Requirements for a Reference Dataset for Multimodal Human Stress Detection

Bhargavi Mahesh
Dept. of Computer Science
Bonn-Rhein-Sieg University of Applied Sciences
Sankt Augustin, Germany
bhargavi.mahesh@smail.inf.h-brs.de

Teena Hassan
Electronic Imaging Department
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Germany
teena.hassan@iis.fraunhofer.de

Erwin Prassler
Dept. of Computer Science
Bonn-Rhein-Sieg University of Applied Sciences
Sankt Augustin, Germany
erwin.prassler@h-brs.de

Jens-Uwe Garbas
Electronic Imaging Department
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Germany
jens.garbas@iis.fraunhofer.de

***Abstract*—Stress is necessary for optimal performance and functioning in daily life. However, when stress exceeds person-specific coping levels, then it begins to negatively impact health and productivity. An automatic stress monitoring system that tracks stress levels based on physical and physiological parameters, can assist the user in maintaining stress within healthy limits. In order to build such a system, we need to develop and test various algorithms on a reference dataset consisting of multimodal stress responses. Such a reference dataset should fulfil requirements derived from results and practices of clinical and empirical research. This paper proposes a set of such requirements to support the establishment of a reference dataset for multimodal human stress detection. The requirements cover person-dependent and technical aspects such as selection of sample population, choice of stress stimuli, inclusion of multiple stress modalities, selection of annotation methods, and selection of data acquisition devices. Existing publicly available stress datasets were evaluated based on criteria derived from the proposed requirements. It was found that none of these datasets completely fulfilled the requirements. Therefore, efforts should be made in the future to establish a reference dataset, satisfying the specified requirements, in order to ensure comparability and reliability of results.**

***Index Terms*—stress detection, reference dataset, requirements analysis**

I. INTRODUCTION

Stress is identified as one of the top ten social determinants of health disparities [1]. Organisations such as the World Health Organisation [2], [3], American Psychological Association [1] and Occupational safety and health administration [4] are raising awareness about negative impact of stress on health, and its associated costs to society. A conservative estimate of the cost of work-related stress in the European countries alone was reported to be €20 billion in 2002 [4].

The term 'stress' was introduced by Hans Selye through the General Adaptation Syndrome concept which stated that the energy of a person to adapt to an alarming situation is finite and its magnitude is person-dependent [5]. This concept has evolved over the years from stress being a mere response to a perceived threat to being a health condition. Manifestation of stress begins with cognitive appraisal of a threat or a challenge in an attempt to overcome it and regain the stable state of homoeostasis [5]. Therefore, stress is essential to survival. However, frequent, intense, and consistent exposure to stress results in suppressed immunity and prolongs recovery [4]. Therefore, in order to limit such exposures and facilitate faster stress recovery, monitoring of stress levels is essential.

Biologically, the sympathetic nervous system (SNS) and the Hypothalamus-Pituitary-Adrenal (HPA) axis play a major role in eliciting stress responses [6]–[8]. The HPA axis transmits neural and chemical messages in preparation to deal with the threat. Neural messages are quick and short-lived. But, chemical messages (hormones), which are transmitted through blood, last longer in the body and cause a lasting impact. Cortisol, often regarded as the biomarker for stress, acts as a feedback to suppress HPA activity [9]. However, cortisol, which is transmitted through blood, also stimulates chemical activity that results in irreversible biological changes [8]. Elevated levels of such chemicals are known to contribute to cardiovascular diseases [10], which are responsible for a significant number of deaths every year [11].

An automatic stress[1] detection system can identify stress build-up and help in managing stress levels. With the growing trend in the use of wearable sensing devices and quantified self-applications, such stress detection systems could play a major role in real-time stress monitoring while performing daily activities. Two major hurdles in building a stress detection system are the lack of a common notion of ground truth for stress and the insufficient consideration of inter- and intrapersonal variability. Algorithms and computation capabilities are evolving towards taking such variations into account. In order to build robust and reliable stress detection systems, it is necessary to develop, validate and benchmark the

[1] In this work, we focus on acute stress.

