Contents

I Group Report 3

1 Introduction 3

2 Data Collection 3

3 Data Synchronization and Data Labelling 3

4 Modalities and Features 3

5 Modelling and Optimization 4

5.1 Data Preprocessing 4

5.2 Model Design 4

5.3 Feature Selection 5

5.4 Training and Parameter Optimization 5

6 Evaluation and Results 6

II Individual Report 7

7 Critical Discussion and Challenges 7

8 Sensors and Modalities 7

9 More Modalities 8

10 Conclusion and Results 8

Part I
Group Report

1 Introduction

Stress, which is a state of physical, mental, or emotional strain or tension that results from demanding or adverse circumstances [1], is increasingly becoming a problem in our society. The aim of our project was to build an automatic, continuous stress detection system that is based on non-invasive methods. Beyond simply detecting stress, our system was desired to be able to distinguish between lower and higher levels of stress. Therefore, our system was trained to be able to detect three different levels: low stress, medium stress, and high stress. Stress detection systems like ours could have important applications in a variety of settings, including health and safety, educational, and more recreational purposes.

As signals, galvanic skin response (GSR) and Electrocardiogram (ECG), two physiological signals that are commonly used for stress detection and that can be easily measured with the Empatica sensors [2], were chosen. In addition, auditory data was modelled, specifically participants' voice during our experimental tasks. The rationale behind choosing these sensors was that they are non-invasive and allow continuous, real-time tracking.

For our system, the steps gone through are data collection, refining the labels of our data, feature selection and optimization, modelling and parameter optimization, and evaluating the models.

2 Data Collection

A mental arithmetic task was used to elicit three different levels of stress in the participants. The three tasks each lasted two minutes, and were always preceded by a rest phase of at least one minute. After each task, the participants gave a self-report of their stress levels during the task, on a range from one (very low stress) to ten (very high stress). All participants performed first the 'high stress' task, then the 'medium stress' task, and then the 'low stress' task.

In the high stress condition, the task was to count down from 1022 in steps of 13 (e.g. 'One thousand and twenty-two, One thousand and nine, ...'), starting from scratch whenever a mistake was made. During this condition, five experimenters were present in the room and observing the participant, and the participant was encouraged to try to maintain eye contact with the experimenters. Additionally, the participant was encouraged to try to maintain a faster pace, if possible. Mistakes were indicated by one of the experimenters silently raising their hand.

In the medium stress condition, the task was to count down from 1022 in steps of three. The participants were unobserved by the experimenters, and were encouraged to look wherever they wanted. Mistakes were still indicated by a raised arm by one of the experimenters, and the participant had to start from scratch if a mistake was made. The participant was encouraged to maintain a comfortable speed.

Finally, in the low stress condition the task was to count up from 1000 in steps of one. Experimenters not taking measurements left the laboratory, so that only about three people remained in the room. Participants were encouraged to maintain a very leisurely pace, and to close their eyes if they felt comfortable doing so.

A total of eight participants were used for the final dataset. In addition to these eight participants, two participants were run for a pilot run, which was different from the final run in the order of conditions (low, medium, and high stress, rather than high, medium, and low), and in the length of each condition (3 minutes rather than 2). The data from the pilot run was not included in the models because of the condition order difference and the fact that the quality of the audio recordings was poor. However, the pilot run helped to conduct the final experiment in a controlled way, thus avoiding errors.

3 Data Synchronization and Data Labelling

After collecting the data, the raw data streams had to be synchronized in order to allow for a meaningful classification model. The HR data consisted of one data point per second, the EDA data of four data points per second. Therefore, the mean for each subsequent set of four data points was calcuated for the EDA data stream. In terms of audio, the pitch was extracted from the signal of each audio file (see Chapter 4). The pitch consisted of the frequencies present and their corresponding magnitudes for each frame of the audio clip. In order to match the frames with the seconds of the audio clip, the hop length was set to match the sample rate. Therefore, each frame corresponds to a second within each audio file. This meant that fusing the audio features with the features of the other modalities was a trivial task.

