Stress and Amusement Detection with Wearables

Hunter Gregory, Grace Langford, Jaravee Boonchant

September 2020

1. Introduction

In the last ten years, there have been significant strides in the development of wearable devices. Wearable devices have existed for centuries, dating back to the invention of an air conditioned top hat in the Victorian era, but more recently the industry has seen an emergence of devices that track affect and activity using physiological and motion data[5].
Using wearables to detect and predict stress has the potential to minimize and prevent the negative health consequences associated with chronic stress. Stress responses are triggered by the Sympathetic Nervous System (SNS) (within the Autonomic Nervous System (ANS)) as reactions to environmental stimuli that are perceived as threatening. The release of cortisol and adrenaline causes accelerated heart rate, increased respiration, dilated pupils, perspiration, elevated body temperature, increased blood pressure and many more physical reactions[12]. The incremental effect of cortisol release over time can have detrimental effects on the human body. Repeated episodes of stress also encourage memory loss by weakening the hippocampal brain cell, resulting in "brain shrinkage." Studies have shown that it is linked to brain atrophy, compromises many physiological systems, and is highly correlated with chronic illness and reduces life expectancy[12].
Stress response acts through three main forms; physiological, behavioral, and affective[12]. Wearable devices attempt to detect stress through observable physiological data. However, because stress is highly subjective and presents differently in individuals, stress reactions are difficult to directly discern through observable physiological responses. This lack of a standard definition of stress presents as a major obstacle in the effective computing of stress responses with wearable devices[11].
The Wearable Stress and Affect Detection (WESAD) dataset was created in an attempt to provide a standardized multimodal dataset to aid in the development of empathetic machines based on human-computer interaction models[11]. We build on the work of the WESAD authors by considering the classification problem between stress and amusement using physiological data from wearable devices.

The aim of this paper is to use the WESAD dataset to investigate the following questions:

- Is sensor data useful in discriminating between stress and amusement conditions and if so how?
- Which types of sensor data are most useful in discriminating between stress and amusement (alone or in combination)?
- Can we detect stress only using the wrist-worn wearable device?
- Can we quantify the heterogeneity across individuals in the response to stress vs. amusement?

Our analysis found that we can classify stress vs. amusement with high accuracy using a combination of wrist-worn and chest-worn device data in combination. Classifying solely based on wrist-worn device data is less accurate. The most heterogeneity between individuals in response to stress vs. amusement was related to heart rate variability. Our model found that all chest and wrist sensor data have characteristics that contribute significantly to detecting stress vs amusement, except from electrodermal activity (EDA).
Section 2. of our paper will elaborate on the WESAD dataset, feature extraction from the data, and results of exploratory data analysis. In Section 3, we will describe our methodology for the logistic regression model

with features from both RespiBan and Empatica 4 (Full Model), as well as the logistic regression model with only features from Empatica 4 (Wrist Model). We will also analyse the results of both models in this section. In Section 4, we will provide model validations and sensitivity analysis and discuss the limitations and future investigations of our analysis. Lastly, in Section 5. we will develop conclusions of our analysis. The code will be attached in the Appendix.

2.Data and Exploratory Data Analysis

2.1 Data

The WESAD dataset is composed of multivariate time-series data of sensor modalities of 17 participants recorded from a chest-worn (RespiBan Professional) device and a wrist-worn device (Empatica E4) that were synced up together. Two participants did not complete the study and were therefore omitted from the data, leaving a remainder of 15 participants for our analysis. The study was conducted on graduate students and excluded pregnant women, heavy smokers, people with mental disorders, as well as people with chronic and cardiovascular diseases. Participants provided demographic information, a pre-study assessment, and a post-study survey to highlight any biases or errors in the reported data. The original authors of this dataset wanted to detect affective states with wearable devices. Participants were exposed to stimuli that evoked four different affective states– baseline, stress, amusement, meditation. Physiological metrics were used to measure ANS functioning, which can then be used to detect stress[11]. The physiological metrics and their corresponding sampling rates, in hertz (Hz), are provided in the Table 1 below:

