Stress Detection in Lifelog Data for Improved Personalized Lifelog Retrieval System

Van-Tu Ninh, B.Sc.

A Dissertation submitted in fulfillment of the

requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

Faculty of Engineering and Computing, School of Computing

Supervisors

Assoc. Prof. Cathal Gurrin

Assoc. Prof. Sinéad Smyth

July 2023

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Sign: Student No.: 18214653 Date: 18/07/2023
(Van-Tu Ninh)

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Assoc. Prof. Cathal Gurrin and Assoc. Prof. Sinéad Smyth, for their invaluable guidance, expertise, and unwavering support throughout my Ph.D. journey. Their wisdom, encouragement, and patience have been instrumental in shaping this research and molding me into a better researcher.

I would like to express my deep appreciation and gratitude to Assoc. Prof. Minh-Triet Tran for his continuous support and encouragement throughout this challenging and rewarding endeavor. I am truly grateful for the opportunities provided to collaborate on research projects and for his insightful feedback, which has significantly enhanced the quality of my work.

In memory of my grandfather, Quy Ninh, your unwavering belief in my abilities and your gentle guidance shaped the person I have become today. This thesis stands as a testament to his enduring legacy and the profound impact he had on shaping my educational journey.

In memory of my dad, Thy Ninh, your unwavering support, encouragement, and belief in my abilities have been a driving force behind my academic pursuits. This achievement is a tribute to your enduring influence on my life.

To my mother, Lap Dang, your constant encouragement and belief in my abilities have been instrumental in shaping me into the person I am today. This accomplishment is as much yours as it is mine, and I dedicate this thesis to my loving mother with heartfelt gratitude.

To my fiancée, Tuyen Truong, your presence in my life has brought me immense joy and support. Your unwavering belief in me, patience, and understanding during challenging times have been a constant source of motivation. I am grateful for your love, encouragement, and for always standing by my side.

I would also like to extend my heartfelt thanks to my friends for their unwavering

support, friendship, and countless moments of laughter and inspiration. I want to acknowledge and express my gratitude to all the individuals who have contributed to my research in various ways, whether through their collaboration, insightful discussions, or assistance. Your contributions have enriched my work and shaped its outcomes.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

GSR	Galvanic Skin Response
EDA	Electrodermal Activity
SCR	Skin Conductance Response
SCL	Skin Conductance Level
BVP	Blood Volume Pulse
PPG	Photoplethysmography
ECG	Electrocardiogram
TEMP	Skin Temperature
CNS	Central Nervous System
PNS	Peripheral Nervous System
ANS	Autonomic Nervous System
SNS	Somatic Nervous System
SOTA	State-of-the-art
REC	Research Ethics Committee
LSC	Lifelog Search Challenge
VR	Virtual Reality
LMRT	Lifelog Moment Retrieval Task
NLP	Natural Language Processing
CNN	Convolutional Neural Network
DNN	Deep Neural Network
TF	Term Frequency
IDF	Inverse Document Frequency
LRT	Lifelog Retrieval Task

Abstract

Van-Tu Ninh

Stress Detection in Lifelog Data

for Improved Personalized Lifelog Retrieval System

Stress can be categorized into acute and chronic types, with acute stress having short-term positive effects in managing hazardous situations, while chronic stress can adversely impact mental health. In a biological context, stress elicits a physiological response indicative of the fight-or-flight mechanism, accompanied by measurable changes in physiological signals such as blood volume pulse (BVP), galvanic skin response (GSR), and skin temperature (TEMP). While clinical-grade devices have traditionally been used to measure these signals, recent advancements in sensor technology enable their capture using consumer-grade wearable devices, providing opportunities for research in acute stress detection. Despite these advancements, there has been limited focus on utilizing low-resolution data obtained from sensor technology for early stress detection and evaluating stress detection models under real-world conditions. Moreover, the potential of physiological signals to infer mental stress information remains largely unexplored in lifelog retrieval systems. This thesis addresses these gaps through empirical investigations and explores the potential of utilizing physiological signals for stress detection and their integration within the state-of-the-art (SOTA) lifelog retrieval system. The main contributions of this thesis are as follows. Firstly, statistical analyses are conducted to investigate the feasibility of using low-resolution data for stress detection and emphasize the superiority of subject-dependent models over subject-independent models, thereby proposing the optimal approach to training stress detection models with low-resolution data. Secondly, longitudinal stress lifelog data is collected to evaluate stress detection models in real-world settings. It is proposed that training lifelog models on physiological signals in real-world settings is crucial to avoid detection inaccuracies caused by differences between laboratory and free-living conditions. Finally, a state-of-the-art lifelog interactive retrieval system called LifeSeeker is developed, incorporating the stress-moment filter function. Experimental results demonstrate that integrating this function improves the overall performance of the system in both interactive and non-interactive modes. In summary, this thesis contributes to the understanding of stress detection applied in real-world settings and showcases the potential of integrating stress information for enhancing personalized lifelog retrieval system performance.

Chapter 1

Introduction

1.1 Lifelogging: Continuous Self-Tracking

The idea of creating an archive of a personal life experience and knowledge storage for later usage and sharing originates from “The Memex Concept” of Vannevar Bush in 1945 [13], which is later well-known as lifelogging. The idea of lifelogging has been popularized gradually in everyday discussion, and the acceptance of using technology as an augmented memory started to grow [14]. However, it was not until the release of an improved version of the MyLifeBits developed by Gemmell et al. in 2006 [15] that lifelogging began to become an active research topic. MyLifeBits was a research project of Microsoft Research Lab which was proposed and led by Gordon Bell in 2001 with the target of implementing ”The Memex Concept“ described in the essay ”As We May Think“ written by Bush in 1979. In particular, ”The Memex Concept“ proposed a new device in which individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility” [13]. The first version of the MyLifeBits system was released in 2002 in an attempt to digitize all possible data from the daily life of Gordon Bell, including web pages, telephone, radio, television, chat sessions, and a combination of media files in his personal computers [16]. However, due to the limitations of sensor and computing technologies at that time, only media content and activities could be digitized, which was not enough to capture all aspects of an individual’s life. Nowadays, with the rapid development of wearable sensors and ubiquitous computing, more types of data can be recorded to capture different aspects of an individual in daily life continuously and digitize them for storage instantaneously. This development facilitates researchers to conduct more research on the application in the lifelogging field owing to the availability of massive multi-modal data captured from multiple wearable sensor sources. Research in lifelogging applications can be expanded into other domains that result in the creation of many open interdisciplinary research directions. For instance, lifelogging devices can capture human activities during the day, facilitating epidemiological studies such as potential risks to children's health via the high frequency of non-core food advertisements that expose them over a period of time [17], health-related behavior analyses (e.g. obesity, diet, daily activities) [18, 19], exploring solutions for societal issues like privacy-related concerns [20], etc. As lifelogging research is emergent at the same speed as the development of wearable-device technology and ubiquitous computing, some new terminologies have also developed as follows:

- **Lifelogging**: According to Dodge and Kitchin [21], lifelogging is referred to as "a form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences, captured multi-modally through digital sensors and stored permanently as a personal multimedia archive" [22]. In short, it is a form of self-capturing common and social activities of an individual digitally during the day that results in the form of a lifelog data archive.

- **Lifelogger**: is the people who log as many moments, activities, and experiences in their life as possible. Gordon Bell is one of the typical lifeloggers who is commonly known as the first lifelogger pioneering in recording his life in digital form for decades [23].

- **Lifelog**: is the result of data gathering from the lifelogging activity. The lifelog data is actually multi-modal data that spans different types (e.g. vision, speech, bio-signals), formats (e.g. structured, semi-structured, or unstructured), and contexts (e.g. different environments, activities, people, or locations).

- **Quantified-Self** [1]: is defined as the practice of self-tracking, measuring, and quantifying all aspects of an individual in daily life using technology such as smartphones, wearable sensors, wearable health monitoring devices, activity trackers, etc. The quantified-self can be considered as a subset of lifelog which tries to digitize self-experience. However, it focuses more on providing analyses and insights of an individual based on the input data to either answer self-questions or explore new self-aspects through self-experience.

Though lifelog data is multimodal data that spans different types, formats, and contexts; it can be divided into six typical types of data, which are: *Vision*, *Hearing*, *Speech*, *Biometrics*, *Location*, and *Activities*. These six types of data are detailedly described as follows:

1. **Vision**: The visual data in conventional lifelog data archives are egocentric images or videos captured automatically by the wearable cameras mounted on either the head or neck of the user that show the user's view at a certain moment. Visual data is an important type of data in most lifelog data archives as it provides insights about an individual's life intuitively, such as the place/environment, the daytime, the people/objects that the user interacts with, the social interactions, etc.

2. **Hearing**: This type of data can be either soundtrack data that records what the user hears during the day or tabular data that summarises the music/sounds that the user listens to at a certain time as in the lifelog data provided in the Lifelog Search Challenge 2019 [2] and 2022 [3].

3. **Conversation**: The conversation data can be in either speech or text formats. The conversation in daily life can be the content of the messages, emails, texts,

[1] https://quantifiedself.com/

[2] http://lsc.dcu.ie/2019/data/index.html

[3] http://lsc.dcu.ie/

and the talks between the lifelogger and others during the day. It can be recorded in the video from the wearable cameras or can be logged the computer usage (e.g. keystroke log, on-screen text).

4. **Biometrics**: The biometrics data are usually recorded automatically from smart bands or smartwatches. Typical biometrics data includes heart rate, respiration rate, galvanic skin response (section 2.2.1.1), skin temperature (section 2.2.1.3), blood oxygen levels, and blood volume pulse (section 2.2.1.2) [24].

5. **Location**: The location data can be the coordinate (latitude and longitude) of the lifelogger acquired from the Global Positioning System (GPS) on his/her smartphones or from the smartwatches (e.g. Garmin watches). The coordinate data can be used to infer the address and the semantic name of the location, which is an important cue for most lifelog application research.

6. **Activities**: The activities data consists of the semantic activity labels such as sitting, standing, running, or lying, etc., estimated by the smart watches or smart bands from the data obtained from the accelerometer and gyroscope sensors integrated inside the devices themselves.

Apart from these data types that can be recorded from an individual by wearable sensors, other data such as browsing histories, read documents, and media files that the lifelogger interacts with on the computer in daily life can also be recorded in the lifelog data archive. However, due to privacy issues, only a few types of data have been released for research. Those six typical data types have been released in publicly available lifelog datasets with strict data anonymisation methods applied in order to ensure the private identity of the lifelogger. Among the aforementioned data types, visual data are mostly exploited in research with many potential applications. For instance, the lifelog images can be used to summarise a day of an individual in the ImageCLEF 2017 Lifelog Challenge [25], understand and identify the Activity of Daily Living (ADLs) in the ImageCLEF 2018 [26] and NTCIR-14 [27] Lifelog

Challenges. Apart from these examples, visual data in combination with other types of data, including hearing, conversation, biometrics, location, and activities, are utilized to develop lifelog retrieval system in the Lifelog Moment Retrieval Task (LMRT) in ImageCLEF Lifelog [25, 26, 28, 29] and the benchmarking Lifelog Search Challenge [30]. The research in the development of lifelog retrieval system is crucial to other interdisciplinary research due to the capability of the system to explore an individual's life through the first-person view (FPV) and its related data. One example is to find the number of moments that children are exposed to non-core food marketing, thereby, helping researchers to have an insight into the frequency of unsuitable advertisements exposed to children that can lead to a potential health problem for children if they consume those kinds of food [17]. Another application of the lifelog retrieval system is that it can be used as a prosthetic memory acting as a smart assistant for the human in the future [31]. Therefore, despite its recent emergence in lifelogging research, the research in the development of lifelog retrieval systems gain much interest from the research community all over the world.

During the process of joining the lifelog research domain and developing competitive state-of-the-art lifelog retrieval systems in the benchmarking Lifelog Search Challenge, I notice that most of the lifelog retrieval systems do not exploit all the available lifelog data. In detail, though six types of data are available in the lifelog dataset, only the visual data in combination with spatial-temporal data (location and time) and activities of the lifelogger are commonly used for retrieval. The hearing, conversation, and biometrics data are rarely exploited in these systems. Apart from the privacy concerns of releasing the conversation data, the hearing and biometrics data are not utilised effectively in these systems. Among these two data types, the biometrics data can be used to provide subjective stress indicators via the physiological signals (e.g. skin temperature, blood volume pulse, and galvanic skin response) [32]. The subjective stress indicators or any emotional-related information is a key factor in the memory-recalling process [33]. This implies that by exploiting this kind of data, the retrieval system can actually be improved to deal with the queries involving mental status or emotional-related information such as *"Find the moments that I was watching a video on Youtube to understand the underlying theory to complete my assignment. I was stressed as the content was hard to understand and the deadline was coming very soon."* This application can help improve the lifelog retrieval system to understand the stress indicators of the query, thereby equipping the existing systems with an empathetic ability to upgrade the retrieval power of the systems. This potential application is also the motivation for us to carry out this research.

1.2 A Brief Introduction to Stress & Stress Detection Challenges

The Yerkes-Dodson law, which is also known as the inverted-U model of arousal, is a model illustrating the relationship between stress and task performance. The theory has been proposed by two psychologists Robert Yerkes and John Dillingham Dodson from their experiment on mice, showing that everyone has a peak level of performance with an intermediate level of stress or arousal [34]. Too little or too much arousal or stress can lead to poor task performance. As can be seen from Fig. 1.1, the Yerkes-Dodson law can be illustrated as an inverted-U shaped curve whose left side of the curve demonstrates low arousal or stress while the right side represents high arousal indicating a poor task performance. The optimal state of arousal is in the middle section of the curve which implies the optimal performance an individual can achieve. However, from Fig. 1.1, we can recognize that although low arousal also results in poor task performance, we are more concerned about the high arousal section causing strong stress and anxiety that can damage both mental and physical health as well as leading to impaired performance.

Stress, in general, is defined as a "non-specific response of the body to any demand upon it" [9, 35]. In medical or biological contexts, stress is simply defined as the physical, mental, or emotional factors that cause bodily or mental tension

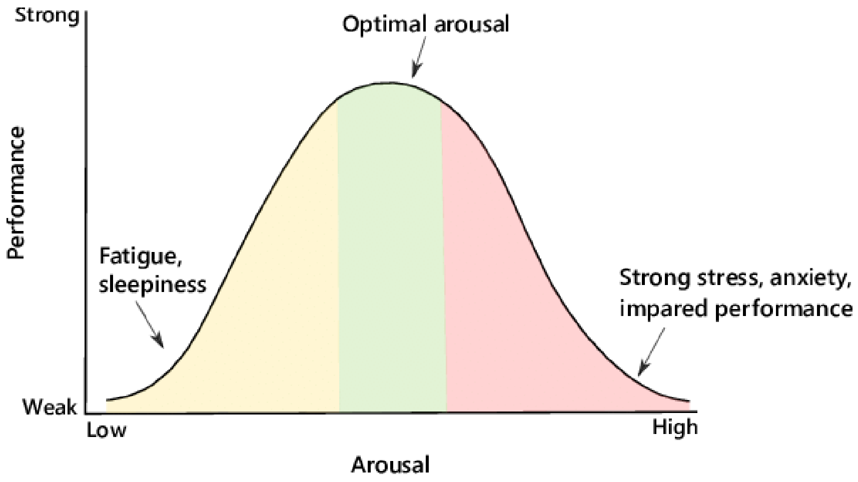


Figure 1.1: Illustration of the Yerkes-Dodson Laws [1].

[36]. Causes of stress also originate from many different sources such as a chemical or biological agent, environmental condition, external stimulus or any event that forces an organism to adapt to new conditions [37]. Stress can be categorized into three main types ordered ascending based on its damaging level: acute stress, episodic acute stress, and chronic stress [38]. Acute stress is the most common and least damaging type, which is experienced as an immediate perceived threat, either physical, emotional, or psychological. While acute stress can be perceived throughout the day, it can evolve into episodic acute stress if the bouts of acute stress occur frequently. Stress is beneficial for human beings as it helps them to recognize and prepare for upcoming potential dangers, thereby, increasing the concentration of an individual to accomplish a task or deal with the danger for a certain period of time. However, at a certain level, when the human nervous system cannot differentiate between emotional and physical danger, it begins to harm both the mental health and physical health of the human. This is the case of chronic stress, which traps a person in a negative situation filled with negative thoughts and worries and occurs repeatedly over a long period of time.

Identifying stress automatically is not a simple task. Although the power of using Machine Learning models in helping predict the chance of having cancer [39] and other diseases such as common flu, heart disease, kidney disease, etc. [40] has been certified; the application of these learning models in mental stress detection faces multiple challenges. The very first challenge when building a stress detection model is to choose appropriate data types to record and measure stress levels effectively. Two main approaches are traditionally used to quantify the effects of stress: questionnaires or surveys (mostly used in the field of psychology) and psychological sensors (commonly used by medical approaches and psychological research) [41]. The second challenge when building a stress detection model is the ability to record high-quality data without hindering individual daily activities [37]. For instance, Heart Rate Variability (HRV) has the highest quality when it is recorded by a chest-worn device. However, it is uncommon for a person to wear a chest-worn device for a long time during a typical day, but using a wrist-worn device might yield low-quality and unreliable data [37]. The third challenge is that the stress monitoring system should be personalized for each individual since different people have different physiological responses to stress according to the conclusion of Philip Schmidt et al. [42]. However, most stress detection models from the research focus much on the improvement of the subject-independent model as it is not a good approach to ask the user to gather enough stress data before using the stress detection models in real-life scenarios. Even so, it is important to evaluate and compare the performance of the personalized stress detection model (subject-dependent) with the general one (subject-independent) to determine a good approach to building the stress detection model, especially for low-resolution physiological signals from consumer-grade wearable data. These challenges are the motivation for us to pursue this research.

1.3 Research Challenges

Researching and developing an optimal stress detection model faces multiple challenges that mostly are related to the experiment design to gather stress data from participants and the stress data annotation process. Resolving these challenges can help researchers gather a reliable stress dataset to conduct experiments to evaluate the models' capability of detecting stress in practice either in the laboratory (constrained environment) or in the wild (unconstrained environment).

1.3.1 Stress Data Gathering Challenge

The main challenge when gathering stress data lies in the experiment design so that participants' stress responses can be triggered during the experiment. Indeed, designing the stress task is challenging since the task could be stressful for some participants but it is not that stressful for others. For instance, in the driving experiment conducted by Neska El Haouij et al. [10], the authors try to capture physiological stress responses from participants when driving through busy roads with high-load traffic or highways. However, it is not always the case, especially for experienced drivers or those whose main job relates to driving. Therefore, appropriate stress tasks should be used for suitable participants in the experiment. According to Sonia J. Lupien [43], a task should be designed to have at least one of four conditions to induce stress responses:

1. **Novel**: The task should be new to the participant, which requires confirmation from the participant that he/she has not ever done this task before.

2. **Unpredictable**: The participants should not be informed about the details of the task until the moment that they join the experiment. Thereby, the participants could not have enough time to prepare for the upcoming threat.

3. **Uncontrollable**: The task should put the participants under pressure (e.g. time pressure) so that the participants can not control their performance and behaviors in their normal state.

4. **Social Evaluation Threat**: The participants should have the feeling that their performance could be judged by others (either positively or negatively).

> The fear of being judged negatively can cause pronounced responses in the different stress systems [44].

The Trier Social Stress Test (TSST) [45] used in the benchmarking stress dataset named WESAD [9] in the laboratory environment is considered a typical example of a good stress task design as the task is social evaluative [46]. The version of the Trier Social Stress Test (TSST) in their experiment consists of public speaking and a mental arithmetic task. Indeed, during public speaking, the participants are told to deliver a five-minute speech in front of a three-person panel. The participants are told to impress the panel as the participants are convinced in the experiment that the three-person panel comes from the human resource department of their research faculty. This task design approach contains the social evaluation threat and the unpredictable factor (the three-person panel coming from the human resource department was not informed before the task) that forces the participant to focus on delivering the best performance on the task. In the experiment, I also focus on the uncontrollable factor and the social evaluation threat design of stress tasks for stress data gathering by asking the participant to try their best in a heavy-workload task under time pressure to achieve the best rank on a public scoreboard. However, the challenge still remains as some participants provide feedback that some tasks in the pilot experiment are not stressful enough for them. This requires us to use multiple self-evaluation methods to correctly choose the appropriate stress tasks for all participants.

1.3.2 Stress Data Annotation Challenge

Annotation of stress data is a real challenge due to the high subjectivity of the label and the continuous nature of the stress event. Indeed, though stress symptoms are defined clearly on most web pages and documents, the perception of stress between each individual is different and vague. The stress scales that I obtain from questionnaires and forms in the experiment might record the subjective evaluation of the participants, which is sometimes not reliable owing to two reasons. The first reason is that the participant might not have a clear perception of their stress status despite the existence of physiological stress responses, as they are not usually aware of the change in their mental status. The second reason is that the participant hide their mental status intentionally, resulting in wrong labels obtained after the stress data-gathering process. These two reasons frequently happen in both data gathering in the laboratory and in the wild. In addition to the vague perception of stress, the continuous nature of stress makes it hard for an individual to actually mark the beginning, the duration, and the end of stress events; especially when collecting stress lifelog data. It is hard to define what moments are stressful and what moments are not. For the annotation of stress data recorded in the constrained environment, two conventional approaches to provide a ground-truth of stress data are either using the study protocol label as the ground-truth (e.g. all the data in the stress session are assigned stress labels and vice versa [9]) or using the self-evaluation forms (e.g. State-Trait Anxiety Inventory, Positive and Negative Affect Schedule, Self-Assessment Manikins) and subjective stress scales [2, 10]. The main difference between these two approaches is that the study protocol labels provide an objective ground truth while the ones obtained from the self-evaluation forms are subjective. Though there are controversies about which approach is correct, in this research, both ways of annotating laboratory stress data are accepted so that all available stress datasets employing either one of the two methods of annotation can be used in the experiment. For the annotation of stress lifelog data, I try to overcome the challenge by using the event marking button on the wearable device in combination with the lifelog images as evidence of the stress events to support participants in their stress-moment annotation process. We insist that this annotation method should be considered approximately acceptable in order to conduct research and examine the true performance of stress detection applied in real life.

1.4 Hypothesis and Research Questions

From the observation that the biometrics data are not exploited in lifelog retrieval systems, I realize that valuable information can be gained from this type of data, such as emotion, mental status, etc. Indeed, from the literature review, the biometrics data in lifelog data archive contain physiological signals recorded from wearable devices, including galvanic skin response, heart rate, blood oxygen level, blood volume pulse, and skin temperature can be used to detect the current mental stress status of an individual precisely in constrained environments. Though evaluating mental stress detection models in unconstrained environments has not yet been validated, potential benefits can be gained by just having a stress detection model detect stressful moments precisely in real life. For instance, by integrating stress information into the lifelog retrieval system, early stress symptoms, and causes can be recognized from the analyses obtained from the personal lifelog retrieval system so that the user could intervene to prevent it from evolving into detrimental chronic stress and anxiety. In addition, I believe that the stress information could also improve the performance of the state-of-the-art lifelog retrieval system when dealing with stress-related/emotion-related queries such as *"Find the moments that I could not focus on my work as I argued with my friends that made me stress and angry at the same time"*. My conjecture is that the stress information can be used as a condition for filtering purposes which helps remove irrelevant results in the ranked list, thereby reducing the time that the user needs to navigate to the correct moment in the interactive lifelog retrieve system and increasing the number of relevant moments retrieved in the automatic mode. To validate my conjecture, I define the following hypothesis for this Ph.D. research as follows:

Hypothesis

It is possible to identify stress moments in lifelog data using the physiological signals captured from readily available lifelog sensors and enhance the performance of the state-of-the-art lifelog retrieval system with the stress-indexed information to address stress-related queries.

In order to either prove or disprove this hypothesis, a number of related research questions have been developed as follows:

- **Research Question 1 (RQ1). Evaluation of Stress Detection Models using Physiological Signals from Consumer-grade Wearable Devices.**

  How successfully can low-resolution physiological signals recorded from consumer-grade wearable devices, unlike traditional clinical devices with high-resolution ones, be used to detect the acute stress of an individual automatically by utilizing learning models?

  From both benchmarking and collected data, I evaluate the performance of the stress detection model trained on low-resolution physiological signals recorded from consumer-grade wearable devices compared to the one trained on high-resolution data captured from traditional clinical-grade devices. Then, I conduct experiments to propose an optimal approach to building a good stress detection model with low-resolution physiological data. To address research question 1, I propose to split it into two sub-research questions. Research question 1 can be addressed by providing answers to these two sub-research questions:

  Research Question 1.1: Can physiological signals recorded from consumer-grade wearable devices be used to develop a stress detection model for an individual?

  Research Question 1.2: Does the subject-dependent stress detection model achieve higher evaluation scores in detecting stress moments than the subject-independent stress detection model as used by the current generation consumer-grade wearable devices?

- **Research Question 2 (RQ2). Evaluation of Stress Detection Models Applied to Lifelog Data**

How successfully can stress detection models using low-resolution physiological signals from consumer-grade wearable devices be applied for lifelog data to detect moments of stress?

I conduct a proof-of-concept study to evaluate the performance of the stress detection model applied in real life. I collect stress lifelog data of three participants who joined the previous experiment in the laboratory environment. I capture all of their activities in daily life with multiple data privacy methods applied to ensure the private identity of the participants. From the collected dataset, I examine the performance of the lab-based stress detection model applied to lifelog data and explain the detection results of the model to understand how the lab-based model works in lifelog data. I then propose solutions to enhance the performance of the stress detection model applied in lifelog scenarios and discuss the limitation of my current approach based on the participants' feedback.

- **Research Question 3 (RQ3). Stress as a Facet of Lifelog Interactive Retrieval System.**

  How can biometric and visual data be used in a lifelog interactive retrieval system to retrieve stress-related moments?

  To either prove or disprove the hypothesis that stress information can actually enhance the performance of the state-of-the-art lifelog retrieval system, my team and I develop a state-of-the-art lifelog interactive retrieval system based on the knowledge about the core features of other systems acquired from the literature review. Finally, using the stress moments detected by the stress detection model, I evaluate and compare the performance of my lifelog interactive retrieval systems with and without the integration of the stress-moment filter function. I propose to split this research question into two sub-research questions. Research question 3 can be addressed by providing answers to these two sub-research questions:

Research Question 3.1: How can the state-of-the-art lifelog interactive retrieval system be designed and developed?

Research Question 3.2: How much benefit can be derived from adding the biometric stress filters to an interactive lifelog retrieval system in a conventional retrieval task?

1.5 Research Contributions

In this section, the key contributions made in this thesis are outlined as follows:

- **Chapter 4 – RQ1:**
  - **Contribution 1:** Based on the experimental results in Section 4.3, I proved that the stress detection model using low-resolution physiological signals recorded from wearable devices used as training data performs as well as the one trained with high-resolution data recorded from traditional clinical devices.
  - **Contribution 2:** Based on the experimental results conducted in Section 4.4, I proved that the subject-dependent model is more accurate in stress detection than the subject-independent one when trained on low-resolution physiological signals' features, which implies that the subject-dependent model is the most optimal approach to training stress detection model.
- **Chapter 5 – RQ2:**
  - **Contribution 3:** I collected longitudinal stress lifelog data (described in Section 5.2.1) and conducted a proof-of-concept study to evaluate the performance of the stress detection model applied to lifelog data.
  - **Contribution 4:** According to the experimental results and the insights gained from feature analyses and model explanation in Section

    5.3.1 and Section 5.3.2, I showed that applying the laboratory stress detection model to predict stressful moments in daily life would lead to inaccurate detection results.

  - **Contribution 5:** From the experimental results in Section 5.3.3, I proposed that the lifelog stress detection model should be trained on physiological signals recorded in unconstrained conditions instead of in the constrained one.

- **Chapter 6 – RQ3:**

  - **Contribution 6:** I developed (with colleagues) a lifelog interactive retrieval system named **LifeSeeker** (Section 6.2.1) and evaluated its performance through multiple annual benchmarking Lifelog Search Challenges. The results from the challenges proved that **LifeSeeker** is one of the state-of-the-art lifelog retrieval systems.

  - **Contribution 7:** Based on the experimental results in Section 6.3, I proved that when integrating a stressful-moment filter into the state-of-the-art lifelog retrieval system, the overall performance of the system increases in both interactive and non-interactive mode.

1.6 Thesis Outline

This thesis is mainly focused on developing a stress detection model in both constrained and unconstrained environments as well as proving that much benefit can be gained by using stress indicators in lifelog retrieval systems. The structure of this thesis is demonstrated in Fig. 1.2. In this chapter, I introduced the motivations, challenges, and contributions of this research. Additionally, I formed the hypothesis based on the literature review, proposed the research questions, and summarised the research contributions. The remainder of this thesis is organized as follows:

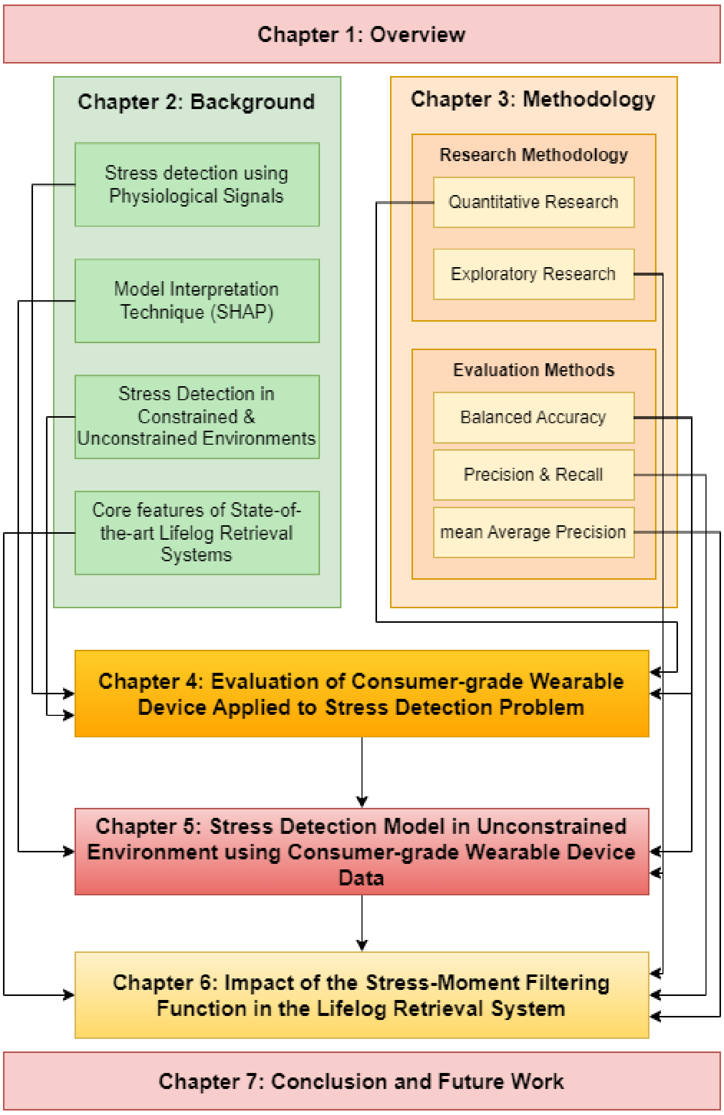

Chapter 1: Overview
Chapter 2: Background
Stress detection using Physiological Signals
Model Interpretation Technique (SHAP)
Stress Detection in Constrained & Unconstrained Environments
Core features of State-of-the-art Lifelog Retrieval Systems
Chapter 3: Methodology
Research Methodology
Quantitative Research
Exploratory Research
Evaluation Methods
Balanced Accuracy
Precision & Recall
mean Average Precision
Chapter 4: Evaluation of Consumer-grade Wearable Device Applied to Stress Detection Problem
Chapter 5: Stress Detection Model in Unconstrained Environment using Consumer-grade Wearable Device Data
Chapter 6: Impact of the Stress-Moment Filtering Function in the Lifelog Retrieval System
Chapter 7: Conclusion and Future Work

Figure 1.2: The structure of the thesis.

- Chapter 2 presents the underlying background of stress detection using physiological signals based on the anatomy knowledge of the human brain and the nervous system. I also discuss the outlines of currently existing work that relates to lifelog moment retrieval tasks. I discuss the conventional information retrieval methodology, review the lifelog benchmarking and the retrieval system from participants, and explain the difference with my system.

- Chapter 3 presents the research methodology and evaluation methods to address three research questions in my research.

- Chapter 4 presents my evaluation of the statistical difference between the performance of various learning models trained on the data recorded from either a clinical device or a consumer-grade wearable device. In this chapter, I compare two conventional approaches to training human-related learning models, which are the subject-dependent and the subject-independent, to select the best approach to building a stress detection model with consumer-grade device data.

- Chapter 5 presents my proof-of-concept study to develop a lifelog stress detection model with physiological data captured by consumer-grade wearable devices in the lifelog data archive. Further analyses and discussions are also presented to provide insights into how the stress detection model works in real life and potential approaches are proposed to enhance the performance of lifelog stress detection model in future work.

- Chapter 6 presents my work in developing the state-of-the-art lifelog retrieval system and the assessment of the performance of the state-of-the-art lifelog retrieval system with the integration of the stress-moment filter function in a conventional retrieval task.

- Chapter 7 presents a summarization of my work in the thesis. Limitations of my research and future work are also presented in this chapter.

Chapter 2

Related Work and Background

2.1 Introduction

This chapter serves as the foundation for our contributions to the fields explored in this thesis. It begins by presenting the background theory concerning the physiological responses of the body to stressors. Subsequently, the detection of stress using these physiological signals is introduced in Section 2.2. Additionally, Section 2.3 provides the background on SHAP value (SHapley Additive exPlanations), which is utilized to explain the detection decisions made by the lifelog stress detection model in Chapter 5. Furthermore, a literature review is conducted on the most recent advancements in stress detection using three benchmarking datasets: AffectiveROAD [10], WESAD [9], and CognitiveLOAD [47]. The selection of stress detection models for our experiments in Chapter 4 and Chapter 5 is based on the insights gained from these literature studies, as presented in Section 2.4.1 and Section 2.4.2. Lastly, Section 2.4.3 presents a comprehensive review of the significant and innovative features identified in the top-3 state-of-the-art interactive lifelog retrieval systems from previous benchmarking Lifelog Search Challenges (LSC).

2.2 Stress Detection using Physiological Signals

To establish the foundation for stress detection using physiological signals, I begin by explaining the mechanism of how the human body responds to stressors and the physiological signals that are triggered by these stressors in Section 2.2.1. These physiological signals comprise; galvanic skin response (GSR), blood volume pulse (BVP), and skin temperature (TEMP). A detailed explanation of how each physiological signal responds to stressors can be found in Section 2.2.1.1, 2.2.1.2, and 2.2.1.3.

2.2.1 The Theory of Physiological Responses Elicited by Stressors

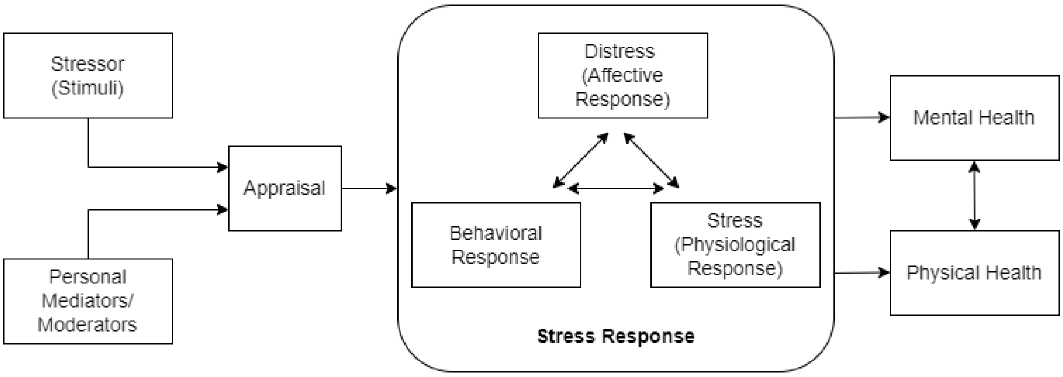

Stressor (Stimuli)
Personal Mediators/ Moderators
Appraisal
Distress (Affective Response)
Behavioral Response
Stress (Physiological Response)
Stress Response
Mental Health
Physical Health

Figure 2.1: An Overview of the Stress Process [2]

As defined by Gillian H. Ice in his book [48], stress is the process of eliciting emotional, behavioral, and/or physiological responses caused by a stimulus conditioned by an individual's personal, biological, and cultural context. Martin Gjoreski illustrated this stress process in his research that is similar to Fig. 2.1 [2]. As depicted in Fig. 2.1, when an individual is stimulated by the stressor, his/her body responds to the stimuli instinctively by his/her personal moderators through an autonomic system that results in a combination of three responses: affective response, behavioral response, and physiological response. These responses affect both individuals' mental and physical health mutually.

