2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 978-1-6654-5381-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/PERCOMWORKSHOPS56833.2023.10150338

Cognitive and Emotional Monitoring with Inexpensive Wrist-Worn Consumer-Grade Wearables

Yunan Wu
Northwestern University
Evanston, IL, USA
yunanwu2020@u.northwestern.edu

Roxana Valdez
Bowdoin College
Brunswick, ME, USA
rvaldez@bowdoin.edu

Clifton Forlines
Northeastern University
Portland, ME, USA
c.forlines@northeastern.edu

***Abstract*—High-cost and physical-limitations limit the use of Electroencephalogram sensing (EEG) with respect to Affective Computing. Wrist-worn Wearable health devices, an intriguing alternative, are becoming cheap, inconspicuous, and commonplace - driving down the cost of heart-rate, accelerometer, respiration, and similar sensing. In this paper, we describe a quantitative empirical study investigating the use of low-cost sensing as a replacement for EEG. Our results show significant correlation between wrist-sensed, low-cost physiological data features and established EEG measures of emotion and cognition. This work demonstrates the value of inexpensive sensors capable of inferring emotion and cognition in a wrist-worn form-factor.**

***Index Terms*—physiological sensing, emotion, cognition, Affective Computing**

I. Introduction and Motivation

Electroencephalogram sensing (EEG) enables many Affective Computing interfaces by providing direct measurements of brain activity and thus insight into user emotion and cognition. This said, EEG's high-cost, long setup times, sensitivity to head and body movement, and uncomfortable form factor limit its applicability in many circumstances.

Consumer-grade, wrist-worn wearable health tracking devices (e.g. from FitBit, Garmen, Apple, Samsung, Google) represent an intriguing alternative to EEG. The large and growing consumer market for these devices is driving down the cost of and increasing the availability of physiological sensing such as heart-rate, accelerometer (ACC), respiration, galvanic skin response (GSR), and the like. While health tracking devices aim to use this data to provide consumers with insights into their health and activities, we hypothesize that these same sensors can be used to provide Affective Computing systems insight into users' cognition and emotional state. Prior efforts into inferring emotion and cognition from physiological data have reported mixed results; however, there are reasons to be optimistic. Prior work has shown a correlation in some circumstances between cognitive load and biometric measurements [1]–[6] as well as detectable changes in physiology based on certain emotions [2], [3], [7]–[15].

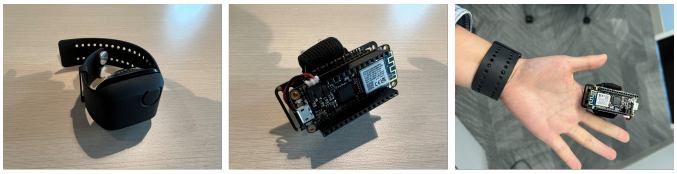
In this paper, we present an exploratory study in which we use low-cost physiological sensing as an alternative to high-cost EEG sensing. After an overview of prior efforts, we then describe the data-collection process and the pairing of physiological measurements with "Ground Truth" labels of cognition and emotion as reported by a commercial EEG device. This labeled physiological data is then used to train a model, the performance of which is evaluated and shown to be in-line with the EEG device itself.

II. Related Work

Emotions are complex, and a major challenge when studying this field is how to elicit specific emotions in a laboratory. Databases such as DEAP [7], MAHNOB-HCI [8], IAPS [16], and IADS [17] facilitate research by providing emotionally labeled stimuli; however, the validity of such collections of labeled stimuli has been questioned. Many have noted that one's emotions are not constant and that they not only vary in intensity but also are effected by one's environment, mood, experiences, recent stimuli, and so on. One study [9] noted the "large variability in emotional ratings [of stimuli among] participants", adding that DEAP videos did not induce strong emotions and that MAHNOB-HCI produced emotions that were unaligned with labels. With respect to IAPS and IADS, the expected and elicited emotions vary significantly [18] and stimuli often fail to elicit any emotional response at all [10].