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

different systems on a reference dataset, which fulfils certain requirements[2].

Several researchers have adopted various data acquisition methods for early detection of stress and have built several stress detection models [12], [13]. These approaches have considered different stress stimuli, used dissimilar sensors, and captured diverse stress response modalities. The differences in data collection approaches of researchers have resulted in inconsistency among the datasets compiled. This inconsistency has further hampered the development of standard stress detection algorithms. Since modelling stress requires complex, multivariate data analysis, it is essential that the dataset includes the important factors that influence the manifestation and measurement of stress response. These factors can be broadly classified with respect to an individual into internal (age, gender, health condition, etc.) and external (stress stimuli, environment, sensors, etc.).

To summarise, the inconsistency among the existing datasets and the necessity to consider the multitude of factors influencing stress responses create a need for establishing a reference dataset for multimodal human stress detection. The objective of this work is to examine clinical and technical literature and to derive a requirements specification for such a reference dataset. Based on the requirements obtained from the results and practices of clinical and empirical research, a set of evaluation criteria are derived and the publicly available datasets are evaluated against it.

In Section II, the different stress assessment methods and the publicly available stress datasets are discussed. In Section III, the requirements for a reference dataset for multimodal stress recognition are specified. This is the major contribution and outcome of this work. Section IV then evaluates the stress datasets based on the criteria derived from the requirements. Section V concludes the paper.

II. RELATED WORK

A. Stress Assessment Methods

Stress research has gained the interest of psychologists, clinicians and computer scientists for over four decades. Researchers are working towards understanding, modelling and classifying stress response patterns. However, there is no consensus on the actual measure of stress. Since there are multiple physiological pathways for the manifestation of stress, it opens up a wide range of modalities through which stress could be measured.

The stress response modalities could be categorised into physiological, psychological and behavioural. Physiological modalities include chemical and physical changes in the body. Clinicians consider increased levels of cortisol and catecholamines in blood plasma as indicators of stress [8]. However, these hormones involve highly invasive sampling process and require medical expertise, thereby, not making them a viable option for monitoring stress on a daily basis. Clinical studies have shown cortisol level variations correlating with Heart Rate Variability (HRV) which can be obtained from electrocardiogram (ECG) [14], [15] and Skin Conductance (SC) which can be obtained from galvanic skin responses (GSR) or electrodermal activity (EDA) [16]. Furthermore, pupil dilation [17], a decrease in skin temperature [18], and an increase in perinasal perspiration [19] are found to be associated with the stressful stimulus.

Psychological studies assess stress through instantaneous self-reporting of feelings using scales such as Perceived Stress Scale [20] and State-Trait Anxiety Inventory [21]. These scales are often used for annotation of stress levels in the collected data. Although self-reports are not regarded as reliable due to their subjective biases and lack of care in reporting, they reflect the perception of one's own psychological state. Together with *personality-related information*, instantaneous self-reports could provide further information on the individual's locus of control, and in turn, on perception of stress [22]. Since physical activity [23], medication and pre-existing diseases [24], and food intake [25] alter stress responses, questionnaires that obtain such information are necessary. Furthermore, information about the task-dependent cognitive, sensory and motor skills of an individual should be collected, in order to support the expected interpersonal variations in the response to identical stimuli.

In addition to physiological and psychological measures, other objective stress assessment techniques are also used. These are categorised as behavioural methods and include observation of changes in facial expressions, body postures, and interaction with computer hardware and devices [26]. However, more research is required to clinically establish their relationship with stress.

Sharma et al. [12] surveyed various stress detection and classification techniques and attempted to empirically rank the stress modalities based on their correlation with stress levels and frequency of usage in stress detection methods. The authors remarked about the ability of fuzzy algorithms to account for the uncertainties in data annotation or stress responses; for example, random heart rate variations. However, the authors did not emphasise on the variety and quality of data collected. Alberdi et al. [13] reviewed the techniques for stress detection in office environments. They highlight the importance of multimodal data quality. However, detailed analysis to derive dataset requirements based on practices in clinical and empirical research has not been performed so far. Our work contributes towards filling this research gap.

