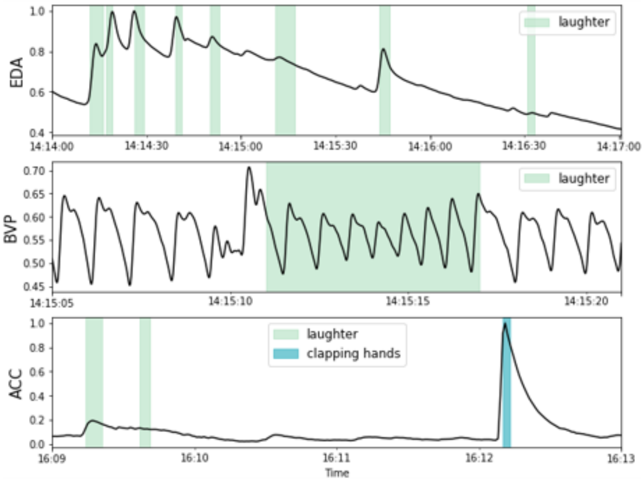


Figure 4.2. Examples of EDA BVP and ACC signals. The laughter episodes are highlighted in the corresponding segments.

over the whole experiments. We normalized the signals to allow a direct comparison among different individuals' physiological responses. We normalized the signals per-participant using the traces collected during the whole experimental procedure. All the signals are normalized using the min-max normalization which brings the signal's amplitude in a range between zero and one [72].

EDA data. We visually inspected the EDA signals, and observed that the data of four participants was significant affected by artifacts, probably due to the wrong placement of the wristband (too tight or too lose). For those subjects we use the data from the dominant hand. We filtered the EDA signal using first order Butterworth low-pass filter with a cut-off frequency of 0.4 Hz similar to [254] to remove noisy high frequency fluctuations. We decomposed the signal using the cvxEDA approach [71]. In this work we used only the normalized mixed-EDA and phasic signals, since the tonic component is not connected to short responses to stimuli and it is thus unsuitable to quantify laughs. More details about the EDA and its characteristics are presented in Section 2.3.3.

BVP data. We filtered the BVP signal with a first order Butterworth FIR filter with a cut off frequency of 5 Hz, similar to [165]. Before extracting the features in the laughter and non-laughter segments we removed the segments that do not confirmed with three rules presented in [255]: the heart rate (HR) should be in the range of 40-180 beats per minute (bpm), the gap between adjacent peaks should be maximum of three seconds and the ratio between the maximum and minimum of the peaks interval should be less than 2.2 seconds [255]. This procedure lead to the elimination of three instances from the non-laughter and 12 instances from the laughter class. More details about the BVP signal and its characteristics are presented in Section 2.3.3.

ACC data. The E4 measures the gravitational force (g) applied to the three axis and limit the range between ± 2g. Changes in the acceleration signal can be observed during motions. To quantify the motion in the laughter and non-laughter segments, we calculated the moving average of the ACC data with the method suggested by Empatica[2].

Features

We extracted 56 features from EDA, BVP and ACC in the laughter and non-laughter segments. Table 5.6 presents an overview of the features.

Examples of these signals and corresponding laughter episodes are presented in Figure 4.2. We can observe that in correspondence of laughs the EDA presents peaks while the BVP pulses' amplitude and the peaks' distance is significantly reduced, these are all expressions of the activation of the SNS.

EDA features. We extracted 22 features from the EDA data proposed in literature [67; 74]. We quantified the skin conductance responses (SCRs) [35] using the number of peaks and the peaks' amplitude[3]. To characterize the changes in the signal we calculated the dynamic change [74], the slope of the signals (estimated with linear regression as in [10]), the absolute value of the slope, the mean and the standard deviation of the first derivative.

BVP features. We extracted 23 features from the BVP signal. We considered time-domain statistical features, heart rate (HR) and heart rate variability (HRV) features, moreover we computed features to describe the BVP pulses. We characterized the BVP pulses using the mean and the standard deviation of the pulses amplitude – difference between the pulse's height and its preceding valley –. To an increment of the amplitude of the BVP signal, corresponds a decrease of the sympathetic arousal [256]. We also calculate the mean and standard deviation of the pulses' length – time interval between two consecutive valleys –. As suggested in [163] we consider the peak-to-peak amplitude variation, such as the difference between the highest and the smallest peaks. This feature should measure the activation of the SNS, in particular as the amplitude decreases, the arousal increases [163]. We computed the HR as $60/\overline{IBI}$, $\overline{IBI}$ is the mean of the

[2] `https://bit.ly/2FndpPE`

[3] We considered an amplitude of the peak of at least 0.01 in the normalized signals as in [67]. If no peak was found in the segment both the number of peaks and the amplitude was set to 0

Sensory channel	Features extracted
EDA (mix-EDA + Phasic) **(22 features)** **From: [67; 68; 73; 74; 75]**	**Time-domain statistical features:** minimum, maximum, mean, standard deviation, difference between maximum and minimum value or dynamic change, slope, absolute value of the slope, mean and standard deviation of the first derivative; **SCR features:** number of peaks, peaks' amplitude;
BVP **(23 features)** **From: [76; 77; 78; 79]**	**Time-domain statistical features:** minimum, maximum, mean, standard deviation, dynamic change, slope, area under the curve, number of peaks, ratio between number of peaks and length of the segment, mean and standard deviation of the first and second derivative, difference between the highest and smallest peak; **Heart rate (HR);** **HRV statistical features:** mean of all NN, standard deviation of all NN (SDNN), standard deviation of differences between NN (SDSD), the square root of the mean of the sum of the squares of differences between NN (RMSSD); **BVP pulse features:** mean and standard deviation of pulses' amplitude and pulses' length;
ACC **(11 features)** **From: [29]**	**Time-domain statistical features:** minimum, maximum, mean, standard deviation, dynamic change, slope, absolute value of the slope, mean and standard deviation of the first and second derivative;

Table 4.1. Overview of the 56 features extracted from the EDA, BVP and ACC signals. NN stands for the distance between BVP adjacent peaks.

interbeat interval (IBI) in the segment [76]. To quantify a short HRV we calculate the distance between adjacent peaks (NN) and then computed the mean and standard deviation of the NN (SDNN), the SDSD and the RMSSD [164][4].

ACC features. We calculated 11 statistical features from the ACC signal.

Laughter recognition pipeline

In order to distinguish between laughter and non-laughter episodes we use a binary classification pipeline described below.

Labeling. We used the 189 identified laughter episodes by the external raters and the corresponding 189 non-laughter episodes randomly extracted as explained in Section 4.3.

[4] If the number of peaks detected in the window was less than four, we set the SDSD and the RMSSD to zero since no variation between difference in adjacent NN could be observed in those cases.

Classifiers. We tested five well-known classifiers. We used linear classifiers – Support Vector Machine with linear kernel and C =100 (SVM-Linear) and Logistic Regression (LR) – , a non-linear classifier – SVM with radial basic function (SVM-RBF) – and an ensemble learning method – Random Forest (FR) – [65]. We consider a Biased Random Guess (BRG) as baseline.

Metrics. To evaluate the performance of the classifiers we considered the following metrics: accuracy, precision, recall and F1 score [65].

Feature selection. We used a filter features selection (FS) approach. Filter FS algorithms select features independently from the specific learning model and usually select the features with the strongest relationship with the output variable [257]. In this work we used the non-parametric Kolmogorov-Smirnov (KS) test which has been successfully used to extract relevant features in an emotion recognition from speech [258]. The aim of the KS test is to reject the null-hypothesis of identical distribution [258]. We applied the KS test to each feature to discard the non-relevant features e.g. the features which present a similar distribution in the two classes. We kept only the features which present a significant difference ($p < 0.05$) in the two classes as in [258].

Validation procedure. To evaluate the performance of the models, we use the *leave-one-subject-out* (LOSO) validation procedure. In this way, laughter and non-laughter episodes derived from the physiological signals of a single subject are not contained in the train and test sets simultaneously. As recommended in literature [227], in the training phase we scaled the features using a standard scaler which removes the mean and scale the features to unit variance[5]. To compare the results with the baseline we used a paired t-test and set the significance level to 0.05 as in [38]. We also considered the corrected threshold for the p-value (α_c) using the Bonferroni correction and report the Cohen's d effect size, as done for the analysis presented in Chapter 3.

4.4 Results and discussion

To investigate the feasibility of recognizing laughter episodes using physiological and movement data. In particular, we analyzed the role of each sensor separately (uni-modal approach) and their combination (multi-modal approach). We analyze which features present significant differences between the laughter and

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Sensors	**SVM-Linear**				LR				RF				SVM-RBF				BRG			
	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
EDA	$67^{*,**}$	$71^{*,**}$	52	56^{*}	67^{*}	$71^{*,**}$	58^{*}	59^{*}	58^{*}	52	46	46	$66^{*,**}$	$71^{*,**}$	50	54^{*}	44	40	42	40
BVP	$66^{*,**}$	61^{*}	64^{*}	57^{*}	$66^{*,**}$	61^{*}	62^{*}	57^{*}	58^{*}	53	60	54	63^{*}	58^{*}	60	54	43	38	42	39
ACC	$75^{*,**}$	$77^{*,**}$	64^{*}	$68^{*,**}$	$75^{*,**}$	$75^{*,**}$	66^{*}	$68^{*,**}$	$67^{*,**}$	67^{*}	54	57	$74^{*,**}$	$71^{*,**}$	59	61^{*}	48	41	45	48

Table 4.2. Performance of all the classifiers for the uni-modal approach. The reported values refers to the mean of the metrics (Accuracy (A), Precision (P). Recall (R) and F1 score (F1)) obtained across the 31 iterations of the LOSO approach. An asterisk (*) identifies the significant difference ($p < 0.05$) from the baseline (BRG). Two asterisks (**) identify the significant difference using the adjusted p-value ($p < 0.016$) from the baseline and a large effect size (Cohen's $d > 0.8$).

non-laughter classes. Lastly, we considered how confounding variables, specifically, movement generated by clapping hands, and cognitive load could affect the laughter recognition task. In this section we report and discuss the obtained results.

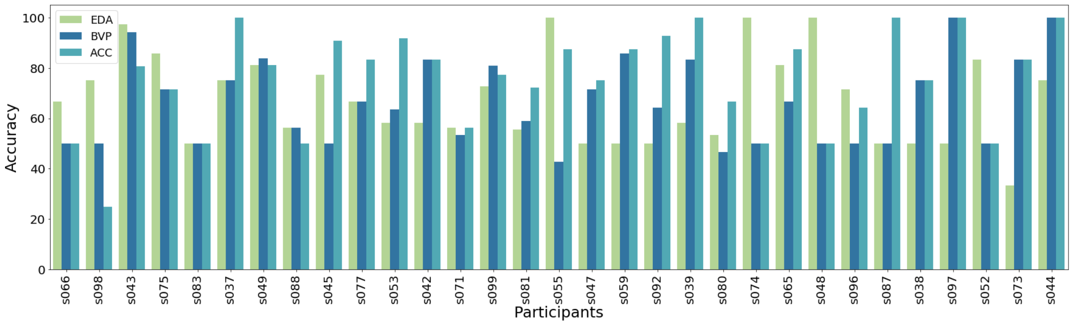


Figure 4.3. Classification accuracy per subject of the SVM-Linear using single sensors i.e., EDA, BVP and ACC.

Uni-modal laughter recognition

In this set of results, only the features extracted from one of the sensors are used as input to the classifiers.

Table 4.2 shows the results obtained for each sensor. We observe that in general the linear classifiers perform better than the others and all their metrics are significantly different and higher than the baseline.

The combination of the SVM-Linear classifier and the features extracted by the ACC lead to the best performing model, with an accuracy of 75%, precision of 77%, recall of 64% and F1 of 68%, outperforming the baseline by respectively

Sensors	**SVM-Linear**				LR				RF				SVM-RBF				BRG			
	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1
EDA + BVP	$69^{*,**}$	$72^{*,**}$	66^{*}	64^{*}	$68^{*,**}$	$73^{*,**}$	65	62^{*}	$69^{*,**}$	62^{*}	65	61^{*}	$68^{*,**}$	67^{*}	66^{*}	63^{*}	48	44	49	46
EDA + ACC	$78^{*,**}$	$84^{*,**}$	72^{*}	$74^{*,**}$	$78^{*,**}$	$78^{*,**}$	71^{*}	72^{*}	$76^{*,**}$	$75^{*,**}$	62	68^{*}	$74^{*,**}$	$75^{*,**}$	62	65	50	51	54	52
BVP + ACC	$77^{*,**}$	$77^{*,**}$	72	71^{*}	$76^{*,**}$	75^{*}	71	$69^{*,**}$	$72^{*,**}$	64^{*}	62	61	70^{*}	62^{*}	60	58	57	55	58	55
ALL	$81^{*,**}$	$82^{*,**}$	$81^{*,**}$	$80^{*,**}$	$79^{*,**}$	$84^{*,**}$	$78^{*,**}$	$77^{*,**}$	$66^{*,**}$	$59^{*,**}$	54	$55^{*,**}$	$73^{*,**}$	$69^{*,**}$	67	$65^{*,**}$	50	47	53	49

Table 4.3. Performance of all the classifiers for the multimodal approach. An asterisk (*) identifies the significant difference ($p < 0.05$) from the baseline (BRG). Two asterisks (**) identify the significant difference using the adjusted p-value ($p < 0.012$) from the baseline and a large effect size (Cohen's $d > 0.8$).

27,36, 19 and 20 percentage points. This underlies the importance of considering body movements in the automatic recognition of laughter [173].

When used alone, the BVP and the EDA features perform better than the baseline but much worse than when the ACC features are used. We believe this could be due to two facts, either the laughter episode was not intense enough to generate a strong physiological reaction that could be captured by the EDA or BVP sensors or some scenes in the videos attracted the attention of the user causing the SNS activation while not eliciting a laughter as discussed also in [176].

We notice a large variability of the performance, in terms of accuracy, when using a single sensor. This variability is reflected not only on the single participant used in the test set, but also on the type of sensor used. For instance, using the ACC only, the accuracy ranges from 25% to 100%. Figure 4.3, shows the classification accuracy per subject obtained using the SVM-Linear and the single sensors. We can observe that some participants obtained roughly similar results with all three sensors (i.d. s083, s049, s099), while others achieve much higher results when using one or two sensors. For example the accuracy obtained when using data from s097, s073, s044 for testing, is much higher with the BVP and ACC rather than with the EDA, while s074, s048 and s052 performs significantly better when using only the EDA. We believe this happens because laughter is not uniquely expressed and people's physiological reaction due to emotions often differ.

However, we notice that groups of users present similar performance using specific sensors. Based on that, we hypothesize that building models which consider the similarity among users could ameliorate the performance, more data is needed to confirm this assumption.

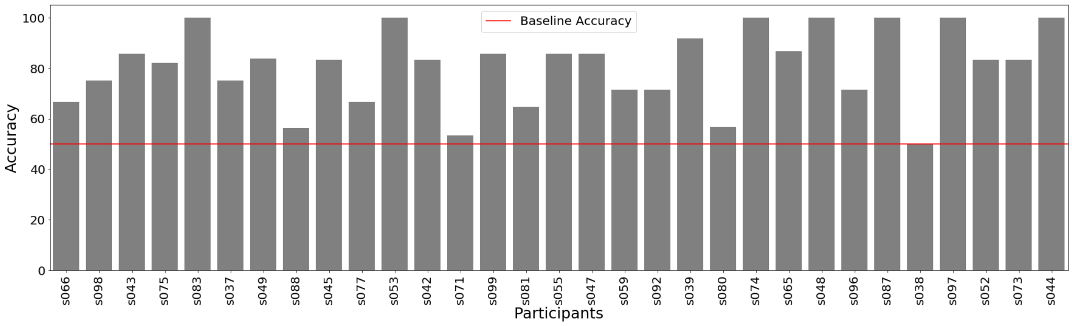


Figure 4.4. Classification accuracy per subject of the SVM-Linear using all the sensors.

Multi-modal laughter recognition

In this section we present the results obtained when combining data from different sensors. To combine data from different sensors, we used a features level fusion approach. Specifically we used the feature concatenation method, which generates a single feature vector from the concatenation of all the features of each sensor [242]. Table 4.3 shows the results for different combination of sensors. We can observe that the best results – shown in bold – are obtained when combining all the three sensors and using SVM-Linear, in particular an accuracy of 81%, a precision of 82%, recall of 81% and F1 of 80% are obtained, the SVM-Linear significantly outperforms the baseline respectively by 31, 35, 28 and 31 percentage points.

This result underlines the importance of considering laughter as a multi-modal rather than an uni-modal expression [173]. Indeed in general we can observe that despite the combination of sensors, the results obtained (especially with linear classifiers) are higher than the ones obtained using a single sensor.

We underline also that the performance obtained using a multi-modal approach are less dependent on the subject used for testing, compared to the ones obtained using the uni-modal approach.

Figure 4.4 shows the per-subject accuracy obtained using all the sensors and the SVM-Linear, it goes from a minimum of 50% to a maximum of 100%. We can notice that only four participants' accuracy is close to the baseline's accuracy, while the others' is much higher, indicating that the improvement given from the fusion of information compensates the poor performance of one sensor for the majority of the subjects.

Sensors	**SVM-Linear**				BRG			
	A	P	R	F1	A	P	R	F1
EDA + BVP	92*	81*	79*	78*	52	44	45	41

Table 4.4. Performance of the SVM-Linear and the baseline for the recognition of laughter episodes from cognitive load.An asterisk (*) identifies the significant difference ($p < 0.05$) from the baseline (BRG) and a large effect size (Cohen's $d > 0.8$)

Significant features

Out of 56 features, 45 presented a significantly different distribution in the laughter and non-laughter classes (tested with the KS test). No features have been discarded from the ACC set, only one from the EDA set (the minimun value of the mixed-EDA), while 10 features of the BVP set did not show significant difference in particular: the area under the curve, the mean of the first and second derivative, the min, mean and max of the BVP, the number of peaks, the mean of their amplitude and the slope. These results confirm the BVP, EDA and ACC as meaningful proxy for detecting laughter.

Laughter vs cognitive load

To understand whether cognitive load could elicit a reaction that could be misinterpreted as laughter, we extracted features presented in Table 4.1 from the EDA and BVP data during the cognitive load task.

The average length of the recorded cognitive load tasks performed by each participant is 21.52 seconds (SD = 10.09 seconds). Given the considered laughter segments' average duration equal to 4.19 seconds, to compare the laughter episodes with the cognitive load task we extracted features from non-overlapping windows of four seconds.

We obtained 175 cleaned instances of cognitive load which we compare against 177 laughter instances.

To understand the difference between cognitive load and laughter we performed two steps. We applied first the KS algorithm then we ran the classification pipeline described before using the SVM-Linear since we already observed that was the best performing classifier.

From the first step we notice that the features that did not show significant differences in the two conditions were only 12 out of 45: from the EDA only the number of peaks of the mixed-EDA and the phasic component and from the BVP: HR, min, mean, max of the BVP, the peaks' distance, the length of the pulses, the SDNN, the mean of the second derivative, the standard deviation of the pulses' amplitude and the slope.

Table 4.4 shows the results of the classification task. We can observe that, using the significant features, the SVM-Linear outperforms the baseline with an high accuracy of 96%, precision of 88%, recall of 80% and F1 of 83%.

These results are particularly encouraging in terms of applicability of the laughter recognition using physiological signals. Indeed they suggest that it is be feasible to use unobtrusive wearable sensors in real-settings to recognize laughter episodes and that these episodes should not be confounded with cognitive load.

Laughter vs clapping hands

To contextualize the laughter reaction in terms of body movements we extract the same set of features from the ACC data presented in Table 4.1 from the clapping hands task. The average duration of the task is three seconds (SD = 1.09 seconds). We tested the difference between the distribution of the laughter episodes and the clapping hands segments using the KS test. We observe that all the 11 features extracted from the ACC present a significant difference in the two conditions.

We can conclude that a laughter episode will unlikely be confounded with a movement as clapping hand, however it might be confounded with less intense movements such as writing or typing.

4.4.1 Summary of the main findings

In the study presented in this chapter, we demonstrated the feasibility of using physiological and body movement data gathered from wristbands to distinguish between laughter and non-laughter episodes. Our main findings are:

- The ACC – proxy for body movement – resulted the most informative sensor in the uni-modal approach. Using ACC features as input to the SVM-Linear classifier allowed to achieve a 75% accuracy in distinguishing laughter from non-laughter episodes.

- When combining the features derived from *all* the sensors, using a feature concatenation strategy (multi-modal approach), as input to SVM-Linear, we obtained an increment of the performance. Specifically, in this configuration laughter episodes can be distinguished from non-laughter episodes with an accuracy of 81%, confirming the multi-modal nature of laughter.

- When distinguishing between laughter and cognitive load segments, we obtained an average accuracy of 92%. This indicates that physiological responses during laughter episodes can be reliably distinguished from the ones during cognitive load.

- All the features from the ACC present a significantly different distribution in the laughter and clapping hands conditions. This indicates that motions generated by laughter are distinguishable from the ones caused by clapping hands.

4.5 Implications

A practical implication from our results is the possibility of using physiological data and body movement data collected using wristbands to recognize laughter episodes. Researchers and practitioners aiming to integrate the expression of laughter into their engagement-aware systems could rely on our approach.