EEG's ability to classify emotions and cognitive states is well established and provides the sensing backbone of numerous research projects. EEG headsets (e.g. [2]) ship with software that (while proprietary) provides validated measurements of emotion and cognition and have been used to investigate such topics as the emotional effects of music [11] and video games [12]. While an EEG's scalp electrodes provide direct measurements of the brain's electrical activity, the technology is relatively high-cost, is prone to failure due to head-movement and eye blinking [19], and is often uncomfortable - qualities that limit EEG's applications.

Of particular relevance to this paper are efforts into using non-EEG physiological sensors to infer emotion and cognition. With respect to these less-expensive options, the results from prior work have been mixed. Magdin et al. [15] measured classification accuracy with a commercial video-based tool and observed a small increase in accuracy when they augmented video with with heart-rate and GSR measurements. Gupta et al. [3] paired EEG measurements with GSR and heart-rate variance (HRV) to measure trust in a virtual agent and cognitive load while interacting with it. While EEG proved useful in predicting cognitive load, they found no predictive



a. Empatica E4 **b.** Emotibit **c.** Placement of E4 and Emotibit

Fig. 1. Wearable devices and the placement of the E4 and the Emotibit.

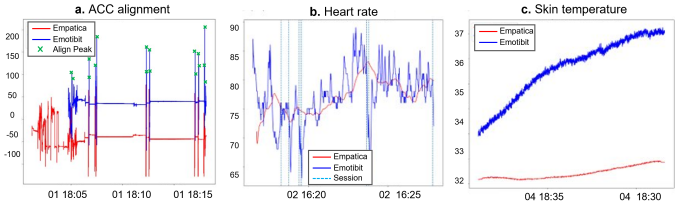


Fig. 2. Examples of the same measurement from the E4 and the Emotibit.

power from physiological measurements. In a larger follow-up study [4], EEG was again helpful in measuring cognitive load, HRV provided no insight and GSR provided only limited predictive power. In contrast, Shi et al. [5] and Khawaji et al. [6] showed preliminary success in estimating cognitive load from GSR signals in some circumstances.

Szwoch [10] reports on a preliminary investigation into physiological signal analysis using participants' blood volume pulse (BVP), skin temperature (T), breathing (Resp), skin conductance (SC), and muscle movement (EMG). Only three features derived from BVP and SC were found to support emotion classification. Godin et al. [9] performed a highly-relevant analysis of both the DEAP and MAHNOB datasets that achieving similar emotional classification results to prior attempts [7], [8] but did so using only a small number of the most relevant physiological-derived features. The authors point out that their (and previous) classification accuracy is "only slightly higher than random classifiers." They continue by noting that even the highest correlations between physiological sensor data and labeled emotion in the datasets were low.

III. Exploratory Study

Our original plan was to perform this exploratory study using a collection of wrist-worn, commercial health-tracking wearables; however, the many devices we investigated all limited direct access to data collected by some or all of their sensors. As such, we settled on two wearable devices that include similarly inexpensive sensors but are targeted at researchers working in this area - the Empatica E4 [20] (Fig.1a) and the Emotibit [21] (Fig.1b).

The E4 shares the same form-factor as many COTS devices and fits on the wrist of one's non-dominant hand (Fig.1c). Its collection of sensors provide real-time measures of heart rate variability (HRV), skin temperature, electrodermal activity (EDA), heart rate (HR), inter-beat interval (IBI), blood volume pulse (BVP), and 6 degree of freedom (DOF) accelerometer (ACC) data. The Emotibit is a flexible form-factor in respect to placement. While the index finger is not ideal for a deployed system (Fig.1c), we found that this location provided the best data quality when paired with the wrist-worn E4 and freed the dominant hand for interacting with experimental activities. The Emotibit provided a similar collection of sensors, adding humidity and a 9-axis IMU.

The use of two physiological devices in this work provided several benefits. Firstly, low-cost sensors often do not have the accuracy of medical-grade devices. For example, as shown in Fig.2b, sensors exhibit noise and drop-outs. The use of multiple sensors measuring the same physiology can help compensate for such missing data and bias from the sensors themselves or their placement. In other cases, "redundant" measurements may be highly-correlated between devices (e.g. skin-temperature in Fig.2c). In this case, different sensor positions result in different temperature profiles which should aid computational modeling.

