Validating Physiological Stress Detection Model
Using Cortisol as Stress Bio Marker

Rajdeep Kumar Nath1, Himanshu Thapliyal1 and Allison Caban-Holt2
1Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY, USA
2The Maya Angelou Center for Health Equity, Wake Forest School of Medicine, Winston-Salem, NC, USA
Corresponding Author: hthapliyal@uky.edu

Abstract—In this work, we have presented the validation of a stress detection model using cortisol as the stress biomarker. The proposed model uses two physiological signals: Galvanic Skin Response (GSR) and Photoplethysmograph (PPG) to classify stress into two levels. GSR and PPG signals were collected from a total of 13 participants along with saliva samples taken at time points throughout the duration of the experiment. We have used 10 out of the 13 participants to train our model. Data from the remaining 3 participants was used to test the robustness of the model in distinguishing stressed states from non-stressed states. We have achieved an overall accuracy of 92% with the model achieving precision, recall and f1-score of 93%, 99% and 96% respectively in predicting the occurrences of stressful events. Results indicate the promise of the proposed methodology in accurately detecting the presence of stressful events by generalizing the test data coming from a subset of population in contrast to the training data.

Index Terms—Galvanic Skin Response (GSR), Photoplethysmograph (PPG), Stress, Machine Learning, Cortisol

I. INTRODUCTION

The negative effects of stress are linked with negative physiological and psychological reactions which can result in the development of hypertension, cardiovascular diseases, anxiety disorders, etc. [1]. In addition, stress is also known to have a negative impact on employee productivity which can impact the economy [2]. Detecting and managing stress is the first step in preventing the negative effects of stress. Stress can be conceptualized as the response of the Autonomic Nervous System (ANS) to various stimuli such as environmental stimulus and physical stimulus, etc. (Figure 1).

As illustrated in Figure 1, the stress response of ANS can be classified into two types: (i) Physiological response (objective response) and (ii) Psychological response (subjective response). Subjective response is the stress perceived by an individual and objective response is the effect of stress on the physiological activities like increased cortisol level, skin conductance, heart rate, sleep abnormality etc. An experiment conducted in [3], showed that perceived stress and objective measures of stress are highly correlated.

A. Related Work

The related work in the area of stress detection can be classified into two main approaches:

- **Approach 1**: This method of classifying stress levels is based on the subjective response of an individual obtained by employing rating questionnaires such as the DASS 21, STAI, POMS etc. Limitations associated with this approach are that these approaches are difficult to implement in real time. Also, the proper interpretation of scores obtained from these measures necessitates the feedback of someone with appropriate training, usually a professional with a background in the psychological science.

- **Approach 2**: This method to classify stress levels is based on using a learning model for training features from physiological signals in combination with the perceived/subjective ratings from individuals to distinguish between stressed and non-stressed states. Although this approach can be implemented in real time, the major challenge associated with this method is the reliability of ground truth [4]. Also, it cannot be claimed with certainty that the stress levels predicted by the machine learning models will perform the same when they are used to predict stress levels of an entirely different population sample.

B. Motivation

In this work, we have attempted to address these limitations in the existing literature by modeling a subject independent stress model. Although work done in [5] addressed the classification of stress using objective response, their study did not address the robustness of the classification model in predicting stress classes when the...
test set comes from different subjects than that of training set. We therefore put forward the following hypothesis.

**Hypothesis 1.** Given two sets of subjects $A$ and $B$ such that $B \not= A$, It is possible to develop a stress model $M$ trained on $A$ with its physiological response that can generalize to the stress responses of $B$ with significantly good accuracy.

In this work, we will present experimental analyses and results to support our hypothesis. We will be training a model with training data $(A)$ which consists of 10 subjects and test the accuracy of the trained model on the test data $(B)$ which consists of 3 subjects. The experimental support for this work has been obtained in collaboration with the Sanders-Brown Center on Aging at the University of Kentucky.

The paper is organized as follows: Section II discusses the experimental steps and the process of data collection. Section III presents the process of data analysis and the stress classification model. Section IV discusses the results and Section V concludes the article.

