Wearable Biosensor-Based Stress Detection to Understand and Improve the Quality of Interactions between Humans and Construction and Built Environments

by

Gaang Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Civil Engineering)
in the University of Michigan
2022

Doctoral Committee:

Professor SangHyun Lee, Chair
Professor Clive Rahul D'Souza
Professor Vineet Rajendra Kamat
Assistant Professor Neda Masoud
Associate Professor Carol Menassa

Gaang Lee

gaang@umich.edu

ORCID iD: 0000-0002-6341-2585

Dedication

To my wife GaRam and baby OnDaam

Acknowledgements

Firstly, I would like to thank my enthusiastic advisor, Dr. SangHyun Lee, for all his support and guidance that he has given me during my PhD study over the past five years. I also wish to acknowledge thoughtful advice from the other members of my dissertation committee, Dr. Clive Rahul D'Souza, Dr. Vineet Rajendra Kamat, Dr. Neda Masoud, and Dr. Carol Menassa. I would also like to thank the scholars who have directly influenced my study, Dr. Changbum Ryan Ahn, Dr. Byungjoo Choi, and Dr. Houtan Jebelli. I have been very fortunate to have been able to discuss my research with these great scholars.

Next, I would like to thank my colleagues. My colleagues who I spent countless hours with at the University of Michigan discussing our research in general: Dr. Daeho Kim, Dr. Meiyin Liu, Dr. Kwonsik Song, Dr. Kwangbok Jeong, Dr. Dongmin Lee, Dr. Jinwoo Kim, Neil Karr, Alan Yin, Sehwan Chung, Hoyoung Lee, Francis Baek, Juhyeon Bae, Patrick Wen, Jiu Sohn. Their assistance, cooperation, and experience were essential for the completion of my PhD study.

Next, I also would like to thank my municipal and industry partners, Ypsilanti Township, Clack East Tower Senior Apartment, Barton Malow Construction Co., and Liberty Mutual Insurance for helping me with field data collection. There help and support were critical for the demonstration of my PhD study.

Finally, I would like to thank my family for their unwavering support and patience. This dissertation would not have been completed without their support.

Table of Contents

List of Tables

List of Figures

Abstract

Ensuring human health, safety, comfort, and productivity is a key factor in the management of both construction and built environments (CBEs); human workers are the most important resource at construction sites and the operation of most built environments places the highest priority on serving people optimally. However, the "one-size-fits-all" approach, widely applied in current CBE management practices, is not effective for serving humans in CBEs because every individual has unique characteristics and thus differently interacts with CBEs even under an identical setup. Wearable biosensors have great potential to continuously and less-invasively monitor stress as the indicator of individuals' quality of experience during their daily work and lives, thereby enabling more individual response-aware CBE management. However, still there is a lack of field-applicable means (1) to detect stress from biosignals in an noise-robust and scalable manner; and (2) to provide stress cues (e.g., location and impact of stress cases) useful in understanding related stressors and designing effective interventions, despite these means' necessity in realizing the wearable biosensors' potential in CBEs. To fill these gaps, five interrelated studies were conducted (1) to denoise both stationary and non-stationary artifacts in biosignals collected during people's daily work and lives in CBEs; (2) to reliably assess generalizability of machine learning models for tasks monitoring human responses from biosignals; (3) to advance model generalizability across different subjects and contexts in detecting stress using a wearable biosensor; (4) to distinguish and locate stress responses related to environmental features; and (5) to differentiate stress types into positive (i.e., eustress) and negative types (i.e., distress) in terms of their impact on individuals. These studies can be the first steppingstones for operating and managing CBEs while actively accounting for individuals' responses. Such individual response-aware CBE operations will significantly contribute to improving human safety, health, and comfort in CBEs and ultimately promoting not only the performance of the construction industry, but only people's quality of life in built environments.

Chapter 1 Introduction

1.1 Background

Although the construction industry has been adopting a wide range of heavy equipment and robots nowadays, human workers remain one of the most important resources at construction sites. Human workers/managers are and will remain as collaborators and supervisors of these new equipment and robots, dealing with dexterous and intellectual tasks and creative decision making at construction sites. Besides, the purpose of most construction projects is to advance the quality of people's experience in the built and natural environment. Therefore, operating construction and built environments (CBEs, e.g., cities, buildings, construction sites) in a way that ensures people's safety, health, comfort, and productivity is key to success in the management of both construction and built environments.

In this regard, comprehensive efforts have been made to promote people's quality of experience in CBEs. Advanced machines, equipment, construction methods and regulations have been introduced to construction sites to relieve workers' physical and cognitive demands and risks to their safety and health. Also, improved design standards and operation and maintenance guidelines and techniques have been adopted to better satisfy citizens and occupants within their cities and buildings. However, there is still much room to promote people's quality of interaction with CBEs. Construction workers are often exposed to excessively demanding physical and mental tasks as well as harsh environments which pose risks to their health and safety. For example, 68% of construction workers suffer from excessive stress as a result of working at construction sites (Campbell 2006) and, relatedly, it has been reported that construction workers are 1.7 times more likely than those in other industries to suffer from emotional and mental disorders (Petersen and Zwerling 1998). Also, many construction workers suffer excessive levels of heat stress at construction sites. The construction industry constituted 31% of the heat-related illnesses claims recorded between 1995 and 2005, which was the highest in the United States industrial sectors (Bonauto et al. 2007). Similarly, between 2000 and 2010, the construction industry caused 36.8%

of occupational heat-related fatalities. This portion was higher than any other industries in the United States (Gubernot et al. 2015). Physical overexertion is another common issue at construction sites. Construction work typically involves physically demanding tasks often performed in harsh environmental conditions, which can cause physical fatigue and lead to poor judgment, poor quality of work, increased risk of accidents, and reduction in productivity. 20% to 40% of construction workers exceed generally accepted physiological thresholds for manual work in their daily routines (Abdelhamid and Everett 2002).

The quality of interactions between humans and their daily built environments, such as buildings and transportation systems, also need to be improved. For example, only about 38% of building occupants in North American commercial buildings are satisfied with their indoor thermal environments (Karmann et al. 2018). ASHRAE 55 guidelines require building heating, ventilating, and air conditioning (HVAC) systems to satisfy more than 80% of building occupants (ASHRAE 2017). However, a case study reported that only 8% of the surveyed buildings met the ASHRAE requirement (Karmann et al. 2018). People are not satisfied by the outside of buildings as well: more than half of citizens perceive the low quality of community transportation environments, such as buses and sidewalks (Bureau of Labor Statistics 2019).

Although there are different roots for low qualities of experience in CBEs, a commonly observed one across different CBEs is the "one-size-fits-all" operation approach. Current management and operation practices in CBEs mainly depend on the one-size-fits-all approach to ensure quality of experience for a hypothetical average person established by aggregating a group of people's responses. For example, Occupational Safety and Health Administration (OSHA) recommends assessing workplaces' heat stress by the ACGIH TLV and Action Limit (ACGIH 2019) to prevent heat-related illnesses among construction workers. These two lines on the two-dimensional plane of environmental and metabolic heats (i.e., wet-bulb-globe temperatures (WBGT) and metabolic rate) determine whether the workplace heat stress is safe, alarming, or dangerous based on the two heat factors: WBGT and metabolic rate. Also, outdoor sidewalks are designed and maintained using a predetermined design guideline and inspection checklist, which have been developed to guarantee that a hypothetical average person representing the majority of people feels comfort while walking over the sidewalks (Story et al. 1998). For providing thermally comfortable indoor environments, buildings' heating, ventilation and air conditioning (HVAC)

systems operate based on the predicted mean vote (PMV) model, which was established to make sure that a hypothetical average person is under the thermal comfort zone (Kim et al. 2018).

This one-size-fits-all approach may be effective for ensuring the quality of experience for those who are similar to whoever is imagined as the normative "one-sized" individual. However, every individual has unique characteristics (in terms of age, gender, physical and cognitive capabilities, prior experiences, etc.) and thus has different interactions with CBEs even under an identical setup. Therefore, the one-size-fits-all approach is not effective for ensuring diverse individuals' quality of experience in CBEs and, more importantly, often exposes populations underrepresented by the hypothetical average person to uncomfortable and risky circumstances. For example, under a work setup assessed as safe by the ACGIH TLV and Action Limit (ACGIH 2019), some construction workers whose characteristics (e.g., heat acclimation) are more subject to heat than the average might suffer excessively high physiological strain, which poses risk to their health and safety (Quandt et al. 2013). According to Arbury et al. (2014), 69% of heat-related fatalities occur during the first 3 days of a workers' job, which means that workers with a lack of heat acclimatization are not sufficiently considered in the current heat stress management practice. Also, citizens' experience on the same sidewalk varies depending on their characteristics (e.g., age, and disability), and some might suffer physical and psychological stress and discomfort while walking over a sidewalk (Lockett et al. 2005). A nation-wide survey conducted by the U.S. Census Bureau shows that seniors and the disabled, populations that are underrepresented by the "one-size," perceive a much lower quality of transportation environments, such as sidewalks and transit systems, than young adults do (Bureau of Labor Statistics 2019). This survey result has been echoed by many previous studies, revealing that those populations suffer physical, cognitive, and emotional barriers during outdoor trips in their daily communities that have been managed by the existing one-size-fits-all urban maintenance approach (Lockett et al. 2005; Michael et al. 2009; Rosenberg et al. 2009).

1.2 Stress as an indicator of an individual's quality of interaction with CBEs

Given individual variability, understanding each individual's quality of interaction with CBEs and reflecting this understanding in CBE operations is essential to effectively improve human quality of experience in CBEs. Stress, defined as a person's state of physiological tension resulting from

an interaction between a person's capability and an environment's demands (Lazarus and Folkman 1984), is a great indicator of the quality of interaction between an individual and their surroundings. Stress, in nature, reflects different individuals' unique characteristics (Ursin and Eriksen 2004) and unique interactions with their surroundings. For instance, if a person is climbing a slop and the physical demand is high enough to a pose challenge or threat to the person, a high level of stress occurs. On the other hand, if another person is walking up the same slop, but this person's physical capability is stronger so that the physical demand does not pose any challenge or threat, only a low level of stress ensues.

Also, stress can represent a wide range of environmental features in CBEs that pose challenges or threats to individuals regardless of whether the types of features are related to physical, cognitive, or emotional demands (Chang et al. 2013; Healey and Picard 2005; Setz et al. 2010). For example, when a construction worker is suffering high levels of physical pain, severe cold or heat, high stress ensues. Also, people feel stress when they experience environmental barriers that pose physical discomforts for their body or fall risks. Like these physical stimuli, when a person conducts a task with high cognitive workload, watches filth, or feels scared and unpleasant, high stress also occurs.

In this regard, detecting individuals' stress and examining where and when they suffer high levels of stress during their daily interaction with CBEs enables us to understand what underlies an individual's low quality of experience and ultimately to plan individual response-aware interventions that improve the quality of interaction between individuals and CBEs. Additionally, given that stress responses manifest as precursors in advance of detrimental outcomes on people's safety, health, and productivity (Choudhry and Fang 2008; Ferguson 1984; Shackleton 2021), detecting stress can allow preventive interventions to be conducted before actual harm occurs.

1.3 Potential of wearable biosensor to detect stress in CBEs

Due to the usefulness of stress as the indicator of individuals' quality of experience, there have been previous attempts to understand human stress in CBEs using manual surveys, such as questionnaires and structured interviews. For example, different scales, such as the Stress and Adversity Inventory (Slavich and Shields 2018) and the Perceived Stress Scale (Cohen et al. 1994), have been developed and applied to detect stress from self-reports of stressor exposure or perceived

stress in construction workers' daily work (Iremeka et al. 2021; Sun et al. 2022). Also, to understand people's environmental stressors in outdoor/indoor built environments, which negatively affect their daily mobility and mood, individual interviews have been conducted to compile a list of potential environmental stressors in interviewees' communities (Gallagher et al. 2010; Lockett et al. 2005; Rosenberg et al. 2012).

Although these manual surveys have contributed to understanding an individual's overall quality of experience and typical stressors in CBEs, there are several limitations in practice. First, manual surveys are in nature discontinuous and sporadic with long intervals. Therefore, these surveys might not support timely interventions, which are critical in some cases. For example, if a worker is suffering stress due to an excessive level of heat stress for a prolonged period, it can cause a serious detrimental outcome such as heat shock and death without a timely intervention. Also, surveys require subjects to recall their past in a post-hoc manner which may inflect results with recall-bias. Lastly, participating in manual surveys can be cumbersome to subjects and interfere with their daily work and lives. Therefore, there is a need of a continuous, recall bias-free, and minimally invasive means to detect people's daily stress in CBEs.

Advancements in wearable biosensing have made wearable biosensors affordable and accessible enough to apply extensively in people's daily lives (Mancino 2017). Wearable biosensors can detect stress by continuously collecting biosignals such as electroencephalogram (EEG), electrodermal activity (EDA), photoplethysmography (PPG), and skin temperature (ST). Perceiving and processing a stressor and peripheral sympathetic arousal induced by stress innervate physiological activities such as cerebral and cardiovascular activities and skin eccrine sweat production. These arousal-innervated central and peripheral physiological reactivities manifest as specific patterns in biosignals such as EEG, EDA, PPG, and ST (McCorry 2007).

Therefore, by measuring biosignals and analyzing their implicit patterns, human stress can be detected continuously without depending on invasive surveys or human recollection. Also, recent advanced machine learning techniques make it possible to learn and automatedly practicalize stress-sensitive biosignal patterns for field application (Elzeiny and Qaraqe 2018). Since these wearable biosensors- and machine learning-based techniques can detect stress in a continuous, recall bias-free, and minimally invasive manner, they have better field-applicability than the aforementioned survey-based methods.

1.4 challenges of applying wearable biosensors for understanding stress in CBEs

Despite these potentials, there are mainly three notable challenges in applying wearable biosensors for detecting stress during people's daily interactions with CBEs (Figure 1.1). First, artifacts in biosignals collected by wearable biosensors in the field significantly diminish the reliability of wearable-based stress detection (a in Figure 1.1). Specifically, artifacts with non-stationary characteristics (e.g., motion artifact in EEG, respiratory artifact in EDA) might not be effectively denoised by current signal processing techniques such as frequency range filtering techniques and blind source separation. These techniques pre-determine and apply a stationary artifact template in their denoising (Castellanos and Makarov 2006). These denoising techniques perform well in alleviating stationary artifacts whose signal characteristics are pre-defined, such as ocular (Nguyen et al. 2012), muscular artifacts in EEG (Chen et al. 2017) and power line interference in EDA (Bornoiu and Grigore 2013). However, as it is practically impossible to pre-define the signal characteristics of non-stationary artifacts due to their inherent variability and unpredictable nature, artifact template-based denoising techniques are not suited for non-stationary artifact removal.

The second challenge is that current wearable-based stress detection techniques are limited in scalability and thus might not be applicable to scrupulously understanding multiple people's responses to environments in a CBE scale (b in Figure 1.1). Biosignals show highly variable patterns even under an identical psychophysiological status according to different subjects and contextual setups, such as the type of related stimuli, ambient temperature, humidity, and light (Picard et al. 2001). However, since most current wearable biosensor-based techniques apply machine learning to only model stress-specific patterns in biosignals by learning existing training datasets, these techniques might not be effective in buffering such individual and contextual variabilities. In turn, the current techniques cannot reliably monitor an unseen person's responses in an unseen context without having collected labeled data from the unseen person and context.

Lastly, the current wearable-based stress detection techniques let us know the level of stress, but do not provide information about the circumstances of stress, such as the stressor and the impact on an individual, which are critical in designing effective stress-relief interventions (c in Figure 1.1). As aforementioned, there are a wide range of potential stressors in people's daily work and lives in CBEs and the impact of an identical stressor can vary by different individuals and even their daily condition. When stress relief interventions are designed to be specific to the stressor and its impact on each individual, the interventions' effectiveness can be guaranteed. However,

little is currently known as to how to non-invasively get such stress-related circumstantial information in people's daily work and lives in CBEs.

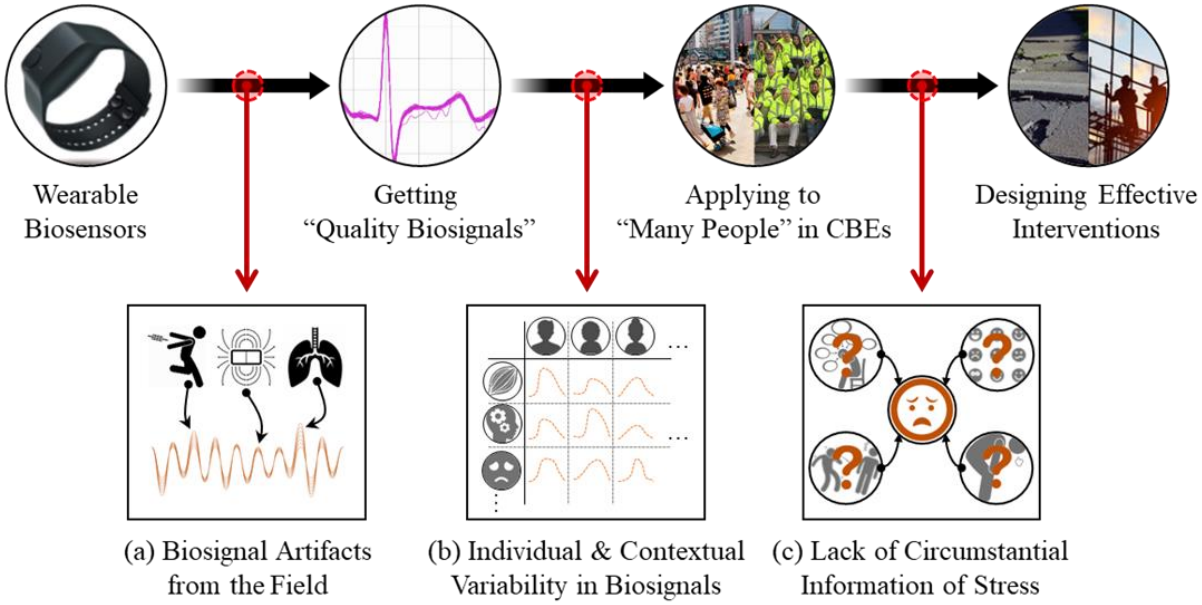


Figure 1.1 Challenges in applying wearable-based stress detection to understand and improve the quality of experience in CBEs.

1.5 Research goals and approaches

Against this background, the overarching goal of my PhD research is threefold: 1) to develop field-applicable denoising techniques that effectively alleviate not only stationary artifacts, but also non-stationary artifacts in biosignals collected in the field; 2) to advance the generalizability of wearable biosensors- and machine learning-based stress detection so that they can reliably work for unseen people under unseen contexts in a scalable manner; and 3) to non-invasively acquire circumstantial information about stress while detecting stress, thereby enabling the design of circumstance-specific effective stress relief interventions in CBEs.

- **To denoise both stationary and non-stationary artifacts in biosignals collected during people's daily work and lives in CBEs:** Pre-determining templates for non-stationary artifacts is not feasible. The most feasible way to alleviate the unpredictable nature of non-stationary artifact is to collect and leverage another signal associated with the artifact to cancel as a noise reference. To realize the potential of noise reference-based approach, we need to

examine what signals can be collected as noise reference, how to collect the noise reference signals in a field-applicable way, and how to subject the noise reference signals from raw biosignals to effectively address the non-stationary artifact to denoise.

- **To reliably assess the generalizability of machine learning models for tasks monitoring human responses from biosignals**: To develop a generalizable stress detection technique, it is critical to reliably assess the technique's generalizability with limited data in the development process. However, current validation methods tend to overestimate generalizability because they do not ensure testing with data collected from unseen people and unseen contexts. Therefore, a new validation method for reliable generalizability assessment needs to be developed.

- **To advance model generalizability across different subjects and contexts in detecting stress using a wearable biosensor**: Given that the physiological mechanism underlying biosignals' reactivity to stress is commonly sympathetic arousal (Cacioppo et al. 2007), detecting stress from different people in different contexts can be seen in a set of similar tasks that can share previously learned information, despite individual and contextual variabilities in biosignals' reactivity to stress. From here, transfer learning might be able to buffer individual and contextual differences in biosignals so that we can reliably detect stress across different people and contexts (Saeed et al. 2018; Taylor et al. 2020). However, what transfer learning technique is effective in buffering the individual and contextual differences in biosignals and how the transfer learning could be optimized to capture subject- and context-independent stress-specific feature vectors from complex biosignal datasets needs to be investigated.

- **To distinguish and locate stress responses related to environmental features:** Exclusively identifying stress caused by interactions with environments and detecting the locations of the related environmental stressors can help to identify stressors, relieve stress, and ultimately improve quality of experience in CBEs. However, in people's daily work and lives, a wide range of other stressors are mixed up with environmental stressors in CBEs. Therefore, there is a need to study how to distinguish stress responses to environmental features in CBEs from other confounding stressors and to spatially track environment-specific stress responses.

- **To differentiate stress types into positive (i.e., eustress) and negative types (i.e., distress)**: Understanding whether an individual experiences positive or negative stress is essential in designing effective interventions. According to how the person experiencing a stressor appraises it and their related coping resources, high stress typically falls into either eustress or distress, whose impacts are quite opposite from one another (LeBlanc 2009). Eustress can improve overall worker performance by enhancing self-efficacy, focus, and motivation (Lazarus and Folkman 1984), while distress is a root cause of detrimental consequences known to be "stress induced," such as chronic lethargy, depression, and reduced task-focus (Lazarus and Folkman 1984; Tomaka et al. 1993). Therefore, distinguishing eustress and distress enables us to make selective interventions that alleviate only distress while maintaining or promoting the benefits of eustress, thereby effectively advancing the quality of interactions between humans and CBEs. However, there is notable paucity of research into applying wearable biosensors to differentiate these stress types.

1.6 The structure of the dissertation

This dissertation is a compilation of the studies conducted to achieve the constructed research objectives. Seven chapters constitute this dissertation. Chapters 1 and 7 provide the introduction and conclusion of this work. Chapters 2 through 6 introduce each of the studies corresponding to the aforementioned research approaches. The following is the list of the chapters.

Chapter 1: Introduction. This chapter introduces the background, challenges to address, objectives, and approaches of the entire research effort.

Chapter 2: Noise reference signal-based adaptive denoising for non-stationary artifacts in biosignals collected in the field. This chapter develops two adaptive denoising techniques that collect and leverage a noise reference signal to effectively alleviate non-stationary artifacts in EDA and EEG, two useful signals for understanding human stress. Validations of the proposed denoising techniques are conducted by comparing their denoising performance with advanced benchmarks.

Chapter 3: Subject- and context-independent validation method to assess generalizability of machine learning models for monitoring human responses from biosignals. Considering individual and contextual variabilities in biosignal patterns, this chapter proposes a new subject-

and context-independent validation method for more reliable generalizability assessment: the leave one subject and context out cross validation (LOSCOCV). The performance of the proposed LOSCOCV in estimating the generalizability of models for tasks monitoring human responses from biosignals is statistically compared with existing, widely applied validation methods.

Chapter 4: Deep learning domain adaptation-based subject- and context-independent stress detection. This chapter proposes a subject- and context-independent stress detection technique that adopts a generative adversarial network (GAN)-integrated deep learning model to buffer domain differences between different subjects and contexts. The results of testing the proposed stress detection techniques in both in-lab and field setups are presented.

Chapter 5: Geographic information system (GIS)-based stress hotspot detection. This chapter presents a wearable biosensor- and hotspot analysis-based technique to detect stress hotspots as locations of environmental stressors. This technique statistically detects hotspots if the density of detected high stress in a given area is abnormally high, given that environmental stressors located on the hotspots might lead to such abnormal spatial stress concentration. The proposed technique is tested by applying it to seniors' daily trips in their urban areas.

Chapter 6: Mobile electroencephalography (EEG)-based stress type classification. This chapter develops a mobile EEG-based technique that differentiates stress types between low stress, distress, and eustress during people's daily interactions with CBEs. The results of the test conducted in a real construction task-similar indoor setup are presented.

- ***Chapter 7: Conclusions and recommendations.*** This chapter provides conclusions that can be drawn from the research. Several recommendations for future work stemming from this research are also provided.

Chapter 2 Noise Reference Signal-based Adaptive Denoising for Non-stationary Biosignal Artifacts in the Field

2.1 Introduction

This study aims to address the first agenda of this research: denoising non-stationary artifacts as well as stationary artifacts from biosignals collected in the field. Despite great potential, the application of wearable biosensors to detect human stress as an indicator of the quality of experience in CBEs has been limited because diverse artifacts in biosignals significantly compromise the reliability of biosignal-based stress analysis (Boucsein 2012; Heikenfeld et al. 2018). Many denoising techniques using an artifact template pre-determined based on the target artifact's stationary signal characteristics (e.g., frequency, amplitude, and morphology) have been developed and applied. However, these techniques might not adequately attenuate non-stationary artifacts, such as an EEG's motion artifacts and an EDA's respiratory artifacts, whose signal characteristics vary unpredictably over time.

The most feasible approach to alleviate such non-stationary artifacts with unpredictably varying signal characteristics is collecting and leveraging a noise reference signal that is correlated with the artifact to denoise and thus reflects the artifact's varying signal characteristics. Adopting the noise reference-based approach, this study presents two different denoising techniques to alleviate non-stationary artifacts in two biosignals most useful for understanding human stress in daily interactions with CBEs: electrodermal activity (EDA) and electroencephalogram (EEG).

2.2 Denoising EDA's respiratory artifacts

I adopted the noise reference-based adaptive denoising to improve the quality of EDA signals collected in the field. There are several stress-responsive biosignals that can be collected by biosensors, such as EDA, EEG, photoplethysmography (PPG) and electrocardiography (ECG). EDA has its own unique strengths for detecting stress in people's daily lives, compared to other

biosignals such as EEG, ECG and PPG. EDA indicates the changes in electrical conductance on the human skin stimulated by activities of the eccrine sweat gland (Boucsein 2012). While measuring EEG requires attaching multiple electrodes on scalp, EDA can be more simply and less-invasively measured by contacting two small dry-type electrodes on skin on wrist. This decreased invasiveness is one of the critical determinants for selecting biosignals considering that the signal should be continuously collected in people's naturalistic daily lives. Although ECG and PPG, which can indicate human stress like EDA, can be also simply measured on skin on chest, wrist or finger, EDA can be more accurate than ECG and PPG to measuring stress because ECG and PPG's reactivities to sympathetic arousal are mixed up with ones to parasympathetic arousal induced by relaxation (Boucsein 2012; Greco et al. 2017; Setz et al. 2010). In this regard, several studies have shown feasibility of wearable EDA sensor-based techniques to detect problematic features in the people's surroundings (such as CBEs) by monitoring people's stress (Kim and Fesenmaier 2015; Yadav et al. 2018).

Despite the great potential for wearable EDA sensing to understand the human-CBE interaction, the accurate measurement of stress from EDA remains challenging due to significant artifacts in EDA from people's daily work and lives (Sweeney et al. 2012) To measure human stress, electrodermal responses (EDR), the EDA's specific reactivity to sympathetic arousal, are first identified (Boucsein 2012; Greco et al. 2017; Setz et al. 2010). Since noises in EDA can be mis-detected as an EDR, they distort the stress measurement using EDA. Therefore, an appropriate denoising step is essential for accurate stress measurement. Artifacts in EDA can be categorized into two types: extrinsic artifacts and intrinsic artifacts (Boucsein 2012). Extrinsic artifacts are defined as artifacts that are generated from environments outside of human body (Heikenfeld et al. 2018). For example, electromagnetic field-related artifacts can be caused by alternating current frequency in the power lines and surrounding uncontrolled electromagnetic environments (Bornoiu and Grigore 2013). Also, variability in temperature and humidity on skin causes undesired drifts in EDA (Jebelli et al. 2018). Furthermore, heavy movements of users destabilize the contact between electrodes and skin, thereby generating noises (Drachen et al. 2010; Heikenfeld et al. 2018). Besides the extrinsic artifacts, several human physiological activities other than ones related to stress, such as activation of muscles and thermoregulatory sweating, also cause undesired modulation in EDA (Boucsein 2012), which is called intrinsic artifacts. One significant intrinsic artifact is respiratory artifact (Ksander et al. 2018; Schneider et al. 2003). When people

have irregular respiration such as deep breath and cough, it induces sudden increases in free-circulating adrenaline, thereby producing responses of eccrine sweat grands (Boucsein 2012). As a result, such irregular respiration-induced responses cause artifacts shaped similar to EDR in EDA and make stress overestimation (Hygge and Hugdahl 1985). In particular, the respiratory artifact in EDA might be more serious when EDA is collected under ambulatory settings than stationary settings because irregular respiration is more likely to happen when people have continuous physical activities than when they are still (Bradley and Esformes 2014).

To date, several techniques have been developed to alleviate artifacts in EDA for more reliably stress measurement. Most of the techniques depend on a pre-determined template of noise signal characteristics such as frequency, magnitude, amplitude, and morphology to suppress EDA's extrinsic artifacts that are in general stationary types. For example, a low pass filter with 0.25 Hz and smoothing techniques such as exponential smoothing and moving average filter have been widely applied to mitigate artifacts with higher frequency range than desired EDA signals, which are caused by electromagnetic fields or instability of electrode contacts (Boucsein 2012; Dube et al. 2009). Also, a high pass filter with 0.05 Hz has proven effective to suppress lower frequency artifacts (i.e., drift) introduced by variability in electrode impedance, humidity and temperature on skin (Jebelli et al. 2018). Recently, a couple of studies have been conducted to better alleviate motion artifact. Since part of the motion artifact has similar frequency range with the desired EDA signals, the denoising techniques have primarily depended on prior knowledge about the morphological characteristics of noise-free EDA signals (Chen et al. 2015; Shukla et al. 2018). These techniques have been validated to effectively suppress the motion artifact. Overall, the extrinsic artifacts are appropriately alleviated by previous denoising techniques based on artifact signal templates. However, theses denoising techniques are limited to attenuate intrinsic respiratory artifact because the respiratory artifact is a typical non-stationary artifact whose signal characteristics are not stationary and more importantly not distinguishable from characteristics of EDA's legitimate reactive waves to stress (i.e., EDR) (Boucsein 2012; Schneider et al. 2003). Given that the respiratory artifact is also very significant in EDA collected from people's daily lives (Bradley and Esformes 2014), there is a growing need for a new denoising technique that can alleviate the respiration artifact as well as other extrinsic noises in EDA. To fill the gap, this study proposes a noise reference signal-based denoising technique to alleviate the EDA respiratory artifact and tests its performance.

2.2.1 Proposed EDA denoising technique

The proposed denoising technique contains two modules: extrinsic noise attenuation and intrinsic respiratory artifact attenuation (Figure 2.1). To attenuate extrinsic noises, several filters such as high pass filter, moving average filter, and wavelet filter are first applied. Then, intrinsic respiratory artifact is alleviated by referencing a simultaneously collected respiratory artifact-correlated signal. In this study, the authors use PPG as the respiratory artifact-correlated signal. PPG, which refers to changes in the intensity of human skin's absorption and reflection of illuminated light (Shelley and Shelley 2001), well represents volumetric flow of blood beneath the target skin. Since intrathoracic pressure modulated by respiratory activity is represented in the blood volumetric flow, respiratory activity can be monitored by analyzing PPG (Garde et al. 2014; Ugnell and Öberg 1995). As irregular respirations inducing respiratory artifacts in EDA (e.g., sudden deep breath and cough) can be expressed by sudden abnormalities in respiratory rate or volume of air breathed (Schneider et al. 2003), the authors suggest using PPG as a respiratory artifact reference signal. The PPG signal is assumed to be simultaneously collected along with EDA by a multimodal biosensor like a wristband.

- **Extrinsic artifact removal**

First, several filtering techniques are applied to suppress extrinsic noises in EDA. Specifically, a high-pass filter with the cut of the frequency of 0.05 Hz was applied to mitigate low-frequency noises caused by variation in temperature, humidity, and the EDA sensor's electrode impedance (Jebelli et al. 2018). Then, high-frequency noises introduced by severe motions and electromagnetic interference are suppressed by applying a moving average filter (Bornoiu and Grigore 2013). Lastly, to mitigate parts of motion artifacts, which the moving average filter could not sufficiently attenuate due to the similarity between the frequency range and the desired EDA signal, wavelet decomposition-based filter technique is applied (Chen et al. 2015). Specifically, this filter technique conducts stationary wavelet transformation to decompose EDA signal into multiple wavelets. Then, the threshold of the wavelets' coefficients is adaptively determined based on the statistic estimation of the coefficients' distribution to filter out abnormal wavelets that might represent noises in EDA. Using the determined thresholds, wavelets representing cleaned EDA are selected, and denoised EDA signal is obtained by applying inverse wavelet transformation.

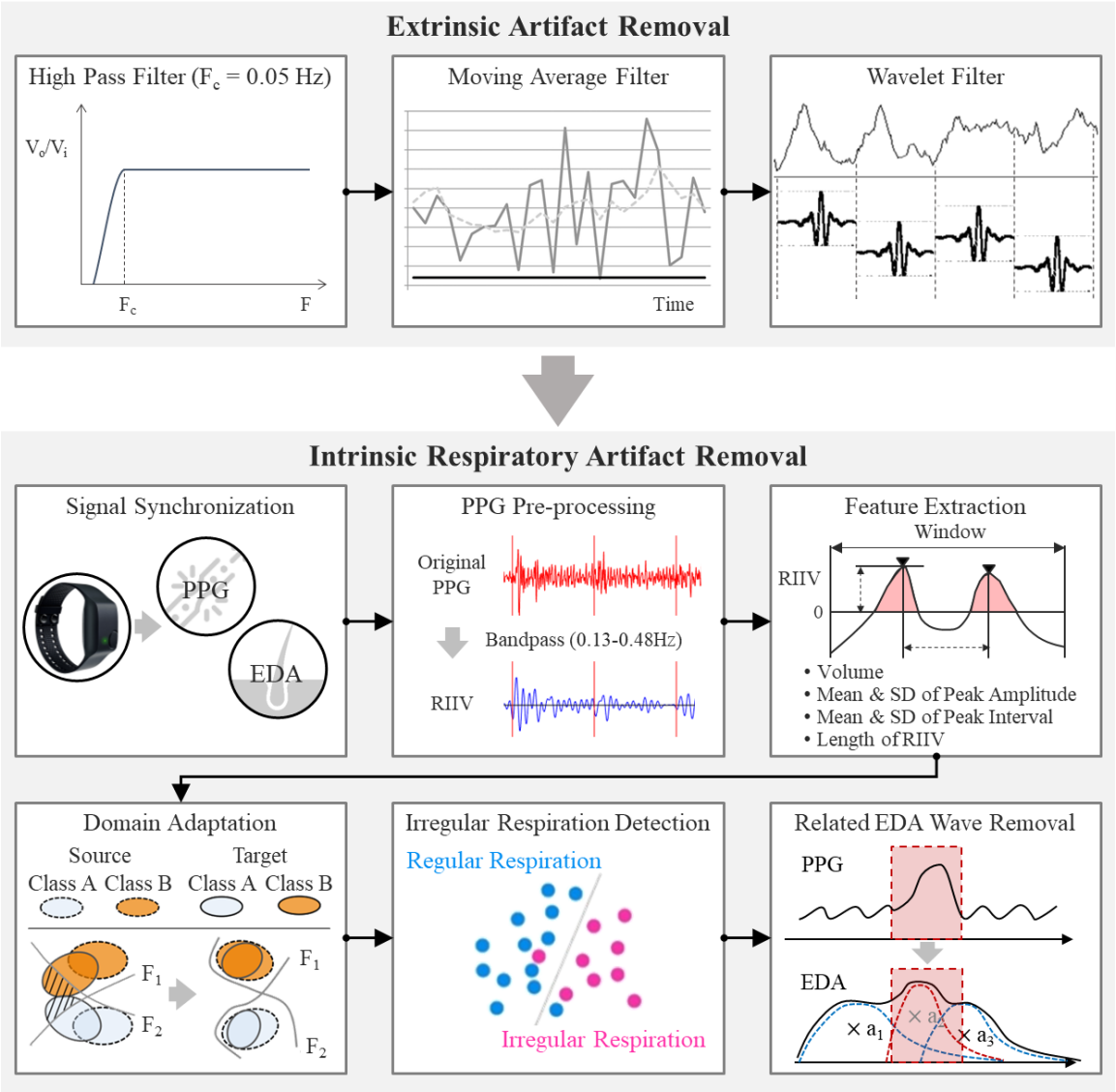


Figure 2.1. Overview of the proposed denoising technique for EDA

- **Intrinsic respiratory artifact removal**

To reference PPG signal for denoising EDA, the two signals are timely synchronized. Then, a classification model to detect irregular respiration from PPG is trained by applying machine learning algorithms. Before training the classification model, a band-pass filter within the frequency range of 0.13 – 0.48 Hz is first applied to alleviate noises, and extract respiratory-induced intensity variations (RIIV) from raw PPG signal (Ugnell and Öberg 1995). The extracted RIIV is segmented into samples by a moving window with the length of five consecutive breaths (i.e., five peaks in RIIV) and one breath shift (Ayappa et al. 2009). Having consecutive breaths as

a sample, the authors extract 11 features (Table 2.1) that have proven useful in detecting the abnormality of the breaths (Ayappa et al. 2009; Blain et al. 2010).

Table 2.1. Features extracted from respiratory-induced intensity variations (RIIV)

Feature Name	Description
Integral of RIIV +	Integral of RIIV in positive area
Integral of RIIV -	Integral of RIIV in negative area
Mean of Peak Amplitudes	Mean of peaks' amplitude
SD of Peak Amplitudes	Standard deviation of peaks' amplitude
Mean of Peak Intervals	Mean of intervals between consecutive peaks
SD of Peak Intervals	Standard deviation of intervals between consecutive peaks
Mean of Trough Amplitudes	Mean of troughs' amplitude
SD of Trough Amplitudes	Standard deviation of troughs' amplitude
Mean of Trough Intervals	Mean of intervals between consecutive troughs
SD of Trough Intervals	Standard deviation of intervals between consecutive troughs
Length of RIIV	Length of RIIV line in a window

With the features extracted from RIIV, several machine learning algorithms [i.e., Logistic Regression (LR), Decision tree (DT), Gaussian support vector machine (Gaussian SVM), and k-nearest neighbors (KNN)] are applied to train a classification model. Specifically, subject-independency (i.e., how accurately a model works for general public, not only for a specific group of people who are involved in the labeled data collection) of classification model is considered important in this study because it is unfeasible to train a classification model specific to each user by collecting labeled irregular respiration data for everyone. To train a subject-independent classifier, domain adaptation is applied. Domain adaptation is a transfer learning technique that enables a model learned on labeled data in a source domain to conduct a similar task on unlabeled data in a different target domain (Daume III and Marcu 2006). The techniques find an optimal feature map that minimizes discrepancy in data distribution between source and target domains assuming that the two domains would share a model (e.g., classification boundary and prediction trend line) on the optimized feature map (Glorot et al. 2011). In this study, a group of subjects who are involved in the collection of labeled irregular respiration data is the source domain, and a new

person who is not involved in the data collection is the target domain. Once an irregular respiration classifier is trained on labeled data from a group of subjects, the domain adaptation can make the trained classifier subject-independently works for a new unlabeled person. Several domain adaptation algorithms such as subspace alignment (SA) (Fernando et al. 2013), transfer component analysis (TCA) (Pan et al. 2011), information-theoretical learning (ITL) (Shi and Sha 2012), domain-adversarial training of neural network (DANN) (Ganin et al. 2016), and multisource domain-adversarial network (MDAN) (Zhao et al. 2017) are attempted in the irregular respiration detection task. To select the best combination of classification and domain adaptation algorithms, subject-independency as well as accuracy of the attempted algorithms are compared using the leave-one-subject-out-cross-validation (LOSOCV). The LOSOCV shows algorithms' accuracy and subject-independency by excluding one subject's data from training phase, and using this subject's data as a validation set (Rice and Silverman 1991).