Prior work has pointed out the difficulty of synchronizing physiological signals from multiple devices [3]. Because our devices were worn on the same hand, we were able to register timestamps from these independent devices by aligning peaks in their ACC data (Fig.2a).

As discussed previously, our goal was to attempt to model the cognitive and emotional measurements provided by a COTS EEG headset using only low-cost physiological data. We chose the Emotiv EPOC X [2] as our ground-truth device. The EPOC X is a 14-channel wireless EEG that is paired with software libraries to provide six calculated inferences of emotion and cognition - engagement, excitement, focus, interest, stress, and relaxation.

1) Participants: We recruited 7 volunteers (gender: 4 females and 3 males, Age: 31 ± 9) from within our organization. One volunteer participated on three different days and two volunteers on two different days, for a total of 11 sessions.

2) Stimulus: Each session consisted of two activities. The first activity was passively watching two music videos, which were selected to elicit different emotional responses. One contained calming music with scenes of nature, the other heavy-metal music with frenetic imagery. The second activity was a standard assessment of working memory in cognitive neuroscience named N-back [22]. In this game, participants view an image sequence and respond whenever the currently presented image is the same as the one presented N steps ago in that sequence. Our game sessions included 1, 2, and 3-back sequences giving us a range of cognitive demands.

3) Summary of the Collected Data: Overall, we collected multivariate time series data from each session for all variations of the game and video activities. Recorded physiological variables included 3-axis Accelerometer (ACC), Blood volume pulse (BVP), skin temperature (TEMP), electrodermal activity (EDA), heart rate (HR) and inter-beat interval (IBI) signals from the E4 and ACC, EDA, 3-axis Gyroscope (GYR), 3-axis Magnetometer (MAG), Photoplethysmography (PPG), TEMP and Humidity signals from the Emotibit. EEG derived "Ground Truth" scores for engagement, excitement, frustration, focus, stress and relaxation were collected from the Emotiv and its included software libraries.

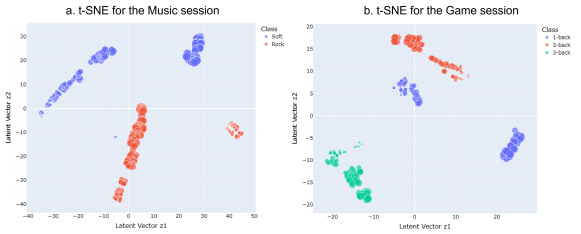


Fig. 3. t-SNE plots of the Music Video session and the Game session in a 2D visualization. The datapoints were sized by the timestamp. The smaller the size, the earlier the time.

IV. Analysis and Results

A. Data Segmentation and Physiological Settling Time

As our sensors were running independently over the course of each session, the first step in our analysis was to accurately align and segment our data so that we could focus on meaningful portions and ignore periods during or near transitions between tasks.

Participants were asked to raise their hand before and after each activity - creating a large spike in acceleration that is easily identified (Fig 2a). These peaks allowed us to align the independent time stamps reported by each device and served as a course break between session activities.

With this coarse segmentation in place, we initially applied an unsupervised learning algorithm, t-distributed stochastic neighbor embedding (t-SNE), to the entirety of all segments in order to visualize the extracted physiological features in a 2D space, as shown in Fig.3a (additional details on feature extraction are in the next section). Clear clusters emerged in the plots; however, rather than the expected two clusters for our two video conditions and three clusters for our three gaming conditions, we observed four and six clusters respectively. Only after visualizing the timestamps of these datapoints did it become clear that the extra clusters consisted of data from the start of each session (Fig.3b). We hypothesize that this period represents a "Physiological Settling Time" during which a participant's physiology is adjusting to new stimulus. As such, we adjusted the start and end time point in the t-SNE plot so that there were only two clusters for video and three clusters for gaming data, giving us an optimal boundary between each task. These boundaries were then applied to our "Ground Truth" EEG measurements before further analysis occurred.

B. Feature Extraction and Data Cleaning

With respect to feature extraction from our multivariate time-series sensor data, in all cases we first applied an eight-second sliding window with a time step of one second. We chose this window by ablation studies after taking into consideration the response delay of physiological signals.

For ACC, GYR, and MAG data, we extracted basic statistical features in each axis, including mean, standard deviation, number of peaks and entropy. Similarly, we extracted basic statistical features from the TEMP and humidity.

