# Detecting Stress Using Wearables

Matty Pahren, Scott Heng On, Ethan Shen

## 1 Introduction

Affect recognition seeks to detect a person's affective state. One of these states is stress, and long-term stress is known to have severe implications on human health and well-being. This study uses the Wearable Stress and Affect Detection (WESAD) dataset, which was introduced in Schmidt et al. [1], and achieves classification accuracies of 86.2% on wrist data and 90.8% on all data for the binary classification problem (stress vs. amusement). In conducting this analysis, we aim to determine whether sensor data are useful in predicting stress, and if so, which predictors are most significant when discriminating between stress and amusement. Furthermore, we seek to understand which types of sensor data are most useful in predicting stress, alone or in combination, and to determine if stress can be detected only using the wrist-worn wearable, which is more convenient to wear than a chest-worn device. Finally, we quantify the heterogeneity across individuals in their responses to stress.

## 2 Data

For this study, we utilized the Wearable Stress and Affect Detection dataset that contains 63 million samples of raw physiological and motion signal data collected from 15 test subjects [[1]]. The data was collected using 2 devices: a chest-worn device (RespiBAN) and a wrist-worn device (Empatica E4). The RespiBAN chest data includes electrocardiography (ECG) data in mV, electrodermal activity (EDA) in microseconds, electromyography (EMG) data in mV, skin temperature (TEMP) in °C, displacement of the thorax for males and the abdomen for females induced by inhaling or exhaling (RESP) as a percentage, and 3-axis accelerometer (ACC) data in g - the acceleration of gravity. All RespiBAN data were recorded at 700 Hz. The E4 wrist data collected includes blood volume pulse (BVP) data (64Hz), ACC data (32Hz), EDA in microseconds (4Hz), and TEMP in °C (4Hz). Each sample also references a label corresponding to the study protocol condition, with each label referring to a different condition (1- baseline, 2- stress, 3- amusement, 4- meditation). These conditions were recorded by having participants exposed to sources of stimulation that would elicit the desired condition. The labels 0, 5, 6, 7 indicate that the data should be ignored. The dataset also included information about each subject, such as height, weight and age, as well as personal responses to PANAS and SSSQ questionnaires.

Considering the nature of the WESAD dataset and the scope of our study, we decided to filter out samples that did not represent stress or amusement, and standardized sampling rates for all RespiBAN and E4 variables to 4Hz. We did this by downsampling data collected at a more frequent rate, and we chose to use the recorded signals from these time periods instead of a mean since we planned on calculating mean windows of our data later. This downsampling allowed for more feasible computation with a dataset of reasonable size and standardized windows for multiple variables. Even though we lost some information by downsampling, we thought this was a good trade-off so we could test more models with a smaller dataset.

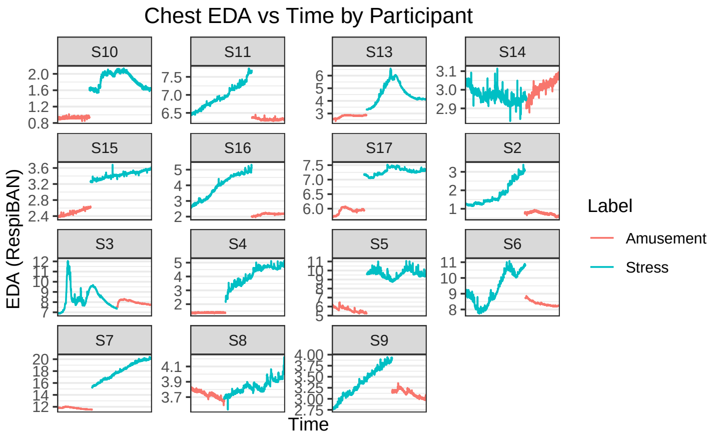


Fig 1.1 Plot of RespiBAN (Chest) EDA samples collected over time by participant and label (2− stress, 3−amusement)

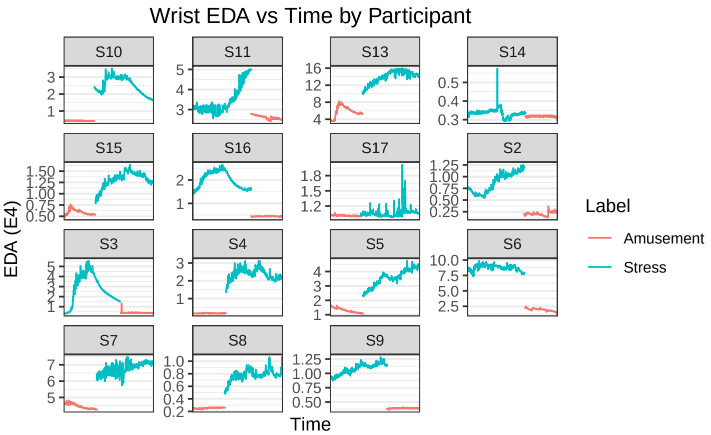


Fig 2.1 Plot of E4 (Wrist) EDA samples collected over time by participant and label (2− stress, 3−amusement)

For example, Figure 2.1 and 2.2 show plots of the EDA signal data by participant and by condition (2-stress, 3-amusement), showing different values when the participant experienced different conditions, and this difference is captured both in the chest and wrist data. Since we only took samples for 2 conditions, we could desirably observe a relative equal distribution of samples in both groups. However, raw signal data only provides limited information and difficult interpretability, and therefore from these raw data signals, we performed data and domain-driven feature engineering to generate potential predictor variables for our model that were easily interpretable and well represented the information in the data. We generated x features, equally between all the raw signal variables.