B. Public Human Stress Datasets

Five public datasets have been considered for evaluation. The description of each dataset is provided in this Section. Later on, in Section IV, we compare these datasets to examine the extent to which they fulfil the requirements for a reference dataset.

1) Drivedb: Healey and Picard [27] compiled the Drivedb dataset[3] and it is one of the earliest works in automatic

[2]Such a reference dataset need not necessarily be a single, generic dataset that covers all stress situations or use cases.

[3]can be accessed from https://www.physionet.org/pn3/drivedb/

TABLE I
OVERVIEW OF EXISTING PUBLIC DATASETS FOR HUMAN STRESS DETECTION

Dataset	#Subjects	Stress inducing methods	Physiological and behavioral modalities available	Data annotations available
Drivedb	9	Driving tasks with varying cognitive load	EMG, GSR, ECG, HR, and respiration	Stressor-based markers
SWELL-KW	25	Knowledge work with time pressure & email interruptions	ECG, EDA, computer logging, facial expressions, body postures	NASA-Task Load Index, Rating Scale Mental Effort, Self-Assessment-Manikin Scale (SAM), Internal Control Index, 7-point Likert scale
SUS	35	Aircraft communication training, roller coaster, free fall, doctor-patient conversation	Speech	Stressor-based
Distracted Driving Dataset	68	Simulated driving with distractions and startling event	EDA (palm & perinasal), HR, RESP, facial video, operational theater video, driving performance	Stressor-based, NASA-Task Load Index, (State-) Trait Anxiety Inventory Type A/B Personality
WESAD	15	Trier social stress test	ECG, EDA, EMG, BVP, body temperature and acceleration	Stressor-based, Positive and Negative Affect Schedule, State-Trait Anxiety Inventory (STAI), Short Stress State Questionnaire, SAM

stress detection. It is by far the most re-used public dataset [28], [29]. The dataset was collected for detecting drivers' overall stress levels. Stress responses were captured while driving on planned routes with varying cognitive load. Video of the driver was captured to manually estimate stress level based on head movements and confirm the cognitive load. They include physiological modalities such as respiration, electromyogram (EMG), electrocardiogram (ECG), heart rate (HR) and galvanic skin response (GSR), captured in an ambulatory environment. Although their initial results motivated future research in stress detection, the dataset had its own limitations such as the inability to measure a driver's response time to stress stimulus due to the unsynchronised video and sensor clocks, and the lack of information on the cognitive state of the driver through self-reports.

2) SWELL Knowledge Work (SWELL-KW) Dataset: Koldijk et al. [26] published the SWELL-KW dataset[4] which was collected for studying the stressful behaviour of knowledge workers using a context-aware pervasive system. Time pressure and interruptions were used as stressors in an office work scenario. Computer interactions, facial expressions and body postures, and physiological modalities such as ECG and SC were captured. Several subjective, self-reports of stress were collected through questionnaires for use as ground truth. The initial results obtained on this dataset demonstrated the ability to distinguish stressful from normal work conditions. However, due to unreported cognitive capabilities of participants, the authors had difficulty in isolating its influence on their actual stress level.

3) Speech Under Stress (SUS) Datasets: Hansen et al. [30] collected three datasets, namely Speech Under Stress Conditions (SUSC), Speech Under Simulated and Actual Stress[5] (SUSAS), and DERA License Plate (DLP) datasets, with the primary goal of developing robust speech processing algorithms to study the effects of stress and emotion on speech. Single-word utterances were recorded during aircraft communication and other activities that differed from activities of daily living. Apart from speech, they did not consider any other stress modality. This prevents a broader study of interpersonal differences in stress responses.

4) Distracted Driving Dataset: This dataset[6] was collected by Taamneh et al. [31] to study driving behaviours under distracting stressors such as cognitive, emotional, sensorimotor, and startling event, that often result in vehicle accidents. It includes data from 68 subjects and consists of stress response modalities such as heart rate, respiration rate, facial expressions, gaze, and EDA from palm and perinasal areas. Several questionnaires are used to obtain self-reports of task load, cognitive state, and personality type. A limitation of this dataset is the lack of HRV data, which is a major indicator of stress [15].