Applying the trained classifier, irregular respirations are subject-independently detected from PPG. Respiratory artifact induced by the irregular respiration is detected and attenuated in EDA. First, EDRs are identified by applying a EDA decomposition technique based on sparse representation (Chaspari et al. 2016). This technique, which identifies EDRs using a dictionary prepared based on prior knowledge about EDR's morphological characteristics, showed higher accuracy of stress measurement than previous EDR identification technique (Chaspari et al. 2016). After identifying all EDRs, EDRs caused by irregular respiration are detected and removed as a respiratory artifact following the rule introduced in Schneider et al. (2003). In the rule, EDRs are first clustered based on their morphology and closeness in their occurrence timing. Specifically, the consecutive two EDRs are clustered if the following three rules are satisfied: i) following EDR's peak amplitude is less than that of the preceding EDR, ii) the following EDR's onset amplitude is higher than that of the preceding EDR, and iii) the onset of the following EDR is within the latency window X from the onset of the preceding EDR; here X is calculated by multiplying the peak amplitude of the preceding EDR by a pre-defined constant (10 seconds / μS). Each cluster is regarded as one reaction to one cause. Then, clusters appearing 1-5 seconds after the occurrence of irregular respiration are collected as a respiratory artifact; there exists the latency interval between irregular respiration and EDA's reactivity (Venables and Christie 1980). Although Schneider et al.'s (2003) rule used an objective threshold of impedance, the authors do not apply the threshold because impedance value can be variant by different conditions of EDA

collection (e.g., areas of electrodes, and distance between electrodes) (Boucsein 2012). Finally, the denoised EDA is reconstructed by combining EDRs not contained in the clusters which are detected as a respiration-induced noise (Figure 2.2).

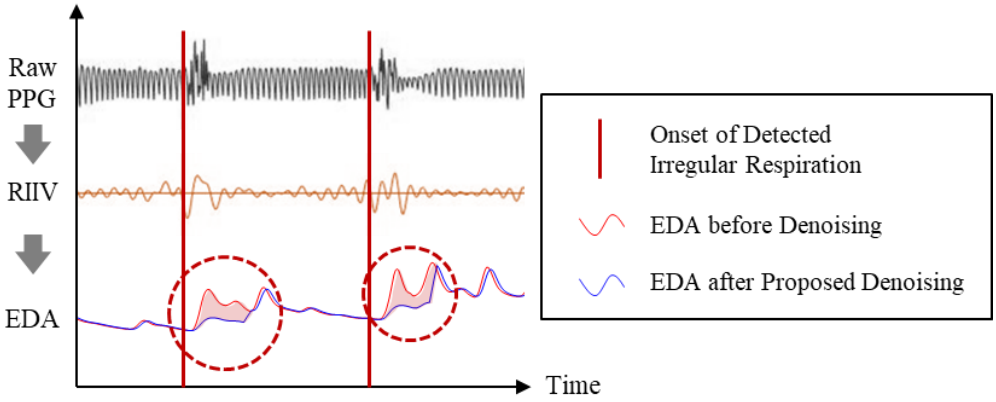


Figure 2.2. Attenuation of respiratory artifacts in EDA

2.2.2 EDA denoising performance test

The proposed technique's denoising performance was compared with advanced benchmark denoising techniques by using data collected from in-lab and field settings as shown in Figure 2.3. Through the in-lab data collection, PPG signals were collected by controlling subjects' respiration and manually labeled as "irregular respiration" and "regular respiration" (a in Figure 2.3). The labeled PPG signals were used to train the subject-independent irregular respiration classifier, which is essential to the intrinsic respiratory artifact attenuation in the proposed technique. A field data collection was also conducted to collect EDA and PPG signals labeled as the level of stress (high and low stress), which was used to examine performance of the proposed denoising technique compared to previous benchmark denoising techniques (b in Figure 2.3). Specifically, the field data collection was executed with senior subjects over the age of 65. Since older adults are likely to have more irregular respiration during physical activities due to their impaired respiratory function (Sharma and Goodwin 2006), EDA collected from older adults might contain more significant respiratory artifact than one from younger people. Therefore, collecting data from the senior population could give change to more explicitly compare the performance to alleviate respiratory artifact in EDA. Also, seniors are among the demographic groups who would benefit the most from wearable EDA sensing because they have more stressful interactions with the built

environment in their daily trips due to their physical and cognitive impairment (Lockett et al. 2005; Rosenberg et al. 2012). The data collection was conducted from Jul. to Sep. of 2018. The protocol for the data collection was approved by the University of Michigan Institutional Review Board (IRB00000245). Using EDA collected by the field data collection, the measurement quality of stress metrics and accuracy of stress classifier based on denoised EDA were compared between the proposed technique and benchmarks as indices of denoising performance (c-f in Figure 2.3).

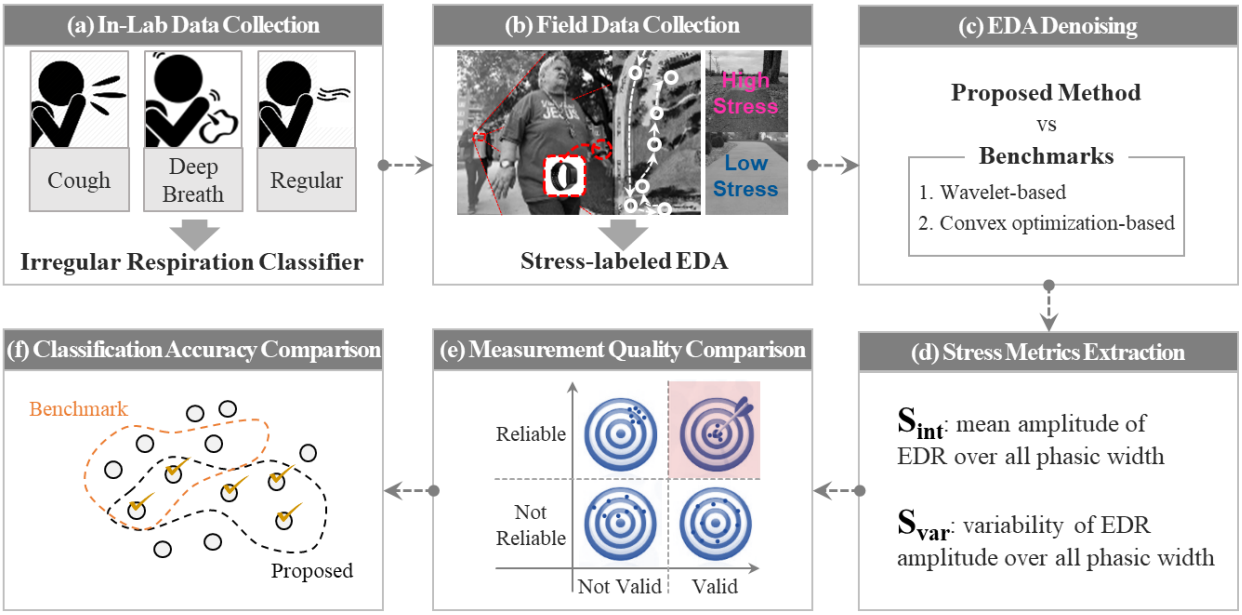


Figure 2.3. Overview of the denoising performance test

- **In-Lab Data Collection**

Ten graduate students in the University of Michigan participated in the in-lab data collection as subjects (Table 2.2). No subject reported any mental or physical health issue that would affect their respiratory activity. To collect PPG signals clearly divided into regular and irregular respiration frames, subjects were asked to breathe regularly or irregularly such as deep breath and cough with a simple wristband-type biosensor. Since the irregular respiration classifier should reliably work for PPG collected in an ambulatory context in which there might be a lot of noise, half of the data was collected while subjects walked. As with irregular respiration, cough and deep breath were simulated because they are the main causes of the respiratory artifact in EDA (Boucsein 2012). Specifically, in the regular respiration session, subjects were asked to breathe regularly while

walking (six minutes) and sitting (six minutes). During the following irregular respiration session, subjects first simulated a cough for five seconds every 20 seconds for six minutes (three minutes of walking and three minutes of sitting). They also took a deep breath for five seconds every 20 seconds for six minutes (three minutes of walking and three minutes of sitting). Samples whose last peak overlapped with the simulated irregular respirations were labeled as irregular respiration. Subjects' PPG signal was collected with sampling rate of 64 Hz.

Table 2.2. Demographic information of 10 subjects in the in-lab data collection

Statistics	Age (years)	Height (cm)	Weight (kg)	Gender
Mean (SD)	30.0 (2.7)	170.1 (10.1)	63.4 (13.4)	Male 7 / Female 3

- **Field Data Collection**

The field data collection was conducted in cooperation with East Clark senior residence in Ypsilanti Township, Michigan. A total of 25 senior residents of the facility were recruited (Table 2.3). Subjects were asked to report any physical or mental health issues that can affect their respiratory activity and physiological reactivity to stress, but none of them reported such issues. The authors gathered the senior subjects' EDA and PPG signals while they experienced different levels of stress. Specifically, the authors designed a route so that subjects naturally experienced 15 pre-designated environmental stressors (e.g., stairs, side-sloped and unpaved sidewalk, and narrow pathway) while ambulating along the route. The designated environmental stressors were determined by studying previous research efforts identifying different types of environmental stressors older individuals suffer from in the current built environment (Lockett et al. 2005; Rosenberg et al. 2012). Given that people can accurately recall their physiological status such as stress and emotion within a maximum of 15 minutes (Jacobs et al. 2005; Nezlek et al. 2008), the length of the route was set to limit the duration of each trial to 10 minutes to ensure that subjects correctly recall perceived stress after a trial. While subjects moved along the route and experienced the stressors, their EDA and PPG signals were collected by the wristband-type biosensor which was used in the in-lab data collection. The sampling rate for EDA and PPG were respectively 4 Hz and 64 Hz. Also, their videos were taken during trials, which was referenced to label collected EDA signal later. After each trial of the route, the authors asked subjects to self-report their perceived stress on each environmental stressor in binary (i.e., high stress and low stress). To

enhance their understanding of stress, examples of stress were provided before the self-report. Also, the authors showed pictures of each environmental stressor to help subjects recall his/her experience.

Table 2.3. Demographic information of 25 subjects in the field data collection

Statistics	Age (years)	Height (cm)	Weight (kg)	Gender
Mean (SD)	68.4 (4.8)	165.8 (10.9)	83.6 (20.3)	Male 4 / Female 21

Based on the collected videos and subjects' self-reports, collected EDA was labeled as stress level in binary. First, EDA signal recorded while subjects faced unintended events that could affect their perceived stress (e.g., interacting with automobiles and other passengers, near fall unrelated to environmental stressors) were excluded. Then, EDA collected in intervals when a subject passed over environmental stressors that the subject reported as high stress was labeled as "high stress," while EDA in other intervals was labeled as "low stress."

- **Comparison of Denoising Performance with Previous Denoising Techniques**

The stress-labeled biosignals collected by the field data collection were used to compare denoising performance of the proposed technique with benchmark denoising techniques. The labeled signals were first denoised by the proposed technique and benchmark techniques respectively (c in Figure 2.3). Two advanced denoising techniques [i.e., the wavelet-based technique (Chen et al. 2015) and convex optimization-based technique (Greco et al. 2016)] were applied as benchmarks in this study. Using the denoised signals, the authors compared i) measurement quality of stress metrics, and ii) accuracy of stress level classifiers to see the performance of the proposed denoising technique. Since the purpose of denoising EDA is to accurately measure stress from EDA, the measurement quality of stress metrics calculated from denoised EDA can be an important index of denoising performance. Specifically, two EDA-based stress metrics suggested by Chaspari et al. (2016) were calculated: mean amplitude of EDR over all phasic width (S_{int}) and variability of EDR amplitude over all phasic width (S_{var}) (d in Figure 2.3). These stress metrics showed better performance to measure stress than previous metrics such as mean and mean frequency of EDR (Chaspari et al. 2016). The length of the window to calculate the stress metrics was set by 10 seconds because EDA's reactivity to stress occurrence generally spans 10 seconds (Singh et al.

2014). The authors compared the validity and reliability—two of the most important criteria of measurement quality (LoBiondo-Wood and Haber 2014) (e in Figure 2.3). The validity means how closely correlated a stress metric score is with stress subjects perceive, while the reliability indicates how consistent a stress metric score is (LoBiondo-Wood and Haber 2014). As an index of the validity, this study used the point-biserial correlation coefficient (BCC) between continuous stress metrics (i.e., Sint and Svar) and the level of stress (Calkins 2005). For measuring reliability, the intraclass correlation coefficient (ICC) has been widely used (Srivastava 1984). The ICC describes how strongly scores of stress metrics (i.e., S_{int} and S_{var}) are similar within a class and different between classes (i.e., high and low stress). For both coefficients, the closer the value is to 1, the better the measurement quality is.

After comparing the quality of stress metrics, the authors examined how the proposed denoising technique affects the performance of machine learning models to detect high stress as well (e in Figure 2.3). Accuracy of machine learning models trained using denoised signals has been widely used to examine performance of denoising techniques (Li et al. 2015; Zhu and Fujii 2017) (f in Figure 2.3). In this study, having the two stress metrics (i.e., S_{int} and S_{var}) as a feature, Gaussian SVM was applied to train a model to measure the level of stress in binary (i.e., high and low stress) because Gaussian SVM has outperformed other machine learning algorithms (e.g., DT and KNN) in the stress classification task using biosignals (Jebelli et al. 2019; Jebelli et al. 2018). Then, the accuracy of the trained model was compared between the proposed denoising technique and benchmarks by 10-fold cross validation, which has been widely used to evaluate the performance of trained machine learning models in previous studies into biosignal-based stress detection (Setz et al. 2010; Sun et al. 2010).

2.2.3 Results

Through the in-lab data collection, a total of 3,961 samples labeled as irregular (1,075) or regular respiration (2,886) were collected. Several machine learning and domain adaptation algorithms were tested, and the best combination of these algorithms was selected by comparing accuracy of LOSOCV. As a result, the multisource domain-adversarial network (MDAN) outweighed other tested algorithms with 0.849 accuracy, 0.828 precision, and 0.800 recall (Table 2.4). Figure *2.4* shows the performance of MDAN in classifying two subjects' (subject #1 and #2) PPG signals to regular and irregular respirations. The two dimensions of these figures are standard deviation of

peak amplitudes and trough intervals that were selected as the most meaningful two features. The backward-elimination wrapper technique (Kohavi and John 1997) was applied to select the most meaningful features. In general, the values of irregular respirations are higher than regular ones in both of the two features, but the distributions of two classes are different for each subject as shown in Figure *2.4*. This figure demonstrates that MDAN can accurately classify respirations by buffering the discrepancy in distributions between different subjects. Based on this result, the MDAN was used to detect irregular respiration as the first step of the proposed denoising.

Table 2.4 LOSOCV accuracy of irregular respiration classifiers

Classifier Learning Algorithm	Domain Adaptation Algorithm	Accuracy (LOSOCV)
Logistic Regression	TCA	0.737
	SA	0.649
	ITL	0.721
	unapplied	0.567
Gaussian Support Vector Machine	TCA	0.745
	SA	0.660
	ITL	0.682
	unapplied	0.609
K-nearest Neighbors	TCA	0.737
	SA	0.735
	ITL	0.656
	unapplied	0.592
Neural Network	DANN	0.708
	MDAN	0.849

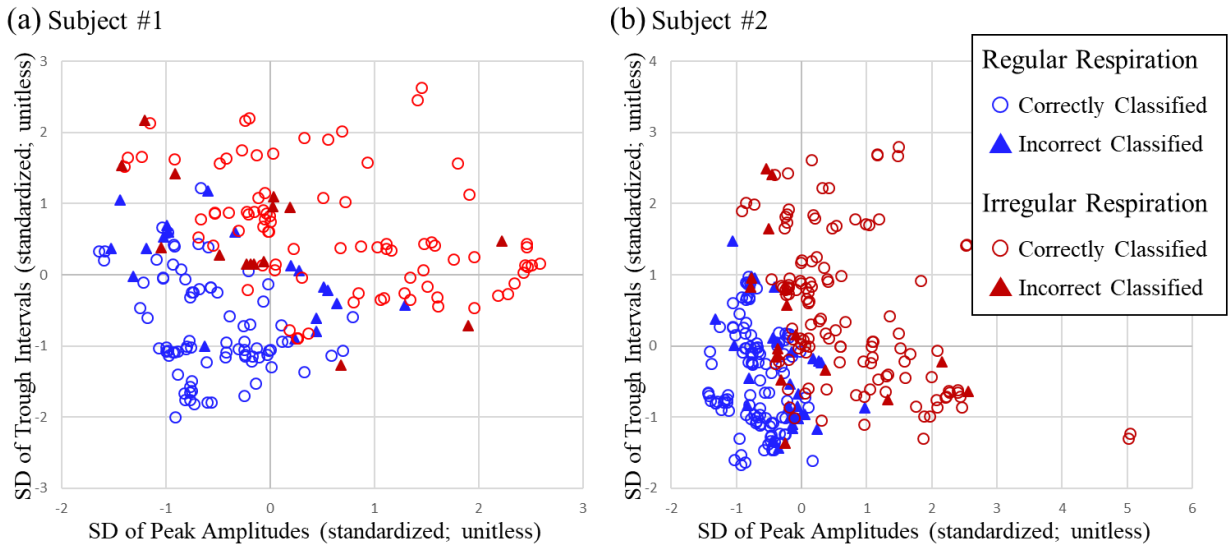


Figure 2.4. Respiration classification performance in two dimensions; (a) Subject #1; and (b) Subject #2

The EDA signal collected by the field data collection was denoised by identifying and discarding irregular respiration using the irregular respiration classifier trained based on the selected MDAN algorithm. As a result of the field data collection, a total of 16,319 samples labeled as high stress (1,499) or low stress (14,820) were collected. After denoising the EDA signal, two measurement quality criteria (i.e., the validity and reliability) of the two stress metrics (i.e., S_{int} and S_{var}) were calculated from the denoised EDA. The calculated validity and reliability were statistically compared with ones from EDA denoised by benchmark techniques (i.e., wavelet-based and convex optimization-based techniques) using the bootstrap resampling technique (Diamantopoulos et al. 2012; Dragoi et al. 2003) that has been widely used to examine statistical difference between two correlation coefficients by generating distributions of the two correlation coefficients through bootstrap resampling. The size of the bootstrap was set equal to original sample size (with replacement) (Kuhn and Johnson 2013), and the number of sampling were determined as 90 based on the statistical power analysis (0.99 power, 0.01 alpha, medium effect size of 0.5). As the result of statistical comparison of the two quality criteria, both S_{int} and S_{var} had statistically higher validity and reliability with the proposed technique than benchmark techniques (Table 2.5). This means that the proposed denoising technique better attenuates the distortion of stress measurement caused by noises in EDA. To see the importance of this improvement in the

application of EDA, the authors also investigated the performance improvement of classification models trained using these two stress metrics (i.e., S_{int} and S_{var}) in classifying the level of stress in a binary manner, which is the most fundamental task in stress measurement as detailed in the following paragraph.

Table 2.5. Statistical comparison in quality of stress metrics between the proposed denoising technique and previous techniques

Stress Metrics	Measurement Quality Criteria	Proposed Technique	Benchmark Technique		Statistical Comparison	
			#1. Wavelet-based Technique	#2. Convex Optimization-based Technique	w/ #1 T-value	w/ #2 T-value
Sint	Validity (BCC)	0.236	0.164	0.164	26.69*	30.22*
	Reliability (ICC)	0.293	0.154	0.153	27.69*	32.27*
Svar	Validity (BCC)	0.230	0.165	0.165	23.45*	26.13*
	Reliability (ICC)	0.281	0.155	0.154	24.15*	27.51*

*: p-value ≈ 0

The authors trained the classification models having the S_{int} and S_{var} as features to detect level of stress by applying Gaussian SVM, and the performance of classifier was statistically compared between the proposed technique and benchmarks. Since the number of low stress samples (90.8%) were much more than high stress samples (9.2%), the trained model might be biased towards the low stress class. To avoid this failure, the low stress samples were first randomly undersampled to make the number of two classess same before training. For avoiding bias introduced by the random undersampling, the classifers' performance was calculated by averaging 10 trials of the random undersampling and 10-fold cross-validation. To statistically compare the performance of classifiers, the authors used McNemar's method (Dietterich 1998) that has been widely used to statistically compare performance of two classifiers using chi-squared test (Saha et al. 2014; Trakoolwilaiwan et al. 2017). The comparison showed that the classification model trained using EDA denoised by the proposed technique have statistically higher accuracy than models using EDA denoised by previous techniques (Table 2.6). Specifically, the increase in accuracy (4.1%, 0.609 → 0.634) was mainly due to the increase in the performance to detect the high stress class (the increase in F1 score of the high stress class: 11.1%, 0.562 → 0.626), not that of the low stress class (the increase in the F1 score of the low stresss class: 0.0%, 0.643 → 0.641).

This means that the proposed technique more benefits classification model's performance to detect high stress, which is more critical than detecting low stress in most practical applications.

Table 2.6. Performance of classification models to detect high stress using the proposed denoising technique and previous techniques

Performance Index	Proposed Technique	Benchmark Techniques		Statistical Comparison	
		#1. Wavelet-based Technique	#2. Convex Optimization-based Technique	w/#1	w/#2
Accuracy	0.634	0.604	0.609	n01 = 10,921 n10 = 12,490 Z-value = 52.37*	n01 = 10,514 n10 = 11,954 Z-value = 46.05*
F1-Score (High stress)	0.626	0.558	0.562		
F1-Score (Low stress)	0.641	0.641	0.643		

Note: n_{01} = # of samples that a classifier based on a benchmark technique correctly classified, but one based on the proposed technique misclassified; n_{10} = # of samples that a classifier based on the proposed techniques correctly classified, but one based on a benchmark technique misclassified.
*: p-value ≈ 0

2.2.4 Discussion

To materialize the potential of wearable EDA sensor-based stress measurement to monitor interaction between humans and CBEs, this study proposes a denoising technique that attenuates both extrinsic artifacts and intrinsic respiratory artifacts, a representative non-stationary EDA artifact. As a part of the denoising technique, several machine learning and domain adaptation algorithms were tested to train a classification model that subject-independently detects irregular respiration from PPG collected by a wearable biosensor. As shown in Table 2.4, the models trained without domain adaptation showed a lower LOSOCV accuracy (0.567 – 0.609) than models with domain adaptation (0.659 – 0.849). This result implies that domain adaptation decreases the subject-dependency of the trained models. This finding can provide a new research direction to improve the practicality of biosening-based approaches to understand diverse human status (e.g., emotion, mood, and physical exertion). Many biosensing-based approaches first train a model to understand human status by applying supervised learning algorithms. Because of people's distinct physiological activities, such dependency on supervised learning algorithms brings one practical hurdle that labeled data should be collected from all targeted people. The finding of this study demonstrates the potential of domain adaptation to eliminate the need for collecting labeled data

from all targeted individuals in the biosensing-based approaches by buffering the individual differences in physiological activity. In this regard, research to investigate whether domain adaptation can assure subject-independency in biosignals-based tasks, other than detection of irregular respiration, is worth consideration.

Out of the attempted algorithms, the multisource domain-adversarial network (MDAN) showed the best LOSOCV accuracy. It can be explained by the fact that there are multiple source domains, not limited to the one used in this study. To detect irregular respiration in a new person, labeled data from multiple people was used, and each of them has distinct PPG reactivity to irregular respiration. The MDAN was proposed to deal with such "multisource" domain adaptation problem by finding one optimal feature map on which distributions of all domains (i.e., multiple source domains and a target domain) are most similar to each other (Zhao et al. 2017). On the other hand, other domain adaptation techniques do not differentiate subjects in the source, so that they might not account for the difference in reactivity to irregular respiration between the source subjects, which can negatively affect the performance of the domain-adapted model.

The proposed denoising technique showed statistically higher validity and reliability of stress metrics than benchmark techniques. This improvement in the quality of stress metrics was mainly because the proposed denoising technique alleviated stress overestimation in samples belonging to the low stress class, which were caused by respiratory artifact. Figure *2.5* shows subject #17's distribution of standardized scores of the two stress metrics. Compared to the convex optimization-base denoising technique, the proposed technique made the distribution of low stress samples skewed more left. Consequentially, the distribution of high stress samples skewed relatively right through standardization. This result is consistent with the observation during the field data collection. The authors observed that the older subjects often had irregular respiration like deep breath and cough regardless of their stress when moving along stressless parts of the route. Such irregular respiration might cause respiratory artifacts shaped like EDR in EDA, which the proposed denoising technique alleviated. Such improvement observed in stress metrics was naturally followed by increase in accuracy of classification model trained using the stress metrics (4.1% increase in accuracy, 11.1% increase in F1 score of the high stress class). As the overlap between two classes of samples declined by the proposed denoising technique, Gaussian SVM more easily optimized a decision boundary to divide data into the two classes. Such an effect of the proposed denoising could be bigger in people's actual lives. Unlike the field data collection in

this study where the duration of ambulation was limited less than 10 minutes, people's daily trips in actual lives are often much longer. In such cases, people might have more frequent irregular respirations in physical tiredness, which induces more respiratory artifacts in EDA.

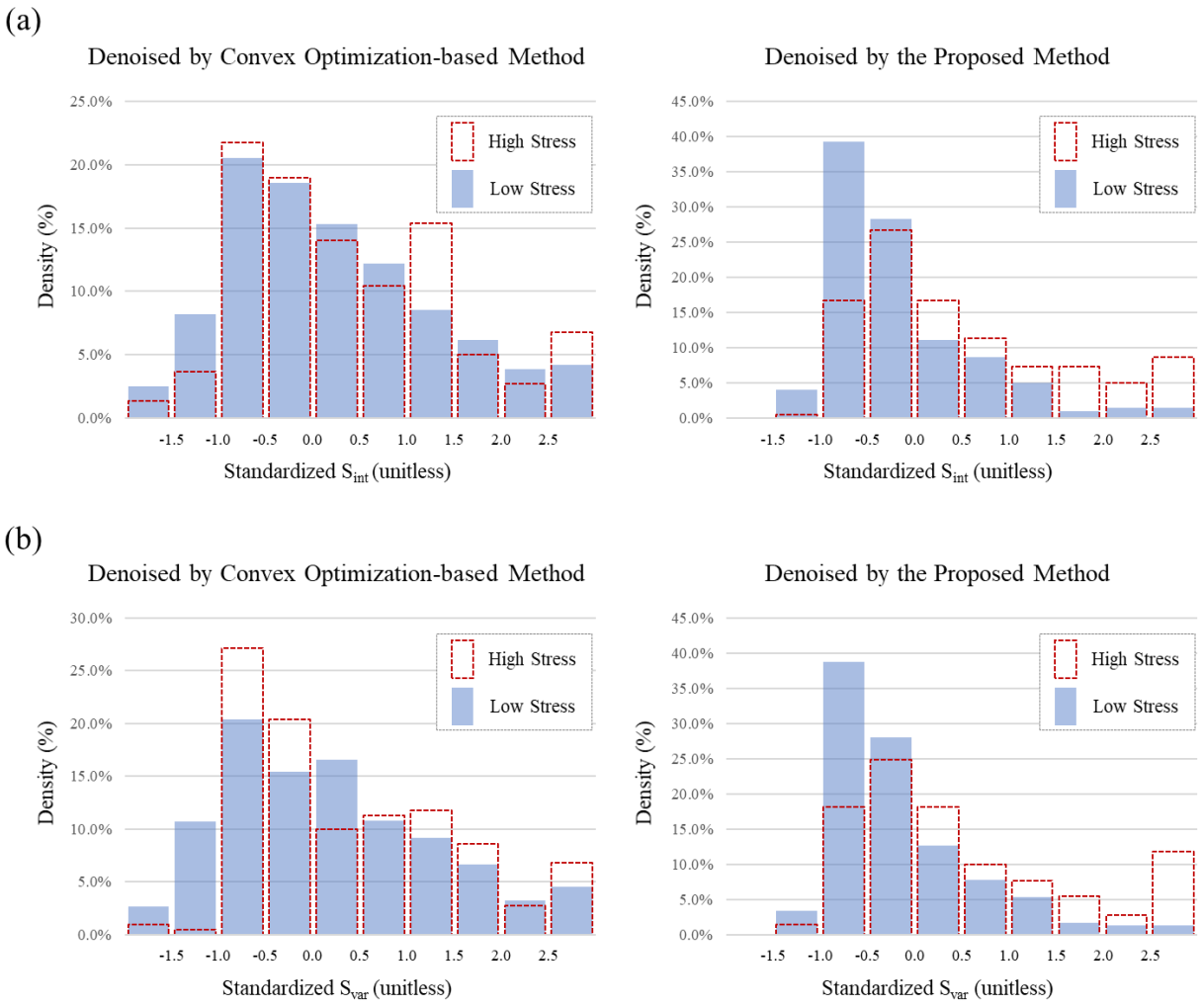


Figure 2.5. Comparison of distribution of stress metrics between the proposed technique and previous technique (convex optimization-based);
(a) distribution of S_{int} value; (b) distribution of S_{var} value

The aforementioned results indicate that the proposed PPG signal-based denoising is effective to attenuate artifacts in EDA, including intrinsic respiratory artifact with non-stationary signal characteristics that are also indifferentiable with legitimate EDRs. The proposed noise-reference signal-based denoising can have extended applicability for improving the quality of biosignals collected by wearable biosensors from fields. Besides the respiratory artifact in EDA,

this noise reference signal-based denoising approach can be applied to other types of artifacts in other signals, which could not be well attenuated by the denoising techniques based on a fixed noise signal characteristic template. For example, muscular artifact, another type of intrinsic artifact in EDA, can be better attenuated by referencing acceleration data than previous denoising techniques such as wavelet and convex optimization-based techniques. Also, EEG's motion artifact, the most significant type of artifact in EEG collected in the field, might be effectively alleviated by this noise reference-based adaptive denoising, about which the following subsection 2.3 describes.

Through this study, the authors found that the proposed denoising technique can improve the accuracy of stress measurement using EDA collected by a wearable biosensor, thereby facilitating wearable biosensing-based continuous, less-invasive, and less-laborious monitoring of the interaction between people and their surroundings, such as CBEs. The wearable biosensing-based scalable monitoring can contribute to advancing the quality of the human-CBE interaction by enabling effective interventions to address people's stress caused by the interaction.

Despite the contribution of this study, there are several limitations in this study that need to be addressed by future studies. First, there is a chance for the proposed denoising technique to degrade measurement quality by underestimating actual stress in a specific case. If a person has several irregular respirations while he/she feels high stress, then some of the actual EDRs caused by high stress levels can be misinterpreted as a respiratory artifact and attenuated by the proposed technique. In general, since multiple EDRs continuously occur when people feel high level of stress, the underestimation effect of the proposed technique is expected to be ignorable. However, the significance of the unintended impact should be thoroughly investigated in future research. Second, although this study demonstrated that the proposed technique improves the accuracy of stress measurement by alleviating intrinsic respiratory artifact as well as extrinsic artifacts in EDA, other types of intrinsic noises remain and can spoil EDA collected from fields. For example, muscular activity-related artifact is another representative intrinsic noise in EDA (Boucsein 2012), which this study does not focus on. Given that people have diverse and continuous muscular activities in their daily trips, a future study is needed to figure out how to suppress the muscular activity-related noise for applying wearable EDA sensing into people's daily trips.

2.2.5 Conclusion

Significant intrinsic and extrinsic artifacts in EDA collected by a wearable biosensor from fields hinder the application of EDA-based stress measurement for understanding of interaction between humans and their surroundings (e.g., CBEs). To address this issue, this study proposed a denoising technique that references PPG to alleviate intrinsic respiratory artifact as well as extrinsic artifacts in EDA. The proposed technique first attenuates extrinsic artifacts by applying several filters (e.g., high pass filter and wavelet filter). Then, intrinsic respiratory artifacts are detected and attenuated by using a subject-independent machine learning model that detects noise-inducing irregular respirations from PPG simultaneously collected with EDA. To test the denoising performance of the proposed technique, the authors conducted an in-lab data collection with ten subjects and a field data collection with 25 subjects. The subject-independent irregular respiration classifier trained by applying multisource domain-adversarial network (MDAN) showed 0.849 accuracy in LOSOCV. Also, the validity and reliability of stress metrics (i.e., BCC and ICC) calculated from EDA denoised by the proposed technique were statistically higher than ones from EDA denoised by advanced benchmark techniques. Consequentially, classification models based on the stress metrics showed statistically higher accuracy with EDA denoised by the proposed technique than with benchmark techniques. These results indicate that the proposed denoising technique can improve the stress measurement using EDA by attenuating both intrinsic respiratory artifact and extrinsic artifacts in EDA. This finding demonstrates that intrinsic artifacts with non-stationary signal characteristics can be alleviated by referencing other signals readily acquired using multi-modal wearable biosensors. The proposed denoising technique contributes to monitoring and improving the quality of the human-CBE interaction by enabling wearable EDA sensors to collect high quality signals and accurately measure stress from people's daily trips.

2.3 Denoising EEG's motion artifact

In addition to EDA's respiratory artifact, this study proposes another noise reference signal-based denoising technique to alleviate electroencephalogram's (EEG) motion artifact, the most significant artifact in EEG collected during people's daily work and lives. Different wearable-type biosensors, such as wristbands, rings, and in-ears, have been applied to continuously collect biosignals (e.g., EDA, PPG, and skin temperature (ST)) and monitor human stress during people's daily work and lives in CBEs (Ahn et al. 2019). This aims at understanding and ultimately

improving people's quality of interactions with CBEs. Among these wearable-type biosensors applied in CBEs, mobile EEG devices have a unique capability: capturing brain waves from central nervous system activities (i.e., brain activities). Monitoring biosignals related to the peripheral nervous system (e.g., EDA, PPG, and ST) can monitor the level of stress in a minimally invasive manner (Jebelli et al. 2019). However, by monitoring brain activities, we could achieve richer and more detailed psychophysiological contexts underlying human stress. For example, levels of valence (i.e., from pleasure to displeasure) and hypothalamic–pituitary–adrenal (HPA) axis activation, an emotional and neural constructs important for understanding whether the stress is negatively or positively affecting an individual (Balters et al. 2020; Dickerson and Kemeny 2004; Fischer et al. 1990; Folkman and Lazarus 1985; LeBlanc 2009; Tomaka et al. 1993), can be understood by monitoring brain activities (Russell et al. 1989). Also, brain activity monitoring enables us to track human cognitive processes, such as perceiving sounds and visual cues, salient stimuli detection, information processing, and problem solving (Michel and Koenig 2018), thereby providing in-depth understanding of how humans cognitively interact with their environments. Even though functional near-infrared spectroscopy (fNIRS) can be an alternative to EEG, EEG provides richer information that tracks activities across all brain regions including the limbic area near the center of the brain (Saha et al. 2015), which gives valuable information about physiological homeostasis (Pop et al. 2018). On the other hand, fNIRS sensing only provides information about activities that occur in the outer layer of the brain.

Despite such a unique capability, current mobile EEG sensing technology still suffers from motion-induced artifacts which make it challenging to analyze EEG signals collected in the field, where human body movements can be unpredictable and dynamic. The main source of motion artifacts in EEG is the gradient of the electromagnetic field (i.e., gradient artifacts) (Chowdhury et al. 2014). A gradient artifact is created by voltage induced at the scalp by surrounding magnetic field gradients. With head motions, electrodes move around within their surrounding electromagnetic field, introducing varying levels of electromagnetic interference (EMI) which cause fluctuations in EEG signal magnitudes. Given that EEG motion artifacts are typically much greater in amplitude than clean EEG signals, motion artifacts can lead to serious misinterpretations in EEG signal analysis (Barua and Begum 2014; Seok et al. 2021).

The most commonly applied denoising techniques to alleviate EEG artifacts are based on blind source separation such as independent component analysis (ICA) (Urigüen and Garcia-

Zapirain 2015). Blind source separation-based denoising techniques first separate raw EEG signals into multiple independent components in a way that minimizes mutual information between different components. Then, these techniques filter out independent components whose signal characteristics (e.g., amplitude and frequency) are well matched with a pre-determined artifact source template, as artifact-related components (Castellanos and Makarov 2006). These denoising techniques perform well in alleviating artifacts whose signal characteristics can be predicted and pre-defined as a template, such as ocular (Nguyen et al. 2012) and muscular artifacts (Chen et al. 2017). However, as it is practically impossible to pre-define the signal characteristics of motion artifacts due to their inherent variability and the unpredictable nature of head movements, blind source separation-based denoising is not suited for EEG motion artifact removal (Chowdhury et al. 2014).

To overcome the limitation of blind source separation-based denoising, attempts have been made to leverage motion artifact reference signals simultaneously collected with raw EEG signals, thereby alleviating motion artifacts without depending on a pre-determined noise source template. To collect reference signals, head motion data have been collected using optical motion-tracking sensors (LeVan et al. 2013) and accelerometers (Onikura and Iramina 2015). However, solely depending on these motion datasets might be insufficient for understanding motion artifacts in collected EEG signals because motion artifacts are determined by interactions between motions and their surrounding electromagnetic field (the Hall effect (Hall 1879)), which vary across contexts. In this regard, it has been recently suggested to pair reference electrodes with normal scalp electrodes as a means of collecting motion artifact references (Chowdhury et al. 2014; Luo et al. 2014; Nordin et al. 2018). In this approach, each normal electrode is paired with a reference electrode while keeping the electrical isolation between them, and thus the reference electrode records only the motion artifacts' references. Then, the motion artifact references are subtracted from raw EEG signals to acquire motion artifact-free EEG signals.