In the context of this thesis, we consider laughter as possible additional cue for assessing students and knowledge workers' engagement in particular situations e.g., breaks, meetings, and lectures. In this direction, the engagement-aware system should be able to recognize when the user in involved in one of these situations and "activate" the laughter recognition engine. For instance, with the method presented in Chapter 5, we show that is possible to distinguish when knowledge workers take breaks from when they are involved in a work activity. The system could use the information about when the worker takes a break as a trigger for activating the laughter recognition.

Once information about laughter (presence, absence, number) are available, they could be used as input to the engagement recognition model. For instance, features about laughter episodes could be combined with the EDA features proposed in the study presented in Chapter 3 for recognizing students' engagement during lectures. Further, laughter episodes could be used as an additional cue to understand the overall *classroom emotional climate (CEC)* [259]. For instance, in the work presented in [260], we investigated the relation between the physiological synchrony, measured using EDA, among students and the CEC. We observed that the more students are synchrony with each other, the more positive is the emotional climate in the classroom [260]. In this direction, the automatic recognition of concurrent laughs from the students, or lack thereof, could be exploited as an additional indicator of a positive or negative CEC .

Information about laughter episodes could be also retrieved to the user, for tagging particularly enjoyable moments, e.g., through a “laughter counter” – to make users gain awareness of their emotional expressions [241].

Given the positive effect of laughter and humor on mood, systems might also incorporate the ability of making the user laugh [246; 261]. Virtual agents or applications running on mobile and wearable devices, could for example send funny videos or jokes with the aim of assessing [261] or ameliorating the mood of the user.

However, people have different perception of what is funny and what is not, so in this case an automatic laughter recognition engine is fundamental for the agent to verify the success of the intervention and adapt the contents of the intervention to the users’ humor [246].

4.6 Limitations

In this chapter, we presented promising results towards the recognition of laughter episodes using physiological and body movement data. However further work is needed to overcome the limitations of our current approach.

A limitation of our approach regards the setting of the study. We tested the current solution in laboratory settings, thus neglecting the noise introduced in real-settings. However, in our scenario participants were sitting in front of a screen watching funny videos, this could be a similar setting as workers watching videos during a work break, so equivalent responses could be expected. Future research should validate our method in real-settings.

Another limitation of our approach derives from the fact that we excluded from the recognition task all the moments when the participants were moving, those moments might reduce the performance in terms of the ACC. However, we performed this step to ensure the reliability of the physiological data not being affected by motion artifacts.

Lastly, we observe that physiological data are suitable for recognizing more intense laughs which also present longer duration.

Including additional information, in terms of e.g., short recordings of audio and facial expression (if possible), could be explored in future research to overcome this limitation.

4.7 Summary of Chapter 4

In this chapter of the thesis we presented our novel method for recognizing a multi-modal behavioural expression of engagement: *laughter*.

Building upon the existing literature describing laughter as a multi-modal expression involving, among others, physiological and movement reactions, we proposed to recognize laughter combining data from EDA, BVP and ACC sensors gathered from wristbands.

To evaluate our approach we collected the *USI_Laughs* data set using an experimental procedure in a laboratory setting with 34 participants.

We ran an extensive data analysis and showed that most of the features extracted from the considered modalities present a significant difference in laughter and non-laughter segments, making those modalities a reliable source of information for recognizing laughter episodes.

We analyzed the role of uni-modal and multi-modal approaches for recognizing laughter. We observed that the best performance is obtained when the features from all the modalities are used as input to the linear SVM classifier, achieving an accuracy of 81% which outperforms the baseline of 31 percentage points. These results confirm the multi-modal nature of laughter.

We contextualized laughter reactions and demonstrated their differences from tasks that could elicit similar responses e.g., clapping hands and cognitive load. In particular we reported that all the features extracted from the ACC present a significant difference between laughing and clapping hands.

Furthermore we show that, despite cognitive load activate the SNS, as laughter should do, the responses generated by the two events are highly distinguishable. Indeed, using BVP and EDA significant features as input to a SVM with the linear kernel, laughter and cognitive load segments could be distinguished with an average accuracy of 92%. Results of this study have been published in [B].

We believe that our findings can open plenty of new possibilities for the integration of laughter recognition into engagement-aware systems.

Chapter 5

Flow and Activity Recognition in the Workplace Using Context, Physiological Data, Movement Data, Laptop Usage and Phone Usage

Flow is a positive state of mind occurring when people are totally immersed and deeply engaged in an activity [32]. Research demonstrates that flow experienced at work is positively linked to job performance and well-being [26; 92].

Recent research shows that physiological parameters – from cardiac activity and electrodermal activity (EDA) – can be sued to measure flow [192; 41; 118; 45; 47]. Even though the relationship between physiological parameters and flow during daily activities is still debated [41], flow has been for instance linked to a moderate level of arousal, neither too high (as stress or anxiety) or too low (as boredom or relaxation) [41; 118]. Researchers have also shown that context, described for instance in terms of time of the day and activity, is a predictor of perceived engagement and flow [50; 132; 128]. For instance, Debus et al. [132], show that during programming activities flow occurs more often than during personal or administrative activities. Nielsen et al. [128] demonstrate that planning, problem solving, and evaluation activities are significant predictors of flow, whereas brainstorming activities are not [128]. Additional discussion about flow at work is presented in Section 2.1.3.

Building upon previous research, we propose to combine phyisological data (EDA and BVP data) gathered from unobtrusive wristbands with context information (type of activity, time of the day, day of the week) collected using self-reports. We experimented with several fusion strategies based on traditional machine learning and deep learning. Our results show that when the type of activity, the raw EDA and BVP data are used in input to a *sensor based late fusion (SB-LF)* strategy implemented using a *convolutional neural network (CNN)*, it is possible to distinguish between levels of flow with a balanced accuracy of 71%.

Despite the importance of flow at work, few approaches have been proposed for its automatic recognition using sensor data [155; 45; 47; 41] and most of the existing studies are conducted in laboratory settings with simulated work activities [45; 41] which might be not representative of the actual work ecosystem. In contrast to previous work [45; 41], we validated our approach using data that we collected “in-the-wild”, during actual work days and work activities.

Besides recognizing flow, we aim also to recognize activities in the workplace. Previous work suggests that the activity is a relevant context information to consider when creating interacting systems [27]. Further, the results obtained from the before-mentioned analysis indicates the type of activity as relevant for recognizing flow at work.

Motivated by these reasons, we move a step towards the automatic recognition of activities at work. Specifically, we aim to distinguish between two types of workplace activities such as *work* and *break* activities. More details about definitions of these type of activities and existing methods for their automatic recognition are also discussed in Section 2.6 and in in Section 2.7.

As discussed in Section 1.2.2, activity recognition in the workplace is a challenging problem. People use different tools (both physical and digital), perform several tasks, work alone and with others and in different locations (e.g., at home, in the office).

To characterize the variability of work and breaks activities, we propose to quantify cues such as the physiological activation, the physical movement, the laptop and phone usage. To this end, we leverage data gathered from personal devices i.e., wristband, laptop and smartphone, that people can use anywhere and at any time.

Existing methods in assessing breaks, rely on simple heuristics, assuming that a break takes place when, e.g., there is no laptop activity [3; 61], the user is not sitting at her desk [56; 51], or she is engaging with social media, news, or shopping websites [51; 199; 200]. However, browsing a news website, not being at one’s own desk, or not interacting with a laptop, often do not imply that the user is taking a break. Thus relying on fixed heuristics is limiting for dealing with inherent intra-activity variability.

To overcome this limitation, in this study, we take a data driven approach based on machine learning. Further discussions on the advantages of using this method are presented in Section 1.2.2.

Our results show that using features from EDA, ACC, phone and laptop usage, in input to a gradient boosting classifier, breaks can be distinguished with F1 score of 69% and work activities with a F1 score of 94%, corresponding to an improvement against baseline methods of 12-54 and 5-10 percentage points respectively.

To summarize, in this chapter we present two major contributions: (1) a method for automatically recognizing flow during work activities based on the combination of context and physiological data; and (2) a method for recognizing activities in the workplace leveraging data from multiple sensors.

The first contribution targets the second research question of this thesis (*RQ2: How can information about context a) be fused with physiological data and b) impact the recognition of workers' engagement?*). Results of the first contribution have been accepted as conference paper [C] at the upcoming ACII conference (September 2021). Part of the text written in this chapter is reported from [C].

The second contribution targets the third research question of this thesis, *RQ3: How can activities in the workplace be accurately recognized?* Results of this work have been published as journal paper [D] in the PACM IMWUT (September 2020). Part of the text written in this chapter is reported from [D].

To evaluate our approaches we conducted a user study *in-the-wild* with 13 knowledge workers from Academia and collected the *WorkplaceDataSet*. We gathered continuous streams of physiological and acceleration data (using a wristband) as well as laptop and phone usage. We asked participants to report their activities and perceived flow, while keeping their usual work routines. In order to collect self-reports, we proposed to use a multi-device strategy. Specifically, we provided participants with multiple instruments we designed for this study such as a laptop widget, a mobile application, a paper-and-pen diary, and a situated self-reporting (SSR) device. We then asked participants to use the one(s) they preferred. We analyzed users' attitude towards the different data entry modalities and reported our observations. In each of the contributions of this chapter we used a subset of the *WorkplaceDataSet*.

The remaining part of this chapter is structured as follows. We discuss existing literature in Section 5.1. Section 5.2 describes the data collection. In Section 5.3 we discussed our observations on the use of a multi-device strategy for collecting self-reports. Section 5.4 and Section 5.5 provide an overview of the data analysis method and results obtained for respectively the automatic recognition of flow, and workplace activities. In Section5.6 we discuss the implications of our results and in Section5.7 we present the limitations of the proposed methods. We conclude with a brief summary of the chapter.

5.1 Related work on methods for collecting self-reports, recognize flow and activities in the workplace

In this section we review existing literature. Specifically, we review literature targeting data collection strategies for collecting self-reports in the workplace, as well as activity and flow recognition in the workplace. Additional discussion about the existing literature in this topic is presented in Chapter 2.

Methods for collecting self-reports in the workplace

Most of the existing approaches designed to collect self-reports of knowledge workers actively request user inputs at specific moments of the day. Notifications to remind the user to log self-reports are typically sent either through the laptop [15; 61], or through the phone [262] and are usually sent at pre-defined moments [15; 51], or triggered by specific events [263]. This approach is widely adopted but presents some drawbacks. First, a notification sent at an inappropriate moment can disrupt the user and cause time loss and frustration [264; 263]. Second, as discussed by Lathia et al., random or context-based sampling could bias self-reported data [265]. Third, sending notifications at pre-determined moments may hamper the possibility to capture the duration of a work activity as perceived by the users, who, as reported by Luo et al., tend to have different preferences in setting the duration of their work activities [197]. To overcome these limitations, we propose to let users decide when and for how long to log their activities by starting and stopping an incremental timer.

Existing approaches for gathering self-reports at the workplace often use a single device, e.g., the laptop [15; 61; 197] or the phone [262]. However, taken individually these devices present several disadvantages, as summarized in Table 5.1. To allow for flexibility and personalization, we instead adopt a multi-device strategy to collect users' input.

Also the authors of [266] adopted a multi-device strategy and collected self-reported mood in a participatory manner. The *Quantified Workplace* system developed by the authors and described in [266] allows workers to report their mood either using the personal smartphone or using shared tablets placed in the office. The authors reported a significant preference of the workers towards the use of tablets as input device compared to smartphones [266]. Using commonly placed devices for gathering self-reports is a promising approach for quantifying the workplace in terms of e.g., behavioural dynamics across employees, which is the goal of the work presented in [266]. In this work we aim of gathering self-reports about individual's activities when they occur and as they are defined by users. For this reason we use tools that workers can easily carry or can place on their work station. Specifically, we rely on the use of paper-and-pen diary, phone, laptop and we design a single-purpose, situated self-report device, inspired by Paruthi et al.'s work [63], which we called *Devo*.

Flow recognition during work activities

Flow is a subjective experience. The gold standard methods for measuring flow in psychology research thus, consist of using questionnaires and experience sampling methods (ESM) [121]. Despite the important knowledge gained with these approaches, these methods prevent a continuous and unobtrusive measurement of flow levels experienced during daily activities as well as the creation of adaptive systems that can promote flow. To overcome this limitation, Peifer [118] discusses the potential of leveraging physiological information, in combination with self-reported measurements, as proxy for a continuous assessment of flow.

Recent studies in psychophysiology research have investigated the physiological indicators of flow [118; 125; 124; 126].

Electrodermal activity and cardiac activity measurements are in general the most used physiological parameters considered in psychophysiolgy studies of flow [124; 118; 125] as also discussed in Section 2.1.3. Being EDA a direct measure of the activation of the SNS and the cardiovascular measurements, that can be derived from BVP, providing information of both branches, they represent promising proxies of flow[124]. For instance, Peifer et al. observed an inverted-U relation between physiological arousal (derived from cardiovascular indices) and flow [118]. In particular authors observed experience of flow happening in correspondence of a moderate level of activation of the sympathetic nervous system measured with the low frequency of the HRV [118] in a lab study. Authors argue that the highest levels of flow could be identified at moderate levels of arousal while the lowest levels of flow were identified during lowest (boredom) or highest (stress) levels of arousal [118]. Gaggioli et al. found a positive relation between cardiovascular measurements in terms of LF/HF with flow during daily activities performed in a natural environment by 10 participants [125]. Being the ratio between LF and HF a proxy for the balance of sympathetic and parasympathetic nervous systems, their findings seem in line with the ones of Pefeir et al. [118].

Even though the above-mentioned studies tried to connect physiological responses to flow, they are not specifically targeting work activities, and do not consider a data-driven machine learning (ML) approach as we do in this work.

ML techniques have the potential to discover additional relations between

Strengths	Weaknesses
Phone	
Phones account for mobility and flexibility [63], thus they could be suitable when people work in different locations (for example when reading on public transportation).	Being multi-purpose devices, phones represent a source of distraction in the workplace and workers tend to not use them when they want to focus [267]. So using this device might cause a collection of fewer labels and risk to increase workers' distraction.
Laptop	
The Laptop is a widely used device for knowledge workers so it is easy to reach especially when workers deal with type of activities that require the use of this device (for example for coding).	Might miss the opportunity of gathering information when the user is working in different locations or using other devices or tools (e.g. tablets, books). Given the large amount of application already being used for working purpose, the worker might forget to use also the applications for gathering labels if not reminded. However, sending a reminder on such a device can disrupt the workflow and cause stress and frustration.
Situated Self-Report (SSR) device	
The physical presence of an SSR can act as reminder for workers to enter self-reports [63]. Being single-purpose devices, SSRs reduce the time of usage and the user's burden [63].	Being located in specific positions, SSR devices might not be used by workers when changing environment (for example when working from home).

Table 5.1. Strengths and weaknesses of devices typically used to collect self-reports at the workplace.

physiological information and flow [47], allow a continuous automatic assessment and enable systems to adapt to the users' inferred flow. ML approaches have been recently applied to physiological signals gathered from wearable devices to infer flow in gaming [192; 268] and few during work activities [47; 41; 45].

The authors of [45] and [41] conducted laboratory experiments in which participants were asked to perform simulated work activities. However, being flow connected to the type of activity the individuals perform and to the motivations they have to address them, building models using physiological reactions occurring during simulated work activities might not generalize when applied to actual work tasks. Further, the laboratory setting prevents to take into account the noise and variability that happen in real-life settings. For instance, data collected in laboratory settings is not affected by the noise in sensors' recordings due to device malfunctioning, the missing data due to users forgetting to wear the device, the situations that generates physiological responses similar to flow (e.g., physical activity, eating), the absence of these aspects limit the ecological validity of the results.

In the study presented in this chapter, we use a data set collected *in-the-wild* while workers performed their daily tasks. Recently Rissler et al. [47] conducted a *in-the-wild* study to recognize flow during work activities. They gathered HRV measurements from 9 knowledge workers using a chest band, and sent notifications at random times during the day for collecting ground-truth about flow level. They processed hand-crafted features extracted from 134 five-minute windows using shallow classifiers and achieved 70.6% accuracy in distinguishing low and high levels of flow.

In contrast to the approach presented in [47], we use a data set that contains data from 390 unique activities – corresponding to 284 hours – collected from 13 knowledge workers. To have a more complete picture of the activation of the SNS and PSN, we combine cardiovascular and electrodermal activity measurements collected using a wrist-worn device. We further investigate the role of context information in the recognition of flow and experiment with different fusion strategies based on classical machine learning and deep learning methods.

Activity recognition in the workplace

Many existing approaches for the passive monitoring and automatic recognition of activities in the workplace consider digital activities as a proxy for work activities [61; 53]. However, as also pointed out by Koldijk et al. [53], knowledge workers typically use several physical tools and digital applications to accomplish their tasks and do not think about their intended activities in terms of applications used [53]. Instead, they might use the same application in different type of activities [53]. In addition to that, considering only the laptop usage as main source of information prevents the possibility of recognizing activities that do not imply the use of the laptop, such as face-to-face meetings, reading books, or taking a break [53; 55].

Recently, Avrahami et al. [55] explored the use of a RF-radar placed under the desk for the recognition of activities that do not involve the use of the laptop, e.g.,*reading papers*. Oliver et al. [54] leveraged data collected using microphones, cameras and laptops as input to Layered Hidden Markov Models to infer the type of activity the user was performing. These approaches focus on recognizing activities that happen in the office or at the desk and do not explicitly target the recognition of breaks. On the contrary, we focus on the automatic recognition of breaks and work activities, without making assumptions about their type and location. We use multiple sensors collected using personal devices to account for the different characteristics of work and break activities.

In contrast to external and internal interruptions [196], which are not triggered or desired by the user, work breaks are defined as moments in which workers *voluntarily* pause their work [195]. Breaks play a fundamental role in workers' productivity and overall well-being [3], yet there is no a unique definition for them [195]. Indeed, knowledge workers may categorize several different activities as work breaks, and activities categorized as breaks might vary significantly across individuals [195].

Most of the existing approaches use simple heuristics to determine when a person is taking a break or not [56; 3; 200; 51; 61; 199]. However, such heuristics are not validated by letting workers confirm whether the system correctly assessed when they were taking a break [56; 3; 51; 199] nor is the correctness of the assumptions made verified otherwise [200]. Thus, a direct comparison of the performance obtained in our work with existing approaches is difficult to perform.

The main cues used in previous work for determining whether a worker is taking a break are: laptop inactivity [3; 61], absence from the desk [56; 51] (sensed using cameras [51] or Bluetooth beacons [56]), physical movement [56] (assessed by counting the number of steps), and time spent on *digital breaks*, defined as moments when workers engage with websites and applications related to social media, news or shopping [51; 199; 200]. Considering such cues only separately can lead to errors. For instance, laptop inactivity can or cannot indicate that the user is taking a break. As phrased by Tseng et al.:"*Physical breaks necessarily result in periods of computer inactivity; however not all the periods of inactivity are physical breaks*" [3]. Similarly, absence from the workstation does not necessarily imply a break from work, as workers might change their work location or be involved in scheduled and unscheduled meetings at different locations. The use of specific websites or applications might vary depending on the context. Workers might use messaging applications, social media or news for both work and personal purposes [269]. Considering their usage as the only hint for detecting breaks can thus be misleading.

Given the complex nature of breaks as perceived by workers [195], in this study, we explore the role of multiple cues derived from different sensory channels for recognizing breaks. In particular, we investigate the combination of laptop and phone usage, physical movement and physiological responses. Only two of the above mentioned approaches use the input of multiple sensors to detect breaks. Kaur et al. propose to determine that users are taking a break when both absence from the desk and digital breaks are detected [51]. Combo et al. [56] propose to use absence from the desk and movement to detect a physical break. The validity of these heuristics in real settings has not been validated, though. Instead, we extensively evaluate the performance of our approach using data collected in-the-wild and also investigate how different sensory cues contribute to the detection accuracy.

5.2 Data collection of the WorkplaceDataSet

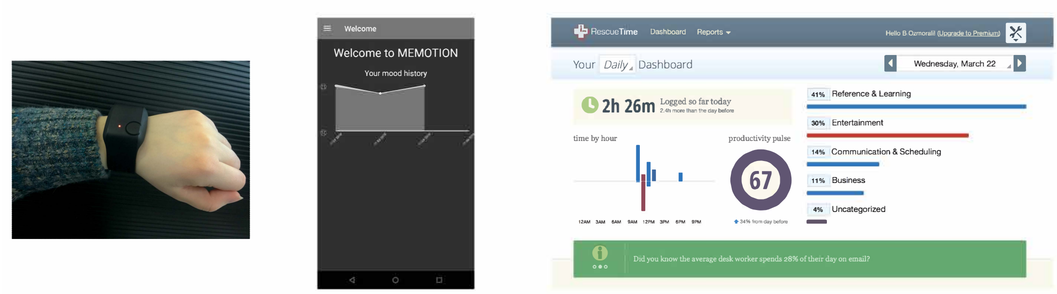
In this section we describe the participants, the tools we used and the procedures we followed for collecting the *WorkplaceDataSet*.

Participants

We recruited 14 knowledge workers from Academia employed as Professor (1), Post-Doc (1), Researcher (1) and PhD student (10) at the Computer Science department of our university. One participant failed to install our data collection tools and did not provide valid data. Our data set thus contains data from 13 users (9 males and 4 females, 7 of age in the range 20-30 and two in 30-40). At the onset of the study, we organized a workshop during which we explained the data collection procedure and asked participants to sign an informed consent agreement.