Table 1: Breakdown of sensors used in the WESAD study

Sensor	Device	Sampling_Rate_Hz
Electrocardiogram (ECG)	RespiBan	700
Electrodermal Activity (EDA)	RespiBan	700
Electromyogram (EMG)	RespiBan	700
Respiration	RespiBan	700
Body Temperature	RespiBan	700
3-Axis Accelerometry (ACC)	RespiBan	700
Blood Volume Pulse (BVP)	Empatica E4	64
Electrodermal Activity (EDA)	Empatica E4	4
Skin Temperature	Empatica E4	4
3-Axis Accelerometry	Empatica E4	32

2.2 Feature Extraction

Schmidt et al performed classification on WESAD for baseline versus stress versus amusement as well as stress versus non-stress. For classification, they calculated features using a sliding window on the sensor data with a window size of 60 seconds for all sensors besides the accelerometer, for which they used a five second window. Since accelerometry data requires a smaller window size due to high variability and rapid changing quality, we decided to follow Schmidt et al's precedent. We experiment with different window shifts, but start with a ten second shift.

We processed the data using NeuroKit2, a Python module that was developed to allow researchers and clinicians to access biosignal processing routines and analyze physiological data in a comprehensible and computationally efficient way[10]. This package has been utilized for a multitude of studies and academic journal articles (ex:Jun et al.[7]).

All features were inspired by Schmidt et al. For body temperature (TEMP), we calculate mean, standard deviation, minimum, maximum, and slope from window start to finish over the raw signal. For the accelerometry (ACC), we calculate mean, standard deviation, and the absolute integral over time for all dimensions

and for the magnitude of the raw signal. For the other signals, we clean and then extract features using NeuroKit2. For RESP, we calculated the mean and standard deviation for inhalation and exhalation duration, the inhalation/exhalation ratio, volume (amount of air taken in), and breath rate (breaths per minute).
For electrodermal activity (EDA), we calculated the mean, standard deviation, minimum, maximum, slope, and features for the skin conductance level (SCL) and skin conductance response (SCR) components of the signal. EDA reflects the skin's response bursts and recovery to impetus. As described in Schmidt et al, the "SCL represents a slowly varying baseline conductivity, while the SCR is a short term response to a stimulus." We calculated correlation of SCL over time as well as the mean and standard deviation of SCL and SCR. We also calculated the number of segments, sum of startle magnitudes and startle response durations as well as the area under SCRs. See Healey et al. for a deeper description[6]. Additionally, we had to upsample the EDA wrist data from 4 Hz to 8 Hz due to a signal processing frequency requirement in Neurokit2. We performed linear interpolation to upsample.
We also convert BVP and ECG signals to Heart Rate (HR) and Heart Rate Variability using Neurokit2. We then calculate mean and standard deviation of HR and HRV, the percent of large intervals (RR intervals differing more than 50 ms), triangular interpolation index, root mean square of RR intervals, and ultra low (0.01-0.04 Hz), low (0.04-0.15 Hz), high (0.15-0.4 Hz), and ultra high (0.4-1.0 Hz) frequency bands' power spectral density.
Finally, we had to forgo EMG data since we couldn't resolve Neurokit2 processing errors with this signal.
In addition, we mean-centered and standardized all features we chose for a better interpretation of the models.
The dataset we use can be recreated on the Duke Compute Cluster by running the following command: sbatch /hpc/group/sta440-f20/hlg16/affect-detection-with-wearables/create-wesad-csv.sh. This will output the data to /work/hlg16/WESAD-one-sec-shift-round2.csv

2.3 Exploratory Data Analysis

First, we discarded HR/HRV frequency features since only high and ultra high bands were detected. We also immediately removed standard deviations, maxima, and minima since all were highly correlated with means.
As seen in the boxplot in Figure 1, mean chest heart rate variability (from ECG) increases overall when subjects are amused. We include this feature and the same one for BVP, yet the other HR/HRV features are highly correlated (e.g. 0.74 correlation between ECG mean HRV and ECG percent large intervals), so we do not include them. Similar boxplots and Pearson correlations were used throughout exploration.
For ACC, out of all dimensions, we included only the mean magnitude since the magnitude is an informative summary and generalizes better when comparing wrist and chest data. There doesn't seem to be any distinction between the two states in accelerometry. For TEMP, we keep slope and mean, as they seem to be higher for stressed states, and the correlation between the two is 0.01.
For RESP, we include breath rate and volume. There seems to be an increase in breath rate for amusement condition overall, and a decrease in volume. Inhale duration and exhale duration were clearly highly correlated with volume, so we omitted those.
For EDA, number of responses/segments is much higher when amused, so we include this and forego its very correlated counterparts (sum of startle magnitudes, sum of startle response durations, area under SCRs). We also included slope to provide information about the trend in EDA, and forego its highly correlated counterpart (SCL correlation with time). Finally, we didn't include EDA/SCL/SCR mean, since these are not as informative as the trend (via slope) and startle characteristics.
Figure 1 below shows an example of a feature that appears to have strong distinctions between stress and amusement responsesand Figure 2 below shows an example of a feature that appears to have weak distinctions between stress and amusement responses.

