Into the Wild: The Challenges of Physiological Stress Detection in Laboratory and Ambulatory Settings

E. Smets, W. De Raedt, and C. Van Hoof

Abstract —**Stress and mental health have become major concerns worldwide. Research has already extensively investigated physiological signals as quantitative and continuous markers of stress. In recent years the focus of the field has shifted from the laboratory to the ambulatory environment. We provide an overview of physiological stress detection in laboratory settings with a focus on identifying physiological sensing priorities, including electrocardiogram, skin conductance and electromyogram, and the most suitable machine learning techniques, of which the choice depends on the context of the application. Additionally, an overview is given of new challenges ahead to move towards the ambulant environment, including the influence of physical activity, lower signal quality due to motion artifacts, the lack of a stress reference and the subject-dependent nature of the physiological stress response. Finally, several recommendations for future research are listed, focusing on large-scale, longitudinal trials across different population groups and just-in-time interventions to move towards disease prevention and interception.**

Index Terms—**Physiology, stress, electrocardiogram, skin conductance, machine learning**

I. Introduction

In the 21st century, stress and mental health at work have become major concerns for organizations worldwide [1]. The American Psychological Association states that in 2015 in the US 24% of adults reported extreme stress [2]. In 2013 in Europe 51% of the working population reports that cases of work-related stress are common in their workplaces, with most important causes of stress being job reorganization or job insecurity and hours worked or workload [3].

Research has already extensively discussed the negative consequences of stress [4, 5, 6, 7]. For example, observational data suggest an average 50% increased risk for coronary heart disease among persons with high stress [8]. Stress can also have a negative impact for companies. It has been shown that people perform worse under excessive stress [4]. Many studies have tried to estimate the cost of stress and while quantitative data is scarce, a report of the European Agency for Safety and Health at Work states that in 2002 the cost of work-related stress for Europe was estimated at €20 billion [9]. A more recent study in 2013 estimates the cost of work-related depression in Europe at €617 billion annually [10].

To tackle the problem of stress, the first step is to measure it. Currently the most widespread method and gold-standard to assess stress is by means of questionnaires, e.g. the Perceived Stress Scale [11]. However, these questionnaires are qualitative, time-consuming and reflect subjective responses collected during spot-checks. Therefore, research has focused on finding objective, continuous and quantitative physiological markers of stress [12], by exploiting the sympathetic nervous system's (SNS) fight-or-flight response [13].

Research focusing on the physiological detection of stress has been conducted under different types of conditions. Historically, most research has been conducted in laboratory settings, where both stressor (timing, frequency, duration) and context can be rigorously controlled [14]. With the increasing use of wearables, many opportunities are emerging for continuous, ambulatory monitoring of stress and research in this field has increased substantially over the last five years. In ambulatory monitoring two types of conditions are used, either context specific, e.g. stress monitoring while driving a car [15] or in a call center [16], or daily living settings in which different types of stressors can influence the subject and the context of the stressors is unknown.

In controlled settings the potential of physiological signals to detect psychological stress has already widely been demonstrated. Several stress inducement stimuli have been used, e.g. the Stroop color word test, mental arithmetic, public speaking, computer work or a cold pressor test [17]. Also multiple physiological signals, features and machine learning techniques have been investigated. In ambulatory settings however, research on physiological stress detection is still very recent and few large-scale and long-term studies exist to confirm the potential of these signals in real-life situations. Additionally, several new challenges, such as the presence of physical activity and technical limitations of wireless wearable devices, arise when moving out of the lab and into ambulatory settings.

This review aims to provide an overview of past research on physiological stress detection in laboratory settings with a focus on identifying physiological sensing priorities and the most suitable machine learning techniques. Additionally, we aim to list the challenges ahead to move towards the ambulant environment with recommendations for future research.

The research was partly funded by a PhD grant of the Flanders Innovation & Entrepreneurship Agency (VLAIO).

E. Smets is with KU Leuven, Leuven, Belgium and imec, Leuven, Belgium (e-mail: elena.smets@imec.be). W. De Raedt is with imec, Leuven, Belgium (e-mail: walter.deraedt@imec.be). C. Van Hoof is with imec, Leuven, Belgium; KU Leuven, Leuven, Belgium and Holst Centre, Eindhoven, The Netherlands (e-mail: chris.vanhoof@imec.be).

