Stress Detection with Deep Learning Approaches Using Physiological Signals

Fabrizio Albertetti[1], Alena Simalastar[2], and Aïcha Rizzotti-Kaddouri[1(✉)]

[1] School of Engineering, Haute Ecole Arc Ingénierie, Neuchatel, Switzerland
{fabrizio.albertetti,aicha.rizzotti}@he-arc.ch
[2] School of Engineering, Haute Ecole Spécialisée Valais-Wallis, Sion, Switzerland
alena.simalatsar@hevs.ch

Abstract. The problem of stress detection and classification has attracted a lot of attention in the past decade. It has been tackled with mainly two different approaches, where signals were either collected in ambulatory settings, which can be limited to the period of presence in the hospital, or in continuous mode in the field. A sensor-based continuous measurement of stress in daily life has a potential to increase awareness of patterns of stress occurrence. In this work, we first present a data-flow infrastructure suitable for two types of studies that conforms with the data protection requirements of the ethics committee monitoring the research on humans. The detection and binary classification of stress events is compared with three different machine learning models based on the features (meta-data) extracted from physiological signals acquired in laboratory conditions and ground-truth stress level information provided by the subjects themselves via questionnaires associated with these features. The main signals considered in current classification are electro-dermal activity (EDA) and blood volume pulse (BVP) signals. Different models are compared and the best configuration yields an F_1 score of 0.71 (random baseline: 0.48). The importance on prediction of phasic and tonic EDA components is also investigated. Our results also pave the way for further work on this topic with both machine learning approaches and signal processing directions.

Keywords: Physiological monitoring · Stress prediction · Sympathetic and parasympathetic activation · Affective computing · Telemonitoring · Self-management systems

1 Introduction

At the physiological level, stress is an organism's response to some external stimuli, or a challenge. In presence of stressor, the "fight or flight" response is

This project is funded by the University of Applied Sciences and Arts of Western Switzerland.

F. Albertetti and A. Simalastar—Both authors contributed equally to this paper.

R. Goleva et al. (Eds.): HealthyIoT 2020, LNICST 360, pp. 95–111, 2021.
https://doi.org/10.1007/978-3-030-69963-5_7

activated through the sympathetic nervous system (SNS), which results in release of cortisol and adrenaline, leading to heart rate increase, sweating, and increased concentration of all senses on current situation. The parasympathetic nervous system (PSNS) works in concert with SNS. Its main function is to activate the 'rest and digest' response and return the body to homeostasis after the "fight or flight" response. This results in a reversion of the physical effects of SNS activation and particularly in a heart-rate decrease. Both SNS and PSNS represent the autonomous nervous system (ANS).

In a sense, stress is a natural reaction of the organism. However, there exist many studies showing the link between stress and illnesses [18]. This means that it is not the fact of stress that causes problems to the organism, it is the level of stress that might be excessive to an organism, such that the PSNS fails to return it to homeostasis. Such an excessive and often prolonged stress is called a *distress*. Identification of distress is not simple, since asking a person about how she or he thinks, or feel is susceptible to a wide range of biases since humans are very often not even aware of how they are affected by various stimuli or situations. This way, it is important to give an objective quantitative evaluation to the level of stress and study the activation of ANS as the first step towards definition of the border between a positive stimulation of the organism and distress. This may allow to not only detect the stress conditions leading to distress, but potentially reduce the fear of stress and its unnecessary consequences.

The approaches to stress detection can be roughly classified into: 1) those performed in the ambulatory setting during a relatively short period of time, and 2) those that are performed during the long term when the participant continue his/her normal life activities. The signals reflecting the ANS activity, can be divided into *physiological*, such as, for example, electro-dermal activity (EDA) [4], heart rate (HR), heart rate variability (HRV) [15], and levels of cortisol [13,20], and *behavioural*, such as smartphone activity statistics [16], and annotated geolocation. It is clear that experiment settings define the set of signals that must be considered as more reliable for that experiment. It is obvious that *behavioural* signals make much less sense in the ambulatory settings as well as the level of cortisol, since its level is a subject to circadian rhythms. The other *physiological* signals (EDA, HR, and HRV), in contrast, are less reliable in long terms studies since they are often heavily corrupted with the movement artifacts that are difficult to filter out. However, their good classification in the laboratory setting could help to find the means to improve their use in the long term studies.

The ultimate goal of our study is a system for real-life seamless monitoring of stress. Therefore, we have first created a data-flow application suitable for two types of studies that conforms with the data protection requirements of the ethics committee monitoring the research on humans, as described in this paper.