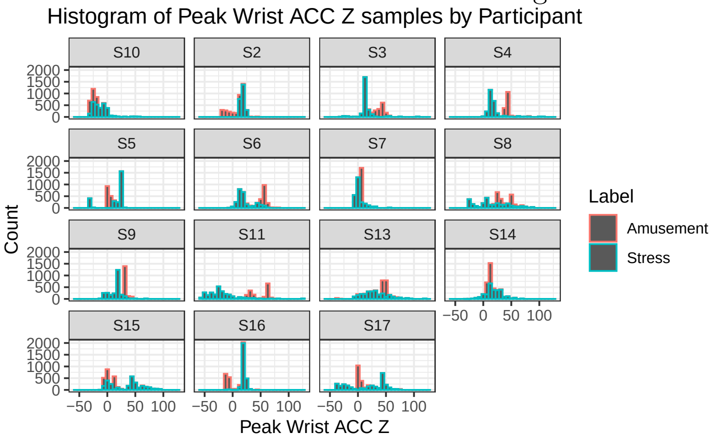


Fig 2.3 Histogram of peak acc wrist samples by participant and label (2− stress, 3−amusement)

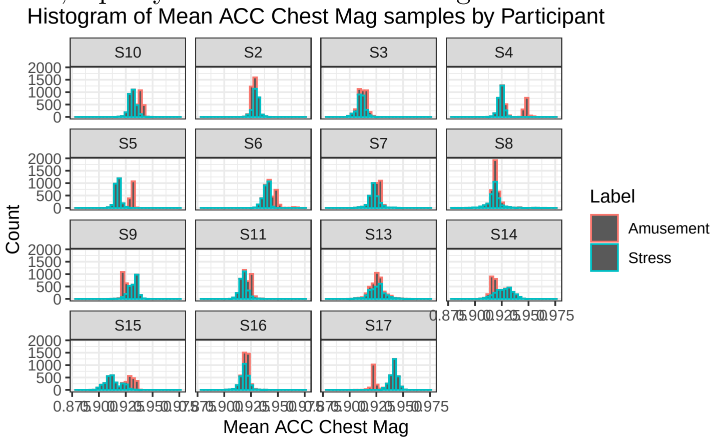


Fig 2.4 Histogram of mean acc chest magnitude samples by participant and label (2− stress, 3−amusement)

Performing exploratory data analysis on the engineered features provided more insightful information related to our study. For example, Figure 2.3 and 2.4 are histograms of Peak Wrist ACC Z and Mean ACC Chest Magnitude samples engineering from the raw signal ACC data. From these plots we can observe that these ACC feature values have different distributions based on the different conditions and could potentially be significant in predicting stress or amusement.

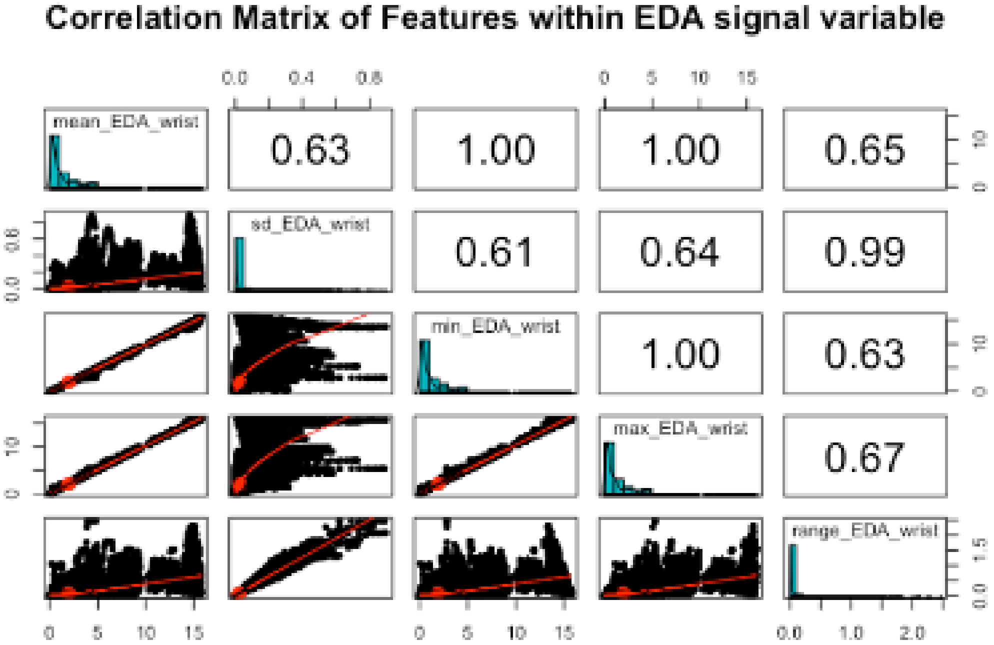


Figure 2.5 Correlation Matrix of Features within EDA signal variable. Matrix shows Features having high levels of collinearity

Other interesting insights like potential issues of multicollinearity were observed when plotting correlation matrices of the features. Figure 2.5 shows features engineered only from EDA and we can see that there exists a high level or collinearity. On the other hand, there was low correlation between features of different signal variables. High collinearity in predictor variables will influence the strategies and thresholds chosen for model selection and interaction effects, which we elaborate further in the paper in relation to our proposed model.

## 3 Literature Review

Stress can impact heart rate, blood pressure [3], and skin temperature. When a threat is presented, it only takes the brain .9 milliseconds, or 1111 Hz, to shift a body into a full stress response [4]. In that time, the brain filters information relevant to our personal safety, redirects energy to only areas that are biologically relevant for survival, and releases stress hormones. While stress can be useful in some situations in that it activates fight-or-flight response and can help us during a dangerous situation, chronic stress can have many negative long-term impacts, even reducing brain mass in some cases.

However, stress is not the only emotion that impacts us physically. In fact, many of our physiological measures change based on our emotional state. L. Shu et. al [5] examine how anger, anxiety, embarrassment, fear, amusement, happiness, and joy have different effects on our cardiovascular, electrodermal, respiratory, and electroencephalographic measures. Specifically, they saw that when patients are amused, heart rate becomes more variable and electrodermal activity, respiration rate and brain activity all increase.

With more advanced technology these days, studies have shown that changes in emotion can be effectively measured by sensors of physiological data [6]. This study found that emotional valence and arousal could all be reliably estimated by using low-cost sensors used to collect data on electroencephalography (EEG), galvanic skin response (GSR), and electromyographic signal (EMG) data.