II. MECHANISM OF THE PHYSIOLOGICAL STRESS RESPONSE

The human body responds to stress using two response systems: a fast response to sudden stress following the sympathetic adrenal medullary (SAM) axis and a slow response to chronic stress following the hypothalamic pituitary adrenal (HPA) axis [18].

The autonomic nervous system exists of three subsystems: the sympathetic nervous system (SNS), the parasympathetic nervous system (PNS) and the enteric nervous system. Most tissues are innervated by both SNS and PNS with opposing effects.

When the brain encounters a stressor, i.e. an internal or external stimulus that disrupts the body's internal balance, the paraventricular nucleus (PVN) of the hypothalamus will activate the SNS. The SNS in its turn will signal the adrenal medulla to secrete epinephrine and norepinephrine (SAM axis) [13]. This activation results in an increase of heart rate, blood pressure, pupil dilation, etc., which is sustained by the presence of epinephrine and norepinephrine. The body is preparing for a 'fight-or-flight' response. Simultaneously, the PVN will release two hormones: corticotropin-releasing factor (CRF) and arginine vasopressin (AVP) [19]. Both hormones are sent to the pituitary gland (hypophysis) through blood vessels, where they stimulate the production and secretion of adrenocorticotropic hormone (ACTH). ACTH in its turn induces the synthesis and release of glucocorticoids from the adrenal cortex (HPA axis). In humans the most important glucocorticoid is cortisol, which has a wide array of regulatory influences. It plays a key role in the central nervous system (CNS), where it is involved in learning, memory and emotion regulation; in the metabolic system, where it regulates the use and storage of glucose; and in the immune system, where it regulates the magnitude and duration of the inflammatory response [20]. Cortisol levels reach a peak in the blood about 30 minutes after acute stress exposure [21].

The stress response system is regulated by cortisol in a negative feedback loop at three levels: 1) the pituitary, where it reduces the release of ACTH, 2) the hypothalamus, where it reduces the activity of the PVN and 3) the hippocampus, which has a stimulating effect on the production of CRF in the absence of cortisol[13]. Through the reduced activity of the PVN both SAM and HPA axes are attenuated and levels of epinephrine will decrease. It is important to notice that the main driver of the fight-or-flight response is the SAM axis and presence of epinephrine. The HPA axis and presence of cortisol do not stimulate the stress response, but rather regulate it. Without cortisol no negative feedback loop would be possible and stress responses would have damaging effects on the body [13].

III. PHYSIOLOGICAL SENSING PRIORITIES

Multiple physiological signals have already been investigated in laboratory conditions and numerous features have been extracted. Below, the most commonly studied signals are discussed, together with an overview of extracted features and their importance (Table 1). We focused mainly on physiological signals that can be measured both in laboratory and ambulatory settings (e.g. electrocardiogram) and less on signals that have solely been used in laboratory settings (e.g. blood pressure). In Table 2 an overview of papers focusing on physiological stress detection is presented. It is a representative yet not exhaustive list of current research using physiological sensors for stress detection up to October, 2018.

TABLE I
COMMON FEATURES FOR PHYSIOLOGICAL STRESS DETECTION

Physiological signal	Feature	Explanation
ECG	Mean HR	Heart rate (mean)
	SDNN	Standard deviation of the R-R peaks
	RMSSD	Root mean square of the successive differences between R-R peaks
	pNN20	Percentage of successive normal sinus R-R intervals more than 20ms
	pNN50	Percentage of successive normal sinus R-R intervals more than 50ms
	HF	High frequency range of the R-R intervals (0.15 - 0.4Hz)
	LF	High frequency range of the R-R intervals (0.04 - 0.15Hz)
	LF/HF	Ratio of low versus high frequency ranges
	SD$_1$	Crosswise standard deviation of the Poincaré plot
	SD$_2$	Lengthwise standard deviation of the Poincaré plot
SC	SCL	Skin conductance level (mean)
	SCR duration	Skin conductance response duration
	SCR magnitude	Magnitude of the skin conductance response
	SCRR	Skin conductance response rate
	OPD	Ohmic Perturbation Duration: period during which a subject remains under stimulation effect and is measured from the start of the stimulus until the recovery initiates
EMG	RMS	Root mean square of the EMG signal
	Gaps	One or more segments (0.2-4 s) in a row with an RMS value below 5% of the RMS reference contraction
	Static load	10th percentile of rank-ordered RMS values
	Median load	50th percentile of rank-ordered RMS values
	Peak load	90th percentile of rank-ordered RMS values
BP	SBP	Systolic BP, pressure when the heart beats
	DBP	Diastolic BP, pressure between heart beats
Pupil	Pupillary dilation	Diameter of the pupil
Eye blinks	Frequency	Frequency of the eye blinks