Collected data

Sensor data. We collected continuous traces of electrodermal activity (EDA), blood volume pulse (BVP), acceleration (ACC), phone usage and laptop appli-



Did you know the average desk worker spends 28% of their day on email?

Figure 5.1. Data collection instruments for gathering sensor data: E4 wristband (left), MEMOTION Android application (center), RescueTime laptop monitoring tool (right).

cation usage, using 3 devices – laptop (La), phone (Ph) and wristband (Wr). Figure 5.1 shows the tools used for collecting sensor data.

Since activities can affect the physiological arousal [217] our expectation is that EDA, and BVP data can help distinguishing work from break activities. Being the level of arousal connected with the level of activation [270], we expect for example higher physiological arousal when workers take breaks and have social interactions. Further, given electrodermal activity and cardiac activities been studied in the physchophysiolgy of flow, as discussed in Section 2.1.4, we consider them as relevant information to use for recognizing flow.

ACC is often used as an indicator of physical activity and to compute, e.g., the number of steps [56; 271]. Since knowledge workers are known to be sedentary [197] and to spend most of their time at their workstations [197],it is reasonable to expect a generally low variation in their physical activity. Moments of stronger physical activity could however hint to physical breaks, e.g., short walks [3; 56].

To gather data from users' phones, we developed an Android application called *MEMOTION*. The application collects user interactions with the screen as a proxy for phone usage [271]. In the workplace environment, the phone is usually not the main work device and it is instead considered a source of distraction [267]. We thus expect a more intense usage of the phone during breaks.

We further collected laptop application usage through the RescueTime monitoring tool[1]. Since the laptop is typically the main work device for knowledge workers, its usage is widely used as proxy for workers' activity [61]. Previous work, for instance, used laptop inactivity as cue to identify breaks [3; 61].

To ensure synchronization among devices we visually verified that the clock

[1] `https://www.rescuetime.com/`

Construct	Item
Satisfaction with performance [272]	During this activity, I satisfied the personal expectations I have of this activity
Flow [26]	a. I was totally immersed in this activity. b. I got my motivation from the activity itself, and not from the reward for it c. I did this activity with a lot of enjoyment
Type of activity	Select the type of activity you just completed

Table 5.2. Questionnaire used in the study to collect ground-truth.

Activity type	Description
E-mail	Dealing with e-mails.
Planning	Editing work items/tasks/todos; creating/changing calendar entries.
Coding	Reading/ editing/ navigating code (and other code related activities).
Meeting	Meeting/call.
Learning	Information acquirement, learning, and knowledge gain. Ex: taking seminars/ online courses, read books, read papers.
Read/Write	Reading/editing documents, project reports.
Research project	Conducting experiments, data analysis, design.
Other	Other work activities not in the list.
Break	-

Table 5.3. Type of activity and descriptions

of the phone and laptop of each participants matched the same time reference. The E4 automatically synchronizes with the internal clock of the laptop, to which it is connected at the beginning of and during the study.

Ground-truth data. Refers to the type of activity performed by the workers as well as the perceived flow level. We used the experience sampling method (ESM), described in Section 2.2, and asked participants to fill-in a questionnaire as common practice in the literature [263; 61; 47]. The questionnaire used in this study is presented in Table 5.2. Specifically, we used the *Work-Related Flow Inventory (WOLF)* questionnaire developed by Bakker [26] to measure perceived flow. It conceptualizes flow at work based on the constructs of (a) *absorption*, which refers to the sense of total involvement in the activity; (b) *intrinsic motivation*, which refers to the need of performing the work activity because of the pleasure and satisfaction it elicits and (c) *enjoyment* which refers to the perceived enjoyment during the activity [26]. From each of the constructs we selected and adapted an item to the work activity as common practice in the literature [273]. More details about flow at work and its component are discussed in Section 2.1.3.

For the type of activity, we asked participants to report the work activity they were performing by selecting among eight categories – *meeting*, *read/write*, *coding*, *learning*, *research project*, *email*, *planning* and *other* –, which were used also in previous work [61; 194] and further added the *break* category. To enable participants to select the appropriate category we provided them with a brief description of the work categories along with examples, similar to previous work [61; 194]. We did not provide an explicit description of break activities because, as reported by Epstein et al. [195], workers might consider different activities as breaks. We thus avoided imposing specific definitions of non-work activities and let users log whatever they consider to be a work break. Table 5.3 summarizes the type of activities and their description.

Lastly, participants reported their satisfaction with own performance using the single item questionnaire adapted to the activity [272], original: “Today, I satisfied the personal expectations I have of my work”, adapted version: “During this activity, I satisfied the personal expectations I have of this activity”.

We adopted a multi-device strategy [274] to enable users to log self-report data. Specifically, we allowed participants to use a paper-and-pen diary and further provided them with three digital tools shown in Figure 5.2: the *MEMOTION* application for the smartphone, a laptop widget, and a single-purpose, situated self-report device called *Devo* described below. Each tool features an incremental timer that users can freely start and stop to log the duration of their activities. Each time the timer was stopped, the questionnaire was displayed to the user.

Since each data entry method has its own advantages, as summarized in Table 5.1, and users tend to have different attitudes towards different devices, we assumed that offering a set of devices rather than a single one could ease data entry and, ultimately, improve data quality.

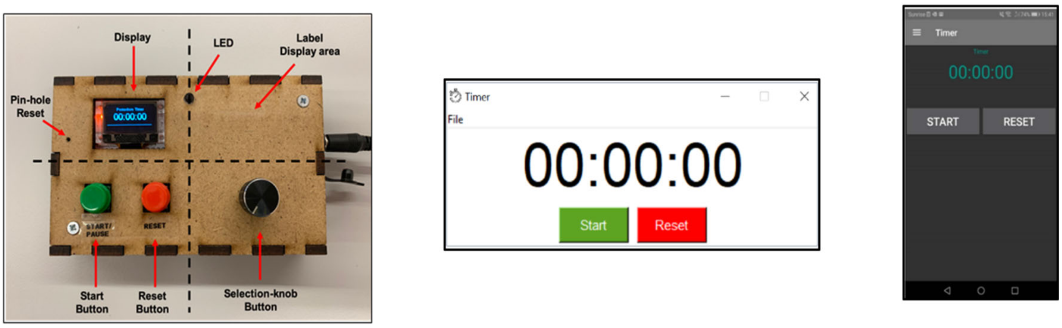


Figure 5.2. Data collection instruments for gathering ground-truth: Devo (left), Laptop widget (center), MEMOTION Android application (right).

Devo

A situated self-report device is a "*situated device intended to be placed in a location to optimize user's self-reporting efficacy*" [63]. Being single-purpose, SSRs have a short access time and represent a low burden for the user [63].

We designed *Devo* according to the design dimensions of *SSR* systems proposed by Paruthi et al. [63]. *Devo* has a simple, tangible interface, shown in Figure 5.2, designed to facilitate the interaction with the user. Data is stored in the internal memory of the device and uploaded every night via Wi-Fi to a remote server.

Devo consists of: a microcontroller equipped with Wi-Fi and Bluetooth modules, a LED, a small OLED screen, two buttons and a rotating knob. We chose a wooden case, as suggested in [63], to give the device a pleasant appearance and long durability. We produced six *Devo* devices at a cost per unit of approximately 30$. While the screen increases the cost of the device, it also makes it able to support multi-items questionnaires and to provide visual feedback to users. Specifically, Devo displayed to users the number of logged activities per day. We asked participants to position *Devo* on their desk and to use it to log their activities. As suggested by Paruthi et al., the physical presence of a SSR device reminds participants to log their activities [63]. To log an activity, workers started/stopped the timer using the (green/red) button below the screen, and answered the questionnaire using the rotatory knob.

Data collection procedure

We asked participants to perform their work activities as usual, log them using their preferred device – phone, laptop, *Devo*, or paper-and-pen diary – during at least one typical working week. We asked them to wear the wristband at the beginning of their work day, keep it at least during the working hours, charge it and upload the data using the E4 Manager by Empatica installed on their laptops. We monitored the quality and quantity of the collected data using daily reports automatically generated by our tools and reached participants for fixing issues when needed.

To evaluate participants' experience with our data collection strategy we used a post-study questionnaire. We asked participants to report which was the device they preferred the most and why. We asked also to report whether logging their activity in the workplace helped them to achieve their goal and feel more productive at work. This is because the use of a timer is also a popular technique to boost productivity and improve the scheduling of tasks at work [275].

5.3 Using multiple devices to log self-reports: collected data and our observations

In this section we describe the amount of self-report data we collected during our study and discuss our experience with the use of a multi-device strategy for gathering self-reports at work.

Collected self-report data

For the 13 participants that collected valid data, the number of days with at least one self-report range from four to 13 (Mean=7, Std=3). In total, the participants logged 625 activities. A subset of these entries (68) were incomplete due to either the participants not completing the manual data entry or to errors in the remote upload. The cleaned data set has 567 entries, of which 73 (13%) are breaks. Taken together, these 567 entries correspond to 401 hours of labelled data.

On average, participants logged their activities for about 5 hours per day, which corresponds to half of a "typical" work day of nine hours. Workers logged activities from 6 a.m. to 11 p.m., the majority of which (95%) between 9 a.m. and 5 p.m. The number of entries (both work or break activities) per participant range from 24 to 82 (Mean=43, Std=19). The duration of individual activities range from 100 seconds to 187 minutes (Mean=42 min, Std=32 min).

The total number of breaks is 73, reported by 10 participants. (Three participants did not log any breaks and their data). The number of logged breaks per participant ranges from 1 to 14 (Mean= 7, Std=3) and their duration ranges from 3.3 to 113 minutes (Mean=42, Std=26).

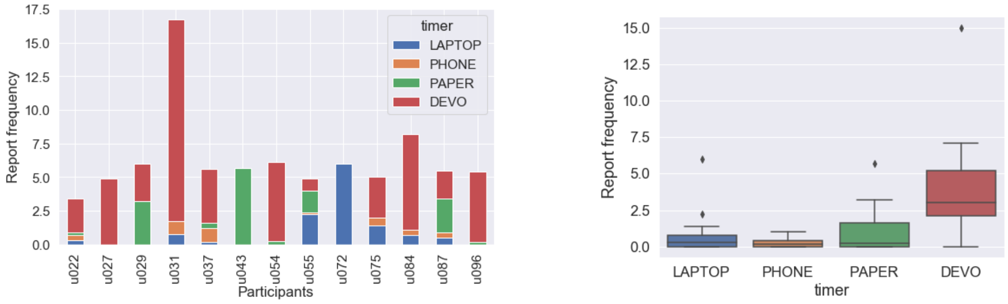


Figure 5.4. Report frequency across devices.

Preferred devices to log self-report data

To evaluate participants' preferences in choosing specific devices to log their self-reports, we used the *report frequency* metric as suggested in [63]. This metric is computed as the ratio of the number of self-reports and the number of days of data collection for each participant. Figure 5.3 shows the report frequency for each participant and per device.

We notice that most of the participants (10) used multiple devices to log their activities. Only three participants used only one device: u043 used only the paper-and-pen diary, u072 only the laptop, and u027 only *Devo*.

These observations hint at the need to account for personal preferences and use multiple devices to collect self-report data at the workplace. Future studies and practical deployments should foresee a setup phase during which participants could test and choose their preferred set of devices, as also suggested in [63].

Figure 5.4 also shows that, on average, our study participants preferred using *Devo* to log their work activities. In their answers to the post-study questionnaire, most of the participants reported that *Devo* was *easier* and *funnier* to use compared to the other data entry devices. They further indicated that *Devo*'s *physical presence* reminded them to use it. One participant wrote: "*[Devo] Was available in no time and funnier than the other tools*". Another participant said: "*I can see it [Devo] and remember to track my activities*".

Figure 5.4 also shows that the phone was the least used device. This is not surprising given the disruptive role of the phone in the workplace [267] and its reduced usage in this context.

Effects of data logging on (perceived) productivity

As part of the post-study questionnaire, we asked participants to rate two items: *I1: "Logging my activities helped me to achieve my goals at work*" and *I2: "Logging my activities made me feel more productive*" using a scale from one to five (from Totally disagree to Totally agree). Most of the participants (7) reported three (equivalent to "Neutral") to *I1* indicating that in general participants felt that logging their activities was neither positively nor negatively contributing in achieving their goals. However, most of the participants (8) reported that logging their activities made them *feel* more productive. This raises the question whether the use of a timer can help improve productivity, as reported in existing work [275], or actually make workers *perceive* themselves as more productive.

5.4 Flow recognition during work activities using context and physiological data

In this section we describe the procedure we followed and the results we obtained in the automatic recognition of flow during work activities using a combination of context and physiological data.

In the literature, flow is described both as a *yes-or-no* state (i.e., a person *is* or *is not* in flow) and as a continuous phenomenon (i.e., “the more the factors of flow are present the higher the experience of flow” [92]). When considering flow as a continuous phenomenon, the flow experience could be divided into *low* and *high* levels [92]. In this study, we adopted the latter characterization and describe the flow experience in terms of low and high levels of flow as in [47]. We refer to users to be *in flow* when they experience a high level of flow.

We first investigated the relation between context information on perceived flow using linear mixed effects model. Then we investigated the feasibility of distinguishing between *low* and *high* flow during work activities, and define the problem as a binary classification task. In the followings we report details about the cleaning and pre-processing procedure, the features we extracted, and classification pipeline, based on classical machine learning and deep learning. Results are discussed in Section 5.4.1.

Data cleaning and pre-processing

For performing the flow recognition analysis, we used a subset of the *WorkplaceDataSet*. Specifically, we focused on work activities (we excluded the breaks)

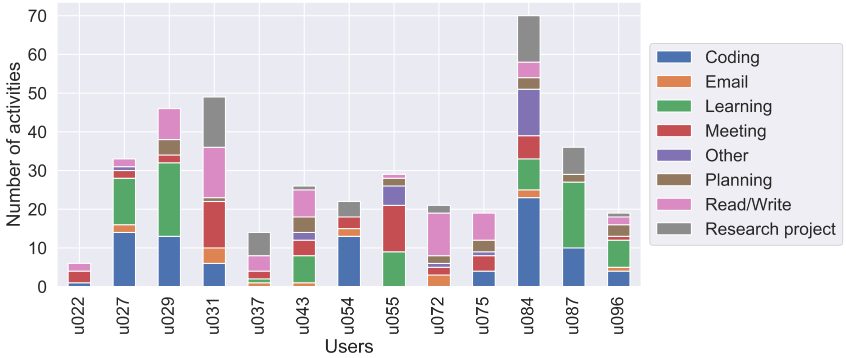


Figure 5.5. Overview of the amount of the collected activities and their type for each participant used for the flow recognition task.

with a minimum duration of 5 minutes and use physiological data i.e., electrodermal activity (EDA), blood volume pulse (BVP) and questionnaires only. We excluded 23 activities due to missing sensor data.

EDA data. We cleaned the EDA signals, sampled at 4Hz, using a first order Butterworth low-pass filter with a cut-off frequency of 0.4 Hz similarly to [254], as we did as well for the analysis presented in Chapter 4 for laughter recognition. We then applied the artifacts detection model *EDArtifact*[2], developed by our research group and presented in [70], which is particularly suited for data sets collected *in-the-wild*. We excluded 67 activities that contained artifacts for more than 50% of the duration of the activity. Most of the activities (75%) contains less than 25% of the time in artifacts (MEAN: 17%, STD: 24%), indicating the overall good quality of the data collected. We decomposed the EDA signal using the *cvxEDA* method [71] as we did in the previous studies.

BVP data. We cleaned the BVP signal, using a first order Butterworth FIR filter with a cut off frequency of 5 Hz, similar to [165]. We down-sampled the BVP signal to 4Hz as in [192].

Segmentation. We segmented the signals gathered during the activities using the timestamps obtained from the timers, then used a five-minute window as in [47] with overlap of 50% to extract the features or the raw data to give as input to the classifiers. We used the five-minutes segments to either compute the features or as a direct input to the CNNs.

Subset of the WorkplaceDataSet. The final data set used for this analysis contains 390 unique activities, corresponding to 284 hours in total and 6712 five-

[2] `https://github.com/shkurtagashi/EDArtifact`

minutes segments, from 13 subjects. The average number of entries per subject is 30 (SD: 17) collected over an average working week. Figure 5.5 presents an overview of the amount of activities and their type per each participant used in this work. Most of the activities are logged between 9 a.m and 6 p.m. and during workdays (from Monday to Friday), only 4 during Sunday. The average duration of an activity is 44 minutes (MIN: 5 min, MAX: 187 min).

Features extraction

We extracted 99 features from physiological signals and context used in previous work [68; 12; 47].

Physiological features. To characterize changes in the arousal levels, we extracted 17 features from the *EDA-mixed* and *phasic* component and 11 from the *tonic* component. Specifically we extracted time-domain statistical features (mean, min, standard deviation), wavelet-based features (e.g., mean and standard deviation of wavelet coefficient and 1Hz, 2Hz and 4Hz). We also extracted features that characterize the skin conductance responses (SCR) [35] (e.g., number of peaks, decay time, rise time) using the *EDAExplorer* [68] publicly available tool[3].

We characterized the cardiovascular responses extracting 11 features from the BVP signal in the time and frequency domain using the *HeartPy*[4] toolkit [80]. In the time domain we extracted features as the heart rate (HR), the mean and standard deviation of RR intervals (meani, SDNN), standard deviation and root mean square of successive differences (SDSD, RMSSD). In the frequency domain, we computed the absolute power of the low-frequency (LF) band (0.04-0.15 Hz) and high-frequency (HF) band (0.15-0.4 Hz) and their ratio (LF/HF), which should reflect the balance between the SNS and PNS [276] and an estimation of the breathing rate [80]. Details about physiological signals and features are also described in Section 2.3.3 of this thesis.

Context features. We considered the context information as categorical variables and used one-hot-encoding as done in [277]. For the type of activity we used the 8 work categories presented in Section 5.2 excluding the “break” activity since we focus on the work activities only. For the time of the day we used 6 categories corresponding to when the end of the activity was recorded: *early morning* (6 a.m. - 8 a.m), *morning* (9 a.m. - 11 a.m), *lunch* (12 a.m. - 1 p.m), *early afternoon* (2 p.m - 3 p.m), *afternoon* (4 p.m - 6 p.m), *evening* ($>$ 7 p.m).

[3] `https://eda-explorer.media.mit.edu/`

[4] `https://github.com/paulvangentcom/heartrate_analysis_python`

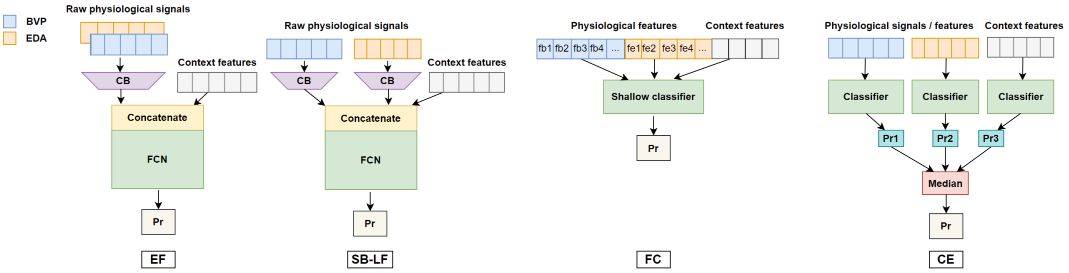


Figure 5.6. Overview of the sensor fusion strategies investigated. CB stands for convolutional base, FCN for fully connected network, Pr for probability. The classifiers in the CE configurations are either shallow classifiers (RF or GB) or CNN.

For the day of the week we used seven categories representing Monday to Sunday.

Linear mixed effects model

We tested the association between each type of context and perceived flow using a linear mixed effects model.

We conducted three tests, using the flow score as dependent variable, the context type as independent variable, and the subject as random effect to take in consideration the correlation among samples due to the repeated measurements, as in [132]. We encoded the context variables using the “dummy coding” procedure, which assigns zeros and ones depending on the group membership as in [132]. For the type of activity we use the *coding* activity as the reference category and set all the samples belonging to that category to zero, while the samples of the other categories were set either to zero or one depending on their occurrence, similarly to [132]. We set *early morning* as reference category for the time of the day variable and *Monday* for the day of the week variable.

Flow recognition pipeline

We describe below the flow recognition pipeline we designed.

Labeling. We derived the *flow score* by averaging the answers to the three items – absorption, motivation and enjoyment – of the *WOLF* questionnaire presented in Table 5.2, as common practice in the literature[12; 47] and as we did for students’ engagement study presented in Chapter 3.

We considered a participant to experience a high level of flow (*high_flow* class) when the flow score was higher or equal to four (corresponding to "Agree" in the 5-points Likert scale) and a low level of flow otherwise (*low_flow* class). We used four as threshold assuming that workers who experience high level of flow would report scores at the higher extremes of the scale, similar to what discussed in [215] and in Chapter 3.

From this procedure we obtained an imbalanced distribution of the two classes, with 144 activities in the *high_flow* class and 246 activities in the *low_flow* class, corresponding respectively to 2532 and 4180 five-minutes windows.

Sensor fusion. Figure 5.6 summarizes the fusion strategies considered in this work. Multi-modal fusion strategies can be divided in *feature fusion* and *classifier ensemble*. More details about these approaches are discussed in Section 2.5.2.

We considered two *feature fusion* approaches implemented using raw signals as input to CNN, namely *early fusion* [44] and *sensor based late fusion* [44], and compared with one based on hand-crafted features as input to shallow classifiers, namely *feature concatenation*. We further considered two *classifier ensemble* methods, one based on CNN and the other using shallow classifiers. Below we describe in details the different fusion strategies.