The stress classification approach presented in this paper is covering the experiments performed in the laboratory setting during which EDA, and HRV signals were collected by means of the Empatica E4 wrist bracelet. The participants of the experiment were induced with four types of stress stimuli, aiming to provoke *emotional*, *intellectual*, and *physical activity* types of stress as well as

pain, alternated with relaxation periods. The signals are annotated with the indicators of relative changes of perceived stress levels provided by each participant. Further, all the signals are processed and vectors of important signals features are extracted. The vectors of signal features with stress indicators are then combined into simple or multiple windows and given as an input to machine learning based models. Furthermore, a comparison of deep learning models is presented.

This paper is organized as follows: Sect. 2 presents related work for stress evaluation, Sect. 3 provides details on the dataset, Sect. 4 presents details of each data-flow step and Sect. 5 discuss experimental results. Finally, in Sect. 6 we summarize the importance of our contribution and suggest some future work.

2 Related Work

Several works proposed in literature aim to detect the condition of stress and estimate the level of mental effort by using wearable sensors and mobile applications, such as in [3] and [12], which have demonstrated that smartphone data can be used for mood classification.

Physiological measures such as EDA, HR and HRV are frequently used in studies related to affection and well-being [16]. [6] proposes a smart-watch based system to collect and analyses biosignal data to detect unobtrusively and at low cost mental stress condition during daily life activities. In particular, EDA has long been used to study a variety of physiological subjects including stress, emotion, depression, anxiety, attention and information processing [7]. In [16] the link between EDA and stress is explored. In the same study, the authors collected data for the analysis and prediction of stress from smartphone logs. [14] proves that the EDA is sensitive to cognitive stress during water immersion while others used derivatives of the BVP signal as in [9] where information on respiratory rate (RR) and HRV is analyzed to obtain reliable interpretation parameters for stress assessment.

Some works have also added other types of data to better support their results as in [2] which adds diameter of the pupil to the characteristics of the user's physiological signals such as blood volume pulse (BVP) providing HRV, galvanic skin response (GSR), i.e. EDA, and temperature of the skin, to provide a system for detecting stress. In [10], a classification method to determine stress on GSR and speech was proposed. In our work we are focusing on signals that can be acquired in a seamless manned in everyday life. Therefore, we are not considering pupil dilatation as a potential physiological measure for our system, even though we admit that it is a useful indicator of stress. Though, we might consider speech recognition as a potential extension of our system in the future.

The wide variety of classification algorithms have been applied to tackle the classification problem. In [2], signal processing techniques were applied to the physiological signals monitored to extract characteristics used by various learning algorithms: Naive Bayes, decision trees, and SVM to classify relaxed states (non-stress) compared to stressed states (stress). In [10], the decision tree, K-means clustering, and support vector machine (SVM) classifiers were proposed. In [6] a

KNN classifier was used to predict stress, from the body temperature, GSR and RR interval. The signals were collected to detect mental stress generated by the subject solving the Tower of Hanoi puzzle. This work [21] used logistic regression to predict the probability of stress state. In [11], the authors use a deep learning model with 7 hidden layers to predict stress state using EEG signal. It is common to classify stress with a binary class as in [8], with an RNN algorithm detecting stress from a voice signal.

The most relevant to our approach is the method of WESAD experiment [17], during which a multi modal data set was collected for stress classification and tested by several algorithms based on physiological data. Data collection was carried out in the laboratory. A binary definition of stress (stress, non-stress), as well as a three-class definition (baseline, stress, and amusement), are tested. However, all the tested algorithms are based on single sequence inputs, such as decision trees, kNN, or AdaBoost.

3 Data Collection

In this section, we describe how the DESY dataset used for the detection and classification of stress was collected.

The signals were acquired from 6 students of our university who have signed the consent form. The study protocol (see Sect. 3.1) was approved by the ethics committee on human resource (CER-VD) [1]. Exclusion criteria, stated in the study invitation, were pregnancy or lactation, major psycho-neuro-endocrinological or cardiac diseases and mental disorders, as well as participants having insufficient knowledge of the project language. All selected subject wore the Empatica E4 bracelet on their non dominant hand and the E4 records BVP (64 Hz), EDA (4 Hz), TEMP (4 Hz), and ACC (32 Hz) were recorded during the whole study. For more details about collected signals see Sect. 4. All the collected data were carefully anonymized.

3.1 Study Protocol

The goal of this experiment was to record physiological signals that will have the least possible movement induced artifacts often corrupting the physiological data collected using wearable technologies. Therefore, this experiment was performed in the laboratory conditions, while the participants were asked to make as little as possible movement with the hand with the bracelet to avoid as much as possible the movement artifacts. As the possible sources of stress we have selected the *emotional arousal*, *intellectual efforts*, *physical exercises*, and *pain*.