In the general biological context of the human nervous system, it is divided into two main parts: the Central Nervous System (CNS) and the Peripheral Nervous System (PNS). While the central nervous system is made up of the brain and spinal cords, the PNS consists of nerves branching off from the spinal cord of the CNS that extend to all parts of the body as illustrated in Fig. 2.2 [49]. These two nervous systems communicate with each other via nerve impulses to execute the commands from the brain to control the body's response to the external environment either voluntarily or involuntarily. Especially, the PNS is considered the most important part of this structure due to its main functions of moderating both conscious and unconscious bodily behaviors [49].

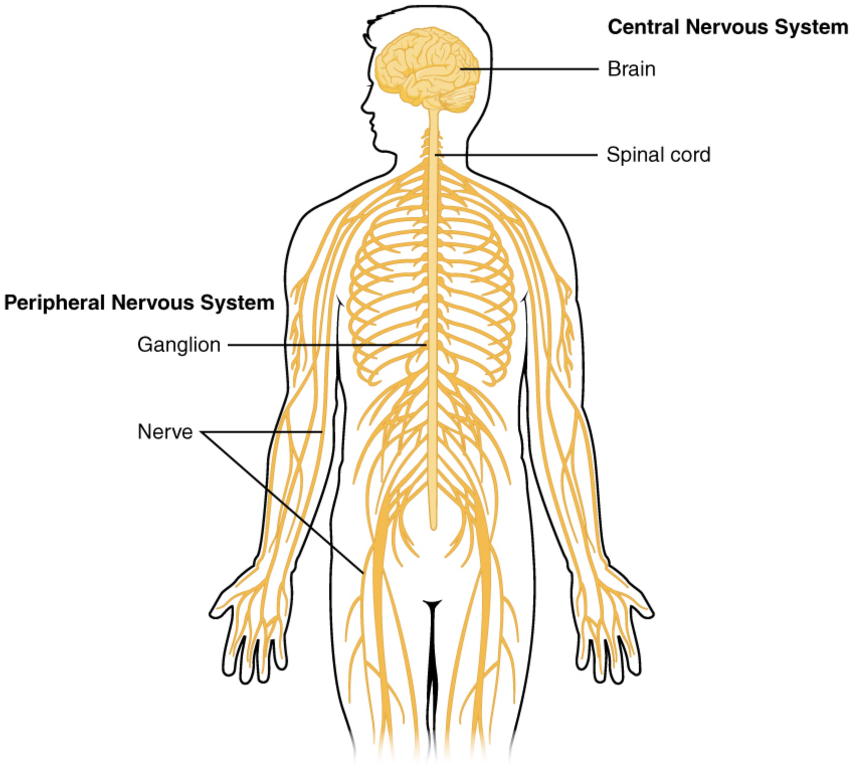


Figure 2.2: An Overview Structure of the Human Nervous System [3].

The PNS is divided into two components including the Autonomic Nervous System (ANS) and the Somatic Nervous System (SNS), which plays the role of regulating the unconscious and conscious bodily behaviors respectively. In terms of consciousness, the unconscious/involuntary responses of the body are the natural reaction of the body to the stimuli without the perception of an individual to control it while the conscious/voluntary ones can be controlled via either the inhibition of skeletal muscles or the encouragement of the behaviors response through the awareness of the individual to the situation. The conscious response of the body (e.g. skeletal movements, reflexes to situations, and external stimuli from the environment) is responsible by the SNS while the involuntary response is moderated by the ANS.

The ANS, which is divided into sympathetic and parasympathetic divisions, plays an active part in regulating the body's immediate reaction to stress exposure [50] without the person's conscious effort. In "fight-or-flight" situations, these two divisions contribute significantly to two different processes, which are known as arousal and recovery, to prepare the body to react with the corresponding stimuli of the situation [49]. Specifically, in the arousal state, the sympathetic nervous system is activated to optimize body function to prepare for the upcoming threat/pressure while the parasympathetic nervous system alters the body functions to help it recover [49]. During the arousal state, the sympathetic nervous system prepares the body for an upcoming threat by increasing heart rate, blood pressure [50, 51], pupil size [52], the level of cortisol [53], stimulating sweat gland secretion [50, 51, 54], expanding the respiratory rate [51], etc. In addition, digestion and urinary activities are inhibited during this time. In contrast to the arousal state (stress state), the parasympathetic nervous system in the ANS would regulate the body functions to their normal states by slowing the heart activity, lowering cortisol levels, inhibiting the sweat gland activity, and increasing the digestive and urinary activity.

All the aforementioned physiological responses can be recorded using clinical-grade devices. However, some of them can only be measured by invasive techniques such as cortisol levels, glucose rate, adrenal amount, etc.; which is not ideal for collecting the data of a normal user in non-clinical experiments. Physiological signals, including heart activity, sweat gland activity, skin temperature, and respiratory rate, can be recorded using non-invasive techniques. Therefore, these signals are widely recorded on non-patient participants in a laboratory experiment to provide data for constructing an automatic stress detection method. These signals can be easily captured in a laboratory environment using clinical-grade devices (e.g. RespiBAN, BioHarness 3 Zephyr) that are tools/sensors with wires and electrodes attached to a part of the body recording high-resolution data. However, thanks to the development of sensors and wearable devices, these physiological signals are now able to be captured in daily life under free-living conditions using consumer-grade wearable devices (e.g. Garmin, Fitbit, Empatica E4), thereby facilitating more research on stress detection in different conditions and environments using more advanced techniques like learning models.

2.2.1.1 Electrodemal Activity (Galvanic Skin Response)

Electrodermal Activity (EDA), to which can be historically referred as Galvanic Skin Response (GSR), Skin Conductance (SC), or Skin Response (SR), is the variation of the electrical characteristics of the skin in response to sweat secretion [55]. Under emotional arousal and stress in response to the context of the environment and events, the eccrine sweat gland activity increases in corresponding with the change in the emotional and stress response [56, 57]. It is worth noting that this activity change is involuntarily elicited by the Autonomic Nervous System (ANS), which means that one could not control this response intentionally. In terms of the method of the EDA signal recording, the variation in the sweat gland activity can be detected by measuring the resistance of the electrical signal between two electrodes applied to the skin (fingers, palms of hands or feet) [56, 58, 59]. Some emblematic devices that can be used to record the EDA signal are ProComp Infiniti, Biopac MP150, Shimmer 3 GSR+, and Empatica E4 wristband [60], which are ranked in the ascending order of portability of the instruments themselves as well as the capability of employing them to record data in daily life.

The EDA signal is the additive result from its two main components: the Skin Conductance Level (SCL) and the Skin Conductance Response (SCR) [57]. For a certain period of time, the Skin Conductance Level (SCL) or the tonic component of the EDA fluctuates slowly while the Skin Conductance Response (SCR) or the phasic component fluctuates faster. The tonic component (SCL) can be considered as the movement baseline of the signal since it does not contain peaks, while the phasic component (SCR) varies in response to the events or stimulus, which is named as Event-Related Skin Conductance Response (ER-SCR). Therefore, the phasic component is usually employed to detect arousal events [56]. However, the phasic component (SCR) sometimes varies without any events, which is known as Non-Specific Skin Conductance Response (NS-SCR). Some statistical features can be possibly extracted to detect stress of an individual, such as the number of SCR peaks, the mean amplitude, the total SCR rise time (sum of duration between the peaks and their corresponding onsets), recovery time, etc. These features were shown to be effective in the stress detection problem in the literature [61].

2.2.1.2 Blood Volume Pulse and Electrocardiogram

Electrocardiogram (ECG) is the measure of the heart's electrical activity via the electrodes placed on the skin at different parts of the body such as arms, legs, and chest [56]. ECG is usually recorded using clinical-grade devices used in laboratory environments, such as the Biopacs MP150, MP35, and Shimmer Sensing 3 [60]. Theoretically, heart activity is considered one of the most important signals for stress detection as it is affected directly by the Autonomic Nervous System [56]. Indeed, one can experience the feeling of fast-beating, fluttering or pounding heart under stressful events. The reason for this feeling is that the body automatically increases the concentration of the individual via the increase of oxygen and energy to the heart, blood flow, and the dilation of the coronary blood vessels to prepare for a fight or flight [62]. In terms of the structure of the ECG, a typical heartbeat is composed of four main components: the baseline, the P wave, the QRS complex, and the T wave [60]

From these components, many statistical features can be extracted to detect stress, such as the mean and standard deviation of the heart rate, Heart Rate Variability (HRV), etc. Heart Rate Variability analysis is the most prominent process that contributes many seminal features for stress detection [63], which extracts features from the NN-intervals. Specifically, the NN-interval refers to the time distance between normal R-peaks (IBI – Inter-beat Interval) in ECG signal since artifacts may arise due to faulty sensors or arrhythmic events [64]. For instance, the value of the RR-interval (IBI) lying outside the range of 300 milliseconds and 2000 milliseconds is considered abnormal since its conversion to heart rate estimation would be either more than 200 beats per minute or less than 30 beats per minute, which is not a valid value for the human heart rate. From the HRV analysis, four types of features are extracted, including time-domain features, frequency-domain features, geometrical-domain features, and non-linear-domain features. The list of the features in each domain can be found in `https://github.com/Aura-healthcare/hrv-analysis`. Each feature in the list has a specific meaning in a stressful context. For instance, in a stressful event in a window size of 60 seconds, it is expected that there would not be many NN-intervals that are longer than 50 milliseconds compared to when the body is at rest or relaxed cause the ANS is expected to increase the heart activity.

Thanks to the development of sensors attached to the wearable device, heart activity can now be captured not only via the ECG signal in a laboratory environment but also can be recorded using Photoplethysmography (PPG) from smart watches and consumer-grade wearable devices. Photoplethysmography (PPG) is a low-cost optical non-invasive technique using a near-infrared light source to measure blood volume pulse [65], which is the variations of skin hue associated with concurrent changes in blood volume in subcutaneous blood vessels during the cardiac cycle [4]. Some popular consumer-grade wearable devices employing this technique are Fitbit, Garmin, Empatica E3 and E4 wristbands.

As depicted in Fig. 2.3, the PPG signal is composed of two phases: the systolic and diastolic phases. The systolic phase (or "rise time") starts with a valley and ends with the pulse wave systolic peak [65] while the diastolic phase is marked at the place where the pulse wave end after following another valley [66]. The RR intervals (Inter-beat intervals – IBI) can be approximated by the Systolic Peak-to-

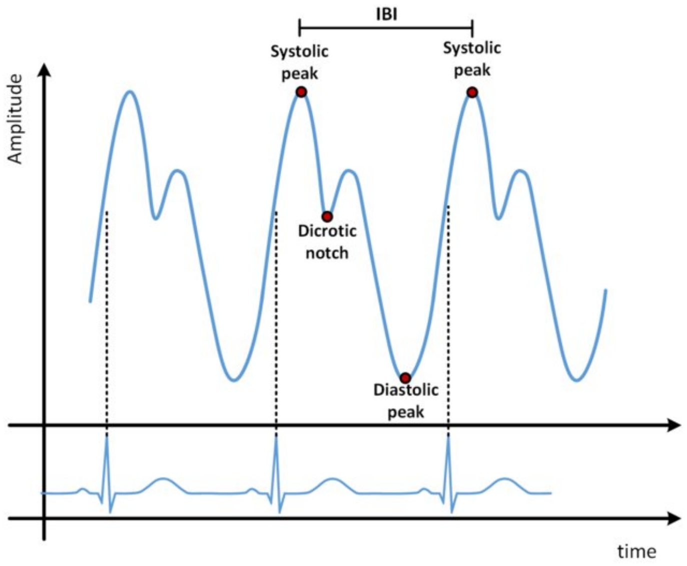


Figure 2.3: The structure of the PPG signal and its relationship to the Inter-beat Interval in the ECG signal [4].

Peak intervals (PP intervals). Therefore, the same method of Heart Rate Variability analysis and feature extraction process employed for the ECG signal can be used for the PPG signal.

2.2.1.3 Skin Temperature

Skin temperature can be measured non-evasively using infrared thermography (IRT) [67]. One well-known consumer-grade wearable device that implements this technology to keep track of personal affective states is the Empatica E4 wristband.

Apart from skin response and heart activity, skin temperature also changes during stress, which can be considered as one of the seminal cues for detecting moments of stress [56]. The skin temperature values can vary between 33 to 35 degrees Celsius [68]. However, under stressful or emotional conditions, the skin temperature can vary either lower or higher than the normal range of each personal baseline skin temperature. In detail, the research from Cornelia Kappeler-Setz showed that in the arousal state, the skin temperature can change from 0.1 to 0.2 Celsius degree [69]. Theoretically, under stress conditions, the activation of the sympathetic nervous system leads to the reduction of the peripheral circulation that results in the reduction in skin temperature [70]. However, the skin temperature does not always decrease under stressful events, which causes an ambiguity concern on the impact of using skin temperature in stress detection problems [71]. For example, Arturas Kaklauskas et al. confirmed that skin temperature rose in the presence of stress [72] while other studies from other research showed that the skin temperature decreased in stressful events [73, 74, 75, 76]. Palanisamy Karthikeyan et al. also observed that the mean skin temperature gradually increased in most subjects during stress state and provided a clue that the variation of the skin temperature under different affective states could depend on the skin property of the race [77]. Despite the uncertainty in stress pattern recognition when using skin temperature, Palanisamy Karthikeyan et al. supported the idea that combining the statistical features of skin temperature like mean skin temperature might enhance the power of stress detection system [77].

2.3 Model Interpretation Technique

Feature contribution is a crucial factor to understand which features are most important to the detection. As the decisions of the models rely on the input features of the data, the contribution of the features would provide insights into the model's decisions. From the interpretation of the model, the underlying rules of the features that the model learns to perform detection can be inferred. The relative feature importance of different features can also be inferred. Utilizing feature importance allows us to obtain insights into the reasons behind the model's performance, whether it is successful or unsuccessful, on any given input data. Thereby, analyses of the distribution of the features can be made to compare the difference between the training and testing data so that further model improvement, data transformation, and data pre-processing can be proposed to enhance the performance of the model. In this section, I present a widely used technique to interpret learning models, which is called SHAP (SHapley Additive exPlanations). The foundation knowledge of the SHAP techniques, which is the Shapley Value, is also presented in this section.

2.3.0.1 Shapley Value

Sharply value is a local explanation technique originating from the cooperative game theory [78] that aims to explain individual detection of the black-box model. Though it is a local explanation technique, the global explanation can be gained by aggregating all of the these individual detection. In general, the main idea of the Shapley value is to compute the contribution of the features that the model uses to detect the results. This is done by considering each feature's marginal contribution in all of its coalitions (subsets of feature combinations containing the targeted feature) that are used to train the model and yield detections. In detail, the marginal contribution of a feature to the detection of the model could be computed by the weighted sum over all possible feature value coalition using the following Shapley value formula:

$$\phi_j(v) = \sum_{S \subseteq \{1 \ldots p\} \backslash \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} (v(S \cup \{j\}) - v(S)) \tag{2.1}$$

In the equation 2.1, S is the subset of features used in the model that does not contain the j-th feature, p is the number of features, $v_(S)$ is the detection for feature values in set S that are marginalized over features not included in set S. The $v_(S)$, in a straightforward representation, could be illustrated as the following formula:

$$v(S) = \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} \tag{2.2}$$

The Shapley value satisfies four properties that can be considered a definition of a fair payout. These four properties are as follows:

1. **Efficiency**: The feature contributions must add up to the difference of detection for x and the average.

$$\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X)) \tag{2.3}$$

2. **Symmetry**: If two feature values j and k contribute equally to all possible subsets of features, the feature contribution of these two feature values must be the same.

$$v(S \cup \{j\}) = v(S \cup \{k\})\ \forall S \subseteq \{1, \ldots, p\} \backslash \{j, k\} \Rightarrow \phi_j = \phi_k \tag{2.4}$$

3. **Dummy**: If the feature j does not change the detection results regardless of the feature combinations, its feature contribution — Shapley value must be equal to 0.

$$v(S \cup \{j\}) = v(S)\ \forall S \subseteq \{1, \ldots, p\} \Rightarrow \phi_j = 0 \tag{2.5}$$

4. **Additivity**: The respective Shapley value of the additive of two features values $val + val^{+}$ is $\phi_j + \phi_j^{+}$

While the main advantage of the Shapley value is that the difference between the detection and the average detection, based on the Efficiency property, is distributed fairly among the feature values; the main disadvantage of this explainable model approach is the high computational cost as all possible coalitions of a feature needs to be taken into account. To consider all the number of subsets containing the feature, the marginal contribution needs to be computed for 2^{n-1} coalitions for n is the number of features in S. An exact Shapley value is computationally expensive to obtain. Therefore, in my work, I only use the approximation solution to compute Shapley values to explain my model's detection, which is discussed in section 2.3.0.3 and section 2.3.0.4.

2.3.0.2 SHAP (SHapley Additive exPlanations)

SHAP was introduced by Lundberg and Lee [79], which represents the Shapley value as an additive feature attribution method that focuses on the local method to explain a detection of a model f based on the locally approximating explanation model g:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2.6}$$

where $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size, and $\phi_j \in \mathbb{R}$ is the Shapley values of feature j.

In addition to fulfilling the four properties inherited from the Shapley values, SHAP also possesses three distinct properties that are advantageous for additive feature attribution methods, which are as follows:

1. **Local Accuracy**: The output of the explanation model g should match the one of the original model $\hat{f}$ when $x = h_x(x')$ where h_x is the mapping function from the simplified input feature x' to the original input.

$$\hat{f}(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j \tag{2.7}$$

2. **Missingness**: A feature value x'_j missing from the coalition vector has the value of 0 and has an arbitrary Shapley value theoretically. However, as it does not hurt the local accuracy property, SHAP enforces its Shapley value to be 0: $x'_j = 0 \Rightarrow \phi_j = 0$.

3. **Consistency**: If the marginal contribution of a feature value varies (increases or stays the same) due to the change of the model, the Shapley values also vary (increase or stay the same) in corresponding with the marginal contribution.

$$\hat{f}'_x(z') - \hat{f}'_x(z'_{/j}) \geq \hat{f}_x(z') - \hat{f}_x(z'_{/j}) \; \forall z' \in \{0,1\}^M \Rightarrow \phi_j(\hat{f}', x) \geq \phi_j(\hat{f}, x) \tag{2.8}$$

where $\hat{f}_x(z') = \hat{f}(h_x(z'))$ and h_x is the mapping function from the coalition vector to the original feature vector.

With a detailed explanation of the variables in SHAP, I can re-define $v(S)$ specific for SHAP as follows:

$$v(S) = \hat{f}_x(z') = \hat{f}(h_x(z')) \tag{2.9}$$

The $\hat{f}(h_x(z'))$ is the conditional expectation function of the original model. This conditional expectation function is designed to develop different strategies to deal with the missing field in the variable z' (the field where $z'_i = 0$). Depending on the type of the model, the corresponding type of $\hat{f}(h_x(z'))$ is used. As in my experiments and analyses in Section 5.3.2, I use an ensemble tree-based model and logistic regression, I focus on using TreeSHAP and LinearSHAP to analyze why the models can work or cannot work on the real-life dataset.

2.3.0.3 LinearSHAP

To interpret the feature contribution and detection explanation of the Logistic Regression model in my experiment, I propose to use LinearSHAP. Linear explanation is valid to use for Logistic Regression as target probability is positively correlated with the features. This implies that by increasing a feature by one point, the target probability is also increased by a certain amount assuming all other features remain the same. In the linear model detection, each individual effect can be computed easily. Considering a linear model for one data instance where x_j is the feature value with $j = 1 \dots p$ and β_j is the its corresponding weight:

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2.10}$$

The contribution ϕ_j of the j-th feature on the detection $\hat{f}(x)$ is:

$$\phi_j(\hat{f}) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j) = \beta_j(x_j - E(X_j)) \tag{2.11}$$

where $E(\beta_j X_j)$ is the mean effect estimate for feature j. The contribution is the difference between the feature effect and the average effect.

2.3.0.4 TreeSHAP

TreeSHAP is a variant of SHAP for tree-based machine learning models proposed by Lundberg et al. [80]. The difference between TreeSHAP and conventional SHAP is the approximation of the value function using the conditional expectation $\hat{f}_x(z') = E_{z_{\bar{S}}|z_S}(\hat{f}(z'))$ instead of the marginal expectation, where $E_{z_{\bar{S}}|z_S}(\hat{f}(z'))$ is computed as in the Algorithm 1 in [80] according to Lundberg et al.

As Lundberg et al. proposed a fast approach to compute the Shapley values in the tree-based model in [80], I use their implementation in my experiment to explain the tree-based model. However, as I use the ensemble tree-based model, a further aggregation step is required to compute the final Shapley value computation for each feature as illustrated in equation 2.12

$$\phi_j = \frac{1}{T}\sum_{i=1}^{T}\phi_j^{(i)} \tag{2.12}$$

2.3.0.5 SHAP Feature Imporance

In a simple explanation, features with large absolute Shapley values are important Therefore, the SHAP feature importance that I use in my experiment to analyze the impact of each feature on the detection results based on the Shapley values is the average absolute Shapley values per feature across the data. Its computation follows the below equation:

$$I_j = \frac{1}{n}\sum_{i=1}^{n}|\phi_j^{(i)}| \tag{2.13}$$

2.4 Related Work

This section presents the relevant work on stress detection in both constrained and unconstrained environments. Section 2.4.1 focuses on related work of stress detection in constrained environments that concentrates on the three benchmarking stress datasets used in many publications in the same research field, while Section 2.4.2 discusses all recent relevant publications of stress detection in unconstrained environments. These discussions provide support for the main research objective of developing an optimal stress detection model applied to lifelog data, as presented in Chapter 4 and Chapter 5. Furthermore, the development of core features in the top-3 state-of-the-art interactive lifelog retrieval systems in previous benchmarking Lifelog Search Challenges that propose significantly novel features is explored to support the research in developing a state-of-the-art lifelog retrieval system. This system will be used in experiments to evaluate the benefits of utilizing detected stress-moment information during the retrieval process, as discussed in Chapter 6.

2.4.1 Stress Detection in Constrained Environment

The context of stress detection in a constrained environment is broad. From my point of view, stress detection in a constrained environment is the one that requires the participants to do a number of pre-defined tasks in a specific environment over a period of time. Therefore, some data collection protocols in previous related work that are claimed to be real-life scenarios are considered experiments in constrained environments in my work.

2.4.1.1 AffectiveROAD Dataset

Neska El Haouij et al. created a dataset named AffectiveROAD [10] using the same experimental setup as of Healy and Picard [81]. However, the AffectiveROAD dataset consisted of low-resolution physiological signals recorded from consumer-grade wearable devices instead of high-resolution data, which was also the focus of their research on the application of using consumer-grade wearable devices' data in constrained real-life scenarios. Daniel Lopez-Martinez et al. proposed a novel personalized machine learning which employed a multi-view multi-task machine learning framework in stress detection [82]. They evaluated the performance of their proposed approach on the AffectiveROAD dataset and achieved the mean accuracy of 83% for binary stress classification. In detail, the best model using a multi-view, multi-task machine learning framework in their work comprised two views and three tasks, which were the number of signals used for learning (Electrodermal Activity and Heart Rate) and the number of normalized spectral clusters, respectively. For the training step, the mean feature vector of each drive in the whole training set (including driving sessions of multiple participants) was computed and clustered into T profiles. The result of this process was that the drives were clustered into T different profiles that shared similar physiological responses to driving-induced affective states. The instances of the drives in each cluster were used to train the personalized stress classifier for that targeted cluster only. They claimed that their proposed approach could "account for inter-subject inter-drive variability in affective responses to the driving experience" [82]. Their proposed multi-view multi-task learning model framework is novel, however, depends much on the hyper-parameters T and could not be extended easily when a new drive is added to the stress corpus due to the need to re-construct the profile.

2.4.1.2 WESAD Dataset

Philip Schmidt et al. [9] created a benchmarking stress dataset in the laboratory, named WESAD dataset, by inducing stress responses from participants in a simulated interview-like situation named Trier Social Stress Test (TSST) [45]. In detail, the participants were exposed to a three-person panel while wearing both a chest-worn clinical-grade device (RespiBan) and a wrist-worn consumer-grade wearable device (Empatica E4 wristband) at the same time to record the physiological responses. Based on the dataset that Philip Schmidt et al. collected, they tried to build and evaluate multiple subject-independent stress detection machine learning models in an instantaneous manner as the step of the window shift that the authors chose was 0.25 seconds. They extracted statistical features from physiological signals, including Electrodermal Activity (EDA), Blood Volume Pulse (BVP), and Skin Temperature (TEMP) with a window size of 60 seconds. Five conventional machine learning models, including Decision Tree, Random Forest, AdaBoost, Linear Discriminant Analysis, and k-Nearest Neighbours were trained in a subject-independent approach, which employed the data of other participants for training and testing on the remaining participant data that was not in the training set. The best subject-independent stress detection model that the authors managed to achieve the mean accuracy of 88.33% (±0.25) [9] was the Random Forest trained on all wrist-worn physio data. Using the same dataset, Kizito Nkurikiyeyezu el al. investigated the difference between the performance of subject-independent and subject-dependent stress detection models trained on chest-worn device's data (high-resolution physio signals) [41]. The results from their experiment indicated that the subject-dependent model outperformed the subject-independent ones in discriminating the stress and non-stress moments, which implied that the inter-subject variability of physiological responses affected the performance of the model. The authors proposed a hybrid calibrated model which incorporated a few personal physiological samples (approximately 50% of personal data) in the generic pool of physiological samples collected from a large group in the training process to mitigate the performance gap between the subject-dependent and subject-independent models. Though the hybrid calibrated model could improve the performance of the subject-independent model, the training approach was not much different from the subject-dependent training approach since 50% of personal physiological samples were used. For the evaluation of the performance of stress detection models using wrist-worn device's data (low-resolution signal), Pekka Siirtola proposed to evaluate the performance of subject-independent stress detection models using the same features as in the preliminary work of Philip Schmidt et al. but employing another evaluation metric – balanced accuracy – to evaluate the performance of the model on imbalanced datasets. Additionally, the author also investigated the effect of the window size on the performance of the subject-independent stress detection model [83]. From their experiment, the author observed that longer window sizes used for feature extraction could lead to a slight increase in the balance accuracy score of the stress detection model. The best model using Linear Discriminant Analysis (LDA) trained on three signals which include Skin Temperature (TEMP), Blood Volume Pulse, and Heart Rate, with a window size of 120 seconds achieved the highest average balanced accuracy score of 87.4% (±10.4). This result is high for a subject-independent stress detection model. However, they suggested that the significant variation in recognition accuracy between study subjects can be alleviated by building subject-dependent stress detection instead. In 2021, Lam Huynh et al. proposed a novel subject-independent stress detection model using Neural Architecture Search (NAS) and evaluated the model on the benchmarking WESAD dataset [5]. The inputs fed into the deep neural network that the authors employed were the filter bank of each physiological signal from a wrist-worn device (EDA, BVP, and TEMP), which was "a quadratic form of signal in the joint time-frequency domain" [5], and mixed features. Instead of constructing the deep neural network (DNN) manually, the authors generated a set of DNNs candidates for each input modality from the search space using the procedure proposed by Xuanyi Dong et al. [84]. The authors randomly searched from 10000 architectures and only used the best ten architectures that had the highest covariance matrices of the gradient for training. An overview of the proposed network is illustrated in Fig. 2.4. By using such a complex architectural design, the authors improved the accuracy of the stress detection model significantly. Their best model using all physio data from wrist-worn devices achieved an accuracy of 92.87%, which was higher than the baseline Random Forest model up by 4.54%. The only drawback of such a complex deep network was the training time, which was approximately 50 hours of training time using Tesla-V100 [5].

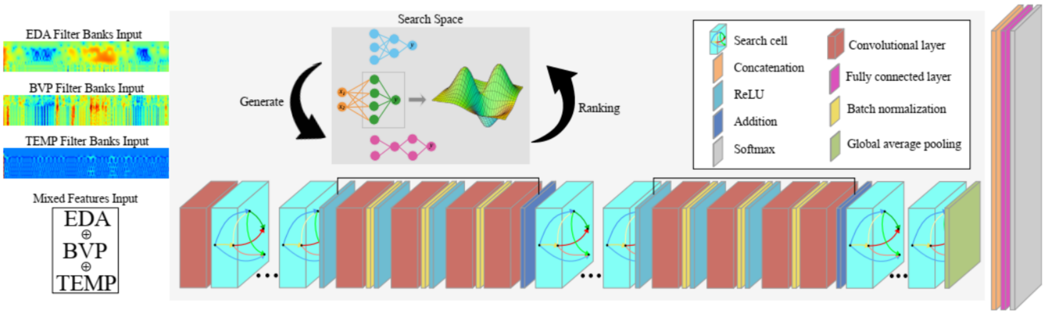


Figure 2.4: The architecture of the Stress Neural Architecture Search [5]

2.4.1.3 CognitiveLOAD Dataset

Another stress dataset capturing low-resolution physiological signals was collected from 21 participants by Martin Gjoreski et al. in the laboratory environment [47]. The participants were exposed to stressors that comprise performing different mental arithmetic tasks at different difficulty levels under time pressure with high rewards. In detail, the stress data were collected following a stress protocol proposed by Dedovic et al [85], which was designed to have three arithmetic tasks with increasing levels of difficulty: easy, medium, and hard. After each task, a false ranking score was shown to the participant to create a competitive stimulus to the participant. The baseline (non-stress) data were recorded on a different day when the participants were relaxed to ensure the consistency of the data. Martin Gjoreski et al. conducted two experiments on this dataset which includes evaluating the effect of window size and the effect of different feature combinations on the performance of the subject-independent stress detection model [47]. The evaluation was performed on nine machine learning models, including Support Vector Machine (SVM), Random Forest (RF), Boosting, Bagging, k-Nearest Neighbors (kNN), Naive Bayes, Decision Tree, Ensemble Selection, and Majority classifier. For the first experiment, the authors employed the window size starting from 30 seconds increasing up to 360 seconds (6 minutes) with the window shift depending on the size of the sliding window (window size - 25 seconds). The authors also drew the same conclusion as Pekka Siirtola [83] that the longer the window size is, the higher accuracy the model can achieve.

Among the classifiers, the SVM model performs well for all window-size, therefore, it was chosen for the second experiment. For the second experiment, the authors compared the performance of the SVM model using a subset of features selected on a sensor-specific base. The proposed feature selection process consisted of two main steps, which were removing correlated and non-informative features. Specifically, non-informative features are the ones that have low information gain as they do not carry many characteristics of the class [86]. In addition, correlated features can deteriorate the model's performance as they contribute less to the decision process and can contain noises. The result from their experiment suggested that using the feature combination from all sensor sources can increase the accuracy of the stress detection model significantly compared to using one sensor modality. The authors achieved an accuracy of approximately 73% for a 3-class problem ("No Stress" v.s "Low Stress" v.s "High Stress") and confirmed that the model can be easily confused "Low stress" with "No stress" and "High stress" as it was difficult to define a strict border between these kinds of events [47].

2.4.2 Stress Detection in Unconstrained Environment

2.4.2.1 Machine Learning Approaches

Martin Gjoreski et al. also collected real-life stress data using a wrist-worn wearable device (Empatica E4) and kept track of their stress status via a combination of a stress log and Ecological Momentary Assessment (EMA) prompting on a smartphone [47]. In case of a stress event, the participants needed to log their stress level on a scale from 1 to 5 (1 to 2 – no stress, 3 to 5 – stress) as well as logged the start and the duration of the stressful situation. The stress level was used to label the logged stressful moments in real-life data to consider if it was actually a stressful event or not. As the stress label was subjectively annotated due to the time lag in the perception of stress, the authors proposed two remedies corresponding to two ways of stress-event annotation scenarios. For the first scenario, at time X, the participant responded to the EMA prompts and confirmed that he/she was suffering from stress, the period of stress label was then extended by 10 minutes before and after the time X since the physiological arousal might start before that certain point of time. For the second scenario, the participant labeled a stress event interval from time Y to time Z. The authors also extended 10 minutes before time Y and after time Z to mitigate the stress perception bias. The rest of the data was considered no-stress and was split into events with a duration of 10 minutes. Using this real-life stress dataset, Martin Gjoreski et al. conducted two experiments to find the optimal approach to building stress detection for in-the-wild data. The first experiment was the aggregation experiment which evaluates the effect of the size of the sliding window used for feature extraction of physiological signals in real-life on the performance of the stress detection model in the wild. From the aggregation experiment, the authors can draw the conclusion that the smaller the aggregation window is, the better performance the model can achieve. Specifically, the decrease of the window size from 17.5 to 10 minutes increased the mean F1-score of the subject-independent stress detection model (Decision Tree) from 84% to 90%. For the second experiment, the authors utilized the best configuration of the stress detection model from the first experiment (Decision Tree with a 10-minute window size of feature extraction) to compare the context-based and a no-context approach applied to stress detection in the wild. The context defined by the authors indicated the consideration of embedding physical activities (lying/sitting/standing/walking/running/cycling) of the participant at a certain time into the training feature while no context implied the direct use of the lab-based stress detection model into the real-life data. The context was an important feature for in-the-wild stress detection problems since physical activity also elicited similar physiological arousal to the one in psychological stress events [47]. To do so, the authors gathered raw acceleration data from 10 healthy volunteers during a 120-minute experiment with a medical expert [87]. The labels of the data included five activities: lying, sitting, standing, walking, and running/cycling. The average activity level was inferred from the activity recognition model trained on their collected datasets and was passed as a feature to the context-based stress classifier [47]. Indeed, from their context-versus-no-context experiment of 10-minute event segments, the context-based classifier provided more accurate stress-event detection than lab-based stress detection one applied to the wild directly. In detail, the context-based classifier achieved a mean F1-score of 90% while the no-context one only provided an F1-score of 47%. The precision and recall of the context-based classifier were 95% and 70% respectively, which indicated that 70% of stress events were detected at a precision of 95%.

2.4.2.2 Deep Learning Approaches

Han Yu et al. investigated the performance of using the Deep Learning model with modality-fusion self-attention mechanism (MFN) [88] applied to stress data in the wild. The real-life stress data collected by the authors contains multimodal physiological data (ECG and GSR sampled at 256 Hz) from consumer-grade wearable devices (IMEC and Belgium) recorded from 41 participants over eight days. In detail, the authors employed the self-attention mechanism in the Transformer model [89] to extract feature representation for the sequential physiological data of the 60-minute window size. The main idea of their approach was to input each physiological signal sequentially into a self-attention network independently to extract the embedding vector of the signal and concatenate these embeddings for training. Thereby, both the learning models can learn not only the information from each sensor source separately but also learn the correlation of both modalities through the fusion embedding. In addition, to mitigate the individual differences in stress dataset caused by differences in human behaviors in real life, the authors proposed to add personalized attention to combine the generalized information of different users with the individual patterns of the targeted subject, which was the main idea of their proposed Modality Fusion Network (MFN). The performance of the MFN network was then compared statistically with the original Transformer merely applied to the stress data. Three different possible scenarios of employing physiological signals (ECG, GSR, and

ECG & GSR) were also considered. According to their experiment, the F1-score of the MFN model was statistically higher than the one that merely applies the Self-Attention Network in all three scenarios at the significance level of 0.01 [88]. Among the three different scenarios, the MFN model utilizing the fusion of both ECG and GSR signals into the training process achieved the highest mean F1-score of 69.3% (±0.6%) The F1-score of the MFN model can also be improved to 77.4% (±0.7%) if the personalized attention layers were added to the model structure.