skin temp ST	Mean	Mean ST
	SD	Standard deviation of the ST
	Slope	Slope of the ST
Respiration	MeanRsp	Mean respiration frequency
	EB1	Spectral power density, summing the energy in band 0-0.1Hz
	EB2	Spectral power density, summing the energy in band 0.1-0.2Hz
	EB3	Spectral power density, summing the energy in band 0.2-0.3Hz
	EB4	Spectral power density, summing the energy in band 0.3-0.4Hz
EEG	Delta	0.5-3.5 Hz frequency range of the EEG signal
	Theta	4-7.5 Hz frequency range of the EEG signal
	Alpha	8-13 Hz frequency range of the EEG signal
	Beta	14-32 Hz frequency range of the EEG signal
	SEn	Sample entropy

A. *Electrocardiogram*

The electrocardiogram (ECG) measures the electrical activity of the heart. A sample of a normal ECG signal is presented in Fig. 1. The P-wave is associated with the contraction of the atria. The QRS complex is associated with the contraction of the ventricles. The T/U waves are associated with the repolarization of the ventricles [22]. The R-R distance (or R-R interval) is the time between two R peaks and is used in the calculation of the heart rate (HR) and heart rate variability (HRV). ECG signals from different individuals can exhibit personalized traits such as the relative timing of the peaks but can also exhibit responses to stress and activity.

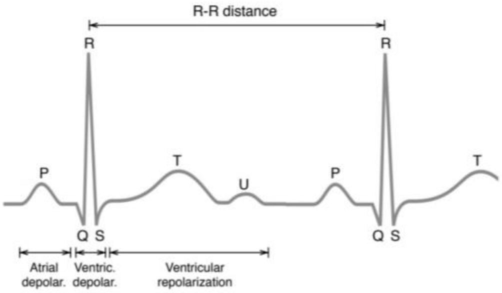

Fig. 1. Sample of a normal ECG [22].

In resting conditions the heart is under inhibitory control of the PNS. This results in a low HR and high HRV, since the PNS adapts the HR to the breathing phase (inspiration versus expiration), i.e. Respiratory Sinus Arrhythmia (RSA) [23]. In stressful situations, PNS control is decreased resulting in a disinhibition of the SNS and an increased SNS activation through the SAM pathway, which causes the HR to increase and the HRV to decrease since SNS modulation of HR reacts too slowly to respond to the respiratory phase.

Multiple features have been computed to evaluate the HRV. In general, these can be subdivided into time and frequency domain features and non-linear HRV features [24]. Time domain features, which decrease when stress increases, include mainly the standard deviation of the R-R peaks (SDNN), the root mean square of the successive differences (RMSSD) and percentage of successive normal sinus R-R intervals more than 20 (pNN20) and 50ms (pNN50) respectively. Frequency domain features focus on the difference between low frequency (LF, 0.04-0.15Hz) and high frequency (HF, 0.15-0.4Hz) signals [24], where LF would represent the slow responding SNS or stress and HF the fast responding PNS or rest [24]. The most common non-linear HRV features include the lengthwise and crosswise standard deviation of the Poincaré plot, which displays the correlation of R-R intervals by assigning each following interval to the former interval as a function value [24].

HR and HRV have already widely been investigated in stress-related research. Karthikeyan *et al.* [25] found a 94.58% classification accuracy for binary stress classification (i.e. rest versus stress during a Stroop Color Word Test) based on HRV, including both time and frequency domain features, and wavelet transforms of the ECG signal. The highest classification accuracies were reached based on the ratio of LF/HF features, the lowest performance was reached using RMSSD features. Klumbies *et al.* [27] showed an average 26.64% increase in HR and a significant decrease in RMSSD during the Trier Social Stress Test as compared with baseline, both for healthy subjects (n=78) and patients with social phobia (n=88). In the study of Gomes *et al.* [26] the three-level self-reported stress (i.e. low, medium and high stress, collected during emergency events) of fire fighters was compared with measures of HR and HRV. The highest correspondence with self-reported stress of 68% was reached for RMSSD, SDNN and HF features [26]. It would be expected that all values decrease as stress increases. However, findings showed that values first increased from low to medium stress and then decreased from medium to high stress [26].