The EDA-based features reflect activities of the sympathetic nervous system, as measured by the skin conductance derived from microscopic changes in the level of perspiration on the skin surface. Previous studies suggest that EDA measurements were closely related to the detection of emotional changes and engagement [23]–[25]. The EDA signal includes two types of changes, a phasic change, in which rapid changes in EDA responses occur and a tonic change, in which smooth and gradual changes in EDA responses occur. We decomposed the EDA signal into phasic and tonic data following [26] and extracted statistical features, such as number of peaks, mean, standard deviation, and min/max from each type of change.

HRV is a physiological phenomenon of variation in the time interval between heartbeats and is often used to evaluate human emotion and cognition performance [27], [28]. We extracted the HRV-based features from both the time and frequency domains. For example, in the time domain, we extracted basic features such as mean, min, and max. Features that reflected long-term effects, such as the standard deviation of successive differences (SDSD) and the number of consecutive interval pairs that differs by more than 50 ms (NN50) could also be extracted. We computed the frequency data by applying the Fast Fourier Transform (FFT) where features, such as the power within different frequency bands and the ratio of the low frequency power to the high frequency power, could be extracted. In addition, we calculated non-linear EDA features, such as SD1 and SD2 and basic statistical features, such as mean, stdev, and number of peaks in the IBI signal according to [29].

In addition to the six measurements of emotion and cognition, our EEG sensor reported continuous measures of signal quality. After calculating physiological features, we discarded time points during which EEG quality fell below 80% to ensure the reliability of our "Ground Truth" labels. Missing data was replaced with the mean value for that feature and all data was scaled to [0,1] using min-max normalization. After all segmentation and feature extraction, we continued our statistical analysis with a total of 4020 data-points for video sessions and 1800 data-points for game sessions.

C. Examining "Ground Truth"

As discussed in II, problems arise with stimulus labels as identical stimulus can evoke different emotions from different people or even different emotions from the same person at different times. As such, we designed our investigation to use established EEG measurements of six emotional and cognitive scores as "Ground Truth" labels. We did not attempt to validate the EEG's measurements themselves; rather, we were interested in whether or not the EEG demonstrated an ability to distinguish among our stimuli using these six measures. If our "Ground Truth" EEG measurements could not themselves correctly discriminate calming vs. rock music videos or three distinct levels of gaming difficulty, then modeling these measurements using low-cost physiological data would be useless.

To test the discriminative power of the "Ground Truth" EEG measurements, we performed an ANOVA on all pairs of stimuli. Table I shows probability-values for each pairwise comparison of mean EEG emotion and cognition scores for

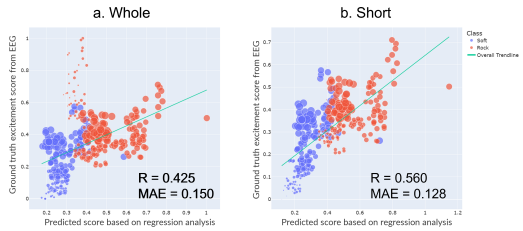


Fig. 4. The correlation plot of a testing participant on the prediction of the *Excitement* score for the Music Video session using the full time series (a) and a short time series that getting rid of the transition time (b). The red circle represent rock music, the blue soft music. The size of the circles stands for the time series. The smaller the size, the early the time.

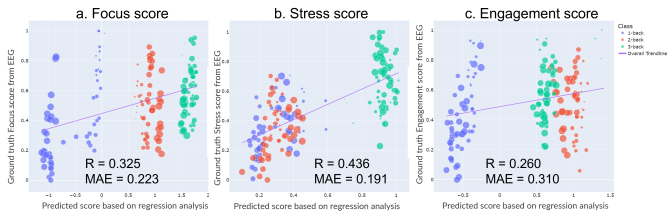


Fig. 5. The correlation plot of a testing participant using the full time series on the predictions of the *Focus* score (a), the *Stress* score (b) and the *Engagement* score (c) for the Game session. The green circle stands for 3-back, the red circle stands for 2-back and the blue circle stands for 1-back. The size of the circles stands for the time series. The smaller the size, the early the time.

each of our experimental activities. With respect to the two video conditions, EEG measures of $Excitement$ and $Interest$ appear significantly different between calming and rock music. With respect to the three gaming conditions, N-Back levels 1, 2, and 3 all appear distinguishable by EEG measures of $Engagement$, $Stress$, and $Focus$. We continued our analysis of low-cost physiological data focusing on these activity-measure combinations.