Han Yu et al. continued to investigate the performance of deep learning models applied to momentary stress detection in the wild by proposing a novel semi-supervised learning method to augment data for training [90]. The idea of their approach was to solve the problem of lacking labels for the training model as real-life labeled data in the wild only occupies a small amount of the stress data due to the conventional stress status response via feedback at a certain point of time [90]. Indeed, based on the analysis of three datasets (SMILE [91] – $n = 45$ for 390 days, TILES [92] – $n = 212$ for 10 weeks, and CrossCheck [93] – $n = 75$ for a year) on which the authors work, the ratio of labeled sequences versus unlabeled sequences of 5-minute window-size is smaller than 0.8%. To utilize most of the unlabeled data, the authors proposed to use the Long Short Term Memory (LSTM) model as an Auto Encoder (AE) to encode each sequence data into a single feature vector in the latent space. The main idea of their proposal was to first pre-train the LSTM-AE model on labeled data, then clustered the labeled samples in latent space using the Gaussian Mixture Model (GMM) into K components based on the Akaike and Bayesian information criterion analysis. Finally, the latent representations of all the unlabeled samples were inferred using the pre-trained model and assigned to the most similar distribution of labeled samples based on the negative log-likelihood values. The selected unlabeled samples were assigned labels based on the most similar samples in the cluster and were used to continue to train the LSTM-AE model. Moreover, the authors also applied the consistency regularization training methods in order to reduce the noise created by the augmented data when the model makes a detection. Using the default configuration of the sequence length of 30 and window shift of 20 minutes for feature extraction based on the logging rate of the stress-logged application, their LSTM-AE model with data augmentation and consistency regularization training achieved the mean F1-score of 66% (±1%), 70% (±1%), and 64% (±2%) on SMILE, TILES, and CrossCheck dataset respectively; which was higher than their baseline LSTM model up to 0.8%.

2.4.3 Lifelog Retrieval System

Since the emergence of the lifelog task challenge at NTCIR-12 [94] with the release of large datasets of lifelog data, many lifelog problems have been introduced to research teams with international participation. This has led to lifelogging becoming a thriving research field. In NTCIR-12 [94], two problems were proposed to explore personal lifelog data which are Lifelog Semantic Access Task (LSAT) and Lifelog Insight Task (LIT). Among the two tasks, the LSAT, which was later known as the Lifelog Moment Retrieval task (LMRT), aimed to develop a methodology and search engine to retrieve the lifelog moments based on the description of the lifelogger (e.g. Find the time when I was looking at an old clock, with flowers visible.). This task has been the core task in many lifelogging research workshops including NTCIR-13 [95], NTCIR-14 [27], four editions of ImageCLEFlifelog from 2017 to 2020 [25, 26, 28, 29], and Lifelog Search Challenge from 2018 to 2020 [30] and attracted many researchers to develop novel remedies to develop state-of-the-art interactive lifelog retrieval system and optimize its functions. Through these lifelogging challenges, many interactive lifelog explorers were proposed with many enhancements. Among these challenges, the Lifelog Search Challenge (LSC) is the benchmarking competitive challenge that requires participants to develop novel features in their interactive lifelog retrieval systems to win. These novel features can be divided into three general categories:

- **Data Indexing**: As the speed of the retrieval algorithm/retrieval engine depends on the way the data is indexed in the database, this section of the lifelog retrieval system needs to be implemented carefully. Any enhancement of this section can lead to the improvement of the retrieval speed of the system.

- **Retrieval Algorithm**: Retrieval algorithm/retrieval engine is the soul of the lifelog retrieval system which reflects 80% of the retrieval power of the system. The improvement of the retrieval algorithm/retrieval engine can boost the precision of the lifelog retrieval system significantly.

- **User-Interface & User-Interaction**: Well-designed interfaces and novel system interactions between the user and the retrieval system can increase the browsing experience of the user, thereby facilitating the user to utilize all the functions of the system and navigating to the desired moments efficiently. It is also more important when the interactive lifelog retrieval system is developed to be used in different kinds of devices (e.g. mobile phones, web applications, and virtual reality environments).

2.4.3.1 Lifelog Search Challenge 2018 (LSC'18)

lifeXplore System [6]

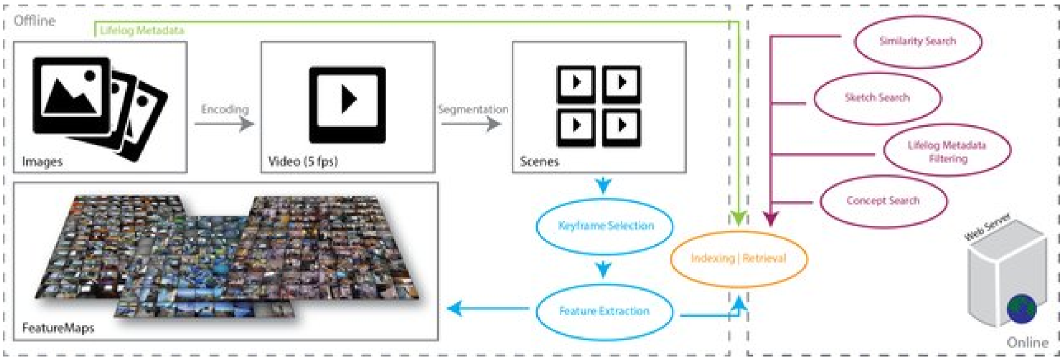


Figure 2.5: The general architecture of lifeXplore interactive retrieval system [6].

- **Data Indexing**: The authors proposed a solution to convert a video retrieval system into an interactive lifelog retrieval system by encoding the sequence of lifelog images of a day into a video with a constant frame rate (5 frames per second). The video was then inputted through a custom shot detection algorithm to group relevant semantic scenes as sub-videos before extracting representative keyframes, thereby reducing the number of items needed to show in the user interface as well as boosting the speed of the retrieval engine. Other metadata such as day of the week, time range, location, etc. were indexed in the database for filter purposes.

- **Retrieval Algorithm**: For retrieval methods, the authors proposed four different methods to search including sketch search, similarity search, concept search, and metadata filtering [6].Both sketch search and similarity search utilized k-nearest neighbor to search for image descriptors that are similar to the query one. Depending on what the user wants to search, different kinds of descriptors are used for different purposes such as HistMap descriptors for searching scenes with similar color distribution, GoogLeNet features for searching scenes with similar semantic content, etc.

- **User-Interface & User-Interaction**: For the design of the user interface, the authors followed the idea of Barthel et al. [96, 97] that utilized a feature map to show the retrieval results efficiently. The feature map displayed the selected keyframes of the generated lifelog videos arranged based on a given similarity criterion such as visual similarity, feature similarity, semantic similarities, etc. An illustration of the feature map interface with their data processing pipeline is shown in Fig. 2.5.

VRLE System [7]

- **Retrieval Algorithm**: For retrieval algorithms of the system, the authors only used the concepts/tags provided by the organizers including the day, time, day of the week, and visual objects for filtering and keywork searching.

- **Data Indexing**: All the metadata such as day of the week, time range,

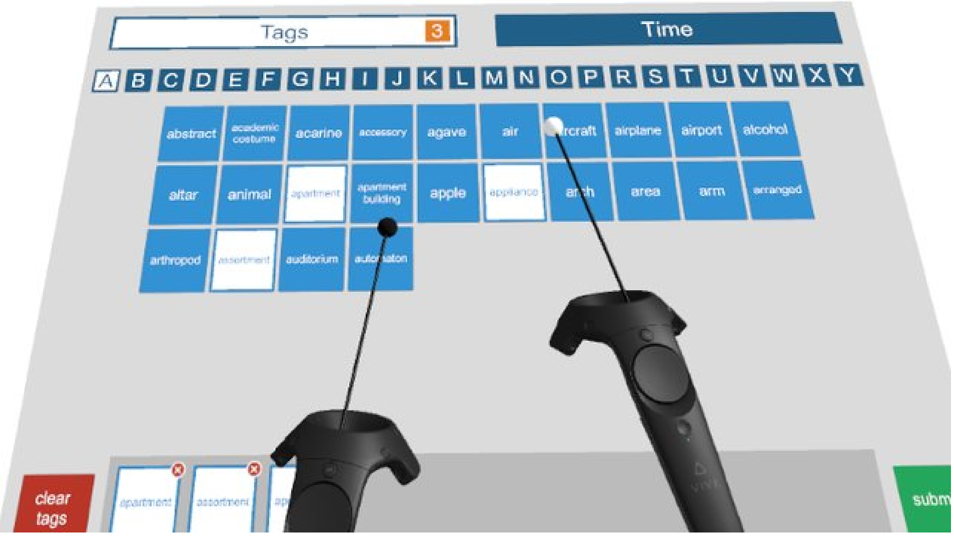


Figure 2.6: The contact-based use-interaction with the retrieval user interface of VRLE lifelog retrieval system in LSC'18 [7].

location, etc. were indexed in a structured database for filtering purpose and keyword searching.

- **User-Interface & User-Interaction**: Duane et al. develop the very-first novel lifelog interactive retrieval system in a virtual reality environment with the proposal of an efficient design of the user interface and multiple user interaction methods [7]. The user interface for the retrieval process is also simple with two menus for tags selection and time selection as can be seen from Fig. 2.6. In total, the authors propose two new user interaction mechanisms in their system: distance-based user interaction and contact-based user interaction. The distance-based approach was functionally similar to using a television remote, which interacted with the retrieval user interface via an interactive beam originating at the top of the user's wireless controller in the virtual environment. The user then only needed to press a button on the controller to select tags and time ranges based on the query description. The same design was used for contact-based user interaction, however, in a much more direct form of interaction like physically touching the interface components instead of remote control interaction simulation. As illustrated in Fig. 2.6, the controllers were designed as drumsticks protruding from the head of each controller [7] in the virtual environment. These drumstick-like controllers imitated the conventional style of browsing on computers or smartphones by interacting with a keyboard or a touchscreen by fingers. Users can use both hands to generate queries in parallel, thereby facilitating dexterous users to use the retrieval system in the virtual environment as fast and precisely as using conventional web-based or mobile interactive retrieval systems.

2.4.3.2 Lifelog Search Challenge (LSC'19)

VIRET [98] & vitrivr [99] Systems

The conversion of interactive video retrieval systems from the Video Browser Showdown (VBS) challenge into interactive lifelog retrieval systems was still effective in the LSC'19. Indeed, two of the best lifelog retrieval systems in the LSC'19, VIRET [98] and vitrivr [99], inherited most features from their precedent video search systems. In LSC'19, most of the state-of-the-art lifelog retrieval systems focused on introducing new query methods and enriching metadata. Specifically, for the VIRET lifelog retrieval system, Jakub Lokoc et al. introduced three different types of query methods in their system including query by color, query by text, and query by images with a filter function in the system to re-select the best-matched images [98]. The vitrivr system also introduced similar query methods with the extension of the query-by-sketch for semantically visual search [99]. For query-by-text methods, all of the systems still formed multiple boolean queries with multiple combinations of AND/OR operators between the tags and visual concepts of lifelog images parsed from the query description manually. For example, the vitrivr system structured its boolean query by splitting it into two different parts: query terms and query containers [99]. The query container composed of multiple query terms which were visual concepts or textual tags connected by a logical AND operator in the late fusion process. Finally, multiple query containers were connected by the logical OR operator. Other metadata such as location, time, and biometrics were used for filtering purpose and boolean query by most of the state-of-the-art systems.

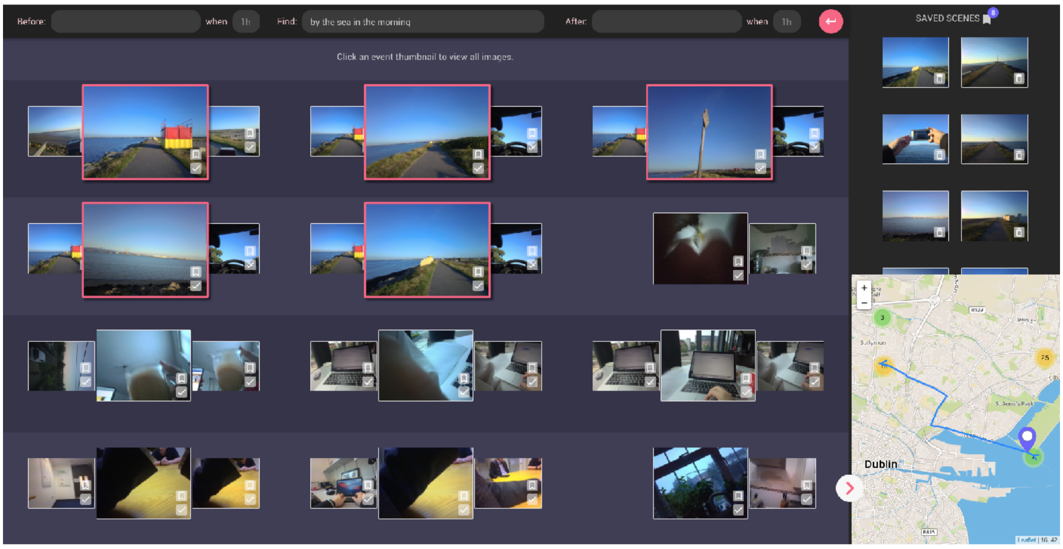
Smart Lifelog Retrieval System with Habit-based Concepts and Moment Visualization [100]

To increase the efficiency of forming a boolean query based on the visual concepts and tags extracted from lifelog images, Nguyen-Khang Le et al. proposed to enrich the metadata by training more concept detectors based on the daily habits of the lifelogger [100]. In detail, the authors extracted a subset from the Open Images V4 [101], which was a large-scale dataset with "unified annotations for object detection and visual relationship detection", to train multiple object detectors that belong to four groups of concepts that appear frequently in the daily life of the lifelogger based on their analysis: Main food, Dessert, Musical Instrument, and Devices. With this approach of metadata enrichment for efficient boolean query formulation, the system was competitively in third place in the LSC'19.

2.4.3.3 Lifelog Search Challenge 2020 & 2021 (LSC'20 & LSC'21)

Myscéal System [8]

- **Data Indexing:** In the Myscéal system, the authors defined the events in their system were sequences of images that are visually identical. Any change of action or shift in the lifelogger's viewpoint was considered an indication of the event change. To group each single lifelog image into events, visual features of each image including SIFT feature [102], embedding feature from VGG16 model [103], and visual concepts were extracted and compared with the image's immediately preceding ones using cosine similarity to determine if they were in the same event. For free-text search implementation, all the metadata, including the location in raw GPS format and time, were all converted into text for free-text retrieval, such as the exact name of the location, the time



Before: when 1h Find: by the sea in the morning After: when 1h

Click an event thumbnail to view all images.

Figure 2.7: The user interface of Myscéal system in LSC'20 [8].

range, the day of the week, etc. The authors also enriched the visual concepts by using an additional pixel-wise object detector called DeepLabv3+ [104] retrained on the ADE20k dataset [105]. Based on the pixel-wise segmentation object detector results, the authors proposed a new TF-IDF weighting named area TF-IDF (aTFIDF) that took the pixel-wise area of the object into account by putting a constant threshold for the area to determine if an object is a visual noise or if it is visually important in the image.

- **Retrieval Algorithm:** Myscéal provided free-text search and filter with re-defined Term Frequency - Inverse Document Frequency (TFIDF) weighting in ElasticSearch database as well as providing a new temporal retrieval function. The free-text search function was done by defining a rule-based part-of-speech tagging from the natural language toolkit [106] with query expansion using Word2Vec [107] and WordNet [108] models to extract keywords for boolean query formulation. In addition, their retrieval engine also supported temporal search, which searched for the exact event based on the description of the event that happened either before or after it.

- **User Interface & User Interaction:** For the user interface, as can be seen from Fig. 2.7, the authors displayed the search results in a straightforward way that showed the retrieved event photo in the middle with the events before and after on both sides of it. The authors also introduced a new method of filter-by-location-area user interaction to search faster in case of the area of the searched event was provided by drawing a rectangle on a geographic map.

These novel features in retrieval engine functions, well-designed user interface, and user interaction resulted in a huge gap in the score of the Myscéal system with other lifelog retrieval systems in both LSC'20 [8] and LSC'21 [109].

2.4.3.4 Lifelog Search Challenge 2022 (LSC'22)

The Contrastive Language-Image Pre-Training model [110] (CLIP model) enabled the mapping of text-based queries and images into the same vector space for embeddings. As a result, the disparity between text-based embedding and image-based embedding was minimized, allowing the contextual information of text-based queries to be utilized for image retrieval. This means that there is no longer a requirement to enhance the metadata of visual concepts from the images in order to perform free-text searches, while still maintaining a high-performing retrieval system. The existence of the CLIP model is indeed a game changer leading to substantial improvement in the performance of lifelog retrieval systems. This was proven in the LSC'22 that the top five best retrieval system uses the CLIP model as the core retrieval engine for the free-text search function.

2.5 Discussion

Based on the aforementioned literature studies, three key issues have been identified, which have subsequently formed the basis of the research conducted in this thesis. These issues can be summarized as follows:

1. **The lack of research on the evaluation of the performance of stress detection models trained on data from consumer-grade wearable**

devices:

Based on the literature review conducted in Section 2.4.1, previous studies utilizing clinical-grade devices and high-resolution physiological signals consistently indicated that subject-dependent stress detection models outperformed subject-independent models. These studies also proposed various solutions to enhance the performance of subject-independent models. However, the evaluation of stress detection models trained on data collected from consumer-grade wearable devices has been lacking. Additionally, there has been no comparison made between the performance of subject-dependent stress detection models and subject-independent models trained on low-resolution data from consumer-grade wearables. Consequently, further research is required to statistically validate the performance of stress detection models trained on low-resolution physiological signals using all three benchmarking stress datasets. This research will facilitate the determination of the optimal approach for constructing stress detection models with low-resolution data.

2. **The lack of research on stress detection models applied to lifelog data/in-the-wild data**:

   The performance evaluation of stress detection models primarily takes place in controlled lab environments, leaving a significant gap in understanding their effectiveness in real-world scenarios. The challenge of effectively training stress detection models with low-resolution physiological signals from consumer-grade wearable devices remains an open problem. Section 2.4.2 highlights that most recent works on in-the-wild stress detection proposed the use of both Machine Learning and Deep Learning techniques to develop subject-independent models capable of detecting stress levels for all individuals in real-life situations. However, these approaches overlooked the variations in physiological responses among individuals and the analysis of results.

While initial research efforts show promising results regarding the performance of stress detection models in real-world settings, most of these models are unable to operate in real time due to the low sampling rate of the devices (1 to 5 minutes per sample). Furthermore, the availability of stress labels for all data points is limited, as they are obtained through questionnaires and self-evaluation forms at specific times of the day [91, 92, 93].

To address these challenges, my research proposes a novel approach to streamline the stress annotation process and conducts experiments to assess the performance of my proposed stress detection model in real-world settings in real time, achieving a detection rate of 20 seconds per detection. In my opinion, resolving this issue necessitates a case study-based solution that accounts for the unique characteristics and requirements of each individual, ultimately leading to the identification of an optimal approach for in-the-wild stress detection.

3. **The lack of exploiting physiological data in lifelog retrieval systems**:

   Despite the extensive research conducted on analyzing lifelog data to gain insights into personal habits [100] and developing lifelog retrieval systems as memory prosthetics [22], the exploration of physiological data for understanding personal physical and mental states remains largely untapped. Through my literature review, I have observed that the physiological data present in lifelog data, such as heart rate, blood volume pulse, and galvanic skin response, can be leveraged to accurately detect stress moments.

   Based on this realization, I hypothesize that mental stress information plays a pivotal role in memory prosthetics, aiding memory retrieval in a faster and more efficient manner. Consequently, I propose that enhancing the current state-of-the-art lifelog retrieval systems by incorporating stress status as a searchable facet can improve their performance. To validate this hypothesis, it is necessary to develop a state-of-the-art lifelog retrieval system and conduct experiments accordingly.

Drawing from the literature studies in Section 2.4.3, I have identified several important functions that would be valuable for my research in developing the aforementioned lifelog retrieval system. These functions include:

(a) **Free-text Search:** A robust free-text search capability is essential for the system to understand the content and context of given queries in order to locate relevant images. This can be achieved through techniques such as the Bag-of-Words model as seen in the initial version of Myscéal [8], or by employing models like CLIP as utilized by Memento [111] or E-Myscéal [112].

(b) **Filter by Text:** Considering the extensive metadata inferred from lifelog data, I am inspired by the query-by-text function of the vitrivr system [99] to define a syntax for forming boolean queries to facilitate filtering, instead of relying on a multi-faceted filter function displayed on the user interface.

(c) **Visual Similarity Search:** Users often desire the ability to search for visually similar images during the browsing process, as these images may contain visual cues associated with specific memories. Furthermore, visual similarity search facilitates the discovery of related or similar images, as well as the identification of patterns or trends within users' life.

(d) **Straightforward User Interface and User Interaction:** The interface should be intuitive and user-friendly, enabling users to easily search and browse their lifelog data. Images should be displayed in a clear and organized manner, accompanied by relevant metadata such as date and location. The interface should support temporal moment (image) browsing, as commonly seen in state-of-the-art systems, to verify the accuracy of the identified moments (images). Interactions should be minimal and straightforward, requiring only a few clicks, taps, or inputs to accomplish tasks such as searching, browsing, and filtering. The system should be designed to be accessible and user-friendly, allowing even novice users to quickly learn and utilize the lifelog retrieval system.

By integrating these critical functionalities, the proposed state-of-the-art lifelog retrieval system can be developed, serving as a foundation for subsequent experiments and evaluations.

2.6 Chapter Summary

This chapter serves as an introduction to the background research and literature review conducted for the research presented in this thesis. The focus of the background theory is on the identification of mental stress through the analysis of physiological signals generated by the body's automatic response system. The theoretical framework encompasses the understanding of the Autonomous Nervous System's behavior within the human neural system, particularly in fight-or-flight situations, by exploring the anatomical aspects of the brain, nerves, and spinal cords throughout the body. Additionally, this chapter provides a comprehensive review of important and innovative features found in state-of-the-art interactive lifelog retrieval systems. These systems have been examined to gain insights into the current advancements and trends in the field. Building upon the background knowledge and literature review, three main problems have been identified as the driving force behind the research conducted in this thesis. These problems include the lack of research in evaluating stress detection models trained on data from consumer-grade wearable devices, the scarcity of studies on stress detection models applied to lifelog data or in-the-wild scenarios, and the underutilization of physiological data in lifelog retrieval systems. These key issues form the central focus of the research conducted in this thesis.

Chapter 3

Research Methodology and Evaluation Methods

3.1 Research Methodology

According to S. Rajasekar et al., research is “a structured inquiry that utilizes acceptable scientific methodology to solve problems and create new knowledge that is generally applicable” [113]. J. Creswell also proposed a definition of research as a logical and systematic process that collects and analyses data to discover new knowledge, find solutions to scientific and social problems, and improve the understanding of a particular topic [114]. Moreover, there are many different approaches and methods to plan, orientate, and conduct the research [114].

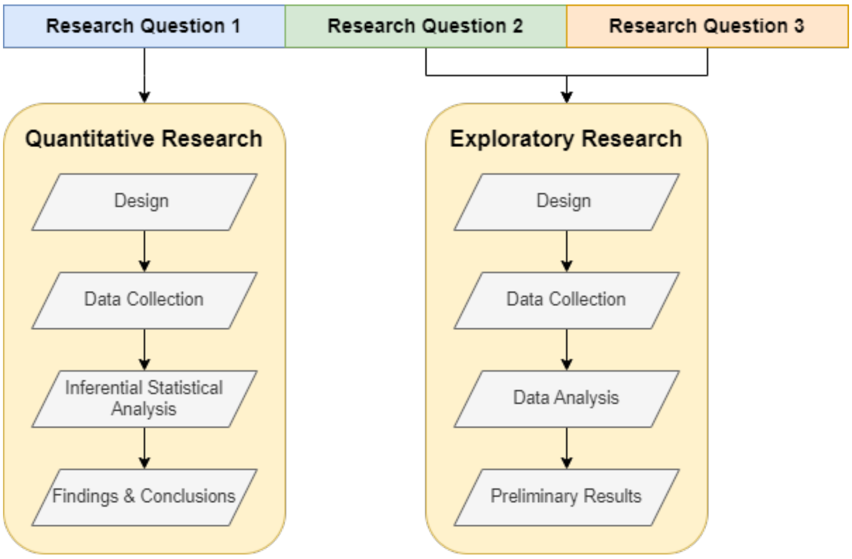


Figure 3.1: Research methodology schema.

Based on this literature knowledge, I choose an appropriate research methodology to address each research question to either prove or disprove the proposed hypothesis mentioned in Section 1.4. As described in Section 1.4, there are three research questions in total. As illustrated in Fig. 3.1, quantitative research is employed to address research question 1 while exploratory research is used to address research question 2 and research question 3. In summary, there are four steps in each research methodology to conduct experiments and research that have similar first two steps of designing and data collecting, as demonstrated in Fig. 3.1. For both research methodologies, the design step would require the literature studies and background research to determine how to approach the problem stated in the research. Then, appropriate data are gathered based on previous studies for experiments and analyses. For research question 1, the quantitative research methodology aims to generalize results from a sample of a targeted population that provides objective analyses using statistical means [115]. Therefore, multiple stress datasets recorded in diverse environments are used in the experiment to augment the sample size, thereby bolstering confidence in the findings and conclusions derived from the analyses conducted to address this research question. For research question 2 and research question 3, the exploratory research methodology is chosen to conduct case-study experiments to gather information from preliminary results to gain a deeper understanding of the research topic and guide further research [116]. The workflow of the research methodology for each research question, as illustrated in Fig. 3.2 is as follows:

- **RQ1**: I divide this research question into two parts. The first part – (RQ1.1) – is to validate the performance of the stress detection model using data from consumer-grade wearable devices compared to the one trained on traditional clinical-grade devices. The second part – (RQ1.2) – is to determine an appropriate approach to training the stress detection model by comparing if a personalized stress detection model is better at detecting stress accurately than a general one. For RQ1.1, the WESAD dataset described in Section

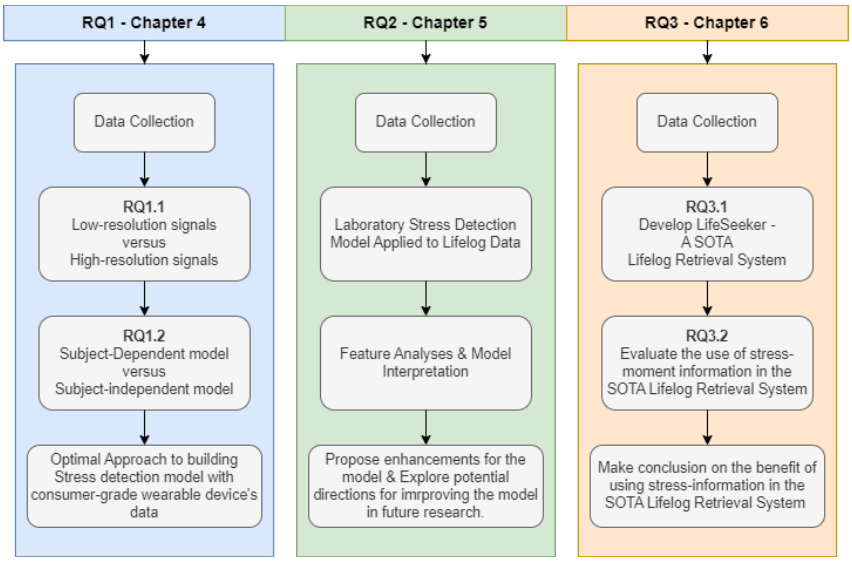

RQ1 - Chapter 4
RQ2 - Chapter 5
RQ3 - Chapter 6
Data Collection
RQ1.1
Low-resolution signals
versus
High-resolution signals
RQ1.2
Subject-Dependent model
versus
Subject-independent model
Optimal Approach to building Stress detection model with consumer-grade wearable device's data
Data Collection
Laboratory Stress Detection Model Applied to Lifelog Data
Feature Analyses & Model Interpretation
Propose enhancements for the model & Explore potential directions for imrproving the model in future research.
Data Collection
RQ3.1
Develop LifeSeeker -
A SOTA
Lifelog Retrieval System
RQ3.2
Evaluate the use of stress-moment information in the SOTA Lifelog Retrieval System
Make conclusion on the benefit of using stress-information in the SOTA Lifelog Retrieval System

Figure 3.2: The workflow of the research methodology for each research question.

4.2.1.1 is used in the experiment as it records the same subject's high-resolution and low-resolution physiological signals contemporaneously is used in the experiment. For RQ1.2, all of the stress datasets described in Section 4.2.1 are used in the experiment as the same consumer-grade wearable device (Empatica E4) is employed during data collection. Based on a sample of data, inferential statistics analysis is applied to estimate the performance of the stress detection model built on different conditions statistically. These conditions including different physiological signal resolutions (RQ1.1 – Section 4.3) and different model training approaches (RQ1.2 – Section 4.4) are analyzed in different experiments. Thereby, an overall statistical conclusion can be drawn to determine the optimal approach to train a stress detection model with a consumer-grade wearable device.

- **RQ2**: I conduct a case study on three randomly selected participants from the list of candidates joining the lab-based stress data collection protocol described in Section 4.2.1.4 to collect lifelog data for the experiment of stress detection in the wild, which is described in Section 5.2.1. As the optimal approach to building the stress detection model with readily available consumer-grade wearable devices has been determined, I examine the capability of detecting stress moments in lifelog using the stress detection model trained on laboratory-based data, which is described in Section 5.3.1. Feature analyses are done in Section 5.3.2 to explain the features' impact on the decisions of the stress detection model. Thereby, further suggestions and improvements on the lifelog stress detection model are proposed in Section 5.3.3 with discussions on the avenues for future research in this field.

- **RQ3**: I conduct experiments and user studies on selected participants joining the lifelog data collection in Section 5.2.1 to determine if using subjective stress indicators as a filter option can actually improve the performance of the state-of-the-art lifelog retrieval system. To do that, I divide this research question into two parts. The first part – (RQ3.1) – is to analyze the system architecture and supported features in the state-of-the-art lifelog retrieval system, which is described in Section 6.2. I develop the lifelog retrieval system named LifeSeeker and have the system evaluated through multiple benchmarking Lifelog Search Challenges (LSC) in 2020, 2021, and 2022. The results from the challenges mentioned in Section 6.2.3 show that LifeSeeker is one of the state-of-the-art lifelog retrieval systems. The second part – (RQ3.2) – is to evaluate the performance of the state-of-the-art lifelog retrieval system without subjective stress indicators through three experiments described in Section 6.3. Preliminary results shown in Section 6.3.2 indicate that the integration of the stress-moment filter enhances the performance of the lifelog retrieval system in both when dealing with stress-related/emotional-related queries. Thereby, further suggestions on the improvement of how this feature can be integrated into lifelog retrieval system effectively in terms of user interaction and user interface are discussed.

3.2 Operating Constraints

For any new research topic, I should define the operating constraints of the research to design and conduct the experiment properly. In this Ph.D. research, I identify these constraints as follows:

- While the labels of the stress-related psycho-physiological signals are subjective as they are recorded based on their emotions and feelings at that moment, they are assumed to be reliable.

- The labels of stress moments used as the ground-truth for the stress detection model applied in real life should be annotated manually by the participants after they were explained the definition of stress. The researcher is not allowed to intervene in the annotation process or influence the participants by any means to achieve the expected results.

- The data collection process must respect the privacy of the participants and comply with the data governance laws. Ethical approvals should be agreed upon by the Ethical Research Committee from the research institution. In this Ph.D. research, I ensure that ethical approvals have been secured for all the conducted experiments from the Dublin City University Ethical Research Committee.

- The data collected at the research institution should be structured into a reusable dataset and support further research in the same research domain while maintaining the private property of personal traits and identities of the participants.

- The targeted participants joining the data collection process for RQ2 and RQ3 should be the lifeloggers or the researchers in lifelogging fields as they have a good understanding of the lifelog sensors to ensure the integrity of the data. Therefore, the number of participants in this dataset is small. It is a trade-off between the reliability of the dataset and the sample size. In this research, I are more concerned about the reliability of the data, which would affect the reliability of my conclusion drawn from experiments conducted on this data.

- The interactive lifelog retrieval system developed for RQ3 after doing background research and further analyses should have a competitive performance compared to the current state-of-the-art interactive lifelog retrieval systems.

These constraints are maintained for this Ph.D. research and act as limiting factors to focus the research effort.

3.3 Evaluation Methods

My proposed evaluation metrics for each research question are as follows:

RQ1: Evaluation of Stress Detection Model trained on Physiological Signals from Consumer-grade Wearable Device.

Due to the imbalanced nature of the stress datasets, the conventional accuracy metrics could not reflect the actual performance of the detection model. In previous works, five evaluation metrics are often used to evaluate the performance of the stress detection model: accuracy, balanced accuracy, precision, recall, and f1-score. Among those evaluation metrics, balance accuracy is the most intuitive evaluation metric to assess the overall performance of the learning models on an imbalanced dataset [117]. Therefore, balance accuracy is employed as the main evaluation metric for stress detection models and inferential statistical analyses are done using this evaluation metric also.

The formula for the balanced accuracy score is:

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2} = 0.5 \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (3.1)$$

where TPR is the True Positive Rate, TNR is the True Negative Rate, and the

Table 3.1: The Template for Binary Confusion Matrix.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TP)	False Positives (FP)
Predicted Negative (0)	False Negatives (FN)	True Negatives (TN)

relation between true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are presented in the confusion matrix in Table 3.1.

RQ2: Stress Detection Model Applied in Lifelog Data: Analyses and Proof of Concept.

In this research question, I still use the balance accuracy score to evaluate the performance of stress detection models applied in lifelog data to assess the general performance of the stress detection models in the wild. However, I also consider precision and recall scores to measure how many stress moments the model retrieves and how many moments are correctly detected as stress. The formula of the precision and recall score based on the confusion matrix in Table 3.1 is as follows:

$$\text{Precision} = \frac{TP}{TP + FP}; \tag{3.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

RQ3: Subjective Stress Indicators as a Facet of Lifelog Interactive Retrieval System.

In this research question, I evaluate the interactive lifelog retrieval systems when conducting a user study by employing the same evaluation metrics used in the benchmarking Lifelog Search Challenge (LSC). In detail, for each query/task, the organizers of the Lifelog Search Challenge rank the system by considering both retrieval speed and the correctness of the submissions into a score as follows:

$$S_q^t = \max\left(0, A_q \cdot \left(\frac{T_q \cdot 0.9^{\omega_q} - 0.5 \cdot \tau_q}{T_q}\right)\right) \tag{3.4}$$

According to the equation 3.4, let A_q be the maximum score of the task, T_q be the maximum provided search time, ω_q be the number of incorrect submissions, and τ_q be the search time that the user required to solve the task; the score linearly decreases from the maximum to half of the points over the allowed search time (and will be zero in the worst case when the user submits too many incorrect submissions) [118]. The final score used for evaluation is the sum of all task scores in the user study session: $S = \sum S_q^t$.

In addition, I evaluate the average precision (AP) and the mean average precision (mAP) of the system by using the ranked list from the log of the system generated when the user submits correctly. In the equation 3.5, the $P@k$ is the precision of the item k (I_k) in the ranked list and the $rel@k$ is 1 if I_k is relevant or 0 otherwise. The $|I_{\text{rel}}|$ is the total number of relevant items in the ranked list of a query/task. The average precision is the mean of all precision values at different levels in the retrieved ranked list of a task while the mean average precision is the mean of all average precision values over all tasks.

$$\text{AP} = \frac{1}{|I_{\text{rel}}|} \cdot \sum_{k=1}^{n} P@k * rel@k \tag{3.5}$$

$$rel@k = \begin{cases} 1 & \text{if } I_k \text{ is relevant} \\ 0 & \text{if } I_k \text{ is not relevant} \end{cases} \tag{3.6}$$

$$\text{mAP(Q)} = \frac{1}{Q} \cdot \sum_{q}^{Q} AP(q) \tag{3.7}$$

Chapter 4

Evaluation of Consumer-grade Wearable Device Applied to Stress Detection Problem

4.1 Introduction

In this chapter, we address research question 1, which is **how successfully can low-resolution physiological signals recorded from consumer-grade wearable devices unlike traditional clinical devices with high-resolution ones be used to detect acute stress of an individual automatically by utilizing learning models?**

To answer this research question, I evaluate the statistical difference between the performance of various learning models trained on the data recorded from both a clinical device and a consumer-grade wearable device. I then compare two conventional approaches to training human-related learning models, which are the subject-dependent and the subject-independent, to apply to the training process stress detection models using consumer-wearable device data. Finally, I compare the performance of each learning model to select the best approach to build a stress detection model with consumer-grade device data. In short, research question 1 can be addressed by providing answers to these two sub-research questions:

- **Research Question 1.1**: Can physiological signals recorded from consumer-grade wearable devices be used to develop a stress detection model for an individual?