Regarding data labelling, the stress self-report values were normalized (re-labelled) into three categories by applying k-means clustering.

4 Modalities and Features

For this project, three modalities (signals) were chosen: Electrodermal Activity (EDA), Electrocardio-

gram (ECG), and auditory signals (speech).
EDA, also known as Galvanic Skin Response (GSR) or Skin Conductance Response (SCR), is the change in the electrical resistance of the skin, which are caused by changes in the skin's sweat level [3]. ECG is the recording of the electrical activity of the heart measured on the body surface [4]. Lastly, speech can be thought of as longitudinal sound waves that are produced by the human vocal cords [5].
For each of the above mentioned modalities, several features were extracted. Concerning EDA, the features were calculated on the raw data. For speech, pitch was calculated, which describes the fundamental frequencies of the sound waves [1]. Of the frequencies present, the maximum, minimum, mean, and standard deviation were extracted. Regarding ECG data, the heart rate (HR) was extracted, which is the speed of the heart beat, measured as the number of contractions of the heart per minute [6].
Importantly, the initial plan was to include features extracted from Heart Rate Variability (HRV) in addition to those from HR, as suggested in [1]. Whilst the sensor captured the full spectrum for HR, the sensor could only obtain a limited number of HRV data, and extracting HRV based on HR is a nontrivial task. As a consequence, it was decided to drop HRV.
For both EDA, HR, and speech[1] we included the following features:

- global mean[2]
- global standard deviation
- global max
- global min
- mean of absolute value of differences between each pair of consecutive time points
- root mean square

This resulted in 36 features overall: six for EDA, six for HR, and 24 for speech. An evaluation of the feature relevance will be discussed in Chapter 6.

5 Modelling and Optimization

5.1 Data Preprocessing

The data consisted of three labeled tasks for each of the eight people that participated in the experiment. Each of these examples was constituted by the HR, GSR and audio time series. The data for all features was normalized by subtracting the time series mean and dividing by the time series standard deviation. This enforced the data from these modalities to have a mean of zero and standard deviation of one.

The training examples were created by extracting the features described in Section 4 on the whole time-series, resulting in 30 examples. Additionally, it was tried to implement a sliding window of a fixed length N and computing the same features within this window. With this approach, $30 \times (120 - N + 1)$ examples can be obtained. The value of N implies the following trade-off: a small N results in local features extracted from a short period of time, while a large N implies having features that are more global. In addition, the number of examples increases as N decreases. However, throughout the course of the project it became apparent that this approach is not trivial. Also, the main focus of this research lies on investigating how the global features correlated with different stress levels. Thus, only the first approach was employed while the idea of the sliding window should be investigated in future research.

5.2 Model Design

The model design was based on a Support Vector Machine (SVM) [7]. The software *Scikit-learn* was employed to train the SVM [8]. A different SVM was trained for each pair of stress levels via a 'one-versus-one' strategy. In particular, $k(k-1)/2$ binary classifiers (where $k = 3$ is the number of classes) were trained on the subset of samples belonging to each pair of stress levels. Training these SVMs implies minimizing the following optimization problem for each $j \in [1, k(k-1)/2]$:

$$\frac{1}{2}\mathbf{w}_j^\top \mathbf{w}_j + C\sum_{i=1}^{N_j} \xi_{ij} \tag{1}$$

subject to the constraints:

$$y_i(\mathbf{w}_j^\top \phi(\mathbf{x}_i) + b_j) \geq 1 - \xi_{ij},\ \ \xi_{ij} \geq 0,\ \ i \in [1, N_j] \tag{2}$$

where $\mathbf{w}_j$ and b_j are the weights of the j-th SVM; N_j is the number of samples belonging to the j-th class pair; C is a regularization hyperparameter trading off the misclassification of training samples against the simplicity of the decision boundary; $\mathbf{x}_i$ and y_i are the input vectors of features and the class label of the i-th pair-wise example, respectively; ξ_{ij} are the slack variables that handle non-separable data; and $\phi(\cdot)$ is the feature map function that allows to enrich the input set of features, ideally by combining them in a non-linear manner.
At prediction time, the outputs were produced through a voting scheme, where each pair-wise classifier casts a vote for one class, and the class with the most votes is selected. One drawback of Support Vector Machines is that they do not naturally provide probability estimates for the posterior probability of class memberships. This difficulty was overcome using the pair-wise