In order to allow each participant to come to his/her baseline condition the experiment was started by filling in a small questionnaire that was used further to define the subjective level of the current health and stress conditions. The participants were asked to answer the following questions using the 5 grade-scale:

- In general, my health is... - *[Excellent (5) .. Poor (1)]*
- I feel energetic... - *[All of the time (5) .. Almost never (1)]*
- Personality: I often stress when unexpected and difficult situations arise - *[Strongly agree (5) .. Strongly disagree (1)]*
- Daily activities: I stressed a lot in the past 24 h - *[Strongly agree (5) .. Strongly disagree (1)]*
- Sleep quality (1): I had trouble sleeping and had many sleep disturbances last night - *[Strongly agree (5) .. Strongly disagree (1)]*
- Sleep quality (2): I did not sleep in the past 24 h - *[Strongly agree (5) .. Strongly disagree (1)]*
- Sleep quality (3): I had trouble sleeping and had many sleep disturbances in the past month - *[Strongly agree (5) .. Strongly disagree (1)]*
- Right now, I fell... - *[Relaxed (5) .. Stressed (1)]*

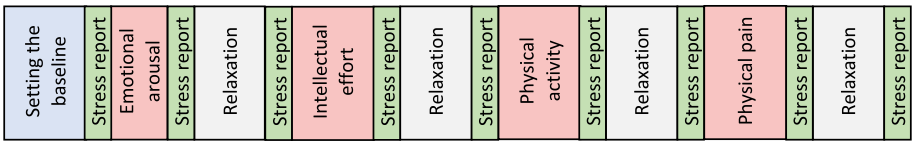


Fig. 1. An example of sequence of stressful and relaxing events with questionnaires. Note that each participant had his own order of stressful events.

To emulate each of the sources of stress each participant was asked to perform different activities. This way:

1. *Emotional arousal* was stimulated by showing a scary video during about 3 min;
2. *Intellectual efforts* was done by solving some riddles that were chosen by each participant randomly from a bunch of riddles printed on a paper (2–3 min);
3. *Physical activity* was represented by series of squats (2–3 min);
4. *Pain* was emulated by letting the participant to put his hand in the icy water for 1–2 min.

Each volunteer was participating in the above described studies with the randomized order of stress test to avoid influence of the order of stressful events on the results of classification. Each stressful period was followed by minimum 7 min of relaxation with some peaceful music and videos of the nature.

After each period (stressful or relaxed) of the experiment, each participate was asked to report their perceived stress level regarding the just finished activity describing it as either of the following: 1) I feel more relaxed, 2) No difference, 3) I feel less relaxed, and 4) I feel more stressed. An example of the sequence of stressful and relaxing events is shown on Fig. 1.

4 Methods

The general architecture of the dataflow in our data processing chain is presented on Fig. 2. It starts with the raw physiological signals collected from bracelet sensors (1), which are further sent to the mobile application (2). Next, the data flow is securely sent to the RedCap platform (3), frequently used and recommended for managing medical data. The data are stored in the cloud for further extraction of various features (4) using our proprietary signal processing and feature extraction algorithms. After signal processing, the features are sent back sent to the REDCap cloud. Further, the data are picked up by a classification algorithm (5) capable of predicting the stressful events (6).

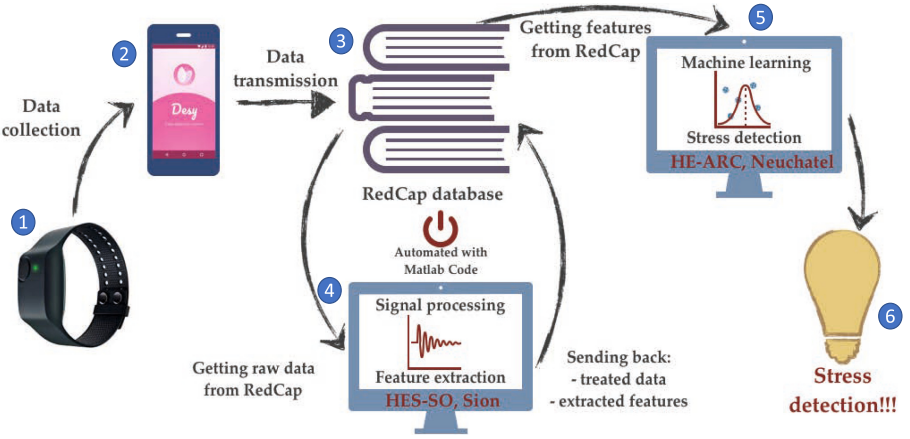


Fig. 2. The general architecture of the dataflow.

4.1 Wearable Sensor

The Empatica E4 bracelet[1], the device that was used for this work, offers the acquisition of physiological signals in real-time. The company has made available the Empatica Connect platform[2], which allows to visualise the graphs corresponding to the different signals. The bracelet works in two modes: (a) *streaming mode*: the bracelet connects via Bluetooth with the mobile application, and (b) *recording mode*: the wristband records the data in the internal memory, while it can record up to 60 h. The Empatica E4 bracelet is equipped with the following physiological sensors:

– EDA Sensor (or GSR Sensor): The skin is the only tissue of human body that is innervated by only SNS branches of the ANS and not by PNS branch fibres.