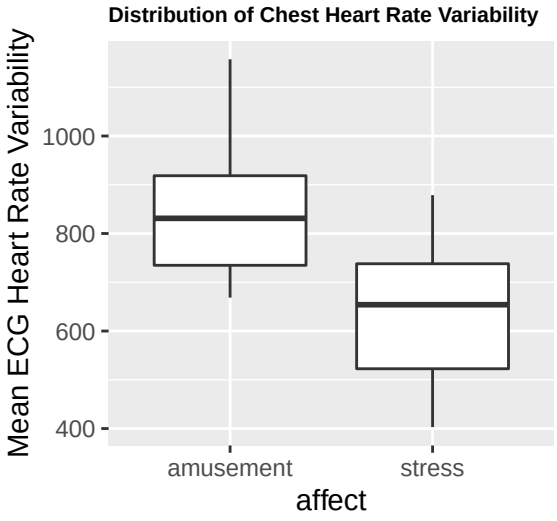

Figure 1

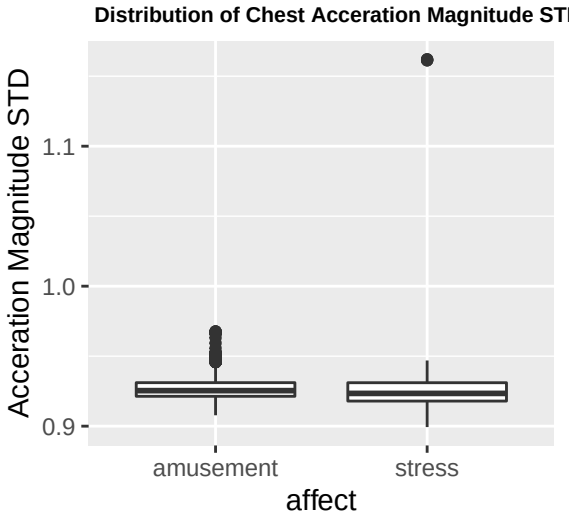

Figure 2

3. Analysis

3.1 Chest and Wrist Model

I. Methodology We randomly split the data into 80% for the training data and the remaining 20% for the test data. We perform a logistic regression model with the chosen 14 extracted features on our training data. The features we used in our model are shown in Table 2 below:

Table 2: Features used in our full model

Feature	Devices	Description
chest_EDA_slope, wrist_EDA_slope	RespiBan, E4	Rate of change in EDA
chest_RESP_breath_rate	RespiBan	Breath per Minute
chest_RESP_volume	RespiBan	Amount volume of air inhale per Minute
chest_SCR_num_segments, wrist_SCR_num_segments	RespiBan, E4	Number of skin conductivity responses (SCR)
chest_ACC_magnitude_mean, wrist_ACC_magnitude_mean	RespiBan, E4	Average magnitude of acceleration
chest_TEMP_mean, wrist_TEMP_mean	RespiBan, E4	Average body temperature
Chest_TEMP_slope	RespiBan	Rate of change of the body temperature from window starts to finishes
chest_ECG_HRV_mean	RespiBan	Average Heart Rate Variability
wrist_BVP_HRV_mean	E4	Average Heart Rate Variability

We hope to find the significant predictors in discriminating between stress and amusement conditions from the logistic model. We constructed the first model including all the features in Table 2. The formula of the model is:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \sum_{i=1}^{13} \beta_i x_{feature_i}$$

π_i = probability of stress for given observation

$1 - \pi_i$ = probability of amusement for given observation
Then, we perform a stepwise backward selection on both of the models. We decided not to include interactions for subject with sensor data or interactions between sensor data for the full model with chest and wrist data because the dimensionality increased dramatically upon adding either, and performing forward or backward selection including either set of interactions produces a model with all coefficients having p-values of 1.