[1] For speech, these global features were computed on each of the four 'local' (i.e. per-second) frequency features described above - mean frequency, maximum and minimum frequency, and standard deviation.

[2] *global* meaning of the whole two-minute time window.

coupling method proposed by Wu *et al.* [9], which extends Platt scaling [10] to produce probability estimates for the multiclass case.

The performance of several kernel functions was explored via cross-validation, and a Radial Basis Function (RBF) kernel was found to have the best validation error:

$$K(\mathbf{x}, \mathbf{t}) = \langle \phi(\mathbf{x}), \phi(\mathbf{t}) \rangle = exp(-\gamma ||\mathbf{x} - \mathbf{t}||)^2 \quad (3)$$

where $\mathbf{x}$ and $\mathbf{t}$ are the vectors of input features for two samples, and γ is a parameter that controls how much influence a single training sample has.

In addition to SVMs, the data was also modelled using Logistic Regression. Logistic Regression is a regression analysis where the prediction variable is categorical. It finds the best-fitting coefficients for a logit transformation of the probability of the category of interest. In particular, the function estimates the logarithm of the chance ratio:

$$\log \frac{p(y=1)}{1-p(y=1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \quad (4)$$

where β are the regression coefficients.

5.3 Feature Selection

The importance of the features was evaluated by a generalized linear model (GLM) with a backward stepwise feature selection algorithm. This algorithm starts with the full model and then iteratively removes the best feature (i.e. the feature whose removal results in the biggest increase in loss) in a greedy way. It assesses the change in R-squared values to rank the variables in importance to the variable of interest. In order to ensure robustness of the result, repeated cross-validation was implemented with ten folds (nine for training and one for testing) and repeated three times. The results are reported in Figure 1.

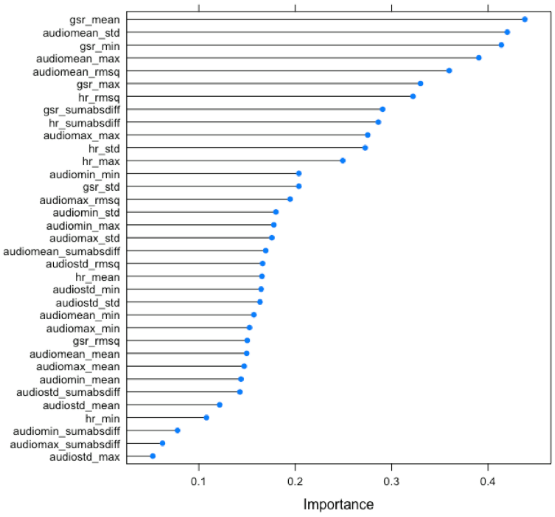


Figure 1: Results of Feature Selection by Stepwise GLM

It is observed that the best features are dominated by the data collected from GSR and audio recordings. This is no surprise as the heart rate recordings by the Empatica were of poor quality. Considering the data set consisted of ten participants only, the ten most relevant features were used in the prediction algorithms.

5.4 Training and Parameter Optimization

The entire data set was split into two parts: one for testing (25 %), which was put in cold storage for later testing; and the remaining 75% was used as a training set to tune our model hyperparameters using 3-fold cross-validation.