Early Fusion (EF). The EF fuses data of all the modalities, independently on the axis or channel, in the first layer [44]. It is the simplest CNN-based fusion strategy in terms of computational cost [44]. This strategy has been adopted in [192] to recognize flow during gaming using EDA and BVP. In this work we stack the EDA and BVP on the depth as in [192] and concatenate the generated features from the convolutional layers with the context features.

Sensor Based Late Fusion (SB-LF). The SB-LF splits the input in branches, corresponding to different sensors, that are processed individually to generate sensor-specific representations and then merged in a second moment [44]. This approach is highly flexible since it allows the architectures of the single branches to be tailored according to the sensors' characteristic [44]. Rastgoo et al. successfully used the SB-LF approach to recognize stress levels of drivers, combining electrocardiogram data with vehicle and environmental data [43]. In this work, we process each signal with separate convolutional layers, and concatenate the generated features with the context features.

Features concatenation (FC). The hand-crafted sensor-specific features, are concatenated with context features in a single vector and used as input to the shallow classifier. Contrary to representation learning methods, this method allows a better interpretation of the results in the feature space. Authors of [41] used the

FC approach to recognize flow levels during simulated work activities. In this work we concatenate hand-crafted features extracted from BVP and EDA with the context features and used them in input to shallow classifiers.

Classifier ensemble (CE). The prediction of different classifiers are aggregated to obtain the final decision. In the Affective Computing field, the CE method has been often shown to perform better than feature fusion strategies [36]. In this work, we combine the predictions of the classifiers trained with each modalities, using the *soft-voting* approach, which often gives better performance and it is more flexible than the *hard-voting* one [66]. We combine the predicted probabilities of the different classifiers using the median.

Shallow classifiers. We tested the performance of a popular shallow classifier i.e., gradient boosting (GB)[5]. We chose this classifier because of its demonstrated high performance and the possibility of getting a class probability [66]. In the training phase we scaled the features using the z-score normalization and to account for the class imbalance, we applied the SMOTE algorithm [83].

Convolutional neural network. We developed a CNN which takes as input the raw physiological signals processed with an instance normalization layer [278]. The CNN is composed by a *convolutional base (CB)* followed by a *fully connected network (FCN)*. The CB is composed by three *convolutional layers (CL)* (64 filters in the first two and 32 in the last layer, kernel size of 10) connected through an average pooling layers (window size of 5) that reduce the dimensionality of the input [31]. For the CLs we used a *parametric leaky ReLU (PReLU)* activation function. We further used a normalization layer that normalizes the activations of the previous layer of each example in the batch independently. The features extracted from the CB are aggregated using a *global average pooling* layer[6]. The context features are concatenated to the output of this layer and the whole vector is then processed by the FCN composed of two *fully-connected layers (FCL)* of 256 and 128 hidden units and *ReLu* activation function. To reduce the risk of overfitting, we used a dropout layer with dropout rate of 0.2 between the two FCL, used a *L2* regularization to all the layers with a rate of 0.0001 and early stopping when the validation error reached the minimum. The batch size is set to 64. To take into account the class imbalance we assigned class weights inversely proportional to the amount of samples per class in the training set. The aim of this procedure is to penalize missclassifications on the underrepresented class to

[5] We use the *Extreme Gradient Boosting (XGBoost)* implementation `https://xgboost.readthedocs.io/en/latest/python/python_api.html`

[6] In case of the SB-LF, the features extracted from each of the branches are first processed with the global average pooling layer and the concatenated

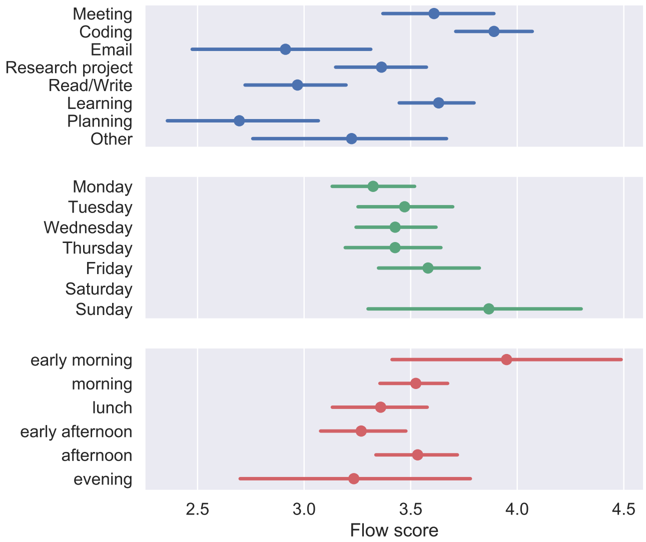


Figure 5.7. Mean and 95% confidence interval to show the relation between context data, in terms of type of activity, day of the week and time of the day, and self-reported flow score.

make the network "pay more attention" to the samples belonging to this class (i.e, the high_flow class) by assigning higher weights to the cost function in their correspondence. We used the *Adam* optimizer with a learning rate of 0.001 which is unaffected by the changed range of the loss introduced by the class weighting. We selected the hyperparameters of the CNNs based on previous work [192]; previous experience; and empirically. The output of the model is provided as input to a *sigmoid* function which returns a value between 0 and 1 that can be interpreted as the probability of the sample to belong to the positive class.

Baselines. We compared the performance of the before-mentioned approaches with a *biased random guess (BRG)*. We also considered models based on single sensors to test the advantages of using fusion approaches and models that have as input context data only to understand the impact of the combination with physiological data.

Evaluation procedure and performance metrics. We used a 5-fold stratified cross-validation approach guaranteeing that the data of the same participant of a given day was not concurrently in the train and test as in [279]. This approach tests the ability of the model to generalize to data of activities performed on unseen days.

We evaluated the performance using the *balanced accuracy (BA)* defined as "the average accuracy obtained on either class" [85]. The BA is a combination of the *sensitivity* and *specificity*, as reported below:

$$BA = \frac{Sensitivity + Specificity}{2} \tag{5.1}$$

Differently from the classical accuracy, the BA is more suitable for evaluating the performance of imbalanced data sets since it takes in consideration also the unrepresented class. The BA is identical to the traditional accuracy if the model performs well on both classes, but it reaches random chance (50%) if the classifier takes advantage only of the imbalance in the two classes for the prediction [85]. We further report the F1 metric obtained for each of the two classes.

We report the performance on activity-basis obtained by computing the median of the probabilities of the windows in which the activity is split. A threshold of 0.5 is then used to assign the sample to the *low_flow* class, this approach is suitable for engagement-aware systems that provide retrospective feedback at the end of the activity. We also report performance on a window-basis, to consider systems that provide real-time interventions.

5.4.1 Results and discussion

In this section we report and discuss the results obtained. We first analyse the relation between self-reported flow and type of context. Then, we analyse the impact of different sensors, contextual information and fusion techniques for the recognition of flow. Lastly we analyze the relation between flow and satisfaction with own performance.

Relation between context and self-reported flow

Results from the analysis described in Section 5.4 show that the type of activity is a predictor of flow. Figure 5.7 shows the variation of the flow score in relation to the type of activity, day of the week and time of the day. Overall, the level of flow is the highest during coding activities. Significant[7] pairwise differences emerged between coding and research project (t = -0.39, p-value < 0.006), read/write (t = -0.52, p-value < 0.006) and other (t = -0.88, p-value < 0.001).

The fact that the *coding* activity is generally more inductive to flow aligns with prior findings [132]. Given participants being researchers from the Computer Science department, their skills might match with the cognitive challenge induced by the *coding* activity thereby promoting flow.

[7] $\alpha = 0.05$, the adjusted p-value based on Bonferroni correction is 0.006

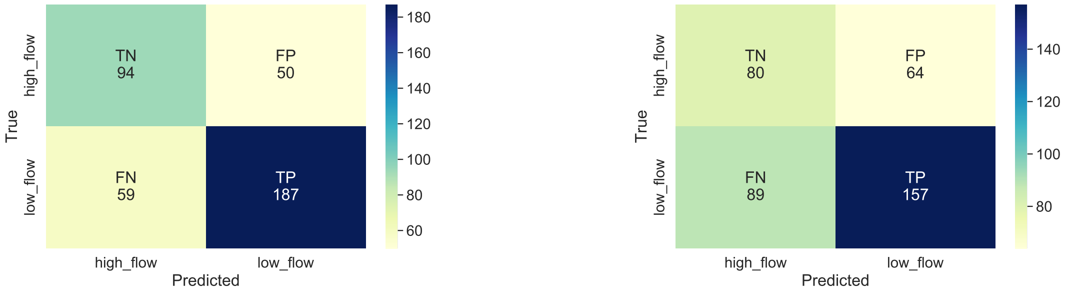


Figure 5.8. Confusion matrix for SB-LF model with (left) and without (right) context.

No significant associations are identified between the variables that quantify the time of the day and the day of the week with the perceived flow. This might due to the fact that workers have different habits and perceive themselves to be more productive in different times of the day [49], this variability across subjects might be reflected as well on the perceived flow.

Comparison of context information and single sensors

Table 5.4 reports the performance of shallow and deep learning models trained with single sensors and different types of context information. In all the cases, except for when EDA is used as input to GB, the type of activity is the most informative context information, in line with the results reported before.

The combination of the type of activity with physiological signals lead to an improvement of the BA of 4 *percentage points (pp)*, $F1_{low}$ of 6 pp and $F1_{high}$ of 5 pp, on average, compared to when no context is considered. We observe that the BVP signal is more informative than EDA and that the performance of the GB and the CNN are similar when this signal is used as input. Interestingly, using only the type of activity as input to the classifiers, the performance in terms of BA are similar to those obtained when combining it with BVP. However, the combination with BVP results in an high improvement (about 10 pp) of $F1_{low}$.

These results hint to the need of combining physiological data together with context information, in particular the type of activity, for improving the recognition performance. Given the relevance of the type of activity, also based on the results presented before, we continue the discussion using that as context information.

Model	**Modality**	**Context type**	**BA**	**F1$_{low}$**	**F1$_{high}$**
CNN	EDA	None	52.35 (3.63)	46.48 (16.38)	48.55 (4.57)
		T	55.81 (5.82)	61.07 (9.11)	48.39 (6.79)
		A	**60.15 (2.27)**	61.62 (4.67)	54.04 (3.09)
		T + A	58.75 (8.76)	65.55 (8.37)	49.98 (11.02)
GB	EDA	None	59.04 (8.18)	70.25 (2.92)	46.92 (14.4)
		T	56.81 (2.09)	68.68 (8.55)	43.47 (7.38)
		A	60.96 (6.13)	71.69 (4.31)	49.81 (9.93)
		T + A	**62.91 (3.18)**	74.00 (4.77)	51.14 (7.84)
CNN	BVP	None	64.25 (3.48)	71.47 (4.17)	55.74 (4.50)
		T	56.90 (8.28)	64.06 (8.79)	48.69 (9.30)
		A	**67.75 (4.02)**	74.69 (3.10)	59.86 (6.14)
		T + A	62.49 (5.62)	71.17 (6.79)	53.08 (8.53)
GB	BVP	None	61.07 (6.50)	69.27 (7.26)	51.95 (7.00)
		T	54.46 (5.10)	66.10 (8.27)	42.43 (6.98)
		A	**67.46 (3.98)**	75.23 (4.37)	59.45 (5.01)
		T + A	59.58 (4.16)	72.36 (3.54)	46.82 (5.55)
FCN	Context	T	50.97 (8.11)	54.48 (10.00)	44.40 (7.88)
		A	**64.84 (5.09)**	63.03 (7.73)	59.79 (5.69)
		T + A	58.91 (4.92)	64.77 (6.34)	50.86 (6.28)
GB	Context	T	49.77 (6.59)	52.07 (8.46)	44.00 (6.55)
		A	**65.91 (4.01)**	64.81 (6.41)	60.56 (5.37)
		T + A	58.08 (7.22)	66.42 (7.31)	48.95 (7.69)
BRG	-	-	51.39 (5.16)	72.74 (2.34)	24.46 (9.71)

Table 5.4. Performance results of shallow and deep learning classifiers trained on single sensors and with different type of context information. Performance are reported on activity-basis and averaged across cross-validation iterations, standard deviation in parenthesis. A stands for activity, T for time.

Model	Fusion strategy	BA	F1$_{low}$	F1$_{high}$
CNN	EF	67.60 (2.53)	74.47 (1.37)	59.58 (4.47)
	SB-LF	**70.93 (4.80)**	**77.71 (4.94)**	**63.15 (7.16)**
	CE$_{CNN}$	64.81 (6.10)	76.93 (3.34)	57.33 (4.81)
GB	FC	62.58 (7.16)	73.54 (1.94)	51.05 (11.83)
	CE$_{GB}$	67.93 (3.68)	71.71 (4.04)	61.3 (4.36)

Table 5.5. Performance results of different fusion strategies based on shallow and deep learning classifiers. Performance are reported on activity-basis and averaged across cross-validation iterations, standard deviation in parenthesis.

Comparison of fusion strategies

Table 5.5 reports the performance of the different fusion strategies investigated. The best performance are obtained when using the SB-LF approach, to which we refer as best model, with a BA of 70.93% (19 pp higher compared to BRG), F1$_{low}$ of 77.71% (5 pp higher compared to BRG) and F1$_{high}$ of 63.15% (39 pp higher compared to BRG). This indicates the importance of processing separately the sensors before concatenating them in a single vector. Further, we observe that *feature fusion* strategies based on CNN perform better than the FC, especially in terms of F1$_{low}$, 8 pp higher in case of EF and 12 pp in case of SB-LF). In this case, deep learning methods are able to better use the complementary information provided by the different sources and create a more efficient representation at the feature level. Regarding the *classifiers ensemble* strategies, the performance between CE$_{GB}$ and CE$_{CNN}$ are similar, with CE$_{GB}$ performing better in the correct identification of F1$_{high}$, 4 pp increment, while CE$_{CNN}$ performs better in terms of F1$_{low}$, 6 pp increment. Compared to when using a single sensor, in particular BVP, the SB-LF achieves an improvement of BA of about 4 pp, F1$_{high}$ of about 2 pp and F1$_{low}$ of about 3 pp higher. This indicates the importance of using the fusion of complementary information to improve the performance.

Figure 5.8, shows the confusion matrices of the SB-LF model with and without the activity type in input. We observe that when the activity type is used, the flow level is correctly identified in most of the activities. When the activity type is not included, a larger number of activities in low flow are miss-classified as high flow (89 instead of 59). These miss-classifications prevent systems that aim to reduce obstacles to flow to know in which conditions it is better to intervene, by for example blocking distractions or sending suggestions. Not including the type of activity causes also an increment of the false positives (FPs), when high flow are predicted as low flow, (64 instead of 50). A large number of FPs prevent the system to correctly identify conditions of high flow and wrongly suggest activities that are not conducive to flow.

We further test the performance of the best model in the recognition of flow levels in single five-minutes windows, to enable real-time recognition of flow. When using this approach we obtain a BA of 66.71%, $F1_{low}$ of 73.26% and $F1_{high}$ of 58.97%. The performance are lower compared to when considering the whole activity of about 4 pp per metric. Further investigation is needed to improve the in-the-moment recognition of flow.

Results obtained by the best model, are comparable to those presented in [47]. The authors achieved an accuracy of 70.6% using HRV features as input to a Random Forest classifier, however, the size of the data set used in that work is significantly smaller than the one we used (134 5-minutes segments against 390 activities for a total of 284 hours).

The size and the diversity of our data set makes our approach and results more robust and generalizable.

Relation between flow and satisfaction with performance

We analyze the relationship between the flow score and the self-reported satisfaction with performance. To take into account the correlated data, the correlation of the samples due to the multiple self-reports from the same individuals, we use the *repeated measures correlation (rmcorr)* technique [280]. The rmcorr is a statistical technique used to identify the relationship between two continuous variables by taking into account the effect of a categorical variable (the different individuals in our case) using a form of analysis of covariance (ANCOVA). In rmcorr different parallel lines are fitted per each subject and the sign of the test is given by the direction of the common slope. We implemente the rmcorr using the Python package pingouin[8]

The rmcorr test returned a positive significant correlation ($p < 0.0001$) with a coefficient of 0.47, indicating that the higher the level of flow workers experience during the activity, the higher the satisfaction with their own performance. This result hints to the importance of designing engagement-aware systems that enable workers to experience high levels of flow with the goal of increasing their satisfaction with performance.

[8] https://pingouin-stats.org/generated/pingouin.plot_rm_corr.html

5.4.2 Summary of the main findings in the recognition of flow using context and physiological data

Results presented in this section indicate overall, that it is feasible to automatically identify flow levels using context and physiological data. The main findings from our analysis are:

- The type of activity resulted to be a predictor of flow while the time of the day and the day of the week not. In general, the knowledge workers we monitored were more likely to be in flow during coding activities compared to when performing other type of activities.
- The type of activity is also the context information, that combined with physiological data, enables the highest increment in the flow recognition performance. Compared to when no context is used, using the type of activity allows an average increment of: BA of 4 pp, F1_{low} of 6 pp and F1_{high} of 5 pp.
- When using data from a single sensor, the BVP resulted to be more informative than EDA.
- The best performance, BA of 70.93% (19 pp higher than BRG), are achieved when raw BVP and EDA are combined together with the type of activity and used in input to a sensor based late fusion strategy implemented using a CNN.

5.5 Automatic recognition of work and break activities using a multi-sensor approach

To investigate the feasibility of distinguishing breaks and work activities, we define the problem as a binary classification task. We use a supervised approach and classify the logged activities into *work* and *break* classes. In the remainder of this section, we describe the pre-processing of sensor data, the classification pipeline, and discuss the results obtained.

Sensor data pre-processing and feature extraction

We removed noise from sensor data using channel-specific pre-processing techniques. We segmented the traces using a five-minute, non-overlapping sliding window, and extracted 106 representative features from EDA, ACC, BVP, laptop application usage and phone usage. We include also the hour of the day as in [4]. The list of all the features used in this work is reported in Table 5.6.

For this analysis, we chose a window of five minutes in line with previous work in this context [53; 200]. Furthermore, RescueTime provides application usage logs at the granularity of five minutes (the exact timestamps of the opening/closing of applications are not available). Thus, application usage could be not be reconstructed for time window smaller than five minutes.

Wristband (Wr). We pre-processed EDA and BVP data using a similar procedure to the one presented in Section 5.4.1. However, in this analysis we did not use the *EDArtifacts* tool for processing EDA due to this analysis being performed after the final implementation of the tool. Features from EDA were extracted using the same methods presented in Section 5.4. From the BVP data we extracted 15 features derived from the literature [76; 77; 78; 79] such as time-domain statistical features, HR, time-domain HRV statistical features and features that describe the BVP pulses. To gather information about participants' variation in physical movement we first calculate the magnitude of acceleration using the Euclidean Norm Minus One as in [281] and then extract seven statistical features [29].

Phone (Ph). We characterized phone usage by extracting the number of times the user unlocks the phone or turns on the screen.

Laptop (La). To data about laptop usage we first re-sample the data to account for missing RescueTime data, which signals no user interaction with the laptop, and we thus label these segments as "laptop inactivity". For each Web site visit and application usage, RescueTime logs: name of the Web-site/application, subcategory or category [9]. To characterized laptop application usage, we considered for how long specific application categories are used within a window and extracted 15 representative features including, e.g., the total time each category of applications is used and the number of categories used (as hint to multi-tasking). To further account for laptop inactivity during the window we subtracted from the total duration of the window the total duration of laptop usage in the window. We considered application categories rather than individual applications to keep a higher level of abstraction, as done in [61] and also to reduce the number of dimensions of the feature space.

[9] The list of categories and sub-categories is presented here: `https://www.rescuetime.com/categories`. Note that a valid RescueTime account is needed to access this list

Sensory channel	**Features extracted**
EDA (mix-EDA + Phasic + Tonic) **(68 features)** **From: [67; 68; 73; 74; B; 75]**	**Time-domain statistical features:** minimum, maximum, mean, median standard deviation, dynamic range, variance,slope, mean and standard deviation of the first and second derivative; **Wavelets features:** mean, standard deviation of wavelets coefficients at 1Hz, 2Hz and 4Hz; **SCR features:** number of peaks, peaks' amplitude, rise time, half-recovery time, width and area under the curve, maximum derivative of SCR;
BVP **(13 features)** **From: [76; 77; 78; 79; B]**	**Time-domain statistical features:** mean, standard deviation, number of peaks, dynamic change; **Heart rate (HR);** **HRV statistical features:** mean of all NN, standard deviation of all NN (SDNN), standard deviation of differences between NN (SDSD), the square root of the mean of the sum of the squares of differences between NN (RMSSD); **BVP pulse features:** mean and standard deviation of pulses' length and amplitude;
ACC **(7 features)** **From: [29; B]**	**Time-domain statistical features:** mean, standard deviation, dynamic range, slope, absolute value of the slope, mean and standard deviation of the first derivative;
Phone usage **(2 features)** **From: [271; 282]**	Number of unlock, number of screen on.
Laptop application usage **(15 features)**	**Duration of category of application:** Social Networking, Utilities, Uncategorized, Software Development, Shopping, News and Opinions, Entertainment, Learning and Reference, Design & Composition, Reference & Learning, Communication & Scheduling, Business, Non used, Most used category, Total duration of the laptop usage, Number of categories used.

