Multi-Modal Acute Stress Recognition Using Off-the-Shelf Wearable Devices

Victoriano Montesinos, Fabio Dell'Agnola, Adriana Arza, Amir Aminifar, and David Atienza[1]

Abstract— **Monitoring stress and, in general, emotions has attracted a lot of attention over the past few decades. Stress monitoring has many applications, including high-risk missions and surgical procedures as well as mental/emotional health monitoring. In this paper, we evaluate the possibility of stress and emotion monitoring using off-the-shelf wearable sensors. To this aim, we propose a multi-modal machine-learning technique for acute stress episodes detection, by fusing the information careered in several biosignals and wearable sensors. Furthermore, we investigate the contribution of each wearable sensor in stress detection and demonstrate the possibility of acute stress recognition using wearable devices. In particular, we acquire the physiological signals using the Shimmer3 ECG Unit and the Empatica E4 wristband. Our experimental evaluation shows that it is possible to detect acute stress episodes with an accuracy of 84.13%, for an unseen test set, using multi-modal machine-learning and sensor-fusion techniques.**

I. INTRODUCTION

Stress represents a global issue [1] and is associated with many diseases. Stress increases the risk of many health pathologies, including depression, sleep disorders, and heart diseases, among others [2]. Stress can also affect workers' performance [3] in high-risk occupations. As a result, the problem of stress detection has attracted a lot of attention over the past few decades due to its negative effects, with special emphasis put on building a pervasive stress monitoring and recognition system.

Despite several advances in the stress monitoring domain, stress detection still represents a challenging task, because of its characteristic response. First, the stress response is highly subjective, i.e., different people react differently to a given stress event [4]. Second, no common and accepted stress measurement standard exists to date [5], [6], [7]. Third, it has a multilevel reactivity, namely, physiological, behavioral, and cognitive [5], [8]. Currently, stress assessments rely on questionnaires and analytic biomarkers obtained from blood and/or saliva samples [6]. The main drawback of these methods is that they do not allow a continuous measure of the stress level. In addition, the latter options, based on blood and saliva samples, are intrusive and costly. Therefore, monitoring biomarkers extracted from stress physiological response represents a viable solution for pervasive stress recognition using wearable sensors.

Today, the use of wearable sensors enables continuous monitoring of several physiological signals and physical activity using, e.g, electrocardiogram (ECG), blood volume pulse (BVP), skin temperature (SKT), respiration (RSP), and electrodermal activity (EDA). These technologies offer a solution to pervasive monitoring, which allows timely and unobtrusive monitoring of a person's physiological state, as opposed to the sporadic entries of questionnaires and biochemical samples [9], [10]. Furthermore, smart wearable sensors enable the implementation of stress detection methods by using embedded machine-learning algorithms [11], [12], which are capable of combining multiple stress biomarkers extracted from different physiological signals.

Stress detection methods based on bio-markers of the physiological stress response have been extensively explored in the literature [7], [13], [14], [15]. The current accessibility of wearable devices (e.g., Shimmer ECG units [16], Empatica E4 wristband [17], and Polar H10 heart rate sensor [18]) has allowed to assess the aforementioned problem using wearable sensors [4], [19], [20]. These studies apply several machine-learning techniques, such as support vector machines [13], [4], [15], decision trees [4], and linear regression models [7].

In [4], the authors present a stress recognition method using Inter-Beat-Interval and EDA signals from the Empatica E4 wristband, reporting a recall of 70% and a precision of 95% using cross-validation. Similarly, in [13], the authors reach an accuracy of 89.7% and a sensitivity of 88.5% using cross-validation, based on biomarkers from EDA, RSP and ECG, and classify three stress levels. In [15], the authors report 94% of accuracy and sensitivity using cross-validation, based on wearable sensors with ECG and RSP signals.

The majority of the previous studies on stress detection, however, report the performance obtained using only cross-validation and training, without any evaluation on unseen test sets. Such results can be misleading, since the machine-learning models tend to suffer from overfitting (high variance) when trained and evaluated on the training set. Hence, these models may fail to generalize when considering unseen data sets [21].

In this article, we aim at developing a multi-modal machine-learning and sensor-fusion technique to recognize stress based on biomarkers extracted from physiological signals, acquired from off-the-shell wearable devices. The main contributions of this work are the following:

- A multi-modal acute stress recognition technique to reliably differentiate between the relax and stress states, based on state-of-the-art off-the-shelf wearable sensors and physiological signals, including ECG, EDA, RSP, BVP, and SKT. We demonstrate the generalization power of our proposed technique by evaluating its detection performance on an unseen test set, reaching an average accuracy of 84.13% for classification between an induced relax state and intense cognitive task.

