Case Study 3: Stress Detection Using Wearables

Eric Tay, Lavonne Hoang, Bill Zhang, Peining Yang

1 Introduction

Affect recognition aims to detect one's affective state through observing their physiological responses to external stimuli. One such state is stress, and long-term stress is known to have severe implications on well-being. In this study, we shall use the Wearable Stress and Detection (WESAD) dataset introduced in Schmidt et al. [1] to predict stress, achieving classification accuracy values of 91.6% on wrist data and 92.9% on all data, for the binary classification of stress and no-stress. This work will follow a similar procedure to [1], but will implement a few changes, as described in Section 3. Most notably, we shall attempt to achieve similar classification accuracy using fewer predictor variables and training samples, reducing storage and computational cost, and therefore potentially allowing our procedure to be implemented on larger datasets with more subjects, which could possibly help our trained model generalize better to a wider population. In carrying out this study, we also aim to achieve the following goals: to determine whether sensor data are useful in predicting stress, to determine which types of sensor data are most useful in predicting stress, alone or in combination, and to determine whether we can detect stress only using the wrist-worn wearables, which is not only more convenient to wear compared to chest-worn wearables, but is also commercially widespread. While the focus of this study will be on stress prediction, we will also explore the predictive ability of our chosen model on the three-class (baseline, stress, amusement) and four-class (baseline, stress, amusement, meditation) classification problem.

2 Data

The WESAD dataset provides sensory information on 15 test subjects collected through wrist and chest-worn wearables. For this study, we will use the labels and raw data in the provided pickle files, which has synchronized the sensory data collected through chest-worn wearables (chest data) and the sensory data collected through wrist-worn wearables (wrist data) together. The chest data includes electrocardiography (ECG) data in mV, electrodermal activity (EDA) data in microseconds, electromyography (EMG) data in mV, 3-channel accelerometer (ACC) data in g (the acceleration of gravity), temperature (Temp) in °C and the displacement of the thorax (male) or abdomen (female) induced by inhaling or exhaling (Resp) in percentage. This data was all sampled at 700 Hz. The wrist data includes blood volume pulse (BVP) data measured from a photoplethysmography sensor at 64 Hz. ACC was sampled at 32 Hz and given in $1/64g$. EDA (microseconds) and Temp (°C) were sampled at 4 Hz. Each sample also contains a label from 0 to 7 which reflects the study protocol condition, with 1 being the baseline, 2 being stress, 3 being amusement and 4 being meditation. Response variables labeled 0, 5, 6, and 7 indicates that data that was not usable. Each subject's personal characteristics (age, height, weight and gender) are also recorded. Age ranges from 24 to 35 years, height ranges from 165cm to 189cm, weight ranges from 55kg to 90kg, and there are 12 males and 3 females in this study.

The strength of the dataset lies in the careful study protocol and validation procedures, which gives us confidence about raw data measurements and labels. Data collection is also extensive for the participants involved. However, our dataset has limitation in terms of its generalizability. The dataset only consists of 15 experimental subjects, and hence a trained model might not be generalizable to a wider population. For example, the ages of participants were between 24 and 35 years, which is not representative. Similarly, the model might not be generalizable to different activities, or a different time of day as the entire experiment only lasted 2 hours. If one were to hope for greater generalizability, we might hence need more training data. Therefore, a model that predicts well using wrist data may be preferable due to the convenience and popularity of wrist-worn wearables. We would also favor models that are computationally inexpensive, to allow for larger datasets in the future.

3 Literature Review

A study involving stress by Seaward (1994) indicates that increased stress could lead to increased heart rate, blood pressure, neurological sweating, and increased metabolic activity [2]. This leads us to believe that the data obtained from wrist and chest-worn wearables would be useful in predicting stress.

Indeed, using this data set, [1] attained good accuracy for the binary classification problem (stress and non-stress), achieving a maximum accuracy of 87.1% for wrist data, 92.8% for chest data, and 92.5% for both types of data. The approach taken in this study was heavily based on [1], with the following changes. Firstly, [1] cited large inter-subject differences in accuracy, emphasizing the need for personalized methods. We hence accounted for person-to-person variability by including person-specific characteristics (age, gender, height, weight), which showed considerable variability. Secondly, [1] conducted extensive feature engineering. Some features were difficult to obtain and the biological motivation was not entirely clear, and hence computational cost was increased. Choosing many variables could also lead to multicollinearity and lower accuracy for algorithms like the Linear Discriminant Analysis (LDA) selected in the paper. Therefore, we attempted to use fewer predictor variables to achieve similar accuracy, implementing simpler feature engineering that was motivated by the biological effects mentioned in [2] (for example, BVP and ECG data were used primarily to calculate mean Heart Rate, which was expected to increases under stress). Thirdly, [1] calculated the slope for Temp and EDA data on a 60 second interval. Instead, we calculated the slope at 1-second interval, which seemed more physiologically reasonable. Fourthly, [1] incorporated data for three states, *baseline*, *stress*, and *amusement*, and mentioned the possibility of using the data involved with the *meditation* state. We use this fourth state for training purposes, and to ensure that our model has predictive power over all possible states. Fifthly, [1] investigated 5 algorithms. The two best performing algorithms were the Random Forest (RF) and the LDA. We hence used these two algorithms in this paper and additionally investigated the Logistic Regression (LR) algorithm which was not explored in [1]. The LR algorithm was chosen due to its low computational cost and ease of interpretation. Lastly, [1] trained the model on the full training set. Instead, we experimented with multiple approaches to subsample the training set, attempting to achieve similar accuracy at lower computational cost.