Our logistic regression model assumes independence of residuals, linearity between log odds and covariates, and no significant impact due to influential points or multicollinearity. The plots to verify these assumptions can be found in Appendix.

II. Results After performing a stepwise backward selection on the models, 9 features appear to be statistically significant in discriminating between stress and amusement conditions in the model (these features are shown below in Table 3). It appears that the model has a 94.2% prediction accuracy (95% CI = (90.76, 96.65)). The confusion matrix shows that the model has the sensitivity rate of 89.01%, while a specificity rate 96.76%.
The ROC curve is plotted to display the trade-off between sensitivity and specificity of the models. We picked the threshold of 0.5 to compare the performance of our predictions to random guesses. The model has an area under the curve (AUC) of 0.9902, which shows that our predictions are better than random guesses.

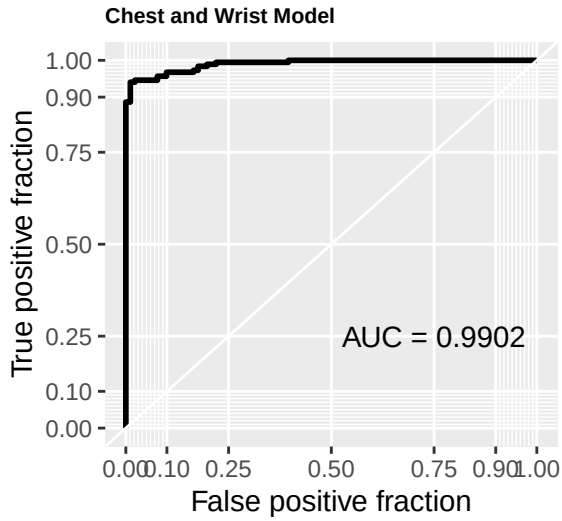

Figure 3

Table 3: Coefficient estimates of the significant predictors in the chest and wrist model.

Terms	Estimate	CI
(Intercept)	1.659	(-152.162, -21.587)
chest_RESP_volume	0.780	(0.381,1.218)
chest_SCR_num_segments	-6.427	(-7.952,-5.136)
wrist_SCR_num_segments	-2.287	(-0.230,-0.141)
chest_ACC_magnitude_mean	1.030	(27.204, 168.772)
wrist_ACC_magnitude_mean	1.391	(0.741,2.115)
chest_TEMP_mean	-0.795	(-1.519,-0.241)
wrist_TEMP_mean	0.703	(0.265,1.178)
chest_ECG_HRV_mean	-5.637	(-6.994, -4.478)
wrist_BVP_HRV_mean	0.838	(0.002,1.700)

Interpretation: Holding other features constant, as $x_{feature_i}$increases by one standard deviation, the odds of stress vs amusement increase by a factor of e^{β_i}). If e^{β_i} is smaller than one, it indicates that the subject is less likely to be stressed. On the other hand, if e^{β_i} is greater than one, it indicates that the subject is more likely to be stressed.
For example: * Holding other features constant, as the average body temperature measured by the RespiBan (`chest_TEMP_mean`) increases by one standard deviation, the odds of stress vs amusement increase by a factor of 0.452 (or $e^{-0.795}$). Since $0.452 < 1$, it means that subject is less likely to be stressed.

- Holding other features constant, as the average body temperature measured by the Empatica 4 (`wrist_TEMP_mean`) increases by one standard deviation, the odds of stress vs amusement increase by a factor of 2.02 (or $e^{0.703}$). Since $2.02 > 1$,it means that subject is more likely to be stressed.

3.2 Wrist model

I. Methodology We wish to determine whether stress (or amusement) can be detected using only the wrist-worn wearable. For this model, we used the same training and test dataset as the previous model. We performed a logistic regression model with the 5 features measured by Empatica 4 (Table 2) on our training data. After performing the stepwise backward selection on the model, it appears that all features are statistically significant except the mean temperature. We also constructed another model including those 5 features with all possible interaction terms. Then, we performed stepwise backward selection on the model with interactions.
The plots to verify logistic regression model's assumptions are included in Appendix.