*This work has been partially supported by the NCCR Robotics through the Symbiotic Drone project, the ONR-G through the Award Grant No. N62909-17-1-2006, and the WiTNESS project (Promobilia Foundation project No. 18079).

[1] V. Montesinos, F. Dell'Angola, A. Arza, A. Aminifar and D. Atienza are with the Embedded Systems Laboratory of the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.

- Comparison between two state-of-the-art off-the-shelf wearable sensors, i.e., the Shimmer3 ECG Unit and the Empatica E4 wristband, in the context of acute stress recognition. Our experimental evaluation shows that acute stress detection based only on the set of physiological signals provided by the Shimmer3 ECG Unit has an average accuracy of 76.19%, while acute stress detection based only on the set of physiological signals of the Empatica E4 wristband has an average accuracy of 78.57%.

II. METHODOLOGY OVERVIEW

The methodology followed in this work consists of three main stages, as depicted in Figure 1. The signal acquisition stage has as output the monitored physiological signals, from which stress biomarkers are obtained in the signal processing stage. Then, a set of features is extracted from these biomarkers, to feed into the last stage, which is the model learning and evaluation phase.

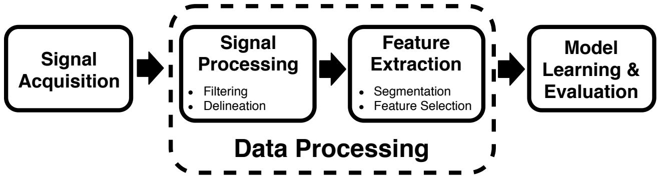

Fig. 1. Overview of our proposed stress recognition methodology

1) Signal Acquisition: The selected physiological signals that can be measured using off-the-shelf wearable devices are ECG, BVP, SKT, RSP, and EDA. Indeed, the physiological stress response triggers the activation in four main body systems that affect the selected signals, i.e., neurohypophysis (SKT, BVP), hypothalamic-pituitary-adrenal axis (ECG, BVP), sympathetic nervous system (SKT, BVP, RSP, EDA), and parasympathetic counterbalance (ECG, BVP, RSP). Therefore, a comprehensive stress model needs to include multiple biomarkers, as shown in [7].

2) Data Processing: The data processing stage is composed of two phases: signal processing and feature extraction. In the signal processing phase, the raw data are filtered and/or delineated to obtain meaningful biomarkers. In the feature extraction phase, the data from the previous phase is segmented into windows (that may overlap), from which different numerical features are extracted (e.g., time and frequency domain features). Additionally, feature selection is performed in order to focus the attention on the features that are important from a physiological perspective, but also to reduce overfitting, model complexity, and execution time.

3) Model Learning and Evaluation: This stage involves training different machine-learning algorithms and assessing the performance of the trained models using the cross-validation and test sets. The evaluation is done twice in order to: (1) select the most important features and classification algorithm using cross-validation on the training set and (2) demonstrate the generalization power of the machine-learning model on an unseen test set.

III. DATA PROCESSING

The first step towards a stress recognition method is to extract meaningful biomarkers that characterize the physiological stress response. Then, the biomarkers have to be combined to distinguish the different body reactions that different stressors may trigger.

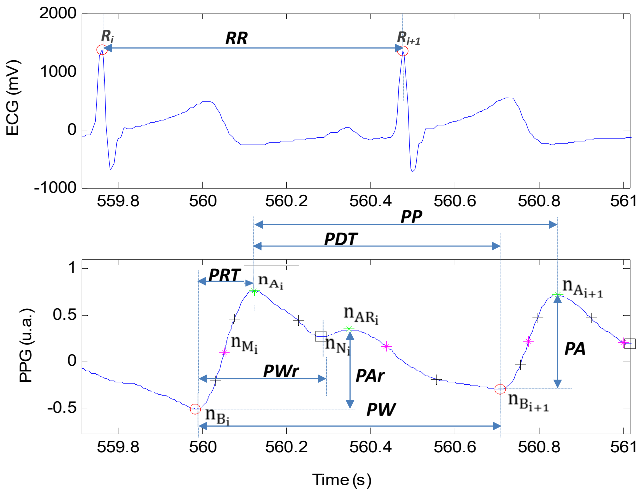

Fig. 2. Detection points in ECG and BVP (PPG) signals.