4 Methods

Beginning with the synchronized WESAD data, we first do feature engineering, and subsequently remove the data with unusable labels. We then do some exploratory data analysis on this data. Following this, we conduct leave-one-out cross-validation (LOOCV) to do preliminary model and variable selection. We then explore different sampling methods to arrive at our final model, variables, and procedure. We use this final method to make predictions for the 3 and 4-class classification problem, and also investigate the importance and effect of each variable on stress.

4.1 Feature Engineering

First, we used a peak detection package (`splus2R`) to obtain the locations of heart beats from the BVP and ECG data. Then, as per [1], we downsampled our data to 4Hz. We then gave each test subject an id, ranging from 1 to 15, and added their personal particulars to the dataset. Finally, for both wrist and chest data, we calculated and stored the magnitude of the 3D acceleration, $ACC_{ij} = \sqrt{ACC_{xij}^2 + ACC_{yij}^2 + ACC_{zij}^2}$, where ACC_{xij} is the ACC value for the channel in the x direction, for subject with id i (subject i) at time j, and other variables are defined analogously. Following this, we removed all axial ACC data. This formed the raw data that we used for further feature engineering, consisting of 347,472 samples. Subject 10 and subject 5 had the least (20932) and most (28284) samples respectively.

As per [1], we calculated values for various predictor variables based on the raw data. A summary of all predictor variables is given in Table 1, along with the predictor categories they belong to. Calculations were done using a sliding window centered at the time of the sample, with a window shift of 0.25 seconds (s). The ACC-variables were calculated with a window size of 5s, as similar window sizes were used for acceleration-based context recognition [3]. The physiological (non-ACC) variables were calculated with a window size of 60s. The window size was chosen following [4]. The exception to this 60s window was slope variables, which used a window size of 1s. The following variables were calculated: Mean, Standard Deviation (SD), Minimum (Min), Maximum (Max), Range, Slope, Heart Rate, Breathing Rate, and Volume. Mean, SD, Min, Max are the associated mean, standard deviation, minimum value and maximum value of the corresponding raw data over the window. Range = Max − Min. Slope = Last value in window − First value in window. Heart Rate = Number of heart beats in window. Breathing Rate = Number of breaths in a window, where locations of breaths were determined using `splus2R`. Volume was determined (to a scale factor) using the absolute area under the curve (obtained using linear

interpolation), with the width set to the number of observations − 1, and the height given by Resp data.

Category	Variables in each Category	Raw Data
Wrist Data		
Wrist ACC	Mean (μ^W_{ACC}), SD (σ^W_{ACC})	ACC
Wrist BVP	Mean (μ_{BVP}), SD (σ_{BVP}), Heart Rate (HR_{BVP})	BVP
Wrist EDA	Mean (μ^W_{EDA}), SD (σ^W_{EDA}), Min (Min^W_{EDA}), Max(Max^W_{EDA}), Range ($Range^W_{EDA}$), Slope (∂^W_{EDA})	EDA
Wrist Temp	Mean (μ^W_T), SD (σ^W_T), Min (Min^W_T), Max (Max^W_T), Range ($Range^W_T$), Slope (∂^W_T)	Temp
Wrist Physio	All variables in Wrist BVP, Wrist EDA, Wrist Temp	
All Wrist	All variables in Wrist Acc, Wrist Physio	
Chest Data		
Chest ACC	Mean (μ^C_{ACC}), SD (σ^C_{ACC})	ACC
Chest ECG	Mean (μ_{ECG}), SD (σ_{ECG}), Heart Rate (HR_{ECG})	ECG
Chest EDA	Mean (μ^C_{EDA}), SD (σ^C_{EDA}), Min (Min^C_{EDA}), Max (Max^C_{EDA}), Range ($Range^C_{EDA}$), Slope (∂^C_{EDA})	EDA
Chest EMG	Mean (μ^C_{EMG}), SD (σ^C_{EMG}), Range ($Range^C_{EMG}$)	EMG
Chest Resp	Volume (V_{Resp}), Range ($Range_{Resp}$), and Breathing Rate (BR_{Resp})	Resp
Chest Temp	Mean (μ^C_T), SD (σ^C_T), Min (Min^C_T), Max (Max^C_T), Range ($Range^C_T$), Slope (∂^C_T)	Temp
Chest Physio	All variables in Chest ECG, Chest EDA, Chest EMG, Chest Resp, Chest Temp	
All Chest	All variables in Chest ACC, Chest Physio	
General		
Personal	Age, Height, Weight, Gender	
All Physio	All variables in Wrist Physio, and Chest Physio	
All Modalities	All variables in All Wrist and All Chest	

Table 1: Summary of Predictor Categories, with the Predictor Variables They Contain and Their Associated Symbols.

