Exploring LAUREATE - the Longitudinal multimodAl stUdent expeRience datasEt for AffecT and mEmory research

Matías Laporte
matias.laporte@usi.ch
Università della Svizzera italiana
Lugano, Ticino, Switzerland

Daniele Gasparini
daniele.gasparini@usi.ch
Università della Svizzera italiana
Lugano, Ticino, Switzerland

Martin Gjoreski
martin.gjoreski@usi.ch
Università della Svizzera italiana
Lugano, Ticino, Switzerland

Marc Langheinrich
marc.langheinrich@usi.ch
Università della Svizzera italiana
Lugano, Ticino, Switzerland

ABSTRACT

Studies in the lab have shown that affect recognition using physiological data is feasible with machine learning methods. Datasets collected in-the-wild can further improve such methods' robustness and applicability. This study presents LAUREATE, a Longitudinal mUltimodal student expeRience datasEt for AffecT and mEmory research. The dataset was collected throughout a university semester with 44 participants (including two lecturers) in two courses totalling 52 sessions, including classes, quizzes, and exams. We recorded participants' physiological signals with a wristband device and collected daily survey answers about participants' behaviour (e.g. study hours, smoking habits, physical activity, caffeine and food intake) and their perceived engagement, attention, and emotions after class. As a proxy for evaluating the quality of the physiological data, we present preliminary findings about the relation between the physiological signals and the different session types.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing*; *HCI design and evaluation methods.*

KEYWORDS

Dataset; Physiological Signals, Students, Wearable, Memory Recall, Affect, Mental Health

ACM Reference Format:
Matías Laporte, Daniele Gasparini, Martin Gjoreski, and Marc Langheinrich. 2022. Exploring LAUREATE - the Longitudinal multimodAl stUdent expeRience datasEt for AffecT and mEmory research. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544793.3563426

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09…$15.00
https://doi.org/10.1145/3544793.3563426

1 INTRODUCTION

The availability of small wearable low-cost sensors, combined with advanced signal processing and machine learning, is driving the revolution in mobile behaviour monitoring. This technological drive enables novel approaches for accurate measurements in sports analytics, well-being, and lifestyle monitoring. To further extend the applicability of wearable sensors in sectors such as mobile health systems, methods for accurately extracting subtle psychological and physiological information from the sensors are required. Such methods have direct applicability in systems for remote monitoring and management of mental disorders. The COVID-19 pandemic considerably increased the demand for such remote solutions [16] because of the reduced capacity for mental health services and the worsening mental health status of the general population (e.g., due to movement constraints and loss of friends and relatives). However, accessing psycho-physiological information in everyday life remains challenging [10] – smartphones can count steps, but they cannot recognise affect.

Many studies in controlled environments have confirmed that some form of affect recognition can be performed with speech analysis, video analysis, or physiological data analysis combined with Machine Learning. Most of the methods that use physiological signals are based on data from electrocardiogram (ECG), electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), galvanic skin response (GSR), electrooculography (EOG) or Blood Volume Pulse (BVP) sensors [25]. However, methods developed in controlled environments often fail when applied in the wild. Several studies have tackled the problem of affect recognition in the wild (e.g., monitoring stress outside the lab [7, 15, 23]), but the number of these studies is limited, and the variability of the affective states monitored in these studies is also limited. The lack of in-the-wild labelled datasets for affect recognition is another factor that contributes to the slow development of robust methods. It is challenging to collect such datasets because it requires participants' adherence multiple times per day for several weeks or months.

In this work, we present LAUREATE, the Longitudinal mUltimodal student expeRience datasEt for AffecT and mEmory research. More specifically, we present an initial exploratory analysis of the physiological data available in the dataset. The dataset, described in more detail in Section 3.1 "Data Collection", includes physiological signals of 42 students and two lecturers over 26 sessions, each within the context of two different courses. With the intention of evaluating the data quality, we will focus on the relationship between physiological signals and their ability to distinguish between different events during sessions, like classes, quizzes and exams. Our assumption is that different sessions cause the students to be in different affective states (e.g., different levels of stress and engagement), causing visible physiological changes measured by wrist devices.