II. Results After performing a stepwise backward selection on both models, it appears that the model with interactions results with a 78.62% prediction accuracy (95% CI = (73.33,83.31)), while the model with only the main effects has a 77.17% prediction accuracy (95% CI = (71.76, 81.99)). The confusion matrix shows that the model with interactions has a significantly higher sensitivity rate (56.04% VS 41.76%), while a lower specificity rate (89.73% VS 94.59%).
The ROC curve is plotted to display the trade-off between sensitivity and specificity of the models. We picked the threshold of 0.5 to compare the performance of our predictions to random guesses. It appears that the model with interactions has a greater area under the curve (AUC) of 0.7932, while the model without interactions has a AUC of 0.7379.

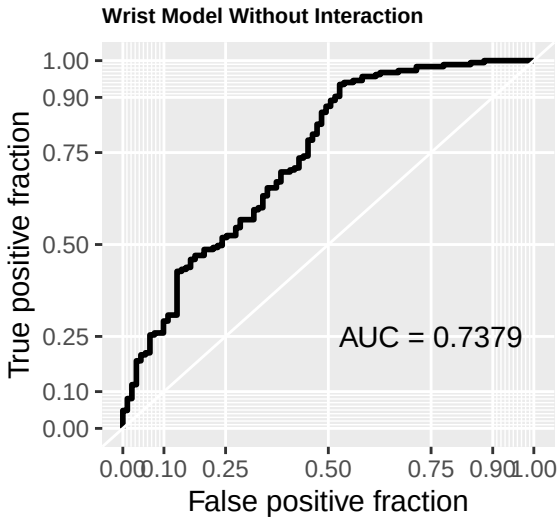

Figure 4

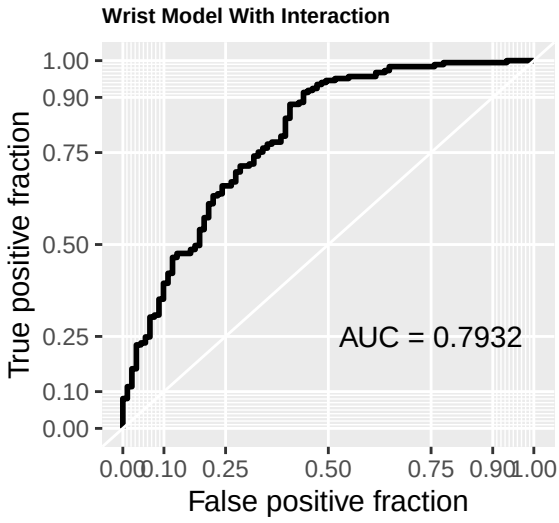

Figure 5

Table 4: Coefficient estimates of the significant predictors in our wrist model with interaction terms.

Terms	Estimate	CI
(Intercept)	0.821	(1.785,2.405)
wrist_ACC_magnitude_mean	0.641	(1.325,2.159)
wrist_BVP_HRV_mean	-0.606	(-0.879,-0.356)
wrist_EDA_slope	0.438	(15.259,39.656)
wrist_SCR_num_segments	-1.409	(-0.132,-0.094)
wrist_TEMP_mean	0.322	(0.067,0.546)
wrist_ACC_magnitude_mean:wrist_BVP_HRV_mean	0.434	(0.151,0.754)
wrist_ACC_magnitude_mean:wrist_EDA_slope	-0.347	(-35.945,-6.410)
wrist_ACC_magnitude_mean:wrist_SCR_num_segments	-1.159	(-0.122,-0.065)
wrist_ACC_magnitude_mean:wrist_TEMP_mean	0.555	(0.234,0.859)
wrist_BVP_HRV_mean:wrist_EDA_slope	0.293	(1.198,37.508)
wrist_BVP_HRV_mean:wrist_TEMP_mean	-0.567	(-0.772,-0.380)
wrist_EDA_slope:wrist_TEMP_mean	0.163	(-0.485,20.795)

Interpretation:

- When all features are at their mean values, as average body temperature (`wrist_TEMP_mean`) increases by one standard deviation, the odds of stress vs amusement increase by a factor of 1.366 (or $e^{0.312}$).
- When the average magnitude of acceleration (`wrist_ACC_magnitude_mean`) is 1 standard deviation above its mean, as average body temperature (`wrist_TEMP_mean`) increases by one standard deviation, the odds of stress vs amusement increase by a factor of 2.38 (or $e^{0.312+0.555}$).