Finally, we removed unusable data with labels 0, 5, 6, 7. We then relabeled the data as follows: Labels 1 (baseline), 3 (amusement), 4 (meditation) were changed to 0 (non-stress), and 2 (stress) was changed to 1. We shall relabel the data appropriately when dealing with the 3 and 4-class classification problem.

4.2 Exploratory Data Analysis of Final Dataset

The final dataset had 179817 samples, of which subject 1 and subject 9 had the least (11556) and most (12292) samples respectively. Each sample/observation had an id, label, and consisted of all the variables listed in Table 1. In general, successive samples were taken from the same subject 0.25s apart. For our data, Heart Rates ranged from 47 to 149, and Breathing Rates ranged from 6 to 28, which were reasonable. Our predictors generally had different scales, and future work could look into scaling these variables, especially for algorithms whose predictions would change as a result (e.g. LDA). 22.17% and 77.83% of our samples were in a state of stress, and non-stress respectively. Our data was hence unbalanced.

4.3 Model and Variable Selection

Following this, we conduct LOOCV. For each fold, one subject is held to be the test subject, and the data on the other 14 is used as the train data. We fit 3 models on the train data, and obtain predictions on the test data, over each selection of predictive categories (listed in Table 1), once with the variables in the Personal category included, and once without them included. For simplicity, we shall henceforth refer to predictive categories as though they were variables. For example, if the model is fit using the Wrist ACC category, the predictor variables used are μ^W_{ACC} and σ^W_{ACC}.

The 3 models selected correspond to the three algorithms chosen, RF, LDA, and LR, which we describe in section 4.4. We additionally provide the results for a baseline model that guesses 0 (non-stress, the majority class) all the time. We record both classification accuracy (the error being the misclassification error) and the Area Under the Receiver Operating Characteristics curve (AUC). The AUC is used as a metric due to the class imbalance of the data. We also record the variable importance of each variable for the RF model, by using the Mean Decrease Accuracy criteria, which measures the decrease in predictive accuracy due to the exclusion of said variable. The procedure described above was repeated again, this time just taking a random subsample of 1000 samples from the training data to train the model, to explore a method with lower computational cost.

From our classification results, we first decide if our final model will be trained on the full training data, or just a random subsample of 1000 observations. Having made this decision, we will then experiment with other sampling methods that may better fit the data. To handle the class imbalance, we will try downsampling the training data

such that there is an equal proportion of observations in each class (stress and non-stress). Downsampling reduces class imbalance by randomly sampling observations without replacement from the majority class. We should also note that the samples are only taken from 15 individuals, over the span of 2 hours. As such, the assumption of independent observations is violated. To attempt to reduce the dependence between observations, we shall also try to thin the data. We note that physiological variables were calculated using a sliding window of 60s, and hence took every 240^{th} sample in our dataset to ensure no two observations were calculated using the same raw data.

Finally, we shall assess all the procedures and models used, and select the model, predictive category, and procedure that gives the highest accuracy and AUC values. We shall make 3 selections, one for when only wrist data is available, one when only chest data is available, and one when both forms of data are available.

4.4 Metrics and Algorithms

We first define the AUC. The True Positive Rate (TPR) is the fraction of positive (stress) samples that our model predicts correctly. The False Positive Rate (FPR) is equal to the number of negative (non-stress) samples our model predicts to be positive, divided by the total number of negative samples. The Receiver Operating Characteristics curve plots the TPR against the FPR. The AUC establishes a tradeoff between the two, ensuring that we maximize the TPR while minimizing the FPR. The higher the AUC, the better the classification, with a random guesser having AUC = 0.5, and a perfect classifier having an AUC of 1.

We will not provide explicit formulae for every variant of the model, as there are 32 different subsets of predictive variables used to train each model. However, we will give a description of the algorithm used for each model below, incorporating details specific to our implementation of these algorithms.