2 RELATED WORK

A recent review by Siddiqui et al. [21] lists 50 datasets from affective computing studies containing multimodal data such as facial expressions, speech, and physiological signals. In this section, we focus on datasets containing physiological signals. Datasets like Ascertain [22] and Amigos [14] are emotion recognition datasets where participants watched affective multimedia in short sessions. Laughter [4] is another slightly different dataset aiming to recognise laughter using non-invasive wearable devices. Some datasets focus on monitoring driving workload [2, 9, 20] where participants drove a predefined route with different stressful or stress-free sections (e.g., crowded vs free highway) in real life [9, 20] or in a simulator [2]. There are also datasets for monitoring cognitive load, where participants played games on a smartphone or performed specific tasks designed to induce various cognitive load levels [6]. Similarly, studies have used math problems combined with time pressure [7] and the Trier Social Stress Test [19] to induce different stress levels in participants. All of these datasets have been widely used to develop affect recognition models in a controlled environment. To further improve the applicability of those models, longitudinal datasets from less controlled environments are needed.

After reviewing related literature on affective datasets [21, 25], we believe this is the largest dataset for physiological data from the Empatica E4 device. There are similar multimodal datasets, like CLAS [12], DEAP [11], MMSE [26] and WESAD [19]. Although CLAS and MMSE benefit from having a larger number of participants, our dataset contains longer-length repeated sessions per participant. WESAD and DEAP, while comparable in the number of subjects, were done in a controlled laboratory setting, whereas our work was carried out in a classroom. Gjoreski et al. [7] released a dataset of only five subjects for monitoring stress in the wild using the Empatica wrist device. Di Lascio et al. [3] also measured students' and lecturers' physiological data. Although their dataset has a higher diversity of courses (five instead of two), their experiment only lasted three weeks, while ours lasted 13 weeks.

Perhaps one of the most relatable datasets in the area of the student experience is StudentLife [24]. In this study, 48 students from one course answered daily and weekly surveys over the course of 10 weeks about their lifestyle, mental well-being, and academic performance. The dataset also comprises sensor data from smartphones. Nonetheless, our dataset differs from it by adding physiological data from students and lecturers.

3 THE DATASET

In this section, we describe the steps followed for the data collection and feature extraction from the physiological signals.

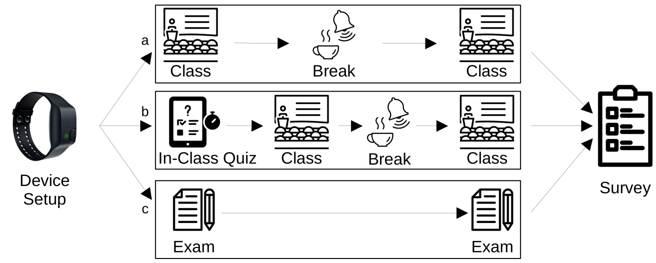


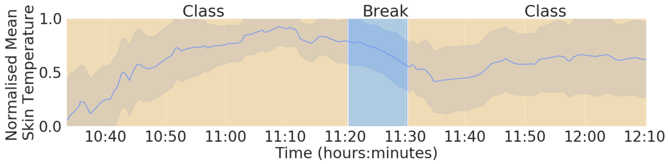
Figure 1: Data collection procedure. Each session begins with the device setup. There are three types of sessions depending on their blocks (*class, quiz, exam,* and *break*): a) *class*-only sessions separated by a *break*, b) *class* sessions preceded by a *quiz*, and c) *exam* sessions. A survey followed all sessions.