There has also been research done on the appropriate time intervals to use when measuring different body measures such as motion, respiration, electrodermal activity, and heart rate. Different studies use various sampling rates of the raw signals and window sizes for feature engineering, and it appears that there is no consensus on the best-suited way to choose these values. Arif and Kattan [7] used a five second rolling window to analyze ACC data, and Uy et. al. [8] also used a five second rolling window for measuring Galvanic Skin Response, Blood Volume Pulse, and Respiratory Variability. Since we found multiple sources using a five-second window for predictors similar to the ones in our dataset, we decided to use this approach as well.

# 4 Methodology

## 4.1 Feature Engineering

As mentioned earlier, the brain takes .9 milliseconds to shift a body into a full stress response, which suggests that these data should be sampled at approximately 1111 Hz to properly identify a state of stress. However, the highest-resolution data is sampled at 700 Hz, which suggests that these data are not granular enough to fully capture a shift to stress. Thus, with computational cost in mind, we downsampled our data to 4Hz. We assigned a unique ID to each of the 15 subjects, and also added their personal characteristics. Since we are primarily interested in differentiating between stress and amusement, we only kept samples that were labeled 2 (stress) or 3 (amusement). We then relabeled amusement as 0 and stress as 1.

Next, we calculated features for the different modalities. Segmentation of the physiological signals was done using a sliding window, with a window shift of 0.25 seconds. The ACC and physiological features were both extracted using a 5 second window size [7][8]. To calculate some BVP and ECG features, we utilized peak detection algorithms to identify the time between each successive peaks and used the peak intervals to calculate the mean, standard deviation and root-mean squared of heart rate variability within each assigned windows.

To calculate the respiration variables, we had to separate out inhales and exhales. Inhale data corresponded to respiration values greater than zero and exhale data corresponded to respiration values less than zero. For five-second intervals where someone either did not inhale or exhale, we coded the corresponding observations to equal zero instead of NA. Additionally, we found total respiration range which took both inhale and exhale values into account when calculating the maximums and minimums. Next, we calculated the inhale to exhale ratio by dividing our mean inhale column by our mean exhale column. Since this would have produced some infinite values due to the fact that periods where someone was not exhaling were set to zero, we set these inhale/exhale ratios to equal 10. Additionally, we had a couple of intervals where the inhale/exhale ratio was very large, sometimes even greater than 1500. This was happening because some five-second windows in our data could have consisted of the participant inhaling for most of the time, thus we would be dividing by an extremely low exhale value. So, we re-coded any ratio greater than 10 to equal an upper bound of 10. This is a fair assumption, since it would not be physically possible for someone to inhale incredibly more than they exhale. We next calculated a breath rate by first summing the number of rows in one breath cycle, which is defined as one full inhale and one full exhale. Then, we took this sum and divided it by four, since the observations in our dataset were taken every quarter of a second, to get the time in seconds of one full breath cycle.

We then calculated the mean, standard deviation, maximum, minimum and range of ACC, inhale, exhale, breath rate, and the other physiological signals; a summary of all predictor variables is given in Table 1.

## 4.2 Model Selection

We first split our data into a 80/20 train/test split, perform 5-fold cross validation on the training data and lastly assess model performance by predicting on the testing data.

The extracted features serve as predictors for our classification models. We propose two different models, both using different types of sensor data: features from the wrist-worn device, and features from both chest-worn and wrist-worn devices. Since our research goals are concerned with the predictive power and interpretability of our models, logistic regression was chosen over other classification algorithms. To determine our final model, we performed variable selection by first including all relevant features and *participant* and then used backwards selection with AIC as the selection criterion. Additionally, we iteratively removed variables that had p-values greater than 0.1. Interaction effects between the remaining predictors and with *participant* were fit to determine if they would improve the predictive power of the models. We also noted that the models with the subjects' personal characteristics did not increase accuracy, F1 score or AUC values, and also led to increased computational cost. Thus, we did not include any personal characteristics in the final models.

The final models are below (Model 1 was fitted using only the wrist-worn data and Model 2 was fitted using all the sensor data):

Table 1: Summary of Sensor Data Categories, with the Extracted Predictors and Respective Symbols

| Metric | Features |
|---|---|
| **Chest Sensor Data** | |
| Chest ACC | Mean ($\mu^C_{ACC}$), SD ($\sigma^C_{ACC}$) |
| Chest ECG | Mean ($\mu^C_{ECG}$), SD ($\sigma^C_{ECG}$), Mean HRV ($\mu^C_{HRV}$), SD HRV ($\sigma^C_{HRV}$) |
| Chest EDA | Mean ($\mu^C_{EDA}$), SD ($\sigma^C_{EDA}$), Min ($Min^C_{EDA}$)<br>Max ($Max^C_{EDA}$), Range ($Range^C_{EDA}$) |
| Chest EMG | Mean ($\mu^C_{EMG}$), SD ($\sigma^C_{EMG}$), Min ($Min^C_{EMG}$)<br>Max ($Max^C_{EMG}$), Range ($Range^C_{EMG}$) |
| Chest Resp | Mean ($\mu^C_{inhale}$), SD ($\sigma^C_{inhale}$)<br>Mean ($\mu^C_{exhale}$), SD ($\sigma^C_{exhale}$)<br>Min ($Min^C_{resp}$), Max ($Max^C_{resp}$), Range ($Range^C_{resp}$)<br>Mean ($\mu^C_{breath}$), SD ($\sigma^C_{breath}$), Min ($Min^C_{breath}$)<br>Max ($Max^C_{breath}$), Range ($Range^C_{breath}$), I/E Ratio |
| **Wrist Sensor Data** | |
| Chest Temp | Mean ($\mu^C_T$), SD ($\sigma^C_T$), Min ($Min^C_T$), Max ($Max^C_T$), Range ($Range^C_T$) |
| Wrist ACC | Mean ($\mu^W_{ACC}$), SD ($\sigma^W_{ACC}$) |
| Wrist BVP | Mean ($\mu^W_{ECG}$), SD ($\sigma^W_{ECG}$), Mean HRV ($\mu^W_{HRV}$), SD HRV ($\sigma^W_{HRV}$) |
| Wrist EDA | Mean ($\mu^W_{EDA}$), SD ($\sigma^W_{EDA}$), Min ($Min^W_{EDA}$), Max ($Max^W_{EDA}$), Range ($Range^C_{EDA}$) |
| Wrist Temp | Mean ($\mu^W_T$), SD ($\sigma^W_T$), Min ($Min^W_T$), Max ($Max^W_T$), Range ($Range^W_T$) |