A. Signal Processing

The ECG, BVP, RSP, SKT and EDA signals are processed to extract time series of biomarkers (parameters) from which the features are extracted later.

The preprocessing and delineation of both the ECG and the BVP signals is done as presented in [22]. From ECG, the *RR* intervals (time between two consecutive R peaks) are obtained, as shown in Figure 2. The delineation of the pulse onsets in the BVP is done slightly different than in [22]: the pulse onsets are detected finding the minimum values that are below the standard deviation of the segment of signal being delineated (windows of 5 seconds are used to segment the signal). Once the onsets are detected, we use the method followed in [22] to delineate the signal between two consecutive onsets. The parameters obtained from BVP are also presented in Figure 2 and listed as follows:

- Pulse Period (PP): Time interval between two consecutive P peaks (n_{A_i}).
- Pulse Amplitude (PA): Amplitude difference between the pulse peak (n_{A_i}) and the pulse onset (n_{B_i}).
- Pulse Wave Rising Time (PRT): Time interval between n_{B_i} and n_{A_i}.
- Pulse Width until Reflected Wave (PWr): Time interval between n_{B_i} and n_{N_i}.
- Pulse Amplitude of the Reflected Wave (PAr): Amplitude difference between the reflected wave's peak (n_{AR_i}) and the pulse onset (n_{B_i}).
- Ratio between the Pulse Amplitude and the Pulse Amplitude of the reflected wave (PA/PAr).
- Pulse Decreasing Time (PDT): Time interval between n_{A_i} and $n_{B_{i+1}}$.
- Pulse Width (PW): Time interval between n_{B_i} and $n_{B_{i+1}}$.
- Slope of the Pulse Wave (k): The slope transit time divided by the difference in amplitude between 1/4 and 3/4 of the pulse wave.

The RSP signal is band-pass filtered (cut-off frequencies of 0.03 and 0.9 Hz) and down sampled to 4 Hz with a 50th order median filter, in order to remove the noise (motion artifacts) [23]. From the filtered signal, the respiration rate (RSP_{rate}) and period (RSP_{prd}) are obtained using the differences between peaks, which are computed using the first order derivative.

The SKT signal is smoothed with a first order median filter. Previous assessments of the SKT signal [4], [7] showed that the patterns of slope give more meaningful information (resulting in improved classification accuracy) than the patterns of mean for this signal. Thus, the parameters obtained are the gradient (ΔT) and the total power (T_{Pt}) from the PSD.

The EDA signal is divided into two main components: the skin conductance level (SCL) and the skin conductance response (SCR). First, the EDA signal is filtered using a sixth order Butterworth lowpass filter. [24] present a deconvolution method that transforms the filtered signal into a driver with intensified SCRs; then, it separates the driver into tonic and phasic components, and convolves each component to obtain the SCL and SCR, respectively. We follow this approach to obtain the SCL, SCR and driver phasic component from the EDA signal.

B. Feature Extraction

A total of 96 features to explore as stress markers are extracted from the set of physiological signals considered in this paper: 16 from ECG, 17 from RSP, 56 from BVP, 5 from EDA and 2 from SKT. All the signals are segmented into 60-second windows with an overlap of 30 seconds from which the features are extracted. We first, explore that wide variety of features to investigate which one gives discriminatory information. Finally, we perform feature selection to considerably reduce the number of features (see Section III-C).

1) ECG Data: From RR intervals, time and frequency domain heart rate variability (HRV) and non-linear features from Poincare plot are extracted [25]. The time domain features include the mean, standard deviation (STD), standard deviation of the differences between adjacent intervals (SDD), number of interval differences of successive intervals greater than 50 ms ($NN50$), and the proportion derived by dividing $NN50$ by the total number of intervals ($PNN50$). The frequency domain features are obtained from the Lomb-Scargle PSD, with the following frequency components: very low frequency (VLF, less than 0.04 Hz), low frequency (LF, 0.04 to 0.15 Hz) and high frequency (HF, 0.15 to 0.4 Hz). The features are: VLF power normalized ($nVLF$), LF power normalized (nLF), HF power normalized (nHF) and power ratio LF/HF. The normalization of each frequency component is done by the total frequency power minus the VLF components. Moreover, additional features are generated based on the Poincare (Lorenz) analysis (for the first and second order difference) using fluctuations on the longitudinal axis (L), fluctuations on the transverse axis (T) [26]: cardiac sympathetic index (CSI: L/T), modified cardiac sympathetic index (CSI_{mod}: L^2/T) and cardiac vagal index (CVI: $\log_{10}(L \cdot T)$).