Although current paired electrode-based techniques have shown promising potential to suppress EEG motion artifacts (Nordin et al. 2018; Nordin et al. 2019), the applied reference subtraction algorithm in these current techniques might not be adaptive enough to alleviate motion artifacts resulting from unpredictable and irregular motions in real field applications. Specifically, the current technique first identifies motion artifact-governing frequencies from the reference and cancels raw EEG signals under the identified frequencies via assuming that EEG signals under

these frequencies are too governed by motion artifacts and thus do not contain any useful information about brain activity (Nordin et al. 2018). This frequency-based dichotomous approach might work when motion artifacts and clean EEG signals are clearly differentiated in frequency range. However, in applications to construction workers, motion artifacts and clean EEG signals often share a wide range of frequencies (1-10 Hz) due to the irregularity of the motion (Gwin et al. 2010; Islam et al. 2020; Shukla et al. 2020). Therefore, the current techniques might remove a significant portion of waves in EEG signals useful in understanding workers' brain activities, thereby compromising the reliability and validity of the following EEG analysis. To overcome this limitation, this study aims to develop a more adaptive reference subtraction that can seek out motion artifacts mixed with clean EEG signals over a wide range of frequencies and parallel the new reference subtraction with the paired electrode-based motion artifact reference recording, thereby enabling applications of mobile EEG at construction sites. To this end, this study first proposed an EEG motion artifact denoising technique that parallels a new adaptive reference subtraction with the paired electrodes. Then, to validate the proposed denoising technique, the denoising performance is compared with an advanced existing paired electrodes-based denoising technique on an EEG dataset collected in a lab setup where human natural motions are carefully reproduced.

2.3.1 Proposed EEG motion artifact removal

- **Acquisition of motion artifact reference**

The authors adopted the paired electrode approach (Chowdhury et al. 2014; Luo et al. 2014; Nordin et al. 2018) to record motion artifact references. A commercially available EEG device (i.e., Mentalab Explore) was customized. This device provides up to eight EEG channels on a flexible cap and has a sampling rate ranging from 250 Hz to 1000 Hz. The authors customized this device to apply our own adaptive EEG motion artifact removal technique. The authors paired electrodes: The noise-reference electrodes were flipped and attached to their paired scalp electrodes using an insulating tape, and then a conductive fabric was layered over the noise-reference electrodes to short the noise-reference electrodes and ground them like the human scalp does for the normal scalp electrodes (Figure 2.6). In this setup, the noise-reference electrodes remain electrically isolated from the scalp but experience similar motion artifacts to what the normal scalp electrodes experience, allowing the noise-reference electrodes to record only motion artifacts.

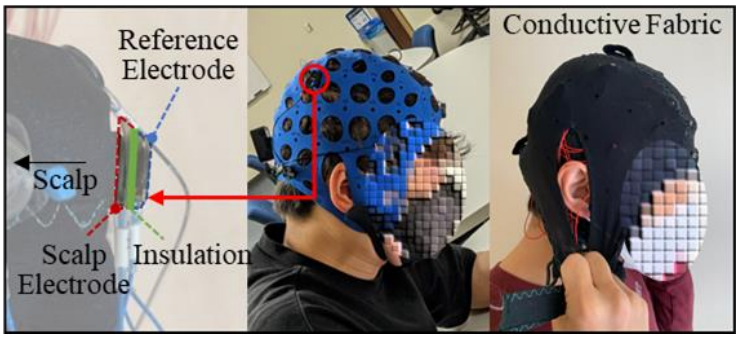


Figure 2.6. Customized mobile EEG device and details of the paired electrodes

- **Subtraction of motion artifact from raw EEG by a constrained independent component analysis with motion artifact reference**

To adaptively alleviate motion artifacts using the references collected by the paired reference electrodes, a constraint independent component analysis (cICA; Figure *2.7*) was applied. First, the raw EEG signals are synced with the motion artifact references collected by the noise-reference electrodes. Then, the cICA is applied to identify and filter out one IC whose signal shape is similar to the provided motion artifact reference, assuming that the IC results from the motion artifact. This technique is based on an empirical finding; since paired normal and reference electrodes experience almost identical electromagnetic gradients from movements, motion artifacts implicit in EEG signals collected by a normal scalp electrode can have similar signal shapes in time domains with the motion artifact reference collected by the reference electrode paired with the normal electrode (Chowdhury et al. 2014), while the amplitude scale might be different due to different electric conductance of the circuits of the normal scalp electrode and the reference electrode.

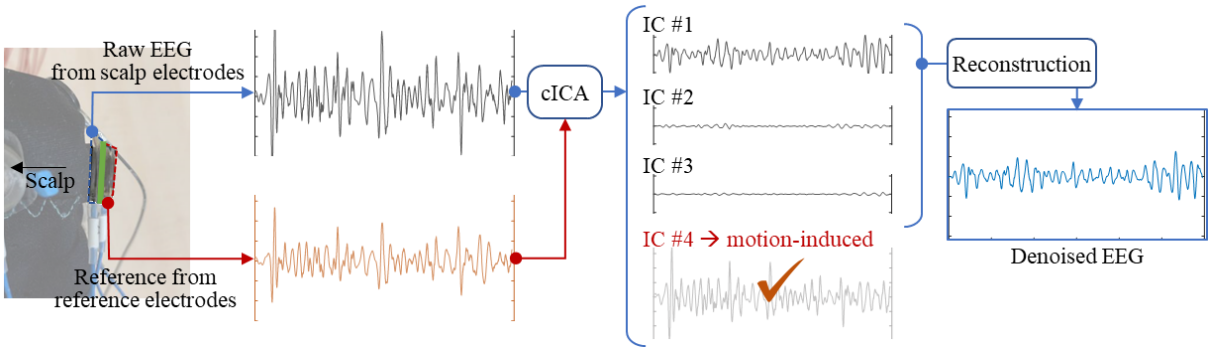


Figure 2.7. Overview of the proposed cICA-based reference subtraction

To this end, the authors designed the cICA to be sensitive to motion artifact references by having two conditions reflected as constraints in the process of identifying ICs: (i) minimized mutual information between ICs and (ii) one IC with similar signal shape to the provided motion artifact references in a scaleless manner. Specifically, the authors combined the ICA with a reference (ICA-R) (Lu and Rajapakse 2006) and the Fast ICA (Hyvärinen and Oja 2000) to realize the aforementioned cICA. Once the ICA-R extracts a weight vector corresponding to the IC similar with the provided motion artifact reference, a modified version of the Fast ICA is applied to calculate the other weight vectors corresponding to the other ICs that share minimal mutual information with the determined motion artifact-similar IC, thereby finalizing the demixing matrix. Then, by multiplying the demixing matrix and the raw EEG signals, ICs are acquired. Among the ICs, the IC identified as motion artifact-similar is linearly subtracted from the raw EEG signals, thereby denoising motion artifacts. Through this procedure, motion artifacts mixed with clean EEG signals across a wide range of frequencies are expected to be addressed.

2.3.2 The proposed denoising technique validation

In validating biosignal denoising techniques, it is challenging to obtain a ground truth/noisy biosignal pair, which is essential to comparing the ground truth with denoised signals. As a means of acquiring the ground truth EEG signals together with ones contaminated by motion artifacts, the authors applied the phantom head approach (Oliveira et al. 2016). In this approach, an EEG sensor is put on a phantom head whose shape and electrical conductivity are similar to a real human head. Then, the phantom head is set to move so as to collect a pure EEG motion artifact. The collected EEG motion artifact is linearly combined with a prepared clean ground truth EEG to make a semi-simulated noise EEG. With the pair of the ground truth EEG and the semi-simulated noise EEG, denoising performance can be quantified by comparing between the ground truth EEG and a denoised version of EEG that is acquired by applying a denoising technique to test on the semi-simulated contaminated EEG. This study statistically compared the denoising performance of the proposed technique with the most advanced existing paired electrode-based denoising technique (Nordin et al. 2019) on the dataset prepared by the phantom head approach.

- **Phantom head generation**

A phantom head was first created to authentically imitate the shape and conductivity of a human head (a in Figure 2.8). Once the skull and its inner part were created with a mixture of dental plaster, water, and sodium propionate the authors duplicated human scalp skin with 1.2-1.5-mm thick conductive gelatin. The conductivity of the plaster mixture and conductive gelatin skin were set to 0.0004 S/m and 0.3 S/m, respectively replicating the conductivity of a real human head. Also, unlike previous phantom-bead EEG denoising validation studies where an artificial regular and consistent motion was applied (Nordin et al. 2018; Oliveira et al. 2016), we intended to test real human motions. To do so, a steel plate was embedded in the bottom of the head through which real human motions could be delivered to the head.

- **Motion artifact collection and generation of semi-simulated EEG contaminated by motion artifact**

The customized mobile EEG sensor with paired electrodes was put on the phantom head and EEG signals were collected in a nosy setup with motions to collect EEG motion artifacts with the motion artifact reference. In the data collection, two electrode pairs were installed on two nodes on prefrontal cortex area (i.e., F3 and F4; b in Figure 2.8), where brain activities related to workers' stress and emotions (Hwang et al. 2018) can be monitored. The sampling rate was set by 250 Hz. Here, motions like those occurring in a construction worker's daily work at sites were provided. Specifically, a hands-free-camera-gimbal-style steel rack was connected to the steel plate embedded in the bottom of the phantom head, so that a person could comfortably wear the phantom head (d in Figure 2.8). A research staff member conducted a sandbag carrying task, a typical daily life task, to elicit natural motions while wearing the phantom head on their back. Specifically, the staff member conducted four sessions of sandbag carrying, each of which took six minutes, so that a total of 24-minute-long EEG motion artifact signals were collected. A 24-minute timespan was determined sufficient for acquiring enough samples via statistical analysis (i.e., paired t-test), which is elaborated on in the following paragraph. In each sandbag carrying session, the research staff member carried moderately weighted sandbags (10 kg) between two spots 10 m apart. Their pace was controlled to 30, 25, 20, and 15 seconds per carry over the four sessions to introduce variability in motion intensity.

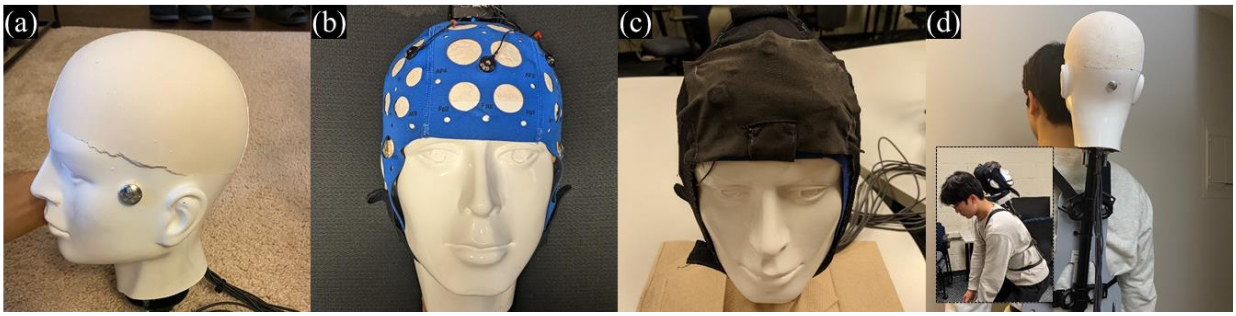


Figure 2.8. Phantom head used in this study;
(a) created phantom head; (b) normal scalp EEG equipped on the phantom head;
(c) paired reference EEG setup; (d) phantom head worn by a staff member

Once the motion artifacts were collected, they were linearly combined with a clean ground truth EEG signal to make a semi-simulated motion artifact-contaminated EEG signal. Since the motion artifacts originate from human motions which are totally independent from other legitimate brain activities, this linear superposition assumption can be valid (Islam et al. 2015; Islam et al. 2020; Shahbakhti et al. 2021). A publicly available EEG signal dataset, collected from 50 subjects under a well-controlled shielded noise-free lab condition, (Klados and Bamidis 2016) was used as a ground truth EEG signal. This EEG signal dataset has been widely used as a ground truth EEG among multiple denoising validation studies (Issa and Juhasz 2019; Mohammadpour and Rahmani 2017; Saini and Satija 2019). First, the collected motion artifact was down-sampled from 250 Hz to 200 Hz to sync with the used ground truth EEG whose sampling rate is 200 Hz. Then, a bandpass filter (0.5-40 Hz) and a north filter (50 Hz) were applied on the motion artifacts just as they were applied on the ground truth EEG. Then, the beginning and ending segments of the 24-minute-long motion artifact signals were removed to minimize the impact of the bandpass-filter-induced distortion and the middle 1250-second-long motion artifact signal data was divided into 50 25-second-long segments. These 50 segments were paired and linearly combined with 50 different subjects' ground truth EEG signals acquired from the ground truth dataset. Through these procedures, the authors had a total of 100 25-second-long semi-simulated motion artifact contaminated EEG signal samples (50 segments and 2 channels per each segment).

- **Statistical comparison with the existing technique**

As a validation, this study statistically compared the performance of the proposed denoising technique with the most advanced paired electrode-based adaptive denoising (Nordin et al. 2019)

as a benchmark. To this end, the proposed denoising technique and the benchmark were applied to the 100 noisy EEG signal samples. From each signal sample, three denoising performance metrics (i.e., signal-to-noise ratio (SNR_{dB}), root mean square error (RMSE) and cross correlation (CC)) were calculated by comparing denoised signals with ground truth EEG signals. Then, the three denoising performance metrics were statistically compared between the proposed technique and the benchmark by conducting paired t-tests. According to statistical power analysis (desired power: 0.95, alpha: 0.05, effect size: 0.35), the number of samples (100) was confirmed sufficient to conduct the planned paired t-tests. The higher denoising performance is represented by higher SNR_{dB} and CC, and lower RMSE.

2.3.3 Results

Figure 2.9 displays the different denoising results between the proposed technique and the benchmark on a sample of EEG. This visualization shows that the signal desnoised by the proposed technique is more similar to the ground truth EEG than the one denoised by the benchmark. This observation coincides with the results of the three paired t-tests; the proposed denoising technique showed statistically higher SNR_{dB}, CC, and lower RMSE than the benchmarks (p-values: almost zeros).

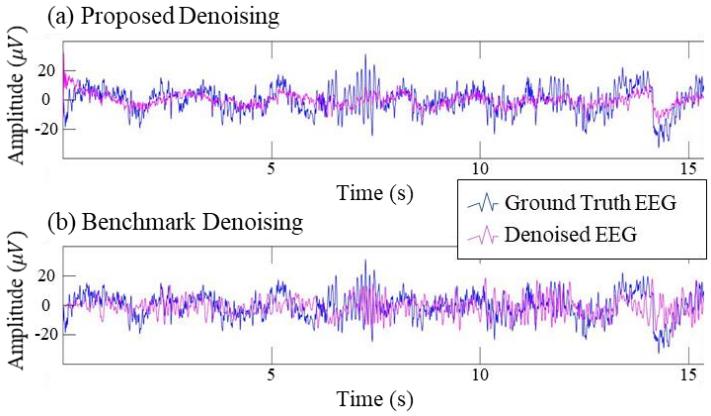


Figure 2.9. Denoising results of the proposed technique and the benchmark

Figure 2.10 shows box plots of the denoising performance metrics and the summary of the results of the three paired t-tests. In all the three compared denoising metrics, the proposed technique showed statistically higher denoising performance than the benchmark. These results

indicate that the proposed denoising technique better denoises EEG motion artifact induced by real human motions occurring during a daily life task (i.e., material handling task).

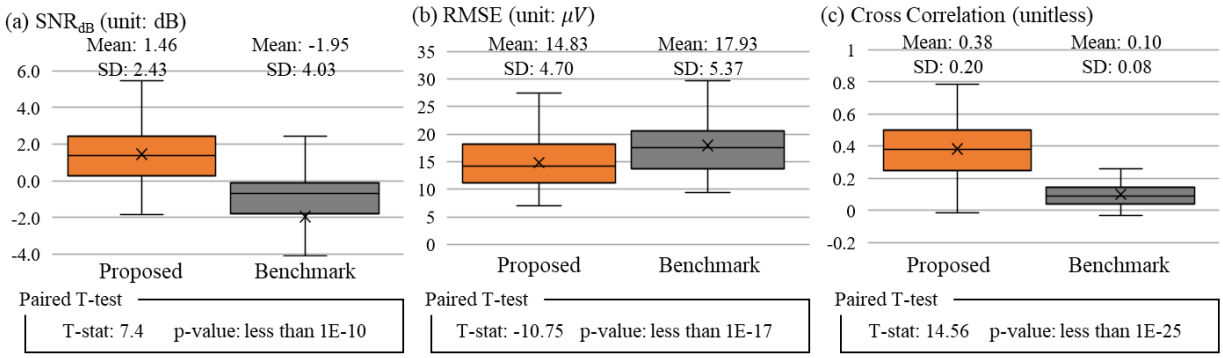


Figure 2.10. Comparison of three denoising performance metrics and the results of paired t-tests

In particular, CC shows denoising performance difference between the two denoising techniques more clearly than SNR_{dB} and RMSE. Given that CC measures how correlated the ground truth EEG and the denoised EEG are in the time domain, while SNR_{dB} and RMSE quantify how close data points of the ground truth EEG and the denoised EEG are, this result means that while the EEG signal denoised by the proposed technique well reflects the ground truth EEG signals' 'up and down patterns,' the benchmark just reduces the amplitude of the noisy EEG to fit the scale of the ground truth EEG but fails to restore the signal patterns of the ground truth.

2.3.4 Discussion

The results of the series of t-tests show that the proposed denoising technique better denoises EEG motion artifact induced by real human motions. To understand the underlying reason why the proposed technique shows better denoising performance than the benchmarks, the authors plotted a segment of the collected motion artifact reference in the frequency domain (Figure 2.11).

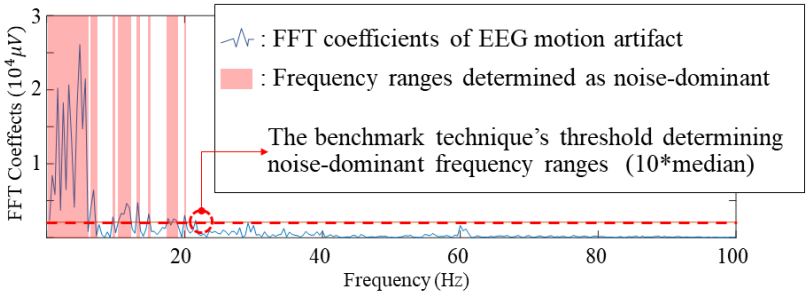


Figure 2.11. Frequency domain plot of the collected motion artifact reference

As shown in this figure, the frequency ranges determined as motion artifact-governing and thus to be removed are widely distributed ranging from 0 to 20 Hz (red marked area). These ranges are overlapped with frequency ranges useful in understanding human psychophysiological responses, such as delta (1-4 Hz), theta (4-7 Hz), alpha (8-12 Hz), and part of beta (12-30 Hz) (Holder et al. 2010). Since the benchmark denoising technique cancels the signals on the frequency ranges determined as motion artifact-governing, the informative waves in EEG signals will be eliminated, thereby compromising the following EEG analysis results.

Also, the authors examined how similar the motion artifacts are to the motion artifact references by visualizing the signals collected by a normal electrode and its paired reference electrode. Across most of the segments of collected EEG signals, the motion artifact and its reference showed a similar signal shape, but their amplitude scales differed (Figure 2.12). Through this visual investigation, the authors confirmed that the aforementioned conditions in the proposed cICA (i.e., similar signal shape in time domain with different amplitude scales) can be assumed.

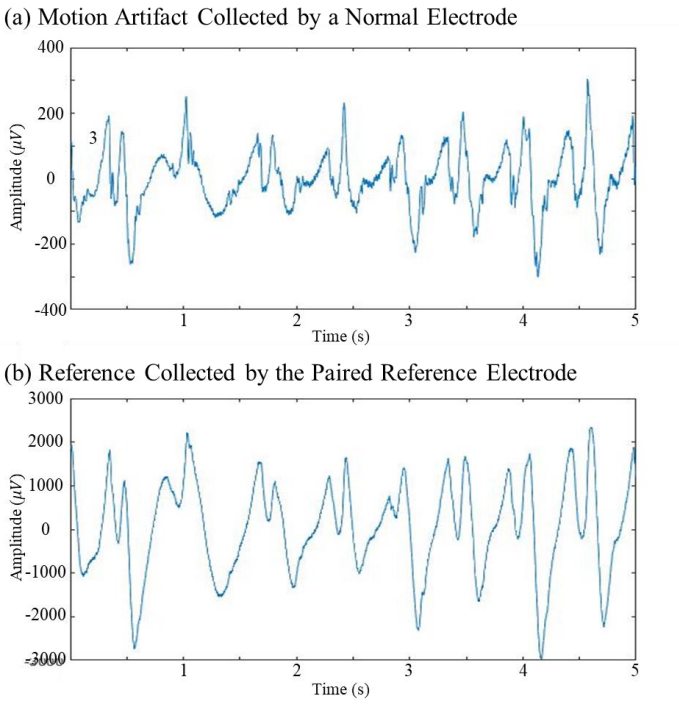


Figure 2.12. Motion artifact and its reference

These results demonstrate that the proposed denoising technique is more effective in alleviating EEG motion artifacts induced by real human motions than the existing paired electrode-

based technique. The motion artifact is one of the most significant hurdles hindering deployment of mobile EEG at people's daily interactions with CBEs. The finding of this study can therefore significantly improve mobile EEG's field applicability, and thus contribute to monitoring human brain activities during their daily work and lives, which can help with the management of peopls' quality of experiences in CBEs. Despite such significance, there are limitations in this study that should be addressed in future studies. First, even though the authors found that the proposed technique shows statistically higher denoising performance than the existing techniques, it needs to be additionally examined how significantly the improved denoising affects performance in tasks using EEG signals (e.g., EEG-based stress level classification and EEG-based cognitive load measurement). Also, this study tested two-electrodes-array only, but the required number of electrodes varies according to the type of EEG field applications. Therefore, whether the proposed technique's denoising performance is independent from the number of electrodes should also be investigated in a future study.

2.3.5 Conclusions

Paired electrodes-based reference recording has great potential to alleviate unpredictable motion artifacts in EEG signals collected by a mobile EEG device in the field. However, existing frequency-based dichotomous reference subtraction algorithms might not be effective in alleviating motion artifacts induced by real workers' motions during their field work because workers' irregularal motions spread the motion artifacts over a wide range of frequencies including frequency ranges of meaningful brain activities-induced EEG waves. To address this limitation, this study proposes a more adaptive denoising technique that conducts a constraint independent component analysis (cICA)-based reference subtraction. To compare the denoising performance of the proposed technique with an existing frequency-based dichotomous technique, the authors collected pure EEG motion artifacts by applying a construction task (i.e., material handling task)-induced real human motions on a mobile EEG-equipped phantom head and generated semi-simulated motion artifact-contaminated EEG signals by linearly combining the collected motion artifacts with clean ground truth EEG signals acquired from a publicly available EEG dataset. Then, three denoising metrics that quantify similarity between the ground truth and denoised EEG signals (i.e., SNR_{dB}, RMSE, and CC) were measured and compared between the two denoising techniques. The results showed that the proposed technique's denoising performance is statistically

higher than the existing technique. The proposed denoising technique can significantly improve the field applicability of mobile EEG, thereby enabling us to monitor human brain activity during their daily interactions with CBEs (e.g., working at construction sites and walking over streets). The EEG-based field brain activity monitoring can provide useful understanding of human psychophysiological responses to CBEs, such as stress, so that their safety, health, and productivity can be better managed.

Chapter 3 Subject- and Context-Independent Validation Method to Assess Generalizability of Machine Learning Models for Monitoring Human Responses

3.1 Introduction

The previous chapter addressed the first research agenda: acquiring high quality biosignals from the field, which is an essential step for field applicable stress detection. Another important step for field application is validating generalizability (i.e., a model's ability to work well "in general" even for situations quite different from those reflected in the training dataset (Mohri et al. 2018)) of the stress detection, so as to guarantee its general performance in field applications. Specifically, generalizability across different people and contexts should be ensured because field applications require a stress detection technique to work reliably for new individuals and contexts not considered during development.

However, generalizability is challenging to achieve in practice. While biosignals' reactivity to human psychophysiological responses (e.g., stress) might vary by individual characteristics (e.g., age, work experience, and health condition) and contextual factors (e.g., temperature, humidity, and work type) (Picard et al. 2001), it is practically impossible to collect and learn vast biosignal datasets incorporating all possible variations in individual and contextual factors. The common approach to developing generalizable psychophysiological monitoring, therefore, is to test multiple models with different structures and parameter setups expected to learn biosignal patterns that are truly generalizable and select the best model by estimating the tested models' generalizability. In this approach, accurately estimating models' generalizability with a limited dataset are crucial in developing field-applicable psychophysiological monitoring. Previously, different validation methods (e.g., k-fold, and leave one subject out) have been applied to assess the generalizability of human response monitoring techniques. However, these validation methods might overestimate the generalizability of a model because their training and testing datasets share the same people and/or the same contexts. Consequently, the model might show much lower performance in real field applications than the reported performance in the development phase.

With this background, as the first effort to ensure generalizability of the wearable biosensor-based stress detection, this chapter proposes and tests a new independent subject and context testing approach which makes sure that training and testing datasets are collected from different subjects and contexts.

3.2 Machine learning to monitor human psychophysiological responses from biosignals from wearable biosensors

As wearable technology advancements have made wearable biosensors compact and accessible enough to extensively apply at people's daily interactions with CBEs, their potential for a continuous and minimally invasive psychophysiological monitoring in the field has been paid attention. When people get an abnormal psychophysiological response such as high levels of stress, heat stress, and fatigue, such anomalies are represented physiologically such as in neural and cardiovascular activities and skin sweat production (Critchley 2002; McCorry 2007). Therefore, wearable biosensors can monitor workers' psychophysiological responses by continuously measuring biosignals (e.g., electroencephalography (EEG), electrocardiogram (ECG), photoplethysmography (PPG), and electrodermal activity (EDA)) that indicate certain psychophysiological anomalies. For example, when a person is stressed, their sympathetic nervous system is aroused and innervates heart activity and skin sweat glands in a specific pattern (Cacioppo et al. 2007). This physiological reactivity well manifests in ECG, EDA, and PPG. Also, under a high level of heat stress, a person's blood flow near the skin and their sweat rate accelerate to quickly transfer heat from the body to the skin and, ultimately, to the air (Takeda and Okazaki 2018)—a process which is straightforwardly represented in EDA, PPG, and ST.

Previously, collected biosignals were manually analyzed by experts based on their knowledge and experience, or a couple of metrics were calculated from the biosignals and analyzed as an index of specific psychophysiological responses (Cui et al. 2020). However, since biosensors and biosensor arrays have been recently advanced to collect multiple biosignals with high resolutions and sampling rates, it has become challenging to apply such manual and fragmented approaches to analyze heterogenous and highly granular biosignals. Further, even when applied, these approaches might not be able to leverage the full potential of huge biosignal datasets (Zhang et al. 2021), thereby limiting monitoring performance. Recent machine learning techniques have shown the potential to more effectively capitalize on huge biosignal datasets. These data-driven

analytics techniques enable us to translate the massive biosignals into valuable and understandable psychophysiological information by learning sophisticated computational models fed with the massive biosignals. Specifically, unlike traditional machine learning techniques that require human experts to handcraft features from raw signals based on their knowledge, recently developed deep learning techniques take the role of extracting meaningful features from raw biosignals that aid in understanding human psychophysiological responses—thereby finding new insights implicit in biosignals themselves and enriching our body of knowledge (Faust et al. 2018).

In this regard, many studies have applied wearable biosensors with traditional machine learning or deep learning techniques to understand human psychophysiological responses. For example, wearable biosensors such as a wristband for measuring EDA, PPG, and ST, and a mobile EEG head-cap have been applied to understand human stress and physical demand levels during ongoing work at construction sites (Jebelli et al. 2019; Jebelli et al. 2018; Jebelli et al. 2018) or regular offices (Sun et al. 2010) and daily outdoor trips (Kim et al. 2020). The collected biosignals were classified as different levels of stress or physical demand by applying the Gaussian support vector machine (SVM) (Jebelli et al. 2019; Jebelli et al. 2018) and the perceptron-based online multi-task learning algorithm (Jebelli et al. 2018). Drivers' mental fatigue has been measured by EDA and EEG with supervised algorithms (Lee et al. 2019; Zhang et al. 2017; Zontone et al. 2020). Human fatigue is another psychophysiological response that wearable biosensors and machine learning have been recently applied toward. For instance, ST and heart rate (HR) were collected and analyzed with SVM and logistic regression models to differentiate levels of fatigue (Aryal et al. 2017). Also, the level of perceived risk was classified as an important factor determining construction workers' safety behaviors by applying a wristband measuring EDA, PPG, ST, and several machine learning algorithms (e.g., SVM, K-nearest neighbors (KNN), and bagging tree (BT)) (Lee et al. 2021). The level of workers' distraction, an important variable affecting their productivity and ability to react to hazards, was gauged by applying a mobile EEG device with SVM (Ke et al. 2021). EEG and SVM were also applied together to predict the mental fatigue level of construction equipment operators who are often exposed to cognitively demanding tasks (Li et al. 2020). Workers' eye tracking data was collected using a glasses-type wearable biosensor and used together with EEG features and SVM algorithm to identify moments of workers' hazard recognition (Noghabaei et al. 2021). As the global average temperature has continued to increase,

workers' heat stress levels have been studied by applying wearable biosensors and machine learning as well (Shakerian et al. 2021).

These machine learning techniques are useful for understanding patterns of human psychophysiological responses from massive biosignal datasets, but it is not an easy task to develop and validate a truly field-applicable machine learning model that can work reliably in diverse situations. If a trained model is too simple, it may be unable to learn sophisticated biosignal patterns, thereby introducing bias—a circumstance called underfitting (Anderson and Burnham 2004). On the other hand, models with too much complexity tend to overlearn their training dataset. This case, called overfitting (Anderson and Burnham 2004), must also be prevented because what we need is a model with high generalizability (i.e., a model's ability to work well "in general" even for situations quite different from those reflected in the training dataset (Mohri et al. 2018)). Specifically, since biosignals' reactivity patterns to psychophysiological responses have wide individual and contextual variability (Picard et al. 2001), preventing overfitting is highly important for tasks that monitor human psychophysiological responses from biosignals. Moreover, given that people are exposed to varied contexts while working in construction sites or walking outside in cities (Biggs et al. 2013), accurately evalualting and advancing models' generalizability is crucial to applying biosensing and machine learning to monitor human psychophysioloigcal responses in CBEs.

3.3 Validation methods to ensure generalizability of the wearable biosensor- and machine learning-based psychophysiological monitoring

Several different validation methods have been introduced to assess a model's generalizability such as k-fold (KFCV), leave-one-out (LOOCV) (Efron 1982), leave-one-period-out (LOPOCV) (Molchanov et al. 2015), and leave-one-subject-out cross validations (LOSOCV) (Esterman et al. 2010). These methods share a common principle to assess generalizability, ensuring independence between training and testing datasets, and can be differentiated according to how they split training and testing datasets and what kind of independence they are specifically designed to ensure. Table 3.1 summarizes different validation methods applied in previous studies which used biosensors and machine learning to understand psychophysiological responses and Figure 3.1 visualizes the difference between validation methods in ways to split training and testing datasets. The most widely applied validation method is k-fold cross validation (KFCV). In KFCV, the dataset is

randomly shuffled and split into k equal sized subsets, one out of the k subsets is retained, and the other k-1 subsets are used as training dataset. Then, the trained model is tested using the retained one subset. Such training and testing procudures are repeated k times so that every subset is used once for testing, and then testing results are averaged. LOOCV is a unique case of KFCV, where k, the number of folds, is set equal to the number of data points so that each fold has only one data point. LOOCV does not have any randomness shuffling mechanism that makes results vary slightly, which can be an advantage compared to KFCV. However, LOOCV is barely applied nowadays mainly due to its need to train and test a model as many times as the number of total data points—making it too computationally expensive for large biosignal datasets (Molinaro et al. 2005).

Table 3.1. Use of different validation methods in previous studies into psychophysiological monitoring from biosignals

Validation Method	Authors (year)	Conducted Task	Collected Biosignals	Reported Accuracy
KFCV	Aryal et al. (2017)	Classifying the level of physical fatigue	HR, ST	0.82
	Jebelli et al. (2018)	Classifying the level of stress	EEG	0.80
	Kim and Choi (2020)	Classifying the levels of valence/arousal	EEG	valence: 0.92 arousal: 0.92
	Lee et al. (2020)	Classifying the level of stress	EDA, PPG	0.93
LOOCV	Anusha et al. (2017)	Classifying stress by stressor type	EDA	0.95
	Herlan et al. (2019)	Classifying sleep and wake	EDA	0.86
LOPOCV	Lee et al. (2021)	Classifying the level of perceived risk	EDA, PPG, ST	0.81
	Udovičić et al. (2017)	Classifying the levels of valence/arousal	EDA, PPG	valence: 0.87 arousal: 0.81
	Jaiswal et al. (2021)	Classifying the level of mental workload	EDA, PPG, ST	0.70 (0.86 by KFCV)
LOSOCV	Šalkevicius et al. (2019)	Classifying the level of anxiety	EDA, PPG, ST	0.80 (0.86 by KFCV)
	Liu et al. (2020)	Classifying the level of mental fatigue	EEG	0.73
	Bianco et al. (2019)	Classifying the level of stress	EDA, HR	0.65 (0.77 by KFCV)
	Greco et al. (2017)	Classifying the level of muscle fatigue	EDA	0.76

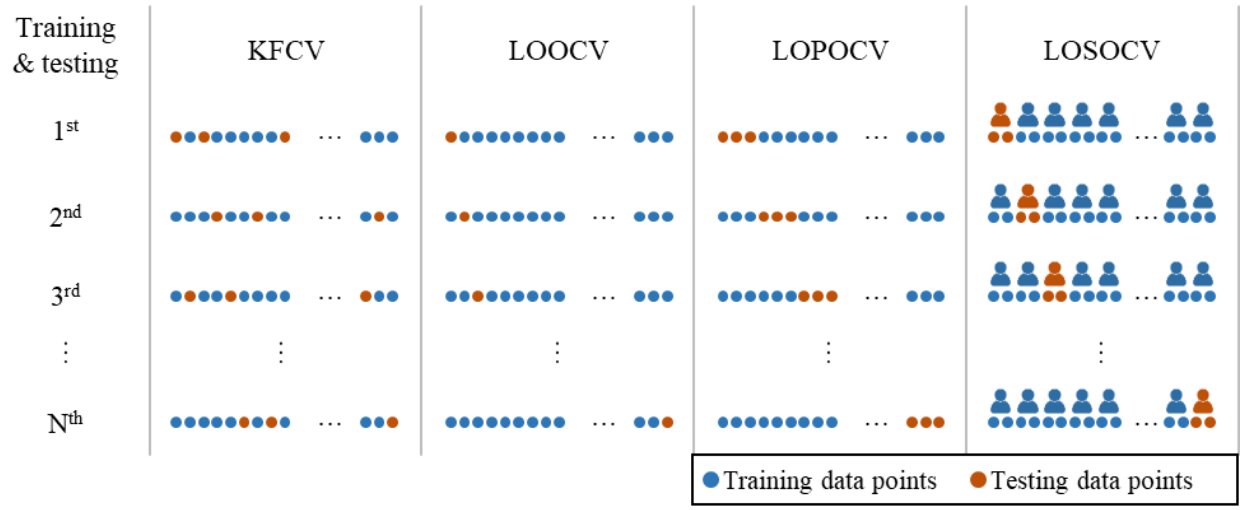


Figure 3.1. Difference in ways to split testing and training datasets between different validation methods

Since both KFCV and LOOCV assume that each data point is unique and independent from other data points, these two validation methods ensure that different data points are used between training and testing datasets. When the assumption of unique and independent data points is valid, these validation methods might perform well in estimating the generalizability of models. In tasks that monitor human responses from biosignals, however, this assumption may not be valid. Biosignals are inherently time-series data, in which the neighboring data points are not clearly independent from each other (Coniglio et al. 2020; Donate et al. 2012; Mozetič et al. 2018; Süzen and Yegenoglu 2019), meaning there is a temporal correlation among data points. Therefore, if training and testing datasets are randomly split without consideration of the data's temporal sequence, testing datasets are not truly unseen during model training. Ultimately, such cases, validation is likely to show overestimated generalizability (Lee et al. 2021).

Leave-one-period-out cross validation (LOPOCV) (or leave-one-session-out or KFCV without shuffling) was introduced to overcome the aforementioned limitations of KFCV and LOOCV. Like KFCV, LOSOCV splits a dataset into k subsets, but keeps the dataset's temporal sequence so as to minimize temporal dependence between subsets, thereby ensuring temporal independence between training and testing datasets. In other words, LOPOCV tests models using data collected from an "unseen time period" to more accurately estimate the generalizability of a model dealing with time-series data such as biosignals. Due to this difference from KFCV, LOPOCV can report more reliable generalizability estimation than KFCV (Lee et al. 2021).

Despite this improvement, LOPOCV still has a significant limitation in estimating the generalizability of human response monitoring from biosignals. Biosignals' reactivity to human psychophysiological responses might vary by individual characteristics (e.g., age, prior experience, and health condition) (Picard et al. 2001). Therefore, truly generalizable, and scalable human psychophysiological response monitoring from biosignals requires learning biosignals' general response patterns through buffering their variability between subjects. Thus, ensuring subject independence between training and testing datasets is essential for generalizability assessment, which is not guaranteed with LOPOCV, KFCV, or LOOCV.

Leave-one-subject-out cross validation (LOSOCV) has been proposed to address the drawback of the aforementioned validation methods. With LOSOCV, one subject, instead of a period or a fold, is excluded in the training phase. The left-out subject is used as a testing dataset, thereby enabling testing of trained models using data collected from an "unseen person." Because of its superior performance in estimating generalizability of human response monitoring, LOSOCV has steadily become a validation standard for studies that apply biosensors and machine learning, although LOPOCV can still be valid for subject-specific model validation (Lee et al. 2021).

However, LOSOCV still might not be an optimal method to assess the generalizability of biosensing- and machine learning-based human response monitoring. In real applications, developed models need to reliably perform not only for unseen people but also under "unseen contexts." In this study, the context is defined as the surrounding circumstances considered relevant to a person's psychophysiological responses, thereby impacting collected biosignals' patterns, as adopted from the definition of Dey (2001). Biosignals have significant contextual differences in addition to subject differences. For example, environmental factors such as ambient temperature, humidity, and radiation might affect how a person's body physiologically processes the psychophysiological responses (Heikenfeld et al. 2018). Besides the environmental factors, other contextual factors can also influence the representations of psychophysiological responses in biosignals (Spencer et al. 2019; Wusk et al. 2019; Yadav et al. 2019). In the case of stress, a representative human psychophysiological response widely studied, the type of stressor (e.g., cognitive load-related and physical exercise-related) is an important contextual factor in determining how human organs, such as the brain and heart, respond (Schlotz 2013). Because of potential contextual variability among biosignals, generalizability across different contexts needs

to be validated. LOSOCV is not a sufficient method, therefore, since it ensures subject independence between training and testing datasets but does not consider contextual independence.