3.1 Data Collection

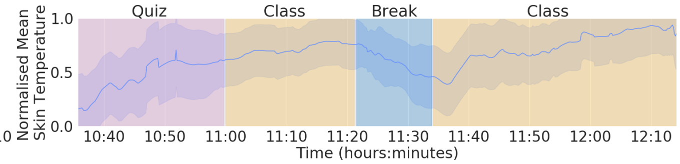
We conducted a longitudinal study throughout a university semester (~14 weeks) with 44 participants (including two lecturers) in two courses. One course was at the Bachelor's level, and the other at the Master's level of our university. The data were collected during 26 sessions of each course, including the different classes' examinations (in-class quizzes, and midterm and final exams). The data-collection procedure is explained in Figure 1. Prior to the start of each session, we provided participants with an E4 wristband from Empatica. The E4 is a wrist-located device, similar to a smartband or smartwatch, that unobtrusively measures four different physiological signals of the users through its sensors: electrodermal activity (EDA), blood volume pulse (BVP), acceleration (ACC) and skin temperature (ST). We fitted the participants with the device and started recording their data. As lecturers gesture more with their hands and walk while presenting, they used one device on each wrist to reduce possible signal artefacts caused by movement. It is important to note that *exam* blocks always took the entire session length, and *quiz* blocks were always at the beginning. Therefore, there are three subtypes of *class* blocks, depending on their context: *class-start*, *class-after-quiz* and *class-after-break*, as shown in Figure 1.

The students answered pre-study and post-study surveys about their habits and personality, including the Big Five questionnaire. They also answered daily surveys about their behaviour (e.g., nutrition, physical activity, sleep quality, and study habits) and, on lecture days, about their post-class emotional states (e.g., positive and negative affect, engagement, satisfaction). Daily surveys were administered via a mobile phone application. Lecturers answered post-class surveys about their emotions, their engagement, and the engagement they perceived from the students. We also captured audio of the lecturer, video recordings of the classrooms, and a video stream of the slides for each lecture.

After each recording session, we connected the devices to a computer for data synchronisation with Empatica's servers, a necessary step for accessing the data. We then downloaded the data and started the processing step. The procedure included data cleaning, feature extraction, and data normalisation. Each student had the possibility of visualising their own processed data through an interactive dashboard.



(a)



(b)

Figure 2: Normalised mean skin temperature (with shaded 95% confidence interval) of all participants in two different sessions. The data is min-max normalised per participant per session. Figure 2a is a *class*-only session, whereas Figure 2b starts with a *quiz* and is followed by a *class* block (i.e., *class-after-quiz*). In both cases, a *break* separates both *class* blocks.

The full dataset contains over 1400 hours of physiological data from 44 individuals, 3600 completed daily surveys from students, 70 hours of audiovisual recordings from the classes, and students' grades for quizzes, assignments and midterm and final exams, among others.

Ethics delegates at our Faculty examined and authorised the experimental protocol. Students were compensated for their participation in the study. Instructors from the course had no access to any of the data from the students. Participants signed an informed consent agreeing to share their data anonymously. An alphanumeric identifier has been used to match identities with their data.

3.2 Feature extraction and data normalisation

In addition to the raw signals from the Empatica device (EDA, BVP, ST and ACC), the dataset contains 157 features extracted using 2-minute sliding windows. The features include statistical descriptors of the raw signals and their derivatives and expert-based features important for affective computing studies. The expert-based features include HRV descriptors in time and frequency domains and features extracted from decomposed EDA signals, which describe the signal's phasic and tonic components (i.e., the faster and slower components of the EDA signal, respectively). A more detailed description of the features and the feature-extraction process is available in Gjoreski's Ph.D. thesis [5, Ch. 3].

In this work, we analyse only some of those features: the average ST, the average HR, the standard deviation of the R-R intervals (SDNN – an HRV feature), the power of the EDA peaks (average amplitude of the peaks in one window, describing the phasic component), and the Inter-Subject Correlation (ISC) based on the average EDA signal (related to the tonic component). Higher inter-subject correlation is related to synchronised engagement between students [17].

