Biomedical Physics & Engineering Express

ENDORSED BY IPEM Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT

Physiological Signal Based Work Stress Detection Using Unobtrusive Sensors

To cite this article before publication: Anusha A S *et al* 2018 *Biomed. Phys. Eng. Express* in press https://doi.org/10.1088/2057-1976/aadbd4

Physiological Signal Based Work Stress Detection Using Unobtrusive Sensors

Anusha A S, Jose Joy, Preejith S P, Jayaraj Joseph and Mohanasankar Sivaprakasam

Abstract—*Objective*: Work stress is identified as the "health epidemic of 21st century" by WHO because, when left unchecked, it wreaks havoc on human mind and body by accelerating the onset and progression of several health disorders. Hence, the evolution of strategies for early detection of mental stress is pivotal. The study presented here is one step towards the goal of developing a physiological parameter based psychological stress detection scheme which can further be incorporated into a wearable vital signs monitor. *Approach*: A group of 34 subjects (14 females and 20 males, age: 21.4 ± 1.7 years; mean ± SD) volunteered to participate in a pilot laboratory intervention that emulated real-life job stress scenarios by incorporating stress factors like mental workload, time pressure, performance pressure and social evaluative threat. Electrodermal Activity (EDA), Electrocardiogram (ECG), and Skin Temperature (ST) were monitored throughout the experiment to capture sympathetic activation during stress. Stress response elicitation was validated using salivary cortisol levels. A total of 61 features were extracted from these signals and four classifiers were investigated regarding their ability to detect 'stress' using single and multimodal schemes. A fusion framework that combined the benefits of feature fusion and decision fusion was employed to generate classifier ensembles for multimodal stress detection schemes. As the generated datasets exhibited a class imbalance issue, three separate schemes for class imbalance rectification viz., undersampling, oversampling and SMOTE were investigated concerning their ability to yield the best classification performance. While ECG based performance analysis were restricted to data segments of 300 s duration to conform to international guidelines for short term HRV analysis, non-overlapping EDA and ST data segments of durations 300 s, 180 s, 60 s and 30 s were examined to determine the optimum data length that can generate best results. *Main Results* : EDA gave a superior performance for 60 s windows while ST performed best with data segments of duration 30 s. A comparative study was performed with 25%, 50%, 75% and 90% overlapping data segments as well. However, overlapping did not enhance the performance of the classifiers significantly.While EDA emerged as the best single modality, the highest stress recognition accuracy of 97.13% was yielded by a combination of EDA and ST with data segments of 60 s duration. Furthermore, the differential effect of 'physical' and 'psychological' stressors on EDA and ST was analyzed. *Significance* : The results clearly suggest that these physiological parameters can not only reliably detect psychological stress but can also discriminate it from physical stress.

Index Terms—classifier ensemble, multimodal stress detection, physiological parameters, stress response elicitation, wearable

1 Introduction

Human stress is typically a state of physiological and psychological hyperarousal which is triggered when there is a disparity between situational demands and an individual's capacity to handle them. Such a disparity will challenge the state of complex equilibrium or homeostasis of the human body which is re-established through a series of innate adaptive responses coordinated by the central nervous system (CNS), collectively referred to as stress response. However, extremely intermittent, incessant or intense triggering of the stress response system will often lead to overstimulation of various target organs of the human body thereby causing numerous disorders like gastrointestinal, cardiovascular, musculoskeletal, epithelial, and more importantly mental disorders [1]. As many of these damaging diseases of gradual accumulation can be either inflicted or made far worse by stress, WHO has identified stress as the "health epidemic of 21st century".

The concept of stress has revolutionized the way in which medical community has been perceiving various diseases and their progression. The basic tenet upon which disease theory was based for generations - "one germ, one disease, one treatment" - is no longer valid. There is a significant paradigm shift due to the realization that an individual's mental state can have an enormous impact on the functioning and well-being of almost every cell in the body. This clearly explains the increased interest of the research community in the early detection of stress and its negative vital signs which lead to deleterious diseases.

The study reported herein is a progression towards the goal of developing a wearable device for "everyday life application" which can continuously monitor stress levels of an individual based on various physiological signals. Such a device can give an actionable feedback to the user regarding stressful situations and help them to calm down and refocus by employing various coping techniques. The utility of such a device can also be extended to psychotherapy and biofeedback therapy where a greater control of one's health is achieved by harnessing the power of the mind.

As job stress is a significant component of everyday stress, stressors close to real-life job stress scenarios were

- *Anusha A S and Mohanasankar Sivaprakasam are with the Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India. E-mail: ee14d016@ee.iitm.ac.in.*
- *Jose Joy, Preejith S P, Jayaraj Joseph and Mohanasankar Sivaprakasam are with the Healthcare Technology Innovation Centre, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India. E-mail: preejith@htic.iitm.ac.in .*

chosen for inducing psychological stress on subjects in a laboratory setting. Mental workload, time and performance pressure and social evaluation threat by superiors and co-workers are considered to be major contributing factors for stress in workplace. Hence stress tests were designed to include these components. Mental workload refers to the capability of a person to complete a task with some amount of the mental effort [2]. Salivary cortisol served as a reliable biomarker in assessing the capability of these stressors to elicit a stress response in subjects. A group of 34 subjects (14 females and 20 males, age: 21.4 ± 1.7 years; mean ± SD) volunteered to participate in this pilot study. Three physiological signals viz., Electrodermal Activity (EDA), Electrocardiogram (ECG), and Skin Temperature (ST) were recorded from each subject throughout the experiment. Feature extraction and classification were performed using each of these signals and their combinations in order to determine the best modality that can discriminate stressful phases from non-stressful phases. Furthermore, the capability of the chosen modality to distinguish a psychological stressor from a physical stressor is analyzed and reported here.

The following section gives an extensive summary of early research in psychological stress detection. The biology of stress response and its effects on EDA, ECG and ST are detailed in Section 3. A comprehensive description of stress experiments conducted in the laboratory setting is presented in Section 4. Subject characteristics, experiment procedures, and protocols are elaborated in Section 5 while Section 6 explains the feature extraction and classification methods employed for stress detection. Study results and inferences are presented in Section 7. Finally, Section 8 summarizes and concludes the work.

2 Previous Work

A review of the literature concerning human stress indicates that the field has been plagued by an overabundance of attempts to quantify stress. Stress response measurements as reported in Psychology and Engineering Sciences are summarized below.

2.1 Studies in Psychology

A common theme that cuts across the literature concerning human stress in the field of Psychology is the use of questionnaires, also known as stressor scales, for assessing stress levels. While some of them like Social Readjustment Rating Scale (SRRS) and Life Events Checklist (LEC) assess a subject's exposure to stressful life events and thereby obtain an overall index for the level of experienced stress, some others like Perceived Stress Scale (PSS) aim to measure the degree of uncertainty, uncontrollability, and overloading that a person feels about his/her life. Although conducting a perceptual study using such questionnaires enables the first-hand classification of stress levels as high, moderate, or low, they are very generalized and have the disadvantage of different individuals reading differently into each question and therefore reply, based on their own interpretation of the question [3].

Another strong focal theme emerged in stress measurement, when psychophysiologists incorporated hormonal and physiological variations during stress response in their studies. They believed that the assessment of endocrine stress axes was the most direct way of measuring stress response. This is normally done via. measurement of the hormone - cortisol. Accordingly, Kirschbaum et al. studied the variations in the salivary cortisol levels in a group of 20 men when subjected to a brief psychosocial stressor and observed that cortisol levels significantly elevated after stressful sessions of the experiment [4]. However, a similar study by Morris et al., examining the cortisol responses in 102 adults with varying risks of depression to a psychosocial stress task, noted the absence of a potent cortisol response to stress among individuals with a history of depressive disorder in comparison to normal subjects [5]. This indicates that cortisol based stress detection may fail with non-responders. Furthermore, the elaborate procedures involved in the collection and assessment of cortisol makes it unsuitable for continuous real-time stress monitoring.

Eminent psychologists like Wolfram Boucsein and Richard Lazarus have done extensive research on EDA as a sensitive and valid indicator of stress reaction. Most of these reported studies use electrical stimuli or movies to elicit a stress response [6], [7]. A recent review article on psychophysiological biomarkers of workplace stressors gave evidence of a negative correlation between workplace stressors and heart rate variability (HRV) which suggests work stress reduces vagal activity [8]. Although there is no universally recognized standard for stress evaluation at present, many researchers have reported studies that used HRV to measure stress, operating under the assumption that HRV is indeed a definitive physiological biomarker of stress. A recent review that examined literature providing a rationale for selecting HRV as a reliable indicator of psychological stress, highlighted the current neurobiological evidences that supported the use of HRV for the objective assessment of stress [9]. Most of these studies correlated variations in HRV metrics with low parasympathetic activity which is characterized by a decrease in the normalized power of high-frequency band and an increase in the low-frequency band. Neuroimaging studies have also suggested that HRV may be linked to cortical regions like the ventromedial prefrontal cortex that are involved in the appraisal of a stressor.

2.2 Studies in Engineering Sciences

As dynamic changes in the autonomic nervous system (ANS) in response to a stressor cannot be altered at will, the subsequent physiological signals can be used as reliable indicators of stress. Most of the reported studies in the field of Engineering used these physiological signals for stress detection. While some studies adopted a single sensor modality, some others focussed on a multimodal approach in order to obtain a better and more precise understanding of human stress phenomenon during different situations. Just as in the field of Psychology, an enormous amount of literature on the utility of HRV in detecting psychological stress is available in Engineering Sciences as well. Several studies reported in the last two decades had investigated HRV based assessment of chronic and acute stress on healthy as well as diseased subjects of different age groups. Long term and short term HRV derived from ECG or PPG had been

TABLE 1
PHYSIOLOGICAL SIGNAL BASED STRESS DETECTION METHODS REPORTED IN THE LITERATURE OF LAST DECADE

Reference	Signals	No.of Subjects	Stressor	Target	Classifier	Best Accuracy	
Setz et al. [11]	EDA	33	Mental arithmetic task	Stress and Cognitive load	LDA, SVM, NCC	82.8%	Single Sensor Modality
Keshan et al. [12]	ECG	17*	Real-time driving task	Low, Medium, and High Stress levels	NB, LR, MLP, Decision Tree	88.2%	
Karthikeyan et al. [13]	ST	60	Stroop color word test	Normal, Low, Medium, and High Stress levels	PNN	88.8%	
Jun and Smitha [14]	EEG	10	Stroop colour word test and Mental arithmetic	Low, High and No Stress levels	SVM	75.0%	
Healey and Picard [15]	EDA, ECG, EMG, RSP	27	Rest, Highway, and City driving	Low, Medium, and High Stress levels	LDA	97.4%	Multi-sensor Modality
Katsis et al. [16]	EDA, ECG, EMG, RSP	10	Simulated car racing	High stress, low stress, Disappointment, Euphoria and Neutral	SVM	86.0%	
Zhai and Barreto [17]	EDA, BVP, PD, ST	32	Computerized Paced Stroop Test	Stress and Relaxation	SVM	90.1%	
Lee et al. [18]	PPG, EDA, ST	80	Stroop color task with visual and auditory stimulus	Stress and Unstress	MLP, GRNN, ANFIS	96.7%	
Liao et al. [19]	HR, ST, EDA, Finger Pressure	5	Asynchronous math task and audio task	Stress	DBN framework	92.0%	
Engert et al. [20]	Cutaneous temperature of facial regions	15	Trier Social Stress Test , Cold Pressor Test	Baseline, Anticipation, Stress and Recovery	Univariate ANOVAs, MPA	56.0%	Imaging Modality
Chen et al. [21]	Tissue oxygen saturation (StO2)	21	Trier Social Stress Test	Normal and Stress levels	Binary Classifier	88.1%	

** The study used ECG signals of 17 automobile drivers obtained from MIT-BIH PhysioNet Multi-parameter Database*

analyzed. A wide range of devices from commercial off-the-shelf wrist wearables to smart phones, CE marked medical devices and advanced biomedical amplifiers had been employed for physiological data collection. Time and/or frequency domain metrics of HRV in both linear and/or non-linear domains had been explored. Descriptive statistical tests, correlation or machine learning algorithms had been utilized for stress detection. A recent and systematic review of such homogenously designed HRV based stress studies in the field of Engineering Sciences can be found in [10]. Recently, imaging techniques have also been utilized for non-contact stress detection. Table 1 summarizes the results of some of the most relevant literature of last decade, focussing on physiological signal based stress detection. The abbreviations and acronyms used in the table are defined in Appendix A.