D. Machine Learning Models

With segmented and trimmed time-series data matched with EEG-derived measures of emotion and cognition, we began the construction of our ML models – mapping low-cost physiological data to our "Ground Truth" measurements. A each time point in our data-set, we had a total of 184 physiological features and the EEG's corresponding cognitive (i.e., *Focus*, *Stress* and *Engagement*) and emotional (i.e., *Excitement* and *Interest*) scores. Our approach to model building was three-fold - select the most important features, train a linear regression model on those features, and evaluate model performance. All analysis in the following sections was performed using the scikit-learn packages in Python 3.6.

1) Feature Selection: An ensemble gradient boosting regressor was used to select the top-k important features in the data. We chose mean square error (MSE) as the loss function and performed a cross-validated grid-search to determine the number of trees, the learning rate, and the tree depth. This process resulted in an importance score for each feature that indicates its value in building decision-trees within the model.

We found that HRV-related features were good indicators of the *Focus* score while EDA-related features were good predictors of the *Excitement* and *Stress* scores. Additionally, the *Engagement* score and the *Interest* score were more correlated with ACC-related features and HR-related features.

2) Model Training and Performance: We investigated several techniques, but settled on linear regression due to its solid performance in this task and the relatively small size of our data-set. We predicted *Excitement* and *Interest* for the music video activities as well as *Focus*, *Engagement*, and *Stress* for the game activities from the top-k selected features.

The results of these tests are shown in Table II, column Whole. The average and the standard deviation of each metric were calculated based on Leave-one-person-out approach.

These results indicate that the predictions derived from the low-cost physiological sensors were correlated with the ground truth as reported by the EEG headset. Specifically, for the music video activities, the prediction for the *Excitement* score achieved a correlation (R) of 0.509 and a mean absolute error (MAE) of 0.150, followed by the prediction for the *Interest* score with a R of 0.330 and an MAE of 0.226. For the Game activities, the prediction for the *Focus* score achieved a R of 0.374 and an MAE of 0.159, followed by predictions for the *Engagement* score with a R of 0.340, an MAE of 0.226 and for the *Stress* with a R of 0.316 and an MAE of 0.268.

Fig.4a shows a correlation figure for one of our participants in the music video activity, although the following description applies generally to almost all of our collected data. Upon examining Fig.4a, we again noticed that the largest errors occurred at the beginning and end of each activity. Fig.4b shows an improved correlation figure in which a shorter time period was used, effectively removing the physiological transition time from the analysis. Table II, column Short, shows the improved results when only data from this shorter time period was used for modeling. The improved correlation scores were statistically significant for *Excitement* and *Interest* in the music video activity and for *Stress* in the game activity.

V. Discussion

We collected low-cost bio-metric measurements from two devices along with established high-cost EEG "Ground Truth" measurements and built Machine Learning models to map one to the other. Our results suggested that features extracted from low-cost, wrist-worn sensors have the potential to be used as valid predictors of emotional and cognitive states.

A. Emotional and Cognitive Predictions

The Emotiv headset gave six scores in real-time as the ground truth and after the ANOVA test in Table I, only the *Excitement* score and the *Interest* score for the Music Video session and the *Focus*, the *Stress* and the *Engagement* scores for the Game session showed significant differences in music types and n-back difficulty, so we used the extracted features to predict these five scores. Specifically, as shown in Table II, the linear regression model predicted the *Excitement* score for the music session with a correlation of 0.509 and an MAE of 0.150. The results of one randomly selected participant were plotted in Fig.4a and all other participants

TABLE I
THE ANOVA TEST RESULTS ON THE GROUND TRUTH SCORES. THE COMPARISONS ARE SIGNIFICANT IF $p < 0.05$, HIGHLIGHTED IN BOLD.