[1] https://www.empatica.com/en-eu/research/e4.
[2] https://www.empatica.com/connect.

Activation of the SNS provokes activation of sweat gland and thereby reducing skin electrical resistance and increasing conductance, whose fluctuating changes are measured by the EDA sensor in μSiemens. Consists of a tonic (referred to as skin conductance level (SCL)) and a phasic (skin conductance response (SCR)) component.
- PPG Sensor (Photoplethysmography Sensor) measures the blood volume pulse (BVP) from which two important signals can be derived: (1) heart rate (HR), and the inter-beat-interval (IBI). The blood volume pulse is measured in nanoWatts, heart rate HR in beats per minute (bpm), while IBI is measured in time periods between two consecutive beats.
- Infrared Thermopile: measures the temperature of the skin and contains the data measured in celsius degrees.
- 3-axis Accelerometer (ACC): measures activity based on motion, contains the data of the 3-axis channels (x, y, and z) accelerometer sensor. It measures continuous gravitational force (g) applied to each of the three spacial dimensions.

As was already mentioned earlier, in our study, we have used the signals acquired only with EDA and PPG sensors. We believe that skin temperature is greatly influenced by the temperature of the environment and therefore without knowing the real environmental conditions it would be difficult to receive a meaningful informations out of that particular signals. ACC signal, in contrast, is very useful, especially for classification of different types of stress, in particular differentiation between physical activity, with intensive movements or pain, with abrupt movement, and emotional/intellectual stress, with minimal movements. However, in this work we aim at binary classification, and therefore, the ACC signal is not used in the current study. Once we extend the use of our prediction model to the 5-class stress identification, this signal will be used.

4.2 Mobile Application

In order to perform experiments we have developed our proprietary mobile application for Android mobile platform with a user-friendly frontend and a backend performing three basic features, such as:

1. Data collection from the Empatica E4 bracelet and its temporal storage at the smartphone;
2. Questionnaire, allowing to collect the perceived level of stress by each participant;
3. Secure transmission of the recorded data into the RedCap database.

As temporary storage before sending data to cloud REDCap database at the end of each experiment an intern database (SQlite) was used. Once the experiment was over, all the signals were converted into the comma-separated values (CSV) format and were sent to the cloud for further processing.

4.3 REDCap Database

There exist several secure solutions to support human health data collection and storage. REDCap[3] is one of those with the further advantage of being available for free for research purposes. REDCap is a secure web application for building and managing online surveys and databases. REDCap provides multiple useful features, including secure mailing facilities supporting exchange of big data among researcher participants, as well as a built-in project calendar, a scheduling module, and ad hoc reporting tools. One feature of interest is the REDCap Mobile App interface that allows collecting data offline, for example, by a mobile device when there is no Wi-Fi or cellular connection, and then, later, sync data back to the server.

The logical portions of data in REDCap are grouped as 'Instruments'. The instruments of the DESY database in REDCap can be classified into communications with participants of the experiment (e.g. consent form), and 'instruments folders' containing raw signals, features extracted from the raw signals, and questionnaires, providing ground-truth information.

4.4 Signal Processing

Signal processing was automatized, such that in one click all the available signal processing techniques are performed on the raw data following the steps:

1. Getting the raw data from the REDCap;
2. Signal processing, analysis and restoration;
3. Features extraction;
4. Sending processed signals and features to REDCap.

It is quite often that the signals recorded by Empatica E4 bracelet are incomplete, such that some data are lost. This usually happens if the signal quality was not good enough, for example, due to the weak connection of the PPG sensor with the wrist. However, since feature extraction is done by small portions from a part of signal selected by a window of as parameterized size sliding over the signal with a regular step, it is crucial to have a complete signal. Therefore, we have developed several methods for restoration of lost data based on another available signal.

After analyzing the collected data, we have discovered that the most corrupted physiological signal among those collected from E4 is the BVP signal of the PPG sensor. However, it is rare that the data are lost from more than 20 s, especially when the experiments are performed in laboratory setting with minimum movements during the experiment. Therefore, first of all we have implemented an algorithm for HR signal extraction from BVP E4 signal by using a simple Fast Fourier Transform (FFT) for a sliding window size (i.e. 30 s, 1 min etc.) of the signal with a variable step that can be chosen according to the need, e.g. 1 s, 5 s, 30 s, etc.

[3] https://www.project-redcap.org/.