Most of these reported studies used stressors with a mental work load component while some of them [11], [20], [21] included social evaluation. Although some studies attempted single physiological parameter based stress detection, they all agreed upon complementing it with other parameters to develop a more robust and reliable scheme mainly because a single sensor modality might be unsuitable for high stakes situations (eg. EDA may be a reliable indicator of stress but it is easy to fake an EDA response by simply clenching a fist). The imaging techniques based stress detection, on the other hand, require a high-performance computing system for real-time applications and might not be suited for continuous, long-term stress monitoring. As stress is highly subjective, a validation of stress response elicitation is indeed important; but it is missing in most of these reported studies.

2.3 Affective Multi-modal Databases

To the best of our knowledge, there are 5 publicly available multimodal affective databases which include physiological and/or behavioural responses of human subjects. Table 2 summarizes features of these datasets. The abbreviations and acronyms used in the table are defined in Appendix A.

As can be seen from Table 2, three out of five available databases (eNTERFACE'06, MAHNOB-HCI-Tagging and DEAP) are based on elicitation of different human emotions using evocative video clips, images or music. According to Lazarus, although psychological stress can be defined as a part of a larger set - the emotions, its motives, beliefs, appraisals, physiological response patterns and coping processes are entirely different [28]. Furthermore, eventhough these databases are interesting and informative, the physiological and psychological variations induced on a subject in a stressful work environment are different from what is experienced while watching a movie clip or listening to music. This reasoning applies to the automobile driver dataset in Physionet as well, although it aims at identification of different levels of stress using variations in physiological signals.Thus, the only publicly available database which is similar to the database generated and reported in this study (hereafter referred to as Stress Sense database) in its objective of detecting work stress is the most recent SWELL database. Stressors close to real life job stress scenarios

TABLE 2
MULTIMODAL DATABASES WITH PHYSIOLOGICAL AND/OR BEHAVIOURAL RESPONSES OF HUMAN SUBJECTS

Database (Year*)	No.of Datasets	Stimulus	Target	Recordings	Validation of Target Elicitation
Driver Stress - Physionet (2008) [15],[22]	17	Real-time driving task around the Boston area	Low, Medium and High stress levels	Physiological signals (EDA, ECG, EMG, RSP)	1.Self assessment using questionnaire 2.Video-based annotations**
eNTERFACE'06 (2006) [23]	5	Emotionally evocative images from the IAPS [24]	Calm, Positive excitement and Negative excitement	1. fNIRS data of frontal brain 2. Physiological signals (EEG, EDA, PPG, RSP)	Validation against IAPS annotation
MAHNOB-HCI-Tagging (2011) [25]	30	1.Emotionally evocative video clips . 2. Images and short videos of human actions with a tag attached to assess participant agreement.	Neutral, Anger, Disgust, Fear, Joy, Sadness, Surprise, Scream, Bored, Sleepy, Unknown, Amusement, Anxiety	1. Audio recordings of video stimuli 2. Camera views of test setting 3. Eye gaze data 4. Physiological signals (EEG, ECG, EDA, RSP, ST)	Self-assessment of emotion using SAM
DEAP (2012) [26]	32	Music video clips	Valence, Arousal, Dominance, Liking	1.Facial video 2.Physiological signals (EEG, EDA, ST, RSP, ECG, PPG, EMG, EOG)	Self-assessment using SAM
SWELL (2015) [27]	25	Writing reports and preparing presentations on a given topic with time pressure and email interruptions.	Neutral, Stress	1. Computer interactions 2. Facial video and Body posture 3. Physiological signals (ECG, EDA)	Self assessment using questionnaires

**Year when database was made publicly available. ** The stress ratings from the study are not available.*

were used for inducing psychological stress on subjects in a laboratory setting in both cases. The distinction of our dataset from SWELL is summarized in Section 5.5.

In an attempt to emulate a real-life job stress scenario, we combined several factors like mental workload, psychosocial elements, time pressure and performance pressure in our psychological stress experiments. The effectiveness of stressors was validated using salivary cortisol levels. The existing studies have already shown the capability of multimodal sensor schemes to detect stress. However, incorporating numerous sensors will make a system complex and sometimes practically implausible. Since we are aiming at a wearable "for everyday application", a minimally obtrusive sensor set up is most suited for comfort reasons. Hence we chose physiological signals like EDA and ST which could be easily measured from hands or peripheral body parts along with ECG which is widely accepted as a means to detect stress. The effect of addition of each of these physiological signals to the stress detection scheme was objectively analyzed to determine the ideal scheme for a wearable scenario. The effect of different class imbalance rectification techniques, varying window sizes and overlap were examined to determine the best scheme. Furthermore, the utility of the chosen modalities to discriminate a psychological stressor from physical stressor was investigated. We have not found a comparable study in the existing literature.

3 EDA, ECG, and ST as Signatures of Stress Response

The prompt and dynamic changes that human body encounters when a stress response is initiated are due to the activation of the autonomic nervous system (ANS). ANS has two branches - sympathetic and parasympathetic - which work together but in antagonistic ways to maintain homeostasis or balance. As ANS connects, like a hard-wire neuron system, to every other target organ of the body, excitation of sympathetic branch increases the activity of target organs while parasympathetic branch tries to slow things down to return them to a soothing state. Accordingly, ANS regulate various physiological measures like EDA, heart rate, ST, brain activity, etc. during stress response [29].

EDA is an all-encompassing term used to indicate electrical properties of skin. The most common measure of EDA is skin conductance which refers to the susceptibility of skin to conduct electricity. It is based on the sweat gland activity that is triggered in response to a stimulus. Unlike other target organs of the human body which are connected to both sympathetic and parasympathetic nervous system, skin with sweat glands and blood vessels are exclusively innervated by the sympathetic branch [6]. This makes EDA an ideal and unperturbed measure of sympathetic activation and hence, the stress response. Skin conductance exhibits a general rising trend with time, especially in humid environments, called the tonic skin response. However, phasic skin responses, which are transient peaks in response to certain stimuli like stressors, are used to mark sympathetic activation. Recorded conductance values are typically in the range of micro siemens (μS).

ECG of a participant was continuously recorded during the stress experiment in order to monitor the cardiac activity. This allowed extraction of two predominant parameters for stress detection viz., heart rate (HR) and heart rate variability (HRV) [30]. When a person encounters a stressor, HR increases. HRV, which is the measure of variation in the time interval between consecutive R-peaks of ECG, can capture variations in ANS activity which has both sympathetic and parasympathetic modulations. This way, examining HRV provides insight to the stress state of an individual whose HR increases during sympathetic activation

and slows down during parasympathetic activation.

As literature indicate the effect of acute psychosocial stress on ST [31], we included it as a physiological readout parameter of stress response in conjunction with other measures. Investigations illustrate that ST depends on the blood flow in the underlying blood vessels, which is regulated by the sympathetic nervous system. An acute stress activates sympathetic fibres which in turn triggers peripheral vasoconstriction thereby causing a sudden dip in ST. This way, ST is negatively correlated with stress.

4 Stress Test Design

As the study reported herein focussed on work stress, an experimental protocol that could emulate a stressful, real-life office scenario had to be administered to the subjects. Existing studies had reported the use of laboratory stressors like the Stroop Color Word Test, mental arithmetic test, and Trier Social Stress Test (TSST) as most effective means to induce mental workload and therefore, stress. While mental load is undoubtedly a vital stress component in an office scenario, there are also other contributing factors such as social evaluative threat by superiors and colleagues, time pressure and performance pressure which need to be considered. Hence, slight but thoughtful modifications were made to the classical form of these well-established laboratory stressors to include these stress components and to make them more pertinent to a real-life stressful office scenario. We have not found a comparable stressor design in the existing literature. The following sections describe in detail, the design of stressors used in this study.

4.1 Stroop Color Word Test

The Stroop Color Word Test developed by J. R. Stroop is based on a very robust finding that people respond slower when asked to name the colors in which words are displayed if the words form names of other colors than if the color and word are congruous (eg: BLUE printed in black color or RED printed in blue color) [32]. We designed a PC based color word test for our study with 5 different levels. The levels were designed such that the cognitive interference that participants' felt during the task gradually intensified. A sample test GUI is shown in Fig.1. A description of each level follows:

Level 1 : Practice session

Level 2 : Timed Stroop test where the word stimuli were presented on a low-intensity gray background as shown in Fig.1(a)

Level 3 : Timed Super Stroop test in which, not only were the word stimuli inked in incongruent colors, but also, they were presented on colored backgrounds that were incongruent to the color name and the color of the word. The task was to identify the color of the background first and subsequently, the color of the word. This was the computerized version of Super Stroop Test developed by Daniel Jozef in 1969 [33].

Level 4 : Timed Stroop test with distractions on the screen. The GUI of level 4 had two visually distracting components in it for each word stimulus viz., (i) a chromatically illuminated ball bouncing across the test screen. (ii) message pop-ups (which read "Close to deadline", "You received an important email" etc.) flashing on the test screen.

Level 5 : Timed Stroop test with background color change and distractions. Level 5 was a combination of level 3 and 4 with a modification. Here the back ground color was continuously changing for each word stimulus. The subject had to identify the word color correctly, amidst the changing background and other visual distractions of Level 4.

While level 1 allowed the participant to familiarize with the test GUI, levels 2 to 5 were used to gradually increase the stress levels. Each session had 10 questions with 6 options. Participant had to click on the correct option within 5 seconds. An incorrect answer (Fig. 1(b)) or timeout (Fig.1 (c)) at any level would restart the test.

Unlike prior studies which involved subjects performing Stroop test in isolation, this study demanded subjects' participation in front of an audience. The test screen was projected for an audience to emulate a social evaluative threat. In the classical Stroop Test, an incorrect answer or inability to answer within the stipulated time did not call for any penalty. However, the Stroop test used in this study not only demanded that the subject perform the test but also perform it correctly. An incorrect answer or timeout at any level of the test would restart the test which added to stress effects. This way, the test completion will happen only when the subject completes all levels of the test correctly. Furthermore, the subject was told before the test that he/she would be ranked based on the time taken by him/her for the successful completion of the test in comparison with others who already took the test. This was to incorporate a performance pressure component.

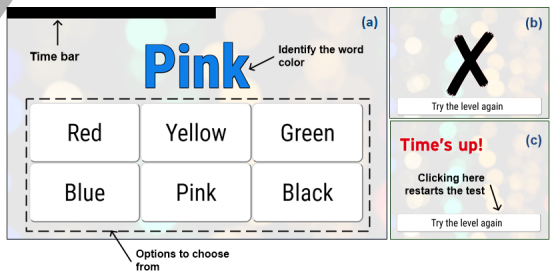

Fig. 1. (a) Stroop Color Word Test screen (b) An error case (c) A time out case

4.2 Mental Arithmetic Test

Mental arithmetic test, though performed in different ways, is the most widely employed laboratory stressor. It is also reported as more effective than other tests in inducing most cardiovascular responses [34], [35].