As laid out above, both SVM and logistic regression are dependent on a set of hyperparameters which are not automatically optimzed by the algorithms. For the SVM, the parameter C controls the trade-off between smooth decision boundaries and accurate classifications. Small values for C indicate a smooth decision boundary where the weight vector has a small norm at the expanse of potential misclassifications. On the other hand, the RBF-specific parameter γ determines the influence of each training example for the subsequent predictions. Thus, high γ means the decision boundary is mostly influenced by the training points closest to that boundary. To find out which values for the hyperparameters lead to best classification, a grid-search with 3-fold cross-validation was conducted. The results are depicted in Figure 2.

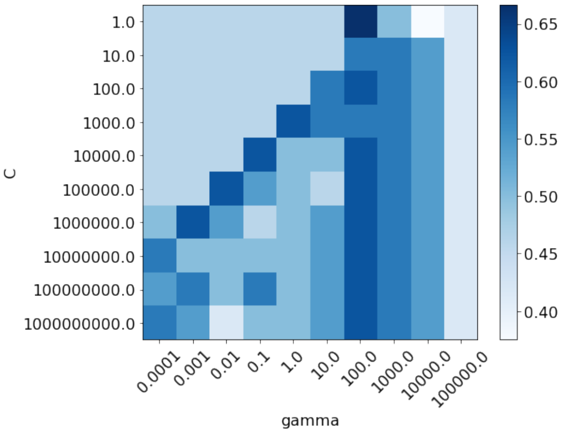


Figure 2: Results of Gridsearch for SVM Parameters

One can see that the best accuracy (for this run) was achieved with $C_{SVM} = 1$ and $\gamma = 100$.

The logistic regression model was also tested with different hyperparameters. Here, however, only C can be altered and by convention C in logistic regression describes the inverse of the regularization strength. Thus, small values indicate high regularization (similar to high C values in a SVM). Figure 3 shows how the test score depends on C. The best accuracy was achieved using $C_{Log} = 1000$. A more in-depth examination of the model performances will be provided in the next chapter.

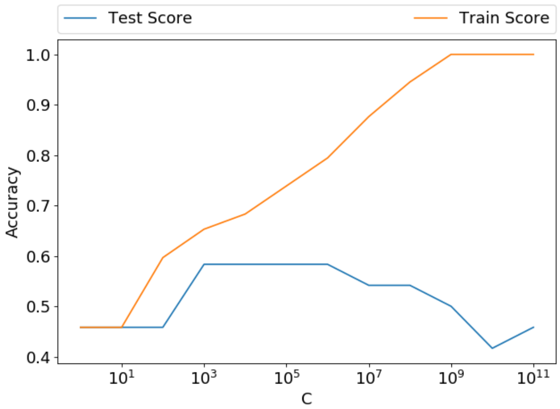


Figure 3: Results of Gridsearch for Logistic Regression Parameters

6 Evaluation and Results

Evaluating the performance of the system is a non-trivial task given how little data can be utilized. It is important to note that the performance is very sensitive to the way the data is split into training and testing. To provide a more robust measure of performance, both models have been run 1000 times. The results are reported in Table 3.

Model	Best	Worst	Average
SVM	87.5 %	25 %	57.8 %
Log. Regression	100 %	12.5 %	55.5 %
Naive Classifier	33 %	33 %	33 %

Table 1: Model Performance on Validation Set

To gain further insights into how the algorithms classify data points, a confusion matrix has been constructed (Table 2, also 1000 repetitions). One can see that the high stress level was classified with the most confidence, whereas the algorithms did not achieve good results for the medium stress level. Here, the models mainly classified these instances as high stress as well.