$$
\begin{aligned}
Y_i &\sim Bernoulli(\pi_i) \\
log(\frac{\pi_i}{1-\pi_i}) &= \beta_0 + \beta_1\ \sigma^W_{EDA} + \beta_2\ \sigma^W_{ACC} + \beta_3\ \sigma^W_{HRV}\ + \beta_4\ \sigma^W_{EDA} * \sigma^W_{HRV}\ + \\
&\beta_5\ \sigma^W_{EDA} * \sigma^W_{ACC} + \beta_6\ \sigma^W_{HRV} * \sigma^W_{ACC}\ + \\
&\sum_{j=2}^{15} \beta_{7j}\ I(subject_i = j)\ + \sum_{j=2}^{15} \beta_{8j}\ \sigma^W_{ACC} * I(subject_i = j),
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
Y_i &\sim Bernoulli(\pi_i) \\
log(\frac{\pi_i}{1-\pi_i}) &= \beta_0 + \beta_1\ \mu^C_{ACC} + \beta_2\ \sigma^C_{ACC} + \beta_3\ \sigma^W_{ACC} + \beta_4\ \sigma^C_{HR} + \beta_5\ \sigma^W_{HR}\ + \beta_6\ \sigma^W_{HRV}\ + \\
&\beta_7\ rms^C_{HRV} + \beta_8\ rms^W_{HRV} + \beta_9\ \sigma^C_{EDA}\ + \beta_{10}\ \sigma^W_{EDA} + \beta_{11}\ min^C_{EMG} + \beta_{12}\ \sigma_{inhale}\ + \\
&\beta_{13}\ \sigma_{exhale} + \beta_{14}\ max_{breath} + \beta_{15}\ ie_{ratio} + \beta_{16}\ \sigma^C_T + \beta_{17}\ min^W_T + \beta_{18}\ \sigma^C_{EDA} * min^W_T\ + \\
&\sum_{j=2}^{15} \beta_{19j}\ I(subject_i = j) + \sum_{j=2}^{15} \beta_{20j}\ rms^W_{HRV} * I(subject_i = j)\ + \sum_{j=2}^{15} \beta_{21j}\ \sigma^C_{HR} * I(subject_i = j),
\end{aligned} \tag{2}
$$

$i$ corresponds to each sample, and $Y_i$ indicates the corresponding state (1 = stress or 0 = amusement). $subject_i$ represents the unique subject ID that was assigned to each of the 15 participants. The reference level for $subject$ is Subject 10 (S10), who had the highest number of samples in the filtered dataset. The remaining levels $j \in (2, 3, \ldots, 14, 15)$ correspond to S2, S3, S4, S5, S6, S7, S8, S9, S11, S13, S14, S15, S16 and S17 respectively. The other predictors correspond to those displayed in Table 1.

### 4.3 Evaluation Metrics

We used accuracy, F1score, and AUC as evaluation metrics. Accuracy is the number of correctly classified observations over the total number of samples. We then define the F1 score and AUC. The True Positive Rate (TPR/Recall) is the proportion of positive (stress) samples that are correctly identified. The False Positive Rate (FPR) is the proportion of negative (amusement) samples that are predicted to be positive over the total number of negative samples. Precision is the proportion of samples that are positive and classified as one over the total number of samples that are predicted to be positive. Thus, the F1 score is the harmonic average of the precision and recall, and is a good choice for imbalanced classification tasks. The Receiver Operating Characteristics (ROC) curve plots the TPR against the FPR. The AUC is thus the area under the ROC curve. Both the F1 score and AUC have a range of [0,1], where a higher value indicates better classification, and a value of 1 indicates perfect classification.

## 5 Results

Table 2: Model Results with 5-fold Cross Validation, with Mean and Standard Deviation reported

| | Cross Validation Data | | | Test Data | | |
|---|---|---|---|---|---|---|
| Model | Accuracy | F1-Score | AUC | Accuracy | F1-Score | AUC |
| Wrist Only | 85.83% ± 0.45% | 79.81 ± 0.77 | 84.17 ± 0.64 | 86.2% | 80.45 | 84.70 |
| Chest & Wrist | 90.88% ± 0.05% | 87.23 ± 0.08 | 90.05 ± 0.13 | 90.83% | 87.27 | 90.19 |