Physiological signals' ranges and values vary between people, which does not directly allow for comparing their signals. Thus, the physiological signals need to be standardised to compare the participants. We normalise the different signals using the following formula, also called *min-max normalisation* [1, Ch. 5]:

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

In the equation, x stands for the starting signal value, x' for the normalised signal value, and *min()* and *max()* for the minimum and maximum signal values for each session for each individual. Signals with negative values were shifted into the positive-value range by adding *abs(min(x))* to every sample in the signal. Therefore, the final values of all signals were between 0 and 1.

4 RESULTS AND DISCUSSION

This section will show our preliminary analysis of the features related to each physiological signal available in LAUREATE. We will evaluate each signal's quality by examining the differences between the different types of sessions and follow them with a brief discussion. First, we will cover the ST signal. Then, we will explore heart rate and HRV features. Finally, we will finish our analysis with the EDA features. The study of the accelerometer signal is out of the scope of the present work.

4.1 Skin temperature

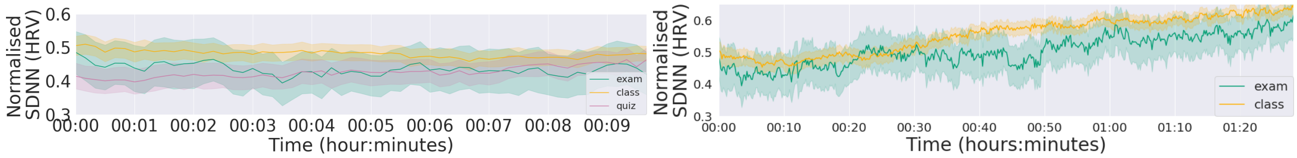
Figure 2 shows the mean skin temperature of all of the participants in two different recording sessions. Because session blocks (in particular, *breaks*) differ in length and the moment they occur, we show their impact by presenting the signal with two examples, as an aggregated visualisation would not allow for this.

Both figures 2a and 2b show the effects of room temperature on skin temperature. At the beginning of the session, when participants enter the classroom, there is a period of adjustment to the new ambient temperature. Data collection was done during the March-June period in the northern hemisphere (i.e., late winter and spring), with classrooms typically heated. The images show that the classroom temperature was effectively higher than the ambient temperature where the students came from (e.g. hallway, outdoors). We can also see the impact of *breaks* on skin temperature: as a consequence of some students leaving the classroom, skin temperature abruptly drops. Nevertheless, the values never reach the initial levels, probably because not all students left the classroom.

Given these results, in some of the following comparisons, we decided to distinguish between the types of *class* blocks: *class-start*, *class-after-quiz* and *class-after-break*. Findings like these demonstrate the need for in-the-wild datasets. For example, many studies suggest the existence of a clear relationship between affective states and skin temperature, e.g., colder extremities, due to blood-flow redirection towards the vital organs during a stressful event. However, our data show that the spatial and temporal contexts also play a significant role.

4.2 Heart Rate and HRV

Figure 3a shows students' mean SDNN (HRV feature) during 10 minutes of each session. Because the shortest quiz is 10 minutes long, we compare the different sessions over that period length. Furthermore, to avoid possible differences arising from students'



(a) **(b)**

Figure 3: Min-max normalised mean SDNN (with shaded 95% confidence interval) of all students. The lines represent the different blocks, i.e. *quizzes*, *exams* and *classes*. Fig. a only shows the mean from the first 10 minutes of the sessions. Only *class-start* blocks are considered for classes, but not *class-after-break* nor *class-after-quiz* blocks. Fig. b shows the mean from 90 minutes of all recording sessions. Only *exam* and *class* sessions (including all the different *class* blocks) are considered.