Stress Classification Outcome

Actual Stress Class	Low	Medium	High
Low	**1,132 (46.8%)** **1,082 (45.5%)**	33 (1.4%) 158 (6.6%)	1,252 (51.8%) 1,140 (47.9 %)
Medium	348 (21.6%) 229 (14.4%)	**9 (0.6%)** **0 (0%)**	1,251 (77.8%) 1,365 (85.6%)
High	312 (7.9%) 503 (12.5%)	216 (5.4%) 154 (3.8%)	**3,447 (86.7%)** **3,369 (83.7%)**

Table 2: Confusion Matrix. Upper Row Corresponds to SVM, Lower Row to Logistic Regression

2182 words

Part II
Individual Report

7 Critical Discussion and Challenges

In todays world recognizing emotional stress has vast applications in the field of education, mental healthcare, driver safety, rehabilitation, marketing and customer care. As computers become more proficient at learning due to advances in technology, the field of affective computing has the potential to play a significant role in these applications. Indeed, a computer gaining the ability to understand emotions and ultimately conveying emotions similar to those of humans is considered a precursor to the creation of conscious machines. This is pivotal as consciousness imparts the ability to generalize. [11]

However, with the expansion in scope of affective computing there also arise potential pitfalls. As computers learn from humans and become better at deducing stress, it may lead to insurance firms using this technology to raise health premiums for people suffering from chronic stress, anger, and depression. The line between doctor-patient confidentiality will be significantly blurred.[11] Another relevant issue is deception. Emotional cues are an important basis for human beings to communicate and build trust with one another, as they are difficult to fake. Stress is a relatively easy emotion to spot amongst humans. A computer that is proficient in understanding and conveying emotion can easily introduce deception into this communication channel since it does not need to be programmed to make its true emotion state obvious, hence affecting users confidence in data.[11]

As a field of study emotional stress has been of interest to human beings since ancient times. Documents detailing mental disorders such as depression can be found all the way back to 1500 BC Egypt.[12] Despite this, it is only recently that a link has been made between the cognitive and physiological aspects of emotional stress and the need to understand both components simultaneously.[11] Cognitive aspects of emotion emphasize understanding the situations that give rise to emotions whereas physiological aspects look at arousal in the body due to emotions.

In this assignment we have built a system that focuses on physiological aspects of stress detection. This was achieved by conducting the Trier Stress Test in lab condition for a willing group of participants. Our aim was to capture stress in 3 modalities: voice, skin, and heart rate. The sensors used in this experiment were the Empatica sensors, which detected heart rate and galvanic skin response, and a voice recorder to capture audio data. We then modelled the collected data using supervised learning techniques and fused them together in an ensemble to complete our system.

8 Sensors and Modalities

Empatica wristband is an advanced sensory device that detects with high-resolution a persons electrodermal activity (EDA), heart rate, and skin temperature. It also has an added benefit that the data collected is stored on a cloud server ready to be downloaded and analysed.[13]The crowning glory for this device was in 2018 when it received FDA approval for medical use in seizure detection. However, despite Empaticas sophistication it suffers from certain drawbacks. One common issue is the frequent drops in signal leading to missing sensory data in our cloud. This compromises the effectiveness of our learning algorithm, as it is not designed to deal with missing data. Empatica also faces issues with regards to the accuracy of its measurements. While reading reviews of the device on Facebook, a common complaint that kept popping up was the high rate of false positives and false negatives in seizure detection. The device had a habit of mistaking strenuous physical activity for seizure, and when attached to the ankle it failed to pick up active seizures. Hence, its effectiveness seems location dependent.

This location dependence has additional consequences especially when measuring skin response. Recent research in multiple arousal theory showed that when arousal is measured as sympathetic nervous system activation indicated by electrodermal activity, the GSR values could differ significantly across the two halves of the upper body[14]. Given that our experiment fell within this scenario, we considered using two Empaticas; one for each wrist, to test if we saw any variations in GSR values. Unfortunately, synchronizing the signals from the two Empaticas became an issue and we could not move forward with this idea.