In general, varying results have been reported related to the use of these time and frequency domain features. In the studies of McNames *et al.* [27] and Billman [28] it was suggested to focus on RMSSD and HF features, rather than others, including the LF/HF ratio, as these are less susceptible to noise and segment duration. Castaldo *et al.* [30] computed 18 HRV features including, time and frequency domain and non-linear features in 42 university students during an oral examination (stress) and after vacation (rest). Results showed significant differences comparing rest versus stress for 12 out of 18 features. Feature selection for machine learning showed that mainly non-linear features were important.

Additional research has investigated the influence of respiration on stress detection using HRV measures. Widjaja *et al.* [29] showed that by removing the influence of the RSA (i.e. the influence of respiration on HR and HRV), the classification accuracy for a binary stress detection could be improved from 57.13% to 97.88%. Additionally, Brugnera *et al.* [30] showed that a stronger cardiovascular response was exhibited when subjects performed a stress test that did not include speech, i.e.

neighbor (KNN) [25] [39]. Less frequently used approaches include Fisher projection and linear discriminant analyses, and neural networks.

In general, SVMs and neural networks are well-performing, but computationally intensive algorithms, especially if the input dataset has a large number of features or observations [78]. Additionally SVMs and neural networks are black box models, which do not allow insight in the feature structure, whereas BN and RFs do allow insight, but can easily be overfitted [78]. The choice of algorithm therefore depends on the application.

Several studies have compared the classification performance of different models for physiological stress detection. Han *et al*. [60] compared the classification accuracies of four models for a three-class stress classification: SVM, Linear Discriminant Analysis (LDA), Adaboost and Nearest Neighbors (KNN). The highest classification accuracy of 84% was reached for the SVM, followed by LDA with 80%, Adaboost with 79% and the lowest accuracy of 72% was reached for KNN [60]. Similar models were compared by Mozos *et al*. [58] where highest accuracies for a binary stress classification were reached for AdaBoost (94%) and SVM classifiers with a radial basis function kernel (93%). The lowest accuracy was reached for KNN (87%) and SVM classifiers with a linear kernel (85%) [58]. Sun *et al*. [36] compared the classification accuracies for a binary stress classification of SVM, BN and decision trees and found the highest accuracy of 92.4% for decision trees. Smets *et al*. [53] compared six classification methods for a binary stress detection problem: logistic regression, SVM, decision trees, RF and static and dynamic BN. Additionally personalized (i.e. including data of one subject) and generalized models (i.e. including data of all subjects) were compared. The highest accuracies were found for personalized dynamic BN (84.6%) and generalized SVM (82.7%) [53]. Castaldo et al. [30] compared five classification techniques for a binary stress detection using HRV features, including Naive Bayes, SVM, multilayer perceptron, AdaBoost and decision trees. The highest accuracies and area under the curve (AUC) were found for the AdaBoost (accuracy = 80%, AUC = 79%) and decision trees (accuracy = 79%, AUC = 83%) methods

In conclusion, multiple ML techniques have been successfully used for physiological stress detection. Currently, there is not one technique that outperforms all others. Therefore, the choice should depend on the application and needs to be tailored to the meet the requirements of each study. Additionally, the size and scale of studies thus far is limited, and more research is needed to determine the best analysis methods in different contexts.

TABLE II
CURRENT RESEARCH ON PHYSIOLOGICAL STRESS DETECTION

Ref.	Laboratory (L) Ambulant (A) Semi-ambulant (SA)	Nr. Of participants	Physiological signals	Analysis technique	Nr of classification levels	Classification performance
[15]	SA (driving on a set route)	9	ECG, SC, EMG, respiration	Fisher projection and linear discriminant	3	97% (accuracy)
[25]	L	10	ECG	KNN	2	94.58% (accuracy)
[26]	A	4	ECG	Correlation	3	68% (correlation)
[29]	L	40	ECG, respiration	LS-SVM	2	97.88% (accuracy)
[35]	SA (controlled driving task)	100	SC	Chi-square test	2	-
[36]	L	20	ECG, SC	Decision Tree BN SVM	2	92.4% (accuracy, decision tree)
[39]	L	10	EMG	KNN	2	90.70% (accuracy)
[40]	L	30	EMG	Friedman test Wilcoxon signed rank test	2	-
[41]	L	23	ECG,EMG, CO_2	Friedman test	2	-
[42]	L	17	ECG, EMG	t-test	2	
[50]	L	65	BP	General linear models	Continuous	
[52]	L	20	BVP, SC, respiration, ST	t-test	2	
[53]	L	20	ECG, SC, respiration, ST	Logistic regression SVM Decision tree Random forest BN (static and dynamic)	2	84.6% (balanced accuracy, dynamic BN
[54]	L	40	ECG, SC, EMG, ST	KNN Probabilistic NN	2	93.75% (accuracy)
[55]	L	32	BVP, SC, pupil diameter, ST	SVM	2	90.10% (accuracy)
[58]	L	18	BVP, SC, sociometric badge (e.g. body movement)	SVM AdaBoost KNN	2	94% (accuracy, AdaBoost)