Table 5.6. Summary of the 106 features extracted from the five sensory channels used in the activity recognition pipeline presented in Section 5.5. NN stands for the distance between BVP adjacent peaks. In addition to the features derived from the sensors we consider the hour of the day as in [4].

Activity recognition pipeline

We report below details about the activity recognition pipeline we designed and implemented.

Labeling and cleaning. We assigned each five-minutes window to the *work* and *break* class using the available self-report data. We assigned each window to one of the two classes if at least 60% of the window belongs to a logged activity (whose start and end were recorded with the timer), similar to [283].

We assigned the label *work* if one of the eight work activities categories was selected and *break* if the “break" label was selected. To ensure the presence of sufficient data of both classes in the test and train sets, we considered for the analysis only data of the participants who logged more than one break, i.e., nine participants. We further discarded 12 activities due to RescueTime installation errors. From this procedure we obtained an imbalanced data set with 4008 instances (an instance corresponds to a five-minutes window), which map to 449 activities logged by the study participants. Of the obtained 4008 instances, 3387 are in the *work* class and 621 in the *break* class.

Imputation. The subset of the data set used in this study has in total 279 instances in which wristband sensor data is missing, probably due to the user not wearing the device. To impute such values in the training phase, we considered the features corresponding to missing values as the target of a regression model (Bayesian ridge regressor) and used the remaining features as input to the model [72]. We then applied the same technique in the test set as recommended in [72]. We implemented this imputation strategy since it is likely to provide more accurate results than the replacement with statistical or constant values [72]. For the implementation we used the *IterativeImputer* class from scikit-learn[10]

Classifiers. We explored several classifiers such as Support Vector Machines [65], Logistic Regression [65], Gradient Boosting [284] and Random Forests [65]. The gradient boosting classifier – instantiated using the XGBoost Python implementation [284] with the default parameters (i.e., learning rate of 0.1, max depth of 3 and 100 estimators) [11], which is used also in [60] – resulted to be the best performing classifier. The results of all classifiers are reported in [D].

Baselines. We compared the performance of the classifiers described above to that of three rule-based classifiers: a *biased random guess classifier (BRG)*, a *time-based predictor (TB)* and a *laptop-inactivity-based predictor (LIB)*. TB considers as breaks all the activities between 12:00 a.m. and 01:00 p.m. (typical lunch time). LIB is based on the approach proposed by Tseng et al. in [3]. The main goal of

[10] https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html#sklearn.impute.IterativeImputer.

[11] https://xgboost.readthedocs.io/en/latest/python/python_intro.html

Classifier	Devices	Pr$_w$	R$_w$	F1$_w$	Pr$_b$	R$_b$	F1$_b$
XGBoost	La	93 (4)	90 (5)	92 (3)	56 (18)	67 (21)	59 (17)
	Ph	94 (4)	84 (4)	89 (4)	47 (13)	**74 (14)**	56 (13)
	Wr$_{EDA+BVP}$	94 (4)	94 (7)	94 (5)	59 (9)	71 (20)	63 (12)
	La + Wr$_{EDA+ACC}$	94 (4)	94 (4)	94 (3)	68 (14)	69(19)	67 (15)
	La + Ph	93 (4)	91 (4)	92(3)	57 (20)	65 (22)	59 (17)
	Ph + Wr$_{ACC+EDA+BVP}$	93 (4)	92 (5)	92 (4)	61 (9)	65 (17)	62 (10)
	La + Ph +Wr$_{\mathbf{EDA+ACC}}$	**94 (4)**	**95 (3)**	**94 (3)**	**71 (11)**	69 (18)	**69 (13)**
BRG	-	84 (6)	84 (3)	84 (3)	16 (8)	15 (5)	15 (6)

Table 5.7. Performance of the automatic recognition of breaks and work activities reported as percentage of mean (standard deviation) of the LOSO iterations, for selected combinations of sensors. The metrics considered are precision (Pr), recall (R), and F1 score ($F1$), while the subscripts w and b indicate the work and break class. Ph stands for phone, La for laptop and Wr for wristband.

the work presented in [3] is to prevent *cyberloafing*, i.e., the use of the Internet for non-work purposes. To this end, Tseng et al. implemented the UpTime system that automatically blocks distracting websites when detecting workers to be back from breaks. UpTime considers as breaks moments in which laptop inactivity is detected for at least five minutes [3]. Similarly, our implementation of the LIB classifies a window as break if the laptop records no user input in the five-minutes window.

Metrics. To test the ability of the classifiers to separate the two classes, we computed the *precision* (Pr), *recall* (R), and *F1 score* ($F1$) – all defined as in [65] – for both the *work* and *break* class. We thereby considered each five-minutes window as a single data point. To reason about what causes classification errors, we reported the confusion matrix.

Features selection. To reduce the dimensionality of the feature space, we used the Kolmogorov-Smirnov (KS) non-parametric test as in the study presented in Chapter 4 and also used in [82].

Validation procedure. To train and test the models we used the *leave-one-subject-out (LOSO)* validation approach. During the training phase, we imputed missing values, re-scaled the features using the z-score normalization calculated from the training samples (different participants and activities) [72], performed feature selection, and – to account for the imbalanced nature of our data set we used the SMOTE algorithm [83].

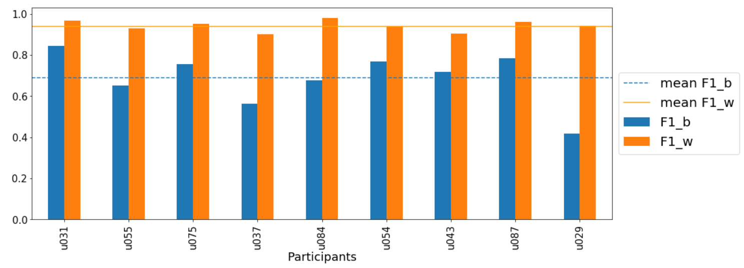


Figure 5.9. Performance for each user when in test set, w stands for work and b for break. Performance, in terms of F1 are consistent across users for the work class. In the break class most of the user presents similar performance to the mean, while two u037 and u029 present lower performance.

5.5.1 Results and discussion

We report in the followings the results obtained using the classification pipeline mentioned above. We discuss how the classification performance vary depending on the sensors considered as input and we report about the outcomes of the feature selection process. We further compare the performance of supervised classifiers to that of rule-based classifiers.

Performance of different combination of sensors and devices

Table 5.7 reports the mean (standard deviation), computed over all iterations of the LOSO procedure, of each of the considered metrics for the best performing classifier (XGBoost) along with a baseline classifier (BRG). For XGBoost, each line reports the results obtained using features computed from specific subsets of sensor traces. Results for all combinations of sensors are presented in [D].

We observe that the best overall results are obtained when using features extracted from all devices but only the *EDA* and *ACC* sensors from the wristband ($La + Ph + Wr_{EDA+ACC}$). We hereinafter refer to this model as the *best model*.

The best model achieves a *F1* for the *work* class ($F1_w$) of 94% – which is 10 percentage higher than the $F1_w$ of *BRG* classifier – and *F1* for the *break* ($F1_b$) class of 69% – which is 54 percentage points higher than the $F1_b$ of the *BRG*.

Comparing the performance of the best model to the results obtained with other combinations of sensors and devices, we observe that using single devices results in similar performance for the *work* class – except for the case in which only phone features (Ph) are used ($F1_w$ is 89%, i.e., five percentage points lower than the best model).

Larger differences are observable when considering the *break* class: the $F1_b$

of the best model is consistently higher than that obtained using features computed from single sensors/devices – of 10 (for the laptop), 13 (for the phone), and six (for the wristband) percentage points –. This underpins our assumption that data from multiple sensors can better describe the characteristics of the breaks. We also observe that when combining the phone with either the wristband or the laptop the $F1_b$ is lower, of seven and 10 percentage points, respectively, compared to the best model. Instead, the best model and the one obtained when using the laptop and the wristband presents similar performance (only two percentage lower $F1_b$).

These results suggest that even though the phone usage data provides additional information to the model, it might not be as informative for the recognition of breaks as data captured from the laptop and wristband. This could be due to the fact that the phone is often used sparingly at work, both during breaks and work activities, or used in equal measure in both situations. This minor role of the phone data implies that users may also decline to let the system track their phone usage, and accept only a small reduction in performance.

Further, we observe that among the wristband's sensors, the *EDA* and *ACC* sensors guarantee the best performance, hinting at the need of including information about the workers' physiological arousal and movement to discriminate work from breaks. Figure 5.9 shows the performance of the model when each user is in the test set. We notice that for all the iterations for the *work* class and for most of the iterations (7) for the *break* class, the performance are consistent across users. This indicates an ability of the model to generalize to unseen users. However, for users *u037* and *u029*, the F1 score of the *break* class, is lower than the mean of 13 and 27 percentage points, respectively. This might be due to the these specific users having work habits significantly different to those of the other participants. This is in line with the observation by Epstein et al. [195], that there is a large subjective variability in the way users define and take breaks. The same activity, e.g., reading news, could indeed be categorized as work by some users and as break by others [195].

Errors caused by interpersonal variability when training user-independent models is a well-known problem in human activity recognition [6]. Further research is needed in this direction, e.g., considering the use of personal or hybrid models [285].

Significant features

Out of the 93 features used, only 89 features were selected at least once using the KS test in the training phase.

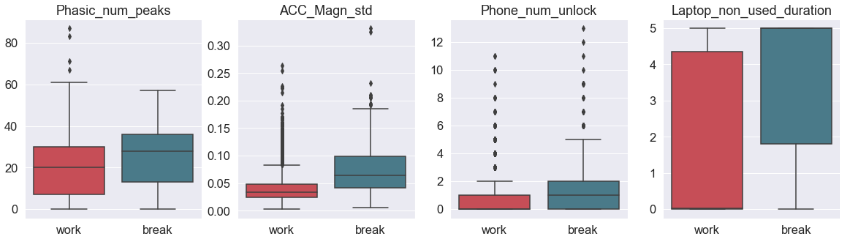


Figure 5.10. Distribution of a subset of the selected features in the two classes. The distribution of all these features in the two classes is significantly different ($p<0.05$) in all the training sets according to the Kolmogorov-Smirnov test.

The four features that are never selected are extracted from the laptop and are related to the duration of usage of the applications in the categories of *Social Networking*, *Entertainment*, *Shopping* and *News*. Interestingly, these are the categories of applications usually included in *digital breaks* by previous work [200; 51]. The fact that these features are never selected means that their duration is never significantly different between work and break activities. We believe that there might be two possible reasons for this. First, RescueTime's application categories are assigned statically, independently of the context in which the applications are used [269], and the same application can be used both for work and personal purposes [269] (for example messaging applications to communicate with colleagues or friends). Second, workers might tend not to log short *digital breaks* either because unconsciously performed or because a general tendency of users to forget to log breaks [195].To alleviate such problems the system could learn, during an initial calibration phase, the habits of specific users and ask them to indicate whether they are taking a break, whenever the likelihood of a digital break taking place is high. A similar strategy has been used in [200], and it could help achieving further performance improvements.

Further, most of the features (61) are selected in all the iterations, while features belonging to the *mixed-EDA* and *tonic* component are selected less often. This implies that the *phasic* component, which reflects the physiological response to stimuli [35], has a higher discriminative role.

Figure 5.10 shows the difference in the distribution in the two classes of exemplary features. Phasic_num_peaks represents the number of peaks of the *phasic* component of the EDA signal and is used as a proxy of physiological arousal [35]. The ACC_Magn_std represents the standard deviation of the vector magnitude of the acceleration and thus captures physical movement. Phone_num_unlock identifies the number of unlock of the phone and is a proxy for phone usage [271].

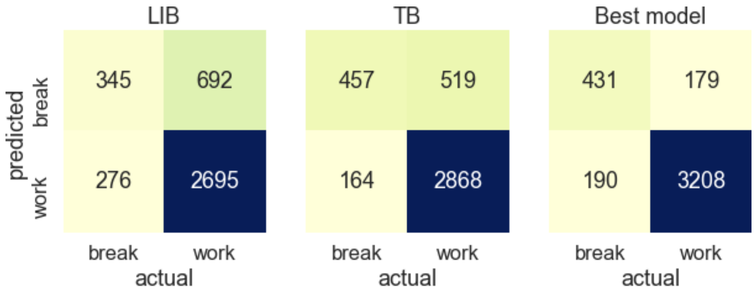


Figure 5.11. Comparison of the confusion matrices of our best model (using XGBoost), the time-based (TB) classifier and the laptop-inactivity-based (LIB) from [3]

Finally, Laptop_non_used_duration represents the percentage of the activities during which workers do not use the laptop. We observe that during break activities there is an increment of the physiological arousal, more frequent interactions with the phone, higher variability in physical movement, and a reduced usage of the laptop. This is in line with the general definition of breaks, which are characterized by workers spending less time at the laptop, changing their location or interacting with colleagues [195; 3; 56].

Classifier	**Pr**$_w$	**R**$_w$	**F1**$_w$	**Pr**$_b$	**R**$_b$	**F1**$_b$
XGBoost	94	**95**	**94**	**71**	69	**69**
TB	95	85	89	47	74	57
LIB	91	80	85	33	56	42

Table 5.8. Comparison of the performance metrics of our best model (using XGBoost), the time-based (TB) classifier and the laptop-inactivity-based (LIB) from [3]. Pr stands for precision, R for recall, while w and b respectively for work and break.

Comparison with rule-based classifiers

Table 5.8 shows the performance of the LIB classifier, which equals laptop inactivity to breaks as in [3], the TB classifier, which always predicts breaks at lunch time, and the best model (XGBoost with input $La + Ph + Wr_{EDA+ACC}$). We notice that the best model achieves better performance compared to the TB and LIB classifiers respectively of 12-27 pp for $F1_b$ and 5-9 pp for $F1_w$.

Figure 5.11 shows the confusion matrices of the considered classifiers. We observe that the LIB miss-classifies work activities often (692 instance, 513 more than the best model) as breaks (false negatives (FN)). The majority of these 692 FN belongs to the type of activity *learning* (201, 145 more than our model) and *meeting* (324, 280 more than our model), which are activities that do not strictly require the usage of the laptop [195; 3]. This confirms that, as also noted by Tseng et al. [3], using laptop inactivity only is a flawed heuristic to determine if a person is working or taking a break.

The TB baseline presents negligible lower number (164, 26 lower than the best model) of breaks predicted as work activities (false positive (FP)) compared to the best model. On the other side, it has a significantly higher number of FN (519, 340 more than the best model), corresponding to moments in which workers were working during the "typical" lunch time. This shows that using only the lunch time to infer a break also has severe limitations.

A system based on the TB and LIB has the advantage of being simpler (both the rule-based classifier could work only with the laptop) than our approach (which needs multiple devices and sensors). However, our approach is more robust and less prone to errors, especially in terms of FN. A large number of FN can have major consequences for personal informatics systems for promoting workers' productivity. For instance, it can make applications to be blocked too often [3], and, consequently, workers' disappointment.

5.5.2 Summary of the main findings in the recognition of work and breaks activities

With the results presented in this section, we demonstrated the feasibility of using data derived from personal devices to correctly distinguish between work and break activities. The main findings from our analysis are:

- The combination of data from different sources allowed an increment of the performance compared to when using a single sensor, especially in the recognition of breaks – $F1_b$ from 6 to 13 pp improvement.

- The best performance are obtained when using EDA data, along with acceleration, laptop and phone usage data, in input to the XGBoost classifier. With this model we achieved a F1 score of 69% for the recognition of breaks and 94% for the work activities.

- The best model, outperformed the rule-based classifiers based on time and laptop inactivity respectively of 12-27 pp for $F1_b$ and 5-9 pp for $F1_w$. Using the best model allowed in particular to reduce the possibility of missclassifying work activities as breaks.

5.6 Implications

Taken together, our findings open up new possibilities in the design and development of engagement-aware personal informatics systems for supporting knowledge workers in the workplace.

Personal informatics systems can leverage information about work and break activities to deliver interventions to promote workers' productivity and wellbeing. For instance, a break recommendation could be triggered when the system detects that the user has taken no breaks for multiple consecutive work windows [200; 199]; to prevent cyberloafing, distracting applications could be blocked when the system recognizes that the user is back from break (i.e., the window state changes from break to work) [3]; and workday summaries could be automatically populated to allow self-reflection [15; 53].

The automatic recognition of work and break activities could also serve as initial step for the automatic segmentation of the work day. For instance, once the work activity is determined (considered as consecutive 5-minutes window recognized as work), the system could interrogate the user about the type of activity performed until that moment. Once this information is available, the system could then "activate" then flow recognition engine and determine the level of flow of the participant during the activity.

Information about flow could be retrieved to the users at the end of the activity or day to allow self-reflection [28]. In this direction, the system could enable workers to identify patterns and habits that lead them to reach, or not, the flow state by using personal analytics [18].

Once several work activities are collected from the users, and corresponding flow levels are automatically inferred, the system could help users to schedule their work and suggest times or activities that are more conducive to flow based on learned patterns [286]. In a real-time scenario, when a prolonged state of low flow is inferred, the system could trigger interventions that reduce obstacle to flow e.g., by blocking distractions [198].

Further, a physical interruptability indicator as the *FlowLight* presented in [287], could be integrated in the engagement-aware system and used to indicate coworkers when the user is in flow and to avoid her to be interrupted.

5.7 Limitations

The contributions presented in this chapter show promising results in the recognition of *flow* during work activities as well as in distinguishing between *work* and *break* activities. However, our approach presents limitations that should be investigated in future work.

An important limitation of our study is that it relies on data collected for an average of 7 days. This short time frame does not allow us to investigate the impact of the novelty effect due to the use of new devices nor the level of fatigue caused by the prolonged usage of multiple self-report devices. More data is necessary to understand the long-term effectiveness of the proposed multi-device strategy.

For gathering ground-truth about flow and activities we rely on timer-based logging, self-initiated by the workers. This strategy avoids disruptions to the users' workflow (which would otherwise be caused by sending notifications as reminders to log activities) and also allows workers to report with full flexibility when and for how long they were working or taking a break. A limitation of this strategy consists in its inability to capture *digital breaks*. Indeed, workers might take short *digital breaks* during their work activities and not log them explicitly, either because not consciously considered as breaks or because of the general tendency of workers to forget to log breaks [195]. To overcome this limitation, researchers could design systems that actively inquire users when specific digital activities are detected in order to better understand their role in the overall workflow.

With the results presented in Section 5.4.1, we showed the importance of integrating information about the type of activity for improving the performance of the automatic flow recognition. However, a limitation of our approach consists in using the answers to self-reports to derive this information. In general, allowing users to complement the automatic assessment of the system with manual inputs could be beneficial for improving the effectiveness of self-tracking [16], it could increase users' awareness and thus the likelihood of behavioural change [16]. However, the need of the model of knowing the type of work activity might be problematic for long-term usage. The type of work activity could be derived from personal calendars, to-do lists, or, automatically recognized using sensors. In the second contribution of this chapter, presented in Section 5.5, we showed the possibility of distinguishing between work and break activities, however, due to the complexity of the work environment, the automatic recognition of the type of work activity is still an open challenge. We believe this is an important problem to be targeted in future work.

Another limitation regards the labeling procedure, we used the same self-reported *flow score* to label all the segments of an activity while the flow state might change over time. This is a known problem as also reported by Lee at al. [41]. As we also discussed in the study presented in Chapter 3, obtaining continuous labels about engagement or flow during an activity is however impractical, because it would require to continuously ask workers (or students) to provide flow labels. This would interrupt individuals' and thus prevent them from sustaining flow during the activity. Understanding when to ask workers to enter labels – and thereby disrupting their activities – is still an open challenge that should be targeted in future work aiming to collect data sets for flow recognition.

Lastly, even though the data set we used contains a large and heterogeneous set of activities, it is obtained by 13 participants only. Future studies, carried with higher number of subjects, are needed to further validate our approach.

5.8 Summary of Chapter 5

In this chapter we presented our observations and findings about the automatic recognition of knowledge workers' flow during work activities as well as about the inference of the activities in the workplace.

Building upon existing literature, we proposed to recognize flow during work activities using a combination of physiological data (i.e., electrodermal activity and blood volume pulse), and context information (i.e., type of activity, time of the day and day of the week) collected using self-reports.

To this end, we experimented with several fusion strategies implemented using shallow classifiers and convolutional neural networks.

We observed that the type of activity is a relevant context information that should be taken into account when recognizing flow during work activities.

Further, we confirmed the importance of using cardiac activity information (represented by the BVP) for the identification of flow during work activities as also discussed in the literature [47; 45].

Our results underlined also the potential of adding information about the physiological arousal (derived from the EDA signal) to have a more complete picture of the flow state and further improve the recognition performance. Indeed, the best performance, BA of 70.93%, is achieved when the raw BVP and EDA signals, together with the type of activity are used in input to a SB-LF strategy implemented using a CNN.

Besides recognizing flow, in this chapter we presented our method for recognizing activities in the workplace. Specifically, we focused on distinguishing work and break activities. For doing so, we proposed to characterize the physiological activation, the physical movement, the laptop and phone usage using sensor data collected using personal devices i.e., wristband, laptop and smartphone.

Concatenating features from the before-mentioned sensory channels, and using them in input to a gradient boosting classifier, we achieved a $F1$ score of 94% for the identification of work activities and of 69% for breaks.