- **Research Question 1.2**: Does the subject-dependent stress detection model achieve higher evaluation scores in detecting stress moments than the subject-independent stress detection model as used by the current generation consumer-grade wearable devices?

As mental stress can be detected via physiological responses, much research has been conducted to experiment with the capability of building stress detection models using the physiological signals recorded from either clinical-grade devices or consumer-grade wearable devices as well as evaluating the performance of these models. The difference between the two kinds of devices lies in the sampling rate of the devices themselves. Conventional clinical-grade devices can provide high-resolution signals due to the high-capacity battery that allows them to capture data at a high sampling rate. However, these kinds of a device are not mobile and convenient enough to capture physiological signals in the wild as they can easily hinder individual daily activities. Though consumer-grade wearable devices manage to record these kinds of signals outside the laboratory environment, the battery issue does not allow them to capture high-resolution signals. However, the low resolution is also good enough to capture minor changes in physiological signals that result in acute stress according to [2, 9, 119, 120]. However, no research has been conducted to measure how useful it is to use low-resolution physiological signals for stress detection compared to the usage of high-resolution signals, except for my previous analyses in [121] with different experiment settings. Additionally, the optimal approach between the two conventional methods of building stress detection with low-resolution physiological data, which is either the subject-dependent method or the subject-independent one, is not determined clearly in any research. These are the core problems that I focus to solve by addressing research question 1 in this chapter.

4.2 Experiment Configuration

4.2.1 Stress Datasets

In this section, I describe the stress datasets that I used in my experiment to evaluate the performance of the stress detection model with low-resolution data captured from consumer-grade wearable devices. I employ four stress datasets recorded in the constrained environment, which are WESAD [9], AffectiveROAD [10], Cognitive Load [2], and my collected dataset at Dublin City University named DCU-NVT-EXP2.

4.2.1.1 Wearable Stress and Affective Detection (WESAD) Dataset

The benchmarking dataset named WESAD [9] consists of four different types of low-resolution physiological data collected from 15 participants (ages of 27.5 ± 2.4 years) under two different study protocols in a laboratory environment by Philip Schmidt et al. The low-resolution physiological signals including accelerometer (ACC), skin temperature (TEMP), Blood Volume Pulse (BVP), and Electrodermal Activity (EDA) were recorded using the Empatica E4 medical-grade wearable sensor while the high-resolution physiological signals of the same types were captured simultaneously using the RespiBAN device with a sampling rate of 700 Hz. The sampling rate of the consumer-grade wearable device for low-resolution signals recording is 64 Hz for Blood Volume Pulse (BVP), 32 Hz for the accelerometer (ACC), and 4 Hz for both Electrodermal Activity (EDA) and skin temperature (TEMP). As illustrated in Fig. 4.1, there are two versions of the study protocols with the same stress and rest conditions but in different orders of execution. The gaps between conditions, which are shown as red boxes in Fig. 4.1, refer to the times when participants filled in self-reports. Each study protocol in the dataset comprises amusement, stress, meditation, and baseline conditions arranged in different orders for each participant. Details of these four affective conditions are as follows:

1. **Baseline Condition**: This condition lasts for 20 minutes and aims to capture the neutral state of the participant. The participant was asked to sit or stand at a table with neutral reading material.

2. **Amusement Condition**: The participant watched a set of eleven funny video clips. A short neutral time period of five seconds was presented between the video clips. The total length for this condition was 392 seconds.

3. **Stress Condition**: The participant was exposed to the Trier Social Stress Test (TSST), where they were required to provide a five-minute speech on their strengths and weaknesses in front of a panel of three human resource specialists. Finally, the participant counted down from 2023 in decrements of 17 and was requested to start over if they made a mistake. The total length of this condition was about 10 minutes.

4. **Meditation Condition**: The amusement and stress conditions were followed by a guided meditation period to "de-excite" the participant back to a neutral affective state.

However, only the amusement, stress, and baseline conditions were used to build and evaluate stress detection models [9].

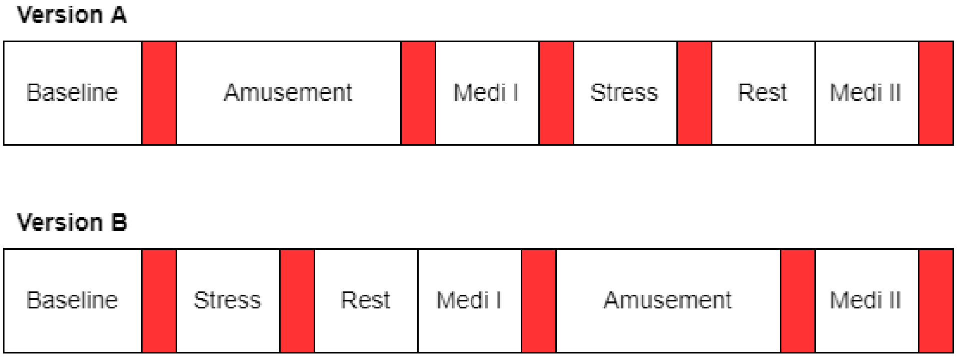


Figure 4.1: The study protocols in the data recording process of the WESAD dataset [9].

The total duration of the study protocol was about two hours. Philip Schmidt et al. collected four subjective self-evaluation reports from each of the participants;

which include the Positive and Negative Affect Schedule (PANAS), State-Trait Anxiety Inventory (STAI), Self-Assessment Manikins (SAM), and Short Stress State Questionnaire (SSSQ); to capture the feelings of the participant after each affective condition. Since previous works on this dataset employed study protocol as the ground truth of both train and test data [9, 41, 83, 121], I also used the same ground-truth construction method as in previous works for consistent comparison of the results. In detail, the baseline and amusement conditions were classified as non-stress while the stress condition is labeled as stress.

4.2.1.2 Cognitive Load Dataset [2]

The dataset was collected by Martin Gjoreski et al. to monitor the physiological signals under the cognitive-load inducing task [2] using both commercial Electroencephalography (EEG) headset and Empatica E4 wristband. The cognitive-load tasks used in their experiment were designed based on the stress-inducing method proposed by Dedovic et al. [85] that required an individual to solve a mental arithmetic task under a certain amount of time and evaluation pressure. The experiment started with a baseline session of 15 minutes on a separate day when the participant was relaxed. They then invited the participant to the laboratory room where the participant joined the arithmetic tasks including easy level (addition and subtraction of two integers), medium level (addition and subtraction of three integers), and hard level (addition, subtraction with multiplication of three integers). The tasks were organized following the four conditions proposed by Sonia J. Lupien [43] to induce a stress response. The authors also announced that there would be monetary prizes for the top three participants to increase the competitiveness of the tasks to stimulate their stress status during the experiment. Each level lasted for five minutes. Therefore, the total duration of the experiment was around 30 to 40 minutes.

For stress level evaluation, the authors employed the Four short State-Trait Anxiety questionnaires (STAI-Y) [122] that required the participant to fill in before the experiment and after each stress session. The answers to the STAI-Y questionnaires were used as the self-evaluated stress level of the data. In detail, the session with the lowest STAI score was labeled as low stress. For each +3 STAI point, the stress label increased by one. In total, there were four stress levels obtained from the stress self-evaluation form including no stress (baseline data), low stress (lowest STAI score), medium stress (lowest STAI score + 3), and high stress (lowest STAI score + 6).

4.2.1.3 AffectiveROAD Dataset [10]

The dataset was collected by Neska El Haouij et al. to identify and validate drivers' state indicators such as stress and arousal. The authors gathered low-resolution BVP and EDA data from two Empatica E4 wrist bands on both arms as well as high-resolution Heart Rate and Respiration data using a chest-worn device named Zephyr BioHarness 3.0. Ten participants, except for one participant having the age of 59 years old, with ages varied between 24 and 34 years old (29.9±3.7) were invited to join in 14 different driving tasks whose driving experience ranges from 5 to 37 years (11 ± 8.37).

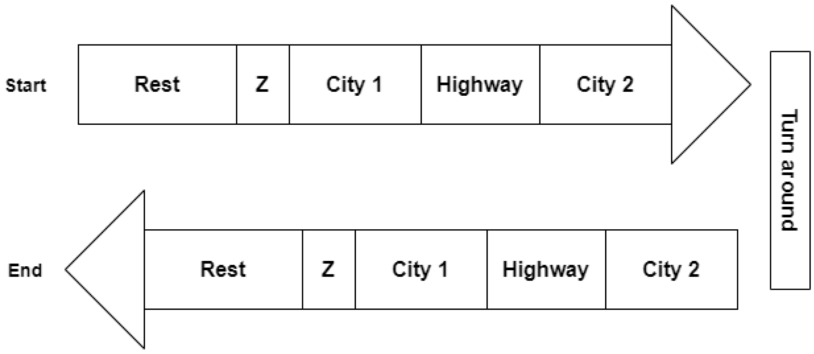


Figure 4.2: The study protocol in the data recording process of the AffectiveROAD dataset [10]

The road path of the study protocol in the data recording process of the AffectiveROAD dataset is depicted in Fig. 4.2. It can be seen from Fig. 4.2 that the driving protocol began and ended with 15 minutes of rest in the institution parking area where the driver sat still in the car, and closed their eyes while the car engine was running. After the rest period at the beginning of the experiment, the participant left the parking and drove through a pre-defined set of routes in daily normal traffic. The Z in Fig. 4.2, was the exiting route of the parking area, which is a part of the Technopole. Then, the driver needed to drive along the streets of the city (City 1), which was presumed to induce high stress due to narrow streets, traffic lights, and high-load traffic. The driver then exited the avenue of the city and continued to drive on a smooth route for around 8 minutes. The driver finally arrived at a roundabout before entering the city driving session again (City 2) for around 10 minutes. However, this city route did not have traffic lights and had many parking lots and restaurants. Finally, the driver arrived at a big roundabout and drove back to the starting point of the experiment following the same routes. The total duration of this study protocol was around one 1 hour and 26 minutes with 30 minutes of rest periods.

The stress level of the participant was rated on a continuous ”stress” metric ranging from 0 (no stressful) to 1 (extremely stressful) by the experimenter using a slider while the experimenter sat in the rear of the car during each driving task. Due to the subject evaluation of stress level, the annotation of each drive was validated again by the participants, which was smoothed using the Hanning filter with a window size of 100s duration, by looking at the synchronized recording video of the drive [10]. In my experiment, the ground truth of the data for stress level based on these continuous metrics was divided into three labels including relaxed, low stress, and high stress by splitting these continuous values into three intervals: [0, 0.33) – relaxed, [0.33, 0.67) – low stress, and [0.67, 1] – high stress.

4.2.1.4 DCU-NVT-EXP2 Dataset

The dataset was collected at Dublin City University with the purpose of investigating the potential of applying a laboratory-based subject-dependent stress detection model in an unconstrained environment. I used the wrist-worn Empatica

E4 device to record stress-related physiological signals from participants. In total, there were 11 participants in my dataset with ages ranging from 20 to 47 (27.8 ± 7.0). The DCU-NVT-EXP2 dataset is the constrained version where participants join the study protocol in a laboratory environment with the requirements to complete a set of pre-defined tasks.

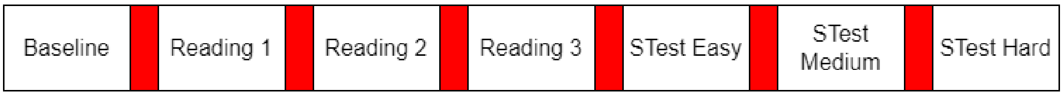


Figure 4.3: The study protocol in the data recording process of the DCU-NVT-EXP2 dataset

As illustrated in Fig. 4.3, the study protocol consists of seven sessions including one baseline task which aims to capture the normal state of the participant. After the baseline session, the participant started to do three reading sessions. The reading tasks were designed by following the International English Language Testing System's reading test format with a time constraint of 15 minutes, which is less than the actual time constraint required to complete an IELTS reading task successfully. After the reading sessions, the participant joined three cognitive-load tasks following the study protocol by Martin Gjoreski et al. in [2]. The ticking clock with time progress was visible to the participant during the sessions (excluding the baseline) to urge the participant to complete the tasks in the specified time constraint. The scoreboard was visible to all the participants to increase the competitiveness for rewards of €100 for the one with the highest score and €20 vouchers for the second-ranked and third-ranked participants. These features in this stress-induced protocol were designed based on four conditions proposed by Sonia J. Lupien [43] to induce a stress response, just as in [2].

The red boxes in Fig. 4.3 between the sessions are the resting stage of 5 minutes which provides a break time for the participant to get back to their normal state. In addition, the participants were asked to evaluate their affective state of themselves before and after each session (excluding the baseline) using the six short State-Trait Anxiety questionnaires (STAI-Y) as in [2] illustrated in Table 4.1. The subjective

Question	**Answer**			
	Not At All	Somewhat	Moderately	Very Much
I feel calm	1	2	3	4
I am tense	1	2	3	4
I feel upset	1	2	3	4
I am relaxed	1	2	3	4
I am content	1	2	3	4
I am worried	1	2	3	4

Table 4.1: Six short State-Trait Anxiety Questionnaires (STAI-Y) and the score for each question used in DCU-NVT-EXP2 experiment

self-evaluation STAI-Y results were assessed as follows:

- The answers from six questions were converted into scores, which were summed up to provide the self-evaluation score of the current stress state of an individual at that moment.

- The self-evaluation scores obtained from 11 participants was then tested to measure if there were any significant changes in the stress state before and after doing the tasks in the session. The significance test was selected based on the normality of the data, which was tested using the Shapiro-Wilk test [123] in advance. Shapiro-Wilk normality test is chosen as it is the conventional approach to test the normality of the data distribution to choose a proper test for hypothesis testing. All the significance levels in these significance tests were 0.05.

- The session was then labeled as stress/non-stress based on the results from the significance test. It is worth noting that the baseline was considered non-stress by default due to the nature of the baseline session in a laboratory stress-induced environment.

It can be seen from 4.2 that distributions of the self-evaluation score differences follow the normal distribution according to the Shapiro-Wilk normality test at the significance level of 0.05. Therefore, I applied the paired t-test to measure the significance difference before and after doing the tasks to validate if the task was

	Normality test		**Paired t-test**		
Session	**p-value**	**Confirmed**	**p-value**	**Significant**	**Label**
Reading 1	0.12	Yes	0.00038	Yes	Stress
Reading 2	0.62	Yes	0.048	Yes	Stress
Reading 3	0.71	Yes	0.011	Yes	Stress
STest Easy	0.35	Yes	0.065	No	Non-Stress
STest Medium	0.052	Yes	0.198	No	Non-Stress
STest Hard	0.26	Yes	0.0016	Yes	Stress

Table 4.2: Labels of sessions in DCU-NVT-EXP2 study protocol based on the statistically significant difference before and after each session

stressful enough to cause a significant change in the self-evaluation of their current affective state. From table 4.2, the statistical test confirmed that the reading sessions actually induce stress in the participants. It could be owing to the time pressure of finishing the task while gaining high scores to achieve the result. However, based on the self-evaluation score, the easy and medium levels of the cognitive-load stress tests were not stressful enough compared to the hard ones. The label of the session was used as the ground truth of all data points during the session time interval.

4.2.2 Bio-signal Statistical Feature Extraction

According to previous work in [9, 41], the statistical features extracted from physiological signals are useful for recognizing stress responses in individuals. The feature extraction method described in [9] and [41] has been shown to be effective when applied to high-resolution signals to detect stress patterns [41, 81]. However, only a few researches have been carried out to verify the efficiency of this feature extraction method when applied to low-resolution signals [9, 82, 120, 124]. In my previous work described in [11], I show that with minor changes and additions to the feature extraction method (e.g. high-pass/low-pass filter parameters, normalization), I can reuse the feature extraction method for low-resolution signals to extract high-quality statistical features that result in accurate stress detection models using data from consumer-grade wearable devices. Therefore, I reuse the statistical feature extraction method that I proposed in [11].

For both EDA and BVP, I extract statistical features using NeuroKit2 package[1] [125] and HRV-analysis library[2] for each 60-second segment. The window shift used in my experiment is 0.25 seconds. The values of the window size and window shift are the same as in the original paper of the WESAD dataset for consistency when comparing the detection results of the models [9]. As the physiological signals vary from person to person, I employ the feature normalization method to reduce the difference in people's physiological responses. In addition, since the signals recorded using consumer-grade wearable devices such as EDA, BVP, etc. contain many types of noise, I utilize different signal processing techniques to remove noises, baseline drifts, and outliers in the raw signal. These steps are combined together to clean the raw signal before extracting statistical features, which is considered to be a bio-signal processing pipeline to improve the quality of the extracted feature.

For the EDA, the raw signal in each 60-second segment is firstly pre-processed to remove motion artifacts using the wavelet-based adaptive denoising procedure as described in [126]. The signal is then filtered by a fourth-order Butterworth low-pass filter with a cut-off frequency of 0.5 Hz to remove line noise. The min-max normalization is then applied to the cleaned signal to remove the inter-individual difference before it is inputted into the NeuroKit2 package for Skin Conductance Response (SCR) and Skin Conductance Level (SCL) decomposition using the cvxEDA method [127]. Other characteristics of SCR including SCR Peaks, SCR Onsets, and SCR Amplitude are also extracted. Finally, the statistical EDA features from three related works [9, 41, 128] are computed, which results in a 36-dimensional vector.

For the BVP, I first clean the raw signal in each window segment by removing the outlier values over the 98^{th} and below the 2^{th} percentile using the winsorization method as in [87] and removing the baseline drift using Butterworth high-pass filter with a cut-off frequency of 0.5 Hz as in [129]. I then apply min-max normalization to the cleaned signal to minimize the physiological signal difference between individuals

[1] https://github.com/neuropsychology/NeuroKit

[2] https://github.com/Aura-healthcare/hrv-analysis

before following the previous research [121] to employ the Elgandi processing pipeline [130] for the photoplethysmogram (PPG) signal cleaning [131] and the systolic peaks detection. The systolic peaks are used to compute a list of RR intervals, which are then pre-processed using the hrv-analysis package to remove outliers and ectopic beats [132] as well as interpolating missing values. The cleaned RR intervals are used to compute the NN-intervals, which are the main items to compute time-domain, frequency-domain, geometrical, and Poincare-plot features. For frequency-domain HRV features, I employ the same parameters of low (LF: 0.04-0.15 Hz) and high (HF: 0.15-0.4 Hz) frequency bands as in [9]. The range of the very-low-frequency band used in my work is the same as in the HRV-analysis package (0.003-0.04 Hz). In summary, I inherit most of the HRV features from [9, 41] and combine them into a 30-dimensional vector.

For the TEMP, the statistical features are extracted on the raw 60-second segment signal as in [9]. The fusion of statistical features from three signal sources is a 72-dimensional vector. The detail of extracted features is shown in Table 4.3.

Table 4.3: List of extracted features. Abbreviations: # = number of, $\sum$ = sum of, STD = standard deviation, RMS = Root Mean Square.

Signal	**Feature**	**Description**
EDA	μ_{EDA}, σ_{EDA}, $\min_{EDA}$, $\max_{EDA}$	Mean, STD, min, max of EDA
	∂_{EDA}	Slope of the EDA
	range_{EDA}, range_{SCR}	Dynamic range of EDA & SCR
	μ_{SCL}, σ_{SCL}	Mean, STD of the SCL
	$\text{corr}(SCL, t)$	Correlation btw SCL & time
	$\#_{Peak}$ $\sum_{SCR}^{Amp}$, $\sum_{SCR}^{t}$	# identified SCR peaks $\sum$ SCR startle magnitudes and response durations
	$\int_{SCR}$	Area under the identified SCRs

Continued on next page

Table 4.3 – continued from previous page

Signal	Feature	Description
	μ_{SCR}, σ_{SCR}, $\max_{SCR}$, $\min_{SCR}$	Mean, STD, min, max of SCR
	$\mu_{\nabla_{SCR}}$, $\sigma_{\nabla_{SCR}}$, $\mu_{\nabla(\nabla_{SCR})}$, $\sigma_{\nabla(\nabla_{SCR})}$	Mean and STD of the 1st and second derivative of the SCR
	μ_{Peak}, σ_{Peak}, $\max_{Peak}$, $\min_{Peak}$	Mean, STD, min, max of SCR Peaks
	kurtosis(SCR), skewness(SCR)	Kurtosis and skewness of SCR
	μ_{Onset}, σ_{Onset}, $\max_{Onset}$, $\min_{Onset}$	Mean, STD, min, max of SCR Onsets
	$\text{ALSC} = \sum_{n=2}^{N} \sqrt{1 + (r[n] - r[n-1])^2}$	Arc length of the SCR
	$\text{INSC} = \sum_{n=1}^{N} \lvert r[n] \rvert$	Integral of the SCR
	$\text{APSC} = \frac{1}{N} \sum_{n=1}^{N} r[n]^2$	Normalized average power of the SCR
	$\text{RMSC} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} r[n]^2}$	Normalized RMS of the SCR
BVP	μ_{HR}, σ_{HR}, μ_{HRV}, σ_{HRV}	Mean & STD of Heart Rate and HRV
	kurtosis(HRV), skewness(HRV)	Kurtosis & Skewness of HRV
	f_{HRV}^{VLF}, f_{HRV}^{LF}, f_{HRV}^{HF}	Very low (VLF), Low (LF), High (HF) frequency band in the HRV power spectrum.
	f_{HRV}^{LFNorm}, f_{HRV}^{HFNorm} $f_{HRV}^{LF/HF}$	Normalized LF & HF band power. Ratio of HRV LF and HRV HF.
	$\sum_{x\in\{\text{VLF, LF, HF}\}}^{f}$	$\sum$ of the freq. components in VLF-HF

Continued on next page

Table 4.3 – continued from previous page

Signal	Feature	Description
	NN50, pNN50, NN20 pNN20	# and percentage of HRV intervals differing more than 50 ms and 20 ms.
	HTI	HRV Triangular index
	rms_{HRV}	RMS of the HRV
	SD1, SD2	Short and long-term poincare plot descriptor of HRV
	RMSSD, SDSD	RMS & STD of all interval of differences between adjacent RR intervals.
	SDSD_RMSSD	Ratio of SDSD over RMSSD.
	RELATIVE_RR (μ, median, σ, kurtosis, RMSSD, kurtosis, skewness)	Mean, median, STD, RMSSD, and skewness of the relative RR.
TEMP	μ_{ST}, σ_{ST}, min_{ST}, max_{ST} range_{ST}, ∂_{ST}	Mean, STD, min, max of ST Range and slope of ST

4.2.3 Train/Test Data Split for Subject-Dependent and Subject-Independent Models Training

For human-related learning models, there are two conventional training approaches which are subject-dependent and subject-independent methods. The description of each training method is described as follows:

- **Subject-independent model**: This method employs the Leave One Group Out (LOGO) strategy when splitting the train and test set. This is commonly used in consumer-grade applications by default when the customers use a model to detect their behavior patterns for the first time due to its practicality and

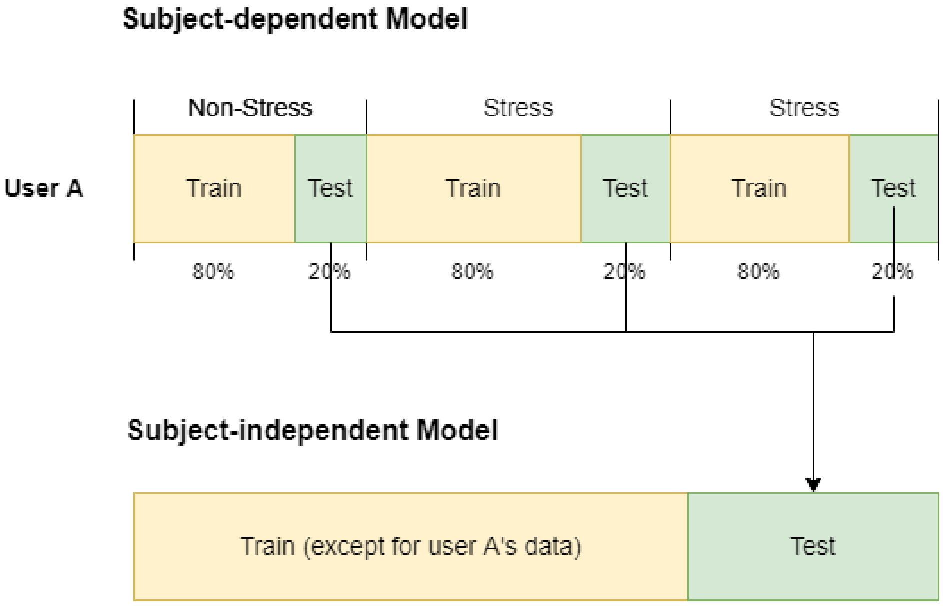


Figure 4.4: Train/test data split for subject-dependent and subject-independent stress detection models in my experiment.

cost-efficiency. In detail, assuming that the dataset contains the data from 15 users, a train/test split of the LOGO method is to use the data of 14 users for training and one of the remaining participants for testing.

- **Subject-dependent model**: In contrast to the subject-independent model, the subject-dependent one employs only the data of a targeted participant for both training and testing. The data of the targeted participant is split into train/test with the percentage of 80%/20% or 70%/30%.

In my experiment, the subject-dependent model splits the train/test data of a targeted participant (e.g. user A) with the percentage of 80%/20%. Specifically, for each stress/non-stress session, the first 80% of the session is used for training and the last 20% of the session for testing. This way of data splitting is to mimic real-life scenarios where testing physiological signals are unseen until the user records and uploads the data for analysis. For a fair comparison between the performance of subject-dependent and subject-independent stress detection models, the test set for subject-independent models evaluation is the same as the one used for subject-dependent models, which is 20% of the test data of a targeted user in the subject-dependent model evaluation (e.g. user A). Details of the data splitting in my experiment are depicted in Fig. 4.4.

4.2.4 Learning Models

In this experiment, I use four conventional Machine Learning models, which are different from each other in terms of the complexity of the model, to build stress detection models. Additionally, I also apply a deep fusion model that I propose in [11], which is described in section 4.2.4.2. The evaluation scores of these models are employed to validate the performance of stress detection models in different contexts including different model training methods and different signal resolutions due to different recording devices.

4.2.4.1 Machine Learning Models

For Machine Learning models, four conventional Machine Learning models including Extremely Randomized Trees (ET), k-Nearest Neighbors (kNN), Linear Discriminant Analysis (LDA), and Logistic Regression (LR) are trained using either high-resolution physiological signals from the chest-worn devices or low-resolution ones from the wrist-worn devices. The reason for selecting these Machine Learning models is that they are commonly used to evaluate the performance of stress detection models in previous related works [9, 41, 120]. In my experiment, I tune the parameters of these models to achieve the best performance on the datasets in general as well as prevent the over-fitting situation and boost the training speed of the model. The final adjusted parameters of each model are displayed in table 4.4 except for the Linear Discriminant Analysis model as all the parameters are set to default values as in the scikit-learn library[3]. Other parameters of each model that are not mentioned in table 4.4 are also set to default values as in the scikit-learn library. It is worth noting that the weight parameter of

[3] https://scikit-learn.org

Logistic Regression and Extremely Randomized Trees are all set to 'balance' mode due to the imbalanced nature of stress datasets, which is inversely proportional to class frequencies in the input data.

Model	Parameters	Values
Logistic Regression	solver max_iter	saga 5000
Extremely Randomized Trees	n_estimators max_depth min_sample_leaf oob_score bootstrap	500 8 8 True True
k-Nearest Neighbors	weights	distance

Table 4.4: Fine-tuned parameters of Machine Learning models in my experiment.

4.2.4.2 Deep Fusion Model

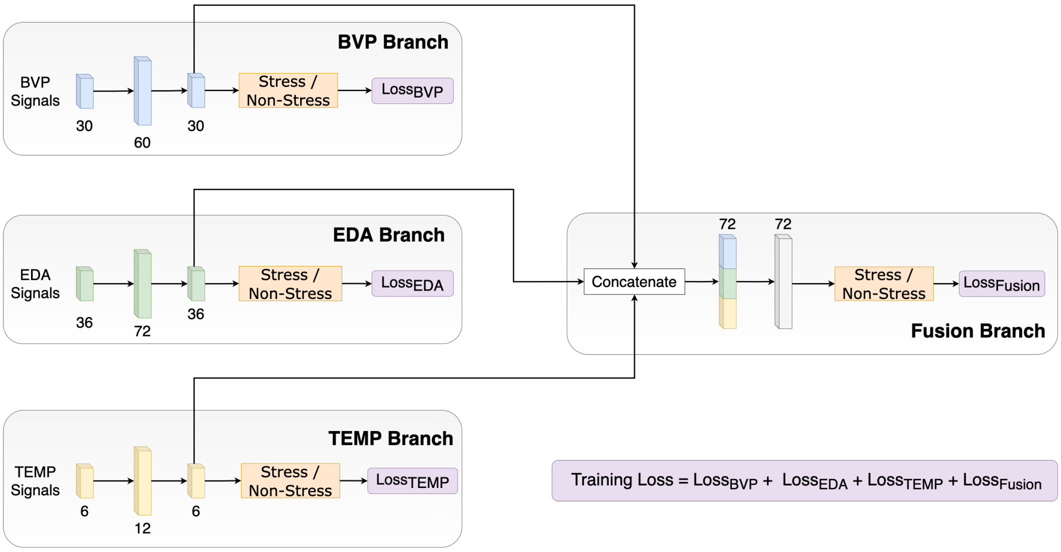


Figure 4.5: The structure of my proposed Deep Fusion model [11]. The numbers in the figure indicate the dimension of the input feature.

In my work reported in [11], I introduce a Deep Fusion model for the improvement of the subject-independent stress detection model. The difference between my proposed Deep Fusion model and conventional Machine Learning models is that it captures not only the local information of each physiological signal separately (EDA, BVP, and TEMP) but also the fusion of these signals. As depicted in Fig. 4.5, the Deep Fusion model contains three distinct embedding modules for each signal and a concatenating layer to learn the jointly encoded features. Each branch for each signal in the model contains three fully-connected layers, which aim to optimize the performance of embedding stages of EDA, BVP, and ST signals prior to the concatenating step. The overall loss used is the sum of losses of all branches and the fusion branch. I also integrate batch normalization and dropout techniques to make the model converge faster as well as address over-fitting concerns. The Deep Fusion model is trained with an Adam optimizer [133] with a learning rate of 0.003 while the dropout level and the batch size are set at 10% and 2048 accordingly.

4.2.5 Evaluation Metrics

To evaluate the performance of stress detection models, five evaluation metrics for stress detection problems including accuracy, balanced accuracy, precision, recall, and f1-score could be reported. However, in the context of conducting statistical analyses in this experiment, I focus on reporting the balance accuracy in this experiment as it is an intuitive metric to evaluate detection results of a binary classification problem when an imbalanced dataset is used for testing based on the analysis of Straube et al. [117]. Indeed, if the balance accuracy score of a model is less than or equal 50%, it means that the model does not perform better than a dummy classifier with random guesses.

4.3 The Influence of Different Recording Devices on the Performance of the Stress Detection Model

In this section, I analyse and compare the performance of stress detection models trained on either chest-worn or wrist-worn devices to address Research Question 1.1, which is **can physiological signals recorded from consumer-grade wearable**

devices be used to develop a stress detection model for an individual?

I choose the WESAD dataset to address this research question as it is the only benchmarking dataset that records both high-resolution and low-resolution physiological signals simultaneously using both clinical-grade devices and consumer-grade wearable one [9]. As mentioned in the section 4.2.4, five stress detection models are employed for each subject in the WESAD dataset, which includes Extra Trees classifier, K-Nearest Neighbours, Logistic Regression, Linear Discriminant Analysis, and my proposed Deep Fusion model. This results in $15 \times 5 = 75$ samples of the stress detection model's performance for each recording device. Additionally, in this experiment, as advised by the analysis of Sirko Straube et al. [117], I use the balanced accuracy (BA) to assess the performance of the models instead of using both evaluation metrics as mentioned in section 4.2.5 due to the imbalance nature of the dataset. Since the analysis focuses on comparing the statistical difference between the models trained on signals from two sources of device captured from the same subject simultaneously, the paired statistic test is used. Therefore, the difference between the BA scores is computed to analyze and compare the influence of different recording devices.

Since there are two conventional approaches to training stress detection models, I assess the statistical difference in the performance of the chest-worn and wrist-worn stress detection models using each training approach independently.

As can be seen from Fig. 4.6 and Fig. 4.7, the difference between the chest-worn and wrist-worn ranges from around -0.2 to 0.5 for subject-dependent models while it is approximately -0.45 to 0.6 for subject-independent models. For subject-dependent models, I can infer from the QQ-plot in Fig. 4.6 (left) that the data does not follow the normal distribution but is right-skewed instead as many data points lie outside the zone of the theoretical line and scatter on the upper plane of the line. This can be intuitively recognized from the corresponding estimated normal curve in Fig. 4.7. In contrast, the QQ-plot, as well as the estimated normal curve in both Fig. 4.6 and 4.7, implies that the distribution of the data might be normal.

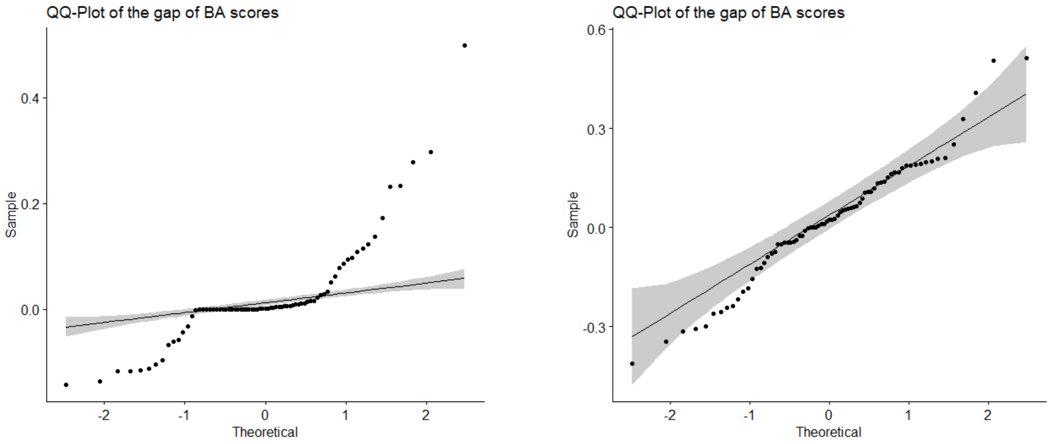


Figure 4.6: The QQ-Plot of the BA score differences between chest-worn and wrist-worn subject-dependent stress detection models (left) and subject-independent ones (right).

Indeed, the normality of the data distribution can be tested using the Shapiro-Wilk test (SW), which is shown to be the most powerful test for data normality [134]. Moreover, the Anderson-Darling test is also utilized to support the conclusion of the normality of this data. In these tests, the distribution of the data is compared with the normal distribution where null hypothesis H_0 assumes that the data comes from the normal distribution.