Session	Class	Engagement	Excitement	Stress	Relaxation	Interest	Focus
Music Video	Soft & Rock	$p = 0.079$	p = **0.001**	$p = 0.204$	$p = 0.900$	p = **0.001**	$p = 0.344$
Game	1-back & 2-back	p = **0.033**	$p = 0.900$	p = **0.001**	$p = 0.666$	$p = 0.155$	p = **0.012**
	1-back & 3-back	p = **0.001**	$p = 0.083$	p = **0.001**	p = **0.009**	p = **0.001**	p = **0.001**
	2-back & 3-back	p = **0.001**	p = **0.031**	p = **0.0029**	p = **0.001**	p = **0.001**	p = **0.001**

TABLE II
LINEAR REGRESSION RESULTS FOR *Excitement*, *Interest*, *Focus*, *Stress* AND *Engagement* SCORE PREDICTIONS ON THE TEST SET (MEAN ± STD BASED ON LEAVE-ONE-PERSON-OUT APPROACH). WHOLE STANDS FOR THE WHOLE TIME SERIES, AND SHORT STANDS FOR THE TIME SERIES AFTER REMOVING THE TRANSITION TIME. * MEANS THE COMPARISON BETWEEN WHOLE AND SHORT IS STATISTICALLY SIGNIFICANT ($p < 0.05$).

Measurement	R		MAE	
	Whole	Short	Whole	Short
Music-Excitement	0.509 ± 0.056	0.615 ± 0.042 *	0.150 ± 0.053	0.131 ± 0.032 *
Music-Interest	0.330 ± 0.026	0.428 ± 0.019 *	0.226 ± 0.041	0.122 ± 0.026 *
Game-Focus	0.374 ± 0.045	0.372 ± 0.031	0.159 ± 0.027	0.158 ± 0.022
Game-Stress	0.316 ± 0.098	0.332 ± 0.087 *	0.268 ± 0.092	0.229 ± 0.078 *
Game-Engagement	0.340 ± 0.057	0.337 ± 0.034	0.226 ± 0.051	0.228 ± 0.039

showed similar results. The figure showed that the model could classify different time points into discriminative rock and soft clusters, with participants in the rock music session predicted to have higher arousal, which is consistent with the physiological phenomenon and previous studies [30]. The size of the circles in the figure represents the time sequence order; the smaller the size, the earlier the time. It is worth noting that the prediction error occurred mainly at the beginning and end of each session, i.e., the transition time between each session, although we have previously segmented the data to try to mitigate this effect, as shown in as shown in Fig.3. This finding is consistent with previous studies that physiological responses often take time to change after the arrival of a new stimulus [31]. Therefore, we further shortened each time session to get rid of these transition time, which gave us a significantly improved correlation of 0.615 and MAE of 0.131. Fig. 4b clearly shows the better prediction results. The same was true for the *Interest* score prediction. After getting rid of the transition time, the model improved significantly by from 0.330 to 0.428 in correlation and from 0.226 to 0.122 in MAE.

For the Game session, the model predicted the *Focus* score, the *Stress* score and the *Engagement* score with a correlation of 0.374, 0.316 and 0.340 and an MAE of 0.159, 0.268 and 0.226, respectively. One reason for the slightly worse results for the Game session compared to the *Excitement* score for the Music Video session was the short duration of the experiment. Each n-back game only took 1-2 minutes, which was not enough for the transition time, so it was difficult for participants to fully engage in each game. Notably, even with a short experimental time, the model predicted a good correlation with ground truth scores and classified the game with different complexity N as shown in Fig.5. However, experiments on shortening each time period showed significant improvements only in the prediction of *Stress* scores (R = 0.332, MAE = 0.229), while the results were even worse in the prediction of *Focus* (R = 0.372, MAE = 0.158) and *Engagement* scores (R = 0.337, MAE = 0.228).

It is interesting to note that our predictions had even more discriminative results than the ground truth when distinguishing between different time periods. For example, as shown in Fig. 4, for this randomly selected test participant, the distribution of the ground truth *Excitement* score was 0.215 ± 0.133 on the soft music session and 0.441 ± 0.160 on the rock music session while the predicted *Excitement* score was 0.265 ± 0.068 on the soft music session and 0.589 ± 0.140 on the rock music session. The means of the predicted scores for different music types were more discriminating, while the standard deviation values were smaller, which indicates that the physiological sensors may be more robust than the EEG headset. The same is true for the Game sessions. For example, as shown in Fig. 5b, the standard deviations of the ground truth *Stress* score on $N = 1, 2, 3$ were 0.162, 0.174, and 0.178 respectively while those of the predicted *Stress* score were 0.053, 0.050, and 0.022.