5) Wearable Stress and Affect Detection (WESAD) Dataset: Schmidt et al. [32] compiled this dataset[7] in order to provide high-quality multimodal data for stress and amusement state (affect) detection. Considering stress, the data is collected from 15 subjects with Trier Social Stress Test as stress stimuli. Physiological modalities such as ECG, EDA, blood volume

[4]can be accessed from http://cs.ru.nl/~skoldijk/SWELL-KW/Dataset.html

[5]can be accessed from https://catalog.ldc.upenn.edu/LDC99S78

[6]can be accessed from https://doi.org/10.17605/OSF.IO/C42CN

[7]can be accessed from https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/

pulse (BVP), EMG, respiration and body temperature were captured along with triaxial acceleration to provide contextual information. A chest-worn device, RespiBAN professional, and a wrist-worn device, Empatica E4, were used for data collection. Four self-reports have been provided along with additional notes wherever available.

III. REQUIREMENTS ANALYSIS

Whether in a laboratory or ambulatory environment, the data collection procedure must be designed to facilitate learning of reliable stress response patterns. In order to ensure reliability and validity of the study outcome using the collected data, we designed a five-part framework. Each part of the framework comprised of questions that initiated the requirements analysis process. We based the analysis on evidence obtained from the clinical and empirical studies. The following framework was used as a guideline to assist the requirements analysis for a reference dataset for multimodal human stress recognition:

I Population selection: the study outcome should be generalisable to a majority of the population.
   - Who should participate?
   - How many should participate?

II Stress stimuli selection: the study outcome should be applicable to real-life situations.
   - Which stress stimuli should be used for inducing stress in a laboratory environment?

III Stress modalities selection: the study outcome should be robust to interpersonal differences in stress responses.
   - Which modalities should be recorded?

IV Sensing device selection and configuration: the study should be practical to enable real-life implementation and reproducibility.
   - Which sensors should be used for data acquisition?
   - What details about sensors should be recorded?
   - How should sensors be configured?

V Self-reported information: the study outcome should be able to capture the influence of internal and external determinants.
   - Which self-report scales can be used for stress data annotation?
   - What prior information about participant's behaviour can influence stress response and recovery?
   - What information about the participant's activity prior to the experiment should be collected?

Based on the questions in this framework, we searched for evidence and results from clinical and empirical research. The findings from the literature search as well as our own conclusions are consolidated into five groups and described in the subsections below. The requirements derived from each of these groups are specified at the end of each subsection. The requirements are coded as "REQ-" followed by the number.

A. *Demographics and size of sample population of the dataset*

It is observed that factors such as age [7], [33] and gender [10], [34] affect stress responses. The sample population should, therefore, include equal proportions of various demographic groups based on age and gender.

Estimation of appropriate sample size that is representative of the population should be carried out in collaboration with statistician [35]. Prior information regarding the stress-related use case and the relevant statistical data should be gathered; for example, assisted daily living may require the ratio of groups of individuals suffering from depression or anxiety disorders to the healthy groups in a large community. Once the statistical parameters are known, the sample size for the experiment can be estimated using Cochran's sample size formula [36] or available sample size calculation tools.

REQ-1: Sample population size should be representative of the target population and should encompass various demographic factors.

B. *Characteristics of the stress stimuli*

Stress responses are known to be stimuli-specific [37]. The chosen stress stimuli should be relevant to day-to-day activities, or be specific to the use case for which the reference dataset is being collected. An effective stress stimulus should be capable of eliciting HPA responses. The four characteristics of stress stimuli that have been found to be essential for effectively eliciting stress responses in any person are novelty, uncontrollability, unpredictability, and socio-evaluative threat [10]. An example of effective stress stimuli is the Trier Social Stress Test [38] that is a combination of cognitive test and public speaking.

REQ-2: Stress stimuli should be relevant and effective.