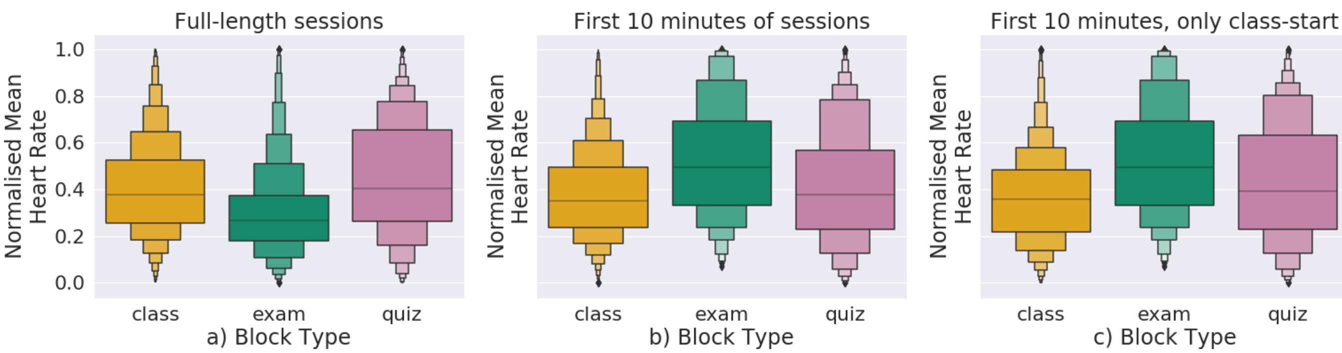


Figure 4: Distribution of mean heart rate values from lecturers' both hands. Values are min-max normalised per lecturer per session. Fig. a shows the comparison between all *classes*, *exams* and *quizzes* blocks. Fig. b shows data from the first 10 minutes of each block type. Fig. c shows data from the first 10 minutes of each recording session, i.e., only *class-start* blocks are considered, and neither *class-after-quiz* nor *class-after-break* blocks are used.

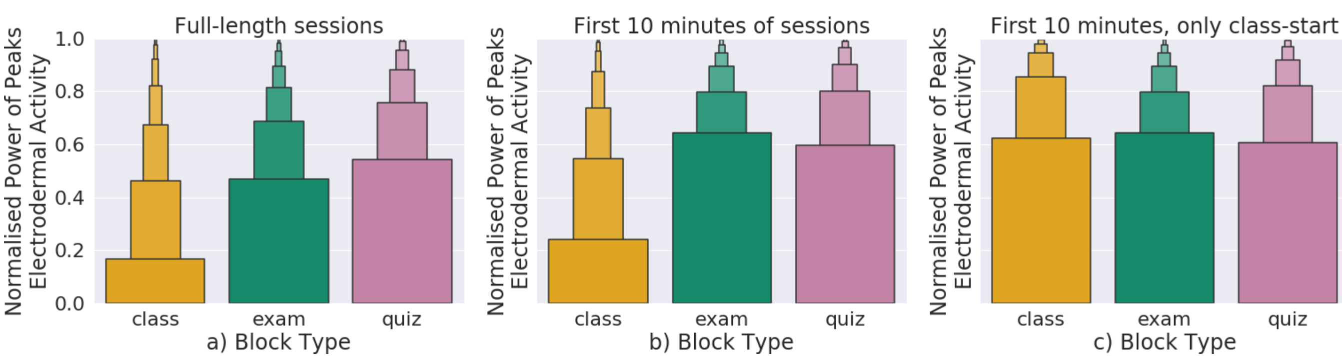


Figure 5: Distribution of the power of the peaks for the electrodermal activity. Fig. a shows the comparison between all *classes*, *exams* and *quizzes* blocks. Fig. b shows data from the first 10 minutes of each block type. Fig. c shows data from the first 10 minutes of each recording session, i.e., only *class-start* blocks are considered, and neither *class-after-quiz* nor *class-after-break* are used.

adaptation to the classroom ambient (i.e., sitting after being physically active while walking to the classroom), we only consider the first 10 minutes of each session. In addition, we only use *class-start* blocks, and we do not use *class-after-quiz* nor *class-after-break* blocks.