B. Threats to Validity and Future Work

As with all investigations, there are threats to the validity of this paper that the reader should be aware of. With respect to internal validity, it became clear to us during the analysis of our data that our session durations were too short. As others have also found [31], physiological measures have a time delay as different systems in our bodies respond at different time constants to stimuli. The need for longer sessions in order to allow physiology to "settle" before measurement must be balanced with the discomfort participants felt while wearing the EEG headset, discomfort that may have itself interfered with emotion and cognition.

With respect to external validity, the relatively small sample size of this investigation (7 individuals, 11 sessions) limits its impact. While we achieved our research goal of demonstrating the value of this sensor suite, clearly a larger sample is needed to confidently provide concrete recommendations for system design. The sample size also limited appropriate modeling methods. While we achieved significant results with a linear regression model, we hoped to improve these results using

Random Forests and 1D Convolutional Neural Networks in our ablation studies. With these techniques, overfitting occurred due to the large hyperparameter space and small sample size. Future work clearly requires collecting data from a larger population. Looking ahead, while we found value in using a pair of partially redundant sensors, future work should include multiple sensors and investigation into their optimal placement on the body for the task at hand. Finally, as others have pointed out [9], emotion and cognition are complex phenomenons that are difficult to label absolutely. Our exploration aimed to model EEG's reported measures, a goal we partially achieved; however, nothing in the nature of low-cost physiological sensors can add to the outstanding questions of the accuracy of the labels themselves.

VI. Conclusion

This paper presented the approach to building the performance of a linear regression model capable of mapping physiological signals measured by low-cost, wrist-worn, consumer-grade devices to cognitive and emotional measurements of excitement, focus, engagement and stress. We are happy to report that the model's inferred measurements were in-line with those reported by a more expensive, more cumbersome EEG headset and at times demonstrated greater discriminatory power than the EEG itself. Through this exploratory work, we feel confident that there is value at the low-end of sensing when it comes to supporting Affective Computing and that wrist-worn devices are an appropriate form-factor for such sensing. We are motivated to continue this line of research through future large-scale research studies.

References

[1] Ravi Naik, Alexandros Kogkas, Hutan Ashrafian, George Mylonas, and Ara Darzi, "The measurement of cog. workload in surgery using pupil metrics: A systematic review and narrative analysis," *J of Surgical Research*, vol. 280, pp. 258–272, 2022.

[2] "Emotiv," https://www.emotiv.com/.

[3] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst, "In ai we trust: Investigating the relationship between biosignals, trust and cognitive load in vr," in *25th ACM Symp on VR Software and Tech*, New York, NY, USA, 2019, ACM.

[4] Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst, "Measuring human trust in a virtual assistant using physiological sensing in virtual reality," in *2020 IEEE Conf on Virtual Reality and 3D User Interfaces*, 2020, pp. 756–765.

[5] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen, "Galvanic skin response (gsr) as an index of cognitive load," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2007, CHI EA '07, p. 2651–2656, ACM.

[6] Ahmad Khawaji, Jianlong Zhou, Fang Chen, and Nadine Marcus, "Using galvanic skin response (gsr) to measure trust and cognitive load in the text-chat environment," 04 2015, pp. 1989–1994.

[7] S Koelstra, C Muhl, M Soleymani, J Lee, A Yazdani, T Ebrahimi, T Pun, A Nijholt, and I Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Trans on Affective Comp*, vol. 3, no. 1, pp. 18–31, 2012.

[8] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[9] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal, "Selection of the most relevant physiological features for classifying emotion," in *Proc of the 2nd Int Conf on Physiological Computing Systems*, Setubal, PRT, 2015, PhyCS 2015, p. 17–25, SCITEPRESS - Science and Technology Publications, Lda.