We observed that combining sensor data allowed to improve the performance in comparison to when using a single sensory channel, especially for the correct recognition of breaks – $F1_b$ from 6 to 13 pp improvement. This indicates the importance of using a multi-sensor approach for better characterize work and break activities. Further, the best classifier (XGBoost), outperformed the rule-based classifiers based on time and laptop inactivity, reducing in particular the chance of misclassifying work activities as breaks. This indicates the potential of a machine learning approach for addressing the activity recognition problem in the workplace.

In this chapter we also discussed our experience with a multi-device strategy to collect self-reports during work activities. We found that workers tend to use multiple devices but have different preferred devices. A multi-device strategy thus seems able to accommodate users' needs. Thereby, we also found that *SSR* devices can be used to replace or complement mobile phones and laptops to gather self-reports in the workplace.

Taken together, our findings can open up new possibilities for the design of engagement-aware systems aiming to support knowledge workers during their daily activities.

Chapter 6

Conclusions and Outlook

In this thesis, we investigated how data derived from personal devices – such as wristbands, laptops and smartphones –, can be used to automatically recognize engagement and activities. We focused on the specific use cases of recognizing students' engagement during lectures and knowledge workers' engagement during work activities.

We conducted three user studies and collected three data sets: *Students Engagement Using EDA (SEED)*, *USI_Laughs* and the *WorkplaceDataSet*.

In the first user study, presented in Chapter 3, we investigated the use of electrodermal activity data, collected with unobtrusive wristbands, to infer students' engagement during lectures. Building upon findings from educational research, we identified three relevant components of students' emotional engagement: momentary engagement, reaction to the teacher, and emotional arousal. We further identified a set of features to be extracted from EDA signals that can characterize these components. We experimented with features used previously in the literature as well as proposed a set of novel features. We observed that the best performance, especially in the recognition of non-engaged students, are obtained when using two of the features we proposed: the *arousing_ratio* and *level_5*, representing the momentary engagement. Specifically, using these features as input to a SVM classifier, we achieved a recall of 81% which corresponds to an improvement of 25 percentage points with respect to a Biased Random classifier used as baseline.

Systems able to identify non-engaged students have several potential applications. They may enable teachers to devise and evaluate methods to (re-)engage students. Students, on the other hand, could use information about their own engagement, or lack thereof, to perform self-reflection and change behavior.

In the second users study, presented in Chapter 4, we investigated if and how accurately laughter episodes can be detected automatically from sensor data collected using wristbands. Specifically, we proposed to quantify laughter through a combination of physiological reactions (using electrodermal activity and cardiac activity data), and body movement reactions. We analyzed the impact of the combination of different data sources. We also tested the robustness of the proposed method against confounding variables such as cognitive load and physical movement.

We observed that laughter episodes, of medium and high intensity, can be reliably distinguished from non-laughter episodes. Specifically, when using features extracted from EDA, blood volume pulse (BVP) and accelerometer (ACC) as input to a SVM classifier, we achieved an accuracy of 81% This corresponds to 6-15 percentage point increment in performance with respect to the case in which data from a single sensor only is used. Further, we showed that the physiological responses generated during laughter episodes are different from the ones generated during tasks inducing a high cognitive load. Indeed, the episodes could be distinguished with an accuracy of 92%. Lastly, body movement reactions during laughs are different from the ones generated by other gestures such as clapping hands, indeed all the features extracted from ACC present a significantly different distribution in the two conditions.

Overall, these findings indicate the possibility of using physiological and body movement data to recognize laughter episodes. Once laughter episodes are recognized, they can be used as additional information for recognizing engagement of students and knowledge workers during activities such as lectures, business meetings or breaks.

In the third user study, discussed in Chapter 5, we analyzed the role of multiple devices to collect self-reports and sensor data from knowledge workers during work days. Further, we investigated how this data can be leveraged to recognize workers' activities and flow levels. We adopted a multi-device strategy to collect self-reports, allowing study participants to choose among a paper-based diary, and three digital devices equipped with a timer: a laptop widget, an Android application and a situated self-reporting device (SSR) called *Devo* we designed. We observed that participants had different preferences in the choice of the device but, overall, *Devo* resulted to be the most used device while the smartphone the least used.

Building upon findings from existing literature, we proposed to combine both physiological parameters and context information to infer flow during work activities. We investigated several fusion strategies based on classical machine learning and deep learning. We observed that combining raw EDA and BVP signals, together with the type of activity using the sensor based late fusion strategy, implemented using a convolutional neural network, allowed to achieve a balanced accuracy of 71%. This method reduced in particular the chance of miss-classifying low flow states in high flow states compared to when the type of activity was not included. This miss-classification error prevents a system aiming to promote flow, to perform appropriate actions such as blocking distractions, when prolonged low flow is assessed.

Besides recognizing flow during work activities, in the third study we also focused on the automatic recognition of work activities from sensor data. As a first step in this direction, we devised a novel method to distinguish between work and break activities. For doing so, we extracted and combined features that quantify cues such as the physiological activation, the physical movement, the laptop and phone usage. We used the features as input to machine learning algorithms.

Our results showed that features from EDA, ACC, laptop and phone usage as input to a gradient boosting classifier, allow to identify work activities with a F1 score of 69% and break activities with F1 score of 94%, outperforming baseline methods by 12-54 and 5-10 percentage points respectively.

Information about work and break could be retrieved to the user to allow self-reflection, or used by an engagement-aware system as an entry point for “activating” the flow recognition or the laughter recognition engine. For instance, when the system recognizes the presence of a work activity (in terms of time between two breaks, or multiple “work windows”), the system could interrogate the user about the type of activity she performed and use the provided information in combination with physiological data to infer the level of flow during the activity.

Summary of contributions

In summary, the main goal of this thesis is to devise methods for recognizing engagement using data gathered from unobtrusive personal devices with the broad goal of contributing to the design of engagement-aware systems for supporting students and knowledge workers. To achieve this goal we addressed the following research questions:

- **RQ1** How can features representing behavioral expressions of engagement be derived from physiological and movement data?

- **RQ2** How can information about context a) be fused with physiological data and b) impact the recognition of workers’ engagement?

- **RQ3** How can activities in the workplace be accurately recognized?

The main outcomes of this thesis are:

- A set of theoretically-motivated EDA features that can be used as proxy for recognizing students' engagement during lectures – related to **RQ1**.
- A novel method based on the combination of physiological and body movement data to recognize laughter episodes – related to **RQ1**.
- A thorough investigation of the impact of different fusion strategies, based on traditional machine learning and deep learning, to combine context and physiological data to infer workers' flow during activities – related to **RQ2**.
- An automatic, multi-sensor approach based on data collected from personal devices to recognize work and break activities – related to **RQ3**.

6.1 Limitations and possible directions for future research

In this section we summarize the main limitations of the work done in this thesis and outline potential directions for future research.

Physiological data as proxy of engagement. In this thesis we extensively leveraged physiological data. From the results obtained in our studies, we conclude that, when adequately processed, physiological parameters represent valid proxies for engagement. However, from the data we collected as well from findings from existing literature, we observed that despite individuals present similar physiological reactions under specific circumstances (e.g., presence of EDA peaks when they laugh), large individual differences are also present (e.g., variability in the EDA intensity or lack of peaks when expected). This inherent intra-subject variability limits the transferability of the model parameters and often causes a reduction in the performance of user-independent models. We believe that future research should focus on understanding how to deal with this individuality and investigate strategies for taking advantage of it. For instance by creating personalized or hybrid models in which data of single or similar individuals are used for training specialized models.

Using multiple sensors and devices. In the work done in this thesis we often relied on the combination of multiple data sources for assessing individuals' level of engagement and the performed activities. We used both manual entries from study participants (e.g., self-reports) as well as sensor data. Confirming existing literature, also in our studies we observed an overall improvement of the performance of the models when combining data. This is due to the complementary information provided by the different inputs. However, we also notice that not always more sensors brought more information to the model while they could rise privacy concerns. We also experimented with a multi-device strategy to collect self-reports, as reported in Chapter 5. This strategy allowed to adapt to study participants' preferences and needs. Despite the advantages of using multiple tools, asking people to charge, wear and use multiple devices can add significant burden in the long-term. This can cause study participants to drop the study and users stop interacting with the system. We believe that future research should focus on guaranteeing a trade-off between invasiveness and completeness of information. Engagement and activity recognition systems should focus on relevant information only and be able to work also in presence of missing data. Further, researchers and practitioners should allow users to decide which tools to use and when to use them, but still guarantee valid assessments.

6.2 Concluding remarks

The technical contributions and insights presented in this thesis show the potential and feasibility of using data derived from mobile and wearable devices to infer the engagement of students and knowledge workers during learning and work activities. Taken together, our findings open up novel possibilities for the design and development of engagement-aware personal informatics systems for supporting students and knowledge workers.

Appendix A

Publications

Full list of publications of Elena Di Lascio (as of August 2021).

Journal articles

- E. Piciucco, **E. Di Lascio**, E. Maiorana, S. Santini, P. Campisi. *Biometric Recognition Using Wearable Devices in Real-life Settings*. In: Pattern Recognition Letters. March 2021. 8 pages.

- **E. Di Lascio**, S. Gashi, J. S. Hidalgo, B. Nale, M. Debus, S. Santini: *A Multi-sensor Approach to Automatically Recognize Breaks and Work Activities of Academic Knowledge Workers*. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 3, Issue 3, September 2020. 21 pages.

- S. Gashi, **E. Di Lascio**, B. Stancu, V. Das Swain, V. Mishra, M. Gjoreski, S. Santini: *Detection of Artifacts in Ambulatory Electrodermal Activity Data*. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 4, Issue 2, June 2020. 31 pages.

- S. Gashi, **E. Di Lascio**, S. Santini: *Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members*. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 3, Issue 1, March 2019. 19 pages.

- **E. Di Lascio**, S. Gashi, S. Santini: *Unobtrusive Assessment of Students' Engagement During Lectures Using Electrodermal Activity Sensors*. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 2, Issue 3, September 2018. 21 pages.

Book chapters

- S. Gashi, **E. Di Lascio**, and S. Santini: *Multi-class Multi-label Classification for Cooking Activity Recognition*. In: Smart Innovation, Systems and Technologies Series of Springer Books, Human Activity Recognition Challenge (ABC2020), August 2020. 13 pages.

Conference papers

- **E. Di Lascio**, S. Gashi, M. E. Debus and S. Santini. *Automatic Recognition of Flow During Work Activities Using Context and Physiological Signals.* To appear in: Proceedings of 9th International Conference on Affective Computing & Intelligent Interaction (ACII). 2021. 8 pages.
- S. Gashi, A. Saeed, A. Vicini, **E. Di Lascio** and Silvia Santini. Hierarchical Classification and Transfer Learning to Recognize Head Gestures and Facial Expressions Using Earbuds. To Appear in: Proceedings of ACM ICMI 2020. 8 pages.
- **E. Di Lascio**, S. Gashi and S. Santini: *Laughter Recognition Using Non-invasive Wearable Devices*. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), May 2019. 10 pages.

Workshop papers, demonstrators and other contributions

- A. Fedosov, B. Stancu, **E. Di Lascio**, D. Eynard, M. Langheinrich: *Movie+: Towards Exploring Social Effects of Emotional Fingerprints for Video Clips and Movies*. In: Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI 2019). 4 pages.
- S. Gashi, **E. Di Lascio**, S. Santini: *Using Students' Physiological Synchrony to Quantify the Classroom Emotional Climate*. In: Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2018). Best Paper Award. 4 pages.
- **E. Di Lascio**: *Emotion-Aware Systems for Promoting Human Well-being*. In: Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2018). 6 pages.
- D. Kumar, **E. Di Lascio**, M. Ahmad, M.Yliniemi: *Calmify: Measuring the Effectiveness of Personalized Meditation Techniques Using Mobile Technologies*. In: Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2018). 4 pages.
- **E. Di Lascio**, S. Gashi, D. Krasic, S. Santini: *In-classroom Self-tracking for Teachers and Students: Preliminary Findings from a Pilot Study*. In: Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2017). 6 pages.

Own publications reporting work described in this PhD Thesis

[A] E. Di Lascio, S. Gashi, S. Santini. Unobtrusive Assessment of Students' Engagement During Lectures Using Electrodermal Activity Sensors. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 2, Issue 3, September 2018.

[B] E. Di Lascio, Shkurta Gashi and Silvia Santini. *Laughter Recognition Using Non-invasive Wearable Devices*. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), May 2019.

[C] E. Di Lascio, S. Gashi, M. E. Debus and S. Santini. *Automatic Recognition of Flow During Work Activities Using Context and Physiological Signals.* Accepted for publication in: Proceedings of 9th International Conference on Affective Computing & Intelligent Interaction (ACII). 2021.

[D] E. Di Lascio, S. Gashi, J. S. Hidalgo, B. Nale, M. Debus, S. Santini. *A Multi-sensor Approach to Automatically Recognize Breaks and Work Activities of Academic Knowledge Workers*. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT), Vol. 3, Issue 3, September 2020.

Bibliography

[1] Corinna Peifer. Psychophysiological Correlates of Flow-experience. In *Advances in Flow Research*, pages 139–164. Springer, 2012.

[2] Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. Deep Learning for Smart Manufacturing: Methods and Applications. *Journal of Manufacturing Systems*, 48:144–156, 2018.

[3] Vincent W-S Tseng, Matthew L Lee, Laurent Denoue, and Daniel Avrahami. Overcoming Distractions During Transitions from Break to Work Using a Conversational Website-Blocking System. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2019.

[4] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing Detailed Human Context in the Wild from Smartphones and Smartwatches. *IEEE Pervasive Computing*, 16(4):62–74, 2017.

[5] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J Oedegaard, and Jim Tørresen. Mental Health Monitoring with Multimodal Sensing and Machine Learning: A Survey. *Pervasive and Mobile Computing*, 51:1–26, 2018.

[6] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *ACM Computing Surveys*, 46(3):1–33, 2014.

[7] Maria Faurholt-Jepsen, Jonas Busk, Maj Vinberg, Ellen Margrethe Christensen, Mads Frost, Jakob E Bardram, Lars Vedel Kessing, et al. Daily Mobility Patterns in Patients with Bipolar Disorder and Healthy Individuals. *Journal of Affective Disorders*, 278:413–422, 2021.

[8] Luca Canzian and Mirco Musolesi. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2015.

[9] Trinh Minh Tri Do and Daniel Gatica-Perez. Where and What: Using Smartphones to Predict Next Locations and Applications in Daily Life. *Pervasive and Mobile Computing*, 12:79–91, 2014.

[10] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. Monitoring Stress with a Wrist Device Using Context. *Journal of Biomedical Informatics*, 73:159–170, 2017.

[11] Javier Hernandez, Rob R Morris, and Rosalind W Picard. Call Center Stress Recognition with Person-specific Models. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2011.

[12] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, 4(3):1–26, 2020.

[13] Ian Li, Anind Dey, and Jodi Forlizzi. A stage-based model of personal informatics systems. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 557–566, 2010.

[14] Akane Sano, Paul Johns, and Mary Czerwinski. Designing Opportune Stress Intervention Delivery Timing Using Multi-modal Data. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 346–353, 2017.

[15] André N Meyer, Thomas Fritz, and Thomas Zimmermann. Fitbit for Developers: Self-Monitoring at Work. In *Rethinking Productivity in Software Engineering*, pages 261–270. Springer, 2019.

[16] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. Semi-automated Tracking: A Balanced Approach for Self-monitoring Applications. *IEEE Pervasive Computing*, 16(1):74–84, 2017.

[17] Mary Czerwinski, Javier Hernandez, and Daniel McDuff. Building an AI That Feels: AI Systems with Emotional Intelligence Could Learn Faster and be More Helpful. *IEEE Spectrum*, 58(5):32–38, 2021.

[18] Rafael A Calvo and Dorian Peters. *Positive Computing: Technology for Wellbeing and Human Potential*. MIT Press, 2014.

[19] Daniel McDuff and Mary Czerwinski. Designing Emotionally Sentient Agents. *Communications of the ACM*, 61(12):74–83, 2018.

[20] Rosalind Picard. Affective Computing. MIT press Cambridge, 1997.

[21] Jianhua Tao and Tieniu Tan. Affective Computing: A Review. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2005.

[22] Arnold B Bakker and Marianne van Woerkom. Flow at Work: A Self-determination Perspective. *Occupational Health Science*, 1(1):47–65, 2017.

[23] Sandra L Christenson, Amy L Reschly, and Cathy Wylie. *Handbook of Research on Student Engagement*. Springer Science & Business Media, 2012.

[24] Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sydney K. D'Mello. Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, 2017.

[25] Sandra L Christenson, Amy L Reschly, and Cathy Wylie. *Handbook of Research on Student Engagement*. Springer Science & Business Media, 2012.

[26] Arnold B Bakker. The Work-related Flow Inventory: Construction and Initial Validation of the WOLF. *Journal of Vocational Behavior*, 72(3):400–414, 2008.

[27] Gregory D Abowd, Anind K Dey, Peter J Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a Better Understanding of Context and Context-awareness. In *International Symposium on Handheld and Ubiquitous Computing*, pages 304–307. Springer, 1999.

[28] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. AffectAura: an Intelligent System for Emotional Memory. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 849–858, 2012.

[29] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. Wearable-Based Affect Recognition—A Review. *Sensors*, 19(19):4079, 2019.

[30] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. Personalized Machine Learning for Robot Perception of Affect and Engagement in Autism Therapy. *Science Robotics*, 3:19, 2018.

[31] Francois Chollet et al. *Deep Learning with Python*, volume 361. Manning New York, 2018.

[32] Mihaly Csikszentmihalyi. *Finding Flow: The Psychology of Engagement with Everyday Life.* Basic books, 1997.

[33] Karen S. McNeal, Jacob M. Spry, Ritayan Mitra, and Jamie L. Tipton. Measuring Student Engagement, Knowledge, and Perceptions of Climate Change in an Introductory Environmental Geology Course. *Journal of Geoscience Education*, 62(4):655–667, 2014.

[34] Chen Wang and Pablo Cesar. Physiological Measurement on Students' Engagement in a Distributed Learning Environment. In *Proceedings of the International Conference on Physiological Computing Systems (PhyCS)*, 2015.

[35] Wolfram Boucsein. *Electrodermal Activity*. Springer Science & Business Media, 2012.

[36] Philipp V Rouast, Marc Adam, and Raymond Chiong. Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Transactions on Affective Computing*, 2019.

[37] Gerhard Hagerer, Nicholas Cummins, Florian Eyben, and Björn Schuller. "Did you laugh enough today?" – Deep Neural Networks for Mobile and Wearable Laughter Trackers. In *Proceedings of the International Speech Communication Association Conference (INTERSPEECH)*, pages 2044–2045, 2017.

[38] Stavros Petridis and Maja Pantic. Audiovisual Discrimination Between Speech and Laughter: Why and When Visual Information Might Help. *IEEE Transactions on Multimedia*, 13(2):216–234, 2011.

[39] Boris Reuderink, Mannes Poel, Khiet Truong, Ronald Poppe, and Maja Pantic. Decision-level Fusion for Audio-visual Laughter Detection. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, 2008.

[40] Jérôme Urbain, Radoslaw Niewiadomski, Maurizio Mancini, Harry Griffin, Hüseyin Cakmak, Laurent Ach, and Gualtiero Volpe. Multimodal Analysis of Laughter for an Interactive System. In *Proceedings of the International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*, 2013.

[41] Matthew Lee. Detecting Affective Flow States of Knowledge Workers Using Physiological Sensors. *arXiv preprint arXiv:2006.10635*, 2020.

[42] Héctor P Martínez and Georgios N Yannakakis. Deep Multimodal Fusion: Combining Discrete Events and Continuous Signals. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 34–41. ACM, 2014.

[43] Mohammad Naim Rastgoo, Bahareh Nakisa, Frederic Maire, Andry Rakotonirainy, and Vinod Chandran. Automatic Driver Stress Level Classification Using Multimodal Deep Learning. *Expert Systems with Applications*, 138:112793, 2019.

[44] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. CNN-based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *Proceedings of the ACM International Symposium on Wearable Computers (ISWC)*, pages 158–165, 2017.

[45] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Michael Thomas Knierim, and Alexander Maedche. Got Flow? Using Machine Learning on Physiological Data to Classify Flow. In *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–6, 2018.

[46] Koji Ara, Nobuo Sato, Satomi Tsuji, Yoshihiro Wakisaka, Norio Ohkubo, Youichi Horry, Norihiko Moriwaki, Kazuo Yano, and Miki Hayakawa. Predicting Flow State in Daily Work Through Continuous Sensing of Motion Rhythm. In *International Conference on Networked Sensing Systems (INSS)*, pages 1–6. IEEE, 2009.

[47] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Nico Loewe, Michael T Knierim, and Alexander Maedche. To Be or Not to Be in Flow at Work: Physiological Classification of Flow using Machine Learning. *IEEE Transactions on Affective Computing*, 2020.

[48] Shugang Zhang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li. A Review on Human Activity Recognition Using Vision-based Method. *Journal of Healthcare Engineering*, 2017, 2017.

[49] Andre N Meyer, Laura E Barton, Gail C Murphy, Thomas Zimmermann, and Thomas Fritz. The Work Life of Developers: Activities, Switches and Perceived Productivity. *IEEE Transactions on Software Engineering*, 43(12):1178–1193, 2017.

[50] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. Bored Mondays and Focused Afternoons: the Rhythm of Attention and Online Activity in the Workplace. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 3025–3034. ACM, 2014.