RF: Our RF model generates 500 decision trees, and aggregates the decisions made by each tree, classifying a sample based on which class got the majority vote. Assume that we have n training samples and p predictors. For each tree, a random bootstrap sample of n samples is used for training, and at each node, $\sqrt{p}$ variables are randomly chosen. The splitting criteria is evaluated on each variable, and the node splits based on the variable that fulfils this criteria the best. In this case, we choose the splits that decrease the Gini index the most. We chose the Gini Index as a splitting criteria for its inherent properties that make it suitable for this purpose, more so than other criteria like Mean Decrease in Accuracy [8]. This algorithm is suitable as it prevents over-fitting, the prediction results are scale-invariant, and we can obtain variable importance. An assumption of this model is that the bootstrap sample is representative of the data, which we try to ensure by generating a large number of trees. More details can be found in [5].

LDA: Let observation i have predictor variables stored in x_i, and its label stored in y_i. We assume that $x_i \sim p(x|y_i)$, for some conditional probability function p. LDA assumes that $p(x|y = 0) \sim \mathcal{N}(\mu_0, \Sigma)$, $p(x|y = 1) \sim \mathcal{N}(\mu_1, \Sigma)$, where the mean and covariance are estimated from the training samples, such that likelihood is maximized. Observation j will then be classified as 1 if $w \cdot x_j > c$, where $w = \Sigma^{-1}(\mu_1 - \mu_0)$ and $c = w \cdot \frac{1}{2}(\mu_1 + \mu_0)$. Two assumptions that may be violated for this model is that of multicollinearity between variables and independence of observations. We have tried to address the issue of independence by subsampling our data and the issue of multicollinearity by exploring both different variables, and computing fewer variables on the same raw data as compared to [1]. Despite these violations, we have still shortlisted this model as it provided the best performance in [1]. More details about the algorithm can be found in [6].

LR: Let observation i have predictor variables stored in x_i, and its label stored in y_i. The probability that i is stressed, $p(y_i = 1)$, is then given by $p(y_i = 1) = (1 + e^{-\beta \cdot x_i})^{-1}$, where β is a vector of coefficients derived using iteratively reweighted least squares. When $p(y_i = 1) > 0.5$, we classify i as being stressed, and non-stressed otherwise. For our implementation, we do not consider interaction and nonlinear effects. Hence, for example, if our predictor category is Wrist ACC, which has two predictor variables, $x_i \in \mathbb{R}^2$. While we could include interaction and nonlinear effects to allow the LR model to be more flexible, this could potentially compromise the simplicity of the model, which is why we chose to include this model. In addition, there are 44 total predictor variables, and hence, the extension of this model is left for future work. More details about the algorithm can be found in [7].