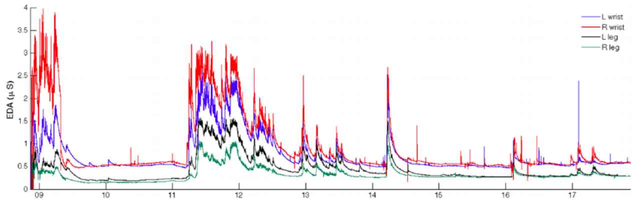


Figure 4: EDA signals from the 4 limbs taken over a 9 hour period. The two wrists and ankles show symmetric result to one another with very similar baselines. We can see right-dominant asymmetric response at 9:15 and 11:15 due to perceived future risk by the participant.[14]

Our experiment also faced issues measuring heart rate (HR). Apart from the constant dropping of HR signal by Empatica, we also noticed that this modality is heavily state dependent. It is very easy to get different HR values if the participant is doing same task standing up or sitting down, in an empty stomach or a full stomach. Hence, the high variation in this modality seems to raise question marks on its usefulness in detecting stress.

9 More Modalities

Facial recognition models for stress detection have seen very favourable results recently. With the help of deep convolutional neural networks, up to 6 emotions have now been recognised with varying degree of success.[15]

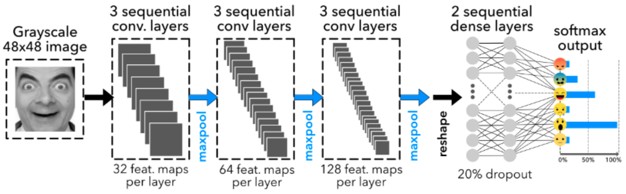


Figure 5: Shows the full deep CNN emotion recognition model built by (Ho, 2016). The model takes grey scale image inputs of 48x48 dimensions and processes it through 3 sequential convolutional layers and then through dense layer with 20 percent dropout to control for overfitting. The output is a softmax layer, which returns the probability between 0 and 1 on which of the 6 emotions is seen in the figure.

A strong advantage of using facial recognition to determine stress is that it can be done in non-invasive ways. A small camera that is out of sight of the individual in focus can be utilized effectively to determine stress levels. Indeed, commercial applications of this modality are already being implemented in the field of driver safety.[16]

Generally speaking, degree of invasiveness can be a strong criterion for determining which modality to apply given the context. In day-to-day life and for long-term stress detection, a less invasive technique would be preferable. For medical diagnostics and short-term situations there is scope for more invasive techniques.

A slightly more invasive modality, which has been used since ancient times for medical diagnostic is tongue colour. It is well know that during stress the body experiences changes in saliva composition, muscle tension and blood circulation which influences tongue appearance.[17] With the advent of computer vision and image segmentation techniques tongue colour can be used as a modality quite cheaply and effectively.[18]

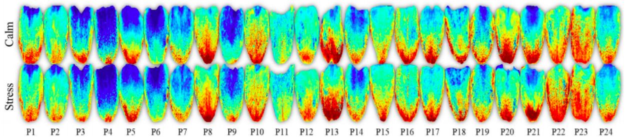


Figure 6: 3 Heat maps showing the amount of a value in the Lab colour space after the calm (top) and stress (bottom) tasks for each of the participants. [18]

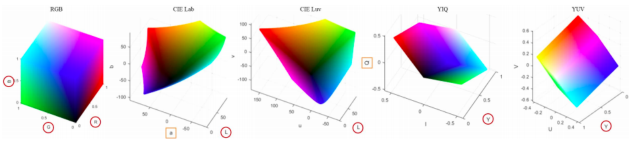


Figure 7: Tongue heat map showing change in stress. Top row indicates before the stress task. Bottom row after. Red shows high degree of correlation with reported stress scores. [18]

In the above experiment tongue images were captured using a simple cell phone camera. Using the GRABCUT segmentation algorithm a heat map of the tongue was formed before and after the stressful event. Colour features of the images were extracted and changes were compared as seen in figure 2. It was observed that the QQ component of YIQ space showed consistent differences before and after arousal of the body.

10 Conclusion and Results

Once the features were created and narrowd down an early fusion strategy was adopted to train the machine learning classifiers. It is important to note that due to the limitations of our dataset arising from small sample and distinct lack of variation in response, deep learning or other data intensive models became entirely ineffective.

