



Human Behaviour Sensing and Profiling in the Wild

A thesis submitted in fulfillment of the requirements for
the degree of Doctor of Philosophy

NAN GAO

M.Sc (Software Engineering)

B.Sc (Software Engineering)

Beijing Institute of Technology (BIT)

School of Computing Technologies

College of Science, Technology, Engineering and Maths

RMIT University

March 2022

This work is dedicated to my child CHENGXI. You have made me better, stronger and more fulfilled than I ever imagined. I love you to the moon and back.

Declaration

I certify that except where due acknowledgement has been made, this research is that of the author alone; the content of this research submission is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

In addition, I certify that this submission contains no material previously submitted for award of any qualification at any other university or institution, unless approved for a joint-award with another institution, and acknowledge that no part of this work will, in the future, be used in a submission in my name, for any other qualification in any university or other tertiary institution without the prior approval of the University, and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of any published works contained within this thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my research submission to be made available on the web, via the University's digital research repository, unless permission has been granted by the University to restrict access for a period of time.

Nan Gao

March 29 2022

Acknowledgements

It is my great pleasure to express my sincere gratitude to my supervisors Prof. Flora Salim, Dr. Wei Shao and Dr. Mohammad Saiedur Rahaman, for their underlying support, encouragement, and continuous efforts to improve my skills throughout my PhD journey. Specifically, I would like to thank Flora for being incredibly patient and showing many possibilities not in research, but also in life. She is the one who told me to aim high and not set limits for myself. I profoundly thank Wei for his patience, encouragement and guidance in my research. Not only did he advise me on the scientific aspects of my research, but he also diligently enhanced my confidence to accomplish my research and set appropriate goals at different stages. I would like to thank Saiedur a lot for his effective supervision, constructive criticism on the scientific research methods and academic writing. Without my supervisors, I would never be who I am today.

I would like to thank Prof. Yun Yang for leading me on the path of scientific research before my doctorate study. I thank my collaborators Prof. Jane Burry, Prof. Simon Watkins, Dr. Max Marschall, Judith Simone Heinisch, Dr. Christoph Anderson, Prof. Klaus David for their great suggestions and feedback on my research. I thank the members of Context Recognition and Urban Intelligence (CRUISE) research group: Shohreh Deldari, Sichen Zhao, Kyle Kai Qin, Kaixin Ji, Yonchanok Khaokaew, Arian Prabowo, Ali Ali, who have provided friendship and support, and with whom I have shared sweat, tears, joy and many wonderful moments during my PhD journey. Especially, I would like to thank my best friend, Eva Jin. I am grateful to the whole world for giving birth to such a warm, lovely, intelligent, talented, sincere, empathetic person and for bringing her into my life. Her mental support and care will always be treasured and remembered.

I would also like to thank my family. To my father, Chengxian Gao and my mother, Xuehua Li, who have always provided me the courage and mental strength while being eight thousand of kilometres away in China. My gratitude to my husband Dongwei Li for his unfailing support and efforts to the family. I also would like to express my thanks to my son, Chengxi Li, you have made me better, stronger, and more fulfilled than I ever imagined. People are mortal, but we can decide how to live. When that moment comes for you, I hope you have already found the career that you truly love and guard it.

Last but not least, I would like to thank my scholarship providers, RMIT University, Aurecon and the Australian Research Council for providing financial support and giving me opportunities to pursue a PhD in Australia. I hope this work would benefit the research community in the areas of ubiquitous computing and human-computer interaction.

Contents

Declaration	iii
Acknowledgement	iv
Statistical Summary	vi
Contents	vii
List of Figures	xiii
List of Tables	xvi
Abstract	1
1 Introduction	4
1.1 Motivation	6
1.2 Research Challenges	7
1.3 Research Questions	9
1.4 Research Contributions	10
1.5 Thesis Organisation	11
2 Data Collection and Ground Truth Validation	15
2.1 Introduction	16
2.2 Related Work	18
2.2.1 Inferring Emotions and Mental State using Sensing Technology	18

2.2.2	Reliability of Self-report Data	19
2.3	Data Collection	20
2.3.1	Participants and Recruitment	20
2.3.2	Experiment Setup	21
2.4	Reliability of Self-Report Data	25
2.4.1	Confidence Level of Responses	25
2.4.2	Completion Time and Reliability	26
2.4.3	Perceived vs Physiological Measurements	28
2.4.4	Discussion	30
2.5	Conclusion	31
3	Modelling User Engagement Behaviours from Wearable and Environmental Sensors	33
3.1	Introduction	34
3.2	Related Work	37
3.2.1	Traditional Methods for Measuring Engagement	37
3.2.2	Engagement Prediction with Sensing Technology	38
3.3	Dataset	40
3.3.1	Participants and Procedures	40
3.3.2	Collected Data	41
3.3.2.1	Physiological and Activity Data	41
3.3.2.2	Indoor Environmental Data	42
3.3.2.3	Ground Truth: Self-report Survey Instrument Data	43
3.4	Data Preprocessing	44
3.4.1	Class Period Segmentation	45
3.4.2	Data Cleaning	46
3.4.3	Data Pre-processing	47
3.5	Feature Extraction	47
3.5.1	EDA-based Features	48
3.5.2	HRV-based Features	49