	RF				LDA				LR			
	Subsample		Full Sample		Subsample		Full Sample		Subsample		Full Sample	
	Acc.	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**
Wrist ACC	80.11%± 4.23%	0.66± 0.10	78.82%± 4.12%	0.66± 0.10	83.38%± 4.42%	0.65± 0.11	83.09%± 4.52%	0.65± 0.11	82.82%± 4.48%	0.62± 0.11	83.02%± 4.52%	0.62± 0.11
Chest ACC	81.71%± 7.60%	0.71± 0.10	79.47%± 7.27%	0.65± 0.10	81.15%± 7.25%	0.64± 0.15	81.11%± 7.36%	0.64± 0.15	80.82%± 6.29%	0.61± 0.13	80.67%± 6.10%	0.61± 0.12
Wrist BVP	81.44%± 8.58%	0.70± 0.12	80.22%± 7.44%	0.68± 0.12	83.94%± 9.78%	0.72± 0.16	83.89%± 9.75%	0.71± 0.17	82.95%± 8.70%	0.67± 0.17	83.36%± 8.81%	0.67± 0.18
Wrist EDA	77.96%± 11.51%	0.66± 0.15	74.83%± 14.59%	0.62± 0.16	79.52%± 6.30%	0.61± 0.15	80.33%± 6.38%	0.62± 0.14	79.46%± 4.32%	0.58± 0.14	79.91%± 4.88%	0.59± 0.13
Wrist Temp	73.64%± 15.08%	0.62± 0.14	71.63%± 13.99%	0.61± 0.12	77.02%± 2.49%	0.49± 0.02	77.76%± 0.79%	0.50± 0.002	77.69%± 0.84%	0.50± 0.003	77.81%± 0.79%	0.50± 0.001
Wrist Physio	87.58%± 7.19%	0.80± 0.14	84.10%± 5.49%	0.74± 0.14	85.72%± 10.78%	0.77± 0.16	85.60%± 12.59%	0.77± 0.17	84.45%± 8.79%	0.73± 0.16	84.26%± 10.09%	0.72± 0.17
Chest ECG	75.77%± 21.390%	0.67± 0.18	74.40%± 22.57%	0.67± 0.19	81.39%± 18.88%	0.66± 0.22	81.31%± 18.91%	0.66± 0.22	79.93%± 18.35%	0.63± 0.21	79.91%± 18.37%	0.63± 0.21
Chest EDA	78.11%± 9.49%	0.68± 0.15	73.42%± 9.01%	0.65± 0.09	82.43%± 7.02%	0.65± 0.18	82.68%± 7.33%	0.65± 0.18	82.08%± 7.07%	0.64± 0.17	82.23%± 6.95%	0.64± 0.17
Chest EMG	74.32%± 4.16%	0.51± 0.05	71.49%± 4.41%	0.50± 0.04	77.18%± 4.77%	0.52± 0.07	77.01%± 4.97%	0.52± 0.08	77.88%± 1.59%	0.51± 0.03	78.00%± 1.75%	0.51± 0.04
Chest Resp	80.59%± 8.99%	0.69± 0.12	79.71%± 8.63%	0.68± 0.11	79.56%± 9.41%	0.62± 0.16	79.68%± 9.54%	0.62± 0.16	79.03%± 7.59%	0.58± 0.13	78.99%± 7.59%	0.58± 0.12
Chest Temp	69.76%± 13.14%	0.51± 0.13	67.58%± 11.34%	0.52± 0.13	77.53%± 1.23%	0.50± 0.01	77.52%± 1.32%	0.50± 0.01	77.51%± 1.27%	0.50± 0.01	77.50%± 1.31%	0.50± 0.008
Chest Physio	85.74%± 16.66%	0.78± 0.18	83.66%± 19.20%	0.76± 0.21	84.22%± 19.05%	0.76± 0.20	84.85%± 19.21%	0.78± 0.21	82.98%± 17.78%	0.74± 0.19	83.16%± 17.97%	0.74.19
All Wrist	89.05%± 6.05%	0.82± 0.13	85.08%± 6.23%	0.76± 0.13	85.96%± 10.24%4	0.77± 0.14	85.71%± 11.62%	0.77± 0.15	86.00%± 9.30%	0.76± 0.14	86.10%± 9.77%	0.76± 0.16
All Chest	89.43%± 8.35%	0.81± 0.18	87.71%± 7.86%	0.78± 0.18	83.77%± 18.49%	0.76± 0.19	84.33%± 18.78%	0.78± 0.19	82.67%± 16.16%	0.74± 0.18	83.31%± 16.40%	0.75± 0.19
All Physio	90.36%± 8.91%	0.83± 0.17	90.09%± 7.31%	0.83± 0.17	82.18%± 19.49%	0.77± 0.19	82.61%± 20.70%	0.78± 0.20	85.46%± 11.87%	0.80± 0.16	85.01%± 13.89%	0.80± 0.17
All Modalities	91.49%± 7.28%	0.85± 0.16	91.05%± 6.72%	0.83± 0.16	85.76%± 14.59%	0.80± 0.18	84.75%± 15.25%	0.81± 0.18	86.38%± 9.21%	0.80± 0.15	86.23%± 10.41%	0.81± 0.15

	Acc.	**AUC**
Baseline (Always predicts non-stress)	$77.84\% \pm 0.81\%$	0.500 ± 0.00

Table 2: Model results with Leave-One-Out Cross Validation. Mean values are shown along with their standard deviations (mean value ± standard deviation).

5 Results

5.1 Testing Accuracy and Model Selection

We display initial results for the 3 models across different procedures (subsample vs full sample) and choices of predictor categories (Table 2). We again reiterate that each row refers to the algorithm being trained, and predictions being made, using all variables belonging to a particular predictor category, as listed in Table 1. We noted that the addition of variables in the Personal category did not increase accuracy nor AUC values, across all 3 models, and across procedures, sometimes even leading to reduced performance. For brevity, we have hence omitted these values and reported the values for when variables in the Personal category were not added. Henceforth, we shall exclude these Personal variables from further models.

We also note that when a subsample of 1000 observations was used for training, the RF model achieves significantly higher accuracy and AUC scores than when the full train set was used, and there was no difference for the LDA and LR model. This could perhaps be due to the fact that observations were only taken for 15 individuals over a span of 2 hours. Hence, the amount of variability in data could perhaps be captured with a small subsample of data. We also speculate that such an approach could perhaps reduce the dependency between observations. We hence also made the decision to use a subsample for training for all further approaches, due to decreased computational cost and better performance in predicting stress.

Finally, we noted that the RF Model achieved the highest accuracy and AUC scores, especially for All Wrist, All Chest, All Physio and All Modalities. We hence selected the RF Model for further analysis.

5.2 Final Model and Procedure

We explored two other sampling approaches, with the LOOCV results summarized in Table 3. First, we downsampled the training data, before taking a random subsample of 1000 observations. Second, we thinned the training data, taking every 240^{th} sample. The resulting sample had around 700 observations for all 15 folds, and was hence not further sampled. From our results, we determined that thinning did not improve predictive power, but downsampling generally served to improve the model's performance.