2) BVP Data: PP intervals features are equal to the aforementioned for RR intervals. From PRT, PWr, PDT and PW parameters, the time domain features include the mean, standard deviation (STD) and standard deviation of the differences between adjacent intervals (SDD), and frequency domain features are the same as for the RR intervals. From the parameters based on amplitude (PA, PAr, PA/PAr and k), only time domain features are obtained: mean, STD and SDD.

3) RSP Data: From the respiration parameters, the following features are obtained: $RSP_{RATE_{MEAN}}$ and $RSP_{RATE_{STD}}$, and frequency domain features: we estimate the mean frequency distribution in the HF band. This estimation is performed in two different ways: 1) by fitting a Gaussian in the HF band (RSP_{HF_Gauss}), and 2) by computing a weighted mean of all the frequency components in the HF band using as weight the power amplitude at each frequency (RSP_{HF_pond}). Finally, we compute the PSD of five HF subbands of equal bandwidth (RSP_{HF_pFi}, with i from 1 to 5). Additionally, for each window of analysis we applied the method proposed in [23] to compute the estimated respiratory frequency (RSP_{eRF}) as the largest peak (RSP_{Pk}) of the Lomb-Scargle power spectral density (PSD) of respiration using a Welch periodogram on 12-second windows with an overlap of 6 seconds. The power around the peak is computed in a bandwidth of 0.04 Hz, and divided by the total power (RSP_{Pt}). This normalized respiratory peak power ($RSP_{Pk_{norm}}$) represents the concentration of power around the RF and is related to the variability of the RF within the interval. Two features are obtained from the filtered skin temperature signal: ΔT, which is computed as the difference between the last (not missing) value of the window minus the first one, and the power, SKT_{Pt}. Regarding the EDA data, SCL_{MEAN}, SCL_{STD}, gradient (ΔSCL) and total power from PSD (SCL_{Pt}) are extracted from the SCL, while total power (from PSD) is obtained from the driver phasic component (SCR_{Pt}).

C. Feature Selection

In order to select the relevant subset of features, we first explore the whole set of features to select the only features that gives discriminatory information for the classification problem applying the Student t-test for paired samples. Next, correlated features (per signal) with a Pearson's correlations coefficient higher that 0.99 are eliminated from the initial set. Finally, we apply a wrapper feature selection method based on the estimation of the feature importance given by a random forest model. Tree-based machine learning algorithms implicitly perform feature selection when building the model. After training the random forest classifier on the training set, the features' weights are obtained. The higher the weight, the higher the importance. Next, based on the weight distribution, the most important features for each classification problem are empirically selected.

IV. MODEL LEARNING AND EVALUATION

After processing the data, three different low-complexity machine learning algorithms are used to build the classification models. The selection of those low-complexity algorithms is done according to the limited-resources of the wearable sensors, i.e., ultra-low-power microcontrollers and low-memory.

The evaluation of the different experiments is done based on three different classifiers:

- k-Nearest Neighbors (kNN): the kNN algorithm classifies each sample based on the labels of its k nearest

neighbors. The label of the new data sample is determined by majority voting among the labels of its k nearest neighbors, according to a certain distance metric.
- Decision Tree (DT): the DT algorithm classifies each sample based on a set of conditions on features, structured as a decision tree. Each data sample is associated with one of the leaves of the tree and each leaf of the decision tree is associated with one of the classes.
- Random Forest (RF): the RF algorithm is a robust state-of-the-art classification algorithm that uses the bagging technique to avoid overfitting. The RF algorithm creates a bag of DTs and performs majority voting among the DTs to classify a new data sample.

Finally, to ensure generalization of our results, each dataset is split into a training and a test set, containing the 70% and 30% of the data, respectively. In all the experiments, it was used leave-one-out cross-validation. The number of folds for cross-validation equals the number of subjects in the training set. In this fashion, all the data from a given subject are put together within a fold to avoid the overlapping of data between training and validation sets through the different folds.