C. *Multiple reliable modalities*

As described in Sec. II-A, there is no consensus on a ground truth for stress data annotation. However, due to different ways of expressing stress among individuals, it is important to capture multiple modalities. As mentioned in Section II-A, HRV [14], [15] and skin conductance (SC) [16] correlate with cortisol levels. Therefore, the datasets should include a minimum of two reliable modalities viz. electrical responses of heart (ECG) and skin (EDA).

REQ-3: Multiple modalities containing complementary information should be recorded; Electrical responses of heart (ECG) and skin (EDA) should be included.

D. *Self-reported information*

Self-reports are used by psychologists to record the instantaneous feelings experienced after the stress stimulus. Self-reports are highly subjective, unreliable, and difficult to verify. Despite this, they are useful tools to gain insights into the psychological response to the stimulus. It is recommended to use validated instantaneous self-reports such as the Perceived Stress Scale [20]. Cognitive, sensory and motor skills of an individual should be reported prior to the experiment in order to relate the perceived stress levels to the complexity of stimulus.

Interpersonal variability in stress responses is influenced by various internal factors such as age [33], gender [10], [34], personality type [22], medical condition [24], and external factors such as exercise, altered salt intake [25], biological rhythm affecting factors such as sleep, shift work, etc. [39]. Medical conditions such as chronic depression suppress the physiological stress response [40]. According to the American Psychological Association [41], the significant natural causes of stress include the individual's socio-economic status, work conditions, and social relationships. Information about these factors that could affect stress responses should be collected from participants before the experiment. This is necessary to analyse and explain the interpersonal differences in the collected stress response data. Self-reports can also be used for this purpose.

REQ-4: Multiple self-reporting methods should be used to record internal and external factors that affect stress response.

E. Data acquisition devices and software

The quality of sensors has a significant impact on the accuracy of the detection process. Although it is hard to obtain gold-standard-equivalent performance from the devices that can be employed in daily life, it is advisable to use clinically validated sensing devices to reduce their negative impact on the performance of stress detection algorithm. Devices should be configured by considering the properties of physiological signals to be recorded. Signals such as skin conductance are known to have latency in the order of seconds, between the stress stimulus and the response [42]. Therefore, care should be taken during device calibration and time-synchronisation. This is crucial to enable fusion of data from multiple sensors.

Considering the impossibility of having a noiseless sensor, it is important to specify the noise characteristics of the used devices. This is essential for choosing the appropriate noise filtering methods, for better interpretation of the stress responses, and for improving the reproducibility of experiment. Stress detection systems that take these noise characteristics into consideration would be able to generalise more robustly to data acquired using sensors that are not part of the dataset that they were trained on. Often stress response measurements need to be taken over multiple days to study intrapersonal variations. In order to avoid inconsistencies in the measurements taken on different days, it is necessary to design a device setup protocol that mentions the calibration parameters. This would help in reducing error and saving time. This would also enable the extension of reference dataset in the future by consistently calibrating the sensors to match the existing dataset characteristics.

REQ-5: Clinically validated data acquisition devices should be used and their noise characteristics should be specified; Device calibration information should be recorded in a device setup protocol.

IV. Evaluation Results

In order to verify whether any of the five publicly available stress datasets listed in Section II-B qualify as a reference dataset, we applied evaluation criteria derived from the five requirements formulated in Section III. These evaluation criteria are defined as follows:

- **Population**: This is derived from REQ-1. It tests whether a balanced representation of different **age groups** and **gender** were considered during participant selection.
- **Stimuli**: This is derived from REQ-2. It tests whether the stress stimuli used in the dataset is **relevant** for daily life, and whether the four **characteristics** of an effective stressor are met.
- **Modalities**: This is derived from REQ-3. It tests whether the dataset includes **multiple** modalities, and whether both the **reliable** modalities, namely ECG and EDA, are included.
- **Self-reports**: This is derived from REQ-4. It tests whether the dataset provides self-reports about **perception** of stress, the **internal** factors such as age, gender and personality type, and any of the **external** factors such as physical activity and food intake.
- **Data acquisition devices**: This is derived from REQ-5. It tests whether the **devices** used are specified, and whether the sensor **noise** characteristics and sensor **calibration** details are provided.