	2-Class				3-Class		4-Class	
	Downsampling		Thinning		Downsampling		Downsampling	
Category	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**	**Acc.**	**AUC**
Wrist ACC	83.76%	0.75	80.92%	0.66	60.88%	0.60	46.84%	0.58
Chest ACC	84.48%	0.81	83.06%	0.73	66.43%	0.65	50.97%	0.61
Wrist BVP	80.49%	0.75	81.33%	0.68	63.20%	0.62	47.37%	0.58
Wrist EDA	85.95%	0.74	77.96%	0.65	64.13%	0.61	48.84%	0.58
Wrist Temp	80.35%	0.70	72.98%	0.61	60.32%	0.59	46.47%	0.57
Wrist Physio	90.10%	0.83	87.57%	0.80	69.35%	0.65	53.18%	0.61
Chest ECG	74.67%	0.78	75.88%	0.66	61.64%	0.65	48.02%	0.60
Chest EDA	84.96%	0.69	80.22%	0.69	63.36%	0.60	48.43%	0.57
Chest EMG	80.82%	0.62	74.71%	0.50	58.25%	0.57	42.92%	0.54
Chest Resp	82.25%	0.81	79.35%	0.68	66.82%	0.66	48.82%	0.60
Chest Temp	77.24%	0.66	69.98%	0.49	59.21%	0.58	43.92%	0.55
Chest Physio	88.42%	0.75	85.44%	0.77	74.83%	0.71	53.39%	0.62
All Wrist	91.63%	0.86	89.33%	0.82	68.58%	0.65	53.48%	0.61
All Chest	88.81%	0.77	88.63%	0.79	74.83%	0.71	54.48%	0.62
All Physio	90.91%	0.84	89.87%	0.82	76.21%	0.72	55.57%	0.63
All Modalities	92.85%	0.87	90.17%	0.84	76.97%	0.73	56.30%	0.64
Baseline	77.84%	0.50	77.84%	0.50	53.13%	0.50	39.17%	0.50

Table 3: Accuracy and AUC values across the 2-Class, 3-Class and 4-Class Classification Problem, for the RF Model.

Our final model hence uses the RF algorithm, and downsamples the training set before taking a random subsample of 1000 observations to train the model. If only wrist data is available, the variables in All Wrist will be used to train the model. If only chest data is available, the variables in All Chest will be used (one can also choose not to downsample in this case). If both forms of data are available, the variables in All Modalities will be used. The accuracy for All Wrist, All Chest, and All Modalities are 91.6%, 88.8% and 92.9%. Comparatively, [1] achieves maximum accuracy values of 88.3%, 92.8% and 92.3% when wrist data, chest data, and both forms of data are used, respectively. We used our best performing predictor category, All Modalities, and performed LOOCV once more to obtain a confusion matrix (Table 4).

Actual / Predicted	Non-Stress	Stress
Non-Stress	138656	10976
Stress	1297	28888

Table 4: Confusion Matrix for the RF model, for our chosen approach and when we train on the predictor category All Modalities.

Using the RF algorithm, and our approach of downsampling before subsampling, we also conduct LOOCV, and obtain predictions for the 3-class and 4-class classification problems, reporting our accuracy and AUC figures in Table 3, along with a baseline that guesses the majority class all the time.

Finally, to assess the significance of each variable, we then used the entire dataset, downsampling it before subsampling, and trained an RF model on it, using variables in All Modalities. From this trained model, we calculate the variable importance for each variable (Table 5). We also randomly select 5 trees, and inspect the splits made at terminal nodes to make assertions about how each variable correlates with stress. We conduct sensitivity analysis by repeating this process for other predictor categories (Wrist Physio, All Wrist, All Chest, All Physio), and also vary the number of variables chosen at each split from 6 to both 3 and 9.

Variables	**Mean Decrease in Accuracy**
HR_{ECG}	24.620%
σ^{C}_{ACC}	23.814%
$Range^{C}_{EDA}$	17.069%
μ^{C}_{ACC}	16.680%
Min^{W}_{EDA}	15.392%
V_{Resp}	15.220%
μ^{W}_{EDA}	14.985%
HR_{BVP}	14.673%
μ^{W}_{T}	13.712%
σ^{W}_{ACC}	13.475%

Table 5: Mean Decrease in Accuracy in Random Forest Model for the 10 most important variables. Predictor category is All Modalities.

6 Discussion

6.1 Model Selection

We shall draw assumptions from the RF model, and conduct sensitivity analysis by checking if these assumptions hold over the LDA and LR model. For this section, we shall mainly

focus on the subsampled training data, for this was the chosen approach, and assertions made below generally hold when training on the full training data. It should be noted that this agreement can also be seen as a form of sensitivity analysis for these assertions.