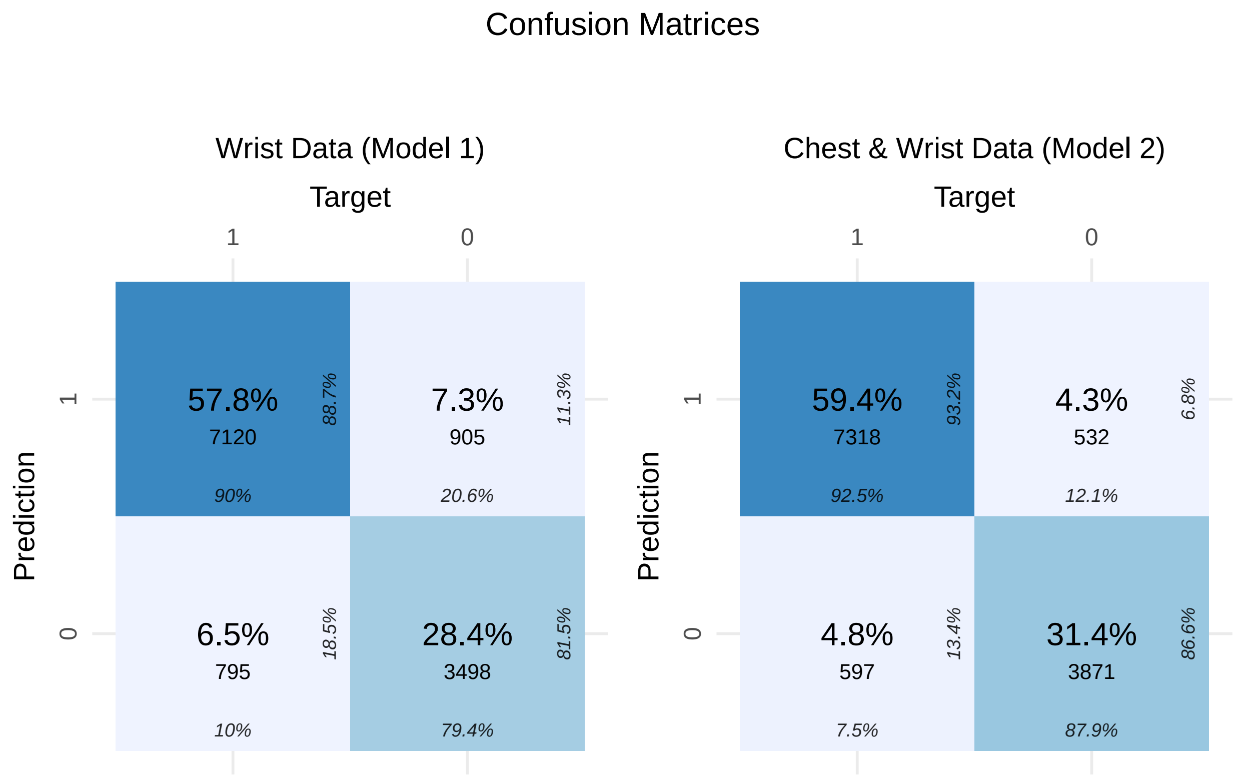
### 5.1 Wrist-worn Data

Using only data from the wrist-worn wearable, our model reports an accuracy of 86.2%, an $F_1$-score of 80.45 and an AUC of 84.7 when predicting on the testing data. This suggests that we can detect stress vs. amusement reasonably well using only the wrist-worn wearable, which may be preferable since a wrist-worn device is minimally intrusive and more convenient than a chest-worn device.

### 5.2 All Sensor Data

Using data from both the chest-worn and wrist-worn wearables, our model reports an accuracy of 90.83%, an $F_1$-score of 87.27 and an AUC of 90.19 when predicting on the testing data. These values suggest that sensor data are useful in discriminating between stress and amusement conditions. These metrics are slightly better than the previous model, which indicates that having data from both the chest-worn and wrist-worn wearables is better than having just data from the wrist-worn device.

### 5.3 Model Diagnostics



#### 5.3.1 Sensitivity Analysis

Since there is no standard sampling frequency in the literature, we first examine the sensitivity of our sampling rate of 4 Hz. Table 3 displays the results of the wrist data and all data models, with the same predictors as Models 1 and 2, when using data that was sampled at 1 Hz. In general, the models using data sampled at 4 Hz perform better on the training data, while the models using data sampled at 1 Hz perform better on the testing data. There is also higher variability in the metrics of the models using data sampled at 1 Hz - this is because we are using a much smaller sample size, which leads to higher sampling variability. However, the difference in metrics is very minimal; thus we can verify that our results are not highly sensitive to different sampling frequencies.

Table 3: Model Results with Data Sampled at 1 Hz

| | Cross Validation Data | | | Test Data | | |
|---|---|---|---|---|---|---|
| Model | Accuracy | F1-Score | AUC | Accuracy | F1-Score | AUC |
| Wrist Only | 85.8% ± 0.56% | 79.73 ± 0.88 | 84.12 ± 0.69 | 86.43% | 80.65 | 84.81 |
| Chest & Wrist | 90.78% ± 0.5% | 87.08 ± 0.66 | 89.95 ± 0.51 | 91.1% | 87.51 | 90.24 |

It is also worth discussing independence when evaluating our logistic regression model. Adding *participant* addresses independence at the subject level, but each individual sample is not independent; presumably, a person's heart rate from one second ago affects their current heart rate.

## 6 Discussion

### 6.1 Interpretation of Predictors

In our wrist-only model, the coefficients for standard deviation of wrist EDA, standard deviation of wrist ACC, and standard deviation of wrist HRV all suggest that the odds of being stressed increase as these measures increase. However, the negative interaction between standard deviation of wrist EDA and standard deviation of

ACC suggests that the odds of being stressed don't increase as much when these two measures increase together, holding HRV constant. For a full list of these estimates, please see Table 4. Also, note that these estimates are in terms of the log-odds.

In the model built with both chest and wrist data, increases in chest standard deviation of EDA, wrist standard deviation of EDA, chest standard deviation of temperature, wrist minimum temperature, standard deviation of inhale, maximum breath cycle time, mean chest ACC, standard deviation of chest ACC, standard deviation of wrist ACC, standard deviation of wrist HR, standard deviation of wrist HRV, root mean square of wrist HRV, and root mean square of chest HRV all increase the odds that a person is stressed while holding other variables constant. On the other hand, increases in minimum chest EMG, standard deviation of exhale, inhale/exhale ratio, and standard deviation of chest HR all lead to a decrease in the odds that a person is stressed, keeping other variables constant. For a full list of these estimates, please see Table 5. Also, note that these estimates are in terms of the log-odds.

In both models, standard deviation of wrist EDA has the largest estimate. In the wrist-only model, its estimate is 63.622, and in the wrist and chest model, it is 1993.249. This means that in the wrist model, we expect that an increase of 0.01 in the standard deviation of EDA will cause the odds that a person is stressed to multiply by $e^0$.63622, or 1.889 holding all else constant. In the wrist and chest model, we expect that the same increase of 0.01 in the standard deviation of wrist EDA will cause the odds that a person is stressed to multiply by $e^1$9.932, or 453,492,822 holding all else constant. In this same model, keeping all other variables constant, a 0.01 increase in the standard deviation of chest EDA will cause the odds that a person is stressed to multiply by $e^0$.283, or 1.328. This suggests that small changes in the standard deviation of EDA can have very large impacts on the odds that a person is stressed versus amused. As this variability increases, it appears that a given participant is much more likely to be stressed.