For the difference of BA scores between chest-worn and wrist-worn subject-dependent models, the p-value of the Shapiro-Wilk is 2.99×10^{-9} while the p-value of the Shapiro-Wilk for the subject-independent ones is 0.17. Additionally, the test statistics of the Anderson-Darling test are 6.11 and 0.59 for the subject-dependent models and the subject-independent ones corresponding. At the significance level of 0.05, the critical value of the Anderson-Darling test is 0.72. From these inferential statistics values, I can conclude that the difference in BA scores between chest-worn and wrist-worn subject-dependent models are non-normal while the ones of the subject-independent case follow the normal distribution with the confidence of 95%. Therefore, an appropriate hypothesis test should be applied for corresponding data distribution. In this case, the Wilcoxon signed rank test is employed to measure the statistical significance of the difference between the performance of chest-worn and wrist-worn subject-dependent models, and paired t-test is applied for the remaining case. The null and alternative hypotheses of the tests are established as follows:

- H_0: $M = 0$ (The capability of stress and non-stress pattern discrimination of both chest-worn and wrist-worn models is the same.)
- H_a: $M \neq 0$ (The difference between the performance of chest-worn and wrist-worn stress detection models is statistically significant.)

The term "statistically significant" that I use in the statement implies the magnitude of the difference in the estimated population of the data based on observed samples. The magnitude can be either large or small based on the confidence interval and the estimated median of the population from the sample data. In the above hypotheses, the variable M indicates the median of the BA score difference between the performance of chest-worn and wrist-worn subject-dependent models while the variable implies the mean of the difference in the subject-independent cases.

Applying the Wilcoxon signed rank test to the subject-dependent data, the p-value obtained from the test with a computed test statistic of 1105 is around 0.03, which is smaller than the significance level of 0.05. This implies that the null hypothesis could be rejected at the confidence level of 95%, which means that the difference between the performance of chest-worn and wrist-worn models is statistically significant. However, the 95% confidence interval obtained from the test is around $[0.0016, 0.0437]$ with an estimated median of 0.0111. This indicates that it is 95% confident that the estimated population of the median of the difference balance accuracy scores ranges around 0.16% to 4.37%. For the subject-independent case, the p-value obtained from the paired t-test with a computed test statistic of 0.98 is around 0.33, which is higher than the significance level of 0.05. This implies that the difference in balance accuracy scores of subject-independent models using different signal resolutions is not statistically significant as there is not enough evidence to reject the null hypothesis. Indeed, the 95% confidence intervals obtained from the test is around $[-0.021, 0.063]$ with the estimated mean from the observed samples of 0.021. This indicates that it is 95% confident that the estimated population's mean of the difference in balance accuracy for subject-independent models using different signal's resolution is around -2.1% to 6.3, which is small. From these inferential analyses, I can conclude that the magnitude of this statistical significance is small, which ranges around 0.16% to 4.37% difference in terms of performance for subject-dependent models while it is around -2.1% to 6.3% for the subject-independent case. It could be considered as the trade-off between the low error due to the granularity level of the recorded signals and the convenience of using the device for mental health monitoring in daily life. However, this trade-off is small enough that could be considered acceptable to apply for stress detection of individuals in daily life.

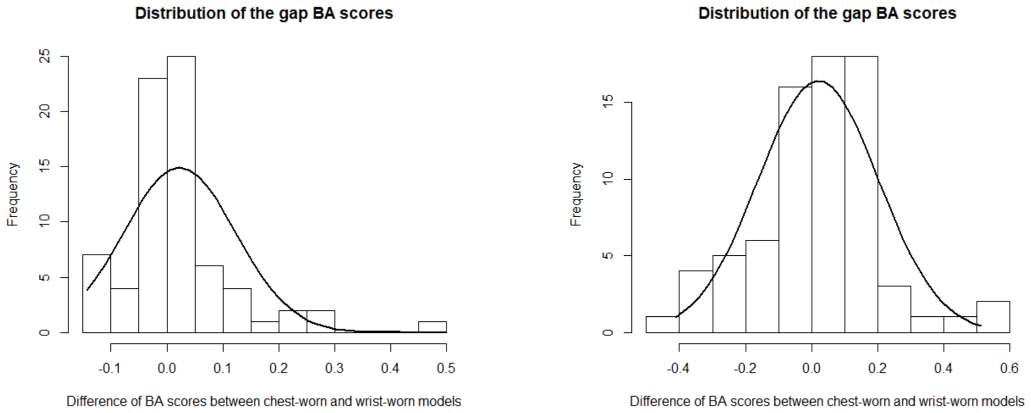


Figure 4.7: The Distribution of the BA score differences between chest-worn stress detection models and wrist-worn ones for subject-dependent (left) and subject-independent (right) training methods.

Indeed, from both histograms in Fig. 4.7, I can recognize that the mode of the difference score distribution is 0 while most of the other different scores lie in the range from $[-0.05, 0.05]$ for subject-dependent models and $[-0.45, 0.5]$ for the subject-independent case. The mean and median of this distribution data also gather around the mode, which is also small. Only a few points lie in the extreme region for both cases. This also indicates that some models do not fit well for some individuals. However, this can be addressed by using ensemble models such as Voting Classifier or Stacking Classifier. In summary, through the statistical analyses that I conducted, I can confirm that low-resolution physiological signals captured by consumer-grade wearable devices can be successfully used to build a high-performance stress detection model.

4.4 A Comparison between Subject-Dependent and Subject-Independent Stress Detection Model

In this section, I analyze and compare the performance of two conventional types of stress detection models training methods using wrist-worn devices' signals to address Research Question 1.2, which is **Does the subject-dependent stress detection model achieve higher evaluation scores in detecting stress moments than the subject-independent stress detection model as used by the current generation consumer-grade wearable devices?**. Therefore, in this section, Only the data captured from consumer-grade wearable devices, which includes WESAD, AffectiveROAD, DCU_NVT_EXP2, and CognitiveDS dataset, are used.

As described in section 4.2.3, the subject-independent models use the data from a pool of users' data to train and then provide the detection of the stress status of a targeted individual. For a fair comparison between subject-dependent and subject-independent models' performance evaluation, I use the same test set for both models as described in section 4.2.3. The same strategy for model performance comparison as in section 4.3. In detail, the difference in balanced accuracy scores (BA scores) between subject-dependent and subject-independent models is computed to compare the difference between the performance of these models.

As can be seen from the QQ-plot on the left side of Fig. 4.8, the distribution of the difference BA scores does not follow the normal distribution as most of the points are not attached to the theoretical line, especially the ones with values

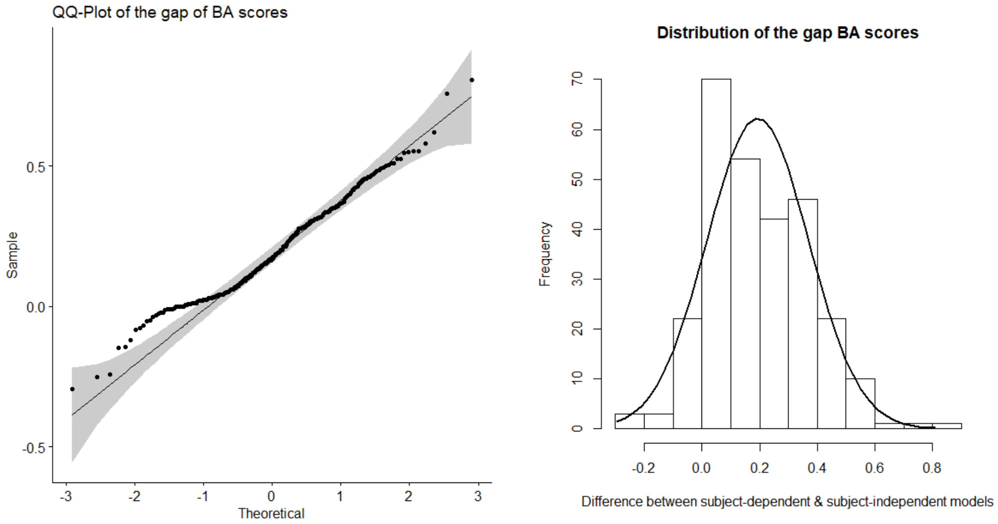


Figure 4.8: **Left**: The QQ-plot of the difference BA scores between subject-dependent and subject-independent models. **Right**: The histogram chart with a normal curve shows the distribution of differences in BA scores.

smaller than 0. Applying the Shapiro-Wilk and Anderson-Darling normality tests on the data, I obtain a p-value of 0.00026 (< 0.05) and a statistical test of 2.21 (> 0.72 at a significance level of 0.05) correspondingly, which indicates the rejection of the null hypothesis of normality of the data. Indeed, the histogram plot with an estimated normal curve on the right side of Fig. 4.8 implies that the distribution of the difference BA scores right-skew, which supports the hypothesis that the subject-dependent models provide more accurate detections that subject-independent ones regardless of the learning models. To measure the statistically significant difference between the performance of subject-dependent and subject-independent models, the non-parametric statistical test for paired samples is employed, which is the one-sided Wilcoxon signed rank test as believes that the subject-dependent is statistically more accurate in stress detection than subject-independent one. The null and alternative hypotheses of the test are stated as follows:

- H_0: $M = 0$ (The capability of stress and non-stress pattern discrimination of both subject-dependent and subject-independent models is the same.)

- H_a: $M > 0$ (The subject-dependent model manages to distinguish stress/non-stress patterns statistically more accurately than the subject-independent one)

In the above hypothesis statements, the variable M indicates the median of the data, where most of the data in the distribution of the population are scattered around this value in the non-normal distribution. Conducting the Wilcoxon signed-rank test on the data, the p-value obtained from the statistic test is 2.2×10^{-16}, which is smaller than the significance level of 0.05. This implies that the null hypothesis could be rejected at the confidence level of 95%, meaning that the subject-dependent models are more statistically accurate in distinguishing stress/non-stress patterns compared to the subject-independent ones. The 95% confidence interval is larger than 0.1733, indicating the median of the population inferred from the observed samples is larger than 17.33%. This value shows that the magnitude of the statistical performance between the subject-dependent and subject-independent models is huge. Therefore, it is reasonable for us to conclude that the subject-dependent models could achieve higher evaluation scores in detecting stress moments than the subject-independent stress detection model as used by the current generation of consumer-grade wearable devices.

4.5 Conclusion

Dataset	**Extra-Trees**	**LR**	**Deep-Fusion**
WESAD	95.78 ± 13.04	96.57 ± 7.71	95.41 ± 6.49
AffectiveROAD	80.69 ± 11.64	75.90 ± 12.66	79.47 ± 8.96
Cognitive-Load	94.42 ± 9.47	92.09 ± 9.08	91.93 ± 8.86
DCU-NVT-EXP2	91.04 ± 8.84	83.48 ± 10.20	83.20 ± 9.98

Table 4.5: Balanced Accuracy Score of the Subject-dependent Stress Detection Model trained on Wrist-worn Data.

Through the experiments conducted to address research 1, I gain insights that only the difference between the performance (balanced accuracy) of the subject-dependent stress detection model using wrist-worn data and the one using chest-worn data is statistically significant. However, the magnitude of the performance difference ranges around 0.16% to 4.37%, which is not huge. It can be considered as the trade-off between the low error caused by the low granularity level of the recorded signals and the potential application of using consumer-grade wearable devices for health monitoring in daily life. In addition, through inferential analyses from my experiments, I manage to prove that the optimal approach to building a stress detection model using low-resolution physiological signals is to employ the subject-dependent training method. Indeed, as can be seen from Table 4.5, the mean balance accuracy score of the subject-dependent stress detection models in the four datasets that I use in my experiment ranges from 75.90%±12.66% to 96.57%±7.71%. The model that achieves the best evaluation score in most cases is the Extra-Trees model with the mean balanced accuracy score varying from 80.69% ± 11.64% to 95.78% ± 13.04%. This shows that it is possible to detect mental stress status with high accuracy by using the low-resolution physiological signals captured from consumer-grade wearable devices. Owing to the high accuracy of detecting stress in the laboratory experiment, it can be concluded that the most optimal approach to training a stress detection model using low-resolution physiological data is to employ the subject-dependent method.

4.6 Chapter Summary

In this chapter, I addressed Research Question 1 which evaluates how successfully low-resolution physiological signals recorded from consumer-grade wearable devices can be applied to the stress detection problem compared to the use of the same kinds of signals captured from clinical-grade devices. According to the experimental results, the stress detection model using low-resolution physiological signals was proven to perform as well as the one using high-resolution, and the subject-dependent stress detection model was more accurate than the subject-independent one. These findings, which are the two main contributions of my research in this chapter, are crucial to providing the conclusion that low-resolution signals from consumer-grade wearable devices are good enough to build a stress detection model and the most optimal training approach is the subject-dependent method. This conclusion is important as it facilitates further research in the same field to be conducted to provide more useful applications and meaningful insights into an individual's life.

Chapter 5

Stress Detection Model in Unconstrained Environment using Consumer-grade Wearable Device Data

5.1 Introduction

In this chapter, we address research question 2, which is **how successfully can stress detection models using low-resolution physiological signals from consumer-grade wearable devices be applied for lifelog data to detect moments of stress?**

I define the term “moment” mentioned in this research question to be the point of time captured by the lifelog camera each 30 seconds (by default in Narrative Clip 2 device), which is also referred to as a single lifelog image. According to the literature review in chapter 2 and the lab-based experimental stress detection results from chapter 4, I know that mental stress can be detected using physiological signals of consumer-grade wearable devices with high accuracy. However, the performance of stress detection applied to real-life scenarios is still vague. One of the reasons is that in most laboratory experiments to collect stress datasets, the researcher would try to restrict the tasks and the activities that the participants can do during the experiment session. For instance, in the data collection protocol of the WESAD [9], AffectiveROAD [10] and Cognitive Load [2] dataset, the stress tasks are designed to avoid high-load physical activities by restricting the participants to do the tasks while standing and sitting at one place. The main reason for this design is to prevent the appearance of physiological responses elicited by physical stress which is the same as the ones elicited by mental stress. However, this is not the case in real life when a user is exposed to different environments under different conditions. Though previous works in real-life stress detection with consumer-grade wearable provides initial results of the performance of the stress detection model in real life, more analyses should be done to understand how the model uses the features to detect stress that results in either low or high evaluation score. Additionally, multiple experiments should be conducted to determine an optimal approach to building personal stress detection in the wild as well as examine the possibility of applying the lab-based model to in-the-wild data to verify if additional data from users are required to enhance the performance of the model.

Therefore, to answer research question 2, I conduct a longitudinal study and collect more stress data in unconstrained environments in daily life using lifelog sensors and consumer-grade wearable devices. As it is a longitudinal study, potential candidates are the ones who join in my laboratory experiment described in section 4.2.1.4. Based on the collected dataset, I conduct the following experiments:

- Evaluate the performance of the lab-based subject-dependent stress detection model of each participant applied directly to real-life scenarios.

- Conduct feature analyses to understand how the models manage to detect stress from the statistical features extracted from the physiological signals by using multiple learning model interpretability techniques. Thereby, I can gain insights into what factors impact the success of detecting moments of stress in daily life.

- Re-train the personal stress detection model on lifelog stress data, compare its performance to the lab-based model applied to lifelog data (real-life data) and analyze the difference between the results.

5.2 Experiment Configuration

In this section, I would describe the configurations for the experiment of assessing the performance of stress detection models in an unconstrained environment using low-resolution signals recorded from wearable devices. These include the description of data collection for the experiment, the learning models used in building stress detection models in the unconstrained environment, and the evaluation metrics used for the models' performance assessment.

5.2.1 Dataset

From the dataset that I collected in the laboratory environment presented in section 4.2.1.4, I conducted a longitudinal study on a subset of participants in the previous experiment based on the following conditions:

- The participants should join the experiment willingly with the insurance of the data privacy and data governance from the researcher and principal investigators.
- As lifelog sensors need to be worn properly to provide good-quality data, the participants are required to have prior experience with lifelog data to mitigate the risk of data failure during the longitudinal study.
- As physical stress and mental stress can elicit the same physiological signals [47], the participants are expected to work in an office environment without many physical tasks or relax without moving around much during the day.
- The daily working task loads of the participants should be stressful at a certain level to capture enough stressful moments.

One of the challenges which is the main difference between laboratory data and in-the-wild data is that I cannot control the behaviors and stress moments of the participants by limiting the task that they have to do. I could only estimate the number of stressful moments that I can capture from the participants via the description of their job on a normal working day. Therefore, an interview with the participant was carried out to understand the participant's daily working tasks so that participants satisfying all those conditions could be selected properly. The fourth and third conditions are crucial to ensure that there are enough stress moments to conduct experiments and evaluate the model.

In total, three participants who satisfied my proposed conditions were invited to join the lifelog stress data collection. The mean age of these three participants was 28.00(±2.83). These selected participants were either postgraduate students or staff at Dublin City University (DCU) who either had prior background knowledge of lifelog data or research in lifelogging topics. After securing the consent from participants, the participants were provided with a description of the longitudinal study protocol of the data collection. In detail, the participants were required to record the lifelog data for six days while maintaining their daily life routine as normal. At first, the participants were introduced to the definition of mental stress, its common signs and symptoms following the information provided by the mind.org.uk organization. This source is good for participants to refer to when doing annotation as it does not only provide the causes of stress but also specifies the signs and symptoms of stress clearly. Particularly, some emblematic signs and symptoms of mental stress that were delivered to the participants were:

- Irritable, angry, or impatient.
- Overburdened or overwhelmed.
- Anxious, nervous, or afraid of past events or upcoming events.
- Like your thoughts are racing and unable to switch them off.

- Unable to enjoy yourself.
- Depressed.
- Neglected or lonely.

Then, two lifelog sensors which were the Narrative Clip 2[1] and the Empatica E4 wristband[2] were provided to the participants to capture lifelog images every 20 seconds and record fine-grained physiological signals continuously. The participants were required to wear these sensors for at least six hours per day. At the end of each day, the participants were required to self-evaluate their stress level during the day subjectively and determine the stress moments by looking at the lifelog images of that day. I believe that the participants can precisely distinguish the ranges of stress events during the day by reminding the location, time, and visual cues shown in the lifelog images on the same day. My assumption is that the annotation of stressful moments on a day should be completed on the same day as the participants might be confused or overwhelmed by new stressful events on the next day, which might lead to unreliable annotation. To facilitate the annotation process, I utilised the event marking function of the Empatica E4 wristband. Whenever the participants started to feel stressed, they needed to press the button on the Empatica E4 wristband once to mark the beginning of the stressful event and then pressed that button again to mark the end of the event. Thereby, the participants could be reminded of the estimated time range of the stress event during the annotation process at the end of the day to increase the reliability of the annotations.

For the annotation process, as I only focus on detecting mental stress, I asked the participants to remove the moments that they do physical exercising, walk, run, or do physical activities. Additionally, all the moments containing private information were removed by the participants when annotating. Details of the number of stress and non-stress moments of each participant on each day are shown in Table 5.1. It

[1] http://getnarrative.com/

[2] https://www.empatica.com/research/e4/

can be seen from Table 5.1 that the gap between the percentage of stress and non-stress moments of User 1 is small (8.84%) while it is extremely large for User 2 and User 3 (90.86% and 59.10%). The difference might be due to the daily activities of User 1, which was more stressful as User 1 is a staff working in the university following a strict schedule while User 2 and User 3 are Ph.D. students, whose working time can be flexible. The nature of the job of User 1 increases the participant's privacy concern when recording lifelog data. Therefore, many images considered private were removed by User 1. These reasons lead to a small gap between the percentage of stress and non-stress moments. In total, I recorded 3557 stress moments (lifelog images) from three participants, which account for 24.83% of the moments captured in the dataset. The number of non-stress moments was 10769, which accounts for 75.17% of the moments in the dataset.

	User 1		**User 2**		**User 3**	
	Normal	Stress	Normal	Stress	Normal	Stress
Day 1	611	80	615	110	725	56
Day 2	145	493	714	57	480	325
Day 3	199	780	741	21	1127	274
Day 4	390	241	1027	49	819	0
Day 5	501	563	967	0	450	246
Day 6	100	166	881	0	277	96
Total	1946	2323	4945	237	3878	997
Perc (%)	45.58	54.42	95.43	4.57	79.55	20.45

Table 5.1: Details of the number of stress and non-stress moments of each participant on each day.

5.2.2 Learning Models

Despite the evaluation of five different conventional Machine Learning models on stress detection tasks, in this experiment, I only use two main conventional Machine Learning models among these five to facilitate the application of model interpretation techniques to explain the results, which are the Extra-Trees classifier and Logistic Regression. In theory, though the model interpretation technique that I employ can be applied to most of the learning models, its disadvantage is the expensive computational cost that results in the consideration of applying it to all the models that I use in Chapter 4. The two conventional models that I choose are both efficient in stress detection as experimented with laboratory-based data and have low computational cost to apply model interpretation techniques as optimal solutions are specifically developed for these models [79, 80]. The importance of the features can also be inferred from these models so that important data distribution analysis can be performed. Apart from these two core models that are used for the model's performance explanation, I also employ my proposed Deep Fusion model to evaluate my proposed state-of-the-art stress detection model [11] applied in the unconstrained scenarios.

5.2.3 Evaluation Metrics

5.2.3.1 Balanced Accuracy

As the number of the binary classes in the real-life stress dataset is also imbalance as can be inferred from Table 5.1, I still employ the balance accuracy metric to evaluate the performance of the in-the-wild stress detection model. The balance accuracy metric is the average of the accuracy of the model for each class in the entire dataset, which is one of the effective metrics to evaluate the performance of the model based on the research conducted by Straube et al. [117]. Balanced accuracy is the main evaluation metric that I use to assess the performance of stress detection models in my experiment.

5.2.3.2 Precision and Recall

Though the main goal is to evaluate the model based on the objective metrics – balance accuracy, I also evaluate the percentage of stress moments that the model manages to detect and how precise the stress moments are detected by the model by employing the precision and recall metrics. For instance, suppose that the precision and recall scores of my in-the-wild stress detection model are $x\%$ and $y\%$, these numbers indicate that the model manages to detect $x\%$ of moments precisely and this $x\%$ of the stress moments accounts for $y\%$ of the actual stress moments of the targeted user in the dataset. From these evaluation scores, I can have an intuition of how successfully the model can detect stressful moments in real-life.

5.3 Experiment Results

This section reports the experimental results of using multiple approaches to building in-the-wild stress detection models as well as doing feature analyses to determine a potential approach to building stress detection models for lifelog data/data in the unconstrained environment.

5.3.1 Laboratory Stress Detection Model Applied to Lifelog Data

	Deep Fusion			**Extra-Trees**			**Logistic Regression**		
	BA	**P**	**R**	**BA**	**P**	**R**	**BA**	**P**	**R**
User 1	98.47	99.52	97.80	96.91	96.71	100	**99.33**	99.27	100
User 2	84.49	85.42	100	**100**	100	100	90.09	90.64	98.73
User 3	64.75	73.42	86.73	**88.58**	92.24	91.13	69.83	76.26	91.67

Table 5.2: The performance of stress detection model with laboratory dataset (DCU-NVT-EXP2).

As mentioned in the introduction of this chapter, at first, I evaluate the performance of the laboratory stress detection model applied directly to lifelog data. To do so, I report the performance of the targeted user's personal stress detection model in the constrained environment (laboratory) to understand how good the model is. As can be seen in Table 5.2, the best stress detection model for User 2 and User 3 employs the Extra-Trees classifier while the one for User 1 utilizes Logistic Regression. In detail, for User 1, the personal stress detection model using Logistic Regression achieves the balance accuracy of 99.33% while it is only 96.91% and 98.47% for the Extra-Trees and Deep Fusion models respectively. The gap between the performance of these models is not large compared to the ones of User 2 and User 3. For User 2 and User 3, the highest balance accuracy scores can be achieved by using the Extra-Trees model, which is 100% and 88.58%

respectively. The gaps of the balance accuracy scores between the best models and other ones of User 2 and User 3 are approximately 12.71% (±2.80%) and 21.38%(±2.63%). Based on the results from Table 5.2, I can conclude that stress detection models can identify stress moments in a real-time manner with high accuracy (varies from 88.58% to 100%) in the laboratory environment.

However, when laboratory stress detection models are applied to lifelog data, the evaluation scores decrease significantly. It can be seen from Table 5.3 that using laboratory stress detection to detect stress moments in daily life results in poor performance. The balance accuracy scores of all the models decrease approximately 64.01% (±1.04%), 45.23% (±7.03%), and 20.03% (±8.96%) for User 1, User 2, and User 3 respectively in average. The laboratory best-performed personal stress detection model for each user decline 65.38%, 54.38%, and 32.43% for User 1, User 2, and User 3. To understand the reason why this significant drop in the performance of the stress detection model applied to lifelog data exists, I need to interpret the models to gain insights of which features affect the decision of the model when making detections.

	Deep Fusion			**Extra-Trees**			**Logistic Regression**		
	BA	**P**	**R**	**BA**	**P**	**R**	**BA**	**P**	**R**
User 1	**34.69**	40.93	42.36	34.05	40.52	42.40	33.95	38.77	36.25
User 2	**47.21**	3.73	23.63	45.62	2.82	13.50	46.06	2.67	10.55
User 3	53.19	22.72	50.75	**56.15**	23.84	68.81	53.71	23.05	52.26

Table 5.3: The performance of laboratory stress detection model applied to the whole lifelog data to detect stress moments.

5.3.2 Analysis on Feature Impacts to the Model Inference

To analyze the difference between the performance of stress detection models in the laboratory and the ones applied to the daily-life scenario, I apply Shapley Additive Explanation (SHAP) to compute the contributions of the features used by the model for inference, which is Shapley values. The Shapley values are computed from the laboratory train data as I would need to have an insight into how the model learns to use the feature to detect stress. In my analysis, I only analyze the laboratory personal stress detection model that has a high balance accuracy score on the laboratory dataset (DCU-NVT-EXP2). In this case, it is the Logistic Regression model for User 1 and the Extra-Trees classifier for User 2 and User 3. I employ LinearSHAP to explain the Logistic Regression model while TreeSHAP is used to explain the ensemble Extra-Trees classifier. One problem when interpreting the Logistic Regression model is the multicollinearity. Though this problem does not affect the performance of the model, it can mislead the analyses of model interpretation. Therefore, I recursively remove the multicollinear features used to train the model by employing the Variance Inflation Factor (VIF) with the VIF's threshold value of 5. The average features' impacts (top-10) on the personal laboratory stress detection model for each user is illustrated in Fig. 5.1 with feature names shown in Table 5.4.

In overall, it can be inferred from Fig. 5.1 that the features computed from skin temperature (max, min, mean, slope) and heart activity (time-domain, frequency domain, and non-linear domain) are most important in distinguishing between stress and non-stress states, while the galvanic skin response's feature impacts weakly to the detection of the models. The feature impact in Fig. 5.1 is important since it provides a list of core features among 72 ones on which the stress detection model relies on to detect stress moments. Based on this information, I plot multiple box plots to visualize the distribution of the core features in laboratory train data and the ones in lifelog data for intuitive comparison. For an intuitive comparison between the data distribution in the laboratory and in daily life, I focus on comparing the median and the interquartile range (IQR). The IQR is the spread of the data, which is a measure of statistical dispersion where most of the data points scatter around [135]. By comparing the IQR, I can have an intuitive insight into the difference in the distribution of the core features in the two datasets.

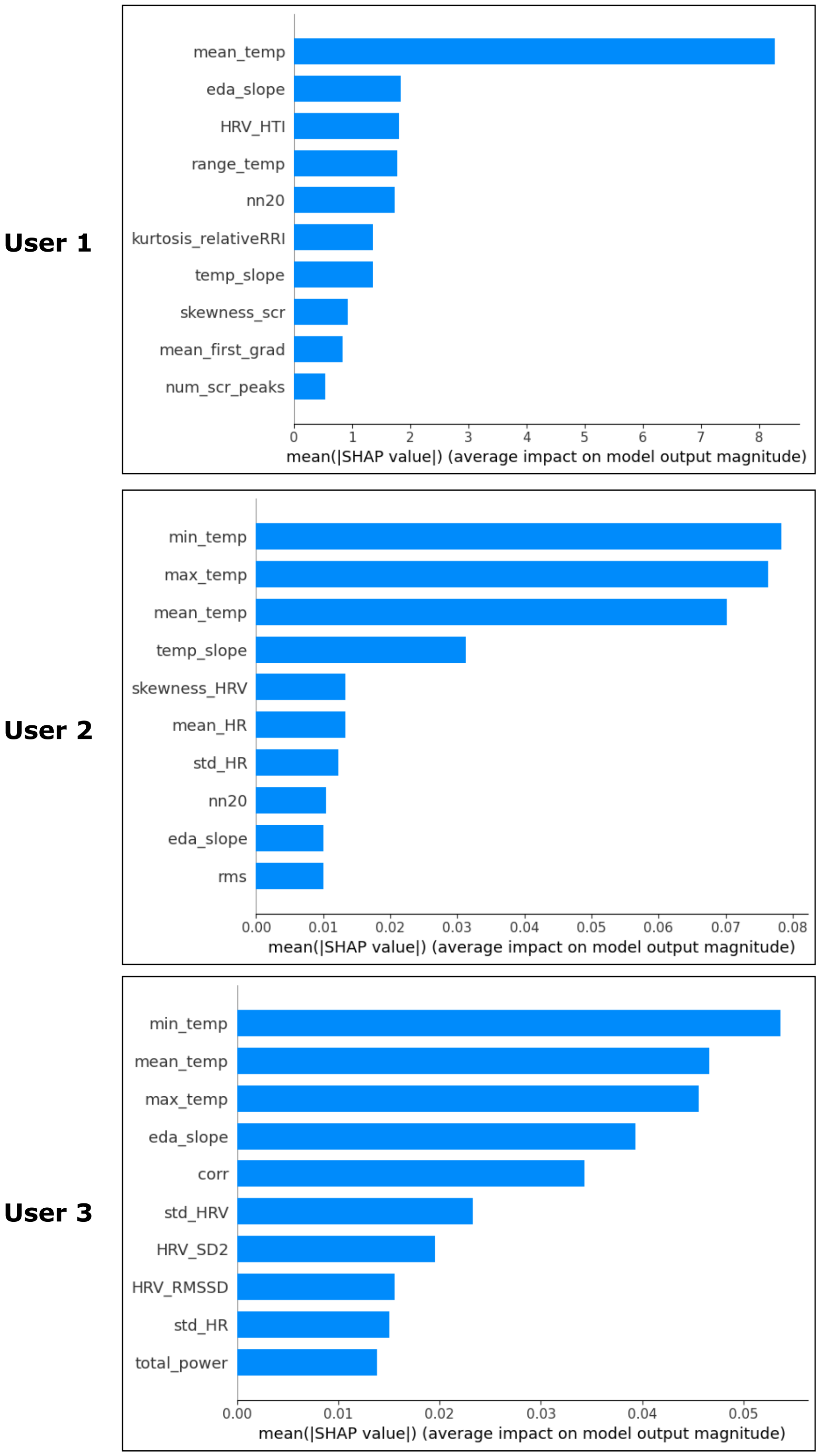
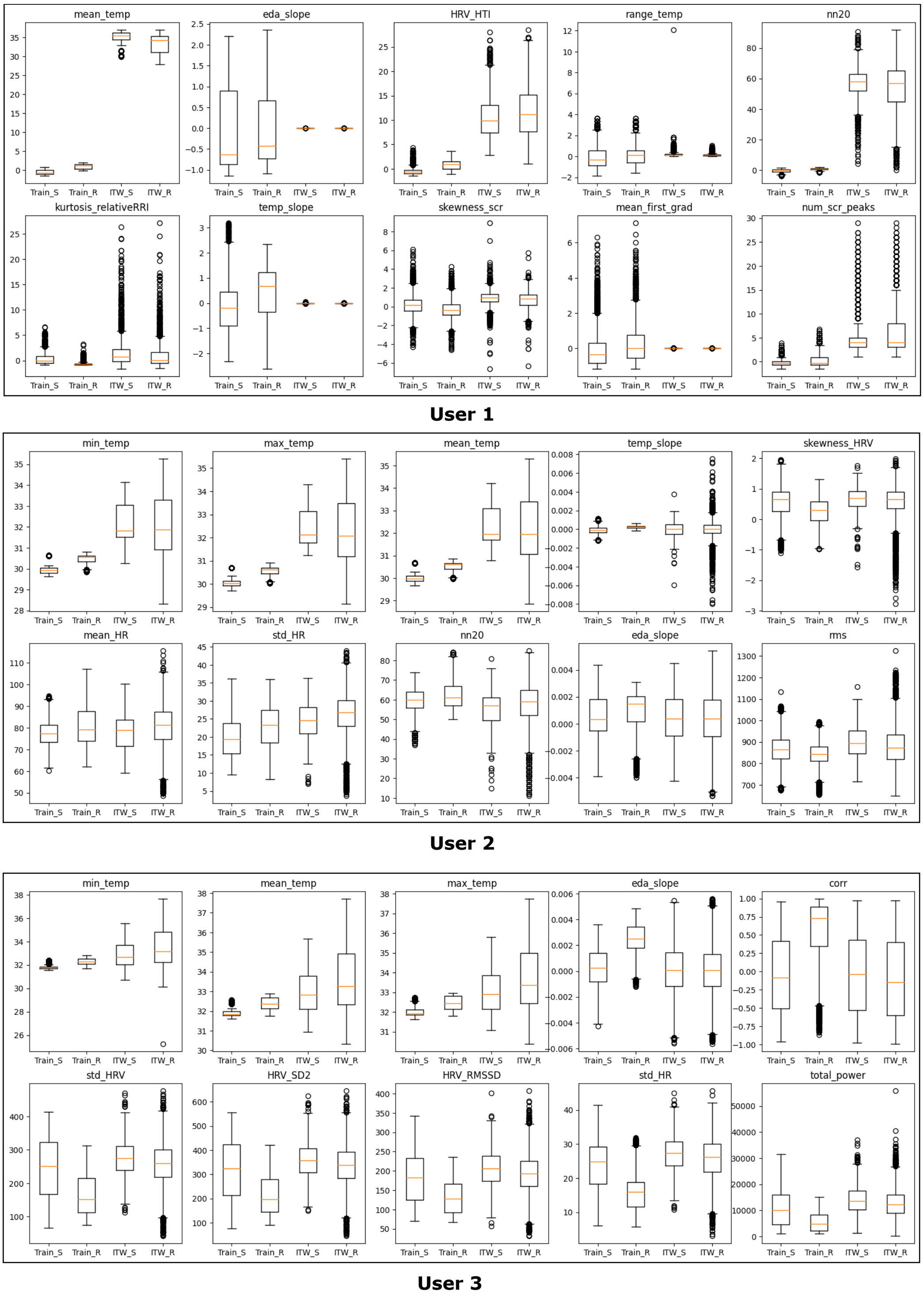


Figure 5.1: Average features' impacts (top-10) on personal laboratory stress detection model for each user computed by aggregating Shapley values (feature contributions).

Signal	Feature Name	Description
EDA	eda_slope	Slope of the EDA
	corr	Correlation btw SCL and time
	eda_dynamic_range	Dynamic range of EDA
	mean_first_grad	Mean of the first derivative of the SCR
	area_of_response_curve	Area under the identified SCRs
	num_scr_peaks	# identified SCR peaks
	skewness_scr	Skewness of SCR
	mean_scl	Mean of the SCL
	std_scl	STD of the SCL
	max_eda	Max of the EDA
	mean_eda	Mean of the EDA
	std_eda	STD of the EDA
	APSC	Normalized average power of the SCR
BVP	nn20	# HRV intervals differing more than 20ms.
	HRV_pNN20	% HRV intervals differing more than 20ms.
	HRV_pNN50	% HRV intervals differing more than 50ms.
	kurtosis_relativeRRI	Kurtosis of the relative RR intervals
	RMSSD_relativeRRI	RMSSD of all interval of differences between adjacent RR intervals.
	skewness_HRV	Skewness of the HRV
	HRV_RMSSD	RMS of all interval of differences between adjacent RR intervals.
	total_power	Total power of the freq. components from VLF to HF
	mean_HR	Mean Heart Rate
	std_HR	Standard Deviation of Heart Rate
	rms	Root Mean Square of the HRV
	HRV_HTI	HRV Triangular index
	HRV_SD2	Long-term poincare plot descriptor of HRV
	std_HRV	STD of HRV
TEMP	max_temp	Max of the skin temperature
	min_temp	Min of the skin temperature
	mean_temp	Mean of the skin temperature
	temp_slope	Slope of the skin temperature
	range_temp	Range of the skin temperature
	std_temp	STD of the skin temperature

Table 5.4: The description of the features shown in Fig. 5.1. Abbreviation: RMS = Root Mean Square, VLF = Very Low Frequency, HF = High Frequency. # = Number of, % = Percentage of.