With reference to Table 2, we note that multiple models and multiple selections of predictor categories achieve accuracy values of above 85% and AUC values above 0.8. This indicates that sensor data is useful in predicting stress, as accuracy is high, TPR is high and FPR is low. We note significant variability in predictive power across different predictor categories. For the RF model, the individual (non-aggregated) predictor categories that seemed to have the most predictive power were Wrist ACC, Chest ACC, Wrist BVP, Chest EDA and Chest Resp. We conduct sensitivity analysis by comparing this with results for the LDA Model and LR Model, which agree on the high predictive power of these predictive variables. The other two models also seem to indicate that Chest ECG is a 6^{th} individual predictive category with significant predictive power. Ultimately, this could indicate that motion, heart rate, respiration and sweating were most predictive of stress, which is reasonable.

We note that all aggregated predictor categories (Wrist Physio, Chest Physio, All Wrist, All Chest, All Physio, All Modalities) had relatively higher accuracy and AUC figures, than individual (non-aggregated) predictor categories. This indicated that combinations of sensor data were more useful in predicting stress than individual sensor data. This conclusion was mostly agreed upon for the LDA and LR models too. In particular, we note that the accuracy for All Wrist, All Chest and All Modalities were 89.05%, 89.43%, and 91.49% respectively. This indicated that while using both forms of data (wrist and chest) predicted stress the best, we can achieve similar accuracy just using wrist data alone, and therefore it is likely that we can detect stress using only the wrist-worn wearable. This claim is supported by the LDA and LR model, along with the results in [1].

We also note, as found in [1], there was considerable person-to-person variability, as evidenced from our relatively high standard deviation figures. Adding variables from the Personal category did not help make better predictions, and perhaps this could only be addressed with more data across more subjects. Another area for future work is the area of “personalized" predictions. Currently, we make the assumption that our test subject is unseen. However, in reality, one may wear a watch for a period of time for calibration before predictions are made. This may allow us to make predictions on more “personalized" variables, like the deviation of Heart Rate from one’s mean Heart Rate, which could be more useful for predictions.

6.2 Final Model

As mentioned in Section 5.2, we experimented with other variants of sampling approaches to achieve better accuracy rates. Thinning (taking every 240th sample) didn’t improve accuracy rates. This could be because by taking a random sample, we perhaps already reduced the dependency between observations. On the other hand, downsampling (randomly sampling observations from the majority class such that there is a equal distribution of samples across classes) seemed to improve our accuracy and AUC. For both approaches, we note the same patterns hold from before. Firstly, that Wrist ACC, Chest ACC, Wrist BVP, Chest EDA and Chest Resp have considerable predictive power (downsampling also seems to enhance the predictive power of Wrist EDA). Secondly, that aggregated predictor categories have more predictive power than individual ones. Lastly, that models trained using only wrist data have considerable predictive power (Our model trained on the variables in All Wrist attains an accuracy of 91.63%).

When comparing the results of this study with that of [1], we note that our RF model achieves higher accuracy for most predictor categories. This could indicate that we could use simpler feature engineering, and training on less training data, to achieve comparable or improved results, providing promise for a simpler, less computationally expensive approach. We note, however, that our proposed model achieves lower accuracy for Wrist BVP, Chest ECG, Chest Resp, Chest Physio, and All Chest. This could indicate that Wrist BVP, Chest ECG, and Chest Resp required more extensive feature engineering as done in [1], such as obtaining statistics for Heart Rate Variability. It should, however, be noted that our LDA model has lower accuracy figures than theirs. As such, the proposed methodology may only be suitable for the RF algorithm. Nonetheless, the approach outlined in this paper achieves the highest accuracy for models trained using only wrist data, and models trained using both chest and wrist data, and is hence suitable for the purpose of stress prediction.

Inspection of our confusion matrix indicates that our TPR is not ideal (72.5%), while our FPR is optimal (0.9%). This limitation persists, despite downsampling already being done. Hence, if our objective is to maximize recall, one may decide to collect more data for when subjects are in a stressed state, or tweak the current procedure

(feature engineering, model selection, sampling approaches) to better achieve this goal.

6.3 Variable Importance Results

From Table 5, we note that variables that are related to heart rate, motion, and sweating seem to be most predictive of stress, while slope-related variables were the least important. When performing sensitivity analysis, we note that across different predictor categories, slope-related variables are indeed insignificant, and that heart rate variables seemed to be the most important (HR_{ECG}, HR_{BVP}), which corroborates with the results in [1]. This provides support for our claim that further feature engineering of the ECG and BVP raw data could increase predictive power. This also indicates that our treatment of slope variables were either inappropriate or insignificant, and one may consider removing this variable in future work.

In inspecting the trees of different forests, we also noted that stress was often correlated with increased heart rate, increased breathing rate, increased sweating and increased temperature, which follows biological sense. These trends are supported by Table 5, which notes the importance of related categories ($HR_{ECG}, Min^{W}_{EDA}, V_{Resp}$, $\mu^{W}_{EDA}, HR_{BVP}, \mu^{W}_{T}$).