We also see a very large negative estimate for the interaction effect between the standard deviation of wrist EDA and minimum wrist temperature. The estimate of -57.13 implies that for a 0.01 increase in standard deviation of wrist EDA and an increase of 1 in minimum wrist temperature, we would expect the odds of being stressed versus amused to multiply by $e^{0.604+19.932-0.5713}$, or 468,567,160, keeping all other variables constant.

It is also interesting to note that in the model built with chest and wrist data, most of our predictors appear to be standard deviation measures, and in the wrist-only model, participant is the only non-standard deviation predictor. This observation suggests that the variability in physiological signals is a more important distinguisher of stress and amusement as opposed to the magnitude of these signals in many cases.

## 6.2 Quantifying Heterogeneity

In the wrist only model, when compared to the baseline S10 and holding all else constant, we expect the odds of being stressed for S17 to multiply by $e^{0.256}$, or 1.29, while we expect the odds of being stressed for S7 to multiply by $e^{-5.765}$, or 0.003, and the odds of being stressed for S3 to multiply by $e^{-4.34}$, or 0.013.

For the sdWACC, when compared to the baseline S10 and holding all else constant, as SD of ACC magnitude increases by one unit, we expect the odds of being stressed for S7 to multiply by $e^{10.803}$, or 49168.08 more and the odds of being stressed for S5 to multiply by $e^{6.297}$, or 542.9 more.

In the all data model, when compared to the baseline S10 and holding all else constant, we expect the odds of being stressed for S8 to multiply by $e^{0.413}$, or 1.51, and the odds of being stressed for S15 to multiply by $e^{0.178}$, or 1.19, while we expect the odds of being stressed for the other 12 subjects to multiply by a value less than 1.

For the sdCHR, when compared to the baseline S10 and holding all else constant, as HR increases by one unit, we expect the odds of being stressed for S11 to multiply by $e^{13.179}$, or 529135.6 more and the odds of being stressed for S17 to multiply by $e^{8.979}$, or 7934.7 more; conversely, as HR increases by one unit, we expect the odds of being stressed for S8 to multiply by $e^{-4.834}$, or 0.008.

Although some variation between subjects is expected, within both models, we see large variability among the odds of being stressed. These results indicate that there is heterogeneity across individuals' responses to stress vs. amusement. Overall, although we only see the differences between S10 vs. other subjects (but not S2 vs. S3 for example), the disparity in the odds of being stressed across the 15 participants is pretty apparent. This could also be due to model instability or a small number of participants - a larger pool of participants could reduce heterogeneity and variability between the estimates.

# 7 Conclusion

While we were able to distinguish between stress and amusement conditions fairly well with our model, it is worth pointing out some limitations. First, we only had data for the 15 participants in this study, and participant ID is an important predictor in our models, so we are unable to generalize the likelihood of being stressed or amused to the people in the population. Also, since all of the data was collected in one sitting during an experiment, our data is very highly correlated and suggests there is heterogeneity across individuals' responses to stress vs. amusement. We can distinguish between stress and amusement in this controlled laboratory setting, but we can't generalize about the differences in these emotions in a person's day-to-day life, or what stress and amusement might even look like if the measurements were taken on a different day. Since it appears that the wrist wearable is sufficient for detecting stress versus amusement in the laboratory setting, further research could study wrist data for individuals over a long time period, perhaps a day or even a week, and see if different emotions are still distinguishable. Also, studies could be done with a larger random sample of individuals to try and see if there are any findings about stressed versus amused physiological responses generalizable to all individuals instead of just at the participant level.

# 8 Appendix

## 8.1 Bibliography

1. P. Schmidt et al. (2018). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI 18.

2. D. Watson, L. A. Clark and A. Tellegen. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology, 54, 6, 1063.

3. H. Yaribeygi et al. (2017). The impact of stress on body function: A review. EXCLI Journal, 2017, 16, pp. 1057–1072.

4. D. Strickland. (2019). .9 Milliseconds. Transitions.

5. L. Shu et al. (2018). A Review of Emotion Recognition Using Physiological Signals. Sensors (Basel), 18, 7.

6. D. Girardi et al. (2017). Emotion Detection Using Noninvasive Low Cost Sensors

7. M. Arif and A. Kattan. (2015). Physical Activities Monitoring Using Wearable Acceleration Sensors Attached to the Body. PLoS ONE 10, 7.

8. G. Uy et al. (2013). Correlation of Stress Inducers and Physiological Signals with Continuous Stress Annotations. WCTP 2012, PICT 7, pp. 178–183.

## 8.2 Model Output

### 8.2.1 Wrist Data Model

### 8.2.2 All Data Model

```
## # A tibble: 14 x 3
##    Participant `Log Odds`     Odds
##    <chr>            <dbl>    <dbl>
##  1 S17               7.57  1948.
##  2 S11               5.94   382.
##  3 S4                2.66    14.4
##  4 S16               2.41    11.1
##  5 S14               2.28     9.82
##  6 S9                2.22     9.26
##  7 S15               2.12     8.37
```

Table 4: Cofficients obtained with Model 1 (Wrist Model)