[51] Harmanpreet Kaur, Alex C Williams, Daniel McDuff, Mary Czerwinski, Jaime Teevan, and Shamsi Iqbal. Optimizing for Happiness and Productivity: Modeling Opportune Moments for Transitions and Breaks at Work. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–15, 2020.

[52] André N Meyer, Chris Satterfield, Manuela Züger, Katja Kevic, Gail C Murphy, Thomas Zimmermann, and Thomas Fritz. Detecting Developers' Task Switches and Types. *IEEE Transactions on Software Engineering*, 2020.

[53] Saskia Koldijk, Mark van Staalduinen, Mark Neerincx, and Wessel Kraaij. Real-time Task Recognition Based on Knowledge Workers' Computer Activities. In *Proceedings of the European Conference on Cognitive Ergonomics*, pages 152–159, 2012.

[54] Nuria Oliver, Ashutosh Garg, and Eric Horvitz. Layered Representations for Learning and Inferring Office Activity from Multiple Sensory Channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

[55] Daniel Avrahami, Mitesh Patel, Yusuke Yamaura, and Sven Kratz. Below the Surface: Unobtrusive Activity Recognition for Work Surfaces Using RF-radar Sensing. In *International Conference on Intelligent User Interfaces*, pages 439–451, 2018.

[56] Scott A Cambo, Daniel Avrahami, and Matthew L Lee. BreakSense: Combining Physiological and Location Sensing to Promote Mobility During Work-breaks. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, 2017.

[57] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School Engagement: Potential of the Concept, State of the Evidence. 2004.

[58] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. Empatica E3—A Wearable Wireless Multi-sensor Device for Real-time Computerized Biofeedback and Data Acquisition. In *Proceedings of the International Conference on Wireless Mobile Communication and Healthcare (MobiHealth 2014)*, 2014.

[59] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2017.

[60] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, 3(2):37, 2019.

[61] A. N. Meyer, L. E. Barton, G. C. Murphy, T. Zimmermann, and T. Fritz. The Work Life of Developers: Activities, Switches and Perceived Productivity. *IEEE Transactions on Software Engineering*, 43(12):1178–1193, 2017.

[62] Stavros Petridis, Brais Martinez, and Maja Pantic. The MAHNOB Laughter Database. *Image and Vision Computing*, 31(2), 2013.

[63] Gaurav Paruthi, Shriti Raj, Seungjoo Baek, Chuyao Wang, Chuan-che Huang, Yung-Ju Chang, and Mark W Newman. Heed: Exploring the Design of Situated Self-Reporting Devices. volume 2, page 132. ACM, 2018.

[64] Michael Kipp. Anvil – A Generic Annotation Tool for Multimodal Dialogue. In *Proceeding of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001.

[65] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[66] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[67] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. Using Electrodermal Activity to Recognize Ease of Engagement in Children During Social Interactions. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2014.

[68] Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. Automatic Identification of Artifacts in Electrodermal Activity Data. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2015.

[69] Noura AlHinai. Introduction to Biomedical Signal Processing and Artificial Intelligence. In *Biomedical Signal Processing and Artificial Intelligence in Healthcare*, pages 1–28. Elsevier, 2020.

[70] Shkurta Gashi, Elena Di Lascio, Bianca Stancu, Vedant Das Swain, Varun Mishra, Martin Gjoreski, and Silvia Santini. Detection of Artifacts in Ambulatory Electrodermal Activity Data. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, volume 4, pages 1–31, 2020.

[71] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering*, 2016.

[72] Andriy Burkov. *The Hundred-page Machine Learning Book*. Andriy Burkov Quebec City, Can., 2019.

[73] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, GC Nandi, and Domenec Puig. Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity. *IEEE Transactions on Affective Computing*, 2019.

[74] Roberto Zangróniz, Arturo Martínez-Rodrigo, José Manuel Pastor, María T López, and Antonio Fernández-Caballero. Electrodermal Activity Sensor for Classification of Calm/Distress Condition. *Sensors*, 17(10):2324, 2017.

[75] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, volume 3, page 13. ACM, March 2019.

[76] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote Measurement of Cognitive Stress via Heart Rate Variability. In *Proceedings of the International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2957–2960, 2014.

[77] Erik Peper, Rick Harvey, I-Mei Lin, Hana Tylova, and Donald Moss. Is There More to Blood Volume Pulse than Heart Rate Variability, Respiratory Sinus Arrhythmia, and Cardiorespiratory Synchrony? *Biofeedback*, 35(2), 2007.

[78] Dongrae Cho, Jinsil Ham, Jooyoung Oh, Jeanho Park, Sayup Kim, Nak-Kyu Lee, and Boreom Lee. Detection of Stress Levels from Biosignals Measured in Virtual Reality Environments Using a Kernel-based Extreme Learning Machine. *Sensors*, 17(10):2435, 2017.

[79] Wahida Handouzi, Choubeila Maaoui, Alain Pruski, and Abdelhak Moussaoui. Anxiety Recognition Using Relevant Features from BVP Signal: Application on Phobic Individuals. *Modelling, Measurement, Control Journal, Modelling, Measurement, Control (2B)*, 2014.

[80] Paul van Gent, Haneen Farah, Nicole van Nes, and Bart van Arem. HeartPy: A Novel Heart Rate Algorithm for the Analysis of Noisy Signals. *Transportation Research Part F: Traffic Psychology and Behaviour*, 66:368–378, 2019.

[81] Pavel Pudil, Jana Novovičová, and Josef Kittler. Floating Search Methods in Feature Selection. Elsevier, 1994.

[82] Alexei Ivanov and Giuseppe Riccardi. Kolmogorov-Smirnov Test for Feature Selection in Emotion Recognition from Speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5125–5128, 2012.

[83] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. 2002.

[84] Nils Y Hammerla and Thomas Plötz. Let's (Not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the Joint conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 1041–1051, 2015.

[85] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The Balanced Accuracy and its Posterior Distribution. In *International Conference on Pattern Recognition (ICPR)*, pages 3121–3124, 2010.

[86] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. Revealing Household Characteristics from Smart Meter Data. *Energy*, 78:397–410, 2014.

[87] Mary Ainley. Students' Interest and Engagement in Classroom Activities. In *Handbook of Research on Student Engagement*. Springer, 2012.

[88] Timothy P Mottet and Steven A Beebe. Emotional Contagion in the Classroom: An Examination of How Teacher and Student Emotions Are Related. ERIC, 2000.

[89] Regalena Melrose. *Why Students Underachieve: What Educators and Parents Can Do About It*. R&L Education, 2006.

[90] Ryan Cain and Victor R Lee. Measuring Electrodermal Activity to Capture Engagement in an Afterschool Maker Program. In *Proceedings of the Conference on Creativity and Fabrication in Education (FabLearn 2016)*. ACM, 2016.

[91] Martin EP Seligman. *Flourish: A Visionary New Understanding of Happiness and Well-being*. Simon and Schuster, 2012.

[92] Stefan Ed Engeser. *Advances in Flow Research.* Springer Science+ Business Media, 2012.

[93] Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov. Measuring User Engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4):1–132, 2014.

[94] Heather L O'Brien and Elaine G Toms. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *Journal of the American society for Information Science and Technology*, 59(6):938–955, 2008.

[95] Jennifer A Fredricks and Wendy McColskey. The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report instruments. In *Handbook of Research on Student Engagement*, pages 763–782. Springer, 2012.

[96] Ellen A. Skinner and Jennifer R. Pitzer. *Developmental Dynamics of Student Engagement, Coping, and Everyday Resilience*, pages 21–44. Springer US, Boston, MA, 2012.

[97] Gale M Sinatra, Benjamin C Heddy, and Doug Lombardi. The Challenges of Defining and Measuring Student Engagement in Science. *Educational Psychologist*, 50(1):1–13, 2015.

[98] Jerome I Rotgans and Henk G Schmidt. Cognitive Engagement in the Problem-based Learning Classroom. *Advances in Health Sciences Education*, 16(4):465–479, 2011.

[99] James A Russell. Affective Space is Bipolar. *Journal of Personality and Social Psychology*, 37(3):345, 1979.

[100] Kristy A Nielson and Timothy J Arentsen. Memory Modulation in the Classroom: Selective Enhancement of College Examination Performance by Arousal Induced After Lecture. *Neurobiology of Learning and Memory*, 98(1):12–16, 2012.

[101] Marian L. Houser and Caroline Waldbuesser. Emotional Contagion in the Classroom: The Impact of Teacher Satisfaction and Confirmation on Perceptions of Student Nonverbal Classroom Behavior. *College Teaching*, 65(1):1–8, 2017.

[102] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. Emotional Contagion. *Current Directions in Psychological Science*, 2(3):96–100, 1993.

[103] John T Cacioppo, Louis G Tassinary, and Gary Berntson. *Handbook of Psychophysiology*. Cambridge University Press, 2007.

[104] John T Cacioppo, Gary G Berntson, Jeff T Larsen, Kirsten M Poehlmann, Tiffany A Ito, et al. The Psychophysiology of Emotion. *Handbook of Emotions*, 2:173–191, 2000.

[105] William James. *What is Emotion?* Appleton-Century-Crofts, 1948.

[106] Walter B Cannon. The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*, 39(1/4):106–124, 1927.

[107] Laurie Kelly McCorry. Physiology of the Autonomic Nervous System. *American journal of Pharmaceutical Education*, 71(4), 2007.

[108] Patrick Charland, Pierre-Majorique Léger, Sylvain Sénécal, François Courtemanche, Julien Mercier, Yannick Skelling, and Elise Labonté-Lemoyne. Assessing the Multiple Dimensions of Engagement to Characterize Learning: A Neurophysiological Perspective. *Journal of Visualized Experiments*, (101), 2015.

[109] Simo Järvelä, Jari Kätsyri, Niklas Ravaja, Guillaume Chanel, and Pentti Henttonen. Intragroup Emotions: Physiological Linkage and Social Presence. *Frontiers in Psychology*, 7, 2016.

[110] Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. Interpersonal Autonomic Physiology: A Systematic Review of the Literature. *Personality and Social Psychology Review*, 21(2):99–141, 2017.

[111] Petr Slovák, Paul Tennent, Stuart Reeves, and Geraldine Fitzpatrick. Exploring Skin Conductance Synchronisation in Everyday Interactions. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI)*, 2014.

[112] R Todd Stephens. Integrating Web 2.0 Technologies within the Enterprise. In *Web Technologies: Concepts, Methodologies, Tools, and Applications*, pages 201–218. IGI Global, 2010.

[113] Gordon B Davis. Anytime/Anyplace Computing and the Future of Knowledge Work. *Communications of the ACM*, 45(12):67–73, 2002.

[114] Sabine Sonnentag. A Task-level Perspective on Work Engagement: A New Approach that Helps to Differentiate the Concepts of Engagement and Burnout. *Burnout Research*, 5:12–20, 2017.

[115] Maslach Christina and Leiter Michael. The Truth About Burnout. *Publish by Jossey-Bass. San Fransisco, CA*, pages 94103–1741, 1997.

[116] Christina Maslach, Wilmar B Schaufeli, and Michael P Leiter. Job Burnout. *Annual Review of Psychology*, 52(1):397–422, 2001.

[117] Wilmar B Schaufeli, Marisa Salanova, Vicente González-Romá, and Arnold B Bakker. The Measurement of Engagement and Burnout: A Two Sample Confirmatory Factor Analytic Approach. *Journal of Happiness Studies*, 3(1):71–92, 2002.

[118] Corinna Peifer, André Schulz, Hartmut Schächinger, Nicola Baumann, and Conny H Antoni. The Relation of Flow-experience and Physiological Arousal Under Stress—Can U Shape It? *Journal of Experimental Social Psychology*, 53:62–69, 2014.

[119] Wilmar B Schaufeli, Arnold B Bakker, and Willem Van Rhenen. How Changes in Job Demands and Resources Predict Burnout, Work Engagement, and Sickness Absenteeism. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 30(7):893–917, 2009.

[120] Wilmar B Schaufeli, Arnold B Bakker, and Marisa Salanova. The Measurement of Work Engagement with a Short Questionnaire: A Cross-national Study. *Educational and psychological measurement*, 66(4):701–716, 2006.

[121] Jeanne Nakamura and Mihaly Csikszentmihalyi. The Concept of Flow. Flow and the Foundations of Positive Psychology. *Springer Netherlands*, 1:239–263, 2014.

[122] Mihaly Csikszentmihalyi. *Beyond Boredom and Anxiety.* Jossey-Bass, 2000.

[123] Mihaly Csikszentmihalyi and Judith LeFevre. Optimal Experience in Work and Leisure. *Journal of Personality and Social Psychology*, 56(5):815, 1989.

[124] Michael T Knierim, Raphael Rissler, Verena Dorner, Alexander Maedche, and Christof Weinhardt. The Psychophysiology of Flow: a Systematic Review of Peripheral Nervous System Features. *Information Systems and Neuroscience*, pages 109–120, 2018.

[125] Andrea Gaggioli, Pietro Cipresso, Silvia Serino, and Giuseppe Riva. Psychophysiological Correlates of Flow During Daily Activities. *Annual Review of Cybertherapy and Telemedicine*, 191:65–69, 2013.

[126] Tahmine Tozman, Elisabeth S Magdas, Hamish G MacDougall, and Regina Vollmeyer. Understanding the Psychophysiology of Flow: A Driving Simulator Experiment to Investigate the Relationship Between Flow and Heart Rate variability. *Computers in Human Behavior*, 52:408–418, 2015.

[127] J. Whitehill, Z. Serpell, Y-C Lin, A. Foster, and J. Movellan. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[128] Karina Nielsen and Bryan Cleal. Predicting Flow at Work: Investigating the Activities and Job Characteristics that Predict Flow States at Work. *Journal of Occupational Health Psychology*, 15(2):180, 2010.

[129] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. Measuring Cognitive and Psychological Engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, 44(5):427–445, 2006.

[130] Joel M Hektner, Jennifer A Schmidt, and Mihaly Csikszentmihalyi. *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage, 2007.

[131] Mihaly Csikszentmihalyi et al. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6(3):281–94, 1977.

[132] M. E. Debus, S. Sonnentag, W. Deutsch, and F. W. Nussbeck. Making Flow Happen: The Effects of Being Recovered on Work-related Flow Between and Within Days. *Journal of Applied Psychology*, 99:713–722, 2014.

[133] Penelope L Peterson, Susan R Swing, Kevin D Stark, and Gregory A Waas. Students' Cognitions and Time on Task During Mathematics Instruction. *American Educational Research Journal*, 21(3):487–515, 1984.

[134] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. Engagement Detection in Online Learning: a Review. *Smart Learning Environments*, 6(1):1–20, 2019.

[135] E Joseph. Engagement Tracing: Using Response Times to Model Student Disengagement. *Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, 125:88, 2005.

[136] Celine Latulipe, Erin A Carroll, and Danielle Lottridge. Love, Hate, Arousal and Engagement: Exploring Audience Responses to Performing Arts. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1845–1854, 2011.

[137] Akhil Mathur, Nicholas D Lane, and Fahim Kawsar. Engagement-aware Computing: Modelling User Engagement from Mobile Contexts. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 622–633, 2016.

[138] Javier Hernandez, Zicheng Liu, Geoff Hulten, Dave DeBarr, Kyle Krum, and Zhengyou Zhang. Measuring the Engagement Level of TV Viewers. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013.

[139] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. In *Educational Data Mining*, 2013.

[140] Sidney D'Mello, Rosalind W. Picard, and Arthur Graesser. Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, 22(4):53–61, July 2007.

[141] M Ali Akber Dewan, Fuhua Lin, Dunwei Wen, Mahbub Murshed, and Zia Uddin. A Deep Learning Approach to Detecting Engagement of Online Learners. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1895–1902. IEEE, 2018.

[142] Chen Wang and Pablo Cesar. Do We React in the Same Manner?: Comparing GSR Patterns Across Scenarios. In *Proceedings of the Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI)*, 2014.

[143] Zexian Wang, Fei Jiang, and Ruimin Shen. An Effective Yawn Behavior Detection Method in Classroom. In *International Conference on Neural Information Processing*, pages 430–441. Springer, 2019.

[144] E Friesen and Paul Ekman. Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Palo Alto*, 3(2):5, 1978.

[145] Paul Ekman. Facial Action Coding System (FACS). *A Human Face*, 2002.

[146] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The Computer Expression Recognition Toolbox (CERT). In *Face and Gesture*, pages 298–305, 2011.

[147] Anna Gruebler and Kenji Suzuki. Measurement of Distal EMG Signals Using a Wearable Device for Reading Facial Expressions. In *International Conference of the IEEE Engineering in Medicine and Biology (EMBC)*, pages 4594–4597. IEEE, 2010.

[148] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. Automatic Smile and Frown Recognition with Kinetic Earables. In *Proceedings of the Augmented Human International Conference (AH)*, pages 1–4, 2019.

[149] Sidney S D'Mello, Patrick Chipman, and Art Graesser. Posture as A Predictor of Learner's Affective Engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[150] Daniele Bibbo, Marco Carli, Silvia Conforto, and Federica Battisti. A Sitting Posture Monitoring Instrument to Assess Different Levels of Cognitive Engagement. *Sensors*, 19(3):455, 2019.

[151] Ardhendu Behera, Peter Matthew, Alexander Keidel, Peter Vangorp, Hui Fang, and Susan Canning. Associating Facial Expressions and Upper-body Gestures with Learning Tasks for Enhancing Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 30(2):236–270, 2020.

[152] Joseph F Grafsgaard, Robert M Fulton, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. Multimodal Analysis of the Implicit Affective Channel in Computer-mediated Textual Communication. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 145–152, 2012.

[153] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. EduSense: Practical Classroom Sensing at Scale. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, volume 3, pages 1–26, 2019.

[154] Ghassem Tofighi, Haisong Gu, and Kaamraan Raahemifar. Vision-based Engagement Detection in Virtual Reality. In *Digital Media Industry & Academic Forum (DMIAF)*, pages 202–206, 2016.

[155] Koji Ara, Nobuo Sato, Satomi Tsuji, Yoshihiro Wakisaka, Norio Ohkubo, Youichi Horry, Norihiko Moriwaki, Kazuo Yano, and Miki Hayakawa. Predicting Flow State in Daily Work Through Continuous Sensing of Motion Rhythm. In *International Conference on Networked Sensing Systems (INSS)*, pages 1–6, 2009.

[156] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, 2017.

[157] Mariam Hassib, Mohamed Khamis, Susanne Friedl, Stefan Schneegass, and Florian Alt. Brainatwork: Logging Cognitive Engagement and Tasks in the Workplace Using Electroencephalography. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (ICMI)*, pages 305–310, 2017.

[158] Nataliya Kosmyna and Pattie Maes. AttentivU: an EEG-based Closed-loop Biofeedback System for Real-time Monitoring and Improvement of Engagement for Personalized Learning. *Sensors*, 19(23):5200, 2019.

[159] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 3–14, 2014.

[160] Alan T Pope, Edward H Bogart, and Debbie S Bartolome. Biocybernetic System Evaluates Indices of Operator Engagement in Automated Task. *Biological Psychology*, 40(1-2):187–195, 1995.

[161] Kimberly A Brownley, Barry E Hurwitz, and Neil Schneiderman. Cardiovascular Psychophysiology. *Handbook of Psychophysiology*, pages 224—26, 2000.

[162] Erik Peper, Fred Shaffer, and I-Mei Lin. Garbage In; Garbage Out – Identify Blood Volume Pulse (BVP) Artifacts Before Analyzing and Interpreting BVP, Blood

Volume Pulse Amplitude, and Heart Rate/Respiratory Sinus Arrhythmia Data. *Biofeedback*, 38(1):19–23, 2010.

[163] Wahida Handouzi, Choubeila Maaoui, Alain Pruski, and Abdelhak Moussaoui. Anxiety Recognition Using Relevant Features from BVP Signal: Application on Phobic Individuals. *Modelling, Measurement, Control Journal, Modelling, Measurement, Control*, 2014.

[164] Dongrae Cho, Jinsil Ham, Jooyoung Oh, Jeanho Park, Sayup Kim, Nak-Kyu Lee, and Boreom Lee. Detection of Stress Levels from Biosignals Measured in Virtual Reality Environments Using a Kernel-Based Extreme Learning Machine. *Sensors*, 17(10), 2017.

[165] Xiao Zhang, Yongqiang Lyu, Xiaomin Luo, Jingyu Zhang, Chun Yu, Hao Yin, and Yuanchun Shi. Touch sense: Touch screen based mental stress sense. *Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 2(2), 2018.

[166] Fernando Silveira, Brian Eriksson, Anmol Sheth, and Adam Sheppard. Predicting Audience Responses to Movie Content from Electro-dermal Activity Signals. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp)*, 2013.

[167] Saša Branković. Assessment of Brain Monoaminergic Signaling Through Mathematical Modeling of Skin Conductance Response. In *Neuroscience-Dealing With Frontiers*. InTech, 2012.

[168] Rui Henriques, Ana Paiva, and Claudia Antunes. Accessing Emotion Patterns from Affective Interactions Using Electrodermal Activity. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Humaine Association (ACII)*, 2013.

[169] Lyndon S Kennedy and Daniel PW Ellis. Laughter Detection in Meetings. 2004.

[170] Kornel Laskowski and Susanne Burger. Analysis of the Occurrence of Laughter in Meetings. In *Conference of the International Speech Communication Association (ISCA)*, 2007.

[171] Janez Zaletelj and Andrej Košir. Predicting Students' Attention in the Classroom from Kinect Facial and Body Features. *Journal on Image and Video Processing*, 2017(1):1–12, 2017.