II. EXPERIMENTAL PROTOCOL AND DATA COLLECTION

A. Experimental Protocol:

For the experimental protocol, the Trier Social Stress Test (TSST) has been selected because of its capability to capture social and psychological stress in a naturalistic environment [6]. Figure 2 illustrates the steps in the experimental protocol. The protocol consists of a waiting period and a Pre Stress (PS) period for 20 minutes, Anticipatory Stress (AS) period of 10 minutes followed by Speech and Math tasks for another 10 minutes. After the stress induction period, there is a recovery period for a total of 20 minutes. Saliva samples are taken for cortisol sampling from individuals five times during the experimental procedure to estimate the objective stress. The waiting period consists of participants completing a demographic questionnaire, signing the consent form and setting up the sensors. In the PS stage, baseline measurements are taken. During the AS period, participants are exposed to anticipatory stress as they are required to prepare the speech mentally for 10 minutes. During the stress induction period, participants are required to speak on a given topic continuously for 5 minutes and provide correct answers to mental math questions while being observed for 5 minutes. The recovery phase is a total of 40 minutes. For individuals in the intervention group- the first 20 minutes of the recovery is a relaxation induction utilizing guided meditation and breathing and nature videos, followed by 20 minutes of quiet rest. In our control group the entire 40 minute recovery period was quiet rest. After the second recovery phase, participants provide a final saliva sample. The second recovery phase is required because the cortisol reading from the saliva sample has a lag of approximately 20 minutes.

B. Data Collection:

Data collected during this study can be broadly divided into two categories as follows

- **Physiological Data:** For measuring physiological data, we have used the shimmer device integrated with iMotions platform. The shimmer sensor is equipped with a skin conductance sensor and an optical sensor for measuring GSR and PPG signals respectively.
- **Cortisol Sampling:** Cortisol, also known as the stress hormone, is measured typically by saliva processing. The HPA axis reacts to stressful situations by the increased secretion of cortisol. Cortisol is used as an objective measure for stress and is regarded as the gold standard for physiological stress measurement. Saliva is collected via a swab which the participant has to keep under their tongue for approximately two minutes, which is then placed in a tube. These samples are then preserved in a temperature of -20 to -30 degree Celsius in a freezer. The samples are then sent out for processing.

III. DATA ANALYSIS AND STRESS CLASSIFICATION

The implementation of the stress detection system is divided into three broad phases: (i) Data Preprocessing, (ii) Feature Extraction and Selection and (iii) Training and Testing. The overview of the system is shown in figure 3

A. Data Preprocessing

The data preprocessing module is composed of Normalization, Filtering and Integration units:

- **Normalization:** This unit normalizes the GSR and PPG values to be within the range 0 and 1.
- **Filtering:** This unit filters the normalized GSR and PPG signals. A Butterworth filter of 5 order is used to filter off signal components that are higher than 10 Hz.
- **Integration:** In this unit, the filtered signals from different time periods of the experiment are integrated for feature extraction.

B. Feature Extraction and Selection

The feature extraction and selection module consists of four functional units: (i) GSR Feature Extraction Unit: Extracts feature from GSR signal, (ii) PPG Feature Extraction Unit: Extracts feature from PPG signal, (iii)
Standardization Unit: Standardizes the train data and test data and (iv) Feature Selection Unit: Selects important subsets of features based on training feature and training label.

1) GSR Feature Extraction: A total of 8 features were extracted from the GSR signal. These features are based on the peak characteristics occurring as a result of arousal. The peaks are detected using a window of 30 second with an overlap of 15 second by applying the `find_peaks` algorithm [7], on the first derivative of the GSR signal. This is illustrated in Figure 4. From the GSR peak, features calculated were (i) peak frequency, that is the number of peaks occurring per second, (ii) mean, rms and range of peak time and (iii) mean, standard deviation, rms and range of the peak values in a window size of 30 second.

2) PPG Feature Extraction: A total of 9 features were extracted from PPG signal based on Heart Rate (HR) and Heart Rate Variability (HRV) features. Two HR features: (i) mean HR and (ii) HR range over 30 second window. Four HRV features: (i) HRV, (ii) HRV range, (iii) HRV standard deviation and (iv) HRV rms value over 30 second window. The rest three features were mean, standard deviation and rms value of systolic peak values over 30 second window.

Systolic peaks were obtained by using the `find_peaks` algorithm. Peaks with height thresholds of 0.4 and widths of 10% of the sampling rate were classified as systolic peaks. Also, the minimum distance between peaks was set as about 65% of the sampling rate. Figure 5 shows the peaks classified as systolic peaks marked by blue cross along with the original PPG signal. Subsequently, these peaks were used to calculate HR and HRV related features.

Hence, a total of 17 features were extracted from both GSR and PPG signal streams. All the features were then integrated into a dataframe. Each sample is also marked by its corresponding labels generated from cortisol analysis together with an ID number which denotes the subject number. After this procedure, the feature set is divided into two sets: train and test data. The ID associated with each sample ensured that no two samples of the same subject are in both the train and the test set.

In the next step, the training feature and the training labels are separated in to two different input lines to the Feature Selection Unit.

3) Feature Selection Unit: The main purpose of this unit is to reduce the number of input features to the model to control over fitting. In this unit, each feature is scored by computing the ANOVA F value between the feature and the target class. These scores are then used to remove features with high auto correlation by removing duplicate scores. These scores are then sorted in descending order.
From the remaining features, the first 5 features based on the score are selected for input to the machine learning model.