| Coefficient | Estimate | Std. Error | p-value |
|---|---|---|---|
| (Intercept) | -0.555 | 0.057 | $<$ 2e-16 |
| $\sigma^{W}_{EDA}$ | 63.622 | 1.184 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ | 1.005 | 0.112 | $<$ 2e-16 |
| $\sigma^{W}_{HRV}$ | 0.022 | 0.015 | 0.134 |
| $\sigma^{W}_{EDA}$ * $\sigma^{W}_{ACC}$ | -4.024 | 0.111 | $<$ 2e-16 |
| $\sigma^{W}_{EDA}$ * $\sigma^{W}_{HRV}$ | 6.108 | 1.013 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * $\sigma^{W}_{HRV}$‘ | 0.05 | 0.014 | $<$ 2e-16 |
| **Participant** | | | |
| S10 | (Reference) | | |
| S2 | 0.007 | 0.078 | 0.924 |
| S3 | -0.59 | 0.077 | $<$ 2e-16 |
| S4 | -0.853 | 0.085 | $<$ 2e-16 |
| S5 | -3.072 | 0.153 | $<$ 2e-16 |
| S6 | -1.674 | 0.088 | $<$ 2e-16 |
| S7 | -5.765 | 0.254 | $<$ 2e-16 |
| S8 | -0.009 | 0.074 | 0.898 |
| S9 | 0.022 | 0.078 | 0.778 |
| S11 | -1.169 | 0.083 | $<$ 2e-16 |
| S13 | -4.34 | 0.122 | $<$ 2e-16 |
| S14 | -0.558 | 0.09 | $<$ 2e-16 |
| S15 | -0.066 | 0.073 | 0.363 |
| S16 | -0.054 | 0.082 | 0.508 |
| S17 | 0.256 | 0.075 | 0.001 |
| $\sigma^{W}_{ACC}$ *** Participant** | | | |
| $\sigma^{W}_{ACC}$ * S10 | (Reference) | | |
| $\sigma^{W}_{ACC}$ * S2 | -0.614 | 0.117 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S3 | -0.644 | 0.114 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S4 | -0.589 | 0.118 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S5 | 6.297 | 0.355 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S6 | -1.059 | 0.116 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S7 | 10.803 | 0.621 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S8 | -0.64 | 0.115 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S9 | 0.262 | 0.149 | 0.079 |
| $\sigma^{W}_{ACC}$ * S11 | -0.552 | 0.116 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S13 | -0.508 | 0.114 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S14 | 0.216 | 0.134 | 0.107 |
| $\sigma^{W}_{ACC}$ * S15 | -0.612 | 0.115 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S16 | -0.612 | 0.122 | $<$ 2e-16 |
| $\sigma^{W}_{ACC}$ * S17 | -0.668 | 0.115 | $<$ 2e-16 |

```
##  8 S2                    1.17     3.22
##  9 S7                    0.756    2.13
## 10 S6                    0.729    2.07
## 11 S13                  -0.485    0.615
## 12 S3                   -1.18     0.307
## 13 S5                   -1.98     0.137
## 14 S8                   -4.46     0.012
```

Table 5: Cofficients obtained with Model 1 - Quantifying Hetereogeneity

| term | Log Odds | Odds |
|---|---|---|
| S2 | 0.007 | 1.008 |
| S3 | -0.590 | 0.554 |
| S4 | -0.853 | 0.426 |
| S5 | -3.072 | 0.046 |
| S6 | -1.674 | 0.187 |
| S7 | -5.765 | 0.003 |
| S8 | -0.009 | 0.991 |
| S9 | 0.022 | 1.022 |
| S11 | -1.169 | 0.311 |
| S13 | -4.340 | 0.013 |
| S14 | -0.558 | 0.572 |
| S15 | -0.066 | 0.936 |
| S16 | -0.054 | 0.947 |
| S17 | 0.256 | 1.291 |
| $\sigma^W_{ACC}$ * S2 | -0.614 | 0.541 |
| $\sigma^W_{ACC}$ * S3 | -0.644 | 0.525 |
| $\sigma^W_{ACC}$ * S4 | -0.589 | 0.555 |
| $\sigma^W_{ACC}$ * S5 | 6.297 | 542.768 |
| $\sigma^W_{ACC}$ * S6 | -1.059 | 0.347 |
| $\sigma^W_{ACC}$ * S7 | 10.803 | 49184.968 |
| $\sigma^W_{ACC}$ * S8 | -0.640 | 0.527 |
| $\sigma^W_{ACC}$ * S9 | 0.262 | 1.299 |
| $\sigma^W_{ACC}$ * S11 | -0.552 | 0.576 |
| $\sigma^W_{ACC}$ * S13 | -0.508 | 0.602 |
| $\sigma^W_{ACC}$ * S14 | 0.216 | 1.241 |
| $\sigma^W_{ACC}$ * S15 | -0.612 | 0.542 |
| $\sigma^W_{ACC}$ * S16 | -0.612 | 0.542 |
| $\sigma^W_{ACC}$ * S17 | -0.668 | 0.513 |

### 8.3 Random Forest Model

In addition to the logistic regression, we also fit several other types of models. The first of those was a random forest model. We built a random forest model using all of our data and just the wrist data, and had great overall performance for both. Our model built with all the data consisted of only three predictors: participant ID, max wrist electrodermal activity, and magnitude of standard deviation of chest movement, and it had a misclassification rate of 0.74%. The model built using only wrist data used five predictors, which included participant ID, minimum electrodermal activity, maximum electrodermal activity, the range of electrodermal activity, and standard deviation of Z-plane movement, and it had an even lower misclassification rate of 0.002%.

Despite the great performance of these models, we opted to go with the logistic regression as our final model since interpretability was important to the goals of the case study. Random forest is great for prediction, however it is difficult to interpret how exactly the predictors affect the response. Therefore, we were willing to sacrifice a little bit of accuracy in order to make more concrete conclusions about how our predictor variables affected stress and amusement.

### 8.4 Multilevel Model

Table 6: Cofficients obtained with Model 2 (Chest and Wrist Model)