The principal component of the mental arithmetic task designed for this study was a computer program that demanded the participant perform mathematical calculations for a stipulated time of five minutes. The test was performed in front of an audience to emulate social evaluative threat. Questions involved four basic mathematical operations - addition, subtraction, multiplication, and division- on up to 3 digit numbers. The participant was not given any options for the correct answer but was instructed to type the answer in the text box provided. Multiple choice questions were avoided to increase the mental work load and to prevent mere guess work. Each question had a time limit of 15 seconds. The answer was evaluated immediately after the attempt and the score was updated. A countdown timer and score card were visible to the participant throughout. Prior

to the test, the subject was informed that he/she would be ranked based on the performance in the test and a comparison of his/her performance to the average performance of others who already took the test would be displayed on test completion which added to the performance pressure. The test GUI is shown in Fig.2.

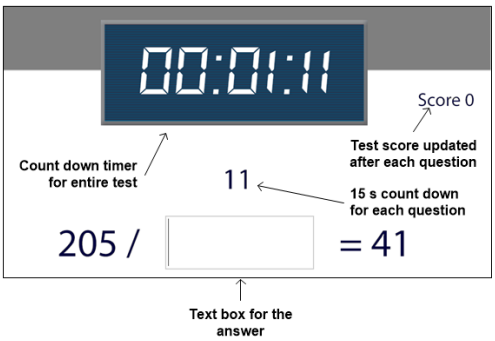

Fig. 2. Mental Arithmetic Test screen

4.3 Trier Social Stress Test

TSST is the most widely used protocol to induce moderate to intense psychosocial stress in laboratory settings [36]. The slightly modified TSST protocol used in this study mainly consisted of an anticipation period and a stress period during which the participant had to deliver a speech and perform mental arithmetic infront of an audience. The presentation was judged by a 4-member panel.

Initially, the participant was taken to an interview room. The room was set up in the following way: the participant sat on one side of a table facing a video camera and four chairs were put on the other side of the table for the panel members as shown in Fig. 3. The audience was seated behind the camera. The camera was used to induce stress more reliably. Once inside the interview room, participant was introduced to the subsequent task and a topic was given on which he/she had to make an oral presentation. A 3-minute preparation time was given to organizing the presentation during which the participant could use a pen and paper. At the end of the preparation phase, audience and the 4-member panel comprising of both males and females entered the interview room unexpectedly. The participant was informed that the presentation had to be given infront of the audience and would be judged by the panel. He/she was also told that the presentation would be videographed. Further, the paper with the presentation outline was taken away from the participant to exploit the vulnerability of stress response. The participant had to speak for 3 minutes and was urged to continue when stopped in less than 3 minutes.

The traditional TSST protocol used in prior studies demanded the participant to deliver an oral presentation in front of an evaluative panel that did not provide any form of feedback. However, in this study, participants had to perform in front of a responsive panel and audience. The panel and audience maintained a neutral expression throughout the presentation. However, a feedback and Q&A session followed, wherein the panel gave feedback on the promptness and believability of the oral presentation, and the audience raised a few queries to the participant. These interactions made the stressor much more relatable to a real-life work presentation scenario.

The presentation was followed by a mental arithmetic task, during which the participant was instructed to perform cyclic subtraction of a 2-digit prime number from a 4-digit number till the first two digit number was reached. This had to be done within a stipulated time of 3 minutes. A prime number was chosen as the subtractor to increase the difficulty level of the task. A countdown timer was set for the participant to know the elapsed time. On every failure, one of the panel members interfered and asked the participant to restart the task thereby enhancing the stress levels. The entire TSST procedure took an approximate duration of 15 minutes.

While oral presentations may be stressful for some people, some will be good at it. Similarly, mental arithmetic or a paced cognitive task may be difficult for some, while easy for others. Hence a combination of the aforementioned psychological stressors was used to ensure stress response elicitation in all subjects in the reported study.

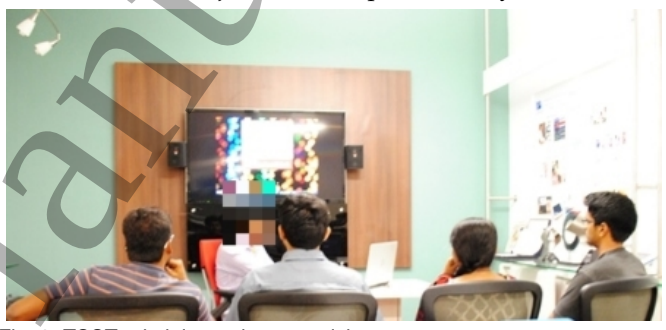
Fig. 3. TSST administered to a participant

4.4 Physical activity

An exhaustive physical activity is often considered a physical stressor because it forces the human body to temporarily deviate from its natural set point (homeostasis) and causes the release of cortisol [37]. However, studies show that regular exercise can significantly reduce the intensity of stress response and cortisol release in humans [38].

To examine the effect of a physical stressor on physiological signals, the participant was instructed to walk continuously on a treadmill (Afton® CP161) at varying speeds for 10 minutes. Speed was set at 3 km/hr for the first 3 minutes, varied to 4 km/ hr for the next 3 minutes followed by 5 km/hr for the last 4 minutes of the experiment.

5 Subjects and Methods

5.1 Subject Characteristics

A group of 34 subjects (14 females and 20 males) volunteered to participate in this pilot study. They were not selected based on any criteria related to their stress levels. All participants were of Indian origin and had a Science and Engineering educational background. All of them were unmarried and lived away from their parents owing to studies/work. While 7 out of 34 subjects were professionals working at the Healthcare Technology Innovation Centre, IIT Madras (with experience ranging from 2 months to 2 years), the rest were university students undergoing post-graduate or undergraduate courses in Engineering. All of

them were in self-reported good health with no signs or history of neurological, psychiatric, endocrine or acute/chronic physical disease. They were not taking any medication and had not reported any habitual drug or alcohol usage or smoking habits. All participants were right-handed and had normal or corrected-to-normal vision. Participants could read, write and speak English and had substantial computer experience.While 19 participants reported having a definite sleep pattern with 6 or more hours of sleep every day, the rest reported having varying sleeping hours (5-8) based on the daily schedule. 4 participants reported eating fast food on a daily basis and 8 on a weekly basis. Rest of them ate fast food occasionally. 22 out of 34 subjects reported indulging in moderate to vigorous physical activities every day. All of them were naive to the Stroop Color-Word test, Trier Social Stress Test (TSST) or similar stress paradigms. All female participants were tested in the luteal phase of their menstrual cycle and none of them reported use of oral contraceptives. Along with anthropometric measures, body composition measures of participants were also recorded prior to stress experiments using BC-545N, a segmental body composition monitor from TANITA® . Some of the relevant physical characteristics of the subjects are summarized in Table 3.

TABLE 3
PHYSICAL CHARACTERISTICS OF STUDY SUBJECTS

Variable	Combined (n=34)	Females (n=14)	Males (n=20)
Age (years)	21.4 ± 1.7	21.9 ± 1.8	21.1 ± 1.5
Height (cm)	166.7 ± 8.6	160.4 ± 4.5	171.1 ± 7.9
Weight (kg)	62.7 ± 10.2	63.6 ± 10.0	62.0 ± 10.3
BMI (kg/m^2)	22.7 ± 4.1	24.8 ± 4.5	21.2 ± 3.1

Values are mean ± SD; n is the number of subjects.

5.2 Instrumentation and Experimental Setup

The instrumentation used in this study enabled a continuous acquisition of EDA, ECG and ST from the participant throughout the experiment session. These signals were conveniently monitored by non-invasive means.

Measure of EDA was obtained using a Galvanic Skin Response (GSR) amplifier from AD Instruments® which has a low excitation voltage (22 mV$_{rms}$). It operates at 75 Hz AC excitation and has automatic zeroing. Bipolar finger plate electrodes made from brightly polished stainless steel were taped around the second phalanx of the index finger and ring finger of the non-preferred hand of the subject for EDA recording.

ECG was continuously acquired using AD8232, a heart rate monitor front end from Analog Devices®. Three acrylic based Ag/AgCl solid gel ECG electrodes were placed on the chest of the subject, two directly below the clavicle, near the left and right shoulders and the third on the left lower abdomen.

ST was acquired from two locations of the non-preferred hand of the subject viz., middle finger and ventral forearm, using two carefully calibrated NTC thermistors from Ohmeda® in a potential divider network. The sensors had a resolution of 0.1 °C. Sensors were calibrated for accurate measurement using 9142 Metrology well from Fluke®over a range of 19 °C to 41 °C and calibration coefficients were derived using Steinhart–Hart equation. The ambient temperature was maintained constant at 27 °C during measurements.

NI USB 6366, a multifunction DAQ device from National Instruments® was used for the continuous acquisition of all physiological signals. The device has 8 differential analog input channels with 16 bit resolution with an update rate of 3.33 MS/s. The platform communicated via a USB connection to an external personal computer where a graphical interface was implemented in LabVIEW to visualize the acquired signals. LabVIEW VIs were also implemented to acquire the mouse click events during the stress tests and screen capture. Video recordings of the participants face and upper body were made with a high-definition Handycam Camcorder (Sony HDR-CX260V).

5.3 Salivary Cortisol as a Biomarker for Stress

When an individual encounters a stressor, the hypothalamic-pituitary-adrenal (HPA) axis is activated which further leads to secretion of cortisol and catecholamine in the body. Although these end products of HPA activation are measurable in urine, plasma, and saliva, the recent years of research has seen the emergence of salivary cortisol as a reliable biomarker for sympathetic activation [39]. Therefore, salivary cortisol levels are used as a means of validating the stress response elicitation due to physical and psychological stressors in this study.

Saliva samples were collected from the participant before and after stress tasks for measuring the cortisol concentration. Both collection and measurement were done using the salivary cortisol test kit from SOMA Bioscience®. The test kit is known to correlate well with values measured in the laboratory ELISA and when running in duplicates usually has within assay CVs of below 10% with a dynamic range of 2 - 40 nM and a sensitivity of 2 nM [40]. The following precautions were taken during sample collection [41]:

(1) Participants were instructed to avoid beverages which are acidic or has high sugar, caffeine or nicotine content during the hour prior to sample collection in order to prevent pH alterations, further leading to bacterial growth in samples.

(2) As glucose intake affects the cortisol response to stress in a predominant way, a heavy meal in the hour prior to sample collection was not allowed.

(3) A strenuous physical activity immediately before the psychological stress test was avoided as it would affect the baseline cortisol levels.

(4) Subjects were instructed to rinse their mouths thoroughly with water 10 minutes prior to the baseline sample collection.

The first sample was taken before the stress task and the second was taken within 20-30 minutes after the onset of stress task. The second sample collection was timed based on the pioneering studies by Kirschbaum and Hellhammer which suggest that the cortisol concentration in saliva peaks within 20-30 minutes after the onset of a laboratory stressor [42].

Apart from stress response validation, salivary cortisol levels were also used to investigate the circadian pattern of basal cortisol secretion on all subjects. Accordingly, 6

samples were collected from each subject, a day prior to the test. The first sample was collected nearly 30 minutes after awakening from sleep so as to capture the morning peak of cortisol secretion in most healthy people [43]. The participants were instructed to start sample collection only if:

(1) They had a quality sleep of normal duration the previous night; and

(2) They woke up between 6.00 h and 8.00 h in the morning.

The other samples were collected at 12.00 h, 16.00 h, 18.00 h, 20.00 h, and just before sleep. Salivary cortisol levels were determined from these samples for all subjects in order to replicate the cortisol circadian rhythm.