V. EXPERIMENTAL SETUP

Our experiment is designed to induce two main states, no-stress and stress, on the participants and measure their response to that stimulus. Additionally, a questionnaire with the subjects' self-reported stress level in a visual analog scale is used to assess the induced stress in each state.

A. Participants

The participants in the experiments include 30 subjects in an age range of 25-35 years. The inclusion criteria are young and healthy subjects, who are non-regular consumer of psychotropic substances, alcohol or tobacco. The exclusion criteria are a body mass index higher than 30, any chronic disease or psychopathology and a stress level higher than 90% on a visual analog scale. Prior to performing the experiment, each subject was asked to give explicit consent after having been explained the details of the experiment. The ethical approval for this study was obtained from the Cantonal Ethics Commissions for Human Research Vaud and Geneva, ethical approval application number PB_2017-00295.

B. Experiment Protocol

The acquisition protocol is limited to a single session for each participant, and its design includes a baseline stage, a relax stage (i.e., no-stress) and a stress stage. The baseline stage takes place first, consisting of a meditation audio [27] to reduce the difference on stress levels that participants may have at the beginning of the experiment. Next, the relax state includes seven scenery clips to induce a relaxing state on the participants. This is followed by two stress stages, comprising four horror clips and a cognitive test (i.e., an arithmetic task) at the end.

Both the scenery and horror clips have been selected from the Emotional Movie Database (EMDB) [28], which is a collection of 40-second film clips (without auditory content) classified on an affective space depending on the emotional stimuli (valence and arousal). Valence indicates if the stimulus is positive or negative, while arousal indicates the level of activation. After the assessment of each clip in the previous affective space, based on the participants' self-report ratings, the authors of [28] conclude that the selected scenery clips produce low arousal and high valence, while horror clips produce a state of high arousal and low valence.

TABLE I
SHIMMER3 ECG UNIT AND EMPATICA E4 TECHNICAL SPECIFICATIONS

Device	Sensor	Acronym	F_s (Hz)
Shimmer3 ECG 	Electrocardiogram	ECG	512
	Respiration	RSP	512
	3-axis Accelerometer	ACC	512
Empatica E4 	Skin Temparature	SKT	4
	Blood Volume Pulse	BVP	64
	Electrodermal Activity	EDA	4
	3-axis Accelerometer	ACC	32

The cognitive test is an arithmetic task used in the final stage of the Trier Social Stress Test (TSST) [29], a widely used protocol in stress research, which has proven to be a reliable stressor to induce acute stress [30]. In the TSST, the subjects are asked to serially subtract 13 from 1022, for a period of 5 minutes. If the subjects make a mistake, they have to start again from 1022. In our case, the subjects enter their responses using a numeric keyboard, and the duration of the cognitive test is limited to 3 minutes. The cognitive test is placed after the horror clips stage. As stress produces a decrease in cognitive skills [31], if the subject already comes from a stress state (e.g., high arousal and low valence), the number of errors in the arithmetic task will be higher and, consequently, the stress level is expected to rise.

In summary, the structure of the experiment is as follows:

1) Sensors' placing and signal stabilization (10 min).
2) Meditation audio (3 min) [27].
3) Scenery clips (4 min 40 s) from the EMDB [28].
4) Horror clips (2 min 40 s) from the EMDB [28].
5) Cognitive test (3 min): from the TSST [29].
6) Sensors' removal (2 min).

Between contiguous stages of the experiment, a self-report questionnaire is completed by the subject. The questionnaire consists of a visual analogue scale for stress (VASS) in which the subjects can select their stress level at that moment.

C. Sensor Placement and Signal Acquisition

During the experiments, the bio-signals are recorded using two off-the-shelf wearable sensors: the Shimmer3 ECG Unit and the Empatica E4 wristband. The specifications of the sensors of each device are shown in Table V-C. The Shimmer's sensors use four electrodes that we placed on the thorax. The lead II configuration is used to measure ECG, while respiration is demodulated from the ECG signal obtained from arm to arm. The Empatica E4 wristband is placed on the left wrist.

In this particular study, we focus on the use of the main biological signals described previously to detect stress levels. Therefore, the acceleration data is not considered, as the subjects are in a static position. However, this information can be considered in future studies as context information to distinguish what biomarker changes are caused by physical activity with respect to those caused by stressful events [4].

VI. Experimental Results

In this section, we evaluate the proposed acute stress recognition method and demonstrate its performance on an unseen test set. The machine-learning models are built and evaluated in Matlab R2018a using MathWorks's Statistics and Machine Learning Toolbox.