Figure 3 shows that students have, on average, a higher heart-rate variability when they are in a *class* block compared to an *exam* or *quiz*. Lower HRV values correlate with periods of higher stress, cognitive load and attention [6, 8, 13], like those happening at the start of an examination (*quiz*, *exam*). The difference decreases at the end of the 10-minute periods, possibly due to students' relaxation after the stressful initial period of facing an examination.

After the findings of Figure 3a, we decided to see if the differences in heart rate variability were maintained over long periods. Figure 3b shows the mean HRV of all students during 90 minutes of a recording session. Therefore, we only consider sessions with *exam*

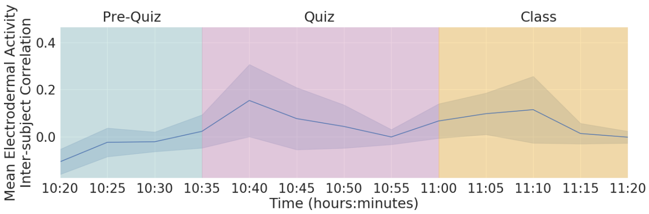


Figure 6: Inter-subject correlation of all students from a session with *pre-quiz, quiz* and *class* blocks. The shaded area represents the standard deviation.

and *class* (*class-start*, *class-after-quiz*, *class-after-break*, and *break* included) blocks. We can observe in Figure 3b that when comparing *classes* and *exams*, HRV is consistently lower in the latter. Still, the HRV values for both conditions increase as time passes.

Figure 4 shows the distribution of the mean heart rate values from lecturers' both hands. The values are normalised for each lecturer per session. Figure 4.a shows the comparison between all *class*, *exam* and quiz blocks, regardless of length and occurrence in a session. Figure 4.b shows the comparison between the first 10 minutes of each session, i.e. Here, we consider the 10-minute periods after a *quiz* or *break* as a *class* block. Figure 4.c only considers *class-start* blocks and not the *class-after-quiz* nor *class-after-break* blocks.

Figure 4 shows that lecturers' mean heart rate values are generally not affected by the block type. The differences in Figure 4.a may be attributed to *exam* sessions being longer than *class* (which are interrupted by *breaks*) and *quiz* sessions. *class* blocks are an active period for lecturers, as they walk and gesture while presenting. On the other hand, *exam* blocks are calmer, as the lecturers only react to questions from the students. When considering only 10 minutes of data, as in Figure 4.b and Figure 4.c, the difference between the block types disappears. A possible reason for this in the *exam* condition is that we are considering shorter length periods. A 10-minute period does not give the lecturers enough time to decrease their heart rate since the start of the exam, which entails prior physical activity (e.g. arriving to the classroom, setting out the desks and exams, sitting students).

4.3 Electrodermal Activity

Figure 5 shows the distribution of the power of the peaks of the EDA signal across all sessions. The categorisation resembles the one found in Figure 4: Figure 5.a shows the mean calculated over the entire length of all blocks (*class, exam, quiz*), Figure 5.b shows the mean calculated over the first 10 minutes of each block type, and in Figure 5.c we only consider *class-start* blocks, discarding *class-after-quiz* and *class-after-break* blocks.

Figure 5.c shows that the first 10 minutes of a recording session are similar for the students, irrespective of their situation (*class*, *exam*, *quiz*). As room temperature impacts sweating rate measurements, we justify this similarity due to the adaptation of the students to the classroom ambient, as previously discussed in Section 4.1 "Skin temperature". Therefore, we believe that Figure 5.a is the most suitable way to compare the power of the peaks of electrodermal activity in longer sessions, i.e. while comparing *classes* and *exams*[1]. As Figure 5.a considers the entire *class* and *exam* sessions length, students had enough time to adapt to the classroom ambient. Therefore, we could posit that the difference seen in the data distribution occurs due to the change in conditions. We can see that in the *exam* condition, the power of the peaks tends to have higher values, representing a situation where the participants are subject to more stressful conditions.