5.4 Study Protocol

The psychological stress experiment protocol as shown in Fig.4, had an approximate duration of 110 minutes with three experiment phases viz., baseline, stress, and recovery. The experiments were conducted in the late afternoons to control for the diurnal variations in cortisol secretion. The experimental schedule is listed below:

(1) The participant, on arrival at the test venue, was asked to relax and wait comfortably in a room for about an hour. The participant was informed that he/she was called in to record some physiological measures as a part of a research work and a consent was taken. However, the stress aspect of the experiment was not revealed. Identity details, anthropometric and body composition measures of the participant were recorded.

(2) After an hour, the pre-stress saliva sample was collected from the participant.

(3) The participant was taken to the interview room and sensors to capture various physiological parameters were attached. He/she was instructed to restrict all voluntary movements of torso and hand on which sensors were attached to minimum. The baseline recording was done for 10 minutes.

(4) At the end of the 10 minutes baseline period, TSST was administered. The judging panel and the audience entered the interview room after the preparation phase of TSST

(5) TSST was followed by task 4.1 or 4.2 or both as decided by the test supervisor who monitors the acquired physiological signals throughout.

(6) The stress phase was followed by debriefing of the participant on the true nature of the experiment. The participant was informed that the objective of the study was to induce stress and the performance was in no way a reflection of his/her aptitude or ability. The panel and the audience left the interview room after debriefing.

(7) The participant was instructed to remain seated for another 10 minutes when the physiological measures during the recovery phase were recorded. A post-stress saliva sample was also taken from the participant.

(8) Sensors were removed at the end of the recovery phase and the participant was allowed to leave the venue.

The physical stress study protocol had an approximate duration of 90 minutes and followed a similar schedule. After a waiting time of one hour, sensors were attached to the subject and a 10-minute baseline recording was done. This was followed by a physical stress phase and a recovery period, each of 10 minutes duration. Saliva samples were collected before and after the physical activity to validate the stress response elicitation. 12 out of 34 subjects (4 females and 8 males) volunteered to participate in the physical stress study. The physical stress test was also conducted in the late afternoons but on a different day.

5.5 Dataset Summary

The 'Stress Sense database' generated and reported in this study, following the above protocol has 34 psychological stress datasets each of an approximate duration of 50 minutes and 12 physical stress datasets each of 30 minutes duration. The physiological data of each subject include EDA, ECG, finger temperature and wrist temperature. Apart from the physiological signals, the database includes video recordings of the stress experiments conducted on all subjects. The study population involves university students as well as working professionals.

The Stress Sense database is superior to the already existing SWELL knowledge worker database in the following aspects:

(1) The psychological stressor employed in the SWELL project uses time pressure as a stress eliciting factor. This is certainly an important stress component in a real-life office environment. However, there are other contributing factors such as mental workload, performance pressure and social evaluation by superiors and colleagues. We have therefore used stressors that includes all these factors along with time pressure to generate the Stress Sense dataset.

(2) The SWELL project uses standard questionnaires to validate the capability of the stressor to elicit a stress response in subjects. Although conducting a perceptual study using such questionnaires enables the first-hand classification of stress levels as high, moderate, or low, they are very subjective. We used an objective measure like salivary cortisol for evaluating stress response elicitation in subjects. Data of only those participants whose post stress cortisol level is more than the pre-stress cortisol level are included in the database which makes it much more reliable.

(3) The Stress Sense database includes datasets of 12 subjects who volunteered to participate in both physical and psychological stress tasks. These 24 datasets can be utilized to study the differential effects of these stressors on recorded physiological signals. The differential effects of these stressors on EDA was analysed by our group and the details can be found in [44].

(4) The rather few systematic studies on the relationship between EDA and ST indicate that EDA parameters like latency are considerably influenced by increases or decreases in skin temperature, which can be explained by the temperature dependency of the acetylcholine transport mechanism [6]. Some of these studies observe that skin temperature is a mandatory measurement in the study of EDA [45]. The simultaneous recordings of EDA and ST available in the Stress Sense database can be used for further research along these lines.

6 Methodology

This section describes various evaluation methods used in the study to detect 'stress' based on the acquired physiological data. After preprocessing the data, standard features

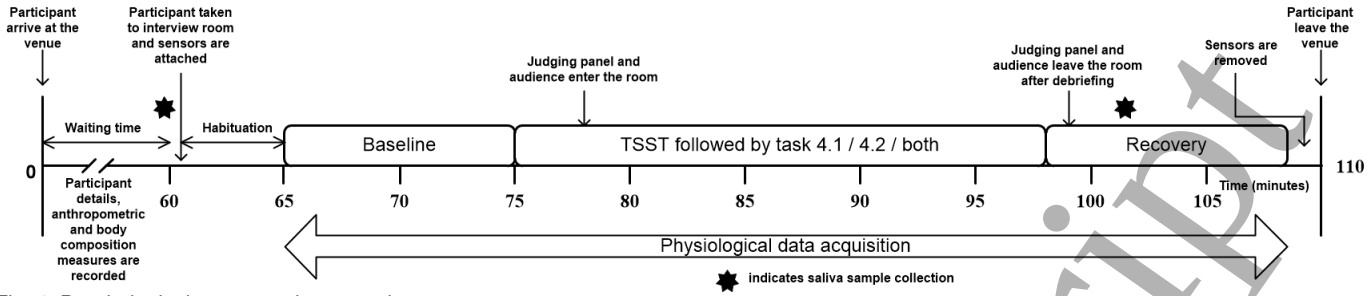

Fig. 4. Psychological stress study protocol

were extracted from the generated HRV while descriptive statistics was employed to find potentially meaningful features from EDA and ST. Feature extraction was followed by feature selection in order to determine the optimum feature subset that yielded the highest leave-one-person-out cross-validation accuracy. Four different classifiers and seven sensor modalities were investigated concerning their ability to classify 'baseline' and 'stress' states. MATLAB® software was used for all data processing unless mentioned otherwise. The following sections describe the methodology in detail.

6.1 Measured Signals and Preprocessing

As already mentioned, 3 physiological measures viz., EDA, ECG and ST (from finger and wrist) were recorded throughout the experiment at a sampling frequency of 1 kHz. Further, LabVIEW® was used to generate and examine plots of each of the physiological measure over time so as to ensure that data were recorded properly without any intermittent loss during the experimental sessions. This way, any issues with the dataset that would later invalidate experimental results (e.g., if a sensor electrode went off contact with the subject during the experiment) could be identified and removed. As the experimental protocols adopted in the study required minimum/no movement of the non-dominant hand and torso, participants were instructed to restrict all voluntary movements to a minimum during the recordings. However, time synchronized video recordings coupled to simultaneously recorded physiological signals of each subject were analysed manually to get a precise annotation of the movements (if any) performed by the subject and to subsequently identify and eliminate those parts of physiological signals that were potentially affected by motion artifacts. The physiological datasets so generated were further cleaned up using appropriate filters to eliminate the high frequency noise content in each signal. As a sampling rate as high as 1 kHz is not imperative to precisely represent relatively slow signals like EDA and ST [46], these were downsampled to a frequency of 250 Hz before further analyses. A Butterworth low pass filter of cutoff frequency 5Hz was applied on EDA and ST signals, while a 5-15 Hz band pass filter was implemented on ECG signal.

6.2 Feature Extraction

After preprocessing, the physiological data recorded from each subject were divided into 'baseline', 'stress' and 'recovery' phases using time stamps. These were labelled as '1', '2' and '3' respectively. Further, the recovery section was removed as the objective was to distinguish 'stress' from 'baseline'. A total of 61 features (30 from EDA, 10 from ECG and 21 from ST) were calculated using the original and derived signals. The following sections describe feature calculation from each physiological signal in detail.

6.2.1 EDA

An EDA signal is generally characterized by its tonic and phasic components where tonic component refers to gradual variations in the skin conductance levels in the absence of any external stimuli and phasic EDA represents sudden variations in response to an external or internal stimulus [6]. As such, the phasic component qualifies as an appropriate measure of sympathetic activation. However, a strict demarcation between these two components is not straightforward because tonic EDA generates a constantly varying baseline among individuals as well as within an individual. This makes mere averaging performed over the whole signal inadequate to extract a precise measure of tonic EDA. On the contrary, such an averaging will lead to an over-estimation of true tonic component as it also contains phasic component which artificially elevate the measure. In an attempt to overcome this challenge, the noise filtered EDA (referred to as 'filtered_EDA' hereafter) was divided into small windows, each of 15 seconds duration and the least-squares straight line fit to the data was computed. The resulting least-square fit was subtracted from the data to generate a 'detrended_EDA' signal. Further, a 'diff_EDA' signal was obtained as the difference between filtered_EDA and detrended_EDA. The 15 seconds window was chosen because tonic EDA typically changes over a period of few tens of seconds to minutes [6].

Although standard parameters like amplitude, latency time, recovery time etc., can be used to characterize an EDA signal, these are defined with respect to a single stimulus like a startle event. Since stressors used in this study cannot be categorized as startle events, these features become insignificant. Therefore, descriptive statistics was used to explore the filtered_EDA signal and its derivatives (detrended_EDA and diff_EDA) to determine potentially useful features for distinguishing 'stress' and 'baseline'. 10 features extracted from each window of these signals are summarized in Table 4. Thus a typical window of EDA data will have 30 features. Four different window sizes viz., 30 s, 60 s, 180 s and 300 s were examined to determine the optimum data length for EDA data.

6.2.2 ECG

The cardiovascular reaction induced by stress is most efficiently measured using two parameters derived from ECG viz., HR and HRV. While HR can be determined by computing the frequency of R-peaks in an ECG signal, HRV is a measure of oscillations in the interval between consecutive

TABLE 4
DESCRIPTIVE STATISTICAL FEATURES EXTRACTED FROM EDA

Feature	Unit	Definition
mEDA	μS	Arithmetic mean of data values
varEDA	μS^2	Expectation of the squared deviation of data values from mean
ranEDA	μS	Difference between maximum and minimum data values
ctEDA	-	Ratio of range to absolute mean
cvEDA	-	Ratio of standard deviation to mean
quant_10, quant_25, quant_50, quant_75, and quant_90	μS	Thresholds calculated such that 10%, 25%, 50%, 75% and 90% of data values are smaller than respective thresholds.

heartbeats. Both these measures require precise detection of R-peaks from the ECG signal. Accordingly, an algorithm based on that of Pan and Tompkins [47] was implemented in-house for R-peak detection with a GUI that allows to manually add missed R-peaks and also remove if false R-peaks were detected. ECG from all datasets were individually analyzed using this algorithm in order to make sure that all R-peaks were accurately identified.

A window size of 5 minutes was used for ECG data to conform to international guidelines for short term HRV analysis [48]. As the sympathetic and parasympathetic branches of the ANS exhibit different latencies in affecting the HR (1 second for the parasympathetic and 2-3 seconds for the sympathetic branch), their influence can be separated by means of frequency analyses [49]. Therefore some frequency domain HRV measures were also used apart from the time domain parameters. HRV features used in this study is indicated in Table 5.

TABLE 5
HRV FEATURES IN TIME AND FREQUENCY DOMAINS

Feature	Unit	Definition	
mRR	ms	Arithmetic mean of RR intervals	Time
medRR	ms	Median of RR intervals	
mHR	bpm	Mean heart rate	
SDRR	ms	Standard deviation of RR intervals	
RMSSD	ms	Square root of the mean of squares of differences between successive RR intervals	
RR50	count	Number of pairs of adjacent RR intervals differing by more than 50 ms	
pRR50	%	RR50 count defined as a percentage of total RR intervals.	
LF	ms^2	Power in low frequency range (0.04-0.15 Hz)	Frequency
HF	ms^2	Power in high frequency range (0.15-0.4 Hz)	
LF/HF	-	Ratio of power in LF range to HF range	

6.2.3 Skin Temperature

The preprocessed finger and wrist temperature data (hereafter referred as 'finger_TEMP' and 'wrist_TEMP') were also analysed as potential physiological read out parameters of stress response. Seven descriptive statistical features shown in Table 6 were extracted from every window of both these signals. Furthermore, a 'diff_TEMP' signal generated as the difference between finger and wrist temperatures was also used for feature extraction. Thus, every window of wrist and finger temperature data yielded 21 features. Four different window sizes viz., 30 s, 60 s, 180 s and 300 s were examined to determine the optimum data length for ST data.