[172] Pascal E Fortin and Jeremy R Cooperstock. Laughter and Tickles: Toward Novel Approaches for Emotion and Behavior Elicitation. *IEEE Transactions on Affective Computing*, 8(4):508–521, 2017.

[173] Sarah Cosentino, Salvatore Sessa, and Atsuo Takanishi. Quantitative Laughter Detection, Measurement, and Classification – A Critical Survey. *IEEE Reviews in Biomedical Engineering*, 9:148–162, 2016.

[174] Willibald Ruch and Paul Ekman. The Expressive Pattern of Laughter. In *Emotions, Qualia, and Consciousness*. World Scientific, 2001.

[175] Radoslaw Niewiadomski, Maurizio Mancini, Giovanna Varni, Gualtiero Volpe, and Antonio Camurri. Automated Laughter Detection from Full-body Movements. *IEEE Transactions on Human-Machine Systems*, 46(1):113–123, 2016.

[176] Willibald Ruch. Exhilaration and Humor. *Handbook of Emotions*, 1, 1993.

[177] Mary Payne Bennett and Cecile Lengacher. Humor and Laughter May Influence Health: III. Laughter and Health Outcomes. *Evidence-Based Complementary and Alternative Medicine*, 5(1), 2008.

[178] Helena Kangasharju and Tuija Nikko. Emotions in Organizations: Joint Laughter in Workplace Meetings. *The Journal of Business Communication*, 46(1):100–119, 2009.

[179] Katsuya Fujii, Plivelic Marian, Dav Clark, Yoshi Okamoto, and Jun Rekimoto. Sync Class: Visualization System for In-Class Student Synchronization. In *Proceedings of the Augmented Human International Conference (AH)*, 2018.

[180] Qiaosi Wang, Shan Jing, David Joyner, Lauren Wilcox, Hong Li, Thomas Plötz, and Betsy Disalvo. Sensing Affect to Empower Students: Learner Perspectives on Affect-Sensitive Technology in Large Educational Contexts. In *Proceedings of the ACM Conference on Learning @ Scale (L@S)*, pages 63–76, 2020.

[181] Kaśka Porayska-Pomsta, Manolis Mavrikis, Sidney D'Mello, Cristina Conati, and Ryan SJd Baker. Knowledge Elicitation Methods for Affect Modelling in Education. *International Journal of Artificial Intelligence in Education*, 22(3):107–140, 2013.

[182] Steven S Coughlin. Recall Bias in Epidemiologic Studies. *Journal of Clinical Epidemiology*, 43(1):87–91, 1990.

[183] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and Localization of Student Engagement in the Wild. In *Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018.

[184] Aaqib Saeed, Stojan Trajanovski, Maurice Van Keulen, and Jan Van Erp. Deep Physiological Arousal Detection in a Driving Simulator Using Wearable Sensors. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 486–493, 2017.

[185] Thomas Plötz and Yu Guan. Deep Learning for Human Activity Recognition in Mobile Computing. *Computer*, 51(5):50–59, 2018.

[186] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal Deep Learning for Activity and Context Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, 1(4):1–27, 2018.

[187] Jiacheng Liao, Yan Liang, and Jiahui Pan. Deep Facial Spatiotemporal Network for Engagement Prediction in Online Learning. *Applied Intelligence*, pages 1–13, 2021.

[188] Hao Zhang, Xiaofan Xiao, Tao Huang, Sanya Liu, Yu Xia, and Jia Li. An Novel End-to-end Network for Automatic Student Engagement Recognition. In *IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 342–345, 2019.

[189] Tao Huang, Yunshan Mei, Hao Zhang, Sanya Liu, and Huali Yang. Fine-grained Engagement Recognition in Online Learning Environment. In *IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 338–341, 2019.

[190] Abhay Gupta, Richik Jaiswal, Sagar Adhikari, and Vineeth N Balasubramanian. DAISEE: Dataset for Affective States in E-learning Environments. *arXiv*, pages 1–22, 2016.

[191] Abhinav Dhall. Emotiw 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. In *International Conference on Multimodal Interaction (ICMI)*, pages 546–550, 2019.

[192] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 2108–2110, 2019.

[193] Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities, author=Chen, Kaixuan and Zhang, Dalin and Yao, Lina and Guo, Bin and Yu, Zhiwen and Liu, Yunhao. *arXiv e-prints*, pages arXiv–2001, 2020.

[194] Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. Understanding Personal Productivity: How Knowledge Workers Define, Evaluate, and Reflect on their Productivity. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–12, 2019.

[195] Daniel A Epstein, Daniel Avrahami, and Jacob T Biehl. Taking 5: Work-breaks, Productivity, and Opportunities for Personal Informatics for Knowledge Workers. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2016.

[196] Gloria Mark, Victor M Gonzalez, and Justin Harris. No Task Left Behind? Examining the Nature of Fragmented Work. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2005.

[197] Yuhan Luo, Bongshin Lee, Donghee Yvette Wohn, Amanda L Rebar, David E Conroy, and Eun Kyoung Choe. Time for Break: Understanding Information Workers' Sedentary Behavior Through a Break Prompting System. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2018.

[198] Gloria Mark, Mary Czerwinski, and Shamsi T Iqbal. Effects of Individual Differences in Blocking Workplace Distractions. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–12, 2018.

[199] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. A Conversational Agent in Support of Productivity and Wellbeing at Work. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2019.

[200] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. Design and Evaluation of Intelligent Agent Prototypes for Assistance with Focus and Productivity at Work. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 2020.

[201] André N Meyer, Thomas Fritz, Gail C Murphy, and Thomas Zimmermann. Software Developers' Perceptions of Productivity. In *Proceedings of ACM International Symposium on Foundations of Software Engineering (SIGSOFT)*, pages 19–29, 2014.

[202] Seung Hyun Cha, Joonoh Seo, Seung Hyo Baek, and Choongwan Koo. Towards a Well-planned, Activity-based Work Environment: Automated Recognition of Office Activities Using Accelerometers. *Building and Environment*, 144:86–93, 2018.

[203] Akane Sano and Rosalind W Picard. Stress Recognition Using Wearable Sensors and Mobile Phones. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 671–676, 2013.

[204] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W Picard. Improving Students' Daily Life Stress Forecasting Using LSTM Neural Networks. In *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4, 2019.

[205] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. Predicting Students' Happiness from Physiology, Phone, Mobility, and Behavioral Data. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015.

[206] Akane Sano, Andrew J. K. Phillips, Amy Z. Yu, Andrew W. McHill, Sara Taylor, Natasha Jaques, Charles A. Czeisler, Elizabeth B. Klerman, and Rosalind W. Picard. Recognizing Academic Performance, Sleep Quality, Stress Level, and Mental Health Using Personality Traits, Wearable Sensors and Mobile Phones. In *BSN*, pages 1–6, 2015.

[207] Phuong Pham and Jingtao Wang. AttentiveLearner: Improving Mobile MOOC Learning via Implicit Heart Rate Tracking. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 367–376, 2015.

[208] Jingjing Chen, Bin Zhu, Olle Balter, Jianliang Xu, Weiwen Zou, Anders Hedman, Rongchao Chen, and Mengdie Sang. FishBuddy: Promoting Student Engagement in Self-Paced Learning through Wearable Sensing. In *In Proceedings of the IEEE International Conference on Smart Computing (SMARTCOMP 2017)*, pages 1–9, 2017.

[209] Kasia Muldner, Michael Wixon, Dovan Rai, Winslow Burleson, Beverly Woolf, and Ivon Arroyo. Exploring the Impact of a Learning Dashboard on Student Affect. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 307–317, 2015.

[210] Brandon Booth, Asem Ali, Ian Bennett, and Shrikanth Narayanan. Toward Active and Unobtrusive Engagement Assessment of Distance Learners. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 470–476, 2017.

[211] Ivon Arroyo, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion Sensors Go To School. In *Proceedings of the Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling (AIED)*, pages 17–24, 2009.

[212] Samara Ruiz, Sven Charleer, Maite Urretavizcaya, Joris Klerkx, Isabel Fernández-Castro, and Erik Duval. Supporting Learning by Considering Emotions: Tracking and Visualization a Case Study. In *Proceedings of the ACM International Conference on Learning Analytics & Knowledge (LAK)*, pages 254–263, 2016.

[213] John P Pollak, Phil Adams, and Geri Gay. PAM: a Photographic Affect Meter for Frequent, in Situ Measurement of Affect. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, pages 725–734, 2011.

[214] João Maroco, Ana Lúcia Maroco, Juliana Alvares Duarte Bonini Campos, and Jennifer A Fredricks. University Student's Engagement: Development of the University Student Engagement Inventory (USEI). Springer, 2016.

[215] Claudio Martella, Ekin Gedik, Laura Cabrera-Quiros, Gwenn Englebienne, and Hayley Hung. How Was It?: Exploiting Smartphone Sensing to Measure Implicit Audience Responses to Live Performances. In *Proceedings of the International Conference on Multimedia (ICME)*, pages 201–210, 2015.

[216] Yuning Zhang, Maysam Haghdan, and Kevin S Xu. Unsupervised Motion Artifact Detection in Wrist-measured Electrodermal Activity Data. In *Proceedings of the International Symposium on Wearable Computers (ISWC)*, pages 54–57, 2017.

[217] Jorn Bakker, Mykola Pechenizkiy, and Natalia Sidorova. What's Your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 573–580, 2011.

[218] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. A Wearable Sensor for Unobtrusive, Long-term Assessment of Electrodermal Activity. *IEEE Transactions on Biomedical Engineering*, 57(5):1243–1252, 2010.

[219] David T Lykken and Peter H Venables. Direct Measurement of Skin Conductance: A Proposal for Standardization. *Psychophysiology*, 8(5):656–672, 1971.

[220] Mathias Benedek and Christian Kaernbach. Decomposition of Skin Conductance Data By Means of Nonnegative Deconvolution. *Psychophysiology*, 47(4):647–658, 2010.

[221] Adrián Colomer Granero, Félix Fuentes-Hurtado, Valery Naranjo Ornedo, Jaime Guixeres Provinciale, Jose M Ausín, and Mariano Alcañiz Raya. A Comparison of Physiological Signal Analysis Techniques and Classifiers for Automatic Emotional Evaluation of Audiovisual Contents. *Frontiers in Computational Neuroscience*, 10, 2016.

[222] Carl D Marci, Jacob Ham, Erin Moran, and Scott P Orr. Physiologic Correlates of Perceived Therapist Empathy and Social-Emotional Process During Psychotherapy. *The Journal of Nervous and Mental Disease*, 195(2):103–111, 2007.

[223] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. Experiencing SAX: a Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007.

[224] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[225] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[226] Yutaka Sasaki. The Truth of the F-measure. *Teach Tutor mater*, pages 1021–1032, 2007.

[227] Andreas C Müller and Sarah Guido. *Introduction to Machine Learning with Python: a Guide for Data Scientists*. O'Reilly Media, Inc., 2016.

[228] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, volume 1, pages 1–25, 2017.

[229] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2016.

[230] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, volume 1, pages 1–24, 2017.

[231] Douglas G Altman. *Practical Statistics for Medical Research*. CRC Press, 1990.

[232] Kimberly E Arnold, Brandon Karcher, Casey V Wright, and James McKay. Student Empowerment, Awareness, and Self-Regulation Through a Quantified-Self Student Tool. In *Proceedings of the ACM International Learning Analytics & Knowledge Conference (LAK)*, 2017.

[233] Victor R Lee, Joel R Drake, and Jeffrey L Thayne. Appropriating Quantified Self Technologies to Support Elementary Statistical Teaching and Learning. *IEEE Transactions on Learning Technologies*, 9(4):354–365, 2016.

[234] Elena Di Lascio, Shkurta Gashi, Danilo Krasic, and Silvia Santini. In-classroom Self-tracking for Teachers and Students: Preliminary Findings from a Pilot Study. In *Adjunct Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 865–870, 2017.

[235] Steven Cantrell and Thomas J Kane. Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-year Study. *MET Project Research Paper*, 2013.

[236] Vasileios Triglianos, Cesare Pautasso, Alessandro Bozzon, and Claudia Hauff. Inferring Student Attention with ASQ. In *Proceedings of the European Conference on Technology Enhanced Learning (ECTEL)*, pages 306–320, 2016.

[237] Alison Beard. Leading with Humor. *Harvard Business Review*, 92(5):130–131, 2014.

[238] Wolff-Michael Roth, Stephen M Ritchie, Peter Hudson, and Victoria Mergard. A Study of Laughter in Science Lessons. *Journal of Research in Science Teaching*, 48(5):437–458, 2011.

[239] Daniela Jeder. Implications of Using Humor in the Classroom. *Procedia-Social and Behavioral Sciences*, 180:828–833, 2015.

[240] Khiet P Truong and David A Van Leeuwen. Automatic Discrimination Between Laughter and Speech. *Speech Communication*, 49(2):144–158, 2007.

[241] Monica Perusquia-Hernandez, Mazakasu Hirokawa, and Kenji Suzuki. A Wearable Device for Fast and Subtle Spontaneous Smile Recognition. *IEEE Transactions on Affective Computing*, 8(4):522–533, 2017.

[242] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A Review of Affective Computing: From Unimodal Analysis to Multimodal FusionA Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion*, 37:98–125, 2017.

[243] Simon Flutura, Johannes Wagner, Florian Lingenfelser, Andreas Seiderer, and Elisabeth André. Laughter Detection in the Wild: Demonstrating a Tool for Mobile Social Signal Processing and Visualization. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 406–407, 2016.

[244] Sarah Cosentino, Tatsuhiro Kishi, Massimiliano Zecca, Salvatore Sessa, Luca Bartolomeo, Kenji Hashimoto, Takashi Nozawa, and Atsuo Takanishi. Human-humanoid Robot Social Interaction: Laughter. In *Proceedings of the International Conference on Robotics and Biomimetics (ROBIO)*, pages 1396–1401, 2013.

[245] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying Social Actions with a Single Accelerometer. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 207–210, 2013.

[246] Shiho Tatsumi, Yasser Mohammad, Yoshimasa Ohmoto, and Toyoaki Nishida. Detection of Hidden Laughter for Human-agent Interaction. *Procedia Computer Science*, 35:1053–1062, 2014.

[247] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[248] Reshmashree B Kantharaju, Fabien Ringeval, and Laurent Besacier. Automatic Recognition of Affective Laughter in Spontaneous Dyadic Interactions from Audiovisual Signals. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 220–228, 2018.

[249] Jérôme Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. The AVLaughterCycle Database. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2010.

[250] J Ridley Stroop. Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology*, 18(6), 1935.

[251] Jo-Anne Bachorowski, Moria J Smoski, and Michael J Owren. The Acoustic Features of Human Laughter. *The Journal of the Acoustical Society of America*, 110(3):1581–1597, 2001.

[252] Kevin El Haddad, Hüseyin Çakmak, Emer Gilmartin, Stéphane Dupont, and Thierry Dutoit. Towards a Listening Agent: A System Generating Audiovisual Laughs and Smiles to Show Interest. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2016.

[253] Jérôme Urbain, Hüseyin Çakmak, Aurélie Charlier, Maxime Denti, Thierry Dutoit, and Stéphane Dupont. Arousal-Driven Synthesis of Laughter. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):273–284, 2014.

[254] Kyriaki Kalimeri and Charalampos Saitis. Exploring Multimodal Biosignal Features for Stress Detection During Indoor Mobility. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pages 53–60, 2016.

[255] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring. *IEEE Journal of Biomedical and Health Informatics*, 19(3):832–838, 2015.

[256] Erik Peper, Rick Harvey, I-Mei Lin, Hana Tylova, and Donald Moss. Is There More to Blood Volume Pulse than Heart Rate Variability, Respiratory Sinus Arrhythmia, and Cardiorespiratory Synchrony? *Biofeedback*, 35(2), 2007.

[257] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature Selection: A Data Perspective. *Computing Surveys*, 50(6):1–45, 2018.

[258] Alexei Ivanov and Giuseppe Riccardi. Kolmogorov-Smirnov Test for Feature Selection in Emotion Recognition from Speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5125–5128, 2012.

[259] Maria R Reyes, Marc A Brackett, Susan E Rivers, Mark White, and Peter Salovey. Classroom Emotional Climate, Student Engagement, and Academic Achievement. *Journal of educational psychology*, 104(3):700, 2012.

[260] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. Using Students' Physiological Synchrony to Quantify the Classroom Emotional Climate. In *the Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers (UbiComp 2018)*. ACM, October 2018.

[261] Pegah Hafiz, Raju Maharjan, and Devender Kumar. Usability of a Mood Assessment Smartphone Prototype Based on Humor Appreciation. In *Adjunct Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pages 151–157, 2018.

[262] Alexandros Zenonos, Aftab Khan, Georgios Kalogridis, Stefanos Vatsikas, Tim Lewis, and Mahesh Sooriyabandara. HealthyOffice: Mood Recognition at Work Using Smartphones and Wearable Sensors. In *Proceedings of the International Conference on Pervasive Computing and Communication Workshops (PerComp)*, pages 1–6, 2016.

[263] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, and Paul Johns. Bored Mondays and Focused Afternoons: The Rhythm of Attention and Online Activity in the Workplace. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 3025–3034, 2014.

[264] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, 2(1):1–20, 2018.

[265] Neal Lathia, Kiran K Rachuri, Cecilia Mascolo, and Peter J Rentfrow. Contextual Dissonance: Design Bias in Sensor-based Experience Sampling Methods. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 183–192, 2013.

[266] Akhil Mathur, Marc Van den Broeck, Geert Vanderhulst, Afra Mashhadi, and Fahim Kawsar. Tiny Habits in the Giant Enterprise: Understanding the Dynamics of a Quantified Workplace. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 577–588, 2015.

[267] Jaejeung Kim, Chiwoo Cho, and Uichin Lee. Technology Supported Behavior Restriction for Mitigating Self-interruptions in Multi-device Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, 1(3):1–21, 2017.

[268] Xiaozhen Ye, Huansheng Ning, Per Backlund, and Jianguo Ding. Flow Experience Detection and Analysis for Game Users by Wearable Devices-based Physiological Responses Capture. *IEEE Internet of Things Journal*, 2020.

[269] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. TimeAware: Leveraging Framing Effects to Enhance Personal Productivity. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 272–283, 2016.

[270] Laura Fiorini, Gianmaria Mancioppi, Francesco Semeraro, Hamido Fujita, and Filippo Cavallo. Unsupervised Emotional State Classification Through Physiological Parameters for Social Robotics Applications. *Knowledge-Based Systems*, 190, 2020.

[271] Vedant Das Swain, Koustuv Saha, Hemang Rajvanshy, Anusha Sirigiri, Julie M Gregg, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. A Multisensor Person-Centered Approach to Understand the Role of Daily Activities in Job Performance with Organizational Personas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT)*, 3(4):1–27, 2019.

[272] Anja Baethge and Thomas Rigotti. Interruptions to Workflow: Their Relationship with Irritation and Satisfaction with Performance, and the Mediating Roles of Time Pressure and Mental Demands. *Work & Stress*, 27(1):43–63, 2013.

[273] Reeshad S Dalal, Holly Lam, Howard M Weiss, Eric R Welch, and Charles L Hulin. A Within-person Approach to Work Behavior and Performance: Concurrent and Lagged Citizenship-counterproductivity Associations, and Dynamic Relationships with Affect and Overall Job Performance. *Academy of Management Journal*, 52(5):1051–1066, 2009.

[274] Jens Grubert, Matthias Kranz, and Aaron Quigley. Challenges in Mobile Multi-device Ecosystems. *mUX: The Journal of Mobile User Experience*, 5(1):1–22, 2016.

[275] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. "Don't Waste My Time" Use of Time Information Improves Focus. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1729–1738, 2016.

[276] Fred Shaffer and JP Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, 5.

[277] Varun Mishra, Tian Hao, Si Sun, Kimberly N Walter, Marion J Ball, Ching-Hua Chen, and Xinxin Zhu. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Adjunct Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 1708–1716, 2018.

[278] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[279] Terumi Umematsu, Akane Sano, Sara Taylor, Masanori Tsujikawa, and Rosalind W Picard. Forecasting Stress, Mood, and Health from Daytime Physiology in Office Workers and Students. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5953–5957, 2020.

[280] Jonathan Z Bakdash and Laura R Marusich. Repeated Measures Correlation. *Frontiers in Psychology*, 8.

[281] Dafne van Kuppevelt, Joe Heywood, Mark Hamer, Séverine Sabia, Emla Fitzsimons, and Vincent van Hees. Segmenting Accelerometer Data from Daily Life with Unsupervised Machine Learning. *PloS One*, 14(1), 2019.

[282] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. When Attention is Not Scarce – Detecting Boredom from Mobile Phone Usage. In *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 825–836, 2015.

[283] Andreas Zinnen, Ulf Blanke, and Bernt Schiele. An Analysis of Sensor-oriented vs. Model-based Activity Recognition. In *International Symposium on Wearable Computers (ISWC)*, pages 93–100, 2009.

[284] Tianqi Chen and Carlos Guestrin. XGboost: A Scalable Tree Boosting System. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016.

[285] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing*, (1):1–1, 2017.

[286] Kazuo Yano, Sonja Lyubomirsky, and Joseph Chancellor. Sensing Happiness. *IEEE Spectrum*, 49(12):32–37, 2012.

[287] Manuela Züger, Christopher Corley, André N Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. Reducing Interruptions at Work: A Large-scale Field Study of FlowLight. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 61–72, 2017.