In spite of such limitations of current validation methods, there is still notable paucity in the current literature about how to proactively consider individual and contextual differences among biosignals during validation so as to accurately estimate generalizability in tasks that monitor human responses from biosignals. A couple of studies have proposed a sort of "leave-one-context-out" cross validation to focus on generalizability across contexts (Saeed et al. 2018) or a specific contextual factor (e.g., leave-one-stimuli-out cross validation to ensure independence across visual stimulus to brain activities (Handjaras et al. 2016)), but these efforts cannot reliably estimate generalizability in human response monitoring wherein individual and contextual dependences coexist. Given that studies that apply biosensing and machine learning to monitoring human responses have been extensively conducted both inside and outside the construction domain and that valid generalizability estimation is the most fundamental procedure in all such studies, the need to fill this gap in the body of knowledge is urgent.

3.4 Research objectives and the proposed leave-one-subject-and-context-out cross validation

To fill the knowledge gap, this study's objective is to propose and test a new independent subject and context testing approach, which ensures that training and testing datasets are collected from different subjects and contexts. Specifically, "leave-one-subject-and-context-out cross validation" (LOSCOCV) is proposed as a new validation standard for models that monitor human psychophysiological responses from biosignals and is examined as to how well it estimates the generalizability of machine learning models compared to previous validation methods. To apply LOSCOCV, a dataset is first collected while multiple subjects experience multiple pre-determined contexts. Then, LOSCOCV excludes one subject and one context during the training phase and tests trained models using data from the excluded subject and context (Figure 3.2). Like other validation methods, LOSCOCV repeats training and testing so that every combination of subject and context are used once as a testing dataset. By splitting datasets in this manner, models can be tested with a truly unseen dataset collected from an unseen person and context, which is critical to estimate generalizability, especially for tasks that monitor human responses from biosignals.

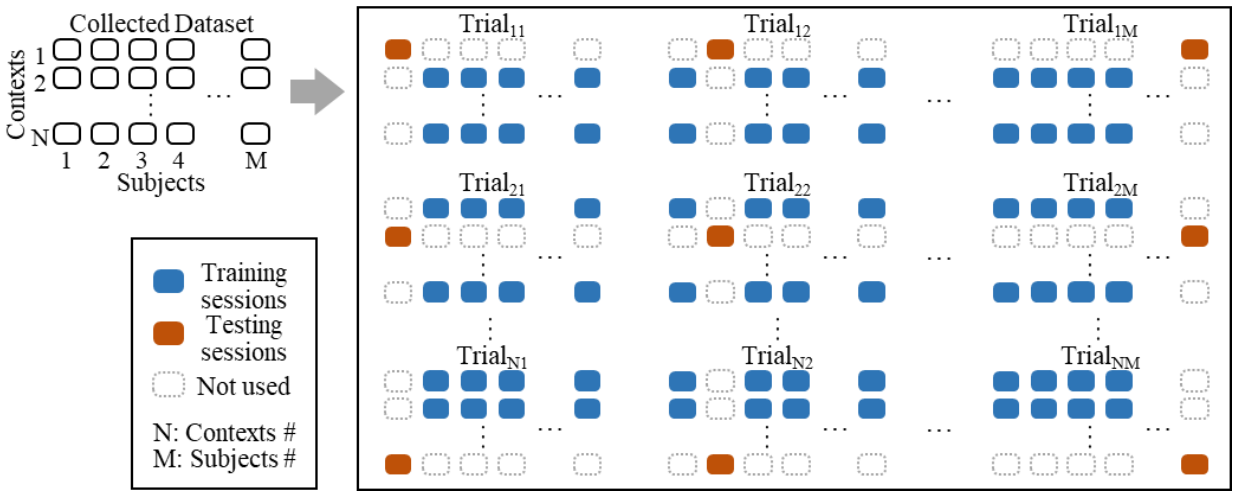


Figure 3.2. Proposed leave-one-subject-and-context-out cross validation

3.5 Test for the generalizability estimation performance of LOSCOCV

To examine the proposed LOSCOCV's performance in estimating generalizability compared to other benchmark validation methods, the authors conducted a test wherein the ground truth of models' generalizability was measured and compared with generalizability estimation made by different validation methods. Since measuring the ground truth of generalizability requires testing models with datasets that incorporate all possible situations, which is not feasible, this test adopted a proxy to examine the generalizability estimation error of different validation methods, inspired by Wu et al. (2021). An in-lab dataset was collected from a relatively limited variability in subjects and contexts. Then, a field data collection was also conducted as an example of field applications which have more individual and contextual variability and thus situations not considered in the in-lab dataset were naturally incorporated. Given the definition of generalizability (i.e., performance in more general situations), these data collection setups might enable to examine generalizability estimation errors of different validation methods by training and validating models using the in-lab dataset and comparing validation results with the models' performance with the field dataset. Following these procedures, the authors compared the proposed LOSCOCV and other benchmark validation methods in terms of their generalizability estimation error. The protocols for the two data collections were approved by the University of Michigan Institutional Review Board (IRB00000245).

3.5.1 General test setup

KFCV and LOSOCV, the two most widely applied validation methods, were applied as benchmarks. The task in the test was to detect the level of human stress, the main psychophysiological response to monitor in this dissertation. Three biosignals, EDA, PPG, and ST, were collected using a multimodal wristband-type biosensor (E4 from Empatica Inc.) and labeled as binary levels of stress (i.e., low and high stress). These three biosignals have been widely collected and analyzed to understand human stress (Jebelli et al. 2019; Nath et al. 2020; Posada-Quintero et al. 2020) because they are well-known indicators of stress-induced arousal in the sympathetic nervous system (Boucsein 2012; McCorry 2007).

In this test, Gaussian SVM and a deep neural network (DNN) were applied as the base machine learning model. Gaussian SVM was selected as a representative of traditional machine learning models because it has been proven most effective out of traditional machine learning models such as K nearest neighbors and bagging trees in understanding human stress from biosignals (Jebelli et al. 2019; Jebelli et al. 2018). 52 features used by Lee et al. (2021), which have proven useful to understand patterns in EDA, PPG, and ST according to the sympathetic arousal, an underlying stress detection mechanism, were fed into the Gaussian SVM. Also, many studies have recently found that DNN can be more effective than traditional machine learning models by learning useful features from biosignals itself without depending on human knowledge-based handcrafted features (Faust et al. 2018). Because of this capability, DNN was applied together with the Gaussian SVM in this study. The DNN's architecture (Figure 3.3) was designed based on DeepER Net introduced by Seo et al. (2019), which showed great performance at monitoring stress from raw biosignals. The designed DNN first splits a segment of biosignals into six temporal localities. Then, these temporal localities are analyzed through multiple convolutional blocks, each of which consists of 1-dimensional convolution, maxpool, batchnorm, leaky relu activation, and drop out layers in order, thereby turning into six sets of morphological features at the end of the last convolutional block. A long-short-term-memory (LSTM) layer processes the six sets of morphological features in a many-to-one manner to understand the temporal sequence thereof. Lastly, the temporal feature extracted by the LSTM layer is transferred to a classification learner, comprised (in order) of a fully connected layer, a leaky relu activation, the other fully connected layer, and a softmax layer.

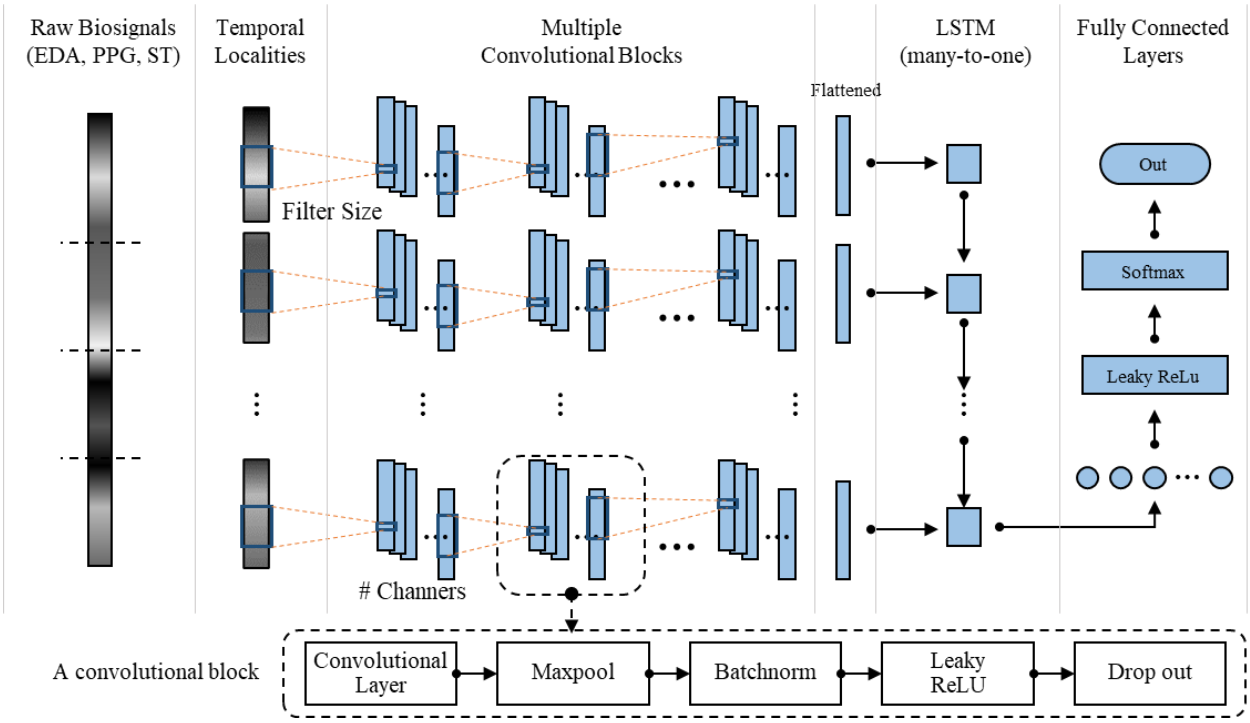


Figure 3.3. Deep neural network architecture applied in this study

The three biosignals (i.e., EDA, PPG, and ST) are first denoised following a procedure introduced by Lee et al. (2020) and then segmented by a window with a 15-second length and a 7.5-second shift so that 15-second-long biosignal segments are fed into Gaussian SVM and DNN models as data points. The window is set at 15 seconds long based on previous studies into stress detection from biosignals collected by a wearable biosensor (e.g., EDA and ECG) (Bornoiu and Grigore 2014; Zontone et al. 2019).

3.5.2 In-Lab data collection

Ten University of Michigan students were recruited as subjects for the in-lab data collection (Table 3.2). Before data collection, subjects were asked to report any physical and mental health issues that affect their psychological and physiological reactivity to stress, and none reported any such issues. Subjects individually participated in three different stress sessions within different contexts over three days. Considering that the type of stressor is one of the most important factors determining how biosignals' reactivity to stress manifests in a given context (Schlotz 2013), three different stressors (i.e., cognitive load-related, emotion-related, and physical exercise-related

stressors) were applied to constitute three different stress contexts for in-lab data collection. While subjects experienced each stress session, their EDA, PPG, and ST signals were recorded with sampling rates of 4 Hz, 64 Hz, and 4 Hz respectively.

Table 3.2. Demographic information of 10 subjects in the in-lab data collection

Statistics	Age (years)	Height (cm)	Weight (kg)
Mean (SD)	29.1 (4.1)	177.5 (6.4)	74.8 (7.4)

At the beginning of each stress session, subjects were asked to have a 10-minute relax period to minimize any prior external factors' impact on their stress and psychophysiological status and also to collect biosignals under low-stress conditions. During the relax period, a nature video clip with sound was provided via a laptop to subjects. Biosignals collected during the last half of the relax period (5 minutes) were labeled as low-stress. After the relax period, subjects experienced different stressors. On the first day, as a cognitive load-related stressor, subjects were asked to do Engle-Friedman's Math Effort Task (Engle-Friedman et al. 2003) with Advanced level (the second highest difficulty). In the task, subjects experienced five total subsessions, in each of which they were asked to continuously solve as many addition problems as possible for 2 minutes. In each problem, subjects added four numbers between 7 and 25 without the use of pencil and paper within 15 seconds. Immediately after each subsession, subjects rated their arousal using the Self-Assessment Manikin scale (Bradley and Lang 1994). Subjects' biosignals collected during subsessions rated as high arousal (6 or more on the 9-point arousal scale) were labeled as high-stress given the significant correlation between levels of stress and arousal (McCorry 2007). To ensure that participants were cognitively challenged during the task, they were asked to try to do their best and informed that they would be able to receive a performance score privately.

On the second day, subjects were asked to concentrate on three unpleasant video clips as an emotion-related stressor after the 10-minute relax period. Three bloody, conflicting, and horror video clips were presented to elicit negative-valence and high-arousal emotions such as anxiety, fear, and anger, which are expected to elicit stress according to the emotional stress model (Bong et al. 2013). Each clip lasted between two and three minutes. After watching each clip, subjects were asked to rate their emotional state while being exposed to the clip using the Self-Assessment Manikin scale (Bradley and Lang 1994). Biosignals were collected during video clips they rated

as low valence (4 or less on the 9-point valence scale), rated as high arousal (6 or more on the 9-point arousal scale), and labeled as high-stress referencing the emotional stress model (Bong et al. 2013).

On the third day, subjects repeated two physical exercises in order: arm curls and sit-to-stands, which constituted the physical exercise-related stressor. These two exercises have been widely used in studies to elicit exercise-induced stress in human subjects (Jeoung and Lee 2015; Naliboff et al. 1988). For every 5 repetitions of the exercises, a researcher asked subjects for their perceived exertion using Borg's perceived rate of exertion (PRE) scale (Borg 1982). The PRE scale has proven useful in gauging stress levels caused by physical activity (Russell 1997), echoing Borg's suggestion to use perceived exertion as an indicator of physical activity-induced stress (Borg 1970). Then, biosignals collected between time intervals in which subjects rated their perceived exertion equal to or greater than 15 (hard) and 20 (i.e., maximal exertion; the termination point) were labeled as high-stress.

3.5.3 Field data collection

Field data collection was also conducted at a building construction site in Ann Arbor, Michigan, to test models in a dataset which had much more variability than the in-lab dataset and to ultimately examine the generalizability estimation error of the validation methods considered in this study. The construction field was selected for the field data collection given that construction tasks naturally expose workers to various stressors. Also, the authors focused on incorporating more variability in the subject group. 15 construction workers with more individual variability in personal characteristics (e.g., age, work experience, and trade) than in the in-lab dataset were recruited as subjects (Table 3.3). The subjects' three biosignals (i.e., EDA, PPG, and ST) were collected using the multi-modal wristbands (E4 from Empatica Inc.), the same one used in the in-lab data collection, during their daily work for two to three hours. Also, all their work and body movements were recorded using a go-pro-type action camera attached to the front of their hardhats. The collected biosignals were labeled as low-stress and high-stress activities based on the recorded videos. Acquiring the ground truth of workers' stress during their actual work without any interference is almost impossible. Therefore, the authors collected and labeled only biosignals that clearly corresponded to low and high stress activities under the assumption that workers would experience low and high stress in such moments. To this end, two research team members who

have construction site field experience separately watched the videos to select subjects' activities that were conducted under obviously high or low stress based on their field experience and knowledge. Then, only data corresponding to activities consistently labeled as low or high stress by these two team members were used in the test. Working with heavy equipment, from a height with a fall risk, and with postural instability were the most common activities labeled as high-stress. These working conditions have been identified as typical stressors in construction sites (Goldenhar et al. 2003; Jebelli et al. 2018). Walking with or without light materials, talking with coworkers, performing plain jobs on ground level, and resting were labeled as low-stress.

Table 3.3. Demographic information of 15 subjects in the field data collection

Statistics	Age (years)	Work Experience (years)	Height (cm)	Weight (kg)
Mean (SD)	36.7 (10.8)	12.0 (9.9)	179.6 (9.8)	95.7 (12.2)

3.5.4 Statistical comparison of validity between different validation methods

To statistically compare the generalizability estimation performance between LOSCOCV and the two benchmark validation methods (i.e., KFCV and LOSOCV), a total of 160 machine learning models, consisting of 80 Gaussian SVM and 80 DNN models, were trained and tested with different hyperparameter setups (Table 3.4). The total number of models, 160, was determined by conducting a statistical power analysis using G*power 3.1 (Faul et al. 2009). A prior power analysis based on obtaining the desired power of 0.95 in mean difference analysis with one group sample, with an alpha of 0.05 and a medium effect size of 0.35 indicates that a reasonable sample size is 145. The 80 Gaussian SVM models were trained with 80 different values of kernel bandwidth from 0.6 to 16.4 in increments of 0.2, which is the range normally tested in finetuning. The 80 DNN models were trained by changing five important hyperparameters (i.e., channel number in each convolutional layer, LSTM hidden unit number, hidden node number in the fully connected layers, drop out rate, and L2 regularization) as shown in Table 3.4. The tested ranges of the hyperparameters were empirically pre-determined.

Table 3.4. Tested hyperparameter setups for deep neural network

	SVM	DNN Hyperparameters				
	Kernel Bandwidth	Convolutional Channel Number	LSTM Hidden Unit Number	Fully Connected Hidden Node umber	Drop Out Rate	L2 Regularization
Tested Values	0.6 – 16.4	20, 25, 30, 35, 40	32, 40	6, 8	0.5, 1	1e-03, 1e-04

With the in-lab dataset, all 160 models were trained and their generalizability was estimated by applying the proposed LOSCOCV as well as the two benchmarks (e.g., KFCV and LOSOCV). Then, the trained models' performance with the field dataset was measured to acquire their ground truth generalizability. Finally, by comparing the different validation methods' generalizability estimations and the trained models' performance with the field dataset, each validation method's estimation error was calculated. Accuracy, high-stress and low-stress F1 scores were measured while assessing the models' generalizability. As the index for each validation method's generalizability estimation error, this study calculated absolute error with Equation 3.1.

$$\begin{aligned} Absolute\ Error\ (\%) \\ = |Estimated\ Generalizability\ (\%) \\ - Ground\ Truth\ Generalizability\ (\%)| \end{aligned} \tag{3.1}$$

Repeated measures analysis of variance (RM ANOVA) was applied to examine the statistical significance of difference in absolute error of generalizability estimation between the different validation methods. RM ANOVA was used to compare the mean among more than two groups when observations were not independent among the groups, which is the case of this study where the authors aimed to compare the generalizability estimation error between three validation methods and the observations between the validation methods were dependent due to the use of same machine learning models. In particular, a two-way RM ANOVA was applied to additionally examine whether there is an interaction between the validation methods and the types of machine learning models (i.e., Gaussian SVM and DNN) in determining generalizability estimation error.

3.6 Results

Throughout in-lab data collection, 3,892 and 4,927 data points were collected and labeled as low-stress and high-stress respectively. From field data collection, a total of 9,383 data points were collected of which 4,906 and 4,477 data points were labeled as low-stress and high-stress respectively. Using these two datasets, 160 machine learning models were trained and tested, thereby measuring the absolute error of generalization performance estimation for the proposed LOSCOCV method and the two benchmarks (i.e., KFCV, LOSOCV). Table 3.5 shows descriptive statistics of the different validation methods' absolute errors. LOSCOCV showed much lower absolute error than the two benchmarks across all three performance metrics (i.e., accuracy, F1 scores of the two stress classes).

Table 3.5. Descriptive statistics of absolute errors in generalizability estimation of validation methods

Metrics	Statistics	KFCV	LOSOCV	LOSCOCV
Absolute Error of Accuracy	Mean	20.6%	12.3%	7.2%
	SD	11.2%	6.2%	6.2%
	Median	21.0%	10.5%	5.8%
	Min	1.5%	4.5%	0.7%
	Max	44.6%	33.8%	24.2%
Absolute Error of Low-stress F1 Score	Mean	16.1%	8.9%	5.7%
	SD	8.7%	5.6%	4.8%
	Median	15.4%	7.4%	3.2%
	Min	1.8%	3.9%	0.1%
	Max	39.2%	31.7%	19.6%
Absolute Error of High-stress F1 Score	Mean	29.5%	19.6%	15.5%
	SD	19.3%	12.0%	9.5%
	Median	29.9%	17.7%	13.6%
	Min	0.1%	3.0%	0.7%
	Max	88.1%	60.3%	46.7%

The results of RM ANOVA showed that the difference in absolute error between different validation methods is statistically significant. Table 3.6 displays the results of RM ANOVA. The Mauchly's test for sphericity showed that across all the three performance metrics (i.e., accuracy, high-stress and low-stress F1 scores), the sphericity assumption was violated, and so the Greenhouse-Geisser and Huynh-Feldt p-values need to be considered to examine statistical significance. Regardless of performance metric, the calculated Greenhouse-Geisser and Huynh-

Feldt p-values for the validation method were less than 1e-04, which means the absolute error is significantly different between validation methods. The results of post-hoc paired t-tests that were conducted to compare absolute errors between different validation methods coincide with ones of RM ANOVA, except when comparing DNN models' error in low-stress F1 score between LOSOCV and LOSCOCV DNN (Table 3.7).

Table 3.6. Results of repeated measures analysis of variance

Absolute error in accuracy							
	Sum Sq	DF	Mean Sq	F-stats	P-value		
					Sphericity assumed	Greenhouse-Geisser	Huynh-Feldt
Validation method	0.157	3	0.052	25.424	≈ 0	≈ 0	≈ 0
Validation method × Model type	0.598	3	0.199	96.556	≈ 0	≈ 0	≈ 0
Error (validation method)	0.978	474	0.002	1	0.5	0.5	0.5
Absolute error in high-stress F1 score							
	Sum Sq	DF	Mean Sq	F-stats	P-value		
					Sphericity assumed	Greenhouse-Geisser	Huynh-Feldt
Validation method	0.088	3	0.029	17.596	≈ 0	≈ 0	≈ 0
Validation method × Model type	0.328	3	0.109	65.705	≈ 0	≈ 0	≈ 0
Error (validation method)	0.788	474	0.002	1	0.5	0.5	0.5
Absolute error in low-stress F1 score							
	Sum Sq	DF	Mean Sq	F-stats	P-value		
					Sphericity assumed	Greenhouse-Geisser	Huynh-Feldt
Validation method	0.174	3	0.058	16.512	≈ 0	≈ 0	≈ 0
Validation method × Model type	0.646	3	0.215	61.488	≈ 0	≈ 0	≈ 0
Error (validation method)	1.661	474	0.004	1	0.5	0.5	0.5

*≈ 0: less than 1e-04

Also, the results of RM ANOVA showed that the p-values for the validation methods × tested model types were less than 1e-04, which means that absolute error was associated with the interaction of validation methods and applied machine learning model types. To examine the

interaction effect, in Figure 3.4, the authors draw boxplots to visualize differences in absolute error between different validation methods in three different cases: i) all 160 models, ii) only the 80 DNN models, and iii) only the 80 Gaussian SVM models are considered. The figure shows that the gaps in absolute error between different validation methods are bigger when Gaussian SVM models are applied (gaps in accuracy absolute error between validation methods: 20.7%, 11.7%, and 9.0%) than when DNN models are applied (6.1%, 4.9%, and 1.2%).

Table 3.7. Results of post-hoc paired t-tests

Tested Metric	Tested Model Type	Sample Number	P-value		
			KFCV & LOSOCV	LOSOCV & LOSCOCV	KFCV & LOSCOCV
Accuracy	All	160	≈ 0	≈ 0	≈ 0
	Gaussian SVM	80	≈ 0	≈ 0	≈ 0
	DNN	80	≈ 0	0.0019	≈ 0
High-stress F1 score	All	160	≈ 0	≈ 0	≈ 0
	Gaussian SVM	80	≈ 0	≈ 0	≈ 0
	DNN	80	≈ 0	≈ 0	≈ 0
Low-stress F1 score	All	160	≈ 0	≈ 0	≈ 0
	Gaussian SVM	80	≈ 0	≈ 0	≈ 0
	DNN	80	≈ 0	0.164	≈ 0

*≈ 0: less than 1e-04

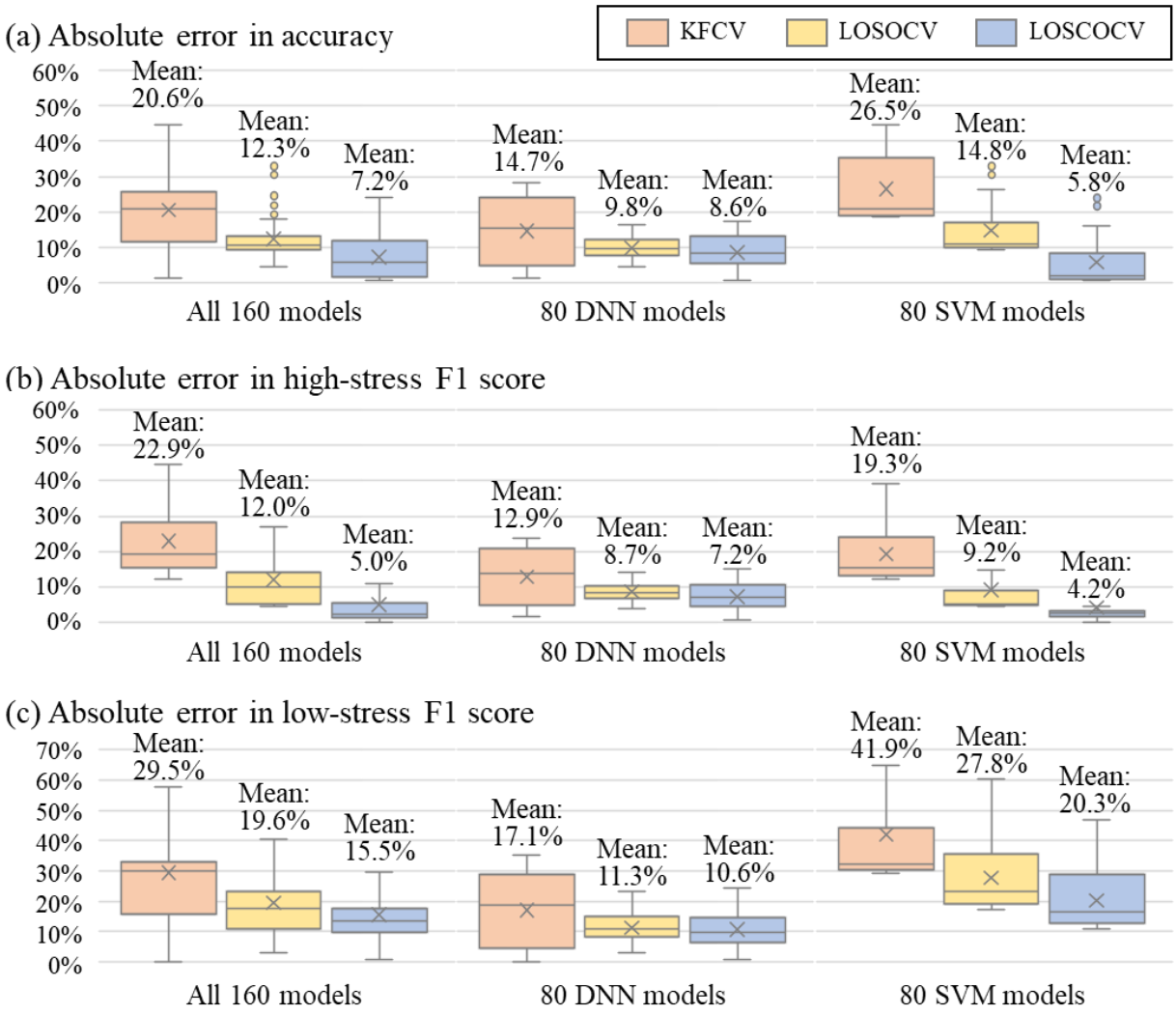


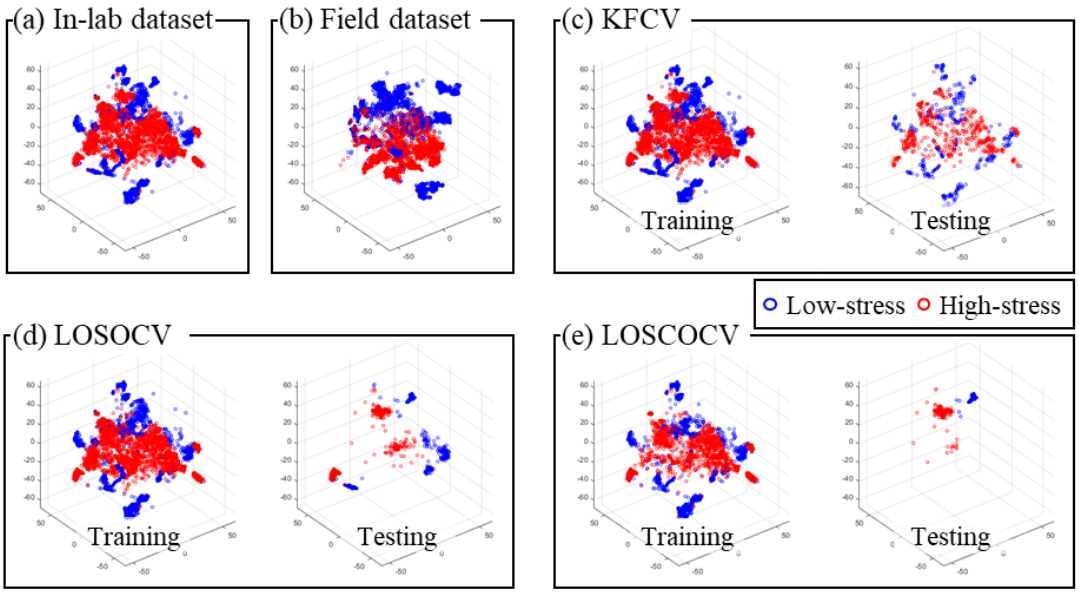
Figure 3.4. Absolute errors in generalizability estimation by performance metrics and tested model types

3.7 Discussion

To reliably estimate generalization performance of machine learning models for tasks that monitor human psychophysiological responses (e.g., stress) from biosignals, subject- and context-independence between training and testing datasets must be ensured, which current validation methods lack. Therefore, this study proposes a new subject- and context-independent validation method, leave-one-subject-and-context-out cross validation (LOSCOCV) and tests its generalizability estimation performance compared to two benchmark validation methods, k-fold and leave-one-subject-out cross validations (KFCV and LOSOCV) that have been most widely

applied in current studies. The results showed that the proposed LOSCOCV estimates the generalizability of machine learning models with statistically higher performance than the two benchmarks.

The authors examined the underlying reason why the proposed LOSCOCV method better estimated machine learning models' generalizability than the two benchmarks. To do so, the authors visualized different datasets (i.e., in-lab and field datasets and the training and testing datasets split by validation methods from the in-lab datasets) on the trained feature vector and examined how well the testing and training datasets reproduced independency between the in-lab and field datasets (Figure 3.5). Specifically, a DNN model was used for this visualization, not a Gaussian SVM model, because it is possible to intuitively examine how a DNN model analyzes datasets by visualizing them on the feature vector extracted by the trained feature extractor—a meaningful intermediate output of the trained DNN model. Therefore, comparing data distributions of the in-lab and field datasets on the learned feature vector enables us to see how generalizable the trained model would be. Also, by comparing the distributional difference between training and testing datasets split by a validation method with one between the in-lab and field datasets, it would allow to guess how closely the validation method estimates the generalizability. Such a visualization-based interpretation approach cannot be applied to Gaussian SVM because Gaussian SVM does not have any intermediate output of the training result to visually investigate.



* 3 dimensions of this figure were calculated for visualization purpose from the original feature vector of a DNN model using T-Distributed Stochastic Neighbor Embedding, a non-linear dimensionality reduction method.

Figure 3.5. Data distributions of training and testing datasets in validation methods and the field dataset (testing subject: #1 and testing context: emotion-related stress context for LOSOCV and LOSCOCV)

Feature vector was first extracted by the multiple convolutional blocks and LSTM layer of a DNN model trained using the in-lab dataset (hyperparameters setup: 20 channels in each convolutional layer, 32 LSTM hidden units, 6 hidden nodes in the fully connected layers, 1.0 drop out rate, and 1e-03 L2 regularization). T-Distributed Stochastic Neighbor Embedding, a non-linear dimensionality reduction technique developed to visualize high-dimensional data (Bakker and Heskes 2003), was then applied to reduce the dimension of the extracted features to three for visualization.

Given that the data distributions of the different stress levels on the learned feature map are quite different between the in-lab dataset (a in Figure 3.5) and the field dataset (b in Figure 3.5), the trained DNN model might not be successful at buffering individual and contextual differences and, therefore, may not be truly generalizable. On the other hand, it is obvious that KFCV will overestimate generalizability because the data distributions are similar between training and testing datasets (c in Figure 3.5), meaning that KFCV's way to split training and testing datasets fails to duplicate the difference between the analyzed dataset and datasets in field applications. Although LOSOCV showed much higher difference between training and testing

datasets than KFCV, a considerable portion of data points in the training and testing datasets are positioned in similar areas on the feature space (d in Figure 3.5), thereby inducing generalizability overestimation. Unlike the two benchmark validation methods, LOSCOCV's training and testing datasets show quite different distributions (e in Figure 3.5). The significant difference in training and testing independency between LOSOCV and LOSCOCV might be because LOSCOCV ensures context independency in addition to the subject independency between training and testing datasets, which is not guaranteed in LOSOCV.

According to the results of RM ANOVA, the absolute error in generalizability estimation is associated with the interaction between the validation methods and the applied model types (e.g., Gaussian SVM and DNN). As shown in Figure 3.5, bigger gaps in absolute error are observed between different validation methods when Gaussian SVM models are applied than when DNN models are. This discrepancy can be explained by DNN models' superior ability to prevent overfitting. In the applied DNN architecture, multiple measures aimed to prevent overfitting were introduced such as the batchnorm, pooling and dropout layers in each convolutional block, and L2 regularization. Thanks to these anti-overfitting measures, the trained DNN models are less likely to overfit the training dataset. Consequently, even in cases where training and testing datasets are somewhat dependent, the degree of generalizability overestimation might be less when DNN models are applied than when Gaussian SVM models are applied. A Gaussian SVM with an appropriate kernel bandwidth has proven effective in alleviating the overfitting issue (Jebelli et al. 2018), but since a wide range of kernel bandwidth was tested in this study (from 6.0 to 16.4), some of the trained gaussian SVM models came to be highly overffitted to the training dataset, thereby making a significant difference in the generalizability estimation between model types.

The proposed LOSCOCV method showed statistically higher generalizability estimation across all the observed metrics (i.e., accuracy, high-stress and low-stress F1 scores), yet showed quite large absolute error in estimating low-stress F1 scores (mean: 15.5%) unlike the other metrics (mean absolute errors in accuracy and high-stress F1 score: 7.2%, 5.7%). This estimation performance gap between the low-stress F1 score and the other metrics might result from the distributional characteristic of the analyzed stress datasets where the low-stress data points have smaller variability than high-stress ones. In both in-lab and field data collection, low-stress data points were collected in a few very limited conditions (e.g., when subjects relaxed or lightly walked) to ensure subjects were, indeed, under low-stress conditions, while high-stress data points

were collected from a relatively wider range of conditions. For instance, subjects might naturally experience varying levels of cognitive load and displeasure during the math task and emotional video clips. Also, during the physical exercise, high-stress data points were collected from varying levels of physical exertion (i.e., 15 to 20 by Borg's PRE). Consequently, low-stress data distribution from this part of the study comes to be tight compared to the high-stress one. On the other hand, the space on the feature map where the tight low-stress data distribution is positioned varies depending on the surrounding environmental factors such as ambient temperature, humidity, and radiation. In this regard, it is extremely challenging to accurately estimate a model's low-stress F1 score in the field dataset by conducting cross validation using the in-lab dataset because the in-lab dataset was collected in a controlled indoor setup where environmental factors are quite stable, while the field dataset was collected from an outdoor construction site whose environmental factors are quite different from the controlled indoor setup and also change dynamically. However, this result indicates a drawback of the in-lab dataset used in cross validation, not of the proposed LOSCOCV method. Also, it is noteworthy that if environmental factors are considered in the in-lab dataset for establishing contextual independency, LOSCOCV's generalizability estimation performance might be significantly increased, but the benchmarks would not because LOSCOCV is the only validation method that ensures contextual independency between testing and training datasets.

This study shows that the proposed LOSCOCV method can provide more reliable generalizability estimation of machine learning models than currently applied validation methods such as KFCV and LOSOCV for tasks that monitor human status from biosignals. This is explained by LOSCOCV's capability to test generalizability across not only subjects, but also contexts. This unique strength of LOSCOCV might be more significant when wearable biosensors and machine learning are applied for monitoring human psychophysiological responses to CBEs. This is because while interacting with CBEs, people are exposed to varied contexts it varies by context how psychophysiological responses manifest in biosignals.

Given that ensuring generalizability is critical in field application and that accurately tracking generalizability is fundamental to efforts to advance generalizability, the LOSCOCV method can significantly contribute to helping researchers develop truly field applicable human psychophysiological monitoring. Understanding human psychophysiological responses in a continuous and less-invasive manner is critical to advancing their quality of experience in CBEs.

Therefore, the findings of this study can contribute to promoting the quality of interaction between humans and CBEs by enabling wearable biosensing- and machine learning-based psychophysiological monitoring to be applied to them during their daily work and lives in CBEs.

Despite its contributions, this study has several limitations that need to be addressed by future studies. First, environmental factors such as ambient temperature, humidity, and radiation were not incorporated into the in-lab datasets used in testing different validation methods. Given that these factors may be important in duplicating the contextual difference of biosignals in real field applications, a future study incorporating the environmental factors in establishing contextual difference in an in-lab dataset can further verify the results of this study. Second, the labeling method used for the field stress dataset may be limited in fully measuring construction workers' stress levels. This study selectively collected and labeled biosignals corresponding to activities labeled as low or high stress consistently by two independent observers. Workers' stress levels are a subjective construct, which can also be affected by outwardly invisible internal factors, such as the recollection of previous unpleasant experiences. Therefore, incorporating a minimally invasive means of hearing workers' perceptions might help us more accurately label their stress levels.

3.8 Conclusion

Given that biosignals' patterns according to human psychophysiological responses, such as stress, have individual and contextual variability, the objective of this study was to propose leave-one-subject-and-context-out cross validation (LOSCOCV) as a new validation method that tests machine learning models using data collected from unseen people and contexts, thereby more validly assessing a models' generalizability. The proposed LOSCOCV method's generalizability estimation performance was compared with currently applied validation methods by conducting a test where machine learning models were developed to detect construction workers' stress levels from biosignals collected during their ongoing work. As a result, the proposed LOSCOCV method showed statistically lower error in estimating machine learning models' generalizability than the benchmarks across all tested performance metrics (i.e., accuracy, high-stress and low-stress F1 scores). The results indicate that testing machine learning models using unseen subject and unseen context dataset is crucial to assessing generalizability, and so the proposed LOSCOCV method can be more valid than currently applied machine learning validation methods for tasks that monitor human responses from biosignals. Generalizability is the most critical quality of field

applicable human response monitoring, and validly assessing generalizability is fundamental to efforts to advance machine learning models' generalizability. The finding of this study can therefore significantly contribute to the field applicability of wearable biosensors and machine learning to monitor human psychophysiological responses, ultimately and overall advancing their health, safety, comfort and productivity in CBEs.