25% of the data that was earlier kept in cold storage was now taken out and models with best hyperparameter values were run on this test set. The test set cosnsisted of: 3 low response, 2 medium response and 1 high response. Additionally for this individual report a random forest classifier was also attempted.

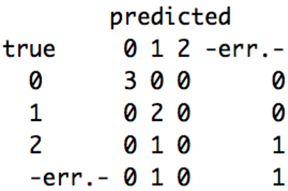
The results are benchmarked in table 3.

From the results of the benchmarking table it looks like SVM and logistic regression failed to improve on the naive classifier which guessed the most common label in the test set (low stress). SVM and Logistic regression predicted the low stress for all stress levels similar to naive classifier. However, random forest was quite successful on this dataset. The result must be accepted with caution however, as the small sample size of our dataset meant the results are very sensitive to splits. It is conceivable that an increase in the sample size would change the results entirely.

An attempt to generalise the results was made by adding the EDA and HR response from the MIT driver stress database.[19] However, due to different type of experiment, different labeling scheme and different lab environment a fusion of these two datasets was unfortunately unviable.

Model	Classification Error	Confidence Interval
SVM	50%	±0.400%
Random Forest	16%	±0.298%
Log. Regression	50 %	±0.400%
Naive Classifier	50 %	

Table 3: Benchmark Performance on Test Set



```
        predicted
true     0 1 2 -err.-
  0      3 0 0      0
  1      0 2 0      0
  2      0 1 0      1
  -err.- 0 1 0      1
```

Figure 8: Confusion Matrix of Random Forest Predictions on Test Set

References

[1] A. Alberdi, A. Aztiria, and A. Basarab, “Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review,” *Journal of biomedical informatics*, vol. 59, pp. 49–75, 2016.

[2] “Empatica e4: the wearable device for researchers that need access to real-world physiological data,” 2015.

[3] Z. A. Villarejo MV, Zapirain BG, “A stress sensor based on galvanic skin response (gsr) controlled by zigbee,” *Sensors (Basel, Switzerland)*, vol. 12, pp. 6075–6101, 2012.

[4] J. D. Bronzino, *Biomedical engineering handbook*, vol. 2. CRC press, 1999.

[5] D. B. Fry, *The physics of speech*. Cambridge University Press, 1979.

[6] H. Tanaka, K. D. Monahan, and D. R. Seals, “Age-predicted maximal heart rate revisited,” *Journal of the American College of Cardiology*, vol. 37, no. 1, pp. 153 – 156, 2001.

[7] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, Sept. 1995.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] T.-F. Wu, C.-J. Lin, and R. C. Weng, “Probability estimates for multi-class classification by pairwise coupling,” *J. Machine Learning Research*, vol. 5, pp. 975–1005, Dec. 2004.

[10] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, pp. 61–74, MIT Press, 1999.

[11] R. Picard, *Affective Computing*, vol. 1. MIT Press, 1997.

[12] B. Ebell, *The Papyrus Ebers : the greatest Egyptian medical document*, vol. 1. Copenhagen : Levin Munksgaard, 1937.

[13] “Empatica e4: the wearable device for researchers that need access to real-world physiological data.” https://empatica.app.box.com/v/E4-TechSpecs.

[14] P. et al., “Multiple arousal theory and daily-life. emotion review,” *Emotion Review*, pp. 1–14, 2015.

[15] J. Ho, “Facial emotion recognition.” https://github.com/JostineHo/mememoji.

[16] e. a. Hua Gao, “Detecting emotional stress from facial expressions for driving safety,” 2014.

[17] M. Bakke, “Salivary cortisol level, salivary flow rate, and masticatory muscle activity in response to acute mental stress: a comparison between aged and young women,” *Gerontology*, vol. 50, pp. 383–392, 2004.

[18] *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, vol. 50, 2004.

[19] “Stress recognition in automobile drivers.” https://physionet.org/pn3/drivedb/.