It was observed that datasets did not always publish the complete information collected while creating the dataset. For example, Drivedb dataset does not provide the instantaneous self-reports of the participants. Therefore, only the publicly accessible information about the datasets were considered during the evaluation. The evaluation criteria, as well as the evaluation results, are summarised in Table II. Some of the key results of the evaluation are listed below:

- Population: The Distracted Driving dataset has an equal proportion of male and female subjects, and includes two age cohorts. However, the other datasets did not satisfy these criteria.
- Stimuli: The stimuli considered in Drivedb, SWELL-KW, Distracted Driving, and WESAD datasets are relevant to daily living. However, the stimuli considered in SUS do not include daily life scenarios.
- Modalities: Drivedb, SWELL-KW, Distracted Driving, and WESAD datsets include multiple modalities. However, Distracted Driving and SUS datasets do not have either or both of the reliable modalities viz. ECG and EDA.
- Self-reports: SWELL-KW, Distracted Driving, and WESAD datasets provide subjective reports of personality and perceived stress. Information on smoking, intake of caffeine, and physical activity performed prior to the experiment is present in WESAD. However, the subjects in SWELL-KW were informed not to smoke or drink coffee three hours prior to the experiment. Drivedb and SUS do not provide
- Data acquisition devices: Drivedb [43], SWELL-KW [44], WESAD, and Distracted Driving [31] datasets mention the devices used for data acquisition. SUS mentions that microphones and telephone receivers were used for

data acquisition, but do not provide further details. None of the datasets provided noise and calibration information for the devices used.

It can be seen that none of the datasets satisfy the evaluation criteria completely. Consequently, they do not fulfil the requirements for a reference dataset.

TABLE II
RESULT OF EVALUATION OF PUBLIC HUMAN STRESS DATASETS

Evaluation criteria		Dataset				
		Drivedb	**SWELL-KW**	**SUS**	**Distracted Driving**	**WESAD**
Population (REQ-1)	Age groups	×	×	×	✓	×
	Gender groups	×	×	×	✓	×
Stimuli (REQ-2)	Characteristics	✓	✓	✓	✓	✓
	Relevance	✓	✓	×	✓	✓
Modalities (REQ-3)	Multiple	✓	✓	×	✓	✓
	Reliable	✓	✓	×	×	✓
Self-reports (REQ-4)	Perception	×	✓	×	✓	✓
	Age	×	×	×	✓	✓
	Gender	×	×	×	✓	✓
	Personality	×	✓	×	✓	✓
	External	×	✓	×	×	✓
Data acquisition devices (REQ-5)	Devices	✓	✓	×	✓	✓
	Noise	×	×	×	×	×
	Calibration	×	×	×	×	×

V. CONCLUSION

Stress is a highly complex and subjective phenomenon. In order to build reliable, automatic stress detection systems, it is necessary to use a reference dataset that captures all the important aspects of the stress phenomenon. This paper specifies the requirements that should be fulfilled by a reference dataset for multimodal human stress detection. The requirements are derived by reviewing clinical and technical literature and are based on clinical practices and empirical evidence. A set of evaluation criteria is defined based on the requirements, and the existing publicly available human stress datasets are evaluated against it. It was found that none of these datasets fulfil all the requirements to qualify as a reference dataset. Therefore, future efforts should aim at establishing such a reference dataset, by considering the proposed requirements. The requirements listed in this paper are not exhaustive. Future work could also focus on extending the requirements by examining, for example, use case specific as well as chronic stress-related aspects.