Figure 6 shows the inter-subject correlation (ISC) of the entire class for the EDA signal during one recording session. The session shown in the plot started with a *quiz* (including its preparation phase, i.e. *pre-quiz*) and was followed by a *class* block. During the first part of the *class* block, the lecturer focused on explaining the correct answers for the quiz.

Figure 6 shows an increment in the ISC of the students' sweating rate in two moments of the recording session. Moments of higher ISC are related to the synchronised engagement between the participants [17]. The first peak in the signal occurs at the start of the *quiz*. The increase in the signal is not visible immediately due to two possible reasons. The first possibility is that electrodermal activity responses are not instantaneous and occur sometime after the stimulus. Nevertheless, the response's delay is only a couple of seconds [18], so it does not fully explain the delay seen in Figure 6. The second and more substantial reason is that we calculated the ISC values every 5 minutes to reduce processing time, as we offered students visualisations of their data after each session. Therefore, the response (increased sweating rate) to the stimulus (start of the quiz) may be seen up to 5 minutes later. The second peak of the ISC occurs during the explanations of the correct answers. We speculate that the students were highly engaged during these periods, resulting in an increased ISC.

5 LIMITATIONS AND FUTURE WORK

The objective of the present work is to briefly introduce the dataset to the research community. To demonstrate its possibilities, we showed some examples with tentative explanations for the causes of the results we visualised. Nevertheless, these preliminary analyses lack the necessary rigour to make any stronger statement or conclusion.

We believe that the dataset offers many possibilities in several domains, including: human augmentation (e.g., memory augmentation interventions), affective computing (e.g., modelling stress, cognitive load and emotions), privacy (e.g., user identification), modelling behaviour (e.g., exercise level monitoring, sleep and food intake analysis) and learning performance (e.g., grade prediction), context recognition (e.g., class vs exam vs breaks), multimedia signal analysis (e.g., speech and video processing), and the development of multimodal machine learning algorithms (e.g., combining wearable, audio and video data).

6 CONCLUSION

In this work, we introduced the LAUREATE dataset, a Longitudinal multimodAl stUdent expeRience datasEt for AffecT and mEmory. We described the data collection procedure and the dataset's contents in Section 3. We briefly showed the results of our preliminary

[1]Because *quizzes* are at most 20 minutes long, their distribution doesn't vary much between the different conditions.

analysis and explained possible reasons for them in Section 4. We consider that these results are a good indicator of the quality of the data and its potential for the research community. Furthermore, we believe that the dataset is versatile with applications in several domains, some of them covered in Section 5.

The dataset still needs to be fully anonymised and cleaned before release. We plan to make the dataset available upon request after the signature of a sharing agreement. Please contact the main author for updates on this matter.

ACKNOWLEDGMENTS

We thank study participants for volunteering to be a part of the experiment.

REFERENCES

[1] Andriy Burkov. 2019. *The Hundred-Page Machine Learning Book.* Andriy Burkov, Polen.

[2] Malcolm Dcosta. 2018. SIMULATOR STUDY I: A Multimodal Dataset for Various Forms of Distracted Driving. (2018). https://doi.org/10.17605/OSF.IO/C42CN

[3] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (Sept. 2018), 103:1–103:21. https://doi.org/10.1145/3264913

[4] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2019. Laughter Recognition Using Non-invasive Wearable Devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'19)*. Association for Computing Machinery, New York, NY, USA, 262–271. https://doi.org/10.1145/3329189.3329216

[5] Martin Gjoreski. 2020. *A Fusion of Classical and Deep Machine Learning for Mobile Health and Behavior Monitoring with Wearable Sensors.* Ph. D. Dissertation. Jožef Stefan International Postgraduate School, Ljubljana, Slovenia.