Chapter 4 Deep Learning Domain Adaptation-Based Subject- and Context-Independent Stress Detection

4.1 Introduction

As the first effort toward ensuring generalizability of wearable biosensor-based stress detection, the previous study presented a new subject- and context-independent validation method, the leave one subject and context out cross validation (LOSCOCV). Then, I found that the proposed LOSCOCV can more reliably assess machine learning models' generalizability in tasks monitoring human responses from biosignals. Following up this effort, the current chapter proposes a new transfer learning-based stress detection technique that actively advances models' generalizability performance by buffering domain differences between different people and contexts.

Achieving a high level of generalizability is challenging with the current machine learning-based stress detection techniques. Biosignals' reactivity to human stress vary by individual characteristics (e.g., age, work experience, and health condition) and contextual factors (e.g., temperature, humidity, and work type) (Picard et al. 2001). However, the current machine learning techniques just model stress-specific patterns in biosignals by learning existing training datasets. Therefore, these techniques might not work accurately for a new person and/or new context that was not part of the training datasets. Previous wearables- and machine learning-based stress detection studies have echoed this issue by showing results that model performance in the between-subjects (e.g., leave one subject out cross validation (LOSOCV)) and between-contexts tests (e.g., leave one context out cross validation (LOCOCV) is significantly lower than in the within subject and context test (e.g., k-fold cross validation (KFCV)). For example, in (Bianco et al. 2019), an ensemble model's stress level detection accuracy in KFCV and LOSOCV was 77.3% and 63.2% respectively. In (Saeed et al. 2018), the Kappa score, an accuracy index for multi-class classification, of a neural network model in classifcaying the level of stress ranged 0.71-0.86 and 0.04-0.64 in KFCV and LOCOCV respectively. Such subject- and context-dependency significantly hinders scalable field deployment of wearable biosensors- and machine learning-

based stress detection because it entails collecting labeled biosignals for all targeted people and contexts.

4.2 Transfer learning to advance generalizability across subjects and contexts

Recent research in transfer learning has paved a way for overcoming subject- and context-dependency in wearable biosensors- and machine learning-based stress detection. Transfer learning is a manlike learning strategy that trains models to solve a task by leveraging information gained from previously trained models for similar tasks while effectively buffering differences between previous and current tasks (Murre 2014). If detecting stress from a person in a context is viewed as a task, detecting stress from a different person in a different context can be seen as a similar task which can share previously learned information given that the physiological mechanism underlying the biosignals' reactivity to stress is commonly sympathetic arousal (Cacioppo et al. 2007), despite individual and contextual variabilities in biosignals' reactivity to stress. From here, transfer learning might be able to buffer individual and contextual differences in biosignals so that previously trained models can perform reliably for new people in new contexts. As such, attempts have been made to apply transfer learning to subject- and/or context-independent stress detection techniques.

One representative transfer learning technique applied to advance subject- and context-independency in detecting people's stress is multi-task learning (Jaques et al. 2017; Jebelli et al. 2018; Lee et al. 2020; Sakri et al. 2018; Stewart et al. 2020; Taylor et al. 2020). Multi-task learning is one type of transfer learning wherein multiple models are learned for multiple similar tasks while sharing information across tasks through similarity constraints (Taylor et al. 2020). Here, sharing information indicates that a model for a task is learned using labeled data points collected from other tasks through buffering differences between tasks. Thus, multi-task learning can reduce the burden of collecting labeled training data for detecting a new person's stress in a new context by leveraging data collected from other people and contexts in a subject- and context-independent manner. In fact, previous studies have proposed multi-task learning-based stress detection techniques and have found them effective in detecting stress for multiple different people while needing considerably less labeled data from each person than typical machine learning algorithms (Jaques et al. 2017; Jebelli et al. 2018; Lee et al. 2020; Sakri et al. 2018; Stewart et al. 2020; Taylor et al. 2020). However, multi-task learning still requires collecting some labeled data for all targeted

people and contexts though the required amount of labeled data to train reliable models might be much less than normal machine learning algorithms. When a model learns a data point collected from a task other than the task it was assigned to, the extent to which the model can learn from this new data point depends upon the similarity between the two tasks (i.e., the data point's task and the learning model's task) (Saha et al. 2011). Since task similarity is measured by comparing the learning models' paramaters or distributions of labeled data between tasks, at least some labeled data must be collected from every task. Multi-task learning, therefore, cannot eliminate the key barrier to wide field application of wearable biosensor stress detection: needing labeled data collection from all targeted people and contexts.

To ensure generalization across subjects and contexts for scalable application of wearable biosensors- and machine learning-based stress detection, the authors propose applying domain adaptation instead of multi-task learning. Domain adaptation is another type of transfer learning which can learn a robust model by leveraging labeled training data from other source tasks that are similar to the target task without having any labeled data from the target task (Duan et al. 2012). Domain adaptation is based on an empirical finding that the difference between tasks (domains) manifests as the discrepancy in data distribution between tasks. So, if we find an optimal feature vector on which the data distribution discrepancy among different tasks is minimized, tasks might be able to share one common classification model on the optimal feature vector. In this sense, domain adaptation is an optimization problem seeking an optimal feature vector to minimize discrepancies in data distribution among target and source tasks while maximizing the classification performance in source tasks.

If a person/context combination is viewed as a task (or, domain), domain adaptation can be applied to develop a model that detects a person's stress in that context using labeled stress data collected from different people and contexts. In other words, it might be possible to propose a subject- and context-independent stress detection technique in which domain adaptation can provide a stress detection model for a new person and new context by leveraging labled people and context data we already have. This technique can significantly advance field applicability of wearable biosensors- and machine learning-based stress detection by rendering labeled data collection for new people and contexts unnecessary.

Noting this potential, previous studies have attempted to apply domain adaptation to advance generalization in detecting stress (or, similar psychophysiological status) across different

people (Jaques et al. 2017; Lan et al. 2019; Saeed et al. 2018). Still, there are knowledge gaps to address to realize the full potential of domain adaptation for subject- and context-independent stress detection. First, the previous studies have conducted feature extraction and domain adaptation in two separate steps which might lead to suboptimal solutions because domain adaptation optimization objectives are not considered in the feature extraction step. Second, the previous studies have applied single source domain adaptation wherein a single source task is paired with a target task. This type of domain adaptation might not be suitable for the task of detecting stress. Given the variabilities in biosignals' reactivity to stress according to different people and contexts, incorporating such variabilities into source domains is key to ensuring trained models' performance in target domains (Picard et al. 2001). Therefore, it is necessary to conduct a "multi source" domain adaptation having enough numbers of different people and contexts as the labeled source domains. Third, although subject- and context-independencies are equally important for wearable biosensors- and machine learning-based stress detection to be applied extensively in the field, previous studies have focused on only one out of the two (i.e., subject-independency (Lan et al. 2019; Taylor et al. 2020), context-independency (Saeed et al. 2018)). No studies have yet considered whether domain adaptation can work in cases where both individual and contextual differences coexist.

To fill these gaps, the objective of this study is to present and test a subject- and context-independent stress detection technique that conducts "end-to-end" "multi source" domain adaptation. To achieve the objective, this study first proposes a deep neural network-based domain adaptation as the core of the proposed stress detection. This domain adaptation learns not only subject- and context-independent feature vectors from raw biosignals collected from multiple source domains, but also a classifier as a set in an end-to-end manner. Subject- and context-independency of the proposed stress detection technique was assessed through a series of tests wherein subject- and/or context-independencies between training and testing datasets were ensured.

4.3 Proposed subject- and context-independent psychophysiological monitoring technique

To reliably detect an unseen person's stress within an unseen context in a subject- and context-independent manner, the proposed technique first prepares a labeled stress dataset collected from multiple people and contexts as labeled source domains. Then, with a new person in a new context

(i.e., an unlabeled target domain), the proposed technique collects unlabeled data from the new person and context. Using data from both the one target domain and multiple source domains, the proposed technique learns a subject- and context-independent feature vector and a stress classification model simultaneously in an end-to-end manner by applying a specially designed deep neural network (DNN). Then, on the learned independent feature vector, the learned classification model classifies stress levels for the new person in the new context. In this manner, it might be possible to reliably detect a new person's stress in a new context without additional labeled data collection.

The key component in the proposed stress detection process is the specially designed DNN. A unique strength of applying DNN for domain adaptation is that we can conduct feature extraction, domain adaptation, and stress classifier learning all together in an end-to-end manner, thereby effectively finding an optimal subject- and context-independent feature vector and a stress classifier as one set. Also, it is possible to straightforwardly account for multiple different source domains in the network's design. (See subsections 4.3.1 and 4.3.2)

4.3.1 Model Architecture

The proposed DNN aims to simultaneously learn a stress classifier and a domain-adapted subject- and context-independent feature vector from raw biosignals. To this end, the proposed model architecture is comprised of i) a feature extractor (G_f in Figure 4.1), ii) a stress classifier (G_s in Figure 4.1), and iii) a domain classifier (G_d Figure 4.1), inspired by Ganin et al. (Ganin et al. 2016).

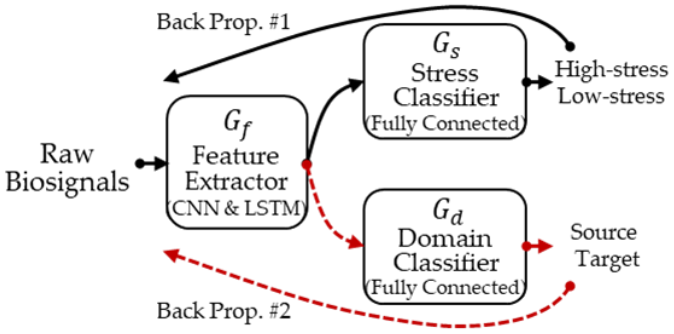


Figure 4.1. Overview of the proposed deep neural network architecture

The stress classifier learns how to successfully classify different levels of stress in the labeled source domains using features extracted by the feature extractor. The feature extractor

learns how to extract a feature vector satisfying two different objectives according to the domain adaptation's general setup. The feature vector should: i) be useful in classifying different stress levels in source domains and ii) minimize the discrepancy in data distribution of target and source domains. The first learning objective can be digested by training along the first forward/backward propagation path toward the stress classifier's output (Back Prop #1 in Figure 4.1). To enable the feature extractor to learn the second objective, a domain classifier is introduced to determine which domain each data point is from. The domain classifier is paired with the stress classifier in the architecture as shown in Figure 4.1, constituting the second forward/backward propagation path (Back Prop. #2 in Figure 4.1). Training along the second path to the domain classifier's output is intended to maximize the error of domain classification. This learning direction is based on an empirical finding: if we learn a feature vector that maximizes domain classification error, then the learned feature vector is domain-adapted such that the discrepancy of data distribution of different domains is minimized and thus different domains can share a common stress classifier on the feature vector (Ben-David et al. 2007). Training for the proposed DNN's three components is described in detail in the next subsection (4.3.2). Once training is done, i) the trained feature extractor extracts a domain-adapted subject- and context-independent feature vector and ii) the stress classifier classifies the level of stress in the target domain on the extracted feature vector.

The feature extractor is structured to learn the optimal feature vector from three raw biosignals (i.e., EDA, PPG, and ST) by deploying multiple 1-dimensional convolutional blocks and one long-short-term-memory (LSTM) layer as shown in Figure 4.2. 1-dimensional convolutional blocks have been found useful to extract meaningful features indicating morphological characteristics of each temporal biosignal locality (Seo et al. 2019). An LSTM layer can obtain sequential information about features extracted from each temporal locality by convolutional blocks, which is crucial for profiling biosignals' different patterns (Masood and Alghamdi 2019; Seo et al. 2019).

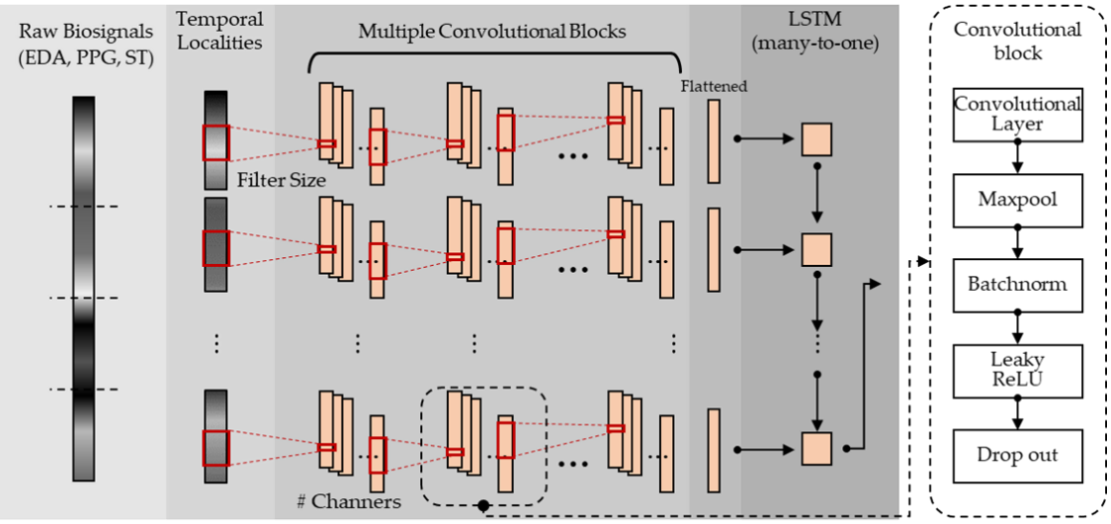


Figure 4.2. Detailed structure of the feature extractor (G_f)

After following denoising procedures introduced in (Lee et al. 2020), the three raw biosignals are first segmented by a window with a 13-second length and a 7.5-second shift, so that 13-second-long biosignal segments are fed into the proposed DNN as data points. The feature extractor first splits these segments into five 3-second-long temporal localities by applying a split window with a 3-second length and a 2.5-second shift. These five temporal localities are then fed into the 1-dimensional convolutional blocks. Each convolutional block consists of a 1-dimensional convolution, maxpool, batchnorm, leaky relu activation, and drop out layers in order as shown in Figure 4.2. According to the different sampling rates (PPG: 64 Hz, EDA and ST: 4 Hz), the number of convolutional blocks to apply differs among the three biosignals (PPG: 5, EDA and ST: 3). Once the convolutional blocks extract morphological features from each temporal locality of the three biosignals, the extracted morphological features are gathered and transferred through a flattened layer to an LSTM layer wherein sequential feature information is obtained. The LSTM layer outputs the final feature vector in a many-to-one manner. The stress classifier and the domain classifier are comprised of a fully connected layer, a leaky relu activation, the other fully connected layer, and a softmax layer in order.

4.3.2 Training algorithm

The stress classifier's learning objective can be expressed by Equation 4.1: minimizing stress classification error. The feature extractor has two learning objectives expressed by Equations 4.2

and 4.3 respectively: minimizing stress classification error while maximizing domain classification error.

$$\min_{G_s} \mathcal{L}_s \left(G_s \left(G_f(\mathbb{x}) \right) \right) \tag{4.1}$$

$$\min_{G_f} \mathcal{L}_s \left(G_s \left(G_f(\mathbb{x}) \right) \right) \tag{4.2}$$

$$\max_{G_f} \mathcal{L}_d \left(G_d \left(G_f(\mathbb{x}) \right) \right) \tag{4.3}$$

where $\mathbb{x}$ denotes input raw biosignals, G_f is the feature extractor, G_s is the stress classifier, G_d is the domain classifier, $\mathcal{L}_s$ is the prediction loss of stress classification, and $\mathcal{L}_d$ is the prediction loss of domain classification. Here, inspired by Ganin et al. (Ganin et al. 2016), the authors introduce an additional learning objective for the domain classifier (Equation 4.4: minimizing domain classification error) for adversarial training between the feature extractor and domain classifier.

$$\min_{G_d} \mathcal{L}_d \left(G_d \left(G_f(\mathbb{x}) \right) \right) \tag{4.4}$$

This learning objective aims to effectively train the feature extractor to learn a domain-adapted subject- and context- independent feature vector minimizing the discrepancy in data distribution of target and source domains. If a fixed domain classifier is paired, it is likely that the feature extractor learns to extract a feature vector that fools the fixed domain classifier rather than an optimized domain-adapted feature vector. To prevent the former case, the domain classifier must be continuously updated according to the feature extractor's updates in an adversarial manner.

Table 4.1 describes how the model is trained in each iteration. To achieve the four learning objectives during training, each batch should contain data points from the sources ($\mathbb{x}_i^{s_j}$) and target ($\mathbb{x}_i^t$) so that $\mathcal{L}_s$ and $\mathcal{L}_d$ can be calculated in each iteration. Each batch is populated by sampling the same number of data points from each domain (Lines 2-3 in Table 4.1. Data points from sources have a stress label (the binary level of stress; $y_{s,i}$) and a domain label ($y_{d,i}$) while those from a target have only a domain label.

Per each training iteration, the forward/backward propagations are conducted along with the two paths introduced in Subsection 4.3.1 (i.e., toward the outputs of the stress classifier and domain classifier), outputting two losses ($\mathcal{L}_s$ and $\mathcal{L}_d$) in the process. Specifically, data points

sampled from sources are first used to conduct the forward propagation along with the first path toward the stress classifier's output, thereby calculating $\mathcal{L}_s$ (Line 9). Note that source data points are also later used to calculate $\mathcal{L}_d$.

Although only a single domain classifier is displayed in the proposed model's overview (Figure 4.1), to effectively deal with multiple source domains, the authors actually train as many "binary" domain classifiers as there are source domains (Zhao et al. 2017). In practical applications, the number of source domains could be huge and so it might be difficult to train a single domain classifier to classify data into a huge number of domains. A competitive domain classifier is essential for effective adversarial training with the feature extractor. Therefore, the authors deploy multiple binary domain classifiers to determine whether a data point comes from a target or a source: Data points from the target are first paired with data points from all source domains, thereby making as many "mini-mini batches" (i.e., a mini batch in a mini batch) as the number of source domains Figure 4.3. So, the ratio of target data points to source data points is 1:1 in each mini-mini batch. These mini-mini batches are used to train the feature extractor and the binary domain classifiers in an adversarial manner. Specifically, the feature extractor learns all mini-mini batches while domain classifiers are paired with mini-mini batches one by one so that each domain classifier learns only one mini-mini batch of data points from the source domain that the domain classifier is assigned to (Figure 4.3).

Here, two types of losses are calculated from the domain classifiers' output for adversarial training. First, to train domain classifiers that satisfy Equation 4.4, cross-entorpy loss is calculated in each domain classifier ($\mathcal{L}_{ce,d_j}$) (Lines 12-13). Also, the other loss ($\mathcal{L}_{se,d}$) is calculated to train the feature extractor to satisfy Equation 4.3 (Line 14-15). This loss ($\mathcal{L}_{se,d}$) quantifies how far the softmax output of the domain classifiers is from a dummy domain label (0.5, 0.5) (y_d^d) using square error (SE). This calculation is based on intuiting that a domain classifier softmax output of 0.5/0.5 means that source and target data distributions are so similar on the feature vector that domain classifiers cannot be inclined toward either side in determining whether data points are from a source or target. Adversarial training is realized via these antagonistic two losses.

Also, to prevent abandoning any source domains during feature extractor training, the authors apply the log-sum-exp trick (Zhao et al. 2018). Once all SE losses are calculated from all domain classifiers ($\mathcal{L}_{se,d_j}$), one final loss representing all SE losses ($\mathcal{L}_{se,d}$) is calculated using the

Equation described on Line 19 in Table 4.1. In this way, the larger the error from one domain classifier is, the larger $\mathcal{L}_{se,d}$ is.

Once all losses are calculated, they are normalized by the number of data points and then the backward propagations are conducted to calculate the gradients of each weight (Lines 22-27).

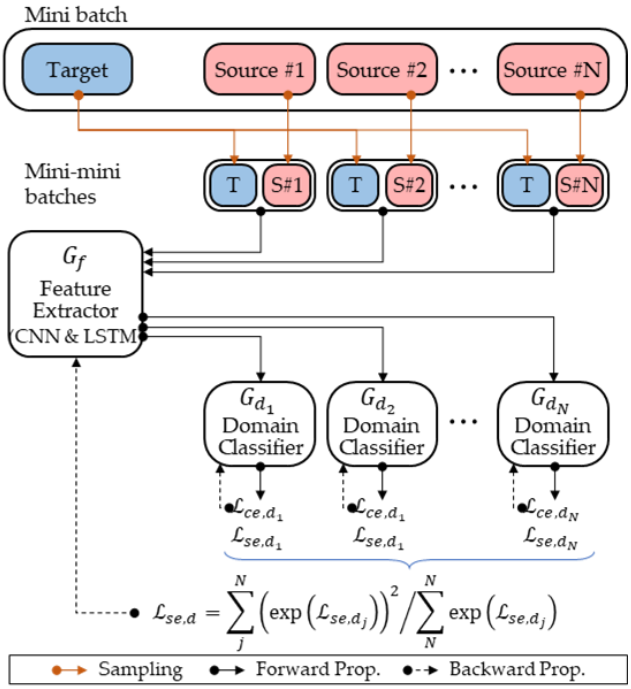


Figure 4.3. Adversarial training between feature extractor and multiple binary domain classifiers

To adaptively balance training pace between the three components (G_f, G_s, G_d), an adaptive control introduced by Ma et al. (Ma et al. 2019) modified for this study's task and applied (Lines 31-46). Essentially, guide learning curves for the losses ($\mathcal{L}_{s\_g}$, $\mathcal{L}_{ce,d\_g}$, $\mathcal{L}_{se\_g}$) are predetermined and the training balance is adaptively controlled by regularly moderating update steps of the three components (k_1, k_2, k_3) during training so that their learning curves can follow the guide learning curves. Specifically, power exponential functions are tuned in the fine-tuning phase as guide learning curves ($\mathcal{L}_{s\_g}$, $\mathcal{L}_{ce,d\_g}$, and $\mathcal{L}_{se,d\_g}$). Then, update steps (k_1, k_2, k_3) of the different components are adaptively updated at every c iterations following the procedures described in Lines 31-46. For example, if $\mathcal{L}_s$ is much higher than the guide curve $\mathcal{L}_{s\_g}$ at the current iteration, K_1 is increased so that the feature extractor learns more weighting gradients from $\mathcal{L}_s$. On the other hand, if $\mathcal{L}_{ce.d}$ is

much lower than the guide curve $\mathcal{L}_{ce.d\_g}$ at the current iteration, K_2 is decreased so domain classifiers are updated with a smaller step.

RMSprop is applied to update weights using the calculated gradients (Line 49-52). Specifically, the weights of G_f ($\mathbb{w}_{G_f}$) are updated twice per iteration using a gradient from $\mathcal{L}_s$ ($\nabla\mathbb{w}_{G_f}$) and another from $\nabla\mathcal{L}_{se,d}$ ($\nabla\mathbb{w}_{G_f}^d$) because G_f is trained for two learning objectives: to minimize stress classification error and to maximize domain classification error.

Table 4.1. Stochastic training algorithm for the proposed DNN

1. ## Input:
2. ## Data points from Target $= (\mathbb{x}_i^t, y_{d,i})_{i=1}^n$
3. ## Data points from Source$_j$ $= \left(\mathbb{x}_i^{s_j}, y_{s,i}, y_{d,i}\right)_{i=1}^n$ (j $= 1,2,\cdots,$ N)
4.
5. ## Loss calculation
6. For j from 1 to N do:
7. For i from 1 to n do:
8. ## Forward propagation toward the output of stress classifier
9. $\mathcal{L}_s \mathrel{+}= \ell_{ce}\left(G_s\left(G_f\left(\mathbb{x}_i^{s_j}\right)\right), y_{s,i}\right)$
10.
11. ## Forward propatation toward the output of domain classifiers
12. $\mathcal{L}_{ce,d_j} \mathrel{+}= \ell_{ce}\left(G_{d_j}\left(G_f(\mathbb{x}_i^t)\right), y_{d,i}\right)$
13. $\mathcal{L}_{ce,d_j} \mathrel{+}= \ell_{ce}\left(G_{d_j}\left(G_f\left(\mathbb{x}_i^{s_j}\right)\right), y_{d,i}\right)$
14. $\mathcal{L}_{se,d_j} \mathrel{+}= \ell_{se}\left(G_{d_j}\left(G_f(\mathbb{x}_i^t)\right), y_d^d\right)$
15. $\mathcal{L}_{se,d_j} \mathrel{+}= \ell_{se}\left(G_{d_j}\left(G_f\left(\mathbb{x}_i^{s_j}\right)\right), y_d^d\right)$
16. For end
17. For end
18.
19. $\mathcal{L}_{se,d} = \sum_j^N \left(\exp\left(\mathcal{L}_{se,d_j}\right)\right)^2 \Big/ \sum_N^N \exp\left(\mathcal{L}_{se,d_j}\right)$ ## log-sum-exp trick
20.
21. ## Backward propagation to calculate gradients
22. $\nabla\mathbb{w}_{G_s} \leftarrow \nabla\mathcal{L}_s/\left(n*(N-1)\right)$
23. $\nabla\mathbb{w}_{G_f} \leftarrow \nabla\mathcal{L}_s/\left(n*(N-1)\right)$
24. For j from 1 to N do:
25. $\nabla\mathbb{w}_{G_{d\_j}} \leftarrow \nabla\mathcal{L}_{ce,d_j}/n$
26. For end
27. $\nabla\mathbb{w}_{G_f}^d \leftarrow \nabla\mathcal{L}_{se,d}/\left(n*(N-1)\right)$
28.
29. ## Adaptive control of learning steps of components
30. If mod(iteration #, C) $== 0$
31. For $\mathcal{L}$, $\mathcal{L}_g$, k, α in [$\mathcal{L}_s$, $\mathcal{L}_{ce,d}$, $\mathcal{L}_{se\_d}$], [$\mathcal{L}_{s\_g}$, $\mathcal{L}_{ce,d\_g}$, $\mathcal{L}_{se,d\_g}$],
32. [k_1, k_2, k_3], [α_1, α_2, α_3]
33. If k ≥ 1
34. if $(\mathcal{L}-\mathcal{L}_g)/\mathcal{L}_g > \alpha$
35. k += 1
36. else if $(\mathcal{L}_g-\mathcal{L})/\mathcal{L}_g > \alpha$
37. k = 0.5

If end
else
if $(\mathcal{L} - \mathcal{L}_g)/\mathcal{L}_g > \alpha$
$\mathrm{k} = 2$
else if $(\mathcal{L}_g - \mathcal{L})/\mathcal{L}_g > \alpha$
$\mathrm{k} = 1/(\mathrm{k}^{-1} + 1)$
If end
For end
If end

Weights updates
$\mathbb{w}_{G_f} = RMSprop(\nabla \mathbb{w}_{G_f}, \mu \times \mathrm{k}_1)$
$\mathbb{w}_{G_s} = RMSprop(\nabla \mathbb{w}_{G_s}, \mu \times \mathrm{k}_1)$
$\mathbb{w}_{G_d} = RMSprop(\nabla \mathbb{w}_{G_d}, \mu \times \mathrm{k}_2)$
$\mathbb{w}_{G_f} = RMSprop(\nabla \mathbb{w}^{d}_{G_f}, \mu \times \mathrm{k}_3)$

n: number of data points sampling from each domain; N: number of source domains; $\mathbb{x}_i^t, \mathbb{x}_i^{s_j}$*: raw biosignals of a data point from the target and jth source respectively;* $y_{s,i}, y_{d,i}$*: stress and domain labels of a data point;* y_d^q*: dummy domain label;* ℓ_{ce}*: cross-entropy loss;* ℓ_{se}*: square error loss;* $\mathbb{w}$*: weights;* k_1, k_2, k_3*: learning steps for the weight updates from* $\mathcal{L}_s$*,* $\mathcal{L}_{ce,d}$*,* $\mathcal{L}_{se,d}$ *respectively;* $\alpha_1, \alpha_2, \alpha_3$*: learning step control thresholds;* C*: learning step update cycle;* $RMSprop()$*: RMSProp weight updater;* μ*: learning rate.*

4.4 Performance test of the proposed stress detection

The proposed technique's generalization performance across different people and contexts was compared with benchmarks by using two datasets collected from in-lab and field settings respectively (Figure 4.4). In the in-lab data collection, the stress data was collected from 13 people and 3 different contexts to see the performance of the proposed technique in an unseen dataset collected from an unseen person and an unseen context. Field data was collected from workers at a construction site during their ongoing work since construction workers experience stress in a wide range of contexts (e.g., different stressors and varying ambient conditions)—a good setup for testing context-independency together with subject-independency (Jebelli et al. 2018) and an opportunity to examine the proposed technique's performance outside of a lab. Data collection was conducted between Aug. and Dec. 2019. The data collection protocol was approved by the University of Michigan Institutional Review Board (IRB00000245).

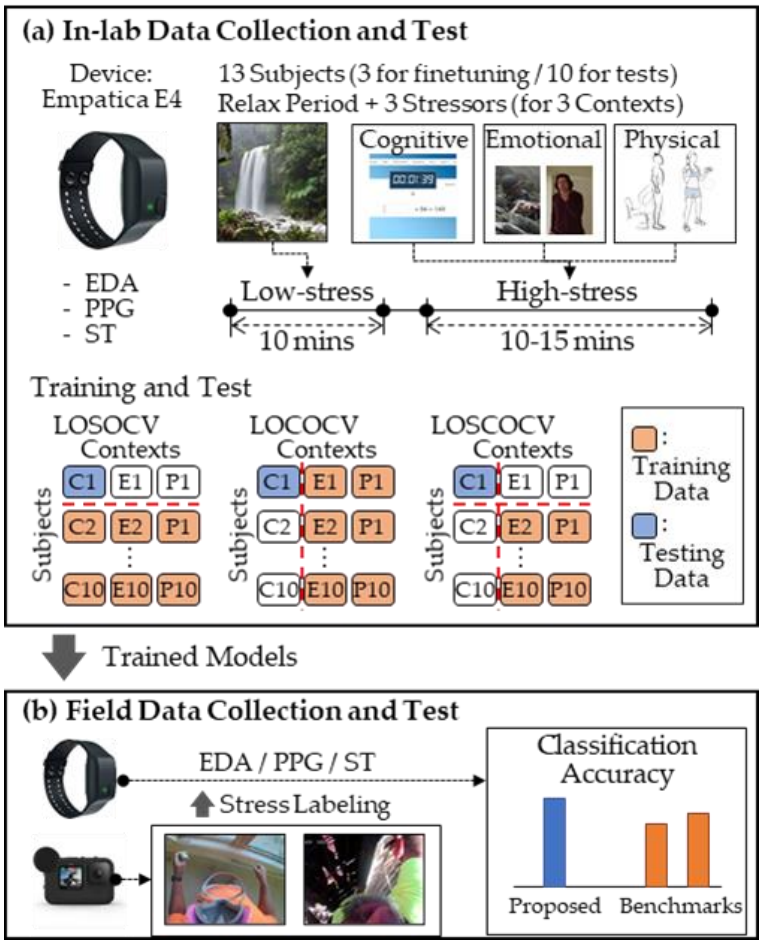


Figure 4.4. Overview of the performance test for the proposed stress monitoring

4.4.1 In-Lab data collection

13 University of Michigan graduate students were recruited as subjects for in-lab data collection (Table 4.2). Subjects were asked to report any physical and/or mental health issues that could affect their physiological reactivity to stress, but none reported such issues. Subjects individually participated in three sessions on three different days each with a different type of stressor. Different stressors (i.e., cognitive load-related, unpleasant emotion-related, and physical activity-related stressors) were used to establish three different stress contexts in the collected data since stressor type is a major contextual factor shaping biosignals' patterns in response to stress i (Schlotz 2013). An off-the-shelf wristband-type biosensor (i.e., E4 from Empatica Inc.) was used to collect subjects' EDA, PPG, and ST signals with sampling rates of 4 Hz, 64 Hz, and 4 Hz respectively.

Table 4.2. Demographic information of 13 subjects in the in-lab data collection

Statistics	Age (years)	Height (cm)	Weight (kg)	Gender
Mean	31.2	177.9	76.6	Female 3
(SD)	(2.1)	(4.9)	(6.0)	Male 10

Before each stress session, subjects had a 10-minute relax period to minimize the impact of prior, external factors on their stress levels and to collect biosignals under low-stress conditions. During the relax period, a nature video with sound was provided to subjects via a computer screen. Biosignals were collected from the last 5 minutes of this period and labeled "low-stress."

As a cognitive load-related stressor, subjects conducted Engle-Friedman's Math Effort Task (Engle-Friedman et al. 2003). This task presents addition problems via a computer screen. For each problem, subjects were asked to add four numbers between 7 and 25 without the use of pencil or paper within 15 seconds. A total of five problem sets, each of which lasted 2 minutes, were presented. Immediately after each subsession, subjects rated their arousal using the Self-Assessment Manikin scale (Bradley and Lang 1994). Subjects' biosignals collected during subsessions rated as high arousal (6 or more on the 9-point arousal scale) were labeled as high-stress given the significant correlation between levels of stress and arousal (McCorry 2007). To cognitively challenge subjects throughout the task, they were asked to try to do their best and informed that they could privately receive a performance score.

In the unpleasant emotion-related stressor session, subjects were asked to concentrate on three unpleasant video clips. The video clips were selected to elicit negative-valence and high-arousal emotions such as anxiety, fear and anger, which are expected to function as stressors according to the emotional stress model (Bong et al. 2013). Specifically, bloody, conflicting, and horror scenes, which are copyright-free and two to three minutes long, were used. Subjects were asked to rate their emotion after watching each clip using the Self-assessment manikin scale (Bradley and Lang 1994). Biosignals were collected while subjects were watching videoclips that they rated as low valence (4 or less on the 9-point valence scale) and high arousal (6 or more on the 9-point arousal scale), and labeled as high-stress referencing the emotional stress model (Bong et al. 2013).

As a physical activity-related stressor, subjects were asked to repeat two physical exercises in order: arm curls and sit-to-stands, which have been widely used as a stress-elicting physical exercise (Jeoung and Lee 2015; Naliboff et al. 1988). After every five repetitions, subjects were asked to self-

report their perceived exertion using Borg's Rating of Perceived Exertion scale (RPE) (Borg 1982) given that the RPE scale has been useful in assessing subjects' stress level during exercise (Borg 1970; Russell 1997). Then, from times when subjects rated their perceived exertion equal to or more than 15 (hard), biosignals were collected and labeled as high-stress. Once subjects reported the highest physical exertion level (i.e., 20: maximal exertion), the exercises were terminated.

4.4.2 Field data collection

Field data collection was conducted at a building construction site in Ann Arbor, Michigan to test the proposed subject- and context-independent technique's performance in a field application beyond the lab. 15 construction workers were recruited as subjects while incorporating individual variabilities (e.g., age, height, and weight) as shown in Table 4.3. For two days, the subjects' three biosignals (i.e., EDA, PPG, and ST) were collected using E4 wristbands during their daily work. Also, all of their work and body movements were video-captured using an action camera attached to the front of their hardhats. The collected biosignals were labeled as low-stress and high-stress activities based on the recorded videos. Since it is difficult to ascertain a ground truth for workers' stress without interfering with their tasks, the authors used only collected biosignals that clearly corresponded to obvious low or high stress activities. To identify these clear moments of low or high stress, two research team members with construction field experience separately watched the videos and selected subjects' activities conducted under obviously high or low stress. Then, only data corresponding to activities consistently selected as low or high stress by both team members were labeled and used. Walking with or without light materials, talking with coworkers, performing plain jobs on ground level, and resting were labeled as low-stress. Working at height with a fall risk and with postural instability were the most common activities labeled as high-stress.

Table 4.3. Demographic information of 15 subjects in the field data collection

Statistics	Age (years)	Height (cm)	Weight (kg)
Mean (SD)	36.7 (10.8)	179.6 (9.8)	95.7 (12.2)

4.4.3 Test of the subject- and context-independency

To test and compare the proposed technique's subject- and context-independent performance with benchmarks, three validation methods were applied while training and testing models using the in-lab stress data. First, leave one subject out (LOSOCV) and leave one context out cross validation (LOCOCV) methods were used to compare generalization across different people and contexts separately. Then, the authors applied a new validation method, leave one subject and context out cross validation (LOSCOCV), introduced in Chapter 3. Finally, how better the proposed method would actually perform in a real field application than benchmarks was examined using the field dataset collected from a construction site.

Two benchmarks were compared with the proposed technique. The first benchmark, the Gaussian support vector machine (SVM), widely acknowledged as the best traditional machine learning algorithm to detect human stress from biosignals (Jebelli et al. 2019), was applied to train a machine learning model on a set of features previously hand-crafted based on related prior knowledge (Lee et al. 2021). The second benchmark, a convolutional and LSTM layers-based DNN model modified from DeepER Net (Seo et al. 2019), has proven useful in detecting different levels of stress from biosignals. This model has nearly the same architecture as the DNN model proposed in this study, except it lacks a domain adaptation structure. Before testing, all models' key hyperparameters (i.e., the proposed model, SVM, and DNN) were fine-tuned using three in-lab subjects' data. Table 4.4 shows the list of hyperparameters fine-tuned or empirically pre-determined.

Table 4.4. Hyperparameter setup for the tested models

- Proposed DNN **Kernel lengths for PPG (in order): 50-20-10-5-3, **Kernel lengths for EDA and ST (in order): 10-7-3, **Channel number: 30, **Maxpool length and stride: 2, **Drop out rate: 0.5, **Leaky Relu Slope: 0.01 (the above convolutional block parameters are shared with the benchmark DNN), *LSTM hidden unit number: 40, Hidden node number in the classifier part: 6, *Learning rate: e-06, *Number of Epoch: 70, **L2 regularization: e-02, *number of data points from each domain in a batch (n): 8, *C: 3, *α_1: 0.1, *α_2: 0.01, *α_3: 0.1
- Benchmark DNN (without domain daptation) *LSTM hidden unit number: 32, Hidden node number in the classifier part: 4, *Learning rate: e-04, *Number of Epoch: 80, *L2 regularization: e-03, *number of data points in a batch: 32
- Benchmark Gaussian SVM *Kernel width: 8.0

**: fine-tuned; **: empirically pre-determined*

4.5 Results

The three biosignals (i.e., EDA, PPG, and ST) collected in-lab were segmented into data points with a 15-second-long length and a 7.5-second-long shift and labeled according to the data collection procedures descrived above. As a result, the in-lab stress dataset was comprised of 3,892 and 4,927 data points labeled as "low-stress" and "high-stress" respectively. This dataset was used to train and test machine learning models via the proposed DNN model and two benchmarks (i.e., SVM and DNN without domain adaptation).

To compare different levels of generalizability between the proposed technique and benchmarks, LOSOCV, LOCOCV, and LOSCOCV were each applied using the in-lab dataset. Table 4.5 compares test performance between the proposed technique and the two benchmarks with all three validation methods. Overall, across all the three validations, the proposed technique outperformed the two benchmarks. The gap in performance between the proposed technique and the benchmarks was bigger when testing generalization across both subjects and contexts (LOSCOCV) than when testing generalization across only subjects or contexts (LOSOCV or LOCOCV). Between the two benchmarks the DNN model showed better performance than the SVM model except in LOSOCV where the SVM model showed slightly better performance.