TABLE 6
DESCRIPTIVE STATISTICAL FEATURES EXTRACTED FROM ST

Feature	Unit	Definition
mTEMP	°C	Arithmetic mean of data values
maxTEMP	°C	Maximum data value
minTEMP	°C	Minimum data value
sdTEMP	°C	Standard deviation from mean
ranTEMP	°C	Difference between maximum and minimum data values
ctTEMP	-	Ratio of range to absolute mean
cvTEMP	-	Ratio of standard deviation to mean

6.3 Class Imbalance and Feature Extension

As the baseline and stress phases of the psychological stress experiment had different durations, there was a problem of class imbalance in the datasets, i.e., the 'stress' class had more data points than the 'baseline' class. Accordingly, 'stress' forms the majority class and 'baseline' forms the minority class. Such an imbalance may bias the classification algorithms for always predicting the majority class thereby giving high accuracy but a low generalization. Hence, three different schemes were examined in this study to identify the best scheme to overcome the class imbalance:

(1) Undersampling where a subset of samples was randomly selected from the majority class to equalise the number of instances coming from each class [50].

(2) Oversampling where samples from the minority class were duplicated cyclically to generate additional instances until class balance is achieved [51].

(3) Synthetic Minority Over-sampling TEchnique (SMOTE) where the minority class was oversampled by creating synthetic data points via linear interpolation between existing samples [51].

6.4 Feature Selection and Classification

Feature extension was followed by a feature selection procedure aimed at obtaining a subset of features by discarding the least relevant ones whilst maintaining the efficacy of the classifier algorithm. Accordingly, scatter plots, which provide a visual representation of the correlation between two variables were generated for all possible combinations of already identified features to obtain a feature subset. This was done on EDA, ECG and ST features separately. Further, feature selection using wrapper method was performed on the generated subsets with leave-one-person-out cross-validation.

In leave-one-person-out cross-validation, the classifier is trained with all data sets except one and is tested on the left out dataset. The procedure is repeated for all datasets and accuracy is estimated as the percentage ratio of correctly classifed data points to the total number of data points [52]. A complete search for the best feature combination was performed separately for EDA, ECG and ST feature subsets. A cross-validation was done for all possible feature combinations of the subsets and the combination that returned the maximum accuracy in distinguishing 'baseline' and 'stress' states was chosen for each classifier. Four classification methods viz., Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) , Support Vector Machine with rbf kernels (SVM), and k Nearest Neighbors (kNN) with k = 3 were investigated based on leave-one-person-out cross-validation accuracy as the performance

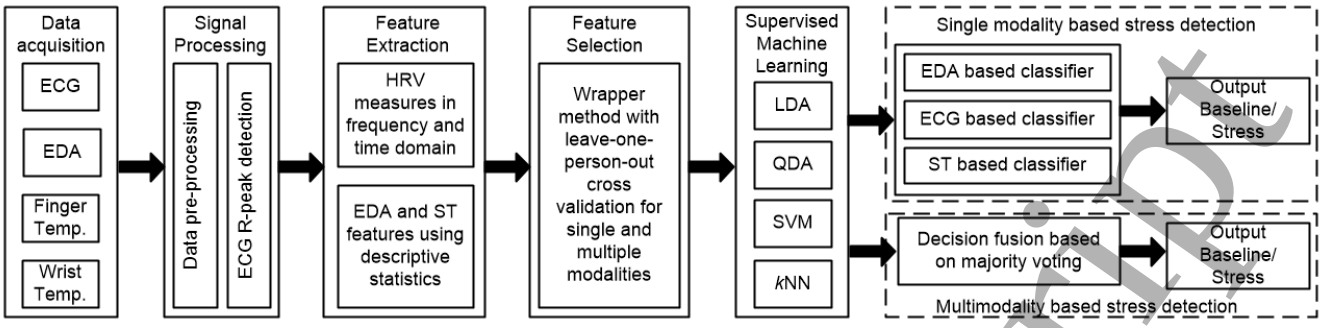

Fig. 5. General architecture of the stress detection scheme employed in this study

index. This way, a classifier system that used an optimum feature set as input and gave superior performance in terms of classification accuracy, was determined for EDA, ECG and ST.

Further, a classification framework that combined the concepts of both feature fusion and decision fusion was developed to generate ensemble classifier systems for multimodal stress detection. Evidently, feature fusion and decision fusion are two well established techniques used in pattern recognition systems to achieve the best possible classification performance for a given problem domain. While methods of feature fusion deal with the selection and combination of features to remove redundant and irrelevant features, decision fusion uses a set of classifiers to provide a better and unbiased result [53]. Accordingly, feature selection using wrapper method was performed to determine the best feature combination that yielded maximum accuracy for each of the four multimodal schemes possible in this study viz., EDA+ECG, ECG+ST, EDA+ST and EDA+ECG+ST. This was done separately for all classifier combinations using LDA, QDA, SVM and kNN. These classifiers were combined to form ensemble classifier systems using decision fusion. The decision fusion scheme is analogous to "seeking opinion from an expert panel" before making an important decision. Each classifier was considered as an "expert" and the decisions of these classifiers were combined using the 'majority voting' approach [54]. This way, the class that received the highest number of votes was chosen. In case of a tie where both classes received an equal number of votes, the corresponding classifiers were identified as candidate classifiers and the one with the highest confidence score among them provided the deciding vote.

An exhaustive two stage analyses involving an initial feature fusion stage followed by a decision fusion stage was done to identify the best classification scheme. Accordingly, all possible feature combinations of each of the physiological signals were considered against all possible classifier combinations in the decision fusion stage. The analysis was repeated for all 4 multimodal schemes. Leave-one-person-out cross-validation accuracy was considered as a performance index to identify the best scheme for stress detection. A schematic representation of the stress detection scheme employed in this study is illustrated in Fig.5.

7 Measurement Results and Discussions

7.1 Circadian Rhythm and Cortisol Validation

The cortisol circadian rhythm is considered to be one of the most distinct and fascinating patterns in human physiology. As mentioned in Section 5.3, salivary cortisol analyses were performed on all 34 subjects over the duration of a day to replicate this pattern. The diurnal variations as can be seen from Fig.6, indicate that cortisol levels are at a peak around 7:00 h in the morning (15.2 ± 0.48, mean ± SEM), steadily decline during the day to decrease to the nadir towards midnight (1.7 ± 0.46, mean ± SEM). Accordingly, all stress experiments were conducted during late afternoons to limit the interference of diurnal variations in cortisol secretion. Furthermore, the rhythm clearly indicates that participants were normal and healthy and not suffering from any form of exhaustion like adrenal fatigue.

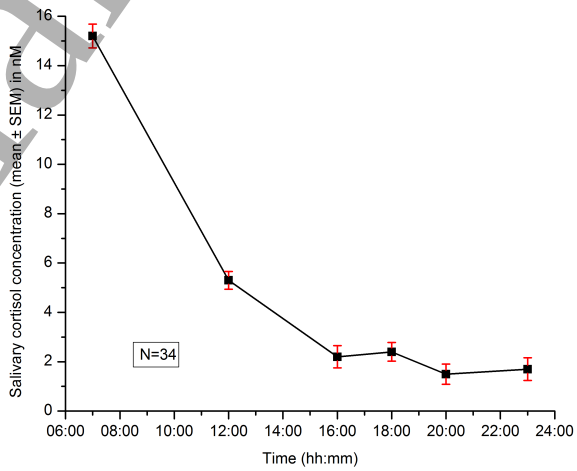

Fig. 6. Circadian rhythm of cortisol secretion in 34 subjects represented as mean ± SEM

Fig. 7 indicates salivary cortisol levels before and after the physical and psychological stress experiments conducted on 12 subjects and 34 subjects respectively.

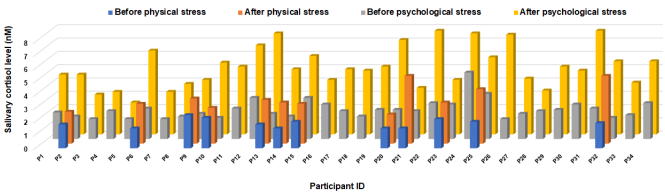

Fig. 7. Salivary cortisol levels of subjects before and after stress tasks. P1 to P20 : Male, P21 to P34 : Female.

As can be seen, both physical and psychological stressors effected an increase in the cortisol levels for all subjects. For the psychological stress task, cortisol levels showed nearly 123% increase on an average. This shows that the participants were "really stressed" during the laboratory intervention. Therefore, it can be concluded that the stress detected in this study is a strong affective state. The physical

stressor, on the other hand, illustrated an average increase of 78% in cortisol levels. Collectively, these findings support the view that in addition to psychological stressors, a moderate intensity exercise can also provoke cortisol secretion, although at a comparatively lesser level. However, regular exercise can effect lesser cortisol release thereby making the body less sensitive to stress [55].

7.2 Classification and Modality Comparison using Learning Systems

Non-overlapping data segments of 5-minute duration were chosen initially for feature extraction. As mentioned in Section 6.4, a subset of EDA, ECG and ST features, with 22, 10, and 16 elements respectively, were generated using scatter plot analysis. Furthermore, feature selection using wrapper approach was employed on these subsets separately. Cross-validation was performed for all possible feature combinations from these subsets and the best feature combination that can distinguish 'baseline' and 'stress' states were chosen for each of the 4 classifiers based on the cross-validation accuracy. The maximum accuracy was achieved by the classifiers when the best feature combination derived using leave-one-out cross-validation was provided as input. As the performance of classifiers are also influenced by the class imbalance rectification technique employed, the above mentioned feature selection and cross-validation was performed separately after employing each of the imbalance correction technique mentioned in section 6.3. A summary of such a comparative study done for EDA based stress detection is shown in Table 7. 'Original' in the Table refers to the original dataset where no class imbalance correction was employed. Accuracy represents the proportion of correctly predicted instances among all instances. Sensitivity is the proportion of samples correctly predicted as "stress" among all stress samples while specificity describes the proportion of correctly classified negatives (correctly predicted as "baseline" among all baseline samples). Precision describes the proportion of "stress" predictions that are actually from stress samples and F1 score is the harmonic mean of precision and sensitivity.

As can be inferred from Table 7, different imbalance rectification techniques can influence the performance of classifiers. Furthermore, *k*NN yielded a maximum accuracy of 90.91% for the EDA based classification done on a dataset where class imbalance was rectified using the SMOTE method.

Similar analyses were done for stress classification based on ECG and ST as well. A comparison of such maximally achieved cross-validation results for stress detection based on EDA, ECG and ST is shown in Table 8. The optimum feature combinations corresponding to the best performing classifiers (*k*NN for EDA and ECG, QDA for ST) are indicated in the 'Best Feature Combination' column. When multiple feature combinations gave the same highest accuracy for a classifier, the one with least number of features were chosen.

While *k*NN yielded a maximum accuracy of 88.03% for ECG based stress classification, ST was slightly more precise with QDA. The SMOTE method outperformed other schemes of imbalance correction in all three modalities. The maximum single modality classification accuracy of 90.91% was given by EDA. Thus, EDA clearly represents the best single modality, followed by ST and ECG. Although *k*NN requires a large data set for precise classification, the classifier performs well when the class boundaries are non-linear. The superior performance of *k*NN with EDA and ECG can be attributed to this non-linearity. The second best performing classifier, QDA, is also non-linear indicating that the problem may not be linearly separable.