[6] Martin Gjoreski, Tine Kolenik, Timotej Knez, Mitja Luštrek, Matjaž Gams, Hristijan Gjoreski, and Veljko Pejović. 2020. Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits. *Applied Sciences* 10, 11 (May 2020), 3843. https://doi.org/10.3390/app10113843

[7] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring Stress with a Wrist Device Using Context. *Journal of Biomedical Informatics* 73 (Sept. 2017), 159–170. https://doi.org/10.1016/j.jbi.2017.08.006

[8] Anita Lill Hansen, Bjørn Helge Johnsen, and Julian F. Thayer. 2003. Vagal Influence on Working Memory and Attention. *International Journal of Psychophysiology* 48, 3 (June 2003), 263–274. https://doi.org/10.1016/S0167-8760(03)00073-4

[9] J.A. Healey and R.W. Picard. 2005. Detecting Stress during Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (June 2005), 156–166. https://doi.org/10.1109/TITS.2005.848368

[10] Douglas Heaven. 2020. Why Faces Don't Always Tell the Truth about Feelings. *Nature* 578, 7796 (Feb. 2020), 502–505.

[11] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (Jan. 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[12] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. 2019. CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE, Varna, Bulgaria, 1–4. https://doi.org/10.1109/BIA48344.2019.8967457

[13] H. C. Middleton, A. Sharma, D. Agouzoul, B. J. Sahakian, and T. W. Robbins. 1999. Contrasts between the Cardiovascular Concomitants of Tests of Planning and Attention. *Psychophysiology* 36, 5 (Sept. 1999), 610–618. https://doi.org/10.1111/1469-8986.3650610

[14] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* 12, 2 (April 2021), 479–493. https://doi.org/10.1109/TAFFC.2018.2884461

[15] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 147:1–147:29. https://doi.org/10.1145/3432220

[16] OECD. 2020. *Health at a Glance: Europe 2020: State of Health in the EU Cycle.* Organisation for Economic Co-operation and Development, Paris.

[17] Pauline Pérez, Jens Madsen, Leah Banellis, Başak Türker, Federico Raimondo, Vincent Perlbarg, Melanie Valente, Marie-Cécile Niérat, Louis Puybasset, Lionel Naccache, et al. 2021. Conscious processing of narrative stimuli synchronizes heart rate between individuals. *Cell Reports* 36, 11 (2021), 109692.

[18] R. W. Picard and J. Healey. 1997. Affective Wearables. *Personal Technologies* 1, 4 (Dec. 1997), 231–240. https://doi.org/10.1007/BF01682026

[19] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, Boulder CO USA, 400–408. https://doi.org/10.1145/3242969.3242985

[20] Stefan Schneegass, Bastian Pfleging, Nora Broy, Albrecht Schmidt, and Frederik Heinrich. 2013. A Data Set of Real World Driving to Assess Driver Workload. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '13*. ACM Press, Eindhoven, Netherlands, 150–157. https://doi.org/10.1145/2516540.2516561

[21] Mohammad Faridul Haque Siddiqui, Parashar Dhakal, Xiaoli Yang, and Ahmad Y. Javaid. 2022. A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database. *Multimodal Technologies and Interaction* 6, 6 (June 2022), 47. https://doi.org/10.3390/mti6060047

[22] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. 2018. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing* 9, 2 (April 2018), 147–160. https://doi.org/10.1109/TAFFC.2016.2625250

[23] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2020. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* 11, 2 (April 2020), 200–213. https://doi.org/10.1109/TAFFC.2017.2784832

[24] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*. ACM Press, Seattle, Washington, 3–14. https://doi.org/10.1145/2632048.2632054

[25] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. 2022. A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. *Information Fusion* 83–84 (July 2022), 19–52. https://doi.org/10.1016/j.inffus.2022.03.009

[26] Zheng Zhang, Jeff M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. 2016. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3438–3446.