3.3 Heterogeneity Across Subjects

To assess the heterogeneity across the 15 subjects in response to stress vs. amusement, we considered the average feature value across all windows for each subject during both stress and amusement states. It is important to look at averages of these features for each individual because by nature, individuals have different baselines for many of these features (i.e. heart rate). For example, if one subject has a low baseline heart rate of 60 and another subject has a resting heart rate of 80, and both of their heart rates increase by 30 bpm after the stress condition, their average heart rates over time will still be very different.

We compared these average values between subjects and looked at the subject with the lowest average and highest average for each feature. The features with the largest range of overall averages are mean heart rate variability measured by ECG (RespiBan) and mean heart rate variability measured by BVP (Empatica E4). Logically, this makes sense as heart rate is very dependent on individual factors (genetics, body size, etc...). Further quantifications of the range of differences between subjects for all features are shown in Table 5 below:

Table 5: Minimum and maximum Subject average for each feature. This serves as a method for quantifying heterogeneity across subjects.

Characteristic	**amusement**, N = 15	**stress**, N = 15
Mean_Chest_ACC	(-1.16, 2.08)	(-1.28, 1.46)
Mean_ECG_HRV	(-0.08, 2.53)	(-1.52, 0.49)
Mean_Chest_EDA_Slope	(-0.89, 0.61)	(-0.47, 0.26)
Mean_Breath_Rate	(-0.32, 0.94)	(-0.67, 0.92)

Characteristic	**amusement**, N = 15	**stress**, N = 15
Mean_Resp_Vol	(-1.43, 1.08)	(-0.91, 1.62)
Mean_Chest_SCR_Segments	(-0.63, 3.73)	(-0.77, -0.15)
Mean_Chest_Temp	(-2.06, 0.92)	(-3.08, 1.38)
Mean_Wrist_ACC	(-0.77, 1.16)	(-0.99, 1.85)
Mean_BVP_HRV	(-0.69, 1.81)	(-1.03, 1.68)
Mean_Wrist_EDA_Slope	(-0.60, 0.38)	(-0.17, 0.27)
Mean_Wrist_SCR_Segments	(-0.90, 3.80)	(-0.87, 0.57)
Mean_Wrist_Temp	(-2.15, 1.97)	(-2.06, 1.01)

4. Discussion

4.1 Model Validation and Sensitivity Analysis

To confirm the assumption of independence of residuals, we plotted binned residuals vs. predicted. Our binned residuals were all incredibly small and less than $1e^{-11}$, which confirms independence. We plotted binned residuals vs. each feature in our model and they are also less than $1e^{-11}$. This confirms our assumption of linearity. All binned residuals fall within SE lines for the most part. We calculated Cook's distance and there are no significant impacts due to influential points.
To explore the sensitivity of our model, we modeled our final chosen variables with a probit model to see how dependent our final model classifications were on our choice to use a logit model. The accuracy of the probit model (93.84%) did not significantly differ from the logit model (94.20%), indicating that either model would have yielded similar classification results. Additionally, we tried multiple window shifts

We also experimented with different window shifts, thinking that a smaller window shift would lead to increased dependence between observations and a larger discrepancy between sample size and effective sample size. We tried using a 30 second window shift, which actually had a slight increase in accuracy for the model using wrist and chest data. A 60 second window shift (no overlap in any observation windows) led to no increase from the 30 second shift.