Three out of four features selected by *k*NN to achieve the highest accuracy with EDA were quantile threshold features. Interestingly, the mean of EDA data values (mEDA), which would probably be one of the first features one would extract, was not selected. This indicates that quantile thresholds are suitable features to distinguish stress from baseline. Quantile measures not only provides some robustness to outliers like a transient peak in EDA, but also are independent of the number of phasic peaks occurring during the respective time period. This is an advantage because individuals differ substantially in their phasic EDA response patterns. The best feature combination for ECG include nine out of ten features extracted. The omission of pRR50 can be attributed to its strong correlation with other measures of vagal ANS activity such HF and RMSSD which are already present in the best feature set [56]. QDA, which is the best temperature based stress classifier chose some of the most straightforwardly calculated time-domain features like mean, minimum, maximum, range and standard deviation. However, ctTEMP and cvTEMP for finger_TEMP, wrist_TEMP or diff_TEMP were not selected indicating that these features are inadequate for the classification of baseline and stress. This could be because relative variations within temperature samples of baseline and stress sections over a duration of 5 minutes are insignificant.

To further improve the stress detection scheme, ensemble classifier systems with multimodal signals were examined. As the number of component classifiers in an ensemble has a predominant impact on the classification accuracy, all 11 classifier combinations using LDA, QDA, SVM and *k*NN were evaluated to determine the best ensemble. This was done for all 4 multimodality schemes possible in this study viz., EDA+ECG, ECG+ST, EDA+ST and EDA+ECG+ST. While ensemble classifier systems did not always gave an accuracy higher than the best single modality classifier, some combinations yielded better results. A summary of the best results is shown in Table 9. Only combinations which yielded higher accuracies than the best single modality are shown in Table 9. The best accuracy for each multimodality scheme is indicated in boldface. SMOTE emerged as the best bias rectifier for all 7 cases.

As can be seen, a QDA-*k*NN ensemble classifier system using EDA, ECG, and ST features as inputs yielded the highest accuracy of 95.86%. This was 4.95% higher than the accuracy of the best single modality classifier. While the best feature combination for QDA in the ensemble had 12 features (cvEDA, quant_25 (filtered_EDA), quant_50 (detrended_EDA); mRR, mHR, RMSSD; sdTEMP, ranTEMP (finger_TEMP, wrist_TEMP), mTEMP, ranTEMP (diff_TEMP)), *k*NN gave the best accuracy with 10 features (varEDA, quant_25 (filtered_EDA), ctEDA, quant_50 (detrended_EDA), varEDA (diff_EDA); SDRR; ranTEMP(finger_TEMP, wrist_TEMP)), sdTEMP

TABLE 7
CROSS-VALIDATION RESULTS FOR EDA BASED STRESS DETECTION WHILE EMPLOYING DIFFERENT CLASS IMBALANCE RECTIFICATION SCHEMES

	Original				Undersampling				Oversampling				SMOTE			
	LDA	**QDA**	**SVM**	***k*NN**	**LDA**	**QDA**	**SVM**	***k*NN**	**LDA**	**QDA**	**SVM**	***k*NN**	**LDA**	**QDA**	**SVM**	***k*NN**
Accuracy (%)	81.05	87.37	77.89	75.79	79.31	75.86	82.76	77.59	82.22	81.34	83.52	84.55	77.27	78.03	71.97	**90.91**
Sensitivity	0.89	0.92	0.91	0.83	0.83	0.86	0.90	0.86	0.88	0.81	0.83	0.80	0.83	0.80	0.79	**0.92**
Specificity	0.62	0.76	0.48	0.59	0.76	0.66	0.76	0.69	0.76	0.81	0.84	0.89	0.71	0.76	0.65	**0.89**
Precision	0.84	0.90	0.80	0.82	0.77	0.71	0.79	0.74	0.77	0.84	0.83	0.86	0.74	0.77	0.69	**0.90**
F1 Score	0.87	0.91	0.85	0.83	0.80	0.78	0.84	0.79	0.82	0.82	0.83	0.83	0.79	0.79	0.74	**0.91**

TABLE 8
BEST CROSS-VALIDATION RESULTS FOR SINGLE MODALITY BASED STRESS DETECTION

Modality	**Best Classifier**	**Best Bias Rectifier**	**Performance measures**					**Best Feature Combination**
			Accuracy (%)	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**	
EDA	*k*NN	SMOTE	90.91	0.92	0.89	0.90	0.91	quant_25 (filtered_EDA, diff_EDA); ctEDA (detrended_EDA); quant_50 (diff_EDA)
ECG	*k*NN	SMOTE	88.03	0.88	0.87	0.85	0.87	LF/HF, LF, HF, mRR, medRR, mHR, SDRR, RMSSD, RR50
ST	QDA	SMOTE	88.79	0.92	0.87	0.79	0.85	mTEMP, maxTEMP, sdTEMP, ranTEMP (finger_TEMP); minTEMP, ranTEMP (wrist_TEMP) ; mTEMP, ranTEMP (diff_TEMP)

(wrist_TEMP), minTEMP (diff_TEMP). However, what is more interesting is that a QDA-*k*NN ensemble classifier system using only EDA and ST modalities yielded an accuracy of 95.45%. The best feature combination for QDA in this ensemble included 10 features viz., quant_50, quant_90 (detrended_EDA), ctEDA, quant_10 (diff_EDA); sdTEMP, ranTEMP (finger_TEMP, wrist_TEMP), mTEMP, minTEMP (diff_TEMP) and *k*NN included 9 features (cvEDA (filtered_EDA), cvEDA, ctEDA, quant_50 (detrended_EDA), cvEDA, quant_25 (diff_EDA); ranTEMP(finger_TEMP, wrist_TEMP), minTEMP (diff_TEMP)). These results indicate that different physiological parameters convey different information about stress and they complement each other when put together there by increasing the stress detection accuracy. Nevertheless, the gains due to addition of a modality (in this case, ECG) are sometimes modest. Hence, it is always advisable to look for an optimum configuration rather than just combining multiple modalities for improved accuracy. In the case of a wearable for "everyday life application", a minimal sensor setup involving only EDA and ST would be most suitable.

TABLE 9
MULTIMODAL STRESS DETECTION USING CLASSIFIER ENSEMBLES

Best Single Modal Accuracy : 90.91 % (EDA)			
Modalities	**Classifier Ensemble**	**Classifier fusion accuracy (%)**	**Increase in accuracy (%)**
EDA+ECG	**QDA, SVM, *k*NN**	**92.42**	**+1.51**
ECG+ST	**LDA, QDA, SVM, *k*NN**	**93.94**	**+3.03**
EDA+ST	**QDA, *k*NN**	**95.45**	**+4.54**
	QDA, SVM, *k*NN	94.53	+3.62
	LDA, QDA, SVM, *k*NN	91.67	+0.76
EDA+ECG+ST	**QDA, *k*NN**	**95.86**	**+4.95**
	QDA, SVM, *k*NN	92.42	+1.51

7.3 Effect of Window Size and Overlap on Classification Accuracy

The European and North American Task force on standards in HRV [48] recommends that the shortest time period for a meaningful evaluation of HRV metrics is 5 minutes. However, no such guidelines exist for the EDA and ST modalities. Hence, we investigated single and multi-modality classification of 'baseline' and 'stress' conditions using these signals with lesser window sizes as well. The objective was to determine the optimum length of data segments for EDA and ST to obtain reliable features and thereby enhance the classification performance. Feature extraction was performed on 180 s, 60 s and 30 s windows of EDA and ST. Feature extension, feature selection and classification were performed as detailed in Section 6. A summary of such a comparitive study done for EDA based stress detection is shown in Table 10.

Similar analysis was done on ST as well. Furthermore, the EDA and ST based multimodality scheme for stress detection were examined for different window sizes. These analyses were performed on all 11 classifier combinations using LDA, QDA, SVM and *k*NN in order to determine the best ensemble classifier for EDA+ST based stress detection. A summary of best results is presented in Table 11.

As can be seen, EDA gave the highest cross-validation accuracy for data segments of 60 s duration while ST performed best with 30 s windows. Both modalities outperformed the respective 300 s equivalents with EDA yielding a 2.22% increase and ST, a 3.19% increase in the cross-validation accuracy. Thus, it can be concluded that for the application discussed in this study, the optimum data length for EDA is 60 s, while for ST it is 30 s.

Even with reduced window sizes, EDA emerged as the best single modality for stress detection with an increased accuracy of 93.13%. The *k*NN classifier algorithm yielded this accuracy with the best feature combination consisting of 6 features, majority of which are quantile measures thereby reducing subjectivity. Furthermore, an ensemble classifier obtained by fusing decisions of QDA, SVM and *k*NN through majority voting yielded the highest cross-validation accuray of 97.13% for the EDA+ST modality employing 60 s windows. Thus, it can be concluded that EDA and ST together can more reliably detect stress with data segments of 60 s duration.

TABLE 10
EFFECT OF WINDOW SIZE OF EDA ON CROSS-VALIDATION ACCURACY WHILE EMPLOYING DIFFERENT CLASS IMBALANCE RECTIFICATION SCHEMES

Window Size	Original				Undersampling				Oversampling				SMOTE			
	LDA	QDA	SVM	*k*NN	LDA	QDA	SVM	*k*NN	LDA	QDA	SVM	*k*NN	LDA	QDA	SVM	*k*NN
	cross-validation accuracy (%)															
30 s	68.75	71.88	74.72	83.46	72.64	71.63	71.49	81.38	67.66	68.81	80.59	82.80	65.70	70.23	69.08	**85.45**
60 s	71.45	68.86	73.18	72.78	66.31	68.98	71.12	77.87	70.46	69.44	82.21	86.41	68.93	69.05	66.71	**93.13**
180 s	65.00	66.11	68.33	68.33	68.64	72.03	70.34	**87.97**	65.70	69.42	84.56	71.90	66.94	71.07	65.29	71.49

TABLE 11
BEST CROSS-VALIDATION RESULTS FOR EDA, ST AND EDA+ST MODALITIES WITH REDUCED WINDOW SIZES

	Best Data Length	Best Classifier	Best Bias Rectifier	Accuracy (%)	Best Feature Combination
EDA	60 s	*k*NN	SMOTE	93.13	quant_25, quant_50 (filtered_EDA); ctEDA, cvEDA, quant_50 (detrended_EDA); quant_50 (diff_EDA)
ST	30 s	*k*NN	UNDER SAMPLE	91.98	maxTEMP (finger_TEMP); mTEMP, minTEMP (wrist_TEMP) ; minTEMP (diff_TEMP)
EDA + ST	60 s	QDA, SVM, *k*NN		97.13	**QDA [**quant_10, quant_25 (filtered_EDA), quant_25 (detrended_EDA, diff_EDA); mTEMP (diff_TEMP)]. **SVM [**quant_10 (filtered_EDA, diff_EDA); mTEMP, minTEMP, maxTEMP (wrist_TEMP), maxTEMP (diff_TEMP)]. ***k*NN [**quant_50, ranEDA (filtered_EDA), quant_25, quant_50 (detrended_EDA), varEDA (diff_EDA); maxTEMP (finger_TEMP), maxTEMP, minTEMP (wrist_TEMP)]

In order to assess the generalizability of this best performing ensemble classifier model, we used the model for predicting stress on a completely new dataset with 10 subjects. This dataset was not used either during the testing or training of the developed model, thereby making it entirely new to the model. The class imbalance in these test datasets were corrected before classification. The test results are summarized in Table 12. S1 to S10 indicate subject IDs. As can be seen, while tested on 10 unseen users, the model yielded an average accuracy of only 93.28%. The worst performance, viz., 83.33%., was obtained for subject S1. The best performance was reached for subjects S4 and S6. The standard deviation between classification performances on different users was within 5.32%. Thus it can be concluded that whether a model performs well on a new, unseen user may depend on the similarity of the new subject to previous subjects.