[10] Wioleta Szwoch, "Emotion recognition using physiological signals," in *Proc of the Mulitimedia, Interaction, Design and Innnovation*, New York, NY, USA, 2015, MIDI '15, ACM.

[11] A Bhatti, M Majid, S Anwar, and B Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Comps in Human Behavior*, vol. 65, pp. 267–275, 2016.

[12] A. Khan and S. Rasool, "Game induced emotion analysis using electroencephalography," *Comps in Biology and Med*, vol. 145, 2022.

[13] Muhammad Khateeb, Syed Muhammad Anwar, and Majdi Alnowami, "Multi-domain feature fusion for emotion classification using deap dataset," *IEEE Access*, vol. 9, pp. 12134–12142, 2021.

[14] S Dargan and M Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities," *Expert Systems with Applications*, vol. 143, pp. 113114, 2020.

[15] Martin Magdin, Michal Kohútek, Štefan Koprda, and Zoltán Balogh, "Emosens – the proposal of system for recognition of emotion with sdk affectiva and various sensors," in *ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part I*. p. 400–411, Springer-Verlag.

[16] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al., "International affective picture system (iaps)," *The center for research in psychophysiology, University of Florida*, 1999.

[17] MM Bradley and PJ Lang, "The international affective digitized sounds: affective ratings of sounds and instruction manual (technical report no. b-3). university of florida," *NIMH Center for the Study of Emotion and Attention, Gainesville, FL*, 2007.

[18] Guillaume Chanel, Julien Kronegg, Didier Grandjean, and Thierry Pun, "Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals," in *Int workshop on multimedia content representation, classification and security*. Springer, 2006, pp. 530–537.

[19] Shi-Yun Shao, Kai-Quan Shen, Chong Jin Ong, and Einar PV Wilder-Smith, "Automatic eeg artifact removal: a weighted support vector machine approach with error correction," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 336–344, 2008.

[20] M Garbarino, M Lai, D Bender, R W Picard, and S Tognetti, "Empatica e3—a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *4th Int. Conf. on Wireless Mobile Comm. and Healthcare (MOBIHEALTH)*. IEEE, 2014, pp. 39–42.

[21] "Emotibit," https://www.emotibit.com/.

[22] Linette Lawlor-Savage and Vina M Goghari, "Dual n-back working memory training in healthy adults: a randomized comparison to processing speed training," *PloS one*, vol. 11, no. 4, pp. e0151817, 2016.

[23] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[24] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim, "n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild," *Proc of the ACM on Interactive, Mobile, Wearable and Ubiquitous Tech*, vol. 4, no. 3, pp. 1–26, 2020.

[25] Leslie Potter, J Scallon, Daniel Swegle, Trevor Gould, and GE Okudan Kremer, "Establishing a link between electrodermal activity and classroom engagement," in *IIE Annual Conf. Proc*. Institute of Industrial and Systems Engineers (IISE), 2019, pp. 988–993.

[26] Simon Föll, Martin Maritsch, Federica Spinola, Varun Mishra, Filipe Barata, Tobias Kowatsch, Elgar Fleisch, and Felix Wortmann, "Flirt: A feature generation toolkit for wearable data," *Comp Methods and Programs in Biomed*, vol. 212, pp. 106461, 2021.

[27] Kathy Y Liu, Thomas Elliott, Melanie Knowles, and Robert Howard, "Heart rate variability in relation to cognition and behavior in neurodegenerative diseases: A systematic review and meta-analysis," *Ageing research reviews*, vol. 73, pp. 101539, 2022.

[28] Mara Mather and Julian F Thayer, "How heart rate variability affects emotion regulation brain networks," *Current opinion in behavioral sciences*, vol. 19, pp. 98–104, 2018.

[29] Fred Shaffer and Jay P Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[30] Damian A Ritossa and Nikki S Rickard, "The relative utility of 'pleasantness' and 'liking'dimensions in predicting the emotions expressed by music," *Psychology of Music*, vol. 32, no. 1, pp. 5–22, 2004.

[31] Kenneth J Hunt, Simon E Fankhauser, and Jittima Saengsuwan, "Identification of heart rate dynamics during moderate-to-vigorous treadmill exercise," *Biomedical engineering online*, vol. 14, no. 1, pp. 1–13, 2015.