The main goal of this study is to distinguish between the relax state, i.e., scenery clips (SC) and the two stress states that are assumed to produce disgust, i.e., horror clips (HC), and cognitive load, i.e., cognitive test (CT). Table II shows the summary of the participants' reported stress level for each stage. The stress level between the stages is analyzed using the Student t-test for paired samples. As expected, the values for HC and CT stages are higher than in SC (p-value $< 10^{-9}$). Moreover, there are statistically significant differences between HC and CT (p-value $= 0.0194$). These results show that three main states were generated in our study, i.e., relax (SC), disgust/low stress (HC) and cognitive load/moderate stress (CT). However, the wide range and standard deviation of stress levels reported among participants in the same stage of the experiment suggests that there are inter-individual different induced states in the same task.

TABLE II
Summary of participants' reported stress level

Stage	Mean	Min	Max	Std
SC	14.27	0.00	43.00	13.49
HC	54.53	10.00	100.00	26.95
CT	63.03	11.00	100.00	25.67

A. Performance Evaluation Using Both Sensors

Firstly, we explore the wide set of features to remove the non-informative and correlated ones reducing the initial set from 96 to 65 features. Then, we evaluate the performance of our multi-modal stress detection algorithm considering the entire set of physiological signals acquired using both the Shimmer and Empatica sensors.

1) Classification Algorithm Selection: In our first set of experiments, we analyze the accuracy of different classification algorithms in each problem. The results in Table III demonstrate that the random forest (RF) algorithm outperforms the decision tree (DT) and the k-nearest neighbors (kNN) algorithms.

TABLE III
Initial evaluation using cross-validation on the training set

	RF		DT		kNN	
Problem	Mean	Std	Mean	Std	Mean	Std
SC vs HC	64.29	14.81	67.69	20.15	61.56	14.89
SC vs CT	85.71	12.58	83.67	13.95	60.88	16.07
HC vs CT	70.00	21.21	57.14	16.17	49.52	8.65

2) Feature Selection: After training a random forest classifier for each classification problem, the feature weights are obtained. Figure 3 shows the sorted weight values per classification problem. Based on the weight distribution, the ten most important features are selected for each classification problem, i.e., the ten features with the highest weights in the entire feature set.

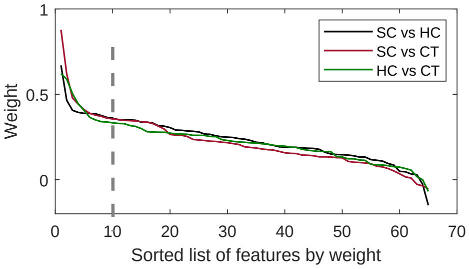

Fig. 3. Sorted list of feature weights per problem

As we can observe in Tables III and IV, the performance increases for the majority of the problems after selecting the most relevant features, according to the cross-validation results. Thus, we consider the reduced feature set when assessing the performance throughout the experiments.

TABLE IV
Evaluation on cross-validation and test set with selected feature set per problem

	Cross-Validation						Test set		
	RF		DT		kNN		RF	DT	kNN
Problem	Mean	Std	Mean	Std	Mean	Std	Acc	Acc	Acc
SC vs HC	63.27	16.18	64.63	22.19	62.24	19.05	65.08	53.17	54.76
SC vs CT	88.10	12.84	84.69	16.64	62.59	7.11	84.13	74.60	62.70
HC vs CT	71.90	24.21	61.43	18.52	50.95	7.00	83.33	77.78	53.33

Using random forest's feature importance estimation, which is based on each feature's weight, the ten most important features for each classification problem are the following:

- SC vs HC: T_{Pt}, PAr_{SDD}, ΔSCL, PAr_{MEAN}, $PP_{CSI_{mod}}$, $PP_{L/T}$, PP_{MEAN}, PRT_{MEAN}, SCL_{MEAN}, RSP_{HF_pF5}.
- SC vs CT: ΔSCL, $RSP_{RATE_{MEAN}}$, RSP_{HF_pF5}, RSP_{HF_Gauss}, PA_{SDD}, PA_{MEAN}, SCL_{STD}, RSP_{HF_pond}, PAr_{SDD}, RSP_{Pk}.
- HC vs CT: PP_{MEAN}, ΔSCL, PWr_{SDD}, PRT_{MEAN}, RSP_{Pk}, PP_{CVI}, SCL_{STD}, PA/PAr_{SDD}, T_{Pt}, PAr_{MEAN}.