REFERENCES

[1] American Psychological Association, “Stress and health disparities: Contexts, mechanisms, and interventions among racial/ethnic minority and low-socioeconomic status populations,” Tech. Rep., 2018. [Online]. Available: http://www.apa.org/pi/health-disparities/resources/stress-report.pdf

[2] World Health Organisation, “Comprehensive mental health action plan 2013-2020,” pp. 1–44, 2016. [Online]. Available: https://www.who.int/mental_health/publications/action_plan/en/

[3] ——, “Guidelines for the Management of Conditions Specifically Related to Stress,” *Assessment and Management of Conditions Specifically Related to Stress: mhGAP Intervention Guide Module*, pp. 1–273, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24649518

[4] EU-OSHA, “Calculating the cost of work-related stress and psychosocial risks,” Tech. Rep., 2014. [Online]. Available: https://www.healthy-workplaces.eu/es/tools-and-resources/publications

[5] H. Selye, “Stress and the general adaptation syndrome,” *British Med. J.*, vol. 1, no. 4667, pp. 1383–1392, 1950. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6649565

[6] B. M. Kudielka, H. Hellhammer, and C. Kirschbaum, “Ten years of research with the trier social stress test–Revisited,” *Social neuroscience: Integrating biological and psychological explanations of social behavior*, no. October 2015, pp. 56–83, 2007.

[7] P. Foley and C. Kirschbaum, “Human hypothalamus-pituitary-adrenal axis responses to acute psychosocial stress in laboratory settings,” *Neuroscience and Biobehavioral Reviews*, vol. 35, no. 1, pp. 91–96, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.neubiorev.2010.01.010

[8] Z. Djuric, C. E. Bird, A. Furumoto-dawson, G. H. Rauscher, M. T. Ruffin, R. P. Stowe, K. L. Tucker, and C. M. Masi, “Biomarkers of Psychological Stress in Health Disparities Research,” *The open biomarkers journal*, pp. 7–19, 2010.

[9] S. H. Golden, G. S. Wand, S. Malhotra, I. Kamel, and K. Horton, “Reliability of hypothalamicpituitaryadrenal axis assessment methods for use in population-based studies,” *European journal of epidemiology*, vol. 26, no. 7, pp. 511–525, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21533585

[10] S. J. Lupien, F. Maheu, M. Tu, A. Fiocco, and T. E. Schramek, “The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition,” *Brain and Cognition*, vol. 65, no. 3, pp. 209–237, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278262607000322

[11] A. Steptoe and M. Kivimäki, “Stress and Cardiovascular Disease: An Update on Current Knowledge,” *Annual Review of Public Health*, vol. 34, no. 1, pp. 337–354, 2013. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-publhealth-031912-114452

[12] N. Sharma and T. Gedeon, “Objective measures, sensors and computational techniques for stress recognition and classification: A survey,” *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1287–1301, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260712001770

[13] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046415002750

[14] B. von Dawans, C. Kirschbaum, and M. Heinrichs, “The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format,” *Psychoneuroendocrinology*, vol. 36, no. 4, pp. 514–522, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.psyneuen.2010.08.004

[15] S. S. Dickerson and M. E. Kemeny, “Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research,” *Psychological bulletin*, vol. 130, no. 3, pp. 355–391, 2004.

[16] S. D. Vanderark and D. Ely, “Cortisol, biochemical, and galvanic skin responses to music stimuli of different preference values by college students in biology and music,” *Perceptual and Motor Skills*, vol. 77, no. 1, pp. 227–234, 1993. [Online]. Available: https://doi.org/10.2466/pms.1993.77.1.227

[17] E. Granholm and S. R. Steinhauer, “Pupillometric measures of cognitive and emotional processes,” *International Journal of Psychophysiology*, vol. 52, no. 1, pp. 1–6, 2004.

[18] K. A. Herborn, J. L. Graves, P. Jerem, N. P. Evans, R. Nager, D. J. McCafferty, and D. E. McKeegan, “Skin temperature reveals the intensity of acute stress,” *Physiology and Behavior*, vol. 152, pp. 225–230, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.physbeh.2015.09.032

[19] D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis, “Perinasal imaging of physiological stress and its affective potential,” *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 366–378, 2012.

[20] S. Cohen, T. Kamarck, and R. Mermelstein, “A Global Measure of Perceived Stress,” pp. 385–396, 1983.