4.2 Limitations and Future Investigation

This analysis is limited by the mere fact that it only has data on 15 participants and the likelihood that there was human error in wearing the devices properly is high. Subjects 2 and 17 did not have the RespiBan temperature sensor fully attached throughout the entire duration of the study protocol so the chest temperature recordings are not reliable. Subject 3 was sitting in a sunny workplace during the baseline condition and provided a valence level of 7 after the stress condition, claiming that he was looking forward to the next condition. His cheerfulness after the stress condition can likely be attributed to his surrounding environment rather than the elicited stress, hence why the survey is not fully reliable. Subject 6 and 8 experienced unrelated stress in the week prior to the study and reported that the study was rather relaxing or unstressful. Subjects 8 and 16 reported that the stressful condition was too cold. Subject 15 didn't believe the cover story of the stress condition. These reports illuminate the potential biases and confounding factors that limit the generalizability of this analysis.
It is possible that some data derived from the wrist-worn device, Empatica 4, might not be accurate. The study conducted by Milstein and Gordon [10], compared the measurement of electrodermal activity (EDA) and heart rate variability (HRV) obtained by the wrist-worn device, Empatica 4 (E4), and a well-validated MindWare mobile impedance cardiograph device during interactive dyadic state. The result showed that the mean IBIs derived from the E4 was similar to the result by the MindWare device. It also appeared that the E4 was less accurate in deriving HRV data and failed to report reliable EDA data in the study. A positive correlation between wrist movement and missing/poor data was also discovered[10]. This might be something we should take into consideration in future studies.
For future investigations, it might be advantageous to conduct a K-fold cross-validation (CV) as its estimates

have a lower test error variance than the validation set approach (we are using in this study). However, the computational time for K-fold CV is much longer than the test is much longer than the validation set approach.
Additionally, there were 9066 observations in our data in the stress affect and only 4675 in the amusement affect. Due to this class imbalance, a future investigation could be to configure amusement weight in the model to penalize the model less for errors made on stress classifications and penalize the model more for errors made on amusement classifications.

5. Conclusions

We created a model that decently categorizes stress vs. amusement based on physiological sensor data from the wrist-worn and chest-worn wearable devices. However, chest-worn devices are not feasible in practice outside of a lab setting. For this classification to be practically applied, more subjects should be enrolled and another algorithm solely based on the wrist-worn device should be developed.

6. References

1. Amiez, C., Procyk, E. (2019). Midcingulate somatomotor and autonomic functions. Handbook of Clinical Neurology. 166(1), 53-71.

2. Boano, C. A., Lasagni, M., Romer, K., Lange, T. (2011). Accurate Temperature Measurements for Medical Research using Body Sensor Networks. IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops. 2011 March 28-31. Los Alamitos, CA:IEEE Computer Soc.

3. Choi, A., Shin, H. (2017). Photoplethysmography sampling frequency: pilot assessment of how low can we go to analyze pulse rate variability with reliability. Physiological Measurement. 38(3).

4. Delaram, J., Salvi, D., Tarassenko, L., Clifton, D. A. (2018). Validation of Instantaneous Respiratory Rate Using Reflectance PPG from Different Body Positions. Sensors(Basel). 18(11). doi: 10.3390/s18113705.

5. Desjardins, J. (2015, May 20), "The History of Wearable Technology," Visual Capitalist. https://www.visualcapitalist.com/the-history-of-wearable-technology/

6. Healey, J. A., Picard, R.W (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. IEEE Transactions on Intelligent Transportation Systems. 6(20), 156-166.

7. Jun, E., McDuff, D., Czerwinski, M. (2019). Circadian Rhythms and Physiological Synchrony: Evidence of the Impact of Diversity on Small Group Creativity. Proceedings of the ACM on Human-Computer Interaction. 3(60).

8. Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. Biological Psychology. 84(3), 394-421.

9. Makowski, D. (2017). Neurokit2. [Online]. Available: https://github.com/neuropsychology/NeuroKit

10. Milstein, N., & Gordon, I. (2020, July 28). Validating Measures of Electrodermal Activity and Heart Rate Variability Derived From the Empatica E4 Utilized in Research Settings That Involve Interactive Dyadic States. Retrieved September 30, 2020, from https://www.frontiersin.org/articles/10.3389/fnbeh.2020.00148/full

11. Schmidt, P., Reiss, A., Duerichen, R., Van Laerhoven, K. (2018). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. ICMI '18: Proceedings of the 20th ACM International Conference on Multimodal Interaction. Boulder, CO. 2018 October 16-18. (pp. 400-408). https://doi.org/10.1145/3242969.3242985

12. Seaward, B. L. (2006). Physiology of Stress. Managing Stress: Principles and Strategies for Health and Well-Being (5th ed., pp. 34-48). Jones and Bartlett Publishers.

13. Tuck, K. (2007). Power Cycling Algorithm using the MMA73x0L 3-Axis Linear Accelerometer. Freescale Semiconductor.

Appendix