Apart from the raw BVP signal, Empatica E4 bracelet provides the IBI signal, the derivative of the BVP, representing the series of time interval between individual beats of the participant heart, largely used in the HRV analysis (see Sect. 4.5). Since IBI signal is build directly from the BVP signal that has missing data, it also has missing data in the same time periods. However, the reconstruction of IBI signal is simpler than of the raw BVP. Therefore, we have developed another algorithm that is reconstructing the missing parts of the IBI signal from the HR signal extracted using our first algorithm. Such a reconstruction cannot result in an ideal signal. Therefore, we also create a vector of quality of each value of the IBI time series with three quality levels, where level 0 corresponds to an optimal quality (the values provided by Empatica E4), level 2 - are the values reconstructed from the HR, and level 3 - values that have no meaning. Currently, this vector is not used but in our future work it is planned to be used by the classification algorithm to weight the credibility of the data.

4.5 Features Extracted

The main signals used for classification were the IBI and EDA. This section summarizes the features extracted from these two signals.

HRV Analysis: IBI signal analysis is often also called HRV analysis, and it is the study of variations in the instantaneous heart rate time series using the beat-to-beat RR-intervals (the RR tachogram, not to confuse with Respiratory Rate (RR)). There exist three main approaches to HRV analysis: 1) time-domain based, 2) frequency domain based, and 3) geometrical methods. The HR may be increased due to activation of the SNS or decreased due to PSNS (vagal) activity. While, in opposite, the variability of HR is decreasing with the activation of the SNS and increasing with PSNS, leading to the decrease (for SNS) and increase (for PSNS) of standard deviation (STD) of RR-intervals. The balance between the effects of SNS and PSNS, is called sympathovagal balance and is believed to be reflected in the beat-to-beat changes of the cardiac cycle. The time domain features (mostly various calculations of STD of RR-intervals) used in our study are presented in Table 1. While the frequency domain and geometrical domain features are presented in Table 2 and 3, respectively.

Table 1. Time domain features

Feature	Formula
Standard deviation	$SDNN = \frac{1}{N-1}\sqrt{(\sum_{i=1}^{N}(RR_i - \overline{RR})^2)}$
Coefficient of variation	$CV = \frac{SDNN}{\overline{RR}}$
Standard deviation of the average RR interval	$SDSD = \frac{1}{N-1}\sqrt{(\sum_{i=1}^{N-1}(\Delta RR_i - \overline{\Delta RR})^2)}$ $\Delta RR_i = RR_{i+1} - RR_i$
Mean difference of successive NN intervals	$RMSSD = \frac{1}{N-1}\sqrt{(\sum_{i=1}^{N-1}(RR_{i+1} - RR_i)^2)}$
Number of RR intervals	$NN50$
Vagus activity	$pNN50 = \frac{NN50}{(N-1)}$

Table 2. Frequency domain features

Feature	Formula	Meaning
The power of the complete signal	TP	Associated with the hypothalamic-pituitary complex activity
High frequency 0.15–0.4 Hz	HF	Associated with breathing arrhythmia and PSNS activity. Subject to circadian rhythms 24 h signal analyzed
Normalized HF	$nHF = HF/TP$	
Low frequency 0.04–0.15 Hz	LW	Slow waves of first order, SNS activity. Subject to circadian rhythms if 24 h signal analyzed
Normalized LF	$nLF = LF/TP$	Grows with SNS activation, since TP goes down and LF does not change
Index of vagosympathetic cooperation	LF/HF	
Very low frequency 0.015–0.04 Hz	VLF	Psycho-emotional tension
Ultra low frequency <0.003 Hz	ULF	Is measured only for long term signals (≈24 h). Subject to circadian rhythm. Therefore, it was not used in our study
Index of centralization	$IC = (VLF + LF)/HF$	

Table 3. Geometrical domain features

Feature	Formula
Mode of the histogram	Mo
Amplitude of the histogram	AMo
Width of the histogram	$Delta_X(TINN)$
Width normalized value	$Dealta_X/RR$
Index of SNS activity	$SNS_{ind} = AMo/(2 * Mo * delta_X)$
Index of PSNS activity	$PSNS_{ind} = 1/(Mo * delta_X)$

Table 4. EDA time domain features

Description	Feature
SCR amplitude peak counts	$EDA_{peakCount}$
The minimum value found in the section	EDA_{MIN}
The maximum value in the section	EDA_{MAX}
Area under curve	EDA_{AUC}
Mean of first order derivatives	$EDA_{MEAN\_derivative}$
Mean of negative values of first order derivatives	$EDA_{MEAN\_negative\_derivative}$
Hjorth features [19]	$EDA_{complexity}$

EDA Analysis. The overall signal called EDA of electrodermal activity consists of two components. One of the components is the EDA general tonic level which relates to an overall signal level, the most common measure of this component is the SCL and the changes in the SCL are believed to reflect the general changes in autonomic arousal. The value of SCL can vary widely, typically between 2–20 μS, due to environmental and personal factors [5]. The second component is the phasic component and this refers to the fast response variations of the signal in the form of peaks, i.e. the SCRs, and appears either in response to a stimulus or without evident stimulation. These instantaneous peaks can be characterized by a rise time, amplitude and a half recovery time. In healthy adults, the rise time is usually between 1 and 3 s, the amplitude often varies (a minimum is usually between 0.01 and 0.05 μS), and the half recovery time is usually between 2 and 10 s [5]. The example of an EDA signal annotated with stress stimuli is presented on Fig. 3.