Table 4.5. Performance comparison between the proposed technique and benchmarks in three validations

Validation Method	Metric	Applied Techniques		
		Proposed	SVM	DNN
LOSOCV (In-lab)	Overall Accuracy	0.809	0.772	0.766
	High-stress Precision	0.758	0.790	0.720
	High-stress recall	0.935	0.802	0.922
LOCOCV (In-lab)	Overall Accuracy	0.799	0.723	0.791
	High-stress Precision	0.756	0.686	0.743
	High-stress recall	0.915	0.794	0.927
LOSCOCV (In-lab)	Overall Accuracy	0.803	0.702	0.724
	High-stress Precision	0.775	0.650	0.691
	High-stress recall	0.911	0.784	0.901

The authors also compared the models' performance using the field stress data. In this field test, models were re-trained using all in-lab data. For the proposed technique, the in-lab stress dataset was used for source domains and each subject's data in the field dataset was used as a target domain in re-training. The re-trained models were then applied to classify the field dataset's binary levels of stress. By doing so, the authors conducted a hold-out test in which the in-lab and field datasets were used as training and testing datasets respectively. The proposed technique showed significantly higher performance in classifying the level of stress than the two benchmarks (see Table 4.6).

Table 4.6. Performance comparison between the proposed technique and benchmarks in the field test

Metric	Applied Techniques		
	Proposed	SVM	DNN
Overall Accuracy	0.785	0.679	0.638
High-stress Precision	0.713	0.599	0.577
High-stress recall	0.911	0.920	0.905

4.6 Discussion

To advance generalization across different subjects and contexts in detecting stress from biosignals, the authors propose a new technique that applies a DNN model to learn subject- and context-independent feature vector and a classifier simultaneously in an end-to-end manner. The subject- and context-independency of the proposed technique was compared with two benchmarks (i.e., SVM and DNN without domain adaptation) by applying LOSOCV, LOCOCV, and LOSCOCV, and the results indicate that the proposed technique is more effective at buffering individual and contextual differences than the benchmarks, and thereby shows better generalization across different subjects and contexts.

The benchmark techniques showed significantly lower performance when validated by LOSCOCV than by LOSOCV or LOCOCV (SVM accuracy: 0.772 by LOSOCV, 0.723 by LOCOCV → 0.702 in LOSCOCV; DNN accuracy: 0.766, 0.791 → 0.724). This result was expected because ensuring generalization across both different subjects and contexts is more challenging than just across one out of the two. On the other hand, the proposed technique showed stable performance regardless of validation method (accuracy: 0.809, 0.799, 0.803). This can be interpreted that since the proposed

technique actively recognizes and buffers domain differences, the technique's performance is less subject to differences between domains.

The field test results coincide with the LOSCOCV test. Due to having more variability in subject demographic information (e.g., age, weight, and height) than the in-lab dataset (Tables 4.2, 4.3), the field stress dataset likely incorporates more individual differences. Also, most field data was collected while subjects were working outdoors, and thus varying ambient factors such as temperature, humidity, and sunlight were added as different dimensions of the contextual difference, which was not directly incorporated in the in-lab data. Because of these differences between datasets, the domain difference between sources (in-lab data) and targets (field data) might be larger than between sources in the field test, unlike in LOSCOCV where only the in-lab dataset was used. This might explain why field test performance was lower than the in-lab LOSCOCV. However, compared to the benchmarks that each show significant performance decreases (SVM accuracy: 0.702 in LOSCO → 0.679 in the field test; DNN accuracy: 0.724 → 0.638), the proposed technique's performance gap between the LOSCO test and field test was relatively moderate (accuracy: 0.803 → 0.785). This result can indicate that even when the difference between target and source domains is radical and larger than between source domains, the proposed DNN-based domain adaptation technique is effective in buffering domain difference between the target and sources. It may not be rare that field application domains are radically different from training dataset domains. Thus, this finding re-emphasizes that the proposed technique is essential to ensure reasonable stress detection performance in field applications.

To understand how the proposed DNN model advances subject- and context-independency in stress detection, the authors examined how extracted feature vector changes over the course of training. Specifically, the authors compared features extracted from Epochs 5 (beginning of training) and Epoch 70 (after training) using data from a subject's cognitive load-related context. The backward-elimination wrapper method (Kohavi and John 1997) was applied to select the ten most meaningful features from the two feature sets extracted at Epochs 5 and 70 respectively. Then, the three most meaningful features from both sets were identified. The target and source domains were then visualized in three-dimensional space with the three common features serving as axes (Figure 4.5) to intuitively display the feature vector's general changes during training. As shown in Figure 4.5, data distributions of source and target domains are more similar to each other (less domain discrepancy) and the decision boundary between levels of stress is clearer (higher classification performance) at Epoch 70 than at Epoch 5. Therefore, it is more likely for the target and sources to share a stress level decision boundary

at Epoch 70 (after training) then at Epoch 5 (beginning of training). This observation indicates that the proposed DNN model successfully learns satisfying the two learning objectives of the domain adaptation (i.e., maximizing classification performance in source domains and minimizing data discrepancy between different domains) during training, which advances domain-independency in stress level classification.

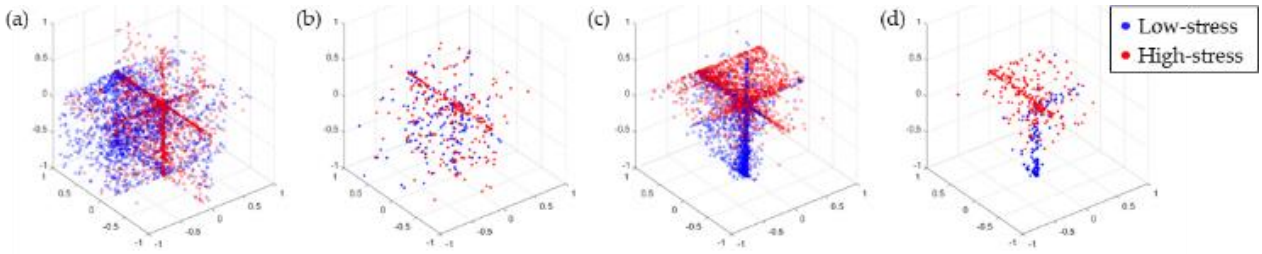


Figure 4.5. Changes in data distribution over training; (a) data from sources at Epoch 5; (b) data from target at Epoch 5; (a) data from sources at Epoch 70; (a) data from target at Epoch 70

It is noteworthy that, regardless of validation method, all tested models showed higher recall values than precision for the high-stress class. These results can be explained by the distributional characteristic of the analyzed stress datasets where low-stress data points have smaller variability than high-stress ones. Low-stress data points were collected under a few limited conditions, such as when subjects relaxed or lightly walked while high-stress data points were collected from more varied conditions. For example, it might be likely that subjects' cognitive load and displeasure levels during the math task and emotional video clips vary over time. Also, high-stress data points collected during physical exercise were from varying levels of physical exertion (i.e., 15 to 20 by Borg's RPE). Consequently, low-stress data points have a relatively tight distribution while high-stress data points have a wide distribution, and the trained models accordingly become relatively specific to low-stress and sensitive to high-stress. When these trained models are applied across domains, they might display a higher tendency to misclassify low-stress data points as high-stress ones than other misclassifications. In some field applications where costly or invasive interventions follow detected high-stress and thus the number of incorrectly detected high-stress cases must be minimized, high-stress precision might be more important than recall. A future study, perhaps one applying class imbalance loss functions and collecting low-stress data points from more varied conditions, might therefore be needed to regulate trained models' inclination toward high-stress recall.

This study demonstrates that the proposed technique advances both subject- and context-independency in detecting human stress from biosignals. Due to individual and contextual variabilities in the biosingals' reactivity patterns to stress, previous techniques have required collecting additional labeled stress data in order to reliably detect stress from new people in new contexts. This study's findings enable us to eliminate the need for the tedious labeled data collection for new domains and thus significantly contribute to advancing the field applicability of the wearable biosensors- and machine learning-based stress detection. Additionally, the finding of this study can be extended to other human response detection tasks that use different biosensors (e.g., mobile-type electroencephalogram (EEG) and sweat patches) because individual and contextual variability is a common issue in biosignal-based human response detection (Picard et al. 2001).

Despite such significant contributions, this study has limitations that should be addressed by future studies. First, although the proposed technique's reported performance is promising, still many hyperparameters (e.g., kernel lengths, channel number, and drop rate) remain untuned. Therefore, a future study should focus on a wider range of hyperparameter tuning. Second, an important strength of the proposed technique is its capability to incorporate individual and contextual variability in multiple different source domains. However, it was not examined how its performance changes according to the extent of individual and contextual differences incorporated in multiple source domains. A follow-up study, therefore, should examine corresponding performance changes while different variability levels are introduced in source domains.

4.7 Conclusion

This study proposes a subject- and context-independent stress detection technique that learns a domain-independent feature vector and classifier simultaneously in an end-to-end manner. The proposed technique was compared with two benchmarks that do not incorporate domain adaptation. In both in-lab and field tests, the proposed technique showed higher accuracy than the benchmarks in classifying stress levels on a testing dataset collected from people and/or contexts not considered during training. These results indicate the proposed technique can significantly advance generalizability in detecting stress from biosinals across subjects and contexts. This study contributes to existing knowledge by providing new means of advancing subject- and context-independency for detecting human responses from biosignals. Also, the proposed stress detection technique can significantly lighten the burden of

labeled data collection, thereby contributing to scalable field application of wearable biosensors- and machine learning-based stress detection in CBEs.

Chapter 5 Geographic Information System (GIS)-Based Stress Hotspot Detection

5.1 Introduction

Previous chapters addressed the first two agendas of this research: to apply wearable biosensors to detect human stress in an artifact-robust and scalable manner. These efforts significantly contribute to the field applicability of wearable-based stress detection. However, just detecting stress may not be enough for understanding how to improve the quality of interactions between humans and CBEs. Designing effective interventions entails reflecting stress-related circumstantial information, such as where an individual is stressed and how the individual is affected in the stressful interaction. However, little is currently known about how to continuously and non-invasively acquire such stress-related circumstantial information in people's daily work and lives, where a wide range of stressors are mixed up and the impact of an identical stressor can vary by different individuals and their time-to-time status. As the first effort to fill this gap, this chapter presents a geographic information system (GIS)-based technique that accumulates people's stress data on a map and identifies locations of stressors from the accumulated stress data. Such location information might be helpful to identify stressors in CBEs and design stressor-specific effective interventions to improve the quality of interactions between humans and CBEs.

The GIS-based hotspot analysis has shown potential to distinguish stress related to environmental stressors and locate the spots of the environmental stressors. The hotspot analysis is to identify spatial "hotspots," defined as locational spots where high values of variables of interest are concentrated, by spatially integrating data using GIS (Anderson 2009). While a single stress occurrence could be caused by diverse stimuli, a stress hotspot, which is a spot where multiple stresses are concentrated with abnormally high density, can be linked to the stressful conditions on that spot (Yang et al. 2017). As such, the stress hotspot can represent the spot where the people's stressful interaction with their surrounding CBEs takes place. By applying wearable biosensors and the hotspot analysis together, it is possible to spatiotemporally detect stressful interactions between humans and their surroundings.

Despite such synergic potential, the full potential of applying both wearable biosensing and hotspot analysis to detect and locate human stress in their interaction with surroundings has not yet been enough pursued. Several studies were conducted to understand the relationship between people's physiological responses and spatial configurations of the built environment, such as sidewalks and touristic places (Hijazi et al. 2016; Kim and Fesenmaier 2015; Shoval et al. 2018). However, their use of GIS has been mainly limited in geocoding and geographically visualizing individuals' or collective physiological features, which required analysts' subjectivity to detect stress hotspots (Hijazi et al. 2016; Kim and Fesenmaier 2015; Shoval et al. 2018). Although a couple of recent studies applied statistical clustering methods to detect stress hotspots (Chrisinger and King 2018; Hijazi et al. 2016), the statistical clustering methods might not reflect uncertainty about the exact position of stress occurrence, which caused by error of GPS and variability in people's physiological latency to stress (Anderson 2009). Also, these previous studies used only one physiological feature (e.g., mean of EDA and heart rate) as an indicator of people's level of stress in their daily trips. Although the features used have a correlation with the level of stress, depending on only one physiological feature might not accurately understand people's stress because the feature could be easily distorted by noises recorded in biosignals, which could not be fully eliminated by current signal processing techniques. Given that different biosignals have different source of noises according to their recording techniques and related organs, leveraging different biosignals can complement each signal's noisy time frames for more accurate stress detection. Therefore, there is a need for combining the GIS-based hotspot analysis with more sophisticated and deliberately designed computational models considering multiple biosignals. To overcome such limitations, a new stress hotspot detection technique, which is objective as well as robust to uncertainty about the exact position of stress, is required.

5.2 Proposed hotspot analysis-based stress hotspot detection

Therefore, the objective of this study is to develop and test a wearable biosensing and hotspot analysis-based framework that detects locations of human stressful interactions with their surroundings. In particular, this study aims to develop a hotspot analysis to objectively detect stress hotspots where people experience stress due to environmental features in CBEs under uncertainty about the exact position of stress occurrence, which is common in the use of wearable-type location sensors and biosensors. Then, the developed objective hotspot analysis is combined with the advanced individual stress detection technique introduced in the previous chapters of this dissertation.

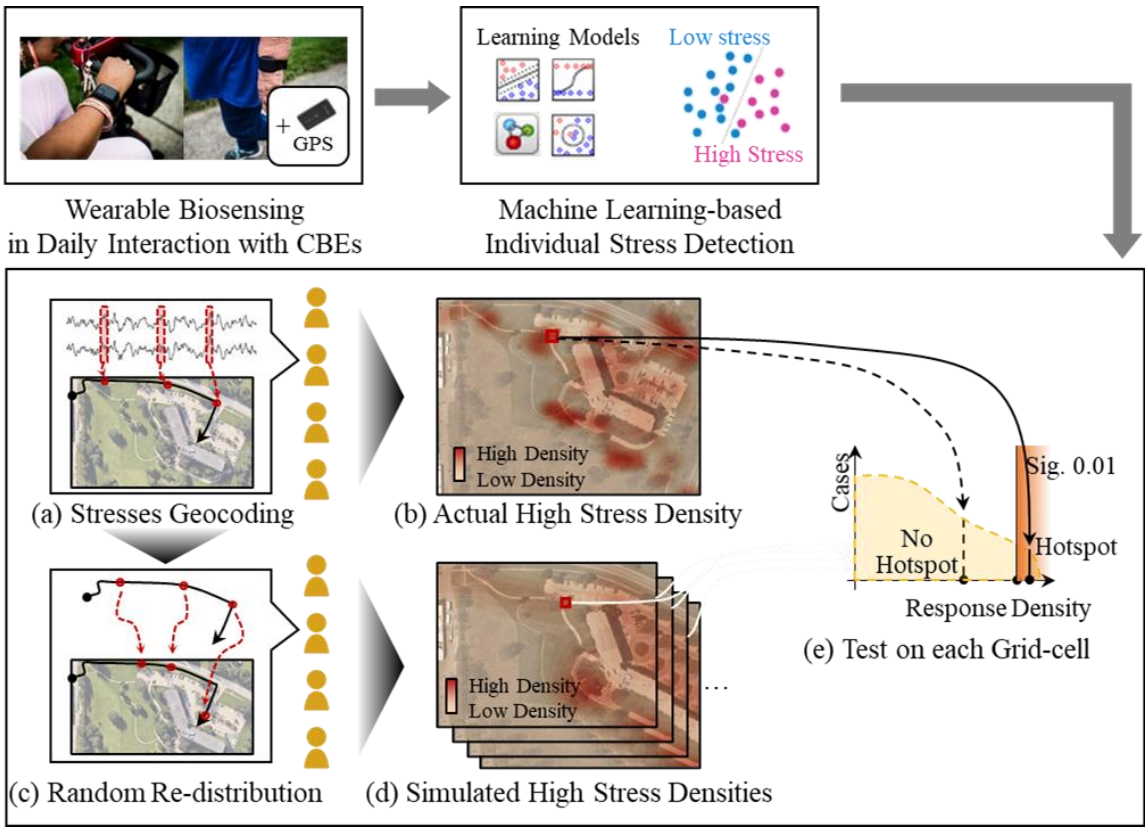


Figure 5.1. Overview of proposed hotspot analysis-based stress hotspot detection

Figure 5.1 shows an overview of the proposed statistical hotspot analysis. First, during people daily interactions with CBEs, their stress-responsive biosignals (e.g., EDA and PPG) are collected together with location data such as global positioning system (GPS) by wearable biosensors. Then, the continuous biosignals are segmented by a window with ten-seconds length and one-second shift. A machine learning-based technique classifies individuals' stress levels into low or high from the ten-second-long segments of biosignals.

The classified stress samples are fed into the proposed statistical hotspot analysis. The individuals' stress samples are geocoded based on GPS data simultaneously collected with biosignals (a in Figure 5.1). In particular, the GPS data are recorded based on the world geodetic system 1984. To conduct the hotspot analysis, the GPS data should be projected (i.e., a transformation of the latitudes and longitudes of locations from the surface of a non-plane shape such as sphere or ellipsoid into locations on a plane (Snyder 1989)). This study uses NAD 1983 State Plan Michigan South as a projected coordinate system because this system induces the least distance distortion by projection for the south area in Michigan state where most of the data were

collected. Each stress sample is allocated on the location recorded by GPS at the start moment of the sample. With such allocation, each stress sample can be considered as the response to its allocated location. The authors apply a grid system whose grid-cell (i.e., unit of location) is set to 1m by 1m. The use of small grid-cells can facilitate in linking detected to people's stressful interactions on the hotspot.

After stress mapping, the authors draw a density map of high-stress (b in Figure 5.1) by applying kernel density estimation (KDE), which is a useful GIS smoothing technique for buffering the uncertainty about the exact position of an event of interest (Anderson 2009; Bíl et al. 2013). KDE bears such uncertainty by providing each grid-cell with a density of the event of interest that considers not only the area of the grid-cell, but also its surrounding areas within a predetermined radius (Seaman and Powell 1996). By applying the KDE with 3-meter radius, each grid-cell had a value of the density of stress samples on the area within 3 meters from the center of that. Previously, analysts detected hotspots by manually observing the result of KDE. Therefore, the hotspot detection results were dependent of the analysts' subjectivity (Anderson 2009). To overcome this limitation, Xie and Yan (2013) and Bíl et al. (2013) proposed to test the significance of each grid-cell using Monte-Carlo method, which uses repeated random simulations to obtain the desired numerical results (O'sullivan and Unwin 2014). The authors applied the Monte-Carlo-based hotspot detection technique as follows.

Based on the definition of the stress hotspot (i.e., a spot on which people's stress are abnormally concentrated), this study tested the null hypothesis (H_0) on every grid-cell: "People's stress is randomly distributed on their trajectories regardless of any locational factors." To test the null hypothesis, the Monte-Carlo method generated cases in these two steps. First, each individual's stress samples were randomly redistributed on his/her trajectory (c in Figure 5.1). The previous studies using Monte-Carlo method to statistically detect car accident hotspots (Bíl et al. 2013; Xie and Yan 2013) just randomly redistributed geocoded accidents in this step because their data did not contain any individuals' trajectory or accident rate. On the other hand, since this study uses data collected from individuals' wearable biosensors, their trajectories and number of stresses can be reflected to generate more sophisticated random distribution. Specifically, stress samples collected from an outing of an individual were counted and then the samples were randomly allocated on points of the outing's trajectory. This redistribution was conducted for all the outings of all the individuals. Second, from the re-distributed stress samples, each grid-cell's density of

stress sample was calculated by KDE, the same as when calculating the actual density of stress samples (d in Figure 5.1). The authors executed these steps for 100 times to gather 100 cases because 100 times of simulation can be enough to assess significance of each grid-cell (Xie and Yan 2013). Consequently, each grid-cell had 100 values of simulated density of stress samples. By counting cases with the density of stress samples over the actual density of stress samples on each grid-cell, it is possible to assess each grid-cell's significance. Grid-cells with significance level higher than predetermined threshold (0.01 in this study, which has been widely used to statistically detect hotspots in previous studies (Bhunia et al. 2013; Pandey et al. 2014)) were detected as a stress hotspot (e in Figure 5.1).

5.3 Pilot study – Seniors' stress in daily trips in Ypsilanti Township, Michigan

The authors conducted a pilot study in which the proposed hotspot analysis was conducted to locate environmental stressors that seniors suffer in their daily outdoor trips. As senior population continues to grow rapidly, the outdoor mobility of senior individuals has become increasingly important to social and economic prosperity. However, the seniors' mobility is limited due to various types of environmental stressors in the current built environment. Therefore, there is an urgent need of a means to effectively identify the senior-specific environmental stressors. This pilot study was conducted in cooperation with the Clark East Tower senior apartment located in Ypsilanti Township, Michigan. Since senior residents in this apartment share many common paths (e.g., sidewalk between the apartment and the closest bus stop) and destinations (e.g., the closest grocery market), the proposed framework can be tested based on a large number of grid-cells on the shared paths. 30 residents living in the senior apartment participated in this pilot study. The informed consent forms were distributed to all the subjects to make them informed about the anonymity of data collection and rights of subjects. The subjects were then asked to report their information such as age, gender, height, weight, and their use of assistive devices (e.g., walker, wheel-chair, and electric scooter) (Table 5.1). The subjects were grouped into 5 groups, and each group separately participated in the data collection for two weeks (Group 1: July 17, 2018 - July 30, Group 2: August 7, 2018 to August 20, Group 3: August 21, 2018 to September 3, Group 4: September 4, 2018 to September 17, and Group 5: September 18, 2018 to October 1). The subjects participated in both types of data collection (i.e., controlled route and daily trip data collection) during two weeks. After identifying stress hotspots based on the collected data, a post hotspot

inspection was conducted for two weeks (December 14, 2018 to December 27) to examine how well the hotspots represent seniors' stressful interactions with built environment. All the data collection protocol in this pilot study was approved by the University of Michigan Institutional Review Board.

Table 5.1. Subjects' demographic information

Subject #	Age (years)	Height (cm)	Weight (kg)	Gender	Assistive Device
1	82	170.2	81.6	Male	No use
2	66	165.1	72.6	Female	No use
3	65	170.2	90.7	Female	No use
4	66	160.0	99.8	Female	No use
5	66	165.1	70.3	Female	Walker
6	65	167.6	69.9	Female	No use
7	67	177.8	93.0	Female	Walker
8	66	158.8	77.6	Female	Walker
9	69	160.0	73.5	Female	No use
10	65	160.0	77.1	Female	Walker
11	77	170.2	63.5	Male	Electric Scooter
12	68	170.2	76.2	Female	Walker
13	85	170.2	81.6	Female	No use
14	69	165.1	100.7	Female	No use
15	65	172.7	64.0	Female	No use
16	65	188.0	77.1	Male	No use
17	69	149.9	87.1	Female	No use
18	65	162.6	98.0	Female	No use
19	66	180.3	99.8	Female	Electric Scooter
20	66	154.9	104.3	Female	Electric Scooter
21	69	165.1	68.0	Female	Walker
22	67	157.5	70.3	Female	Electric Scooter
23	65	167.6	93.4	Female	No use
24	69	157.5	120.2	Female	Walker
25	65	177.8	95.3	Male	No use
26	68	152.5	72.6	Female	Walker
27	79	149.9	45.4	Female	No use
28	66	195.6	145.1	Male	No use
29	66	162.6	92.1	Female	No use
30	71	160.0	59.0	Female	No use
Mean (SD)	68.6 (5.2)	161.6 (28.9)	84.0 (19.3)		

5.3.1 Controlled route data collection

In this pilot study, a controlled route data collection was conducted along with subjects' daily trip data collection. The purpose of the controlled route data collection was to develop and validate a subject- and context-specific stress detection model that can reliably detect stress for the 30 senior subjects in this study. Although I developed a subject- and context-independent stress detection technique and found that it showed much higher stress detection accuracy than previous techniques in the subject- and context-independent test, the final field test showed that the accuracy level is around 80%. Therefore, this pilot study first developed and validated a subject- and context-specific stress detection model whose accuracy can be substantially higher than 80%, thereby focusing more on testing the performance of the proposed GIS-based stress hotspot analysis.

The controlled route data collection was conducted to collect biosignals to be labeled into "high-stress" and "low-stress" for machine learning based classifiers. All 30 senior subjects participated in the controlled route data collection once a week (i.e., total two trials for each subject) to accommodate different response among different days. As a result, the data was collected from different weather conditions (i.e., temperature: 55 to 91°F, humidity: 37 to 100%, sunlight: rainy to sunny). During the data collection, the subjects were asked to move along a predefined route on which there was a series of potential stressful environmental stressors such as "moving along a side-sloped sidewalk" or "climbing stairs" (Figure 5.2). The potential stressful interactions were determined after carefully studying previous research efforts that identified different types of stressful interactions with the built environment (Lockett et al. 2005; Rosenberg et al. 2012). Then, the route was designed to have as many types of environmental stressors as possible, keeping the duration up to 10 minutes considering senior subjects' physical capability and memory to ensure that the subjects complete the route without any harm to their health and correctly recall the perceived stress after a trial. According to previous studies applying ecological momentary assessment for nearly real-time assessment of perceived status (e.g., stress and emotion) in a naturalistic setting (Delespaul 1995; Jacobs et al. 2005; Kasanova et al. 2018; Myin-Germeys et al. 2005; Nezlek et al. 2008), people can accurately recall their status within maximum 15 minutes. Considering seniors' impaired memory, the duration was limited up to 10 minutes to minimize recall bias.

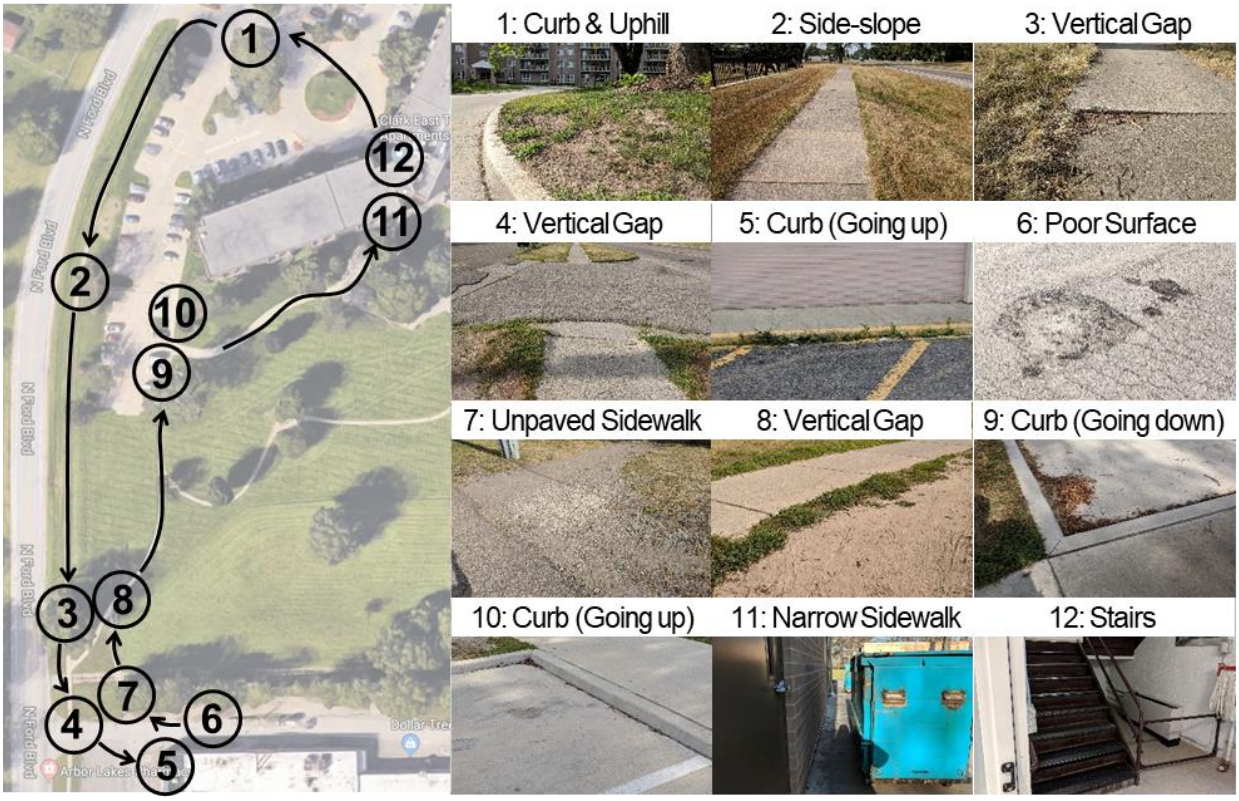


Figure 5.2. Environmental stressors used in the controlled route data collection

While moving along the route and experiencing the potential stressful interactions, the subjects' EDA and PPG signals were collected by the wristband-type biosensor. The sampling rates for EDA and PPG are respectively 4 Hz and 64 Hz. One research staff member accompanied and guided the subjects along the route, and another staff member took a video of the subjects moving, which was referenced when labeling the collected biosignals. After completing the route, the subjects' perceived stress was surveyed. First, the terms "stress" and "stressful interactions" were explained to the subjects. Also, examples of physically and cognitively stressful interactions were provided to enhance their understanding of stress. Then, whether the subjects actually felt stressed or not on each potential location was asked in a binary manner (i.e., true or false). To help the subject recall his/her experience, the authors showed pictures of each potentially stressful location. Based on the video of the subjects moving and their responses to the stress survey, biosignals were labeled. The authors reviewed all the video recordings of trials and excluded video segments recorded when the subjects experienced unintended events that could affect their stress (e.g., interacting with cars or other people, and losing balance while walking). Then, biosignals

collected from stressful location confirmed by the subjects are labeled as “high-stress” and “low-stress,” respectively.

5.3.2 Daily trip data collection

In the daily trip data collection, the subjects were asked to wear two wearable sensors for two weeks whenever they go out outside their home. Along with a wristband-type biosensor, a belt clip-type GPS sensor was used to collect subjects’ locations every second (1 Hz). The authors provided a 30-minute training session per each group to help the subjects understand the use of the wearable sensors before collecting the data. During the daily trip data collection, the authors visited the subjects twice a week to make sure that the subjects were properly using the wearable sensors and to download the stored data in these sensors.

5.3.3 Post hotspot investigation

After detecting stress hotspots based on the collected data, the authors also inspected locations of the detected hotspots while taking videos and pictures. Then, one-to-one interviews with each senior subject were conducted to better understand his/her experience on the hotspots because every subject can be involved in different stress hotspots. Each interview took approximately 30 minutes to complete. During the interview, an annotated map, pictures, and videos of stress hotspots were provided to help the subjects understand each hotspot’s location and recall their experience on it. After that, the subjects were asked which factors made them stressed on the hotspot locations. The authors took detailed notes of the subjects’ answers during the interviews.

5.4 Pilot study result

5.4.1 Best individual stress classifier from controlled route data collection

Through the controlled route data collection, a total of 18,880 samples were collected. Among them, 1,871 samples were labeled into “high-stress” while other 17,009 samples were labeled into “low-stress.” Since the number of low-stress samples was far greater than that of high-stress samples, the resultant classifier may not be accurate for predicting stress, which is a minority class. To avoid such a problem, the “non stress” class is first randomly under-sampled to make the two classes have the same number of samples. 10-fold cross-validation without shuffling was

conducted to select the best classification algorithm. By repeatedly undersampling and conducting the 10-fold cross-validation 20 times, averages of all the tested algorithms' validation accuracy were calculated. Three accuracy metrics were used for comparison: accuracy, precision, and recall. Multiple machine learning algorithms, including gaussian support vector machine, decision trees, bagging tree, and deep learning, were tested and compared, and among the tested algorithms, the bagging tree brought the best validation accuracy, 92.5 %. Therefore, the bagging tree-based model was selected and applied to detect subjects stress from biosignals collected in the daily trip data collection.

5.4.2 Hotspots detected from daily trips

Using the selected bagging tree-based classifier, the 30 senior subjects' EDA and PPG signals collected from their daily trips were classified into "high-stress" and "low-stress." As a result, 2.51 % of samples were classified into "high-stress." There was no meaningful relationship between the subjects' demographic characteristics and their detected stress. After this individual senior subject's stress classification, the multiple stress samples were distributed on the grid-cells to conduct the hotspot detection. The senior subjects' samples were assigned on total 64,914 grid-cells. Since hotspots should be detected based on subjects' multiple experiences, 18,024 grid-cells (27.7% of the total grid-cells) on which more or equal to seven stress samples were located were included in this study. The hotspot detection was conducted on these grid-cells with 0.01 statistical significance level. 0.01 significance level has been widely used to statistically detect hotspots (Bhunia et al. 2013; Pandey et al. 2014). As a result, 40 grid-cells (0.22% of tested grid-cells) were detected as the hotspots. Figure 5.3 shows the locations of the detected hotspots. Out of 40 hotspots, 35 were located near the senior apartments, while the other five hotspots were detected in other areas (e.g., Ypsilanti downtown). More specifically, 12 hotspots were detected on sidewalks around the senior apartment where the senior subjects frequently cross to visit a close commercial building, and 18 hotspots were located on the parking areas of the commercial building and senior apartment. Also, seven hotspots were detected near three bus stops. Last three hotspots are located on an intersection near the apartment.

Figure 5.3. Locations of detected stress hotspots

Table 5.2 shows the number of samples, high-stress samples, passing subjects and stressed subjects (i.e., subjects who had more than one high-stress sample on that hotspot) on each detected hotspot. On average, each detected hotspot has 3.2 stressed subjects with 13.6 stress samples. Particularly, some hotspots were detected based on only two or three stress samples or only one stressed subject. This can be explained due to the fact that the hotspot detection was conducted based on significance test rather than just using a fixed threshold. The significance of density of stress samples on each grid-cell was tested based on the Monte-Carlo method considering subjects' stress proportions and trajectories. Therefore, just the small number of stress samples or stressed subjects on a grid-cell does not necessarily mean that the grid-cell is not a stress hotspot. For example, if a subject who had a low proportion of stress samples became repeatedly stressed on a grid-cell, the grid-cell can be detected as a hotspot.

Table 5.2. Information of detected hotspots

Hotspot	Number of Samples	Number of High-stress Samples	Number of Passing Subjects	Number of Stressed Subjects (Stressed Subject #)
1	259	13	18	4 (#7, #8, #11, #28)
2	444	21	21	2 (#7, #8)
3	77	5	6	2 (#7, #8)
4	79	6	11	2 (#7, #8)
5	83	6	6	3 (#7, #8, #10)
6	103	6	12	5 (#7, #8, #11, #17, #27)
7	603	20	19	8 (#7, #8, #9, #10, #11, #12, #22, #28)
8	513	9	12	3 (#10, #11, #30)
9	514	11	11	1 (#30)
10	59	5	7	2 (#7, #17)
11	111	9	10	2 (#7, #9)
12	372	20	10	3 (#6, #7, #8)
13	165	17	11	3 (#6, #7, #17)
14	157	14	10	2 (#6, #7)
15	206	5	7	3 (#7, #11, #25)
16	393	11	10	2 (#8, #11)
17	299	14	11	5 (#7, #8, #11, #17, #25)
18	215	7	12	3 (#11, #12, #14)
19	392	23	15	3 (#6, #7, #8)
20	645	43	19	6 (#6, #7, #8, #10, #16, #30)
21	549	29	20	6 (#7, #8, #10, #13, #16, #18)
22	674	32	19	5 (#7, #8, #9, #14, #30)
23	840	42	23	3 (#7, #10, #17)
24	624	24	20	4 (#7, #8, #9, #10)
25	1110	43	26	6 (#7, #8, #10, #13, #17, #30)
26	117	9	10	4 (#7, #8, #10, #22)
27	346	11	14	4 (#8, #10, #13, #16)
28	409	13	9	3 (#10, #13, #17)
29	87	5	12	3 (#8, #10, #17)
30	288	14	19	5 (#7, #10, #17, #22, #26)
31	269	12	19	5 (#7, #10, #17, #22, #26)
32	74	3	9	2 (#7, #22)
33	126	5	14	3 (#6, #8, #25)
34	26	4	13	2 (#7, #30)
35	18	2	10	2 (#8, #30)
36	36	6	5	1 (#8)
37	37	3	4	1 (#7)
38	139	6	6	2 (#10, #17)
39	240	3	5	1 (#10)
40	389	11	6	2 (#10, #17)
Mean (SD)	302.2 (250.1)	13.6 (11.1)	12.5 (5.6)	3.2 (1.6)
Median	249.5	11	11	3

5.4.3 Post hotspot investigation in this pilot study

To see how well the hotspots detected by the proposed framework can represent the seniors' stressful interaction with built environment, the authors visited the 40 hotspots to investigate them. The author found two categories of the hotspots: pedestrian-related hotspots and automobile-related hotspots. Hotspots 1-11 and 36 (1-11 and 36 in Figure 5.4) were categorized as pedestrian-related hotspots. These hotspots were associated with the stressful interaction with bad quality of footpaths. For example, Hotspots 1, 2, 6, 9 and 36 (1,2,6,9, and 36 in Figure 5.4) are related to a vertical displacement on sidewalks. On Hotspots 1 and 2, high-stress was detected from four subjects (Subjects 7, 8, 11, and 28). All the four subjects felt high stress while moving by foot or assistive devices (i.e., speed between 1 and 4 km per hour). During the interview, these subjects said that when they moved over the vertical displacement, they should carefully watch their steps. Also, three subjects (Subjects 7, 8, and 11) using a walker as an assistive device mentioned that they felt high stress or discomfort because of vibrations caused by the vertical displacement. In case of Hotspots 7 and 8 (7 and 8 in Figure 5.4), it was found that the senior subjects became stressed due to the unpaved sidewalks. On these hotspots, nine subjects (Subjects 7, 8, 9, 10, 11, 12, 22, 28, and 30) who passed over the spots became stressed. All the high-stress was detected when the subjects were moving by foot or assistive devices, not standing there. Specifically, six out of the nine stressed subjects (Subjects 7, 8, 10, 11, 12, and 22) depend on a walker or an electric scooter for moving. They mentioned that while moving on the unpaved footpath, they suffered from continuous vibration coming from the wheel-based assistive devices. They also said that on rainy or snowy days, the unpaved footpath became muddy, which made passing over it more stressful. Like these two cases, environmental stressors that can induce the seniors' stressful interaction such as a side-sloped sidewalk (Hotspot 3), a wide joint (Hotspots 4 and 5) and an unmowed lawn (Hotspots 10 and 11) were found on the other pedestrian-related hotspots.