From the features listed above, it can be observed that each pair of emotional and/or stress states is best distinguished by certain biomarkers. This may happen because in each problem has the different set features depend on the stress reactivity of each biomarker, i.e, their evolution in time and intensity. Multiple signals and biomarkers are needed to improve the classification performance, as the stress response is multi-dimensional and varies depending on the person's subjectivity and the intensity and type of the stimulus. The results of the VASS scale presented in Table II show the high standard deviations in the participants' evaluation for each stage, i.e., each subject perceives each stimulus differently.

Finally, the results presented in Table IV demonstrate that our proposed multi-modal machine-learning algorithm, based on the random forest algorithm, is able to distinguish between the relaxing task and the intense cognitive task, i.e., SC vs CT, with an accuracy of 84.13%, on average. Moreover, our proposed multi-modal machine-learning algorithm also distinguishes between two different emotion states, i.e., HC vs CT, with 83.33% of accuracy on average.

B. Performance Evaluation Using Each Sensor

In this section, we assess the overall acute stress recognition performance, on both cross-validation (CV) and the test set, for each of the state-of-the-art off-the-shelf sensors.

Table V shows the accuracy obtained from the evaluation on each problem by the Shimmer (*Shim.*) and the Empatica E4 (*E4*) sensors, based on the random forest algorithm. In the context of this article, these results demonstrate that acute stress detection based on the physiological signals provided by the Shimmer3 ECG Unit has an average accuracy of 76.19%, while acute stress detection using only the set of physiological signals of the Empatica E4 wristband has an average accuracy of 78.57%. Moreover, these results demonstrate that using a multi-modal machine-learning technique and combining the information provided by the two wearable sensors, it is possible to improve the stress detection performance to 84.13%.

TABLE V
OVERALL PERFORMANCE ON EACH PROBLEM AND DEVICE USING RANDOM FOREST

	Cross-Validation				Test set	
	Shimmer		E4		Shimmer	E4
Problem	Mean	Std	Mean	Std	Acc	Acc
SC vs HC	62.24	10.63	62.93	20.15	58.73	61.90
SC vs CT	76.19	12.23	77.89	12.14	76.19	78.57
HC vs CT	73.33	17.13	69.05	22.78	64.44	80.00

VII. CONCLUSION

In this article, we have investigated acute stress recognition using off-the-shelf wearable sensors. These results have shown the multi-dimensionality of the stress response that varies due to person's subjectivity and the intensity and type of the stimulus. Accordingly, we have proposed a multi-modal machine-learning technique for stress detection that combines multiple physiological signals from two different wearable devices in order to improve the detection performance. We demonstrated that our proposed multi-modal machine-learning technique is able to distinguish between the induced relax state and the intense cognitive task with an average accuracy of 84.13%, on an unseen test set.

REFERENCES

[1] "STRESS FACTS — Global Organization for Stress." [Online]. Available: http://www.gostress.com/stress-facts/

[2] S. Cohen, D. Janicki-deverts, and G. E. Miller, "Psychological Stress and Disease," *JAMA - Journal of the American Medical Association*, vol. 298, no. 14, pp. 1685–1687, 10 2007.

[3] "WHO — Stress at the workplace," 2010. [Online]. Available: https://www.who.int/occupational_health/topics/stressatwp/en/

[4] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of Biomedical Informatics*, vol. 73, pp. 159–170, 9 2017.

[5] H. Selye, "Stress and the General adaptation syndrome," *British Medical Journal*, pp. 1383–1392, 1950.

[6] E. S. Epel, A. D. Crosswell, S. E. Mayer, A. A. Prather, G. M. Slavich, E. Puterman, and W. B. Mendes, "More than a feeling: A unified view of stress measurement for population science," *Frontiers in Neuroendocrinology*, vol. 49, pp. 146–169, 2018.

[7] A. Arza, J. M. Garzón-Rey, J. Lázaro, E. Gil, R. Lopez-Anton, C. de la Camara, P. Laguna, R. Bailon, and J. Aguiló, "Measuring acute stress response through physiological signals: towards a quantitative assessment of stress," *Medical & Biological Engineering & Computing*, pp. 1–17, 8 2018.