From these signals we can extract several features in the time and frequency domain. For our study we have chosen the most relevant ones, that are presented in Tables 4 and 5.

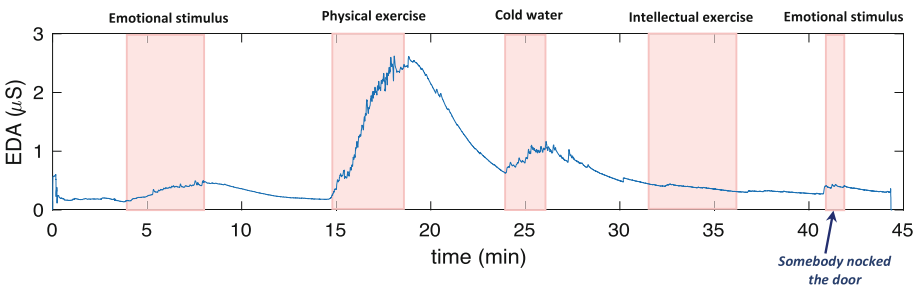


Fig. 3. The EDA signal with four stressful activities and an unexpected stressful even related to the knocking on the door.

Table 5. EDA frequency domain features

Description	Formula
Energy of the signal	EDA_{Signal_Energy}
Summation of FFT harmonics	$EDA_{harmonics\_summation}$
Area under curve of FFT	EDA_{AUC_fft}
Standard deviation of FFT	EDA_{STD_fft}
Mean of FFT	EDA_{MEAN_fft}
Signal values in the frequency domain	$EDA_{coefficients}$

4.6 Machine Learning

Stress can be detected and predicted by machine learning methods with classification or regression models. In the DESY dataset, stress and its predictors are represented as a time series.

For the purpose of binary classification, we decided to compare 3 different methods. First, a decision tree models based on a summarized time window is presented. Second, a recurrent neural network (RNN) capable of handling multiple time windows is tested. And third, an augmented RNN with some convolutional layers first (CRNNs) is tested for a more in-depth extraction of features.

Architecture and Learning Process. The DESY dataset consists of 6 patients, each with a duration of about 44 min. Due to the nature of time series and to the need of a stratified split, we used 4 patients for the training set and 2 patients for the test set, resulting in 28% for the test set, cross-validated (K = 3).

The stress label, as described in Sect. 3, is filled in by participants at the end of each period. All the values in between these periods are linearly interpolated. The decision tree is augmented with gradient boosting and implemented with the CatBoost library. The prediction of a single time window is performed with a maximum depth of 6. The RNN consists of a single layer of LSTM cells, some batch normalization, and a dense layer for the classification task (see Table 6).

Table 6. Architecture of the RNN. 'None' indicates the batch size (set to 256).

Layer	Output shape	# Parameters
LSTM	(None, 64)	25'088
Batch normalization	(None, 64)	256
Dense (sigmoid)	(None, 1)	65
Total parameters: 25'409	Trainable parameters: 25'281	Non-trainable parameters: 128

The CRNN consists of a single layer of the same RNN preceded by 3 convolutional layers (see Table 7). All the methods use the same set features, however with different window strategies.

Table 7. Architecture of the CRNN. 'None' indicates the batch size (set to 256).

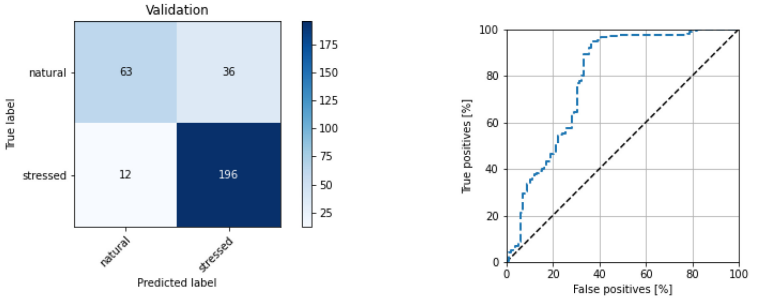
Layer	Output shape	# Parameters
1D convolution	(None, 10, 8)	4608
Batch normalization	(None, 10, 8)	32
ReLU	(None, 10, 8)	0
Max pooling	(None, 5, 8)	0
1D convolution	(None, 5, 16)	1536
Batch normalization	(None, 5, 16)	64
ReLU	(None, 5, 16)	0
Max pooling	(None, 2, 16)	0
1D convolution	(None, 2, 32)	3072
Batch normalization	(None, 2, 32)	128
ReLU	(None, 2, 32)	0
Max pooling	(None, 1, 32)	0
LSTM	(None, 64)	24'832
Dense	(None, 256)	16'640
Batch normalization	(None, 256)	1024
Activation	(None, 256)	0
Dense	(None, 32)	8224
Batch normalization	(None, 32)	128
Activation	(None, 32)	0
Dense	(None, 1)	33
Total parameters: 60'321	Trainable parameters: 59'633	Non-trainable parameters: 688

Experimental Results. The overall results comparing the three different approaches are presented in Table 8. The performance of the best classifier is presented in Fig. 4. The threshold of the classifiers is selected according to the Youden's J statistic.