TABLE 12
PERFORMANCE OF THE BEST CLASSIFIER MODEL ON A NEW DATASET

Modality : EDA+ST, Classifier Ensemble : QDA, SVM and kNN					
Subject ID	**Accuracy (%)**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
S1	83.33	0.67	1.00	1.00	0.80
S2	86.61	0.88	0.86	0.86	0.87
S3	96.67	0.93	1.00	1.00	0.97
S4	100.0	1.00	1.00	1.00	1.00
S5	97.37	0.95	1.00	1.00	0.97
S6	100.0	1.00	1.00	1.00	1.00
S7	90.00	1.00	0.80	0.83	0.91
S8	91.67	0.83	1.00	1.00	0.91
S9	91.67	1.00	0.83	0.86	0.92
S10	95.45	0.91	1.00	1.00	0.95

The best performing mutimodal schemes viz., EDA+ECG+ST with 300 s data segments and EDA+ST with 60 s data segments were further investigated concerning the effectiveness in employing percentage of adjacent window overlap for feature extraction . The objective was to generate more instances with lowest loss of information which may enhance the classification accuracy and status update rate. Accordingly, 25%, 50%, 75% and 90% overlapping windows were considered for feature extraction in both cases. Table 13 illustrates the classification results. As such, window overlapping did not enhance the classification accuracy significantly for either schemes. However, overlapping has the potential to update the output incrementally and more efficiently than fixed window segments.

7.4 Discriminating Psychological and Physical Stress

Based on the analyses above, we can say that psychological stress can be detected reliably using variations in EDA and ST. However, variations in these physiological signals cannot be exclusively attributed to psychological stress. They may vary during a moderate or intense physical activity as well. Since this study aims at developing a wearable device for detecting stress in everyday life, where moderate or intense physical activities are inevitable, it was essential to see if these modalities can distinguish responses to a physical stress from psychological stress. Although there are theories which hypothesise that human brain is capable of categorizing different stressors and choosing context-specific, varyingly intense neuronal response pathways to reinstate the balance [57], these theories are not universally accepted [58].

TABLE 13
EFFECT OF WINDOW OVERLAP ON CLASSIFICATION ACCURACY

		Classification accuracy (%)				
Modality	Window Size	0%	25%	50%	75%	90%
EDA+ST	60 s	97.13	97.16	97.22	97.18	97.34
EDA+ECG+ST	300 s	95.45	95.42	95.46	95.29	95.63

Hence, we chose to investigated this hypothesis by using EDA and ST modalities, which have so far emerged as best indicators of psychological stress, to see if they can be potential physiological measures for classifying 'physical' and 'psychological' stressors. Physical activity described in section 4.4 was used as the physical stressor. The analyses utilized 24 datasets (12 physical and 12 psychological) derived from 12 subjects (4 females and 8 males, age: 21.5 ± 1.7 years, mean ± SD) who volunteered to participate in both physical and psychological stress tasks. As a person sweats more, right after finishing exercising than during exercise, recovery phase features might contain more useful

information for distinguishing physical stress from psychological stress using EDA and ST. Hence, both single and multimodal schemes were analyzed when taking features from different sessions of physical and psychological stress experiments viz., 'stress', 'recovery' and a combination of 'stress and recovery', as classifier inputs. Feature extraction was done for non-overlapping data segments of 1-minute duration. Oversampling was employed to rectify the class imbalance. Feature selection and classification were performed for single and multimodal schemes as mentioned in Section 6. The best classification results are summarized in Table 14.

EDA represents the best single modality to distinguish 'physical' and 'psychological' stressors with the highest accuracy of 94.41% . This was obtained using LDA while taking features of 'stress' phase alone as classifier inputs. The best feature combination included varEDA, cvEDA, quant_90 (filtered_EDA) ; quant_75 (detrended_EDA) and cvEDA (diff_EDA). The results matched with that of a similar study, performed on slightly different datasets, reported in a previous publication from the authors [44]. However, a multimodal scheme employing EDA and ST could classify these stressors with a slightly higher accuracy of 96.53%. This requires an LDA-*k*NN ensemble which takes concatenated stress and recovery phase features as input. While including recovery phase features did not increase the classification accuracy with EDA, it did improve the accuracy with ST and EDA+ST. Hence, it can be concluded that the recovery phase features of ST contain useful information for distinguishing mental stress and physical stress. The results support the hypothesis that sympathetic activation is stressor-specific and can be distinguished using physiological measures.

TABLE 14
BEST CROSS-VALIDATION RESULTS IN CLASSIFYING 'PHYSICAL' AND 'PSYCHOLOGICAL' STRESSORS

	Cross-validation Accuracy (%)			
Modality	**Stress**	**Recovery**	**Stress + Recovery**	**Best Classifier**
EDA	**94.41**	88.36	89.74	LDA
ST	89.29	90.21	**92.27**	*k*NN
EDA+ST	92.44	91.17	**96.53**	LDA, *k*NN

8 Conclusions

Although there are reported literature employing behavioural data (e.g., keyboard use, mouse use, posture, facial expressions, speech etc.), performance data (e.g., logical thinking, attention and working memory) and contextual data (e.g., Calendar, GPS, Ambient sound etc.) for work stress detection, neither of them could deliver high classification results as with physiological signals. The most accurate work stress detection schemes reported in the literature are based on physiological signals, thereby indicating that physiological signal based stress detection is much more accomplished than other modalities [59]. This does not imply that behavioural, contextual or performance data lack the potential to accurately detect stress, as results of the literature prove they do, but that there is still much work to do in this area. Therefore, the current state of the art work stress detection schemes revolves around multimodal physiological signal based multi-level stress detection.

A physiological parameter based scheme for the detection of psychological stress is presented in this paper. Single and multimodal schemes using EDA, ECG and ST were analyzed. Four classifiers viz., LDA, QDA, SVM and kNN were investigated to identify the best single modality based stress detection. Further, a fusion framework that combined the benefits of feature fusion and decision fusion was employed to generate classifier ensembles for multimodal stress detection schemes. Accordingly, all 11 classifier combinations using LDA, QDA, SVM and kNN were evaluated to determine the best ensemble for all possible multimodal schemes viz., EDA+ECG, ECG+ST, EDA+ST and EDA+ECG+ST. All analyses were performed after correcting the class imbalance that existed in the generated dataset. Since the performance of classifiers is also influenced by the class imbalance rectification technique employed, 3 class imbalance correction techniques viz., undersampling, oversampling and SMOTE were investigated in order to determine the best scheme.

The European and North American Task force on standards in HRV recommends that the shortest time period for a meaningful evaluation of HRV metrics is 5 minutes. However, no such guidelines exist for the EDA and ST modalities. Therefore, the current state of the art stress detection schemes does not exhibit any uniformity concerning the length of data segments used for analyses. As a result, overlapping and non-overlapping data segments of varying durations had been reported in the existing literature. In this study, we analysed the classification performance for varying window sizes of EDA and ST data, viz.,300 s 180 s, 60 s and 30 s, and observed that the window size do have an influence on classification accuracy. For the application discussed in the study, the optimum data length for EDA was found to be 60 s, while for ST it was 30 s. Furthermore, a comparative study was performed with 25%, 50%, 75% and 90% overlapping data segments and it was observed that overlapping did not enhance the performance of the classifiers significantly.

Although, existing literature has established that combining different physiological signals viz., EDA, EMG, EEG, ECG, ST etc., improves classification performance, which of these signals will provide the most useful information has not been established yet. To the best knowledge of the authors, there is no existing literature that analyse all possible combinations of these signals to identify the best combination for any application. In this study, we attempt to identify the best physiological signal combination that can further be incorporated into a wearable for work stress detection. As we are aiming at a wearable "for everyday application", a minimal sensor set up is desired for comfort reasons and therefore, incorporating all aforementioned signals, hoping to get the highest accuracy, is not practically plausible. Also, it is difficult to incorporate physiological signals like EEG or EMG to a sensor of wearable form factor. Hence, we chose three signals, viz., EDA, ECG and ST, which could be easily measured from hands or peripheral body parts and objectively analysed the effect of addition of each of these physiological signals to the stress detection accuracy so as to determine the best scheme for a wearable scenario. It was observed that while a multimodal scheme employing all three physiological signals gave the highest accuracy of 95.86%, a combination of EDA and ST could

detect 'stress' with a close enough cross-validation accuracy of 95.45% with 300 s windows. Thus, it is always sensible to look for an optimum configuration of modalities rather than just combining multiple modalities for improved accuracy as sometimes the gains are very modest. This approach is especially important while developing a minimal sensor set up of a wearable form factor "for everyday application".

EDA, ST and EDA+ST modalities outperformed their 300 s counterparts when data segments of reduced window size were used. EDA emerged as the best single modality that can distinguish psychological stress from baseline with a cross-validation accuracy of 93.13% with 60 s data segments. To the best of our knowledge, this is the highest of all single modality based stress detection accuracies reported in the literature. The best performance of EDA can be attributed to the fact that unlike other physiological signals, it is not influenced by the parasympathetic branch of ANS. A classifier ensemble formed using QDA, SVM and *k*NN yielded the highest accuracy of 97.13% for the EDA+ST modality employing data segments of 60 s duration. This accuracy is at par with the highest multimodal accruracy reported in the literature with the entailed advantage of using only EDA and ST modalities. However, it has to be noted that these accuracies were achieved for a binary classification of the existence of stress. As the current state of the art is focused more on predicting multiple levels of stress (e.g., low, medium and high stress levels), extending the developed classification framework for multilevel stress detection is the logical next step.

It was observed that majority of the EDA features selected, either for EDA based stress classification or in multimodal schemes involving EDA, were quantile measures. Interestingly, the mean of EDA (mEDA), which would probably be one of the first features one would extract, was not selected at all. On the other hand, the stress detection schemes involving temperature chose some of the most straightforwardly calculated time-domain features like mean, minimum, maximum, range or SD. The EDA+ST modality also proved to be highly reliable in distinguishing physical stress from psychological stress.

The ensemble classifier systems used in this study for discriminating 'stress' from 'baseline' has entailed advantages. The effect of factors like data distribution changes due to data collection on different days and the inter and intra-individual variability of the features that discriminate the affective state on classification performance can be minimized to an extent using classifier ensembles. Since both EDA and ST can be comfortably measured from wrist or finger, the scheme is remarkably suitable for an unobtrusive wearable system for "everyday life application" which can garner stressful phases over a day and can give a feedback to the user. Such a device can help the user in enhancing productivity at work by employing holidays and breaks more effectively. This study is one step towards that goal. Since the psychological stressors that we employed combined multiple components of stress like mental work load, time pressure, performance pressure and social evaluative threat which are relevant and relatable to real-life work stress scenario, we expect similar results in future long-term experiments in an office setting as well.

This study, however, is not without its own limitations. One important limitation is the small sample size. Further, the stress experiments were conducted in a laboratory setting with constant environmental conditions, in particular with a constant room temperature and with restrictions on movement. However, these restrictions helped us to generate proper, artifact-free datasets that allowed a fair comparison of different physiological measures and draw conclusions regarding the best physiological indicators of stress. An ambient temperature sensor and an activity sensor will help to overcome these limitations to an extent while aiming at long-term studies in real-world settings. The stress recognition accuracy achieved from laboratory experiments under standardized settings are promising. However, replicating this accuracy for a larger, diverse data set in a less standardized real-life scenario is an important area of future work. Further improvements are required in the firmware and harware to develop a wearable stress recognition system that reliably detects stress in real-life and gives feedback to the user.

References

[1] G. P. Chrousos, "Stress and Disorders of the Stress System," *Nat. Rev. Endocrinology,*, vol. 5, no. 7, pp. 374–381, Jul. 2009.