User 1

User 2

User 3

Figure 5.2: Distribution of important features that the lab-based model uses to make detections. The lab-based data's distribution is prefixed by "Train" while the lifelog one is prefixed by "ITW". "S" stands for stress while "R" stands for "Relaxed"

As can be seen in Fig. 5.2, the distribution of the stress and non-stress patterns in lifelog data are different from the ones in the laboratory environment. For instance, for User 1, the decrease in the skin temperature (mean) is the indication of a physiological response to the stressors in the laboratory while it is the opposite in lifelog data. The heart activity features such as nn20, HRV_HTI, and kurtosis_relativeRRI are not distinguishable for stress and non-stress patterns anymore in the lifelog data of User 1 due to different data distribution between different types of environments. In addition, the range of values of the features in lifelog data is wider than the one in the laboratory environment (e.g. nn20, kurtosis_relativeRRI), which implies that the data in the constrained environment is not enough to capture the stress patterns in real-life. The same situation as in User 1 happens to the data of both User 2 and User 3. Indeed, for User 2 and User 3, the drop in skin temperature (max, min, and mean) also indicates stress patterns in the laboratory, however, the distribution of these features is so vague between stress and non-stress class that it is hard to identify the difference from intuitions via box plots. In short, from the visualization of core feature distributions and analyses, I could draw three main conclusions about the difference in the performance of directly applying the laboratory stress detection model to lifelog data as follows:

1. The core features' distributions of the data in the constrained environment are opposite to the ones in lifelog data or do not cover the range of the data in the lifelog data. For instance, for User 1, the decline in the skin temperature is the indication of stress responses in real-world settings while the increase of this signal in the laboratory environment is the relaxed status of an individual.

2. The range of values of the core features is wider than the one in the laboratory stress data indicating that the data in a constrained environment is too strict that it is not enough to capture all the stress patterns in real life.

3. The core features' distributions learned by the laboratory stress detection

model in lifelog data is vague, which indicates that the model might use other sets of features to detect the stressful moments in daily life. Especially, the skin temperature depends much on the environment/room temperature which requires to train the model again with new data in real life under different environmental constraints.

5.3.3 Lifelog Stress Detection Model

Due to my analyses of the difference in the performance of the laboratory stress detection model in lifelog data, I re-train the stress detection model using the same configuration for the laboratory one to learn the new distribution of the physiological signals in the lifelog data. The training set of each user composes of the stress and non-stress moments of the first three days while the test set consists of the data of the last three days. This data splitting results in 1353 stress moments and 955 non-stress moments in the training set for User 1, 188 stress moments and 2070 non-stress moments in the training set for User 2, and 655 stress moments and 2332 non-stress moments for User 3. For the test set, it is 970 stress moments 991 non-stress moments for User 1, 49 stress moments and 2875 non-stress moments for User 2, and 342 stress moments and 1546 non-stress moments for User 3. For each stress/non-stress moment, I extract the corresponding statistical feature for the 60-second window size as in Section 4.2.2. As the size of the training set is small, I augment the training set by extending the stress/non-stress event duration by 5 seconds. As can be seen from Fig. 5.3, for each stress/non-stress moment X, I assign the moments from $X - 5$ seconds to $X + 5$ seconds with window shift of 1 second

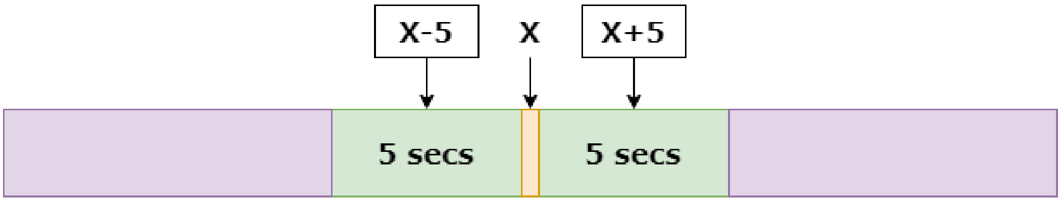


Figure 5.3: Proposed augmentation method to increase the number of data to train lifelog stress detection model.

	Deep Fusion			**Extra-Trees**			**Logistic Regression**		
	BA	**P**	**R**	**BA**	**P**	**R**	**BA**	**P**	**R**
User 1	42.05	42.54	49.83	**43.91**	44.77	58.66	43.64	43.50	46.91
User 2	39.83	0.28	4.08	**42.40**	4.87	6.12	40.90	0	0
User 3	59.15	27.92	42.69	59.61	24.97	57.31	**61.69**	30.15	47.95

Table 5.5: The performance of the laboratory stress detection model applied to the test set of the lifelog data to detect stress moments.

	Deep Fusion			**Extra-Trees**			**Logistic Regression**		
	BA	**P**	**R**	**BA**	**P**	**R**	**BA**	**P**	**R**
User 1	56.19	53.68	79.79	52.10	50.62	93.40	**56.27**	53.86	77.63
User 2	47.14	1.31	20.41	**62.80**	2.85	61.22	48.23	1.47	24.50
User 3	54.10	21.67	40.94	**62.51**	25.05	73.98	54.38	21.69	43.57

Table 5.6: The performance of the re-trained stress detection model applied to the test set of the lifelog data to detect stress moments.

the same annotation as the targeted moment X.

As can be seen from Table 5.5 and Table 5.6, I improve the balance accuracy score by 11.19% (±8.04%) on average considering the difference between the best model in the laboratory applied directly into lifelog data and the model re-trained with real-life data. The best lifelog stress detection model for User 1 is the Logistic Regression model with a balance accuracy score of 56.27% while the best lifelog stress detection model for User 2 and User 3 is the Extra-Trees classifier with an evaluation score of 62.80% and 62.51% correspondingly. The difference in the model type of each user is reasonable as the model is built personally for each user, therefore, the best model among a list of models should be chosen to detect stress accurately. The improvement shows that the stress detection model in daily life should be trained with data recording from unconstrained environments as the one recorded in the laboratory environments does not capture all the conditions in real-life scenarios due to the strict requirements in the lab-based experiment design (e.g. the user is required to perform stress task in the room).

Apart from the balance accuracy score which reflects the capability of distinguishing between stress and non-stress moments, the precision and recall of the best lifelog stress detection model also improve. It can be seen from Table 5.6

that although the recall of these models is considerably higher for all users (approximately 70.94% ± 7.04%), the precision of the models is quite low (approximately 27.25% ± 20.88%). The low precision score of the model indicates that the model tends to detect false positive results. This demonstrates that using learning models to detect stress in daily life requires more research and experiments on other complex models to increase the precision of the models.

In addition to the investigation of the performance of the model, I also try to interpret the model of each user to have insights into the physiological responses of each individual to stress stimuli in real life. I apply SHAP to compute the feature importance as well as the feature contribution in the learning models on the test set to understand how the model makes decisions on stress moment detection so as for us to gain insights into the features of physiological signals that result in stress responses. In lifelog stress detection problem, the model that achieves the best balance accuracy score is the Extra-Tree classifier for User 2 and User 3 while it is the Deep Fusion model for User 1, which can be seen in Table 5.6. I can easily apply TreeSHAP to the Extra-Tree classifier to interpret the ensemble tree-based model while it is harder to apply general SHAP techniques to interpret the Deep Learning model (e.g. KernelSHAP, DeepSHAP [79]) due to high computational cost. For User 1, I use LinearSHAP to interpret the Logistic Regression model to gain insights into the feature interaction for stress detection of User 1. I apply the Variance Inflation Factor (VIF) with VIF's threshold value of 5 to remove the highly correlated features to avoid the multicollinearity problem misleading the model interpretation. From the violin plot in Fig. 5.4, I could draw some insights about the characteristics of each participant corresponding with stress response in daily life learned by the model. The y-axis in Fig. 5.4 is the top-10 features that the model uses to detect stress and is ranked in the descending order of feature importance. Details of the feature names shown in Fig. 5.4 are described in Table 5.4. From the summary plot shown in Fig. 5.4, some observations and insights that I gained are as follows:

- A high feature value of skin temperature is an indication of having stress status

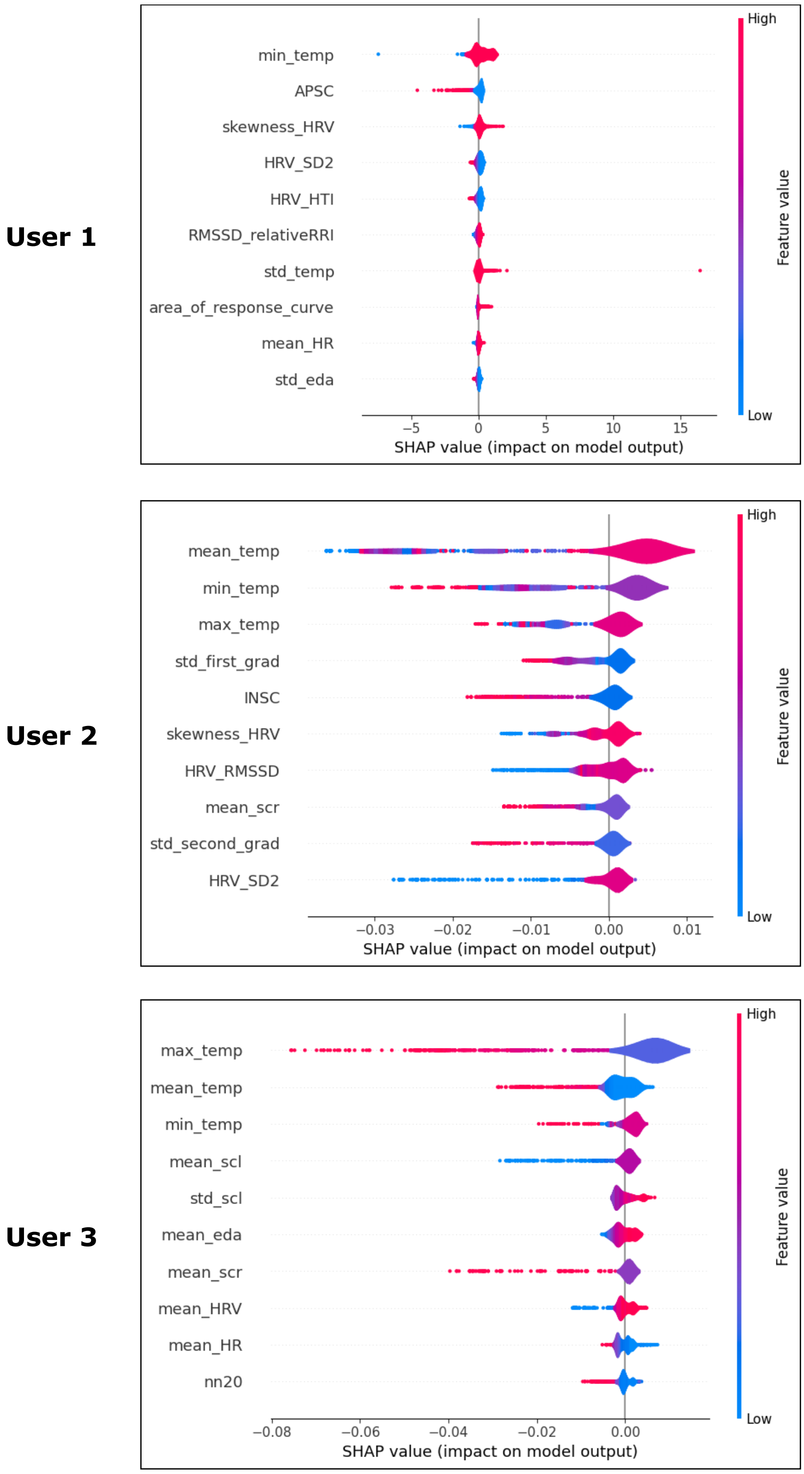


Figure 5.4: SHAP summary plot shows top-10 feature contributions on lifelog stress detection model for each user based on the feature value. Feature points to the left of the 0 SHAP-value on the x-axis indicate relaxation detection and vice versa.

for User 1 and User 2 while a low value of skin temperature is an indication of non-stress status for User 3. The difference between the two groups of users is gender. User 1 and User 2 are female while User 3 is male. This may suggest that physiological responses to stress stimuli are different for different gender. However, more research would be needed to verify my comment.

- For heart activity, the heart activity indicators of the chance of having stress status are different for each user implying that building a personal stress detection model for each individual in daily life is a good practical approach due to the inter-difference in the physiological responses of each person. For instance, User 1 would be detected stress when she has a high heart rate while it is the opposite for User 3. A high value of the skewness of the heart rate variability signals indicates stress responses in females (User 1 and User 2) but it is not for males (User 3).
- For galvanic skin response, I can only draw a simple conclusion that a high value of the raw electrodermal activity signal (mean_eda) could lead the model to be prone to yield stress detection. As the statistical features of Skin Conductance Response (SCR) and Skin Conductance Level (SCL) are not intuitive to explain as well the contribution of these features is only significant for User 3, I cannot provide accurate insights about the effect of the SCR and SCL to the detection of stressful moments.

5.4 Discussion and Contribution

From my experiments and analyses, I show that laboratory stress detection cannot be applied directly to real-life scenarios due to the difference in the data distribution of the core features used to detect stress in laboratory environments. The data in real life is more diverse than the one in the laboratory as multiple different contexts and stimuli are involved in the stress situation instead of a single context caused by a single stress task as in the constrained laboratory environment.

I realize by re-training the model on the lifelog stress data, I could achieve a better model with higher detection accuracy than the laboratory stress detection model. Though the balance accuracy of the model is acceptable, further research should be conducted to improve the precision of the model. Based on the feedback from the participants joining the experiment, through my methods of annotating stress moments by pressing the marking button on the wearable watch and looking back at the lifelog images, they still have difficulties when labeling stress events. In particular, there are some events in which they forget to press the button on the wearable watch to mark the end of the stress event. In these cases, they need to look at the lifelog images to determine their mental status at that time. It would be easy when the stress event with a specific scenario happens once during that day. However, it is not the case. Stress events can be mistaken for non-stress ones due to the same visual information, scenario, and context. For instance, for the same person that the participant meets at different times of the day, the participant can be either stressed or relaxed depending on the context of the meeting. In addition, the lack of content and topic of the conversation owing to privacy issues is the main reason that makes the participants confused when annotating stressful moments. One participant admits that the participant stress factor usually comes from external sources (finance, deadlines, social interactions, etc.). It is hard to capture all of these external factors in real life as all aspects of an individual, including private events, must be recorded in the lifelog data archive. This violates the expectations in my research ethics approved by the Research Ethics Committee (REC) of Dublin City University. Therefore, I have not had the chance to investigate if the lifelog stress detection model could actually improve if more details of the participant's life are used to train and infer instead of depending mainly on the biometrics data and lifelog images. Further research with new experiments would be needed to verify this statement.

I also gain insights about the characteristics of each participant corresponding with stress response in daily life learned by the model. However, as the sample size of my longitudinal lifelog stress dataset is small, I would propose further research work on my approach and feature analyses to gain more insights into different aspects affecting physiological responses to mental stress stimuli. In short, the main contributions of my works are as follows:

1. I collected longitudinal stress lifelog dataset from potential candidates from the previous experiment described in Section 5.2.1 and conducted a proof-of-concept study to evaluate the performance of the stress detection model applied to lifelog data.

2. I showed that applying the laboratory stress detection model to detect stressful moments in daily life would lead to inaccurate detection results and conducted feature analyses using model interpretation techniques to explain why it could not work.

3. I showed that re-training the personal stress detection model on stress lifelog data is a good approach to building a lifelog stress detection model and analysed the characteristics of physiological responses to the mental stress stimuli of each user by interpreting the re-trained model.

5.5 Chapter Summary

In this chapter, I addressed Research Question 2 by investigating the application of the stress detection model in lifelog data after having determined a proper approach to building a stress detection model in a constrained environment with high accuracy in Research Question 1. To do so, I collected the lifelog stress data via a longitudinal study for six days from potential candidates that join my previous laboratory experiment. These participants were selected following strict criteria to ensure the quality of the lifelog stress dataset. In total, I conducted three experiments including evaluating the performance of the lab-based subject-dependent stress detection model of each participant applied directly to real-life scenarios, conducting feature analyses to understand how the models detect stress using multiple learning model interpretability techniques, and evaluating the performance of personal stress detection model re-trained on lifelog data. According to the experimental results, I proposed that the personal stress detection model should be re-trained on the lifelog data instead of directly applying the lab-based stress detection model since the strict constraint in laboratory experiment protocol limits the variety in the individual physiological responses to mental stress under different conditions. Though I build a personal lifelog stress detection model for each user with high balance accuracy and recall, further research should be done to improve the overall performance of the lifelog stress detection model by adding more context to the train data.

Chapter 6

Impact of the Stress-Moment Filtering Function in the Lifelog Retrieval System

6.1 Introduction

In this chapter, I address research question 3, which is **How can biometric and visual data be used in a lifelog interactive retrieval system to retrieve stress-related moments?**

In terms of the use of biometrics data during the retrieval process, we propose to use the mental stress information inferred from the lifelog stress detection model built in Chapter 5 as an indicator for filtering. My hypothesis is that the stress-moment filtering function can be used to increase the performance of the retrieval system when dealing with stress-related/emotional-related queries by removing irrelevant moments and re-ranking the ranked list. To answer this research question, I do the literature review, analyze important functions of state-of-the-art lifelog retrieval systems, and develop a state-of-the-art lifelog retrieval system to compare the performance of the retrieval system with and without the integrating stress-moment filtering function. In short, research question 3 can be addressed by providing answers to these sub-research questions:

- **Research Question 3.1**: How can the state-of-the-art lifelog interactive retrieval system be designed and developed?

- **Research Question 3.2**: How much benefit can be derived from adding the biometric stress filters to an interactive lifelog retrieval system in a conventional retrieval task?

6.2 The Development of the LifeSeeker System

Based on the literature review of state-of-the-art retrieval systems conducted in Chapter 2, a state-of-the-art interactive lifelog retrieval system named LifeSeeker is developed to address Research Question 3.1, which is **How can the state-of-the-art lifelog interactive retrieval systems be designed and developed?** The development of the LifeSeeker system was a collaborative effort between my colleagues and me, with each of us making equal contributions [12, 136, 137, 138]. Specifically, my contributions included proposing the weighted bag-of-words technique, improving the free-text search through the vision-language model, and supporting the design of the user interface and the implementation of the database-indexed structure for filtering.

LifeSeeker system inherits most of the core functions from state-of-the-art systems including:

- Textual concept-based search and filter based on the visual concepts and the metadata extracted from the lifelog image (moment).
- Visual similarity search (Query by Image/Example) to retrieve similar moments based on the matches of visual cues and visual features between the input lifelog image and other lifelog images in the corpus.
- A straightforward user interface and user interaction that supports temporal moments browsing and optimizes the browsing display panel to search for a targeted moment efficiently.

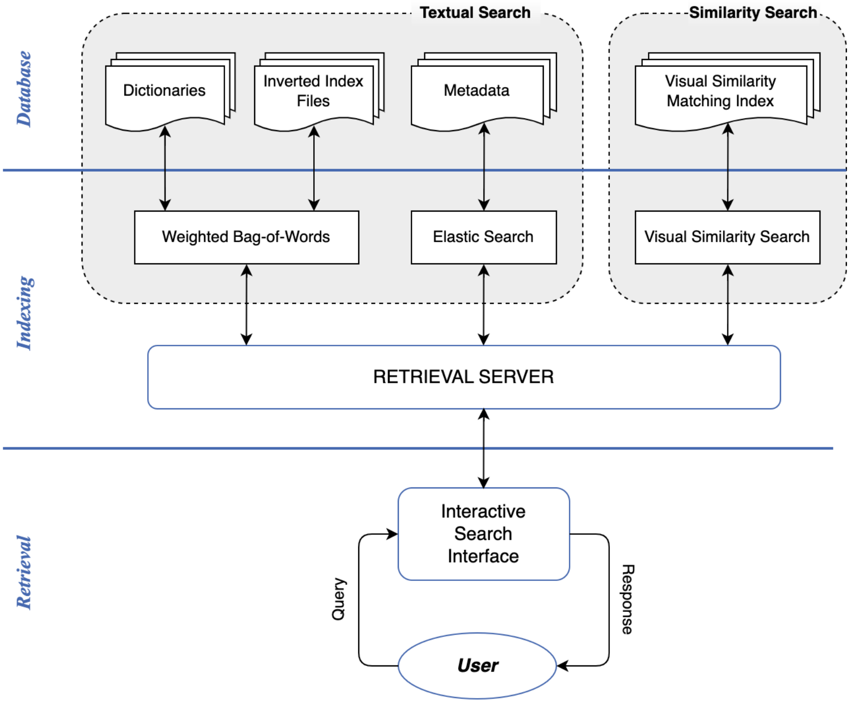


Figure 6.1: The System Architecture of LifeSeeker [12].

LifeSeeker was firstly designed as a concept-based interactive lifelog retrieval system that relies on the analysis of both visual and non-visual content in LSC'21. It is then upgraded into a free-text search in LSC'22. Section 6.2.1 describes the system architecture of the concept-based version of the LifeSeeker– shown in 6.1 – while Section 6.2.2 introduces the enhancement of the LifeSeeker to upgrade it into a free-text search interactive retrieval system. The concept-based version of the LifeSeeker was in the top 3 of the best interactive lifelog retrieval system in LSC'21 while the free-text search version of the LifeSeeker system in LSC'22 was awarded the second-best interactive lifelog retrieval system title whose achieved a competitive overall evaluation score compared to the top-ranked system.

6.2.1 LifeSeeker in the Lifelog Search Challenge 2021: A Concept-based Interactive Lifelog Retrieval System

6.2.1.1 System Architecture Overview

Fig. 6.1 illustrates the architecture of LifeSeeker, which consists of three components: a database, a retrieval engine, and an interactive search interface. The database contains four different types of indexed metadata which are used in three different retrieval methods. In detail, the Textual Search component relies on three dictionaries (metadata-concepts, location, time) and an inverted-index file that maps the moment id, which is also the lifelog image id, in the format of `YYYYmmdd_HHMMSS_000` where `Y`, `m`, `d`, `H`, `M`, `S` is the year, month, day, hour, minute, and second of the moment respectively; with its corresponding dictionary terms. On the other hand, the Elastic Search engine[1] is used to index and retrieve both the metadata provided by the organizers combined with other metadata extracted from the collection. These include place categories and place attributes extracted from PlacesCNN [139], visual object concepts extracted from YOLOv4 [140] pre-trained on COCO dataset [141] and Bottom-up Attention Model [142] pre-trained on Visual Genome dataset [143], and text extraction data (OCR) with other visual concepts extracted using Google Vision API[2] and Microsoft Vision API[3] respectively. For the Similarity Search component, the ranked list of visually similar images of a specific-target photo obtained from the Visual-Similarity Search algorithm (Section 6.2.1.6) is stored in the MongoDB database to boost the speed of the visual similarity search. The LifeSeeker's retrieval server is developed using the Django framework[4] which plays the role of a middleware supporting the communication between the client-side requests (user interface and interaction) and different retrieval modules. In general, the core retrieval engine consists of two components: Textual Search and Similarity Search. In my system, the Similarity

[1] https://www.elastic.co

[2] https://cloud.google.com/vision/docs/ocr

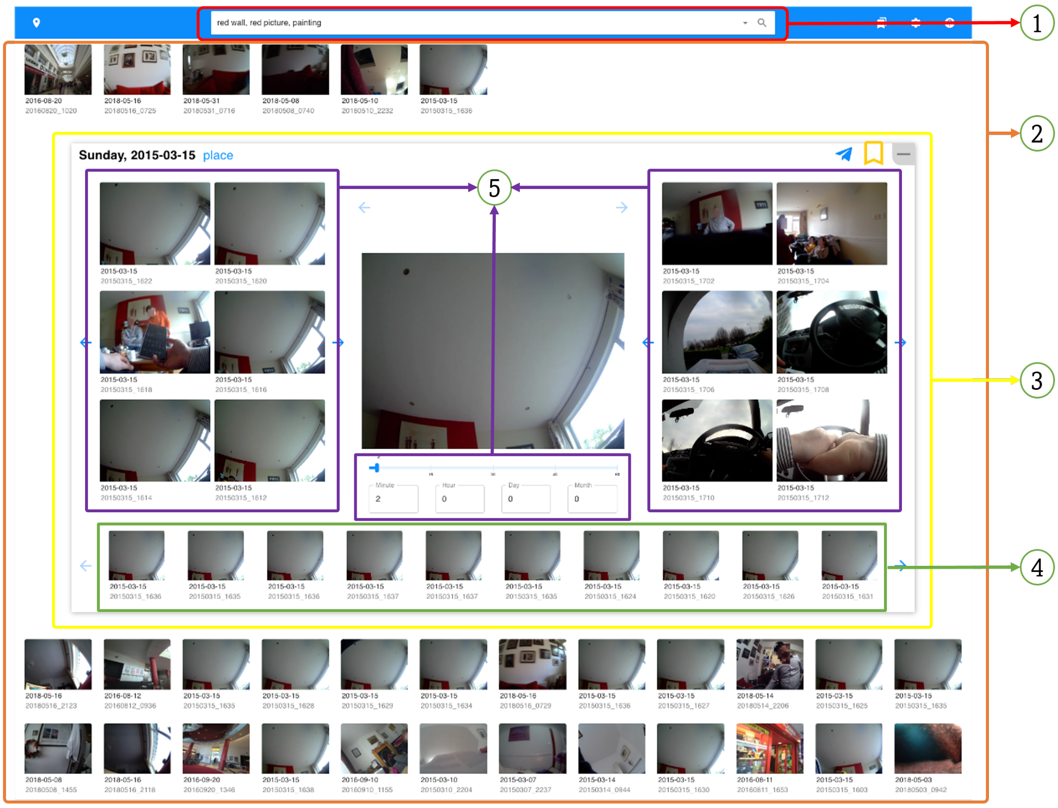
[3] https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision

[4] https://www.djangoproject.com

Search component contains only one module which is the Visual-Similarity Search. This module takes an image as input and returns a ranked list of photos that have similar visual patterns (e.g. objects' edges/angles/orientations) to the input. The Textual Search component of LifeSeeker consists of two retrieval modules: Weighted Bag-of-Words and Elastic Search, which are also two core retrieval modes of the system. The Weighted Bag-of-Words module is an alike free-text search that relies on the matching between the terms automatically parsed from the description of the life moment and the metadata concepts to rank the relevant documents (or relevant lifelog images) based on the cosine similarity score. In contrast, Elastic Search requires the user to input detailed query terms, which are manually parsed from the description, following a pre-defined syntax to form a boolean query. In short, the decision of which component and which module to use for retrieval is determined by the retrieval server based on the input type (full sentence, query terms in a pre-defined syntax, or images). These core retrieval systems are selected based on the core retrieval components of the state-of-the-art lifelog retrieval systems that participated in the Lifelog Search Challenge including the MyScéal [8], Memento [111], VRLE [7], and vitrivr [99] systems. The design of the retrieval server aims to support the simplicity of the user interface and user interaction. The interactive search interface of the LifeSeeker is a web-based application developed using ReactJS framework[5]. The main components of the LifeSeeker's user interface (UI) are the free-text search box, the vertically-scrollable panel displaying the retrieval results, and the detailed box showing related contents of the selected image including visually similar moments, preceding moments, and successive ones. The user provides the query to the system using the search box by entering either query terms following a pre-defined syntax (described in Section 6.2.1.5) or a full sentence describing the desired life moment. Matched lifelog images are then displayed on the vertically-scrollable panel for further browsing or scanning interaction.

[5] https://reactjs.org

6.2.1.2 User Interface and User Interaction



red wall, red picture, painting

Sunday, 2015-03-15 place

Minute 2 Hour 0 Day 0 Month 0

Figure 6.2: The Interactive User Interface of the LifeSeeker Retrieval System.

User Interface: The interactive user interface of LifeSeeker is composed of three main components (Fig. 6.2), which are the free-text search box **(1)**, the vertically-scrollable panel displaying a ranked list of retrieved moments **(2)**, and the moment-detail box **(3)**. The vertically-scrollable panel (**2**) shows a ranked list of retrieved moments obtained from the query submitted in the free-text search box (**1**). Each item in the panel is a square box displaying the lifelog image with minute id (an example of the minute id is shown in the List 6.1) and captured date with format `YYYYmmdd_HHMM` and `YYYY-mm-dd` respectively where `Y`, `m`, `d`, `H`, and `M` denotes the year, month, day, hour, and minute correspondingly. The vertically-scrollable panel displays five rows of images, where each row shows at most 12 lifelog images with date and time information. I show the date and time of the moment as it is considered to be one of the most important pieces of information of a lifelog moment that cannot be recognized visually from the image. For the moment-detail box (**3**), the lifelog image of the selected moment is shown in the middle of the box.

Apart from the lifelog image of the selected moment located in the middle of the moment-detail box (**3**), there are two other essential components; these are the horizontal panel displaying visually-similar images and the temporal browsing panel in component (**4**) and (**5**) respectively. The horizontal panel (**4**) displays at most 10 images, which are visually similar to the selected moment. The temporal browsing panel (**5**) consists of two horizontally-scrollable panels on both the left and the right of the lifelog image in the middle of the moment-detail box (**3**) showing a sequence of moments that happen before (left) and after (right) the selected moment. In addition, a before-and-after time-range controller is placed under the selected moment. This time-range controller is used to adjust the temporal range the user would like to explore from the selected moment. By adjusting the time delta, images before and after the target photo can be adjusted to be temporally nearby or further apart. It is my conjecture that the target memory is usually retrieved by connecting the previous memories, which form a path that leads to the piece of memory during the recalling process. The query “Eating fishcakes, bread and salad after preparing my presentation in PowerPoint" would be a good example to clarify my conjecture. Querying for the moment when the lifelogger had fishcakes, bread, and salad is not enough to uniquely identify the desired moment among a thousand of the same ones without considering the temporal-activity information. Therefore, the temporal browsing panel (**5**) is an essential part of the user interface when dealing with temporal-related queries.

User Interaction: The flow of user interactions can be illustrated via four steps:

1. The user inputs the query into the search box (**1**). The query can be in the form of a full sentence describing the moment or in the form of a sequence of terms following the syntax (6.1). The search box (**1**) also supports the term auto-completion to facilitate the user inputting the query.

2. The user can either scan or browse the ranked list of relevant images displayed on the vertical-scrollable panel (**2**).

3. Any moment for which the user wants to investigate if it is the answer to the query, the user has two options to browse it further; these are left-clicking on the image to open the moment-detailed box (**3**) or hovering on the image while pressing the X key to enlarge the image.

4. In case the user opens the moment-detailed box (**3**) for further browsing, the user can use the temporal browsing panel (**5**) to view the previous/after moments of the selected one by horizontal scrolling the panel as well as adjusting the time delta to view the temporal nearby or further apart moments.

These four steps are performed repeatedly during the search process. It is worth noting that the search box is also capable of performing a filter search by inputting the query terms following the syntax 6.1.

6.2.1.3 Indexing

Since the lifelog dataset is constructed by gathering data from multi-modal sensors (i.e. wearable cameras, biometric devices, GPS, phones, computers), the **Indexing** module requires various sub-modules, each responsible for processing one modality of the lifelog data. I categorize the lifelog data into the followings:

- **Time**: This is one of the most important pieces of information that helps to narrow the search space greatly. For example, knowing when (morning, afternoon, evening) the moment happened can filter out nearly two-thirds of the original amount of images.

- **Location**: Location can be viewed as a summary of a lifelogger in terms of where they were on a daily basis, which might imply the sequence of activities that the lifelogger does throughout the day. It is also useful for adding more context to the query generation process to find more relevant moments (i.e. if finding moments that the lifelogger was eating a sushi platter, the user can add "Asian restaurant" as part of the LifeSeeker input query to obtain more accurate results).

- **Visual concepts**: Images captured from the wearable camera are information-rich, as moments are illustrated in detail (i.e. how the surroundings look like, who appears in that moment, and which objects are seen). However, computers cannot perceive images as humans do.
- **Other metadata**: Apart from the aforementioned data sources, there are other modalities provided in the dataset (e.g. Activity, Biometrics). However, this metadata can be indexed instantly into the search engine without further processing.

6.2.1.4 Weighted Bag-of-Words

I implement a customized Bag-of-Words algorithm that serves both free-text search and filtering. Firstly, three dictionaries are generated from the pre-processed metadata that includes time, location, and visual concepts:

- **Time dictionary**: consists of the information of the month (from January to December), weekday (from Monday to Sunday), and part of the day (early morning, late morning, afternoon, etc.).
- **Location dictionary**: consists of semantic location names, countries, cities, and place categories obtained from PlacesCNN.
- **Visual-concept dictionary**: consists of multiple object labels extracted from deep-vision model pre-trained on MS-COCO and Visual Genomes dataset as well as the ones obtained from Microsoft Vision API.

Listing 6.1: A sample metadata for a lifelog moment generated by the Indexing module

```
"_id": "b00000049_21i6bq_20150225_062023e",
"minute_id": "20150225_0620",
"image_path": "LSC/2015-02-25/b00000049_21i6bq_20150225_062023e.jpg",
"date": "2015-02-25",
"local_time": "06:20",
"day_of_week": "wednesday",
"month": "february",
"year": 2015,
"part_of_day": "early morning",
"gps": [53.3892, -6.15827],
"activity_type": "idle",
"lat": 53.3892,
"lon": -6.15827,
"location_name": "home",
"location_type": "home",
"city": "Dublin",
"country": "Ireland",
"location_address": ["howth junction cottage", "kilbarrack upper",
    "raheny-greendale ed", "dublin 5", "dublin", "county dublin",
    "leinster", "d5", "ireland"
],
"place_category": ["home theater", "television room"],
"microsoft_tag": ["indoor", "desk", "wall", "television", "floor",
    "furniture", "computer monitor", "computer"],
"yolo_concept": ["tv"],
"visual_genome": ["black clock", "white door", "white knob", "black
    television", "wooden floor", "cardboard box", "shelf"
],
"ocr": "stevshark mks"
```

The dictionaries are refined using the **nltk library**[6] so that stop-words are removed. In addition, I also manually filter the dictionaries to remove meaningless terms as well as one-character terms and non-alphabetic characters. Unlike the traditional Bag-of-Words, in my algorithm, I do not consider inverse document frequency (IDF) weighting since occurrences of terms in the corpus are all considered to be equally important; the dictionary weight is used instead. This is because the IDF weighting would reduce the significance of some common terms which frequently appear in the lifelog annotation corpora such as week-of-day, part-of-day, and semantic location labels. The dictionary weight (w) is variable and changes to reflect the importance

[6] https://www.nltk.org

of each dictionary. In this case, Let w_{time}, w_{loc}, and w_{vc} be the weights of the time, location, and visual-concept dictionaries respectively; then I set $w_{\text{time}} > w_{\text{loc}} > w_{\text{vc}}$. The time information is essential to identify a specific moment and filter the results. In addition, the location dictionary is considered more important than the visual concept one (vc), as it would be easier to navigate to the desired moment if the location is given in the query. These weights are combined into a vector and then are multiplied into the L2-norm term frequency vector of the query to amplify the time and location when computing cosine similarity between the query vector and the L2-norm term frequency vector of images in the archive to retrieve relevant images.

6.2.1.5 Elastic Search

A query into Elastic Search can be constructed by combining one or more *query clauses*[7] of various types, thus users can form very complex queries to define how Elastic Search retrieves data. In order to reduce the query analysis time and allow flexibility in controlling how each keyword should behave when retrieving lifelog moments (i.e., which should be used for matching images and which should be used for filtering purposes only), I introduced a syntax-based query mechanism as below:

$$\texttt{<CONCEPTS> ; <LOCATION> ; <TIME>} \tag{6.1}$$

where each query part (`<CONCEPTS>`, `<LOCATION>` and `<TIME>`) corresponds to a category outlined in Section 6.2.1.3. A syntax-based query can be formed by specifying keywords in each part in Syntax 6.1. For instance, the following query is a valid input to LifeSeeker:

```
blue paintings, wall ; conference room ; after 12pm
```

[7] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html

The searching process in Elastic Search mode is done by employing the *query string query*[8] to match <CONCEPTS> and <LOCATION> keywords, while the *term query*[9] and *range query* mechanisms were used to filter images using the given <TIME> keywords.