[21] L. Laux, *Das State-Trait-Angstinventar (STAI) : theoretische Grundlagen und Handanweisung*, 1981.

[22] E. Childs, T. L. White, and H. De Wit, "Personality traits modulate emotional and physiological responses to stress," *Behavioural Pharmacology*, vol. 25, no. 5-6, pp. 493–502, 2014.

[23] 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.

[24] D. H. Hellhammer, S. Wüst, and B. M. Kudielka, "Salivary cortisol as a biomarker in stress research," *Psychoneuroendocrinology*, vol. 34, no. 2, pp. 163–171, 2009.

[25] D. S. Goldstein and I. J. Kopin, "Evolution of concepts of stress," *Stress*, vol. 10, no. 2, pp. 109–120, 2007.

[26] 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*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 291–298. [Online]. Available: http://doi.acm.org/10.1145/2663204.2663257

[27] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[28] L.-l. Chen, Y. Zhao, P.-f. Ye, J. Zhang, and J.-z. Zou, "Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers," *Expert Systems with Applications*, pp. –, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417300477

[29] N. El Haouij, J.-M. Poggi, E. Ghozi Raja, S. Sevestre-Ghalila, and M. Jaïdane, "Random forest-based approach for physiological functional variable selection: towards driver's stress level classification," 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01426752

[30] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 4, 1999, pp. 2079–2082 vol.4.

[31] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "A multimodal dataset for various forms of distracted driving," *Scientific Data*, vol. 4, p. 170110, 2017. [Online]. Available: http://www.nature.com/articles/sdata2017110

[32] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: ACM, 2018, pp. 400–408. [Online]. Available: http://doi.acm.org/10.1145/3242969.3242985

[33] B. M. Kudielka, A. Buske-Kirschbaum, D. H. Hellhammer, and C. Kirschbaum, "HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: Impact of age and gender," *Psychoneuroendocrinology*, vol. 29, no. 1, pp. 83–98, 2004.

[34] B. M. Kudielka and C. Kirschbaum, "Sex differences in HPA axis responses to stress: A review," *Biological Psychology*, vol. 69, no. 1 SPEC. ISS., pp. 113–132, 2005.

[35] B. Röhrig, J.-B. du Prel, D. Wachtlin, R. Kwiecien, and M. Blettner, "Sample Size Calculation in Clinical Trials," *Deutsches Aerzteblatt Online*, vol. 107, pp. 552–556, 2010. [Online]. Available: https://www.aerzteblatt.de/10.3238/arztebl.2010.0552

[36] W. G. Cochran, "Sampling techniques," 1963.

[37] K. Pacák and M. Palkovits, "Stressor specificity of central neuroendocrine responses: Implications for stress-related disorders," *Endocrine Reviews*, vol. 22, no. 4, pp. 502–548, 2001.

[38] C. Kirschbaum, K.-M. Pirke, and D. Hellhammer, "The 'Trier Social Stress Test' - a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, pp. 76–81, 1993.

[39] S. J. Lupien, B. S. McEwen, M. R. Gunnar, and C. Heim, "Effects of stress throughout the lifespan on the brain, behaviour and cognition," *Nature Reviews Neuroscience*, vol. 10, 2009. [Online]. Available: http://dx.doi.org/10.1038/nrn2639

[40] H. M. Burke, M. C. Davis, C. Otte, and D. C. Mohr, "Depression and cortisol responses to psychological stress: A meta-analysis," *Psychoneuroendocrinology*, vol. 30, no. 9, pp. 846–856, 2005.

[41] American Psychological Association, "Stress in America 2011," https://www.apa.org/news/press/releases/stress/2011/, 2012.

[42] C. Setz, B. Arnrich, J. Schumm, R. L. Marca, G. Tröster, and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, 2010.

[43] J. a. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, 2000.

[44] S. J. Koldijk, "Context-Aware Support for Stress Self-Management: From Theory to Practice," *Radboud Repository*, p. 144, 2016. [Online]. Available: http://repository.ubn.ru.nl/bitstream/handle/2066/93611/93611.pdf?sequence=1