[2] S. Hart and L. Staveland, "Development of a Multi-dimensional Workload Rating Scale: Results of Empirical and Theoretical Research. In, PA Hancock and N Meshkati," Human Mental Workload, 1988.

[3] S. M. Monroe, "Modern Approaches to Conceptualizing and Measuring Human Life Stress," *Annu. Rev. Clin. Psychol.*, vol. 4, no. 1, pp. 33–52, 2008.

[4] C. Kirschbaum, J. C. Prussner, A. A. Stone, I. Federenko, J. Gaab, D. Lintz, N. Schommer, and D. H. Hellhammer, "Persistent High Cortisol Responses to Repeated Psychological Stress in a Subpopulation of Healthy Men," *Psychosom. Med.*, vol. 57, no. 5, pp. 468–474, 1995.

[5] M. C. Morris, U. Rao, L. Wang, and J. Garber, "Cortisol Reactivity to Experimentally Manipulated Psychosocial Stress in Young Adults at Varied Risk for Depression," *Depress. Anxiety*, vol. 31, no. 1, pp. 44–52, 2014.

[6] W. Boucsein, *Electrodermal Activity*. Springer Science & Business Media, 2012.

[7] R. S. Lazarus and E. M. Opton, "The Study of Psychological Stress: A Summary of Theoretical Formulations and Experimental Findings," in *Anxiety and Behavior*, C. D. Spielberger, Ed. New York : Academic Press, 1966, ch. 10, pp. 225–261.

[8] T. Chandola, A. Heraclides, and M. Kumari, "Psychophysiological biomarkers of workplace stressors," *Neurosci. Biobehav. Rev.*, vol. 35, no. 1, pp. 51–57, 2010.

[9] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investigation*, vol. 15, no. 3, p. 235, 2018.

[10] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[11] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating Stress from Cognitive Load Using A Wearable EDA Device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, 2010.

[12] N. Keshan, P. Parimi, and I. Bichindaritz, "Machine Learning for Stress Detection from ECG Signals in Automobile Drivers," in *2015 IEEE Int. Conf. Big Data*, 2015, pp. 2661–2669.

[13] P. Karthikeyan, M. Murugappan, and Y. Sazali, "Descriptive Analysis of Skin Temperature Variability of Sympathetic Nervous System Activity in Stress," *J. Phys. Ther. Sci.*, vol. 24, no. 12, pp. 1341–1344, 2012.

[14] G. Jun and K. Smitha, "EEG Based Stress Level Identification," in *2016 IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2016, pp. 3270–3274.

[15] J. A. Healey and R. W. Picard, "Detecting Stress During Real-world Driving Tasks Using Physiological Sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, 2005.

[16] C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis, "An Integrated Telemedicine Platform for the Assessment of Affective Physiological States," *Diagn. Pathol.*, vol. 1, no. 1, p. 16, 2006.

[17] J. Zhai and A. Barreto, "Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables," in *28th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS'06)*, 2006, pp. 1355–1358.

[18] M. H. Lee, G. Yang, H. K. Lee, and S. Bang, "Development Stress Monitoring System Based on Personal Digital Assistant (PDA)," in *26th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS'04)*, vol. 1, 2004, pp. 2364–2367.

[19] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A Real-time Human Stress Monitoring System Using Dynamic Bayesian Network," in *2005 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR) Workshops*, 2005, pp. 70–77.

[20] V. Engert, A. Merla, J. A. Grant, D. Cardone, A. Tusche, and T. Singer, "Exploring the Use of Thermal Infrared Imaging in Human Stress Research," *PLOS ONE*, vol. 9, no. 3, pp. 1–11, Mar. 2014.

[21] T. Chen, P. Yuen, M. Richardson, G. Liu, and Z. She, "Detection of Psychological Stress Using a Hyperspectral Imaging Technique," *IEEE Trans. Affective Computing*, vol. 5, no. 4, pp. 391–405, 2014.

[22] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[23] A. Savran, K. Ciftci, G. Chanel, J. Mota, L. Hong Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut, "Emotion detection in the loop from brain signals and facial images," 2006.

[24] P. J. Lang, "International affective picture system (iaps): Affective ratings of pictures and instruction manual," *Technical report*, 2005.

[25] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Tran. Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[26] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[27] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 291–298.

[28] R. S. Lazarus, *Stress and emotion: A new synthesis*. Springer Publishing Company, 2006.

[29] B. L. Seaward, *Managing Stress*. Jones & Bartlett Publishers, 2013.

[30] J. Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel, "Influence of Mental Stress on Heart Rate and Heart Rate Variability," in *4th European Conf. Int. Federation for Medical and Biological Engineering (ECIFMBE'08)*, 2009, pp. 1366–1369.

[31] C. H. Vinkers, R. Penning, J. Hellhammer, J. C. Verster, J. H. Klaessens, B. Olivier, and C. J. Kalkman, "The Effect of Stress on Core and Peripheral Body Temperature in Humans," *Stress*, vol. 16, no. 5, pp. 520–530, 2013.

[32] M. Hamer, "Stroop Color-Word Test," in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds. New York, NY: Springer, 2013, pp. 1916–1917.

[33] J. Daniel, "Performance in an interference test and some changes in the vegetative functions." *Studia Psychologica*, 1969.

[34] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert *et al.*, "The Montreal Imaging Stress Task: Using Functional Imaging to Investigate the Effects of Perceiving and Processing Psychosocial Stress in the Human Brain," *J. Psychiatry Neurosci.*, vol. 30, no. 5, pp. 319–325, 2005.

[35] R. Li-Mei Liao and M. G. Carey, "Laboratory-induced Mental Stress, Cardiovascular Response, and Psychological Characteristics," *Rev. Cardiovasc. Med.*, vol. 16, no. 1, pp. 28–35, 2015.

[36] M. A. Birkett, "The Trier Social Stress Test Protocol for Inducing Psychological Stress," *J. Vis. Exp.*, no. 56, pp. e3238–e3238, 2011.

[37] A. Luger, P. A. Deuster, S. B. Kyle, W. T. Gallucci, L. C. Montgomery, P. W. Gold, D. L. Loriaux, and G. P. Chrousos, "Acute Hypothalamic–Pituitary–Adrenal Responses to the Stress of Treadmill Exercise," *N. Engl. J. Med.*, vol. 316, no. 21, pp. 1309–1315, 1987.

[38] U. Rimmele, R. Seiler, B. Marti, P. H. Wirtz, U. Ehlert, and M. Heinrichs, "The Level of Physical Activity Affects Adrenal and Cardiovascular Reactivity to Psychosocial Stress," *Psychoneuroendocrinology*, vol. 34, no. 2, pp. 190–198, 2009.

[39] D. Bozovic, M. Racic, and N. Ivkovic, "Salivary Cortisol Levels as A Biological Marker of Stress Reaction," *Med. Arch.*, vol. 67, no. 5, pp. 374–377, 2013.

[40] *The Soma IgA / Cortisol Test With the SOMA Cube Reader*. SOMA Bioscience, Howbery Park, Wallingford, OX10 8BA. UK, 2016.

[41] *Saliva Collection and Handling Advice*, 3rd ed. SalivaBio, Salimetrics, Carlsbad, CA, 2011.

[42] C. Kirschbaum and D. H. Hellhammer, "Salivary Cortisol in Psychobiological Research: An Overview," *Neuropsychobiology*, vol. 22, no. 3, pp. 150–169, 1989.

[43] A. Clow, F. Hucklebridge, T. Stalder, P. Evans, and L. Thorn, "The Cortisol Awakening Response: More than A Measure of HPA Axis Function," *Neurosci. Biobehav. Rev.*, vol. 35, no. 1, pp. 97–103, 2010.

[44] A. S. Anusha, J. Jose, S. P. Preejith, J. Jayaraj, and S. Mohanasankar, "Differential Effects of Physical and Psychological Stressors on Electrodermal Activity," in *39th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS'17)*, 2017. [Accepted for Publication].

[45] T. Deltombe, P. Hanson, J. Jamart, and M. Clérin, "The influence of skin temperature on latency and amplitude of the sympathetic skin response in normal subjects," *Muscle & nerve*, vol. 21, no. 1, pp. 34–39, 1998.

[46] "Methods for Tracing Physiological, Neurological and Other Concomitants of Cognitive Processes," in *A handbook of process tracing methods for decision research: A critical review and user's guide*, M. Schulte-Mecklenbeck, A. Kühberger, and R. Ranyard, Eds. Psychology Press, 2011.

[47] J. Pan and W. J. Tompkins, "A Real-time QRS Detection Algorithm," *IEEE Trans. Biomed. Eng.*, no. 3, pp. 230–236, 1985.

[48] A. J. Camm, M. Malik, J. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, R. E. Kleiger *et al.*, "Heart Rate Variability. standards of Measurement, Physiological Interpretation, and Clinical Use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, 1996.

[49] A. E. Draghici and J. A. Taylor, "The Physiological Basis and Measurement of Heart Rate Variability in Humans," *J. Physiol. Anthropol.*, vol. 35, no. 1, p. 22, 2016.

[50] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *J. Artif. Intell. Research*, vol. 16, pp. 321–357, 2002.

[52] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[53] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification," *IETE Tech. Rev.*, vol. 27, no. 4, pp. 293–307, 2010.

[54] R. Polikar, "Ensemble Based Systems in Decision Making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[55] U. Rimmele, B. C. Zellweger, B. Marti, R. Seiler, C. Mohiyeddini, U. Ehlert, and M. Heinrichs, "Trained Men Show Lower Cortisol, Heart Rate and Psychological Responses to Psychosocial Stress Compared with Untrained Men," *Psychoneuroendocrinology*, vol. 32, no. 6, pp. 627–635, 2007.

[56] R. L. Burr, W. Chen, M. J. Cowan, M. M. Heitkemper, and S. A. Motzer, "Logit50: A nonlinear transformation of pnn50 with improved statistical properties," *J. Electrocardiol.*, vol. 36, no. 1, pp. 41–52, 2003.

[57] T. Steckler, N. H. Kalin, and J. H. M. Reul, *Handbook of Stress and the Brain Part 1: The Neurobiology of Stress*. Elsevier, 2005, vol. 15.

[58] D. S. Goldstein and B. McEwen, "Allostasis, Homeostats, and the Nature of Stress," *Stress*, vol. 5, no. 1, pp. 55–58, 2002.

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

APPENDIX A
LIST OF ABBREVIATIONS AND ACRONYMS

ANFIS	Adaptive Network based Fuzzy Inference System
ANOVA	ANalysis Of VAriance
ANS	Autonomic Nervous System
BMI	Body Mass Index
BVP	Blood Volume Pulse
CNS	Central Nervous System
CV	Coefficient of Variation
DBN	Dynamic Bayesian Network
DEAP	Dataset for Emotion Analysis using Physiological signals
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalogram
EMG	Electromyogram
fNIRS	Functional Near-Infrared Spectroscopy
GRNN	Generalized Regression Neural Network
HR	Heart Rate
IAPS	International Affective Picture System
*k*NN	*k* Nearest Neighbors
LDA	Linear Discriminant Analysis
LFD	Lateral Flow Device
LR	Logistic Regression
MLP	Multilayer Perceptron
MPA	Multivariate Pattern Analyses
NB	Naive Bayes
NCC	Nearest Class Center
PD	Pupil Diameter
PNN	Probabilistic Neural Network
PPG	Photoplethysmogram
QDA	Quadratic Discriminant Analysis
RSP	Respiration
SAM	Stress Assessment Manikins
SD	Standard Deviation
SEM	Standard Error of the Mean
ST	Skin Temperature
SVM	Support Vector Machine
TSST	Trier Social Stress Test
WHO	World Health Organisation