6.2.1.6 Visual Similarity Search

For the visual-similarity search, I utilize the Bag-of-Visual-Words model to transform visual features into a vector representation for the K-Nearest Neighbors algorithm. In general, the algorithm of the Bag-of-Visual-Words model is similar to the traditional Bag-of-Words one used in textual information retrieval except for the creation of the dictionary, which is usually known as the visual codebook. Each item in the visual codebook is called the visual word instead. In the Bag-of-Visual-Words model, the visual codebook can be constructed using the K-Means Clustering approach that clusters the descriptors extracted from Scaled-Invariant-Feature-Transform (SIFT) [144], the Oriented FAST and Rotated BRIEF (ORB) [145], and Speeded Up Robust Features (SURF) [146]. It is worth noting that an image can have many descriptors, therefore, resulting in having many visual words. The choice of the parameter K in the K-Means Clustering algorithm determines the number of visual words in the visual codebook. In LifeSeeker, I use 256-dimensional descriptors of ORB features as inputs for the visual codebook generation process. Due to the huge number of descriptors in a large-scale dataset, I employ the Mini-batch K-Means Clustering described in [147] to reduce the computation cost while gaining asymptotic clustering results compared to the conventional K-Means Clustering approach. The Mini-batch K-Means Clustering is performed with 50 iterations. The value of K used in my case is 4096 as I consider this number of visual words is enough for a visual-similarity search. All the remaining steps including vector quantization and

[8] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html

[9] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-term-query.html

similarity computation are performed as in the traditional Bag-of-Words model. For computing similarities between images, the cosine distance function is employed instead of the Euclidean distance function.

6.2.2 LifeSeeker in the Lifelog Search Challenge 2022: Interactive Lifelog Retrieval System with Enhanced Free-Text Search

The textual search in the version of LifeSeeker that I used to compete in the Lifelog Search Challenge 2021 uses Weighted Bag-of-Words, which is still a concept-based search since it parses the input sentence into terms to match with the ones in provided dictionaries. In the latest version of LifeSeeker, the Weighted Bag-of-Words function is replaced by the CLIP (Contrastive Language–Image Pre-training) model [148], which encodes lifelog images and queries into the same latent space to compute similarity [138]. As the CLIP model uses pre-trained weights to encode the images into a latent space, the Visual Similarity Search using the Visual-Bag-of-Words model with ORB features is also replaced by the CLIP model. Cosine similarity is used to compute the distance between the embedding vectors of the lifelog images and the embedding vectors of either input queries or input images. These are significant enhancements from the previous version of LifeSeeker that result in the performance improvement of the LifeSeeker in Lifelog Search Challenge 2022, which are chosen following the similar function in the state-of-the-art retrieval systems including MyScéal [112] and Memento [149].

6.2.3 Experiment Result

LifeSeeker was benchmarked in the fourth and the fifth annual Lifelog Search Challenge (LSC'21 and LSC'22). The ultimate goal of the challenge is to retrieve relevant lifelog images that match a given query as fast as possible with penalties applied for wrong submissions. The challenge was conducted in an interactive manner, which means that there is a user using the system to perform the search and submit the image that they think best illustrated the query. For each query, the score [150] of one LSC participant retrieving the correct answer at a time t is calculated as follows:

$$S_i = \max\left(0, M + \frac{D-t}{D}(100-M) - W * 10\right) \tag{6.2}$$

where M refers to the minimum score earned, D denotes the query's duration and W represents the number of wrong submissions for each query. M and D are set to 50 and 300, respectively. The score is linearly decreased until the minimum score (50) within the 300-second period. The final score is taken by subtracting each negative submission by 10 points. A participant gets a zero score when the time for the query is over but a positive answer is not found.

Table 6.1: The official scores of the top-5 teams in LSC'21

Team name	**# queries solved**	**Total score**	**Precision**	**Recall**
Myscéal [109]	19	**1604.31**	82.61	82.61
SomHunter [151]	19	1566.32	67.86	82.61
LifeSeeker [12]	**20**	1556.02	76.92	**86.96**
Voxento [152]	18	1466.87	**85.71**	78.26
Memento [111]	16	1238.49	59.26	69.57

Table 6.2: The normalized score of the top-5 teams for each task in LSC'22

Team name	**Ad-hoc**	**KIS**	**QA**
Myscéal [112]	98	**100**	**100**
LifeSeeker [138]	**100**	88	96
Memento [149]	66	92	79
FIRST [153]	51	95	75
Voxento [154]	49	87	56

For the LSC'21, as can be seen from Table 6.1, LifeSeeker is the third best performing system in the challenge which achieved a total score of 1556.02. Though LifeSeeker solves 20 out of 23 queries, LifeSeeker only achieves third place due to the penalties caused by the wrong submissions made by the user. Apart from the official score from the LSC'21, I also evaluate the performance of my system using the precision and recall score. As can be inferred from Table 6.1, LifeSeeker gets the highest recall score of 86.95% while achieving a competitive precision score of 76.92%. This result implies that LifeSeeker is capable of retrieving the desired information up to 87% of the time (4.35% more than that of the second-highest recall system). The precision of the LifeSeeker system is 8.8% lower than the Voxento [152], which achieves the highest precision score. The main cause of the lower precision was the number of incorrect submissions that the user made during the challenge when solving tasks under time pressure. For the LSC'22, apart from evaluating the lifelog retrieval systems by the Known-Item-Search (KIS) search task, the performance of these systems is also evaluated by the Ad-hoc search task and the Question-Answering task. For the Ad-hoc search task, the user needs to find as many images that match the general description of the query as possible under the time constraint while the question-answering task requires the user to find an image that shows the answer to the question with only one attempt. As can be seen from Table 6.2, the LifeSeeker gains the highest score in the Ad-hoc search task which means that the improvement that I apply to the system still achieves the highest recall score. The score in the Known-Item-Search (KIS) task of the LifeSeeker is 88, which is lower than the scores of Myscéal, Memento, and FIRST systems due to the penalties from the wrong submissions made by the user and unsolved queries in the task. This means that as the system manages to retrieve the correct results in the ranked list, the user does not submit the correct image due to the time pressure of the KIS task in the challenge, which results in a low precision score in the KIS task. Indeed, without the time pressure, the user has more time to verify the image to submit the correct one with only one attempt, which is proven by the second-highest score (96/100) in the Question-Answering task in the challenge. In short, through the two benchmarking Lifelog Search Challenges, I manage to develop one of the best interactive lifelog retrieval systems that has a competitive performance with other state-of-the-art systems in the same competition, which can be considered one of the current state-of-the-art interactive lifelog retrieval systems. Through my research of core features to develop the state-of-the-art lifelog retrieval system, I select important core functions from them to integrate into my system that would help find the targeted images that best match the description of the query effectively. Thereby, research question 3.1 is addressed.

6.3 Evaluation of Interactive Lifelog Retrieval System Integrating Stress-Moment Filter

In this section, I conduct experiments to evaluate the impact of the lifeloggers' mental stress information on the performance of the state-of-the-art lifelog retrieval system to provide the answer to Research Question 3.2, which is **How much benefit can be derived from the addition of the biometric stress filters to an interactive lifelog retrieval system in a conventional retrieval task?**. The experimental results of these experiments could help provide the conclusion if the mental stress information is beneficial for the conventional retrieval process task. Thereby, the hypothesis defined in Section 1.4 that the stress-indexed information enhances the performance of the state-of-the-art lifelog retrieval system could be proven.

6.3.1 Experiment Settings

I use the lifelog stress dataset that I collected in the longitudinal study described in Section 5.2.1. The dataset comprises lifelog images and physiological signals captured from consumer-grade wearable devices including blood volume pulse (BVP), galvanic skin response (GSR), and skin temperature (TEMP) recorded by three users during six days. The reason for the choice of the number of participants and the duration of the study is also described in Section 5.2.1 The mental stress information of the lifelog moment is inferred from the most accurate personalized lifelog stress detection model for each user described in Table 5.6 in Section 5.3.3. It is worth noting that the moment that I mention in my work is referred to as the lifelog image with its metadata described in Section 6.2.1.3. In particular, the Logistic Regression model is used for mental stress inference for User 1 while the Extra-Trees classifier is employed for User 2 and User 3.

I employ LifeSeeker in this experiment as it is one of the state-of-the-art lifelog retrieval systems in the benchmarking Lifelog Search Challenge. Since my purpose is to evaluate the impact of the stress-moment-filtering function using the mental stress information inferred from the lifelog stress detection model, I only use stress-related/emotional-related queries in my experiment. For other types of queries, my lifelog retrieval system – LifeSeeker– was evaluated via the benchmarking Lifelog Search Challenges described in Section 6.2.3.

The stress-related/emotional-related queries are constructed by the researcher based on the descriptions of both stressful and relaxed moments made by the user during the data annotation process in the longitudinal study for lifelog stress data gathering. The stress-related/emotional-related queries consist of two parts. The first part of the query is the description of the moment based on the visual cues, the temporal information, the scenario, the environmental attributes and categories, activities, and the social interaction of the user with others (e.g. human-object interaction, social human interaction such as talking with friends, arguing with others, etc.). The second part of the query, which is also the last sentence in the query, is the description of the feeling or mental stress status of the user at that moment. An example of the query is *"I gave a presentation in a competition under a strict time limit and received some tough questions from the judges. I felt nervous since I had to perform well to secure the win for my team."*. The list of queries that I use in my experiment for each user is described in Table A.1.

I integrate stress information into the elastic search and employ the pre-defined syntax described in Section 6.2.1.5. The mental stress information is considered as the additional metadata (other metadata) of the lifelog moments, therefore, it belongs to the field `<CONCEPTS>` in the pre-defined syntax in Section 6.2.1.5. An input example to the system to filter the results is:

```
relaxed ; home ; Monday
```

In total, I conduct two experiments to assess the impact of the stress-moment filtering function on the performance of the lifelog retrieval system:

- **Experiment 1 – Non-interactive Lifelog Retrieval System with Stress-Moment Filter**: To evaluate the performance of the non-interactive lifelog retrieval system with the stress-moment filter, I use the first part of the query which describes the context of the moment including the visual cues, environmental attributes and categories, activities, and social interaction as the input for the free-text search of the LifeSeeker system. Then, the stressful/relaxed information is extracted from the second part of the query to apply to filter the retrieved ranked list. I use the Recall (R@50) and the Average Precision (AP@50) of the top 50 results in the ranked list to evaluate the performance of the non-interactive lifelog retrieval system for every single query. The overall evaluation of the performance of the non-interactive lifelog retrieval system with and without the stress-moment filter function is assessed by the mean average precision (mAP@50) and the mean recall (mR@50) of the top-50 results in the ranked list for all the queries. The reason that I employ the retrieval score at the top-50 results in the ranked list for evaluation is to replicate the process of browsing a list of results in the interactive mode.

- **Experiment 2 – Interactive Lifelog Retrieval System with Stress-Moment Filter**: To evaluate the performance of the interactive lifelog retrieval system with the stress-moment filter, I restrict functions of the LifeSeeker system to assess the impact of using the stress-moment filter during the retrieval process. These restrictions are as follows:

1. The user must search the moments by using the free-text search and filter by elastic search system only. When searching, the user is only allowed to make a free-text search followed by a filter condition, which is considered a valid search. If the user wants to search again, the user needs to repeat this valid search process again. The user must use the stress-moment filter and submit at least one correct answer during the search process in the experiment with LifeSeeker integrated stress-moment filter function.
2. The user can view temporal images of a targeted moment to verify if it matches the query's temporal description. However, the user is not allowed to submit temporal images.
3. Only 100 images are shown to the user on the vertically-scrollable panel.

In addition to the restrictions of the functions of the LifeSeeker system that the user can use in this experiment, I also limit the search time of each query to 180 seconds (3 minutes) and change the evaluation metrics employed to assess the performance of the interactive lifelog retrieval system. The change in the evaluation metrics is due to two reasons. The first reason is that the user in this experiment is also the owner of the data, meaning that the user has a prior perception knowledge of the event in their life based on the description. In addition, the queries and the ground-truth are constructed by having the user describe the events. Therefore, I have to take into account the bias that the user has about the ground-truth. Due to this bias about the ground-truth, the evaluation metrics used in the LSC'21 and LSC'22 as well as the precision score would not be appropriate in this experiment anymore since it is easy for the user to have high evaluation score by finding one example that best matches the description. Therefore, I evaluate the performance of the interactive lifelog retrieval system in this experiment using only the mean recall score and mean search time (in seconds). Overall evaluation of the performance of the systems with and without the integration of the stress-moment filter function is done by the mean recall score of all the queries and the mean search time. Furthermore, the full description of the query is shown to the user before the search time, which is similar to the lifelog challenges in the NTCIR [27] and the ImageCLEF Lifelog [26, 28]. Without these changes, conditions, and restrictions, I cannot assess the contribution of the stress-moment filter function to the performance of the interactive lifelog retrieval system properly. I conduct the experiment with the same set of queries twice on two different versions of the interactive lifelog retrieval system – with and without the stress-moment filter. As I need time for users to forget the way they search for the results in the previous experiment, I set the gap between each experiment session to be two weeks. While not using new queries, such an approach of session gap has been used before in such memory retrieval experiments [155]. In addition, supported by the findings of Caterina Cinel et al. that a session gap of two hours could cause the forgetting effect when joining a new memory retrieval experiment [155], the two-week gap is good enough for the user to have the retrieval-induced forgetting effect [156] to join the next retrieval experiment.

6.3.2 Experiment Results

6.3.2.1 Experiment 1 – Evaluation of the Non-interactive Lifelog Retrieval System integrated Stress-Moment Filter

Table 6.3 demonstrates the evaluation scores for each query of the non-interactive LifeSeeker retrieval system with and without stress-moment filter function. The accuracy column in Table 6.3 shows the percentage of correctly predicted stress/non-stress labels inferred from the lifelog stress detection model. These labels are used as the ground-truth in this experiment despite the potential false-positive stress detection moments made by the model. On average, the mean error rates of the stress/non-stress labels in the ground-truth for User 1, User 2, and User 3 are 21.85%, 19.74%, and 28.96% correspondingly. It can be inferred from Table 6.4 that the stress-moment filter actually increases all the mean recall scores at the top-50 results

User ID	**Query ID**	**Normal System**		**With Stress-Moment Filter**		
		AP@50	**R@50**	**Accuracy**	**AP@50**	**R@50**
User 1	Q1	0	0	100	3.13	2.38
	Q2	60.94	35.71	85.71	76.50	28.57
	Q3	5.88	16.67	33.33	10.08	33.33
	Q4	7.59	4.92	100	14.92	6.56
	Q5	32.14	9.52	95.24	64.29	9.52
	Q6	0	0	100	4.17	1.43
	Q7	74.31	50	50	74.98	50
	Q8	4.59	18.18	72.73	18.10	27.27
	Q9	20.83	22.22	44.44	5.26	11.11
	Q10	59.10	57.90	100	60.26	57.90
User 2	Q1	14.29	7.69	92.31	50	7.69
	Q2	4.67	20	40	8.70	20
	Q3	0	0	91.67	0	0
	Q4	5.51	18.18	54.55	7.02	36.36
	Q5	58.07	50	100	60.57	60
	Q6	2.63	11.11	55.56	46.86	33.33
	Q7	19.92	36.36	100	19.92	36.36
	Q8	20.67	45.83	95.83	47.31	70.83
	Q9	45.14	27.27	72.73	62.01	36.36
	Q10	25	100	100	45	100
User 3	Q1	0	0	80	3.47	40
	Q2	60	13.33	90	44.07	20
	Q3	23.87	14	98	29.63	20
	Q4	46.63	51.85	40.74	40.79	33.33
	Q5	84.28	100	100	94.04	100
	Q6	0	0	66.67	2	11.11
	Q7	81.42	45.35	39.40	76.19	26.74
	Q8	25.14	18.52	62.96	25.45	14.82
	Q9	0	0	61.91	0	0
	Q10	57.20	29.23	70.77	55.66	12.31

Table 6.3: Detailed evaluation scores for each queries of the non-interactive version of the LifeSeeker system with and without stress-moment filter.

User ID	**Normal System**		**With Stress-Moment Filter**			
			Model Detections		**Manual Labels**	
	mAP@50	**mR@50**	**mAP@50**	**mR@50**	**mAP@50**	**mR@50**
User 1	26.54	21.51	**33.17**	**22.81**	33.17	22.81
User 2	19.59	31.64	**34.74**	**40.09**	34.74	40.09
User 3	**37.86**	27.23	37.13	**27.83**	37.13	27.83

Table 6.4: Overall evaluation scores of the non-interactive version of the LifeSeeker system with and without stress-moment filter.

(mR@50) of the non-interactive lifelog retrieval system for all the cases. The increase of the mR@50 is approximately 3.45% on average. The stress-moment filter function also increases the mean average precision score at the top-50 results (mAP@50) of the non-interactive lifelog retrieval system by 6.63% for User 1 and 15.15% for User 2. However, for the queries of User 3, I recognize a slight decrease (approximately 0.73%) in the performance of the non-interactive version of LifeSeeker. These results suggest that the stress-moment filter function can reduce the performance of the non-interactive lifelog retrieval system when the detection error of the lifelog stress detection model increases. Indeed, the mean stress-moment detection error rate of 19.74% improves the mAP@50 by 15.15% while the detection error rate of 21.85% results in an increase in the mAP@50 of about 6.63%. As the detection error rate increases up to 28.96%, the mAP@50 of the non-interactive lifelog retrieval system declines by 0.73%. To verify if the detection error rate affects the mAP@50 of the non-interactive retrieval system, I use the ground-truth data with manually-labeled mental stress information to do the filter. The results shown in Table 6.4 indicate that in a perfect condition when the detection error of the lifelog stress detection model is 0%, the mAP@50 and the mR@50 of the non-interactive system are still the same. Therefore, it is safe for us to conclude that the deterioration of the mAP@50 score of the non-interactive system does not cause by the detection error of the stress detection model. The slight decrease of the mAP@50 is due to the strong dependence on the results obtained from the free-text search using the CLIP model according to my defined search process in this experiment. The stress-moment filter does not help the system find new moments but removes the irrelevant ones with unmatched metadata and re-ranks the retrieval results, thereby, increasing the mean recall score at the top 50 ranked list (mR@50) significantly and the mean precision score at the top 50 ranked list (mAP@50) slightly. In case the results are not already in the ranked list after the free-text search, the stress-moment filter does not help at all. However, when comparing the significant improvement of the mAP@50 score (6.63% and 15.15%) in the case of User 1 and User 2 with the slight deterioration of the mAP@50 score in the case of User 3 (0.73%), I realize that the benefit of using the stress-moment filter in the lifelog retrieval system is considerable. Therefore, I still can conclude that the stress-moment filter function actually enhances the overall performance of the non-interactive lifelog retrieval system. In summary, according to these experimental results, I can conclude that the stress-moment filter function improves the performance of the non-interactive lifelog retrieval system in terms of increasing the mean precision and recall of the system at top-50 retrieval results (mAP@50 and mR@50) when dealing with stress-related/emotional-related queries.

6.3.2.2 Experiment 2 – Evaluation of the Interactive Lifelog Retrieval System integrated Stress-Moment Filter

User ID	Normal System		With Stress-Moment Filter	
	Mean Recall ↑	Mean Time ↓	Mean Recall ↑	Mean Time ↓
User 1	57.73	97.09	**63.12**	**90.49**
User 2	43.34	67.88	**65.04**	**64.78**
User 3	59.12	**97.62**	**59.87**	106.82

Table 6.5: Overall evaluation scores of the interactive version of the LifeSeeker system with and without the stress-moment filter.

In this experiment, I evaluate the performance of the interactive lifelog retrieval system using recall score and the search time as I require the user to search for as many images as possible that match the query description under a time constraint. The reason for the choices of evaluation metrics and the experiment design is that the user has a memory advantage when retrieving their own data. This advantage yields a bias for the user in this experiment that makes the precision scores inappropriate to be used as the main metrics to evaluate the performance of the interactive lifelog retrieval system. For the choice of the evaluation metrics used in this experiment, an increase in the recall score and a decrease in the search time indicate the existence of an improvement in the performance of the retrieval system. From Table 6.5, I recognize that the mean recall score of the interactive lifelog retrieval system with stress-moment filter function increases by 9.27% in average while reducing the mean search time by 4.85 seconds in average. Though the mean recall score increases for all the cases, which implies that the stress-moment filter indeed helps the user to retrieve more relevant results during the retrieval process, the mean search time does not always decrease for all the cases. In particular, while the mean search time required for User 1 and User 2 to search for all possible moments reduces by 6.6 seconds and 3.1 seconds, the one of User 3 increases by 9.2 seconds. The feedback

User ID	**Query ID**	**Normal System**		**With Stress-Moment Filter**	
		Recall ↑	**Time ↓**	**Recall ↑**	**Time ↓**
User 1	Q1	95.24	93.48	57.14	73.54
	Q2	42.86	125.25	42.86	49.84
	Q3	100	88.33	100	102.5
	Q4	90.16	59.38	83.61	81.63
	Q5	57.14	87.17	66.67	72.07
	Q6	16.71	154.77	41.43	92.69
	Q7	66.67	63.88	91.67	68.27
	Q8	0	180	54.55	156.83
	Q9	11.11	37	22.22	94
	Q10	97.37	81.68	71.05	40.85
User 2	Q1	76.92	73	92.31	35.17
	Q2	30	55.67	70	61.71
	Q3	8.33	69	16.67	149.50
	Q4	63.64	114	72.73	92.38
	Q5	50	33.20	20	26.50
	Q6	0	180	55.56	66.40
	Q7	36.36	53.50	54.55	77.50
	Q8	50	110.92	95.83	57.91
	Q9	18.18	42	72.73	51.75
	Q10	100	127.50	100	29
User 3	Q1	20	127	0	180
	Q2	50	65.53	73.33	83.14
	Q3	70	119.63	74	131.65
	Q4	77.78	59.86	62.96	60.88
	Q5	100	11.87	100	19.33
	Q6	33.33	121	11.11	140
	Q7	97.67	87.31	98.84	125.78
	Q8	96.30	127.54	96.30	115.58
	Q9	0	180	19.05	131.13
	Q10	46.15	76.47	63.08	80.66

Table 6.6: Detailed evaluation scores for each query of the interactive version of the LifeSeeker system with and without the stress-moment filter.

that I obtain from User 3 is the user interaction design of the stress-moment filter using pre-defined text-based syntax requires him to re-type the queries many times, which is inefficient in terms of time spent on searching under time pressure. The same feedback is also provided by other users, which is also the explanation for the slight reduction in the mean search time. The suggestions from three users for improvement is to create a separate filter button on the navigation bar of the User Interface would help reduce the number of interactions required to filter, which would help reduce the mean search time. Apart from the overall evaluation scores of the systems shown in Table 6.5, detailed evaluation scores for each query are demonstrated in Table 6.6. It can also be inferred from Table 6.6 that for most of the queries, the recall scores of the system with stress-moment filter function are equal to or higher than the one of the normal system. Therefore, based on the experimental results, I can conclude that the benefit of the stress-moment filter function in the interactive lifelog retrieval system is to help remove irrelevant results in the ranked list leading to a significant increase in the recall score (the number of relevant results) and a slight decrease in the mean search time in overall, thereby, addressing research question 3.

6.4 Chapter Summary

In this chapter, I addressed Research Question 3 by describing the development of my state-of-the-art lifelog retrieval system named LifeSeeker and comparing the performance of this state-of-the-art retrieval system with and without the integration of the stress-moment filter function. I used different evaluation metrics in different contexts (non-interactive and interactive scenarios) to assess the benefit of using mental stress information during the retrieval process. The mental stress information that I used in my experiments is inferred from the personalized lifelog stress detection model built in Section 5.3.3. From the experimental results in both interactive and non-interactive modes, I concluded that the integration of the stress-moment filter function helped improve the performance of the lifelog retrieval system in both scenarios when dealing with stress-related/emotional-related queries. In particular, this newly proposed integration increased the mean recall score (mR@50) and the mean average precision (mAP@50) of the non-interactive lifelog retrieval system while increasing the mean recall score as well as reducing the mean search time required in the interactive mode slightly in overall. Nevertheless, based on user feedback from the experiment, additional research and user studies should be conducted in the future to enhance the user interface's stress-moment filter function and improve user interaction during the retrieval process.

Chapter 7

Conclusion

In this thesis, I proposed the hypothesis that it is possible to identify stress moments in lifelog data using the physiological signals captured from readily available lifelog sensors and enhance the performance of the state-of-the-art lifelog retrieval system with stress-indexed information to address stress-related queries. To validate this, I formed three primary research questions and conduct experiments to find answers for each research question through a series of evaluations to either prove or disprove the hypothesis. In this final chapter, I provide a summarization of how I address these research questions to prove the correctness of my hypothesis.

7.1 Summary

For the first research question, I asked how successfully can low-resolution physiological signals recorded from consumer-grade wearable devices be used to develop a stress detection model, compared to the use of the high-resolution signals captured from traditional clinical-grade devices as the training data. This question is formed because I recognize that, though the physiological data from clinical-grade devices are considered to be of higher quality and resolution, the use of consumer-grade wearable devices is more practical and convenient for everyday use. To address research question 1, I conducted experiments on four stress datasets recorded in the laboratory environment including WESAD [9], AffectiveROAD [10], Cognitive Load [2], and DCU-NVT-EXP2 (Section 4.2.1.4). I trained five stress detection models using five different Machine Learning and Deep

Learning models and carried out statistical analyses on the balanced accuracy scores of these models to compare the stress detection capability of stress detection models trained on either low-resolution physiological signals or high-resolution ones. From the experimental results in Section 4.3, I showed that the performance of the stress detection model trained on low-resolution signals recorded from consumer-grade wearable devices is almost the same as the one trained on high-resolution signals recorded from traditional clinical devices. In addition, the experimental results and statistical analyses in Section 4.4 proved that subject-dependent stress detection models outperformed the subject-independent stress detection ones using low-resolution signals. These analyses and results from experiments in Chapter 4 are strong evidence to conclude that it is feasible and effective to use the data from consumer-grade wearable devices to build a highly-accurate personalized stress detection model, thereby, addressing research question 1.

For the second research question, I asked how successfully stress detection models using low-resolution physiological signals from consumer-grade wearable devices can be applied for lifelog data to detect stressful moments. This question is formed to investigate the possibility of applying the stress detection model to the lifelog data to detect stressful moments in unconstrained environments. To address this research, I conducted a longitudinal study to collect stress lifelog data from three participants who joined my previous experiment that satisfy the experiment requirements of the lifelogging knowledge to ensure the integrity of the collected data (Section 5.2.1). Experimental results from Section 5.3.1 showed that the laboratory stress detection model cannot be used directly to detect stress moments in unconstrained environments due to the difference in the distribution of the statistical features extracted from the signals recorded from two different conditions and environments. In particular, the data collection process in the unconstrained environments did not restrict the location where the participant should be, the task that the participant should do, or the scenario that the participant should involve in while all of these conditions were controlled in the constrained environments. This led to a significant difference in the environment's temperature, humidity, and weather which resulted in different stress responses between the constrained and unconstrained environments. Therefore, it is not appropriate to apply the model directly but a further model re-training step needs to be done. The experimental results in Section 5.3.3 showed that by re-training the model on the lifelog data, the balance accuracy score of the best re-train models – lifelog stress detection models – increased 15.29%(±6.65%) on average compared with the best ones trained on laboratory data. The balanced accuracy scores of the lifelog stress detection models vary from 56.27% to 62.80% indicating that using the physiological signals solely to detect stress moments is not enough. According to the feedback from participants in the experiment, the social interaction, the people they were talking to, the conversation topic, and the context that they were involved are also factors that made participants feel stressed. This is the task that I proposed for future work to experiment with the effect of using both physiological signals and conversation data to train a stress detection model. In summary, through the balanced accuracy scores, I conclude that the personalized stress detection model applied to lifelog should be re-trained on the lifelog data to capture all the context changes in unconstrained environments to detect stress moments more accurately (ranging from 56.27% to 62.80%), which is also the answer to the research question 2.

For the third research question, I asked how biometric and visual data can be used in a lifelog interactive retrieval system to retrieve stress-related moments. This question is formed based on the hypothesis that the stress information inferred from the biometric data by the lifelog stress detection model is useful for the lifelog retrieval system to deal with stress-related/emotional-related queries, however, it has not been exploited in any lifelog retrieval systems yet. I proposed to use the stress information as an indicator to filter for relevant moments. Therefore, I designed experiments to evaluate the impact of the stress-moment filter function on the performance of the state-of-the-art lifelog retrieval system in both interactive and non-interactive modes. To do so, firstly, I did the literature review and analyzed core features in previous state-of-the-art lifelog retrieval systems to develop a state-of-the-art system of my own, which was described in Section 6.2, to conduct experiments. The results from the benchmarking Lifelog Search Challenges in 2021 and 2022 shown in Section 6.2.3 proved that my lifelog retrieval system – named LifeSeeker– was one of the state-of-the-art lifelog retrieval systems. I then integrated the stress-moment filter function into the LifeSeeker system to evaluate its performance compared to the normal LifeSeeker system. For the non-interactive mode, the overall performance of the lifelog retrieval system was assessed by the mean average precision and the mean recall score at the top 50 results in the ranked list (mAP@50 and mR@50). For the interactive mode, the overall performance of the lifelog retrieval system was evaluated by the mean recall score and the mean search time required by the user to find correct images that best illustrated the queries. Through experimental results shown in Section 6.3.2, I concluded that the stress-moment filter function improves the performance of the lifelog retrieval system in both modes despite the detection error rate of the lifelog stress detection model. In particular, the stress-moment filter function improved the mean precision and mean recall of the non-interactive lifelog retrieval system, allowing it to more accurately retrieve stress-related moments. For the interactive lifelog retrieval system, this newly proposed function increased the mean recall score significantly and reduced the mean search time required to search for correct moments slightly. These are the main benefits that I observed through my experiment on the lifelog retrieval system integrated stress-moment filter function, which also provides the answer to research question 3.

As I have addressed the three primary research questions, I can discuss the validity of my hypothesis. Given the limitations of this research, which are described in the following section, I proved that it is feasible to detect stress moments in lifelog data using physiological signals captured from readily available lifelog sensors, and this mental stress information could be used as in the filter function to enhance the performance of the state-of-the-art lifelog retrieval system. Therefore, I consider my proposed hypothesis defined in Section 1.4 to be upheld.

7.2 Limitations

I recognize three main limitations remained in my Ph.D. research as follows:

- **Annotation Process of the Stress Data Collection**: As stress is a subjective experience and different people may experience stress in different ways, this makes it hard to obtain a benchmarking ground truth for stress detection problems. The annotation of mental stress status often relies on self-report methods, such as questionnaires and interviews, which can be subject to bias and error. People may not always be aware of their own stress levels or may not want to disclose them due to stigma or other reasons. Furthermore, stress can manifest in different ways, such as physical symptoms, behavioral changes, or cognitive processes, which can be difficult to assess objectively. Therefore, the limitation in this research when developing a personalized stress detection model is that I have to assume the subjective evaluation and labels of stress are correct as well as try my best to overcome the bias and error caused by subjective self-evaluation using multiple different stress status questionnaires, forms in combination with the supports of multiple annotation tools and devices.

- **Lifelog Stress Detection Model**: As in my experiment of evaluating the performance of the stress detection model applied to lifelog data, I manage to develop a personalized lifelog stress detection model for each individual using physiological signals in the corresponding lifelog data archive with the accuracy ranging from 56.27% to 62.80%. These results are not considerably high in terms of experimental evaluations but are sufficient enough to be used as an indicator for relevant moment filtering to enhance the performance of the state-of-the-art lifelog retrieval systems in overall. However, the lifelog stress detection model would need to be able to make use of all the data it receives from a person's lifelog to enhance performance. The context that the user is in including the people that the user meets, and the conversation topic that the user discusses at that time are important factors to infer the stress status of an individual, according to the feedback from the participants in the experiment. This is also the limitation of my work in this research as I only investigate the performance of lifelog stress detection model using physiological signals solely.

- **Small sample size of the lifelog stress dataset**: One limitation of my research is the small sample size of the lifelog stress dataset due to the limitation of the number of available sensor sets as well as the strict requirements of the participant recruiting process for the longitudinal study and the experiment. In particular, three among eleven participants from my experiment on stress detection in a constrained environment are selected to join the longitudinal study since they have prior background in lifelogging and are familiar with the lifelog wearable sensors. This requirement is necessary to ensure the integrity of the collected data. As the small sample size in the lifelog stress dataset makes it difficult to accurately analyze the data and draw meaningful conclusions from it, it is only possible for us to conduct proof-of-concept research proposing the best approach to building a personalized lifelog stress detection model. I suggest that future work should reuse my work to further improve the performance of the lifelog stress detection model as well as gain more insights into the stress response characteristics between different individuals with different demographic backgrounds.

7.3 Future Work

My research has successfully demonstrated the effectiveness of using stress detection in lifelog retrieval systems as well as proposed an effective approach to building a lifelog stress detection model with physiological signals from consumer-grade wearable devices in lifelog data archive. However, there are still several limitations that need to be addressed in future research. These include improving the accuracy of the lifelog stress detection models, exploring additional applications of stress detection in lifelog retrieval systems, and evaluating the stress-moment filter function integrated into the state-of-the-art lifelog retrieval system via the benchmarking Lifelog Search Challenge. These improvements can lead to several interesting potential directions in the future by further developing the research in this thesis. In particular, the future works that I propose are as follows:

- **Enhancement of Lifelog Stress Detection Model**: As discussed in my research limitation, the accuracy of the lifelog stress detection system can still be improved by taking into account the context of the moments such as the people that the user meets, the discussion topic of the meeting that the user involved, the context of the conversation that the user talks with others, etc. apart from the physiological signals. These are the data that are possible to capture by wearable devices and computer keyloggers. However, the researchers need to ensure the privacy and security of the individuals being monitored as these data can be considered private data that can be used to trace back that individual. By enhancing the stress pattern detection capability of the lifelog stress detection system, potential applications, and software can be developed to gather personal stress-related data of an individual to analyze the causes of stress, thereby, realizing mental stress monitoring in a real-time manner for early stress disorder issue detection.

- **Stress Tracking and Analysis Application**: With the enhancement of the lifelog stress detection model, stressful moments can be detected accurately and instantly which unlocks the potential for stress-tracking and analysis applications to be developed. This application could provide users with detailed insights and analytics about their stress patterns over time. This information can help individuals identify triggers, track their progress in stress management, and make informed lifestyle changes. In order to do so, more research on the understanding of the context of detected stressful moments needs to be conducted. Additionally, this application could also leverage the development of the smart mental health assistant. One approach that I would propose for the application of lifelog stress detection system is that stress moments shown in lifelog images detected from the models can be inputted into an image-to-text generator provided by Midjourney[1] to describe the stressful events/activities. These descriptions can be analysed to provide insights into the causes that make the user stressed. Thereby, prompts can be written to guide the large language model to act as a chatbot to support the mental health of the user.

- **Emotional Lifelog Retrieval System**: Potential applications to the lifelog retrieval system can be developed from the physiological signals in lifelog data archive apart from the stress detection model. In particular, the emotion recognition model can also be trained the same way as the stress detection model using the same kinds of physiological signals [157]. Therefore, my research methodology, experiment designs, and proposed models in my research can be re-used in the research of emotion recognition in lifelog data. With both lifelog emotion recognition and lifelog stress detection models, an emotional lifelog retrieval system can be developed. An emotional lifelog retrieval system is a type of technology that uses either stress or emotional information to identify and retrieve specific moments from a person's personal digital diary or lifelog. This system allows users to quickly and easily access specific moments from their past that may have been particularly emotional, stressful, or significant. This technology has potential applications in a variety of fields such as mental health monitoring,

[1] https://www.midjourney.com/

personal productivity tracking, life satisfaction analyses, etc.