Furthermore, the impact of the phasic and tonic parts of the EDA signal is investigated by their ablation (Table 9). That is the best model is trained and tested without the presence of their related features.

Table 8. Evaluation of the different machine learning models and pre-processing parameters

	Gradient boosting DT				RNN				CRNN			
window size [sec]	30	60	120	180	30	60	120	180	30	60	120	180
step size [sec] 15	15	15	15	15	15	15	15	15	15	15	15	15
total # of windows	889	878	854	830	889	878	854	830	889	878	854	830
postive classes [%]	43	43	44	46	43	43	44	46	43	43	44	46
# time steps	N/A	N/A	N/A	N/A	10	10	10	10	10	10	10	10
AUC	0.63	0.62	0.61	0.62	0.64	0.65	0.71	0.73	0.62	0.63	0.59	0.61
weighted F_1	0.57	0.58	0.56	0.58	0.65	0.61	0.72	0.77	0.67	0.69	0.65	0.68
macro F_1	0.58	0.58	0.55	0.58	0.62	0.58	**0.67**	**0.71**	**0.63**	**0.65**	0.60	0.63
3-fold std macro F_1	0.08	0.04	0.05	0.04	0.02	0.04	0.06	0.05	0.05	0.02	0.04	0.07
	Random baseline macro F_1: 0.48											



(a) Confusion matrix (b) ROC curve (AUC=0.716)

Fig. 4. Classification metrics of the best RNN model

Table 9. Ablation results for the EDA signal with the RNN model of the best macro F_1

	All features except tonic EDA	All features except phasic EDA	All features
AUC	0.65	0.71	0.73
weighted F_1	0.67	0.75	0.77
macro F_1	0.62	0.68	0.71
difference	**−13%**	**−5%**	

5 Discussion

The best model for the binary classification of stress is achieved with a recurrent neural network, and yields a macro F_1 of 71%. In our tests, the lowest score is of 55% (because of the imbalanced nature of the dataset, the random baseline is of 48%). The AUC of the ROC curve of the best model is of 0.716, meaning that if the focus were to detect stress with a better recall, an accuracy of almost

100% could be achieved, traded off by almost 50% of false positives). We did not include other metrics (such as accuracy) because of the imbalanced nature of our dataset.

In almost every situation, RNN and CRNN outperforms decision trees. This result confirms the general idea that time series are better handled by deep learning architectures and more precisely recurrent or convolutional networks, thanks to their capacity to handle sequences of time. However, the hyper-parameters for building the time sequences may highly impact the score (from 0.58 to 0.71 in the same RNN). We were unfortunately not able to acquire more data at this stage of the project, since experimental part was planned for the beginning of March 2020, when global confinement due to COVID-19 has started. Nevertheless, even though we had only six participants, for each of them we have recorded signals, time-series, of about 35 minutes long. The feature extraction algorithms processed these time-series with window size of 1 min and step of half a minutes, thus providing a time-series of training point of 70 for each of six participants. Despite the small size of our training set, the received results are promising, especially considering a deep learning architecture.

As mentioned in Sect. 4, EDA contains information not only related to slow changes, that is the tonic component, but also in the rapid or phasic changes of the signal. We observed that the prediction of stress is strongly based rather on the tonic component, with a drop of 13% on the F_1 score with its ablation.

As future work, globally, we aim at developing a wearable system allowing for seamless monitoring and detection of critical signatures of stress leading to distress. To achieve this goal we still have a long road. First of all, to improve the quality of stress prediction, we intend to continue our project towards implementing a more personalized prediction, since the values contained in physiological signals are specific for each participant. For example, as was already mentioned, the value of SCL can vary widely, typically between 2–20 μS. Therefore, such factors as physical constitution as well as the baseline level of stress of participant must be taken into account. Further, we would like to take into account the ACC signal and provides a five-class definition of stress, differentiating between emotional, intellectual, and physical stress, as well as pain, in contrast to the non-stress conditions.