On the other hand, some hotspots were detected in areas where people had interactions with automobiles such as cars or buses. For instance, several hotspots were located near bus stops frequently used by the senior subjects (i.e., Hotspots 27-29: a bus stop in front of the senior apartment, 37: a bus stop one block away from the senior apartment, and 38-40: a bus stop located on the downtown of Ypsilanti) (27-29 and 37-40 in Figure 5.4). On these hotspots, most of the high-stress samples were collected when the subjects were staying there for a while (at least 30 seconds). The subjects remarked in the interview that long waiting times for buses were stressful

especially when there was no bench where they can sit near a bus stop. They also reported the physical demands of getting on and off buses, which could be stressful depending on their daily conditions like weather. In addition to bus stops, several hotspots were detected on parking areas (i.e., Hotspots 12-26: parking areas of the senior apartments, 30-32: parking areas of the commercial building nearest from the senior apartment) (12-26 and 30-32 in Figure 5.4). It was found through the interview that getting on and off from cars could also be stressful. For senior individuals who depend on assistive devices (e.g., walker), the process of placing and removing their assistive devices was especially stressful. Also, subjects mentioned that interacting with passing cars can be stressful when people's paths overlap with those of cars in parking areas. Lastly, Hotspots 33-35 were detected on an intersection near the senior apartment (33-35 in Figure 5.4). On these hotspots, five subjects (Subjects 6, 7, 8, 25, 30) had high-stress samples there. It is expected that the five stressed subjects were riding vehicles at the moment when they became stressed because their speed was recorded as over 7 km per hour. The subjects remarked that the high stress might be induced by a sudden change of car movement such as stopping short due to signal change at the intersection. Overall, the hotspot investigation showed that the hotspots detected by the proposed framework were well matched with the locations of the senior subjects' stressful interactions with built environment.

Figure 5.4. Pictures of the detected hotspots

5.5 Discussion

Current wearable-based techniques just detect the occurrence or level of stress, without providing any contextual cues about stressors. To design efficient interventions for alleviating people's stressful interactions with CBEs, it is imperative to fill the gap in the wearable-based stress detection. In this regard, this study proposed a GSI hotspot analysis-based technique to detect stress hotspots that are spatially correlated with the locations of environmental stressors in CBEs. Through the post hotspot investigation in the pilot study, the authors found that the detected hotspots are spatially matched with the senior subjects' stressful interactions with their daily built environment. This result demonstrates that the proposed stress hotspot detection has a potential to spatially locate people's stressful interactions with their surroundings, which is highly useful information in identifying and addressing related environmental stressors.

Through the interviews with senior subjects, the author found that the proposed wearable- and GIS hotspot-based technique could allow us to understand "time-dependent" environmental stressors that the site survey might fail to detect, such as interactions with passing vehicles at parking areas and buses at bus stops. It might be hard for the current survey-based approaches to find these time-dependent environmental stressors due to their discontinuity. The result of this study shows that the stress of senior people resulting from these time-dependent environmental stressor could be substantial, which again demonstrates the need for a continuous stress detection technique to understand the quality of interactions between people and CBEs.

Although this study demonstrated the potential of the use of wearable biosensor and GIS-based hotspot analysis to monitor people's stressful interactions with their surroundings (e.g., CBEs) in their daily lives, there are several limitations, which should be overcome. First, the controlled route data collection considered limited types of stressful interactions people experience though there are lots of other types between humans and CBEs. Particularly, the controlled route data collection only simulated physically stressful interaction. However, people's stressful interactions with CBEs can be cognitive or emotional. For instance, complex transit systems can impose excessive cognitive demands onto individuals. Also, some time-specific stressful interactions such as walking in a dark alleyway at night should be considered. If such time-specific stressful interactions are considered in the controlled route data collection and sufficient data is collected in each time period, the proposed technique could be used to detect stress hotspots in different time periods. To overcome these limitations, future studies should collect data for diverse

levels of stress coming from more other types of stressful interactions. Furthermore, more diverse weather conditions should be accommodated for the practical use of the proposed technique.

5.6 Conclusions

This study developed a wearable biosensor- and GIS hotspot analysis-based technique to locate people's stressful interactions with their environments, which can be very useful to identifying underlying stressors and designing effective stressor-specific interventions to reduce people's stressful interactions with their environments. Specifically, in the technique, a machine learning classifier first detects individuals' stress using biosignals collected by wearable biosensors in their daily trips. Then, a hotspot analysis statistically identifies stress hotspots by spatially aggregating individual stress data. A pilot study with 30 senior subjects living in Ypsilanti Township (MI) was conducted to test the proposed technique. As a result, the post hotspot investigation showed that 40 stress hotspots detected by the proposed technique were well spatially matched with the locations of the senior subjects' stressful interactions with their daily built environment. The findings demonstrate that the proposed technique can locate human stress caused by interactions with surroundings. The proposed stress hotspot detection technique can help to identify environmental stressors that reduce the quality of interactions between humans and CBEs and ultimately design effective interventions to promote the quality of people's experience in CBEs.

Chapter 6 Mobile Electroencephalography (EEG)-Based Stress Type Classification

6.1 Introduction

The previous study aimed to locate stressors, thereby supporting identifying stressors and designing stressor-specific effective interventions. This study aims to provide another important circumstantial clue of stress: the type of stress (i.e., positive or negative stress). It is essential to differentiate between stress types when planning effective stress relief interventions because different types of stress have completely opposite impacts on people (LeBlanc 2009; Tomaka et al. 1993).

According to how the person experiencing a stressor appraises it and his/her related coping resources (cognitive appraisal on stressors), stress response typically falls into one of three categories: low-stress, eustress, or distress (LeBlanc 2009). While low-stress, which occurs when a person perceives that the stressor has little to do with his/her current or future well-being, does not have much impact, distress and eustress do and their impacts are opposite from one another (LeBlanc 2009; Tomaka et al. 1993). Eustress occurs when the stressor is assessed as a "challenge," wherein successful coping leads to positive impacts and one's coping resources/capabilities are perceived as able to meet the demand (Lazarus and Folkman 1984; Tomaka et al. 1993). Eustress can improve overall worker performance by enhancing self-efficacy, focus, and motivation (Folkman 1984). On the other hand, if the stressor is assessed as a "threat" whose outcome could be harmful to the person's wellbeing or social status and the demands are perceived to outweigh their coping resources/capabilities, distress ensues and potentially poses detrimental consequences to workers through several pathways such as chronic lethargy, depression, and reduced task-focus (Lazarus and Folkman 1984; Tomaka et al. 1993). As such, distinguishing between stress types to selectively alleviate distress could be effective in managing its various negative consequences while still maintaining the positive impacts of eustress.

Among biosignals collectable using wearable biosensors, electroencephalogram (EEG) can be an excellent means for understanding stress types. Since the brain is the central organ for

cognitive appraisal of stressors, different appraisal processes well manifest in a person's brain activities according to different stress types. For example, people have different emotional (Balters et al. 2020; Tomaka et al. 1993) and cortisol responses (Balters et al. 2020; Dickerson and Kemeny 2004; LeBlanc 2009) according to stress positivity or negativity, both of which are represented through the different patterns of brain activities over multiple brain regions (e.g., prefrontal cortex, hippocampus, and amygdala).

Despite such potential from using mobile EEG sensors, there is little existing research on applying EEG sensors to differentiate stress types in people's daily work and lives in CBEs. Although a couple of studies have applied mobile EEG sensors to understand construction workers' stress in the field, their object was simply to detect stress levels (Arpaia et al. 2020; Jebelli et al. 2018), not to understand stress types, which is essential contextual information for effective stress relief management. To fill this gap, this study proposes a mobile EEG sensor-based stress type classification technique and tests its performance in a daily life-similar setup.

6.2 Proposed mobile EEG-based stress type monitoring

Figure 6.1 describes an overview of the EEG-based stress type classification technique proposed in this study. The first step of the proposed technique is to collect EEG signals from a mobile head-cap (a in Figure 6.1). Then, a series of denoising techniques is applied to alleviate noise in the collected EEG signals (b in Figure 6.1). The denoised EEG signals are fed into a specifically designed deep learning model which will be trained to classify EEG signals as three different types of stress (i.e., low-stress, eustress, and distress) (c in Figure 6.1).

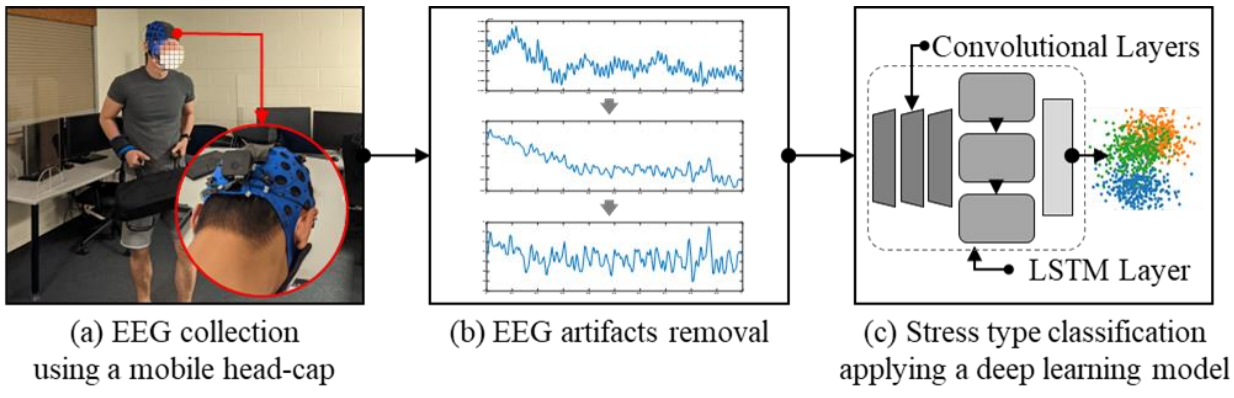


Figure 6.1. Overview of the proposed EEG-based stress type classification technique

6.2.1 EEG signal collection using a mobile head-cap-type sensor

To collect people's EEG signals non-invasively during their ongoing work, this study applies an off-the-shelf mobile head-cap EEG device (Mentalab Explore). This EEG device collects EEG signals using a set of eight dry electrodes that do not require conductive gel, thus providing significantly better wearability than previous EEG devices with gel-type electrodes. The electrodes' locations can be flexibly determined. The authors customized this device to apply the adaptive EEG motion artifact removal introduced in Chapter 2 (a in Figure 6.2). Specifically, one electrode is flipped and isolated from the human scalp and a conductive fabric covers the flipped electrode from the outside, thereby shorting between the flipped electrode and a ground electrode. In this setup, the conductive fabric experiences similar electromagnetic interference to the user's scalp while their head moves and so the flipped electrode solely collects the reference of electromagnetic interference induced by motion.

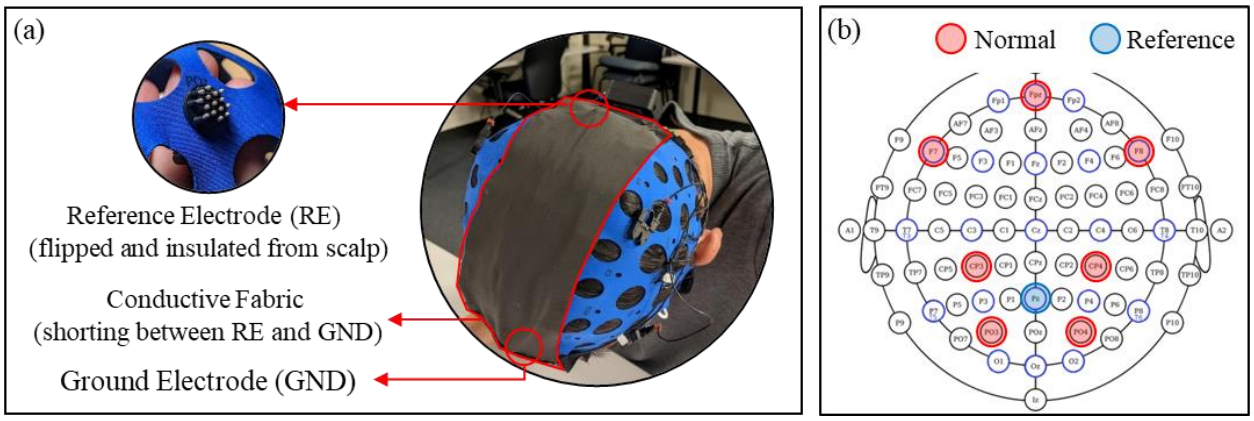


Figure 6.2. Setup for a mobile EEG device; (a) customization for an adaptive motion artifact removal; (b) location of electrodes

Except for the reference electrode, seven other electrodes are located evenly over the entire scalp (Fpz, F7, F8, CP3, CP4, PO3, PO4), as shown in b in Figure 6.2, which differs from previous mobile EEG studies' setups where most of the electrodes are located on the frontal area. This setup is intended to more effectively monitor limbic areas near the center of the brain (e.g., hippocampus and amygdala) (Cannon et al. 2005) which are known to exhibit different activity patterns in accordance with different emotional and cortisol responses (Cannon et al. 2005; Herman and Cullinan 1997). In this study, the sampling rate was set by 250 Hz.

6.2.2 **Denoising *EEG signal***

Since EEG recording is highly vulnerable to various sources of noise, it is essential for a reliable EEG analysis to appropriately denoise collected EEG signals (Jebelli et al. 2017). As a first step of denoising, this study applies low-pass and a high-pass filters with frequencies of 64 Hz and 0.5 Hz respectively. This step aims to alleviate extrinsic noise such as aliasing and drifting (Jebelli et al. 2017). Also, a notch filter targeting frequencies close to 60 Hz is applied to suppress the power line interference-induced noise (Teplan 2002). Following the aforementioned basic filtering techniques, the noise reference-based adaptive EEG motion artifact denoising, developed in Chapter 2, is applied to suppress motion artifacts in EEG, which are the most significant source of extrinsic noise in the field (Nordin et al. 2018). As the final denoising step, intrinsic artifacts caused by ocular and muscular activities are alleviated by applying a hybrid wavelet-enhanced independent component analysis (ICA) (Castellanos and Makarov 2006). This technique first identifies independent components from multichannel EEG signals. Then, the identified independent components are decomposed individually in wavelet coefficient by applying wavelet transformation. Then, the wavelet coefficients are thresholded to exclude real brain activity factors from the independent components based on empirical findings about different signal characteristics between real brain activities and artifacts. After artifacts are reconstructed by combining the thresholded independent components, the reconstructed artifacts are subtracted from the input EEG signals, thereby alleviating intrinsic artifacts. Figure 6.3 demonstrates the overall proposed denoising process.

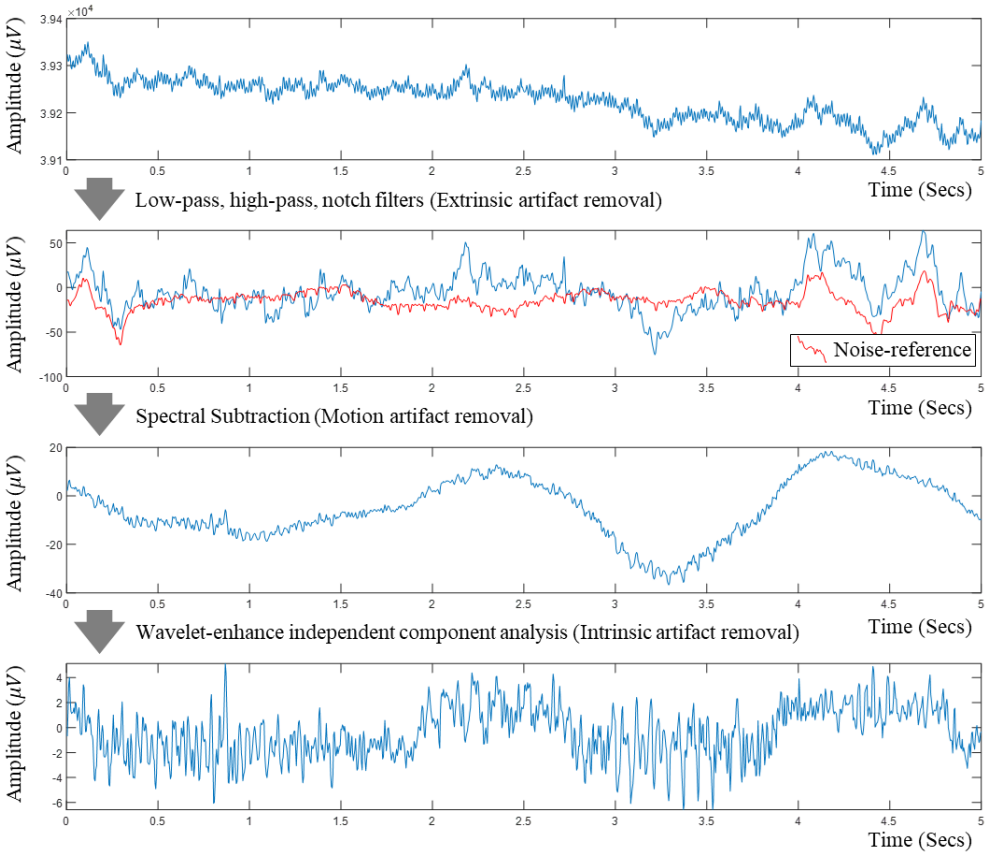


Figure 6.3. Proposed denoising process

6.2.3 Deep learning-based stress type classification

Denoised EEG signals are used as input data for a deep learning-based classification model that is trained to classify EEG signals into three types of stress (i.e., low-stress, eustress, and distress). In previous studies, deep learning models proved useful for understanding EEG patterns according to human psychophysiological status such as stress and emotion (Jebelli et al. 2019; Pandey and Seeja 2019). Therein, a deep learning architecture composed of multiple convolutional and long short-term memory (LSTM) layers is applied. It has also been found that 1-dimensional convolutional layers can extract meaningful features to indicate morphological characteristics of each biosignal locality such as EEG (Seo et al. 2019). An LSTM layer can obtain sequential information about features extracted from convolutional layers—important for practicalizing time series data patterns such as biosignals (Masood and Alghamdi 2019; Seo et al. 2019).

Specifically, denoised EEG signals are segmented by a window with 2-second length and 1.5-second overlap (75%), so the 2-second-long EEG segments (2 seconds * 250 Hz = 500 samples) are fed into the deep learning model as data points. The window length was set based on previous studies where it was found that a 2-second length is enough to see different brain activity patterns by valence and arousal, the important variables for understanding stress types (Lewis et al. 2007). Also, the overlap was determined by 75% because previous studies have shown that 75% overlap can ensure enough independence between neighboring data points while providing a chance to have more data points, thereby enabling us to effectively train models to understand different patterns in human biosignals such as EEG (Rezek and Roberts 1998).

The proposed deep learning model first splits 2-second-long EEG into five 0.4-second-long localities (0.4 seconds * 250 Hz = 100 samples). These five localities are fed into a series of five 1-dimensional convolutional modules. Each convolutional module consists of 1-dimensional convolutional, maxpool, batchnorm, leaky ReLU activation, and drop out layers in order. Once the multiple convolutional modules extract morphological features from each locality of biosignals, the extracted features are transferred through a flattened layer to one LSTM layer where the features' sequential information is obtained. Finally, a classifier part—composed of (in order) a fully connected layer, a leaky ReLU activation, the other fully connected layer, and a softmax layer—is applied to differentiate between three classes of stress types (i.e., low-stress, eustress, and distress). All the hyperparameters of the deep learning model were set by empirically predetermined values as shown in Table 6.1.

Table 6.1. Hyperparameter setup for the deep learning model

Kernel lengths (five convolutional layers in order): 50-20-10-5-3, Channel number: 25, Maxpool length: 2, Maxpool stride: 2, Drop out: 0.5, LSTM hidden unit number: 32, Hidden node number in the classifier part: 6, Learning rate: e-03, Number of Epoch: 150, L2 regularization: e-07, Minibatch size: 64

6.3 Feasibility test

To see the feasibility of the proposed mobile EEG-based monitoring to differentiate between different stress types, EEG signals were collected in a real construction task-similar setup and labeled as stress types (low-stress, eustress, distress). The authors selected a construction task (i.e.,

material handling task) as the base task so as to test the proposed technique under challenging levels of motion artifacts, which is the most significant hinderance to field applications of mobile EEG. Specifically, ten graduate students at the University of Michigan were recruited as subjects in the data collection (Table 6.2). Subjects individually participated in the data collection twice on two different days. On each day, subjects were first asked to have a 10-minute relax session. During the relax session, a nature video with sound was provided via a laptop. This relax session aimed to minimize the impact of external factors on subjects' stress and to collect EEG signals under a low-stress condition. EEG signals collected from the last five minutes of the relax session were labeled as low-stress assuming that the impact of external events prior to data collection were alleviated enough during the first five minutes.

Table 6.2. Demographic information of 10 subjects in data collection

Statistics	Age (years)	Height (cm)	Weight (kg)
Mean (SD)	31.2 (2.1)	177.9 (4.9)	76.6 (6.0)
Maximum	36	187	89
Minimum	29	172	70

After the relax session, subjects were asked to conduct multiple sessions of a material handling task typical of construction work. Specifically, on each day, subjects conducted seven sessions of material handling tasks (14 sessions total over two days), each of which took two minutes. Out of the seven sessions, three sessions were designed to elicit eustress while the remaining four were designed to cause distress. For the three eustress sessions, subjects carried sandbags with a moderate load (10 kg) between two spots 10 m apart. The working pace was controlled at 15-20 seconds per carry during the sessions. The pace was set to keep work intensity within a moderate range (3-6 metabolic equivalent of task (MET)) (Norton et al. 2010), under which people tend to experience eustress symptoms (e.g., plasma norepinephrine decrease. reduction in depression and tension) rather than those of distress (Berger and Motl 2000; So et al. 2017).

During the four distress sessions, subjects conducted the same sandbag carrying task but with additional factors to elicit distress. Two typically distress-eliciting protocols were applied: cold pressure and others' evaluation. During the distress sessions on the first day of data collection,

subjects wore an arm wrap composed of gel packs cooled to 0-5 °C while carrying the sandbags. Cold pressure is one of the most typical stressors to induces distress by directly disturbing human physiological homeostasis (Dickerson and Kemeny 2004; Herman and Cullinan 1997; Sawchenko et al. 2000). On the second day, subjects were asked to follow a set of ergonomic rules during sandbag carrying (e.g., not twisting his/her waist while lifting, keeping his/her elbows to body while carrying a sandbag) and research staff continuously watched how they did and checked every time subjects violated the rules. Also, to encourage subjects to follow the rules, subjects were informed that each session would be evaluated as a success or failure based on whether or not they exceeded five rule violations and that they needed to achieve at least three successful sessions by the end of data collection. Under this setup, subjects might feel judged, which is one of the most typical psychological stressors to elicit a distress response (Sawchenko et al. 2000). To prevent impacts of eustress and distress sessions from mixing, 25 minutes of break time were put between the two sessions. Additionally, the two sessions were counterbalanced to minimize bias caused by the temporal sequence. Specifically, five subjects were exposed to eustress sessions first prior to distress sessions while the other five subjects were exposed distress sessions first.

To confirm what stress type subjects experienced, they were asked after each session to self-report their arousal and valence using the Self-assessment manikin scale (SAM) (Bradley and Lang 1994). A high level of arousal (5 or more on the 9-point arousal scale) is a common symptom of stress regardless of eustress or distress (LeBlanc 2009), negative valence (4 or less on the 9-point valence scale) is a unique symptom of distress, and positive valence (6 or more on the 9-point valence scale) is one of eustress (Balters et al. 2020; Tomaka et al. 1993). Correspondingly, EEG signals recorded during eustress sessions that subjects rated as high arousal and positive valence were acquired and labeled as eustress. Likewise, EEG signals recorded during distress sessions rated as high arousal and negative valence were collected as distress. The data collection protocol was approved by the University of Michigan Institutional Review Board (IRB00000245).

The proposed deep learning model was then trained and tested using these EEG signals labeled as low-stress, eustress, and distress. Specifically, the 10-fold cross validation without shuffling was applied as the testing method. This cross validation tests subject-specific models like normal 10-fold cross validation with shuffling, which has been most widely used in previous studies that use machine learning for human physiological understanding. It was determined that cross validation without shuffling (keeping the temporal sequence while splitting training and

testing subsets) is more valid than cross validation with shuffling to see the generalization performance of models by testing them using data from a truly "unseen" period.

6.4 Results and discussion

Through data collection, a total of 5 hours of labeled EEG signals were collected (about 30 minutes per subject). The numbers of data points for low-stress, eustress, and distress classes were 7,663, 11,987, and 14,666 respectively. Using these EEG signals, the proposed deep learning model was trained and tested. Table 6.3 is the confusion matrix that shows the deep learning model's performance assessed by 10-fold cross validation without shuffling. The overall accuracy of the classification for all three classes was 0.842 and F-1 measures for all three classes, calculated using their precision and recall values, were 0.918, 0.772, and 0.855 (low-stress, eustress, and distress) respectively. These results indicate that the proposed EEG-based technique is feasible to differentiate between different stress types, thereby enabling selective stress relief interventions that only alleviate distress.

Table 6.3. Confusion matrix

Overall Accuracy: 0.842		Actual			Precision
		Low-Stress	Eustress	Distress	
Prediction	Low-Stress	6,692	170	52	0.968
	Eustress	454	8,525	1,124	0.844
	Distress	517	3,292	14,666	0.794
Recall		0.873	0.711	0.926	

Given that this feasibility study only applied empirically predetermined hyperparameter values to the deep learning model without fine-tuning, the authors might need to examine how much the proposed technique's performance can be improved through fine-tuning. Despite the promising results of this study, a considerable number of eustress data points were misclassified as distress as shown in Table 6.3, thereby decreasing the distress class's precision. Also, fine-tuning of the hyperparameters was not conducted in this study. Since accurately detecting distress might be most important stress type classification task, increasing the distress class's precision might be the main focus of fine-tuning in future studies. In addition, a subject-specific model was trained and tested in this feasibility study. Given that subject- and context-independency is important for extensively applying the proposed technique in real field applications in CBEs, a

future study that ensures the proposed technique's generalization across different subjects and contexts is also needed.

6.5 Conclusion

Although there have been efforts to understand human stress during their daily lives in CBEs (e.g., ongoing work at construction sites and trips in cities or buildings) via wearable biosensors, these studies have only provided rudimentary information about individuals' stress (e.g., binary stress levels), which may not be overly useful for designing stress-relief interventions. Since understanding different types of stress is essential information about stress for designing effective stress management which can be specific to the impact of different stress cases, this study examines the feasibility of applying a mobile EEG device to differentiate between different stress types (low-stress, eustress, and distress) while suggesting a series of denoising techniques and a deep learning architecture. The proposed EEG-based technique was tested using EEG signals collected in a real construction task-like setup where the level of actual body movement was carefully simulated. The test's overall accuracy in classifying stress types was 0.842. This promising result demonstrates that the proposed EEG-based technique is feasible to understand different stress types, which is valuable contextual information helping us to understand people's quality of experience. This finding can contribute to promoting the quality of interactions between humans and CBEs by enabling stress impact-specific (or, stress type-specific) selective interventions that only alleviate distress while keeping or promoting benefits of eustress.

Chapter 7 Conclusions and Recommendations

7.1 Summary of research

This research effort began with the following overarching research goals: (1) to develop field-applicable denoising techniques that effectively alleviate not only stationary artifacts, but also non-stationary artifacts in biosignals collected in the field; (2) to advance the generalizability of wearable biosensors- and machine learning-based stress detection so that they can reliably work for unseen people under unseen contexts; and (3) to non-invasively acquire circumstantial information about stress while detecting stress, thereby enabling the design of circumstance-specific effective stress relief interventions in CBEs. Considering these goals, the research had five specific research objectives: (1) to denoise both stationary and non-stationary artifacts in biosignals collected during people's daily work and lives in CBEs; (2) to reliably assess generalizability of machine learning models for tasks monitoring human responses from biosignals; (3) to advance model generalizability across different subjects and contexts in detecting stress using a wearable biosensor; (4) to distinguish and locate stress responses related to environmental features; and (5) to differentiate stress types into positive (i.e., eustress) and negative (i.e., distress) types.

To achieve these objectives, five inter-related studies were conducted. A summary of these studies' results and their implications are as follows.

1. ***Noise reference signal-based adaptive denoising for non-stationary artifacts in biosignals collected in the field***: This study developed two adaptive denoising techniques that collect and leverage a noise reference signal to effectively alleviate non-stationary artifacts in EDA and EEG, two useful signals in understanding human stress. Specifically, to alleviate respiratory artifacts in EDA, photoplethysmography (PPG) was simultaneously collected as a respiratory artifact reference with EDA using a multimodal wristband-type biosensor. To suppress motion artifacts in EEG, reference electrodes were set up by isolating them from the human scalp so that they collect only motion artifact references

while experiencing almost identical movements and electromagnetic gradients with normal electrodes. Then, the artifact references were subtracted from the raw signals by developed algorithms (e.g., sparse decomposition and constraint independent component analysis). The proposed denoising techniques showed statistically higher denoising performance than advanced benchmarks with almost zero p-values in all examined denoising performance indices. These results demonstrate that the proposed noise reference-based adaptive denoising method is effective for alleviating non-stationary artifacts in biosignals collected in the field.

2. ***Subject- and context-independent validation method to assess generalizability of machine learning models for monitoring human responses from biosignals***: Validating generalizability is of the utmost importance in developing field applicable wearable biosensor-based stress detection. As the first effort to ensure generalizability, this study presented a new validation method that more reliably assesses generalizability of machine learning models for tasks monitoring human responses, such as stress, from biosignals. Given the huge individual and contextual variabilities in biosignal patterns, the proposed leave one subject and context out cross validation (LOSCOCV) was designed to ensure that a testing dataset is collected from unseen subjects and unseen contexts not considered in the training phase. The proposed LOSCOCV's generalizability assessment performance was compared with existing, widely applied validation methods (i.e., k-fold cross validation and leave one subject and cross validation) in a test in which training and validation were conducted with an in-lab dataset with relatively low variability. Then, testing was conducted with a field dataset with relatively high variability. The proposed LOSCOCV showed statistically higher generalizability assessment performance than the benchmarks with almost zero p-values. This result shows that testing machine learning models using an unseen subject and unseen context dataset is crucial to assessing generalizability, and so the proposed LOSCOCV can be more valid than currently applied machine learning validation methods for tasks that monitor human responses from biosignals.

3. ***Deep learning domain adaptation-based subject- and context-independent stress detection***: As the first effort to ensure generalizability of stress detection, the previous study presented a new validation method that more reliably assesses the generalizability

of machine learning models for monitoring human responses from biosignals. As a follow-up, this study proposed a transfer learning-based stress detection technique that works with generalizability across different subjects and contexts. The proposed technique conducts domain adaptation, a specific type of transfer learning by adopting a generative adversarial network (GAN)-integrated deep learning structure that actively buffers domain differences between different people and contexts to detect stress in a subject- and context-independent manner. The proposed technique showed much higher generalizability than existing benchmarks (i.e., deep learning without domain adaptation and gaussian support vector machine) in the LOSCOCV and in testing with a real field dataset. This result demonstrates that the proposed technique can significantly contribute to the scalable field application of wearable biosensors- and machine learning-based stress detection.

4. ***Geographic information system (GIS)-based stress hotspot detection***: The current wearable-based stress detection techniques let us know the level or occurrence of stress but do not provide circumstantial information about the detected stress, which is critical in designing effective stress-relief interventions. As the first effort to acquire stress-related circumstantial information, this study presented a geographic information system (GIS)-based technique that accumulates people's stress data on a map and identifies locations of environmental stressors, which is helpful information in identifying stressors in CBEs and ultimately in designing stressor-specific effective interventions. The proposed technique statistically identifies stress hotspots on which individuals' detected stress is concentrated with an abnormally high density by spatially aggregating individual stress data. The result of the pilot study with 30 senior subjects showed that stress hotspots detected by the proposed technique were well spatially matched with the locations of actual environmental stressors the senior subjects suffer in their daily built environment. The proposed stress hotspot detection technique can help to identify environmental stressors that reduce the quality of interactions between humans and CBEs and can ultimately inform the design of effective interventions that promote the quality of people's experiences in CBEs.
5. ***Mobile electroencephalography (EEG)-based stress type classification***: The previous study aimed to locate stressful interactions between humans and CBEs, thereby helping to identify stressors and design stressor-specific interventions. On the other hand, this study focused on differentiating two types of stress, whose impacts on individuals are opposite

from one another, so that we can design stress type-specific selective interventions that alleviate only negative stress (distress) while maintaining the benefits of positive stress (eustress) in CBEs. This study applied a mobile EEG device to differentiate between different stress types (low-stress, eustress, and distress) while suggesting a series of EEG denoising techniques and a deep learning architecture. The result of a test conducted in a real construction task-like setup showed that the proposed EEG-based technique's overall classification accuracy for the three classes task is 0.842. The result confirmed that the proposed mobile EEG-based technique can understand different stress types, which is valuable contextual information for understanding people's quality of experience. This finding can contribute to promoting the quality of interactions between humans and CBEs by enabling selective interventions that only alleviate detrimental distress while keeping or promoting the benefits of eustress.

7.2 Final remark

One of the most important factors in the management of both construction and built environments (CBEs) is to ensure human health, safety, comfort, and productivity; human workers are the most important resource at construction sites and the operation of most built environments places the highest priority on serving people optimally. However, taking care of human health, safety, comfort, and productivity in CBEs is a challenging task because every individual has unique characteristics, such as age, gender, physical and cognitive capabilities, race, and prior experience, and thus has different interactions with CBEs even under an identical setup. Wearable biosensing technology has great potential for monitoring individuals' quality of experience in CBEs through their stress levels, thereby enabling us to have more personalized management approaches toward improving the quality of their interactions with CBEs. My Ph.D. research can significantly contribute to realizing wearable biosensors' potential by advancing the field applicability of wearable-based stress detection and providing means to get stress-related information that supports effective stress relief interventions. The in-depth understanding of the quality of human-CBE interaction, acquired by the proposed wearable biosensor-based framework, will allow us to operate CBEs in an individual response-aware manner. Such individual individual responses-aware CBE operation can effectively and equitably advance diverse people's health, safety,

comfort, and productivity in CBEs and ultimately improve not only the performance of the construction industry, but only people's quality of life in built environments.

7.3 Future research

My future research will focus on realizing the human response-aware CBE operations that sense and interpret individuals' diverse psychophysiological responses and adaptively control/manage diverse CBE features, such as construction fields' collaborative robots, equipment, site lay-out, schedules, and cities' autonomous shuttles and shared micro mobilities, to optimally serve every individual. For example, collaborative construction robots can sense human coworkers' emotions and perceptions and thus learn how to behave to ensure coworkers' trust and comfort. Also, smart and connected city components such as autonomous shuttles are operated to equitably ensure diverse citizens' quality of life in urban infrastructure based on understanding of all the diverse citizens' status and locations. My Ph.D. research on monitoring human stress, one of the most useful psychophysiological responses to understand the quality of interactions between humans and CBEs, is the first steppingstone of the future research direction. In the rest of my research journey, I intend to fully realize such human response-aware smart CBEs for different applications in the construction and built environment management, while answering the below research questions step by step.

1. ***Can wearable biosensors be used to understand more diverse and in-depth psychophysiological responses to the interactions between humans and their surrounding CBEs?*** According to applications, monitoring psychophysiological responses other than stress could be essential to understanding how to improve the quality of human-CBE interactions. For example, As the global average temperature has continued to increase, excessive heat has become one significant hazard for not only workers at construction sites, but also, physically impaired citizens, such as seniors, during their outdoor activities. In this regard, continuously monitoring individuals' heat strain, the personal physiological strain to external heat, is of an urgent need to effectively manage heat-related risks in CBEs. Also, the construction industry is gradually adopting robots as human coworkers. Given that building a trust- and confidence-based strong bond between robots and human workers is critical for maximizing human/robot team performance, we

need to investigate how to measure or indicate human coworkers' trust of robots using wearable biosensors.

2. ***How can the human responses monitored by wearable biosensors be synthesized with other heterogeneous datasets collected from CBEs to acquire comprehensive context-awareness?*** Integrating human responses monitored by wearable biosensors and datasets from CBEs, such as a 4D building information model (BIM), GIS, operation data from construction robots, equipment, autonomous shuttles, and ambient thermal conditions, has the potential to provide clearer insights about what factors are related to low quality of peoples' experiences in CBEs and what interventions need to be made. To realize this potential, we need to investigate what shape of digital platform would be the best to gather and synthesize heterogeneous datasets from multiple sensor arrays and different data sources in an "up-to-current" manner, and how to extract semantic information from massive datasets to diagnose problematic situations in CBEs in order to get insights about the direction of intervention.
3. ***How can we get detailed intelligence about how to intervene in CBEs to improve human safety, health, comfort, and productivity?*** To figure out how to intervene to address an identified problematic situation in a CBE requires simulating how people psychophysiologically and behaviorally respond to diverse what-if CBE operation scenarios. Then, CBEs can operate based on the best scenarios figured out from the simulations. However, there are notable gaps in our body of knowledge about how to establish such simulations of interactions between environments and individuals' psychophysiological systems and how to ensure that the simulations have high levels of quantitative agreement with reality, which is critical to operating CBEs based on simulation results.

Bibliography

[1] Abdelhamid, T. S., and Everett, J. G. (2002). "Physiological demands during construction work." Journal of construction engineering and management, 128(5), 427-437.

[2] ACGIH (2019). "Heat Stress and Strain." TLVs and BEIs, ACGIH Signature Publications, 239–248.

[3] Ahn, J. W., Ku, Y., and Kim, H. C. (2019). "A Novel Wearable EEG and ECG Recording System for Stress Assessment." Sensors, 19(9).

[4] Anderson, D., and Burnham, K. (2004). "Model selection and multi-model inference." Second. NY: Springer-Verlag, 63(2020), 10.

[5] Anderson, T. K. (2009). "Kernel density estimation and K-means clustering to profile road accident hotspots." Accident Analysis & Prevention, 41(3), 359-364.

[6] Arbury, S., Jacklitsch, B., Farquah, O., Hodgson, M., Lamson, G., Martin, H., Profitt, A., Office of Occupational Health Nursing, O. S., and Health, A. (2014). "Heat illness and death among workers - United States, 2012-2013." MMWR Morb Mortal Wkly Rep, 63(31), 661-665.

[7] Arpaia, P., Moccaldi, N., Prevete, R., Sannino, I., and Tedesco, A. (2020). "A Wearable EEG Instrument for Real-Time Frontal Asymmetry Monitoring in Worker Stress Analysis." IEEE Transactions on Instrumentation and Measurement, 69(10), 8335-8343.

[8] Aryal, A., Ghahramani, A., and Becerik-Gerber, B. (2017). "Monitoring fatigue in construction workers using physiological measurements." Automation in Construction, 82, 154-165.

[9] ASHRAE (2017). "Standard 55–2017 thermal environmental conditions for human occupancy." Ashrae: Atlanta, GA, USA.

[10] Ayappa, I., Norman, R. G., Whiting, D., Tsai, A. H., Anderson, F., Donnely, E., Silberstein, D. J., and Rapoport, D. M. (2009). "Irregular respiration as a marker of wakefulness during titration of CPAP." Sleep, 32(1), 99-104.