- **Evaluation of Lifelog Retrieval System integrated Stress-Moment Filter in the Benchmarking Lifelog Search Challenge**: As my work in assessing the impact of the stress-moment filter function in lifelog retrieval system was only completed on private lifelog data, I have not had a chance to evaluate this newly proposed feature in the publicly available lifelog dataset in the benchmarking Lifelog Search Challenge. Hence, my experiment described in Chapter 6 cannot be conducted on other users who do not join my previous longitudinal study to use the retrieval system. The conclusion on the impact of the stress-moment filter function on the lifelog retrieval system would be more reliable when it could be tested on a large number of users who do not own the data. Thereby, the memory bias in my could be ignored resulting in the possibility of evaluating the precision of the retrieval system with a stress-moment filter. However, due to the data privacy expectation from the participants in my experiment, I am not allowed to let others use my lifelog retrieval system to retrieve the data of the participants. Therefore, in future work, I am expecting that this newly proposed feature can be used in the benchmarking Lifelog Search Challenge to evaluate the overall performance (precision, recall, accuracy) of the lifelog retrieval system with the stress-moment filter function in the interactive mode.

7.4 Research Contributions

The key contributions that I made in this thesis can be summarized as follows. In Chapter 4, I addressed Research Question 1, which evaluates the performance of the stress detection model using low-resolution physiological signals recorded from consumer-grade wearable devices. Based on the experimental results in Section 4.3, I showed that the stress detection model, which utilizes low-resolution physiological signals collected from wearable devices as training data, performs equally well when compared to the model trained with high-resolution data gathered from conventional clinical devices. In addition, based on the experimental results in Section 4.3, I showed that the stress detection model, which utilizes low-resolution physiological signals collected from wearable devices as training data, performs equally well when compared to the model trained with high-resolution data gathered from conventional clinical devices. I then employed the findings in Chapter 4 to address Research Question 2 in Chapter 5, which evaluates the performance of stress detection models using consumer-grade wearable devices' data in unconstrained environments. I gathered longitudinal stress lifelog data (as outlined in Section 5.2.1) and conducted a proof-of-concept study to assess the effectiveness of the stress detection model when applied to lifelog data. With the newly collected dataset, I demonstrated that utilizing the laboratory stress detection model for detecting stressful moments in daily life would yield inaccurate results, as supported by the experimental findings and insights obtained from feature analyses and model explanation in Section 5.3.1 and Section 5.3.2. Based on the experimental findings presented in Section 5.3.3, I suggest that the training of the lifelog stress detection model should focus on physiological signals recorded in unrestricted conditions rather than constrained environments. Finally, using the stress/non-stress moments detected from the lifelog stress detection model built in Chapter 5, I evaluated the impact of the stress-moment filtering function in the lifelog retrieval system in Chapter 6. In Section 6.2.1, I developed (with colleagues) an interactive retrieval system for lifelog data called **LifeSeeker** and assessed its performance through annual benchmarking Lifelog Search Challenges. The outcomes of these challenges demonstrated that **LifeSeeker** stands among the top-notch lifelog retrieval systems. Based on the experimental results in Section 6.3, I proved that when integrating the stress-moment filter function into the state-of-the-art lifelog retrieval system, whose stress-moments are detected from the lifelog stress detection model, the overall performance of the system increases in both interactive and non-interactive mode.

7.5 Publication List

Within this section, I present an inventory of all the publications generated during the course of this project. While I acknowledge the significance of each entry in advancing the field of stress detection in lifelog data and its application in lifelog retrieval systems, only a subset of these contributions directly pertains to the research discussed in this thesis. Other publications are supporting findings that are indirectly related to the research in this thesis, therefore, it is worth to also mention them in this section.

- **Chapter 4 – RQ1:**
  - **Van-Tu Ninh**, Sinéad Smyth, Minh-Triet Tran and Cathal Gurrin. "Analysing the Performance of Stress Detection Models on Consumer-Grade Wearable Devices." SoMeT (2021).
  - **Van-Tu Ninh**, Manh-Duy Nguyen, Sinéad Smyth, Minh-Triet Tran, Graham Healy, Binh T. Nguyen, amd Cathal Gurrin. An Improved Subject-Independent Stress Detection Model Applied to Consumer-grade Wearable Devices. In: Fujita, H., Fournier-Viger, P., Ali, M., Wang, Y. (eds) Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence. IEA/AIE 2022. Lecture Notes in Computer Science(), vol 13343. Springer, Cham (2022).
- **Chapter 6 – RQ3:**
  - Thao-Nhu Nguyen, Tu-Khiem Le, **Van-Tu Ninh**, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Sinéad Smyth, Annalina Caputo, and Cathal Gurrin. 2022. LifeSeeker 4.0: An Interactive Lifelog Search Engine for LSC'22. In Proceedings of the 5th Annual on Lifelog Search Challenge (LSC '22). Association for Computing Machinery, New York, NY, USA, 14–19.

  - Thao-Nhu Nguyen, Tu-Khiem Le, **Van-Tu Ninh**, Minh-Triet Tran, Nguyen Thanh Binh, Graham Healy, Annalina Caputo, and Cathal Gurrin. 2021. LifeSeeker 3.0: An Interactive Lifelog Search Engine for LSC'21. In Proceedings of the 4th Annual on Lifelog Search Challenge (LSC '21). Association for Computing Machinery, New York, NY, USA, 41–46.
  - Le Tu-Khiem, **Van-Tu Ninh**, Minh-Triet Tran, Thanh-An Nguyen, Hai-Dang Nguyen, Liting Zhou, Graham Healy, and Cathal Gurrin. Lifeseeker 2.0: interactive lifelog search engine at LSC 2020. In Proceedings of the Third Annual Workshop on Lifelog Search Challenge, pp. 57-62. 2020.
  - Le Tu-Khiem, **Van-Tu Ninh**, Duc-Tien Dang-Nguyen, Minh-Triet Tran, Liting Zhou, Pablo Redondo, Sinead Smyth, and Cathal Gurrin. Lifeseeker: Interactive Lifelog Search Engine at LSC 2019. In Proceedings of the ACM Workshop on Lifelog Search Challenge, pp. 37-40. 2019.
  - **Van-Tu Ninh**, Tu-Khiem Le, Liting Zhou, Graham Healy, Kaushik Venkataraman, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Sinead Smyth, and Cathal Gurrin. A Baseline Interactive Retrieval Engine for Visual Lifelogs at the NTCIR-14 lifelog-3 Task. In NII Conference on Testbeds and Community for Information Access Research, pp. 29-41. Springer, Cham, 2019.
  - **Van-Tu Ninh**, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Mathias Lux, Minh-Triet Tran, Cathal Gurrin, and Duc Tien Dang Nguyen. LIFER 2.0: Discovering Personal Lifelog Insights using an Interactive Lifelog Retrieval System. CLEF, 2019.

- **Others:**

  - **Van-Tu Ninh**, Tu-Khiem Le, Manh-Duy Nguyen, Sinéad Smyth, Graham Healy, and Cathal Gurrin. A Preliminary Assessment of Game

Event Detection in Emotional Mario Task at MediaEval 2021. (2021).

- Tu-Khiem Le, **Van-Tu Ninh**, Mai-Khiem Tran, Graham Healy, Cathal Gurrin, and Minh-Triet Tran. 2022. AVSeeker: An Active Video Retrieval Engine at VBS2022. In MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 537–542.
- Nguyen Thao-Nhu, Tu-Khiem Le, **Van-Tu Ninh**, Ly-Duyen Tran, Manh-Duy Nguyen, Minh-Triet Tran, Binh T. Nguyen et al. DCU and HCMUS at NTCIR-16 Lifelog-4. In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies. NTCIR, 2022.
- Cathal Gurrin, Tu-Khiem Le, **Van-Tu Ninh**, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoč, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schoeffmann. Introduction to the third annual lifelog search challenge (LSC'20). In Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 584-585. 2020.
- Le Tu-Khiem, **Van-Tu Ninh**, Liting Zhou, Minh-Huy Nguyen-Ngoc, Huu-Duc Trinh, Nguyen H. Tran, Luca Piras, M. Riegler, P. Halvorsen, Mathias Lux, Minh-Triet Tran, Graham Healy, Cathal Gurrin and Duc-Tien Dang-Nguyen. Organiser Team at ImageCLEFlifelog 2020: A Baseline Approach for Moment Retrieval and Athlete Performance Prediction using Lifelog Data. Conference and Labs of the Evaluation Forum (2020).
- **Van-Tu Ninh**, Tu-Khiem Le, Liting Zhou, Luca Piras, M. Riegler, P. Halvorsen, Mathias Lux, Minh-Triet Tran, Cathal Gurrin and Duc-Tien Dang-Nguyen. Overview of ImageCLEF Lifelog 2020: Lifelog Moment Retrieval and Sports Performance Lifelog. Conference and Labs of the Evaluation Forum (2020).
- Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, **Van-Tu**

Ninh, T-K. Le, Rami Albatal, D-T. Dang-Nguyen, and Graham Healy. Overview of the NTCIR-14 lifelog-3 task. In Proceedings of the 14th NTCIR conference, pp. 14-26. NII, 2019.

Appendix A

Known-Item Search Stress-related Topics

Table A.1: List of stress-related/emotional-related queries used in the experiments to evaluate the performance of lifelog interactive retrieval integrated stress-moment filter.

User ID	Query	Description
User 1	Q1	I was writing all the business requirements on my laptop and planning for the customer trip to prepare for the upcoming meeting. I feel nervous because the meeting was coming soon.
	Q2	I was cleaning my car door handles because there is a stain on it. I was angry at that time because the stain is hard to clean up.
	Q3	I was playing with my cat at my desk. I am nervous that I can't look after him well as he is too small and weak.
	Q4	I was having a meeting with a man using my smartphone in my office. We are discussing some work I never did before. I feel nervous at that time as the business work is new to me, so there are many tasks I have to do.

Continued on next page

Table A.1 – continued from previous page

User ID	Query	Description
	Q5	I was asking a man in a white shirt for a suggestion in front of the toilet. I was worried about my decision to let an excellent student to start his PhD without the English test.
	Q6	I was watching the Thirteens Lives movies on my computer at home. I really enjoy the movie.
	Q7	I was searching and buying some cats products in Taobao app using my smartphone. I was so relaxed by doing that after a long day.
	Q8	I was having lunch and talking with a Thailand man in Chinese in the meeting room during break time. It was so relaxed to speak Chinese with him.
	Q9	I was watching some cat videos on my phone. It was so relaxed to see cats.
	Q10	I was building the structure of a house with my friend's son. It was a relaxed day.
User 2	Q1	I opened the Bank of Ireland application on the phone to check the balance after receiving an email stating that I had to pay rent. I messaged my boyfriend to give me back my money, then pay the rent. It was such a stressful moment for me.
	Q2	I was in the lab for tutoring. I am kind of relaxed as not many people turned up, so I had a chat with my colleagues.
	Q3	I was working on my computer and arguing with my boyfriend who had a day off that day. He was playing game on the computer and ignore me. I felt stressful

Continued on next page

Table A.1 – continued from previous page

User ID	Query	Description
		and angry at the same time.
	Q4	I was having a discussion about my paper with a colleague (and also a friend) in the meeting room after the biweekly meeting. I felt very relaxed at that time.
	Q5	I was watching some Youtube videos with my boyfriend. I remember it was a new Ed Sheeran song about Pokemons. Then we watched a video about the Japanese egg mascot. We also listened to a Korean acapella group. I felt relaxed at that time.
	Q6	I was playing a mobile game on my phone and lying on the bed. It was a lazy day of mine and I felt very relaxed.
	Q7	I was enjoying a walk in St. Anne's Park with my roommates. I was impressed by the tree with colorful windows on it. It was a cold day but it was great fun there.
	Q8	I was in a biweekly group meeting watching a presentation over Zoom about his new method to improve the Retrieval System. It was an interesting presentation. I enjoy the talk much and felt relaxed as I do not need to do anything in the meeting.
	Q9	I was spinning my pen while coding for a project. My code didn't work at all. I felt really stressed.
	Q10	I was watching a video on a module on Loop about wellbeing. I felt personally attacked on why I'm not so happy so I was so annoyed.
User 3	Q1	I was watching a Youtube video about the Neural Style Transfer using a picture of Van Gogh as an example. I

Continued on next page

Table A.1 – continued from previous page

User ID	Query	Description
		was quite stressful as I spent a lot of effort to understand the content.
	Q2	I gave a presentation in a competition under a strict time limit and received some tough questions from the judges. I felt nervous since I had to perform well to secure the win for my team.
	Q3	I was recording data for my research. I need to do a lot of things at the same time, such as giving task instruction, checking data quality, observing participants' odd behaviors. I also needed to reduce my body movement in order to not causing any distraction to the participant. It was really tense.
	Q4	I remembered that I had a very delicious bowl of "Bun Ca" when gathering with friends on the weekend. I remembered that I was excited to have good food at that time.
	Q5	I was practicing piano and playing a song named "Kiss the rain" after a long break. I felt very relaxed after a long working day as I still managed to play the song smoothly.
	Q6	I was playing a board game with my friends and was splitting cards for a new round. I was excited to play games with others on the weekend after a long tired working streak.
	Q7	I was taking a bus back to DCU after wandering around the city center. I remembered that I was relaxed as I had a big lunch and I felt really full.
	Q8	I was listening to a presentation that introduces the conditions to win the challenge that I joined including the problems, the rules, and the timeline. However, there

Continued on next page

Table A.1 – continued from previous page

User ID	Query	Description
		was a slide that talked about how the participants were ranked did shock me a bit since it considered not only the performance of the solution, but also the execution time and solution presentation.
	Q9	My team and I were discussing with our mentor on what we should try for the upcoming hours in the Huawei university challenge. He wore the Huawei shirt as ours but with a name card and glasses. He did give us many great suggestions and also clarified a few concerns that we had. I was tense at that moment as my team might have a chance to win the challenge but it would require a lot of effort to deliver the best solution.
	Q10	I was chatting to a friend on Zoom to give him some advice for his upcoming interview. Although his work was mostly about content creation and management, the position he applied for required some knowledge about AI in general. I was kind of relaxed at this time.

Bibliography

[1] Aaqib Saeed, Stojan Trajanovski, Maurice Van Keulen, and Jan Van Erp. Deep physiological arousal detection in a driving simulator using wearable sensors. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 486–493. IEEE, 2017.

[2] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. Continuous stress detection using a wrist device: In laboratory and real life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 1185–1193, New York, NY, USA, 2016. Association for Computing Machinery.

[3] Lindsay M Biga, Sierra Dawson, Amy Harwell, Robin Hopkins, Joel Kaufmann, Mike LeMaster, Philip Matern, Katie Morrison-Graham, Devon Quick, and Jon Runyeon. *Anatomy & physiology*. OpenStax/Oregon State University, 2020.

[4] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

[5] Lam Huynh, Tri Nguyen, Thu Nguyen, Susanna Pirttikangas, and Pekka Siirtola. Stressnas: Affect state and stress detection using neural architecture search. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 121–125, 2021.

[6] Bernd Münzer, Andreas Leibetseder, Sabrina Kletz, Manfred Jürgen Primus, and Klaus Schöffmann. lifexplore at the lifelog search challenge 2018. In *LSC '18*, 2018.

[7] Aaron Duane, Cathal Gurrin, and Wolfgang Hürst. Virtual reality lifelog explorer: Lifelog search challenge at acm icmr 2018. In *LSC '18*, 2018.

[8] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. Myscéal: An experimental interactive lifelog retrieval system for lsc'20. *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*, 2020.

[9] Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018.

[10] Neska El Haouij, Jean-Michel Poggi, Sylvie Sevestre-Ghalila, Raja Ghozi, and Mériem Jaïdane. Affectiveroad system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, page 800–803, New York, NY, USA, 2018. Association for Computing Machinery.

[11] Van-Tu Ninh, Manh-Duy Nguyen, Sinéad Smyth, Minh-Triet Tran, Graham Healy, Binh T. Nguyen, and Cathal Gurrin. An improved subject-independent stress detection model applied to consumer-grade wearable devices. In Hamido Fujita, Philippe Fournier-Viger, Moonis Ali, and Yinglin Wang, editors, *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence*, pages 907–919, Cham, 2022. Springer International Publishing.

[12] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Nguyen

Thanh Binh, Graham Healy, Annalina Caputo, and Cathal Gurrin. Lifeseeker 3.0: An interactive lifelog search engine for lsc'21. 2021.

[13] Bush Vannevar. As we may think. *ACM Sigpc Notes*, 1979.

[14] Isabel Pedersen. Ready to wear (or not): Examining the rhetorical impact of proposed wearable devices. *2013 IEEE International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmediated Reality in Everyday Life*, pages 201–202, 2013.

[15] Jim Gemmell, Gordon Bell, and Roger Lueder. Mylifebits: a personal database for everything. *Commun. ACM*, 49:88–95, 2006.

[16] Jim Gemmell, Gordon Bell, Roger Lueder, Steven M. Drucker, and Curtis Wong. Mylifebits: fulfilling the memex vision. In *MULTIMEDIA '02*, 2002.

[17] Louise N Signal, James Stanley, M Smith, MB Barr, Tim J Chambers, Jiang Zhou, Aaron Duane, Cathal Gurrin, Alan F Smeaton, Christina McKerchar, et al. Children's everyday exposure to food marketing: an objective analysis using wearable cameras. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1):1–11, 2017.

[18] Bethan Everson, Kelly A Mackintosh, Melitta A McNarry, Charlotte Todd, and Gareth Stratton. Can wearable cameras be used to validate school-aged children's lifestyle behaviours? *Children*, 6(2):20, 2019.

[19] Qianling Zhou, Di Wang, Cliona Ni Mhurchu, Cathal Gurrin, Jiang Zhou, Yu Cheng, and Haijun Wang. The use of wearable cameras in assessing children's dietary intake and behaviours in china. *Appetite*, 139:1–7, 2019.

[20] Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, and Cathal Gurrin. Dcu at the ntcir-13 lifelog-2 task. NTCIR, 2017.

[21] Martin Dodge and Rob Kitchin. 'outlines of a world coming into existence':

pervasive computing and the ethics of forgetting. *Environment and planning B: planning and design*, 34(3):431–445, 2007.

[22] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. Lifelogging: Personal big data. *Found. Trends Inf. Retr.*, 8:1–125, 2014.

[23] Gordon Bell and Jim Gemmell. A digital life. *Scientific American*, 296(3):58–65, 2007.

[24] Liadh Kelly and Gareth JF Jones. Biometric response as a source of query independent scoring in lifelog retrieval. In *European Conference on Information Retrieval*, pages 520–531. Springer, 2010.

[25] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Giulia Boato, Liting Zhou, and Cathal Gurrin. Overview of imagecleflifelog 2017: Lifelog retrieval and summarization. In *CLEF*, 2017.

[26] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Matthias Lux, and Cathal Gurrin. Overview of imagecleflifelog 2018: Daily living understanding and lifelog moment retrieval. In *CLEF*, 2018.

[27] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Van-Tu Ninh, Tu-Khiem Le, Rami Albatal, Duc-Tien Dang-Nguyen, and Grace Healy. Overview of the ntcir-14 lifelog-3 task. 2019.

[28] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Matthias Lux, Minh-Triet Tran, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. Overview of imagecleflifelog 2019: Solve my life puzzle and lifelog moment retrieval. In *CLEF*, 2019.

[29] Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, and Duc-Tien Dang-Nguyen. Overview of ImageCLEF Lifelog 2020:Lifelog Moment Retrieval and Sport Performance Lifelog. In *CLEF2020 Working Notes*, CEUR

Workshop Proceedings, Thessaloniki, Greece, September 22-25 2020. CEUR-WS.org <http://ceur-ws.org>.

[30] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Loko, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schöffmann. Introduction to the third annual lifelog search challenge (lsc'20). *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020.

[31] Alan F. Smeaton. Lifelogging as a memory prosthetic. In *Proceedings of the 4th Annual on Lifelog Search Challenge*, LSC '21, page 1, New York, NY, USA, 2021. Association for Computing Machinery.

[32] Riccardo Sioni and Luca Chittaro. Stress detection using physiological sensors. *Computer*, 48(10):26–33, 2015.

[33] Donald G MacKay, Meredith Shafto, Jennifer K Taylor, Diane E Marian, Lise Abrams, and Jennifer R Dyer. Relations between emotion, memory, and attention: Evidence from taboo stroop, lexical decision, and immediate memory tasks. *Memory & cognition*, 32(3):474–488, 2004.

[34] Charlotte Nickerson. The yerkes-dodson law and performance. *Simply Psychology*, 2021.

[35] Hans Selye. [stress without distress]. *Bruxelles medical*, 56 5:205–10, 1974.

[36] Willian C. Shield Jr. Definition of stress. `https://www.medicinenet.com/script/main/art.asp?articlekey=20104`, 2018. [Online; accessed 03-July-2020].

[37] Davide Carneiro, Paulo Novais, Juan Carlos Augusto, and Nicola Payne. New methods for stress assessment and monitoring at the workplace. *IEEE Transactions on Affective Computing*, 10:237–254, 2019.

[38] Understanding your stress type how to manage it. `https://www.neurocorecenters.com/blog/understanding-your-stress-type-how-to-manage-it`, 2018. [Online; accessed 03-July-2020].

[39] Asma Abdullah Alfayez, Holger Kunz, and Alvina Grace Lai. Predicting the risk of cancer in adults using supervised machine learning: A scoping review. *BMJ open*, 11(9):e047755, 2021.

[40] Jabir Al Nahian, Abu Kaisar Mohammad Masum, Sheikh Abujar, Md Mia, et al. Common human diseases prediction using machine learning based on survey data. *arXiv preprint arXiv:2209.10750*, 2022.

[41] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. The influence of person-specific biometrics in improving generic stress predictive models. *ArXiv*, abs/1910.01770, 2019.

[42] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. Wearable affect and stress recognition: A review. *ArXiv*, abs/1811.08854, 2018.

[43] Sonia J Lupien, Francoise Maheu, Mai Tu, Alexandra Fiocco, and Tania E Schramek. The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and cognition*, 65(3):209–237, 2007.

[44] Eefje S Poppelaars, Johannes Klackl, Belinda Pletzer, Frank H Wilhelm, and Eva Jonas. Social-evaluative threat: Stress response stages and influences of biological sex and neuroticism. *Psychoneuroendocrinology*, 109:104378, 2019.

[45] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[46] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks*, pages 97–108. IEEE, 2011.

[47] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. Monitoring stress with a wrist device using context. *Journal of biomedical informatics*, 73:159–170, 2017.

[48] Gillian H Ice and Gary D James. *Measuring stress in humans: A practical guide for the field.* Cambridge university press, 2007.

[49] Guy-Evans Olivia. Peripheral nervous system: Definition, parts and function. `https://www.simplypsychology.org/peripheral-nervous-system.html`, April 2021. [Online; accessed 19-September-2022].

[50] Marieke Martens, Angus Antley, Daniel Freeman, Mel Slater, Paul J. Harrison, and Elizabeth M. Tunbridge. It feels real: physiological responses to a stressful virtual reality environment and its impact on working memory. *Journal of Psychopharmacology (Oxford, England)*, 33:1264 – 1273, 2019.

[51] Sylvia D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84:394–421, 2010.

[52] R Bankenahally and Hari Krovvidi. Autonomic nervous system: anatomy, physiology, and relevance in anaesthesia and critical care medicine. *BJA Education*, 16:381–387, 2016.

[53] Fatema Akbar, Gloria Mark, Ioannis T. Pavlidis, and Ricardo Gutierrez-Osuna. An empirical study comparing unobtrusive physiological sensors for stress detection in computer work. *Sensors (Basel, Switzerland)*, 19, 2019.

[54] Katharina Meyerbröker and Paul M. G. Emmelkamp. Virtual reality exposure therapy in anxiety disorders: a systematic review of process-and-outcome studies. *Depression and anxiety*, 27 10:933–44, 2010.

[55] W. Boucsein, D. Fowles, S. Grimnes, G. Ben-Shakhar, W. Roth, M. Dawson, and D. Filion. Publication recommendations for electrodermal measurements. *Psychophysiology*, 49 8:1017–34, 2012.

[56] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, 92:103139, 2019.

[57] Bryn Farnsworth. What is electrodermal activity (eda)? and how does it work? `https://imotions.com/blog/eda/`, Jun 2019. [Online; accessed 21-September-2022].

[58] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010.

[59] Don C Fowles, Margaret J Christie, Robert Edelberg, William W Grings, David T Lykken, and Peter H Venables. Publication recommendations for electrodermal measurements. *Psychophysiology*, 18(3):232–239, 1981.

[60] Shalom Greene, Himanshu Thapliyal, and Allison Caban-Holt. A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine*, 5(4):44–56, 2016.

[61] Ane Alberdi, Asier Aztiria, and Adrian Basarab. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics*, 59:49–75, 2016.

[62] Fight or flight response. `https://www.psychologytools.com/resource/`

fight-or-flight-response, May 2022. [Online; accessed 22-September-2022].

[63] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry investigation*, 15(3):235, 2018.

[64] Luca Citi, Emery N Brown, and Riccardo Barbieri. A real-time automated point-process method for the detection and correction of erroneous and ectopic heartbeats. *IEEE transactions on biomedical engineering*, 59(10):2828–2837, 2012.

[65] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.

[66] Arnold M Weissler, Willard S Harris, and Clyde D Schoenfeld. Systolic time intervals in heart failure in man. *Circulation*, 37(2):149–159, 1968.

[67] Dominic J McCafferty. Applications of thermal imaging in avian science. *Ibis*, 155(1):4–15, 2013.

[68] MT Quazi, SC Mukhopadhyay, NK Suryadevara, and Yueh-Min Huang. Towards the smart sensors based human emotion recognition. In *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pages 2365–2370. IEEE, 2012.

[69] Cornelia Kappeler-Setz. *Multimodal emotion and stress recognition.* ETH Zurich, 2012.

[70] David Y Zhang and Allen S Anderson. The sympathetic nervous system and heart failure. *Cardiology clinics*, 32(1):33–45, 2014.

[71] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank Wilhelm, Tess Osborne, and Jessica Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17):3805, 2019.

[72] Arturas Kaklauskas, Edmundas Kazimieras Zavadskas, Mark Seniut, Gintautas Dzemyda, V Stankevic, C Simkevičius, T Stankevic, Rasa Paliskiene, A Matuliauskaite, Simona Kildiene, et al. Web-based biometric computer mouse advisory system to analyze a user's emotions and work productivity. *Engineering Applications of Artificial Intelligence*, 24(6):928–945, 2011.

[73] Kikuo Asai. The role of head-up display in computer-assisted instruction. *Human Computer Interaction: New Developments; Asai, K., Ed.; IntechOpen: Rijeka, Croatia*, pages 31–48, 2008.

[74] Terence KL Hui and R Simon Sherratt. Coverage of emotion recognition for common wearable biosensors. *Biosensors*, 8(2):30, 2018.

[75] Dongrae Cho, Jinsil Ham, Jooyoung Oh, Jeanho Park, Sayup Kim, Nak-Kyu Lee, and Boreom Lee. Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. *Sensors*, 17(10):2435, 2017.

[76] Bo Zhang. *Stress recognition from heterogeneous data.* PhD thesis, Université de Lorraine, 2017.

[77] Palanisamy Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. *Journal of Physical Therapy Science*, 24(12):1341–1344, 2012.

[78] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997.

[79] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[80] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[81] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[82] Daniel Lopez-Martinez, Neska El-Haouij, and Rosalind Picard. Detection of real-world driving-induced affective state using physiological signals and multi-view multi-task machine learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 356–361. IEEE, 2019.

[83] Pekka Siirtola. Continuous stress detection using the sensors of commercial smartwatch. *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*, 2019.

[84] Xuanyi Dong and Yi Yang. Nas-bench-201: Extending the scope of reproducible neural architecture search. *arXiv preprint arXiv:2001.00326*, 2020.

[85] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, Veronika Engert, Sonia J Lupien, and Jens C Pruessner. The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry and Neuroscience*, 30(5):319–325, 2005.

[86] Houtao Deng, George Runger, and Eugene Tuv. Bias of importance measures for multi-valued attributes and solutions. In *International conference on artificial neural networks*, pages 293–300. Springer, 2011.

[87] Martin Gjoreski. *Continuos Stress Monitoring using a Wrist Device and a Smartphone.* PhD thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, 09 2016.

[88] Han Yu, Thomas Vaessen, Inez Myin-Germeys, and Akane Sano. Modality fusion network and personalized attention in momentary stress detection in the wild. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[90] Han Yu and Akane Sano. Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild. *arXiv preprint arXiv:2202.12935*, 2022.

[91] Elena Smets. Towards large-scale physiological stress detection in an ambulant environment. 2018.

[92] Karel Mundnich, Brandon M Booth, Michelle l'Hommedieu, Tiantian Feng, Benjamin Girault, Justin L'hommedieu, Mackenzie Wildman, Sophia Skaaden, Amrutha Nadarajan, Jennifer L Villatte, et al. Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Scientific Data*, 7(1):1–26, 2020.

[93] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. Crosscheck: toward passive sensing and detection of mental

health changes in people with schizophrenia. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, pages 886–897, 2016.

[94] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. Overview of ntcir-12 lifelog task. In *NTCIR*, 2016.

[95] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, and Duc-Tien Dang-Nguyen. Overview of ntcir-13 lifelog-2 task. 2017.

[96] Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. Graph-based browsing for large video collections. In *International Conference on Multimedia Modeling*, pages 237–242. Springer, 2015.

[97] Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. Navigating a graph of scenes for exploring large video collections. In *International Conference on Multimedia Modeling*, pages 418–423. Springer, 2016.

[98] Jakub Loko, T. Soucek, Premysl Cech, and Gregor Kovalcík. Enhanced viret tool for lifelog data. In *LSC '19*, 2019.

[99] L. Rossetto, Ivan Giangreco, C. Tanase, and H. Schuldt. vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. *Proceedings of the 24th ACM international conference on Multimedia*, 2016.

[100] Nguyen-Khang Le, Dieu-Hien Nguyen, Trung-Hieu Hoang, Thanh-An Nguyen, Thanh-Dat Truong, Tung Dinh Duy, Quoc-An Luong, Viet-Khoa Vo-Ho, Vinh-Tiep Nguyen, and Minh-Triet Tran. Smart lifelog retrieval system with habit-based concepts and moment visualization. In *LSC '19*, 2019.

[101] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit,

et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017.

[102] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[103] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[104] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[105] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[106] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[107] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[108] Peter Oram. Wordnet: An electronic lexical database. christiane fellbaum (ed.). cambridge, ma: Mit press, 1998. pp. 423. *Applied Psycholinguistics*, 22(1):131–134, 2001.

[109] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. Myscéal 2.0: a revised experimental interactive lifelog retrieval system for lsc'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge*, pages 11–16. 2021.

[110] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[111] Naushad Alam, Yvette Graham, and Cathal Gurrin. Memento: A prototype lifelog search engine for lsc'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge*, LSC '21, page 53–58, New York, NY, USA, 2021. Association for Computing Machinery.

[112] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. E-myscéal: Embedding-based interactive lifelog retrieval system for lsc'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*, pages 32–37. 2022.

[113] S. Rajasekar, P. Philominathan, and V. Chinnathambi. Research methodology. 2006.

[114] J. Creswell. Educational reserach : planning, conducting, and evaluating quantitative and qualitative research. 2002.

[115] Stuart MacDonald and Nicola Headlam. *Research Methods Handbook: Introductory guide to research methods for social research.* Centre for Local Economic Strategies, 2008.

[116] family=George given i=T, given=Tegan. Exploratory research | definition, guide, 038; examples.

[117] Sirko Straube and Mario M Krell. How to evaluate an agent's behavior to infrequent events?—reliable performance estimation insensitive to class distribution. *Frontiers in computational neuroscience*, 8:43, 2014.

[118] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, et al. [invited papers] comparing approaches to interactive lifelog search at the lifelog search challenge (lsc2018). *ITE Transactions on Media Technology and Applications*, 7(2):46–59, 2019.

[119] Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors (Basel, Switzerland)*, 19, 2019.

[120] Pekka Siirtola. Continuous stress detection using the sensors of commercial smartwatch. In *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*, pages 1198–1201, 2019.

[121] Van-Tu Ninh, Sinéad Smyth, Minh-Triet Tran, and Cathal Gurrin. Analysing the performance of stressdetection models on consumer-grade wearable devices. In *SoMeT*, 2021.

[122] Laura J Julian. Measures of anxiety. *Arthritis care & research*, 63(0 11), 2011.

[123] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[124] Pekka Siirtola and Juha Röning. Comparison of regression and classification models for user-independent and personal stress detection. *Sensors*, 20(16):4402, 2020.

[125] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, pages 1–8, 2021.

[126] Weixuan 'Vincent' Chen, Natasha Jaques, Sara Taylor, Akane Sano, Szymon Fedor, and Rosalind W. Picard. Wavelet-based motion artifact removal for electrodermal activity. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6223–6226, 2015.

[127] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2016.

[128] Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine*, 16(2):279–286, 2011.

[129] Rahul Kher. Signal processing techniques for removing noise from ecg signals. In *Journal of Biomedical Engineering and Research*, 2019.

[130] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PLoS One*, 8(10):e76585, 2013.

[131] Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen S Quigley, Lisa F Barrett, and Sarah Ostadabbas. An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE journal of translational engineering in health and medicine*, 6:1–11, 2018.

[132] Kamath M. V. and Fallen E. L. Correction of the heart rate variability signal for ectopics and missing beats. In *Heart Rate Variability, eds M. Malik, Camm A. J*, 1995.

[133] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[134] Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33, 2011.

[135] Hans-Michael Kaltenbach. *A concise guide to statistics.* Springer Science & Business Media, 2011.

[136] Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Minh-Triet Tran, Liting Zhou, Pablo Redondo, Sinéad Smyth, and Cathal Gurrin. Lifeseeker: Interactive lifelog search engine at lsc 2019. In *LSC '19*, 2019.

[137] Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh-An Nguyen, Hai-Dang Nguyen, Liting Zhou, Graham Healy, and Cathal Gurrin. Lifeseeker 2.0: Interactive lifelog search engine at lsc 2020. *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*, 2020.

[138] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Sinéad Smyth, Annalina Caputo, and Cathal Gurrin. Lifeseeker 4.0: An interactive lifelog search engine for lsc'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*, LSC '22, page 14–19, New York, NY, USA, 2022. Association for Computing Machinery.

[139] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE*

transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017.

[140] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling cross stage partial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13029–13038, June 2021.

[141] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[142] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[143] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[144] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[145] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[146] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[147] David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178, 2010.

[148] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[149] Naushad Alam, Yvette Graham, and Cathal Gurrin. Memento 2.0: An improved lifelog search engine for lsc'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*, pages 2–7. 2022.

[150] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc Tien Dang Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. [invited papers] comparing approaches to interactive lifelog search at the lifelog search challenge (lsc2018). *ITE Transactions on Media Technology and Applications*, 7:46–59, 04 2019.

[151] Jakub Lokoč, František Mejzlik, Patrik Veselý, and Tomáš Souček. Enhanced somhunter for known-item search in lifelog data. In *Proceedings of the 4th Annual on Lifelog Search Challenge*, LSC '21, page 71–73, New York, NY, USA, 2021. Association for Computing Machinery.

[152] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. Voxento 2.0: A prototype voice-controlled interactive search engine for lifelogs. In *Proceedings of the 4th Annual on Lifelog Search Challenge*, LSC '21, page 65–70, New York, NY, USA, 2021. Association for Computing Machinery.

[153] Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, E-Ro Nguyen, Thanh-Cong Le,

Mai-Khiem Tran, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, and Minh-Triet Tran. Flexible interactive retrieval system 3.0 for visual lifelog exploration at lsc 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge*, pages 20–26. 2022.

[154] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. Voxento 3.0: A prototype voice-controlled interactive search engine for lifelog. In *Proceedings of the 5th Annual on Lifelog Search Challenge*, LSC '22, page 43–47, New York, NY, USA, 2022. Association for Computing Machinery.

[155] Caterina Cinel, Cathleen Cortis Mack, and Geoff Ward. Towards augmented human memory: Retrieval-induced forgetting and retrieval practice in an interactive, end-of-day review. *Journal of Experimental Psychology: General*, 147(5):632, 2018.

[156] Michael C Anderson, Robert A Bjork, and Elizabeth L Bjork. Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of experimental psychology: Learning, memory, and cognition*, 20(5):1063, 1994.

[157] A Bagula, K Kyamakya, and F Al-Machot. Emotion and stress recognition related sensors and machine learning technologies. 2021.