[8] D. H. Hellhammer, A. A. Stone, J. Hellhammer, and J. Broderick, "Measuring Stress," in *Encyclopedia of Behavioral Neuroscience*. Elsevier Ltd, 2010, vol. 2, pp. 186–191.

[9] D. Sopic, A. Aminifar, and D. Atienza, "e-Glass: A Wearable System for Real-Time Detection of Epileptic Seizures in Children," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018.

[10] G. Surrel, A. Aminifar, F. Rincón, S. Murali, and D. Atienza, "Online Obstructive Sleep Apnea Detection on Medical Wearable Sensors," *IEEE Transactions on Biomedical Circuits and Systems*, no. 99, pp. 1–12, 2018.

[11] F. Forooghifar, A. Aminifar, and D. Atienza Alonso, "Self-Aware Wearable Systems in Epileptic Seizure Detection," p. 7, 2018.

[12] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-Time Event-Driven Classification Technique for Early Detection and Prevention of Myocardial Infarction on Wearable Systems," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 5, pp. 982–992, 10 2018.

[13] L.-l. Chen, Y. Zhao, P.-f. f. Ye, J. Zhang, and J.-z. z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, vol. 85, pp. 279–291, 11 2017.

[14] S. Greene, H. Thapliyal, and A. Caban-Holt, "A Survey of Affective Computing for Stress Detection: Evaluating technologies in stress detection for better health." 2016.

[15] L. Han, Q. Zhang, X. Chen, Q. Zhan, T. Yang, and Z. Zhao, "Detecting work-related stress with a wearable device," *Computers in Industry*, vol. 90, pp. 42–49, 2017.

[16] Shimmer, "Shimmer3 ECG Unit. Official Website," http://www.shimmersensing.com/products/shimmer3-ecg-sensor.

[17] Empatica, "The E4 Wristband. Official Website," https://www.empatica.com/en-eu/research/e4/.

[18] "Polar H10 heart rate sensor — Polar Global." [Online]. Available: https://www.polar.com/en/products/accessories/H10_heart_rate_sensor

[19] Z. W. Adams, E. A. McClure, K. M. Gray, C. K. Danielson, F. A. Treiber, and K. J. Ruggiero, "Mobile devices for the remote acquisition of physiological and behavioral biomarkers in psychiatric clinical research," *Journal of Psychiatric Research*, vol. 85, pp. 1–14, 2 2017.

[20] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic Stress Detection in Working Environments From Smartphones Accelerometer Data: A First Step," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1053–1060, 7 2016.

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[22] A. Arza Valdés, "Measurement of acute psychological stress," Ph.D. dissertation, Universitat Autonoma de Barcelona, 9 2017. [Online]. Available: http://www.tesisenred.net/handle/10803/458131

[23] A. Hernando, J. Lazaro, E. Gil, A. Arza, J. M. Garzon, R. Lopez-Anton, C. de la Camara, P. Laguna, J. Aguilo, and R. Bailon, "Inclusion of Respiratory Frequency Information in Heart Rate Variability Analysis for Stress Assessment," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1016–1025, 7 2016.

[24] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity." *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 6 2010.

[25] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology., "Heart rate variability: standards of measurement, physiological interpretation and clinical use." *Circulation*, vol. 93, no. 5, pp. 1043–1065, 3 1996.

[26] F. Dell'Agnola, L. Cammoun, and D. Atienza, "Physiological characterization of need for assistance in rescue missions with drones," in *IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1 2018, pp. 1–6.

[27] T. H. Guys, "MINDFULNESS - 3 MINUTE MEDITATION," https://www.youtube.com/watch?v=evJHBLldMsE, 2016.

[28] S. Carvalho, J. Leite, S. Galdo-Álvarez, and . F. Gonçalves, "The Emotional Movie Database (EMDB): A Self-Report and Psychophysiological Study," *Applied Psychophysiology and Biofeedback*, vol. 37, no. 4, pp. 279–294, 12 2012.

[29] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting." *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[30] B. M. Kudielka, D. H. Hellhammer, and C. Kirschbaum, "Ten years of research with the Trier Social Stress Test," in *Social Neuroscience*, 2007, pp. 56–83.

[31] G. E. Giles, C. R. Mahoney, T. T. Brunyé, H. A. Taylor, and R. B. Kanarek, "Stress effects on mood, HPA axis, and autonomic response: Comparison of three psychosocial stress paradigms," *PLoS ONE*, vol. 9, no. 12, p. e113618, 12 2014.