[60]	L	39	ECG, respiration	SVM Linear discriminant analysis AdaBoost KNN	Binary vs. three-class	84% (accuracy, SVM, three-class) 94% (accuracy, SVM, binary)
[61]	SA (driving on a set route)	13	ECG, SC, respiration, driver events (e.g. GPS)	BN	2	96% (accuracy)
[62]	L + A	21 (Lab) 17 (Ambulant)	ECG, respiration	Decision tree AdaBoost SVM	2	90% (accuracy, lab, AdaBoost) 0.72 (correlation with self-reported stress, ambulant)
[63]	L + A	21 (Lab) 26 (Ambulant)	ECG, respiration	SVM (lab) BN (ambulant)	2	95.3% (median accuracy, lab) 72% (accuracy, ambulant)
[64]	A	10	SC, ST		5	
[65]	A	10	ECG, SC, respiration	Random forest Lasso SVR KNN	continuous	1.5 (mean squared error, SVR)
[66]	A	35	ECG, smartphone (audio, social, physical)	Logistic regression	3	61% (accuracy)
[68]	L	15	ECG, SC, respiration	Nearest mean classifier (NMC) Multilayer perceptron (MP) KNN	3	92% (accuracy, NMC)
[30]	L	42	ECG	Naive Bayes SVM MP AdaBoost Decision tree	2	79% (accuracy, decision tree)

KNN = K-nearest neighbor, LS-SVM = Least squares support vector machines, BN = Bayesian network, BVP = blood volume pulse, NN = Neural network, SVR = Support vector regression, NMC = Nearest mean classifier, MP = Multilayer perceptron

V. AMBULATORY STRESS DETECTION

In recent years, the focus of physiological stress detection has shifted towards the ambulatory environment. The goal is to develop a continuous, non-intrusive stress detection in real-life conditions. The research of Healey and Picard [15] in 2005 was one of the first studies to leave the laboratory and do measurements in a real-world driving task. Although participants (n=9) still had to drive on a set route, this was a first step towards physiological stress detection in uncontrolled conditions. They reached a classification accuracy of 97% in a 5 min interval based on HR, SC, EMG and respiration. In a similar context, Rigas *et al*. [61] measured ECG, SC and respiration to detect stressful driver events in real driving conditions with 13 participants (50 min driving per person). Using a Bayesian Network classifier a classification accuracy of 82% was reached, which was improved to 96% by including driving event information such as GPS data.

In other application domains, Gomes *et al*. [26] found a 68% association of HRV with self-perception of stress for fire fighters in real emergency situations. Ertin *et al*. [62] compared the performance of a newly developed sensor suite called ‘AutoSense’, which measures ECG, SC and respiration, in a laboratory and ambulant environment with 21 subjects. On lab data the classifier reached an accuracy of 90%, whereas on ambulatory data a median correlation of 0.72 with self-reported stress ratings was achieved.

These results are promising towards ambulatory stress detection. However, there are several new challenges as compared to laboratory research that still need to be addressed. The most important challenges are discussed.

A. Physical activity

The presence of physical activity can mask the effect of stress on the physiological signals, both because of sensor inaccuracies due to movement and because of physiological changes that occur due to physical activity [36]. The most common solution is to remove the data with high activity from the dataset. Hovsepian *et al*. [63] used a 3-axis on-body accelerometer to identify low, moderate and high activity levels in 10s windows. If the majority of the 10s windows within a minute was classified as either moderate or high activity, the data of that minute was discarded for further analysis [63]. However, such approach could bias the results, as subjects might generally experience higher or lower stress during high activity periods. Therefore, a more complete solution is to incorporate the physiological effect of physical activity into the classification model. Sun *et al*. [36] suggested an activity-aware stress classifier by adding accelerometer features, e.g. mean, standard deviation and energy of the three axes, to the decision tree model. The classification accuracy including accelerometer