6.4 3-Class and 4-Class Classification Problem

As an extension, we explored the effectiveness of our identified model and approach for the 2-Class, 3-Class and 4-Class classification problem. As expected, prediction accuracy consistently decreases among models trained on the same predictive category, with 2-Class classification having the highest accuracy and 4-Class classification having the lowest. This suggests that with more classification classes, it is more difficult to predict affective state accurately.

For [1], their best performing algorithm for the 3-Class problem was AdaBoost (AB). We note that our model achieves a slightly lower accuracy than theirs, and hence to address the 3-Class problem, our model might not be the most suitable. However, we also note that in general, our RF model achieves comparable accuracy to their LDA model, and performs better than their RF model. This could therefore indicate, that our proposed changes in approach (simpler feature engineering, downsampling followed by subsampling), could also be useful for the 3-Class classification problem. Inspection of confusion matrices indicate that our current approach suffers in accuracy due to the inability of the model to identify Amusement and Meditation states. Since downsampling is already implemented, this could be due to model choice, insufficient feature engineering, or a lack of data.

We note that as before, aggregated predictor categories have higher predictive power than individual predictor categories. As such, aggregated predictor categories may not just be most useful for stress prediction, but for affective state prediction too. Models trained using only wrist data also has similar predictive power to those trained using both wrist and chest data, for the 2-Class and 4-Class problem. This claim does not hold for the 3-Class problem.

7 Conclusion

In conclusion, this work implements a similar procedure to [1], with the most notable changes being simpler feature engineering and a two-step sampling process of our training data (downsampling followed by subsampling). Our proposed methodology achieves better accuracy when training a model either only on wrist data, or both wrist and chest data, and has significantly lower computational and storage cost. As such, it is a viable alternative for the prediction of stress, especially if wrist data can be obtained for more subjects. Future extensions of this approach include further feature engineering of the BVP and ECG raw data, and the application of the sampling procedure to the AB algorithm, when applied to the 3-Class classification problem.

Through our analyses, we conclude that sensory data generated through wrist and chest-worn wearables is indeed useful in predicting stress, although it is limited in predicting affective state. Across different selections of predictor categories, we determine HR_{ECG} and HR_{BVP} to be the individual predictor variables with the greatest predictive power for predicting stress. Ultimately, a combination of both types of sensory data produced the best predictions, but wrist-worn wearables alone also had high predictive powers. This demonstrates the future potential to develop a convenient way of gauging individuals' stress levels using wrist-worn devices only. However, due to the previously mentioned limitations of the data, our model could possibly not be well generalized to a larger population, especially for subjects that are not well represented in the current dataset. A richer dataset with more subjects would hence be useful for stress detection from sensor data.

SUPPLEMENTAL MATERIALS

script.py: File that pulls data from pickle files. Outputs 3 csvs, `data.csv`, `data_BVP.csv`, and `data_ECG.csv` (Python Script). Module versions: `numpy == 1.17.2`, `pandas == 1.0.1`

script-processing.R: File that uses `data.csv`, `data_BVP.csv`, and `data_ECG.csv`, extracts Heart Rate data, and combines data into `data.rda` (R Script). Package versions: `splus2R == 1.2.2`

script-FE.R: File that uses `data.rda` and outputs data that has been feature engineered (R Script). Package versions: `final_data.rda`: `zoo == 1.8-8`, `MESS == 0.5.6`, `splus2R == 1.2-2`, `dplyr == 0.8.5`

script-model.R: File that uses `final_data.rda` and applies most models discussed in this paper (R Script). Package versions: `randomForest == 4.6-14`, `MASS == 7.3-51.5`, `cvAUC == 1.1.0`, `matrixStats == 0.56.0`, `dplyr == 0.8.5`, `caret == 6.0-86`, `ggraph == 2.0.3`, `igraph == 1.2.5`

script-cv-full.R: File that uses `final_data.rda` and applies cross validation on the RF model, when trained on the full train set (R Script). Package versions: `randomForest == 4.6-14`, `MASS == 7.3-51.5`, `cvAUC == 1.1.0`, `matrixStats == 0.56.0`, `dplyr == 0.8.5`

References

[1] P. Schmidt et al. (2018). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 400–408.

[2] B. L. Seaward (1994). *Managing Stress: Principles and Strategies for Health and Well-Being.* Jones and Barlett Publishers.

[3] A. Reiss and D. Stricker (2012). Introducing a New Benchmarked Dataset for Activity Monitoring. *16th International Symposium on Wearable Computers*, 108–109.

[4] S. Kreibig (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3), 394–421.

[5] L. Breiman (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[6] W. N. Venables and B. D. Ripley (2002). Modern Applied Statistics with S. *Springer.*

Venables, and Ripley, B. D. (2002) Fourth edition. Springer.

[7] J. Peng (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), 3-14.

[8] G. Louppe (2015). Understanding Random Forests From Theory To Practice. *arXiv preprint arXiv:1407.7502.*