4) Standardization Unit: This unit transforms the training feature set and the test feature set into a standard form to remove bias in the learning model. This is done by fitting a standard scalar to the train data so that all the features have close to zero mean and unit variance. Subsequently, the train and test set were transformed with the fitted standard scalar to obtain the transformed feature sets.

C. Training and Testing

This module performs the training of the model by fitting a random classifier model to the training set and build the predictive model. The fitted classifier had 100 estimators with the gini function to measure the quality of split and maximum depth was 30. Subsequently, the predictive model was tested with the test feature and its performance was evaluated by comparing the prediction of the model with its actual test labels.

D. Ground Truth Estimation

The ground truth is estimated from the cortisol values obtained from the processing of saliva samples. These cortisol values were analyzed and used to annotate each time point as either stressed (S) or not stressed (NS). For each participant, there were 5 cortisol samples, two of which were from the pre study period and three of were during the study.

Normally, cortisol levels reach their peak in the morning and fall gradually as the day progresses [8]. To reduce the in between differences of each participant’s cortisol value arising from circadian rhythms, the following preprocessing steps were implemented

- The first two samples were averaged and then subtracted from the remaining three samples for each participant.
- The three samples for each participant were integrated and standardized by subtracting the mean from each sample and dividing by the standard deviation.
- A sample \( x \) was classified as stress (S) if \( x > \mu \) and as not stressed (NS) if \( x < \mu \), where \( \mu \) is the population mean.

Analysis showed that the experimental protocol was effective in generating stress response during the stress induction phase which is evident by the increase in cortisol level. The relaxation intervention was effective in producing a calming effect as evidenced by the corresponding decrease in cortisol level. Figure 6 shows the cortisol and GSR trends during each of the Pre Stress (PS), Stress (S) and Relax (R) phases.

<table>
<thead>
<tr>
<th>TABLE I: Performance of the Predictive Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Stress</td>
</tr>
<tr>
<td>Not Stressed</td>
</tr>
<tr>
<td>Macro Average</td>
</tr>
<tr>
<td>Weighted Average</td>
</tr>
</tbody>
</table>

IV. Results

In this section, we will discuss the performance of the classifier with the test data. We have evaluated the model’s performance using accuracy, precision and f1-score of each
of the classes along with their weighted and macro average. This is reported in Table I.

In Table I it can be seen that the model has achieved good precision, recall and f1-score in detecting stressful events. The macro average and the weighted average for precision, recall and f1-score also shows the robustness of the model in detecting stressful events. The overall accuracy of the model is reported as 92%.

The experimental results indicate that our hypothesis has been successfully validated. According to our hypothesis, a learning model $M$ after being trained with features of a set of subjects $A$ (training set) was able to successfully generalize predictions for $B$ (test set) with significant accuracy.

### A. Comparison to Existing Work

Table II shows the comparison of our work with [5] as both have used objective methods to classify stress. As both these investigations were implemented in different technological and environmental configurations, we report the experimental and evaluation variables for further discussion.

From Table II, it can be seen that work done in [5] used 3 signals (ECG, GSR, EEG) for classifying stressful events. However, for everyday monitoring, measuring ECG and EEG are not convenient. Wrist-based devices which capture ECG require additional hardware configuration. EEG measurement requires probes to be placed around the head which is not feasible for supporting long-term real time monitoring. In our work, we have addressed this limitation by using only GSR and PPG signals to estimate the occurrence of stressful events. Sensors for capturing GSR and PPG signals are inexpensive and convenient for real time monitoring. Devices like Empatica E4 wristband has both GSR and PPG sensors which makes it ideal for monitoring stress levels in a non-invasive/unobtrusive manner. Hence, the proposed model will be an ideal fit for integration in smart health devices to facilitate continuous stress monitoring.

Also, the acquisition window for our model is only 30 sec. Hence, the time interval required to obtain feedback from the system regarding the stress levels will also be significantly less.

Furthermore, details of testing the model on a separate group of test subjects, independent from training subjects was not reported in [5]. Our work addresses this limitation by testing the model on subject data that is independent from the training set data.

### V. Conclusion

In this work, we have presented the validation of a generalized physiological stress detection model using cortisol as the bio marker for stress. The proposed model utilizes two input signal streams: (i) GSR and (ii) PPG. The model has been able to achieve an overall accuracy of 92% in classifying stress from non-stressed states.

Such a subject independent stress model will not require training on each device separately in order to be able to recognize individual stress levels. Hence, relating the physiological signals (non-invasive) to the cortisol (invasive) which is regarded as the gold standard of stress measurement will allow for the development of a generalized subject independent stress model.

### VI. Acknowledgments

This work was supported by the Kentucky Science and Engineering Foundation under Grant KSEF-3528-RDE-019.

### References