[11] Balters, S., Geeseman, J. W., Tveten, A.-K., Hildre, H. P., Ju, W., and Steinert, M. (2020). "Mayday, Mayday, Mayday: Using salivary cortisol to detect distress (and eustress!) in critical incident training." International Journal of Industrial Ergonomics, 78, 102975.

[12] Barua, S., and Begum, S. "A review on machine learning algorithms in handling EEG artifacts." Proc., The Swedish AI Society (SAIS) Workshop SAIS, 14, 22-23 May 2014, Stockholm, Sweden.

[13] Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). "Analysis of representations for domain adaptation." Advances in neural information processing systems, 19, 137.

[14] Berger, B. G., and Motl, R. W. (2000). "Exercise and mood: A selective review and synthesis of research employing the profile of mood states." Journal of applied sport psychology, 12(1), 69-92.

[15] Bhunia, G. S., Kesari, S., Chatterjee, N., Kumar, V., and Das, P. (2013). "Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India." BMC infectious diseases, 13(1), 64.

[16] Bianco, S., Napoletano, P., and Schettini, R. "Multimodal car driver stress recognition." Proc., Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, 302-307.

[17] Biggs, S. E., Banks, T. D., Davey, J. D., and Freeman, J. E. (2013). "Safety leaders' perceptions of safety culture in a large Australasian construction organisation." Safety science, 52, 3-12.

[18] Bíl, M., Andrášik, R., and Janoška, Z. (2013). "Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation." Accident Analysis & Prevention, 55, 265-273.

[19] Blain, S., Power, S. D., Sejdic, E., Mihailidis, A., and Chau, T. (2010). "A cardiorespiratory classifier of voluntary and involuntary electrodermal activity." Biomedical engineering online, 9(1), 11.

[20] Bonauto, D., Anderson, R., Rauser, E., and Burke, B. (2007). "Occupational heat illness in Washington State, 1995–2005." American journal of industrial medicine, 50(12), 940-950.

[21] Bong, S. Z., Murugappan, M., and Yaacob, S. (2013). "Methods and approaches on inferring human emotional stress changes through physiological signals: A review." International Journal of Medical Engineering and Informatics, 5(2), 152-162.

[22] Borg, G. (1970). "Perceived exertion as an indicator of somatic stress." Scandinavian journal of rehabilitation medicine.

[23] Borg, G. A. (1982). "Psychophysical bases of perceived exertion." Med sci sports exerc, 14(5), 377-381.

[24] Bornoiu, I.-V., and Grigore, O. "A study about feature extraction for stress detection." Proc., Advanced Topics in Electrical Engineering (ATEE), 2013 8th International Symposium on, IEEE, 1-4.

[25] Bornoiu, I.-V., and Grigore, O. (2014). "Kohonen neural network stress detection using only electrodermal activity features." Advances in Electrical and Computer Engineering, 14(3), 71-79.

[26] Boucsein, W. (2012). Electrodermal activity, Springer Science & Business Media.

[27] Bradley, H., and Esformes, J. D. (2014). "Breathing pattern disorders and functional movement." International journal of sports physical therapy, 9(1), 28.

[28] Bradley, M. M., and Lang, P. J. (1994). "Measuring emotion: the self-assessment manikin and the semantic differential." Journal of behavior therapy and experimental psychiatry, 25(1), 49-59.

[29] Bureau of Labor Statistics (2019). "American Housing Survey (AHS)." <https://www.census.gov/programs-surveys/ahs.html>. (2021).

[30] Cacioppo, J. T., Tassinary, L. G., and Berntson, G. (2007). Handbook of psychophysiology, Cambridge University Press.

[31] Calkins, K. G. (2005). "Applied Statistics: Correlation Coefficients." Andrews University. Retrieved on June, 5.

[32] Campbell, F. (2006). "Occupational stress in the construction industry." Berkshire, UK: Chartered Institute of Building.

[33] Cannon, R., Lubar, J., Thornton, K., Wilson, S., and Congedo, M. (2005). "Limbic Beta Activation and LORETA: Can Hippocampal and Related Limbic Activity Be Recorded and Changes Visualized Using LORETA in an Affective Memory Condition?" Journal of Neurotherapy, 8(4), 5-24.

[34] Castellanos, N. P., and Makarov, V. A. (2006). "Recovering EEG brain signals: artifact suppression with wavelet enhanced independent component analysis." J Neurosci Methods, 158(2), 300-312.

[35] Chang, C.-Y., Chang, C.-W., Zheng, J.-Y., and Chung, P.-C. (2013). "Physiological emotion analysis using support vector regression." Neurocomputing, 122, 79-87.

[36] Chaspari, T., Tsiartas, A., Duker, L. I. S., Cermak, S. A., and Narayanan, S. S. "EDA-Gram: Designing electrodermal activity fingerprints for visualization and feature extraction." Proc., Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE, 403-406.

[37] Chen, W., Jaques, N., Taylor, S., Sano, A., Fedor, S., and Picard, R. W. "Wavelet-based motion artifact removal for electrodermal activity." Proc., Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE, 6223-6226.

[38] Chen, X., Peng, H., Yu, F., and Wang, K. (2017). "Independent Vector Analysis Applied to Remove Muscle Artifacts in EEG Data." IEEE Transactions on Instrumentation and Measurement, 66(7), 1770-1779.

[39] Choudhry, R. M., and Fang, D. (2008). "Why operatives engage in unsafe work behavior: Investigating factors on construction sites." Safety science, 46(4), 566-584.

[40] Chowdhury, M. E. H., Mullinger, K. J., Glover, P., and Bowtell, R. (2014). "Reference layer artefact subtraction (RLAS): A novel method of minimizing EEG artefacts during simultaneous fMRI." NeuroImage, 84, 307-319.

[41] Chrisinger, B. W., and King, A. C. (2018). "Stress experiences in neighborhood and social environments (SENSE): a pilot study to integrate the quantified self with citizen science to

improve the built environment and health." International journal of health geographics, 17(1), 17.

[42] Cohen, S., Kamarck, T., and Mermelstein, R. (1994). "Perceived stress scale." Measuring stress: A guide for health and social scientists, 10.

[43] Coniglio, S., Dunn, A. J., and Zemkoho, A. B. (2020). "Infrequent adverse event prediction in low carbon energy production using machine learning." arXiv preprint arXiv:2001.06916.

[44] Critchley, H. D. (2002). "Electrodermal responses: what happens in the brain." The Neuroscientist, 8(2), 132-142.

[45] Cui, F., Yue, Y., Zhang, Y., Zhang, Z., and Zhou, H. S. (2020). "Advancing Biosensors with Machine Learning." ACS Sensors, 5(11), 3346-3364.

[46] Daume III, H., and Marcu, D. (2006). "Domain adaptation for statistical classifiers." Journal of Artificial Intelligence Research, 26, 101-126.

[47] Dey, A. K. (2001). "Understanding and using context." Personal and ubiquitous computing, 5(1), 4-7.

[48] Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., and Kaiser, S. (2012). "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective." Journal of the Academy of Marketing Science, 40(3), 434-449.

[49] Dickerson, S. S., and Kemeny, M. E. (2004). "Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research." Psychological bulletin, 130(3), 355.

[50] Dietterich, T. G. (1998). "Approximate statistical tests for comparing supervised classification learning algorithms." Neural computation, 10(7), 1895-1923.

[51] Donate, J. P., Sanchez, G. G., and de Miguel, A. S. (2012). "Time series forecasting. A comparative study between an evolving artificial neural networks system and statistical methods." International Journal on Artificial Intelligence Tools, 21(01), 1250010.

[52] Drachen, A., Nacke, L. E., Yannakakis, G., and Pedersen, A. L. "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games." Proc., Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games, ACM, 49-54.

[53] Dragoi, G., Harris, K. D., and Buzsáki, G. (2003). "Place Representation within Hippocampal Networks Is Modified by Long-Term Potentiation." Neuron, 39(5), 843-853.

[54] Duan, L., Xu, D., and Tsang, I. W.-H. (2012). "Domain adaptation from multiple sources: A domain-dependent regularization approach." IEEE Transactions on Neural Networks and Learning Systems, 23(3), 504-518.

[55] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, SIAM.

[56] Elzeiny, S., and Qaraqe, M. "Machine learning approaches to automatic stress detection: A review." Proc., 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), IEEE, 1-6.

[57] Engle-Friedman, M., Riela, S., Golan, R., Ventuneac, A. M., Davis, C. M., Jefferson, A. D., and Major, D. (2003). "The effect of sleep loss on next day effort." Journal of sleep research, 12(2), 113-124.

[58] Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C., and Yantis, S. (2010). "Avoiding non-independence in fMRI data analysis: leave one subject out." Neuroimage, 50(2), 572-576.

[59] Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses." Behavior Research Methods, 41(4), 1149-1160.

[60] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. (2018). "Deep learning for healthcare applications based on physiological signals: A review." Computer Methods and Programs in Biomedicine, 161, 1-13.

[61] Ferguson, W. E. (1984). "STRESS AS A PRECURSOR OF DEPRESSION, PSYCHOSOMATIC ILLNESS, AND SUICIDAL IDEATION AMONG WHITE MIDDLECLASS ADOLESCENTS."

[62] Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. "Unsupervised visual domain adaptation using subspace alignment." Proc., Proceedings of the IEEE international conference on computer vision, 2960-2967.

[63] Fischer, K. W., Shaver, P. R., and Carnochan, P. (1990). "How Emotions Develop and How they Organise Development." Cognition and Emotion, 4(2), 81-127.

[64] Folkman, S. (1984). "Personal control and stress and coping processes: a theoretical analysis." Journal of personality and social psychology, 46(4), 839.

[65] Folkman, S., and Lazarus, R. S. (1985). "If it changes it must be a process: study of emotion and coping during three stages of a college examination." Journal of personality and social psychology, 48(1), 150.

[66] Gallagher, N. A., Gretebeck, K. A., Robinson, J. C., Torres, E. R., Murphy, S. L., and Martyn, K. K. (2010). "Neighborhood factors relevant for walking in older, urban, African American adults." Journal of aging and physical activity, 18(1), 99-115.

[67] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). "Domain-adversarial training of neural networks." The Journal of Machine Learning Research, 17(1), 2096-2030.

[68] Garde, A., Karlen, W., Ansermino, J. M., and Dumont, G. A. (2014). "Estimating respiratory and heart rates from the correntropy spectral density of the photoplethysmogram." PloS one, 9(1), e86427.

[69] Glorot, X., Bordes, A., and Bengio, Y. "Domain adaptation for large-scale sentiment classification: A deep learning approach." Proc., Proceedings of the 28th international conference on machine learning (ICML-11), 513-520.

[70] Goldenhar, L., Williams, L., and Swanson, N. (2003). "Modelling relationships between job stressors and injury and near-miss outcomes for construction labourers." Work & Stress, 17(3), 218-240.

[71] Greco, A., Valenza, G., Citi, L., and Scilingo, E. P. (2017). "Arousal and valence recognition of affective sounds based on electrodermal activity." IEEE Sensors Journal, 17(3), 716-725.

[72] Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., and Citi, L. (2016). "cvxEDA: A convex optimization approach to electrodermal activity processing." IEEE Transactions on Biomedical Engineering, 63(4), 797-804.

[73] Gubernot, D. M., Anderson, G. B., and Hunting, K. L. (2015). "Characterizing occupational heat-related mortality in the United States, 2000–2010: An analysis using the census of fatal occupational injuries database." American journal of industrial medicine, 58(2), 203-211.

[74] Gwin, J. T., Gramann, K., Makeig, S., and Ferris, D. P. (2010). "Removal of movement artifact from high-density EEG recorded during walking and running." Journal of neurophysiology, 103(6), 3526-3534.

[75] Hall, E. H. (1879). "On a new action of the magnet on electric currents." American Journal of Mathematics, 2(3), 287-292.

[76] Healey, J. A., and Picard, R. W. (2005). "Detecting stress during real-world driving tasks using physiological sensors." IEEE Transactions on intelligent transportation systems, 6(2), 156-166.

[77] Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., Li, R., Khine, M., Kim, J., and Wang, J. (2018). "Wearable sensors: modalities, challenges, and prospects." Lab on a Chip, 18(2), 217-248.

[78] Herman, J. P., and Cullinan, W. E. (1997). "Neurocircuitry of stress: central control of the hypothalamo–pituitary–adrenocortical axis." Trends in Neurosciences, 20(2), 78-84.

[79] Hijazi, I. H., Koenig, R., Schneider, S., Li, X., Bielik, M., Schmit, G. N. J., and Donath, D. (2016). "Geostatistical analysis for the study of relationships between the emotional responses of urban walkers to urban spaces." International Journal of E-Planning Research (IJEPR), 5(1), 1-19.

[80] Holder, G. E., Celesia, G. G., Miyake, Y., Tobimatsu, S., and Weleber, R. G. (2010). "International Federation of Clinical Neurophysiology: recommendations for visual system testing." Clinical Neurophysiology, 121(9), 1393-1409.

[81] Hwang, S., Jebelli, H., Choi, B., Choi, M., and Lee, S. (2018). "Measuring Workers' Emotional State during Construction Tasks Using Wearable EEG." Journal of Construction Engineering and Management, 144(7), 04018050.

[82] Hygge, S., and Hugdahl, K. (1985). "Skin conductance recordings and the NaCl concentration of the electrolyte." Psychophysiology, 22(3), 365-367.

[83] Hyvärinen, A., and Oja, E. (2000). "Independent component analysis: algorithms and applications." Neural networks, 13(4-5), 411-430.

[84] Iremeka, F. U., Okeke, S. A., Agu, P. U., Isilebo, N. C., Aneke, M., Ezepue, E. I., Ezenwaji, I. O., Ezenwaji, C. O., Edikpa, E., and Chukwu, C. J. (2021). "Intervention for stress management among skilled construction workers." Medicine, 100(28).

[85] Islam, M. K., Rastegarnia, A., and Yang, Z. (2015). "A wavelet-based artifact reduction from scalp EEG for epileptic seizure detection." IEEE journal of biomedical and health informatics, 20(5), 1321-1332.

[86] Islam, M. S., El-Hajj, A. M., Alawieh, H., Dawy, Z., Abbas, N., and El-Imad, J. (2020). "EEG mobility artifact removal for ambulatory epileptic seizure prediction applications." Biomedical Signal Processing and Control, 55, 101638.

[87] Jacobs, N., Nicolson, N., Derom, C., Delespaul, P., Van Os, J., and Myin-Germeys, I. (2005). "Electronic monitoring of salivary cortisol sampling compliance in daily life." Life sciences, 76(21), 2431-2443.

[88] Jaques, N., Rudovic, O., Taylor, S., Sano, A., and Picard, R. (2017). "Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation." Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing, L. Neil, and R. Mark, eds., PMLR, Proceedings of Machine Learning Research, 17--33.

[89] Jebelli, H., Choi, B., Kim, H., and Lee, S. "Feasibility study of a wristband-type wearable sensor to understand construction workers' physical and mental status." Proc., Construction Research Congress.

[90] Jebelli, H., Choi, B., and Lee, S. (2019). "Application of wearable biosensors to construction sites. I: Assessing workers' stress." Journal of Construction Engineering and Management, 145(12), 04019079.

[91] Jebelli, H., Hwang, S., and Lee, S. (2017). "EEG Signal-Processing Framework to Obtain High-Quality Brain Waves from an Off-the-Shelf Wearable EEG Device." Journal of Computing in Civil Engineering, 32(1), 04017070.

[92] Jebelli, H., Hwang, S., and Lee, S. (2018). "EEG-based workers' stress recognition at construction sites." Automation in Construction, 93, 315-324.

[93] Jebelli, H., Khalili, M. M., Hwang, S., and Lee, S. "A supervised learning-based construction workers' stress recognition using a wearable electroencephalography (EEG) device." Proc., Construction Research Congress, 43-53.

[94] Jebelli, H., Khalili, M. M., and Lee, S. (2018). "A Continuously Updated, Computationally Efficient Stress Recognition Framework Using Electroencephalogram (EEG) by Applying Online Multi-Task Learning Algorithms (OMTL)." IEEE journal of biomedical and health informatics.

[95] Jebelli, H., Khalili, M. M., and Lee, S. (2019). "Mobile EEG-based workers' stress recognition by applying deep neural network." Advances in informatics and computing in civil and construction engineering, Springer, 173-180.

[96] Jeoung, B. J., and Lee, Y. C. (2015). "A Study of relationship between frailty and physical performance in elderly women." Journal of exercise rehabilitation, 11(4), 215.

[97] Karmann, C., Schiavon, S., and Arens, E. (2018). "Percentage of commercial buildings showing at least 80% occupant satisfied with their thermal comfort."

[98] Ke, J., Zhang, M., Luo, X., and Chen, J. (2021). "Monitoring distraction of construction workers caused by noise using a wearable Electroencephalography (EEG) device." Automation in Construction, 125, 103598.

[99] Kim, J., and Fesenmaier, D. R. (2015). "Measuring emotions in real time: Implications for tourism experience design." Journal of Travel Research, 54(4), 419-429.

[100] Kim, J., Schiavon, S., and Brager, G. (2018). "Personal comfort models – A new paradigm in thermal comfort for occupant-centric environmental control." Building and Environment, 132, 114-124.

[101] Kim, J., Yadav, M., Chaspari, T., and Ahn, C. R. (2020). "Environmental Distress and Physiological Signals: Examination of the Saliency Detection Method." Journal of Computing in Civil Engineering, 34(6), 04020046.

[102] Klados, M. A., and Bamidis, P. D. (2016). "A semi-simulated EEG/EOG dataset for the comparison of EOG artifact rejection techniques." Data in brief, 8, 1004-1006.

[103] Kohavi, R., and John, G. H. (1997). "Wrappers for feature subset selection." Artificial intelligence, 97(1-2), 273-324.

[104] Ksander, J. C., Kark, S. M., and Madan, C. R. (2018). "Breathe Easy EDA: A MATLAB toolbox for psychophysiology data management, cleaning, and analysis." F1000Research, 7.

[105] Kuhn, M., and Johnson, K. (2013). Applied predictive modeling, Springer.

[106] Lan, Z., Sourina, O., Wang, L., Scherer, R., and Müller-Putz, G. R. (2019). "Domain Adaptation Techniques for EEG-Based Emotion Recognition: A Comparative Study on Two Public Datasets." IEEE Transactions on Cognitive and Developmental Systems, 11(1), 85-94.

[107] Lazarus, R. S., and Folkman, S. (1984). Stress, appraisal, and coping, Springer publishing company.

[108] LeBlanc, V. R. (2009). "The Effects of Acute Stress on Performance: Implications for Health Professions Education." Academic Medicine, 84(10).

[109] Lee, G., Choi, B., Ahn, C. R., and Lee, S. (2020). "Wearable Biosensor and Hotspot Analysis-Based Framework to Detect Stress Hotspots for Advancing Elderly's Mobility." Journal of Management in Engineering, 36(3), 04020010.

[110] Lee, G., Choi, B., Jebelli, H., and Lee, S. (2021). "Assessment of construction workers' perceived risk using physiological data from wearable sensors: A machine learning approach." Journal of Building Engineering, 42, 102824.

[111] Lee, G., Jebelli, H., and Lee, S. (2020). "Online Multi-Task Learning and Wearable Biosensor-based Detection of Multiple Seniors' Stress in Daily Interaction with the Urban

Environment." The 8th International Conference on Construction Engineering and Project ManagementHong Kong.

[112] Lee, H., Lee, J., and Shin, M. (2019). "Using wearable ECG/PPG sensors for driver drowsiness detection based on distinguishable pattern of recurrence plots." Electronics, 8(2), 192.

[113] LeVan, P., Maclaren, J., Herbst, M., Sostheim, R., Zaitsev, M., and Hennig, J. (2013). "Ballistocardiographic artifact removal from simultaneous EEG-fMRI using an optical motion-tracking system." Neuroimage, 75, 1-11.

[114] Lewis, R. S., Weekes, N. Y., and Wang, T. H. (2007). "The effect of a naturalistic stressor on frontal EEG asymmetry, stress, and health." Biological Psychology, 75(3), 239-247.

[115] Li, J., Li, H., Umer, W., Wang, H., Xing, X., Zhao, S., and Hou, J. (2020). "Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology." Automation in Construction, 109, 103000.

[116] Li, J., Struzik, Z., Zhang, L., and Cichocki, A. (2015). "Feature learning from incomplete EEG with denoising autoencoder." Neurocomputing, 165, 23-31.

[117] LoBiondo-Wood, G., and Haber, J. (2014). "Reliability and validity." Nursing research-ebook: Methods and critical appraisal for evidencebased practice. Missouri: Elsevier Mosby, 289-309.

[118] Lockett, D., Willis, A., and Edwards, N. (2005). "Through seniors' eyes: an exploratory qualitative study to identify environmental barriers to and facilitators of walking." CJNR (Canadian Journal of Nursing Research), 37(3), 48-65.

[119] Lu, W., and Rajapakse, J. C. (2006). "ICA with reference." Neurocomputing, 69(16-18), 2244-2257.

[120] Luo, Q., Huang, X., and Glover, G. H. (2014). "Ballistocardiogram artifact removal with a reference layer and standard EEG cap." Journal of Neuroscience Methods, 233, 137-149.

[121] Ma, X., Jin, R., Sohn, K.-A., Paik, J.-Y., and Chung, T.-S. (2019). "An Adaptive Control Algorithm for Stable Training of Generative Adversarial Networks." IEEE Access, 7, 184103-184114.

[122] Mancino, M. (2017). "Design of an automated system for continuous monitoring of dairy cow behaviour in free-stall barns."

[123] Masood, K., and Alghamdi, M. A. (2019). "Modeling Mental Stress Using a Deep Learning Framework." IEEE Access, 7, 68446-68454.

[124] McCorry, L. K. (2007). "Physiology of the autonomic nervous system." American journal of pharmaceutical education, 71(4), 78.

[125] Michael, Y. L., Keast, E. M., Chaudhury, H., Day, K., Mahmood, A., and Sarte, A. F. (2009). "Revising the senior walking environmental assessment tool." Preventive medicine, 48(3), 247-249.

[126] Michel, C. M., and Koenig, T. (2018). "EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review." NeuroImage, 180, 577-593.

[127] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). Foundations of machine learning, MIT press.

[128] Molchanov, P., Gupta, S., Kim, K., and Pulli, K. "Multi-sensor system for driver's hand-gesture recognition." Proc., 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, 1-8.

[129] Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). "Prediction error estimation: a comparison of resampling methods." Bioinformatics, 21(15), 3301-3307.

[130] Mozetič, I., Torgo, L., Cerqueira, V., and Smailović, J. (2018). "How to evaluate sentiment classifiers for Twitter time-ordered data?" PloS one, 13(3).

[131] Murre, J. M. (2014). "Transfer of learning in backpropagation and in related neural network models." Connectionist Models of Memory and Language (PLE: Memory), 73.

[132] Naliboff, B. D., Goldstein, I. B., Shapiro, D., and Frank, H. J. (1988). "Mental and physical stress as moderators of the postural response in insulin-dependent diabetic patients." Health Psychology, 7(6), 499.

[133] Nath, R. K., Thapliyal, H., and Caban-Holt, A. "Validating Physiological Stress Detection Model Using Cortisol as Stress Bio Marker." Proc., 2020 IEEE International Conference on Consumer Electronics (ICCE), 1-5.

[134] Nezlek, J. B., Vansteelandt, K., Van Mechelen, I., and Kuppens, P. (2008). "Appraisal-emotion relationships in daily life." Emotion, 8(1), 145.

[135] Nguyen, H.-A. T., Musson, J., Li, F., Wang, W., Zhang, G., Xu, R., Richey, C., Schnell, T., McKenzie, F. D., and Li, J. (2012). "EOG artifact removal using a wavelet neural network." Neurocomputing, 97, 374-389.

[136] Noghabaei, M., Han, K., and Albert, A. (2021). "Feasibility Study to Identify Brain Activity and Eye-Tracking Features for Assessing Hazard Recognition Using Consumer-Grade Wearables in an Immersive Virtual Environment." Journal of Construction Engineering and Management, 147(9), 04021104.

[137] Nordin, A. D., Hairston, W. D., and Ferris, D. P. (2018). "Dual-electrode motion artifact cancellation for mobile electroencephalography." Journal of neural engineering, 15(5), 056024.

[138] Nordin, A. D., Hairston, W. D., and Ferris, D. P. (2019). "Human electrocortical dynamics while stepping over obstacles." Scientific Reports, 9(1), 4693.

[139] O'sullivan, D., and Unwin, D. (2014). Geographic information analysis, John Wiley & Sons.

[140] Oliveira, A. S., Schlink, B. R., Hairston, W. D., König, P., and Ferris, D. P. (2016). "Induction and separation of motion artifacts in EEG data using a mobile phantom head device." Journal of neural engineering, 13(3), 036014.

[141] Onikura, K., and Iramina, K. "Evaluation of a head movement artifact removal method for EEG considering real-time prosessing." Proc., 2015 8th Biomedical Engineering International Conference (BMEiCON), IEEE, 1-4.

[142] Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). "Domain adaptation via transfer component analysis." IEEE Transactions on Neural Networks, 22(2), 199-210.

[143] Pandey, M., Singh, V., and Vaishya, R. (2014). "Geomatics approach for assessment of respiratory disease mapping." The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 40(8), 205.

[144] Pandey, P., and Seeja, K. R. (2019). "Subject independent emotion recognition from EEG using VMD and deep learning." Journal of King Saud University - Computer and Information Sciences.

[145] Petersen, J. S., and Zwerling, C. (1998). "Comparison of health outcomes among older construction and blue-collar employees in the United States." American journal of industrial medicine, 34(3), 280-287.

[146] Picard, R. W., Vyzas, E., and Healey, J. (2001). "Toward machine emotional intelligence: analysis of affective physiological state." IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), 1175-1191.

[147] Pop, M. G., Crivii, C., and Opincariu, I. (2018). "Anatomy and function of the hypothalamus." Hypothalamus in health and diseases, IntechOpen.

[148] Posada-Quintero, H. F., Reljin, N., Moutran, A., Georgopalis, D., Lee, E. C.-H., Giersch, G. E., Casa, D. J., and Chon, K. H. (2020). "Mild dehydration identification using machine learning to assess autonomic responses to cognitive stress." Nutrients, 12(1), 42.

[149] Quandt, S. A., Wiggins, M. F., Chen, H., Bischoff, W. E., and Arcury, T. A. (2013). "Heat index in migrant farmworker housing: implications for rest and recovery from work-related heat stress." American Journal of Public Health, 103(8), e24-e26.

[150] Rezek, I., and Roberts, S. J. (1998). "Stochastic complexity measures for physiological signal analysis." IEEE Transactions on Biomedical Engineering, 45(9), 1186-1191.

[151] Rice, J. A., and Silverman, B. W. (1991). "Estimating the mean and covariance structure nonparametrically when the data are curves." Journal of the Royal Statistical Society: Series B (Methodological), 53(1), 233-243.

[152] Rosenberg, D., Kerr, J., Sallis, J. F., Patrick, K., Moore, D. J., and King, A. (2009). "Feasibility and outcomes of a multilevel place-based walking intervention for seniors: a pilot study." Health & place, 15(1), 173-179.

[153] Rosenberg, D. E., Huang, D. L., Simonovich, S. D., and Belza, B. (2012). "Outdoor built environment barriers and facilitators to activity among midlife and older adults with mobility disabilities." The Gerontologist, 53(2), 268-279.

[154] Rosenberg, D. E., Kerr, J., Sallis, J. F., Norman, G. J., Calfas, K., and Patrick, K. (2012). "Promoting walking among older adults living in retirement communities." Journal of aging and physical activity, 20(3), 379-394.

[155] Russell, J. A., Weiss, A., and Mendelsohn, G. A. (1989). "Affect grid: a single-item scale of pleasure and arousal." Journal of personality and social psychology, 57(3), 493.

[156] Russell, W. D. (1997). "On the current status of rated perceived exertion." Percept Mot Skills, 84(3 Pt 1), 799-808.

[157] Saeed, A., Ozcelebi, T., Lukkien, J., Erp, J. B. F. v., and Trajanovski, S. "Model Adaptation and Personalization for Physiological Stress Detection." Proc., 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 209-216.

[158] Saha, A., Konar, A., Chatterjee, A., Ralescu, A., and Nagar, A. K. (2014). "EEG analysis for olfactory perceptual-ability measurement using a recurrent neural classifier." IEEE Transactions on Human-Machine Systems, 44(6), 717-730.

[159] Saha, A., Rai, P., DaumÃ, H., and Venkatasubramanian, S. "Online learning of multiple tasks and their relationships." Proc., Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 643-651.

[160] Saha, S., Nesterets, Y. I., Tahtali, M., and Gureyev, T. E. (2015). "Evaluation of spatial resolution and noise sensitivity of sLORETA method for EEG source localization using low-density headsets." Biomedical Physics & Engineering Express, 1(4), 045206.

[161] Sakri, O., Godin, C., Vila, G., Labyt, E., Charbonnier, S., and Campagne, A. "A Multi-User Multi-Task Model for Stress Monitoring from Wearable Sensors." Proc., 2018 21st International Conference on Information Fusion (FUSION), IEEE, 761-766.

[162] Sawchenko, P. E., Li, H. Y., and Ericsson, A. (2000). "Chapter 6 - Circuits and mechanisms governing hypothalamic responses to stress: a tale of two paradigms." Progress in Brain Research, E. A. Mayer, and C. B. Saper, eds., Elsevier, 61-78.

[163] Schlotz, W. (2013). "Stress reactivity." Encyclopedia of behavioral medicine, Springer, 1891-1894.

[164] Schneider, R., Schmidt, S., Binder, M., Schäfer, F., and Walach, H. (2003). "Respiration-related artifacts in EDA recordings: introducing a standardized method to overcome multiple interpretations." Psychological reports, 93(3), 907-920.

[165] Seaman, D. E., and Powell, R. A. (1996). "An evaluation of the accuracy of kernel density estimators for home range analysis." Ecology, 77(7), 2075-2085.

[166] Seo, W., Kim, N., Kim, S., Lee, C., and Park, S.-M. (2019). "Deep ECG-Respiration Network (DeepER Net) for Recognizing Mental Stress." Sensors, 19(13), 3021.

[167] Seok, D., Lee, S., Kim, M., Cho, J., and Kim, C. (2021). "Motion artifact removal techniques for wearable EEG and PPG sensor systems." Frontiers in Electronics, 2, 685513.

[168] Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., and Ehlert, U. (2010). "Discriminating stress from cognitive load using a wearable EDA device." IEEE Transactions on information technology in biomedicine, 14(2), 410-417.

[169] Shackleton, G. L. (2021). "Towards a biochemical approach to occupational stress management." Heliyon, 7(6), e07175.

[170] Shahbakhti, M., Rodrigues, A. S., Augustyniak, P., Broniec-Wójcik, A., Sološenko, A., Beiramvand, M., and Marozas, V. (2021). "SWT-kurtosis based algorithm for elimination of electrical shift and linear trend from EEG signals." Biomedical Signal Processing and Control, 65, 102373.

[171] Shakerian, S., Habibnezhad, M., Ojha, A., Lee, G., Liu, Y., Jebelli, H., and Lee, S. (2021). "Assessing occupational risk of heat stress at construction: A worker-centric wearable sensor-based approach." Safety Science, 142, 105395.

[172] Sharma, G., and Goodwin, J. (2006). "Effect of aging on respiratory system physiology and immunology." Clinical interventions in aging, 1(3), 253.

[173] Shelley, K., and Shelley, S. (2001). "Pulse oximeter waveform: photoelectric plethysmography." Clinical Monitoring, Carol Lake, R. Hines, and C. Blitt, Eds.: WB Saunders Company, 420-428.

[174] Shi, Y., and Sha, F. (2012). "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation." arXiv preprint arXiv:1206.6438.

[175] Shoval, N., Schvimer, Y., and Tamir, M. (2018). "Real-time measurement of tourists' objective and subjective emotions in time and space." Journal of Travel Research, 57(1), 3-16.

[176] Shukla, J., Barreda-Ángeles, M., Oliver, J., and Puig, D. (2018). "Efficient wavelet-based artifact removal for electrodermal activity in real-world applications." Biomedical Signal Processing and Control, 42, 45-52.

[177] Shukla, S., Roy, V., and Prakash, A. "Wavelet Based Empirical Approach to Mitigate the Effect of Motion Artifacts from EEG Signal." Proc., 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 323-326.

[178] Singh, R. R., Conjeti, S., and Banerjee, R. (2014). "Assessment of driver stress from physiological signals collected under real-time semi-urban driving scenarios." International Journal of Computational Intelligence Systems, 7(5), 909-923.

[179] Slavich, G. M., and Shields, G. S. (2018). "Assessing Lifetime Stress Exposure Using the Stress and Adversity Inventory for Adults (Adult STRAIN): An Overview and Initial Validation." Psychosomatic medicine, 80(1), 17-27.

[180] So, J. H., Huang, C., Ge, M., Cai, G., Zhang, L., Lu, Y., and Mu, Y. (2017). "Intense Exercise Promotes Adult Hippocampal Neurogenesis But Not Spatial Discrimination." Frontiers in Cellular Neuroscience, 11(13).

[181] Spencer, C., Moore, D., McKeown, G., Rutherford, L., and Morrison, G. "Context matters: protocol ordering effects on physiological arousal and experienced stress during a simulated driving task." Proc., 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 1-7.

[182] Srivastava, M. (1984). "Estimation of interclass correlations in familial data." Biometrika, 71(1), 177-185.

[183] Stewart, C. L., Folarin, A., and Dobson, R. (2020). "Personalized acute stress classification from physiological signals with neural processes." arXiv preprint arXiv:2002.04176.

[184] Süzen, M., and Yegenoglu, A. (2019). "Generalised learning of time-series: Ornstein-Uhlenbeck processes." arXiv preprint arXiv:1910.09394.

[185] Story, M. F., Mueller, J. L., and Mace, R. L. (1998). "The universal design file: Designing for people of all ages and abilities."

[186] Sun, C., Hon, C. K., Way, K. A., Jimmieson, N. L., and Xia, B. (2022). "The relationship between psychosocial hazards and mental health in the construction industry: a meta-analysis." Safety science, 145, 105485.

[187] Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., and Griss, M. "Activity-aware mental stress detection using physiological sensors." Proc., International Conference on Mobile Computing, Applications, and Services, Springer, 282-301.

[188] Sweeney, K. T., Ayaz, H., Ward, T. E., Izzetoglu, M., McLoone, S. F., and Onaral, B. (2012). "A methodology for validating artifact removal techniques for physiological signals." IEEE transactions on information technology in biomedicine, 16(5), 918-926.

[189] Takeda, R., and Okazaki, K. (2018). "Body Temperature Regulation During Exercise and Hyperthermia in Diabetics." Diabetes and Its Complications, 89.

[190] Taylor, S., Jaques, N., Nosakhare, E., Sano, A., and Picard, R. (2020). "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health." IEEE Transactions on Affective Computing, 11(2), 200-213.

[191] Teplan, M. (2002). "Fundamentals of EEG measurement." Measurement science review, 2(2), 1-11.

[192] Tomaka, J., Blascovich, J., Kelsey, R. M., and Leitten, C. L. (1993). "Subjective, physiological, and behavioral effects of threat and challenge appraisal." Journal of personality and social psychology, 65(2), 248.

[193] Trakoolwilaiwan, T., Behboodi, B., Lee, J., Kim, K., and Choi, J.-W. (2017). "Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: three-class classification of rest, right-, and left-hand motor execution." Neurophotonics, 5(1), 011008.

[194] Ugnell, H., and Öberg, P. (1995). "The time-variable photoplethysmographic signal; dependence of the heart synchronous signal on wavelength and sample volume." Medical engineering & physics, 17(8), 571-578.

[195] Urigüen, J. A., and Garcia-Zapirain, B. (2015). "EEG artifact removal—state-of-the-art and guidelines." Journal of neural engineering, 12(3), 031001.

[196] Ursin, H., and Eriksen, H. R. (2004). "The cognitive activation theory of stress." Psychoneuroendocrinology, 29(5), 567-592.

[197] Venables, P. H., and Christie, M. J. (1980). "Electrodermal activity." Techniques in psychophysiology, 54(3).

[198] Wu, X., Li, X., Xiong, H., Zhang, X., Huang, S., and Dou, D. (2021). "Practical Assessment of Generalization Performance Robustness for Deep Networks via Contrastive Examples." arXiv preprint arXiv:2106.10653.

[199] Wusk, G. C., Abercromby, A. F., and Gabler, H. C. "Psychophysiological monitoring of aerospace crew state." Proc., Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, 404-407.

[200] Xie, Z., and Yan, J. (2013). "Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach." Journal of transport geography, 31, 64-71.

[201] Yadav, M., Behzadan, A., and Chaspari, T. "Speak Up! Studying the interplay of individual and contextual factors to physiological-based models of public speaking anxiety." Proc., 2019 First International Conference on Transdisciplinary AI (TransAI), IEEE, 52-55.

[202] Yadav, M., Chaspari, T., Kim, J., and Ahn, C. R. "Capturing and quantifying emotional distress in the built environment." Proc., Proceedings of the Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era, ACM, 9.

[203] Yang, K., Ahn, C. R., Vuran, M. C., and Kim, H. (2017). "Collective sensing of workers' gait patterns to identify fall hazards in construction." Automation in Construction, 82, 166-178.

[204] Zhang, K., Wang, J., Liu, T., Luo, Y., Loh, X. J., and Chen, X. (2021). "Machine Learning-Reinforced Noninvasive Biosensors for Healthcare." Advanced Healthcare Materials, n/a(n/a), 2100734.

[205] Zhang, X., Li, J., Liu, Y., Zhang, Z., Wang, Z., Luo, D., Zhou, X., Zhu, M., Salman, W., and Hu, G. (2017). "Design of a fatigue detection system for high-speed trains based on driver vigilance using a wireless wearable EEG." Sensors, 17(3), 486.

[206] Zhao, H., Zhang, S., Wu, G., Costeira, J. P., Moura, J. M., and Gordon, G. J. (2017). "Multiple source domain adaptation with adversarial training of neural networks." arXiv preprint arXiv:1705.09684.

[207] Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. "Adversarial multiple source domain adaptation." Proc., Advances in Neural Information Processing Systems, 8559-8570.

[208] Zhu, X., and Fujii, T. "A modulation classification method in cognitive radios system using stacked denoising sparse autoencoder." Proc., 2017 IEEE Radio and Wireless Symposium (RWS), IEEE, 218-220.

[209] Zontone, P., Affanni, A., Bernardini, R., Piras, A., and Rinaldo, R. "Stress detection through electrodermal activity (EDA) and electrocardiogram (ECG) analysis in car drivers." Proc., 2019 27th European Signal Processing Conference (EUSIPCO), IEEE, 1-5.

[210] Zontone, P., Affanni, A., Bernardini, R., Piras, A., Rinaldo, R., Formaggia, F., Minen, D., Minen, M., and Savorgnan, C. (2020). "Car Driver's Sympathetic Reaction Detection Through Electrodermal Activity and Electrocardiogram Measurements." IEEE Transactions on Biomedical Engineering, 67(12), 3413-3424.