| Coefficient | Estimate | Std. Error | p-value |
|---|---|---|---|
| (Intercept) | -25.94 | 2.374 | $< 2e\text{-}16$ |
| $\sigma^C_{EDA}$ | 28.377 | 1.381 | $< 2e\text{-}16$ |
| $\sigma^W_{EDA}$ | 1993.249 | 34.038 | $< 2e\text{-}16$ |
| $\sigma^C_T$ | 20.924 | 2.342 | $< 2e\text{-}16$ |
| $min^W_T$ | 0.604 | 0.029 | $< 2e\text{-}16$ |
| $min^C_{EMG}$ | -8.065 | 1.173 | $< 2e\text{-}16$ |
| $\sigma_{inhale}$ | 0.354 | 0.016 | $< 2e\text{-}16$ |
| $\sigma_{exhale}$ | -0.235 | 0.019 | $< 2e\text{-}16$ |
| $max_{breath}$ | 0.315 | 0.009 | $< 2e\text{-}16$ |
| $ie_{ratio}$ | -0.21 | 0.014 | $< 2e\text{-}16$ |
| $\mu^C_{ACC}$ | 3.799 | 2.488 | 0.127 |
| $\sigma^C_{ACC}$ | 83.02 | 2.946 | $< 2e\text{-}16$ |
| $\sigma^W_{ACC}$ | 0.093 | 0.009 | $< 2e\text{-}16$ |
| $\sigma^W_{HR}$ | 1.293 | 0.151 | $< 2e\text{-}16$ |
| $\sigma^W_{HRV}$ | 0.143 | 0.016 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ | 0.064 | 0.013 | $< 2e\text{-}16$ |
| $\sigma^C_{HR}$ | -5.687 | 0.73 | $< 2e\text{-}16$ |
| $rms^C_{HRV}$ | 0.013 | 0.004 | 0.002 |
| $\sigma^W_{EDA} * min^W_T$ | -57.13 | 0.983 | $< 2e\text{-}16$ |
| **Participant** | | | |
| S10 | (Reference) | | |
| S2 | -0.407 | 0.328 | 0.215 |
| S3 | -3.423 | 0.333 | $< 2e\text{-}16$ |
| S4 | -2.814 | 0.35 | $< 2e\text{-}16$ |
| S5 | -1.548 | 0.357 | $< 2e\text{-}16$ |
| S6 | -6.138 | 0.338 | $< 2e\text{-}16$ |
| S7 | -3.15 | 0.34 | $< 2e\text{-}16$ |
| S8 | 0.413 | 0.301 | 0.17 |
| S9 | -0.072 | 0.312 | 0.817 |
| S11 | -7.488 | 0.364 | $< 2e\text{-}16$ |
| S13 | -5.21 | 0.385 | $< 2e\text{-}16$ |
| S14 | -0.547 | 0.298 | 0.067 |
| S15 | 0.178 | 0.32 | 0.578 |
| S16 | -1.825 | 0.335 | $< 2e\text{-}16$ |
| S17 | -1.126 | 0.302 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ *** Participant** | | | |
| $rms^W_{HRV}$ * S10 | (Reference) | | |
| $rms^W_{HRV}$ * S2 | -0.064 | 0.017 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S3 | 0.123 | 0.022 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S4 | -0.096 | 0.058 | 0.095 |
| $rms^W_{HRV}$ * S5 | -0.127 | 0.036 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S6 | -0.068 | 0.021 | 0.001 |
| $rms^W_{HRV}$ * S7 | -0.407 | 0.025 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S8 | -0.043 | 0.02 | 0.032 |
| $rms^W_{HRV}$ * S9 | -0.081 | 0.018 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S11 | 0.253 | 0.021 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S13 | -0.308 | 0.041 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S14 | -0.07 | 0.016 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S15 | -0.284 | 0.018 | $< 2e\text{-}16$ |
| $rms^W_{HRV}$ * S16 | 0.02 | 0.021 | 0.331 |
| $rms^W_{HRV}$ * S17 | -0.278 | 0.017 | $< 2e\text{-}16$ |

Table 7: Cofficients obtained with Model 2 - Quantifying Hetereogeneity

| term | Log Odds | Odds |
|---|---:|---:|
| S2 | -0.407 | 0.666 |
| S3 | -3.423 | 0.033 |
| S4 | -2.814 | 0.060 |
| S5 | -1.548 | 0.213 |
| S6 | -6.138 | 0.002 |
| S7 | -3.150 | 0.043 |
| S8 | 0.413 | 1.512 |
| S9 | -0.072 | 0.930 |
| S11 | -7.488 | 0.001 |
| S13 | -5.210 | 0.005 |
| S14 | -0.547 | 0.579 |
| S15 | 0.178 | 1.195 |
| S16 | -1.825 | 0.161 |
| S17 | -1.126 | 0.324 |
| $rms_{HRV}^{W}$ * S2 | -0.064 | 0.938 |
| $rms_{HRV}^{W}$ * S3 | 0.123 | 1.131 |
| $rms_{HRV}^{W}$ * S4 | -0.096 | 0.908 |
| $rms_{HRV}^{W}$ * S5 | -0.127 | 0.881 |
| $rms_{HRV}^{W}$ * S6 | -0.068 | 0.934 |
| $rms_{HRV}^{W}$ * S7 | -0.407 | 0.666 |
| $rms_{HRV}^{W}$ * S8 | -0.043 | 0.957 |
| $rms_{HRV}^{W}$ * S9 | -0.081 | 0.922 |
| $rms_{HRV}^{W}$ * S11 | 0.253 | 1.288 |
| $rms_{HRV}^{W}$ * S13 | -0.308 | 0.735 |
| $rms_{HRV}^{W}$ * S14 | -0.070 | 0.932 |
| $rms_{HRV}^{W}$ * S15 | -0.284 | 0.753 |
| $rms_{HRV}^{W}$ * S16 | 0.020 | 1.021 |
| $rms_{HRV}^{W}$ * S17 | -0.278 | 0.757 |
| $\sigma_{HR}^{C}$ * S2 | 1.639 | 5.149 |
| $\sigma_{HR}^{C}$ * S3 | 2.121 | 8.338 |
| $\sigma_{HR}^{C}$ * S4 | 5.575 | 263.654 |
| $\sigma_{HR}^{C}$ * S5 | -0.310 | 0.734 |
| $\sigma_{HR}^{C}$ * S6 | 6.935 | 1027.876 |
| $\sigma_{HR}^{C}$ * S7 | 4.313 | 74.666 |
| $\sigma_{HR}^{C}$ * S8 | -4.834 | 0.008 |
| $\sigma_{HR}^{C}$ * S9 | 2.379 | 10.793 |
| $\sigma_{HR}^{C}$ * S11 | 13.179 | 529301.616 |
| $\sigma_{HR}^{C}$ * S13 | 5.033 | 153.338 |
| $\sigma_{HR}^{C}$ * S14 | 2.901 | 18.191 |
| $\sigma_{HR}^{C}$ * S15 | 2.231 | 9.308 |
| $\sigma_{HR}^{C}$ * S16 | 4.211 | 67.409 |
| $\sigma_{HR}^{C}$ * S17 | 8.979 | 7931.690 |