Once we will go from laboratory setting to everyday life our dataset will include the contextual data, as well as measurement of cortisol. Measurements of cortisol are quite intrusive. However, since some studies have presented already that slow arousal of the morning cortisol level serves as the indication of "burnout" state [13], its measurement performed with participants of the long experiments will allows us defining the signature of stress events leading to the 'burnout' or distress state. Finally, other machine learning algorithms can be implemented allowing to choose the best ones among them, in order to improve stress detection and monitoring.

6 Conclusion

Stress causes biochemical, physiological and behavioral changes, and can be described as an uncomfortable emotion. The long-term exposure to stress can cause illness. In this paper we have implemented a prediction stress detection system from a wrist-device sensor providing stress relevant physiological signals. We have implemented three classification algorithms providing two-class classification Stress vs Non-Stress. A reasonable prediction can be observed when we apply a recurrent neural network, this model yiels a macro F_1 of 71%. Our work will not stop here, and as described in Sect. 5 we have several perspectives to improve the system.

References

1. La commission cantonale (VD) d'éthique de la recherche sur l'être humain. http://cer-vd.ch/
2. Barreto, A., Zhai, J., Adjouadi, M.: Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In: Lew, M., Sebe, N., Huang, T.S., Bakker, E.M. (eds.) HCI 2007. LNCS, vol. 4796, pp. 29–38. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75773-3_4
3. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Pentland, A.: Daily stress recognition from mobile phone data, weather conditions and individual traits. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 477–486 (2014)
4. Boucsein, W.: Electrodermal indices of emotion and stress. In: Electrodermal Activity, pp. 369–391 (1992). Chapter 3
5. Boucsein, W., et al.: Society for psychophysiological research ad hoc committee on electrodermal measures. Publication recommendations for electrodermal measurements. Psychophysiology **49**(8), 1017–1034 (2012)
6. Ciabattoni, L., Ferracuti, F., Longhi, S., Pepa, L., Romeo, L., Verdini, F.: Real-time mental stress detection based on smartwatch. In: 2017 IEEE International Conference on Consumer Electronics (ICCE), pp. 110–111. IEEE (2017)
7. Doberenz, S., Roth, W.T., Wollburg, E., Maslowski, N.I., Kim, S.: Methodological considerations in ambulatory skin conductance monitoring. Int. J. Psychophysiol. **80**(2), 87–95 (2011)
8. Han, H., Byun, K., Kang, H.G.: A deep learning-based stress detection algorithm with speech signal. In: Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, pp. 11–15 (2018)
9. Hernando, A., et al.: Inclusion of respiratory frequency information in heart rate variability analysis for stress assessment. IEEE J. Biomed. Health Inform. **20**(4), 1016–1025 (2016)
10. Kurniawan, H., Maslov, A.V., Pechenizkiy, M.: Stress detection from speech and galvanic skin response signals. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pp. 209–214. IEEE (2013)
11. Liao, C.Y., Chen, R.C., Tai, S.K.: Emotion stress detection using EEG signal and deep learning technologies. In: 2018 IEEE International Conference on Applied System Invention (ICASI), pp. 90–93. IEEE (2018)

12. LiKamWa, R., Liu, Y., Lane, N.D., Zhong, L.: MoodScope: building a mood sensor from smartphone usage patterns. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, pp. 389–402 (2013)
13. Marchand, A., Juster, R.P., Durand, P., Lupien, S.J.: Burnout symptom sub-types and cortisol profiles: what's burning most? Psychoneuroendocrinology **40**, 27–36 (2014)
14. Posada-Quintero, H.F., Florian, J.P., Orjuela-Cañón, A.D., Chon, K.H.: Electrodermal activity is sensitive to cognitive stress under water. Front. Physiol. **8**, 1128 (2018)
15. Salahuddin, L., Jeong, M.G., Kim, D., Lim, S.K., Won, K., Woo, J.M.: Dependence of heart rate variability on stress factors of stress response inventory. In: 2007 9th International Conference on e-Health Networking, Application and Services, pp. 236–239. IEEE (2007)
16. Sano, A., Picard, R.W.: Stress recognition using wearable sensors and mobile phones. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 671–676. IEEE (2013)
17. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 400–408 (2018)
18. Schneiderman, N., Ironson, G., Siegel, S.D.: Stress and health: psychological, behavioral, and biological determinants. Annu. Rev. Clin. Psychol. **1**, 607–628 (2005)
19. Shukla, J., Barreda-Angeles, M., Oliver, J., Nandi, G., Puig, D.: Feature extraction and selection for emotion recognition from electrodermal activity. IEEE Trans. Affect. Comput. (2019)
20. Stalder, T., et al.: Assessment of the cortisol awakening response: expert consensus guidelines. Psychoneuroendocrinology **63**, 414–432 (2016)
21. Zubair, M., Yoon, C., Kim, H., Kim, J., Kim, J.: Smart wearable band for stress detection. In: 2015 5th International Conference on IT Convergence and Security (ICITCS), pp. 1–4. IEEE (2015)