3.5.3	Accelerometer-based Features	49
3.5.4	Other Features	50
3.6	Prediction Pipeline	50
3.7	Results and Discussion	52
3.7.1	Overall Prediction Results	52
3.7.2	Impact of Sensor Combinations	55
3.7.3	Impact of Class Subjects	58
3.7.4	Discussion	60
3.8	Implications and Limitations	62
3.9	Conclusion	65
4	Understanding Classroom Seating Behaviours from Perceived and Physiologically Measured Student Engagement	66
4.1	Introduction	67
4.2	Related Work	70
4.2.1	Student Engagement in Educational Research	70
4.2.2	Classroom Seating Experience and Student Engagement	71
4.2.2.1	Flexible Seating in the Classroom	71
4.2.2.2	Seating Preference and Student Engagement	71
4.2.2.3	Seating Proximity and Student Engagement	73
4.2.3	Inferring Student Engagement Using Sensing Technologies	74
4.3	Data Collection	75
4.3.1	Student Multi-dimensional Engagement	75
4.3.2	Seating Location	75
4.3.3	Physiological Signals	77
4.4	Overview of Seating Experience, Student Engagement and Physiological Patterns	77
4.4.1	Seating Locations and Seating Preference	78
4.4.2	Student Engagement and Physiological Signals	80
4.5	Results	81

4.5.1	Relationship Between Seating Behaviours and Perceived Student Engagement	82
4.5.2	Relationship Between Seating Behaviours and Physiological Synchrony	85
4.5.3	Relationship Between Seating Behaviours and Physiological Arousal	88
4.6	Limitations and Implications	92
4.7	Conclusion	94
5	Profiling Individual Personality Traits and Response Behaviours from Mobile Sensing Data	95
5.1	Introduction	96
5.2	Related Work	99
5.2.1	Inferring Personality Traits through Mobile Sensing	99
5.2.2	Interruptibility Management and Receptivity	100
5.2.3	Mood Sensing Approaches	101
5.3	Inferring Personality Traits with Mobile Computing	103
5.3.1	Methodology	103
5.3.1.1	Participants and Procedure	103
5.3.1.2	Activity Behaviour Metrics	104
5.3.1.3	Big Five Personality Ground Truth	106
5.3.2	Feature Analysis	107
5.3.3	Predictive Analysis	111
5.4	Inferring Response Behaviours with Mobile Computing	113
5.4.1	Data Collection	113
5.4.1.1	Overview	113
5.4.1.2	Participants	114
5.4.1.3	Collected data	115
5.4.2	Methodology	117
5.4.2.1	Pre-processing Approaches	117
5.4.2.2	Extracted features	118

5.4.3	Understanding the Mood, Usage Behaviours and Notification Response Time of Participants	122
5.4.3.1	Understanding Notification Response Times for Different Participants	122
5.4.3.2	App Categories and Response Time	123
5.4.3.3	Impact of Applications on Notification Response Times	124
5.4.3.4	The Mood of Users and Notification Response time	125
5.4.4	Experiment	127
5.4.4.1	Prediction Pipeline	127
5.4.4.2	Prediction Result with Mobile Data	129
5.4.4.3	Impact of Mood-related Features	131
5.4.5	Implications and Limitations	132
5.5	Conclusion	134
6	Modelling Thermal Comfort with Limited Labelled Data in Smart Buildings	135
6.1	Introduction	135
6.2	Related Work	138
6.2.1	Traditional Thermal Comfort Modelling Methods	138
6.2.2	Transfer Learning Applications	140
6.3	Data Sets Introduction	141
6.3.1	Overview	141
6.3.2	Preliminary Analytics	144
6.4	Methodology	146
6.4.1	Problem Definition	146
6.4.2	Feature Selection	147
6.4.3	Imbalance Class Distribution	149
6.4.4	Thermal Comfort Modelling	151
6.5	Experiment	152
6.5.1	Experimental Setup	152

6.5.2 Overall Prediction Result	157
6.5.3 Impact of Different Feature Combinations	159
6.5.4 Impact of the Number of Hidden Layers	161
6.6 Conclusion	162
7 Conclusion	164
7.1 Research Questions and Answers	165
7.2 Future Research Directions	168
Bibliography	171
Appendix A: Ethics Approval Documents	217
Appendix B: Credits	222

List of Figures

1.1	An illustration of how human behaviours can be inferred from heterogeneous sensing data	5
1.2	Overview of the thesis structure and organisation	11
2.1	Devices and environments for collecting wearable and indoor data	22
2.2	Distribution of responses related to thermal comfort	24
2.3	Distribution of responses related to engagement and emotions	24
2.4	Distribution of the survey responses across hours of the day and days of the week	25
2.5	Distribution of confidence levels for all self-report responses	26
2.6	Confidence levels across all participants	26
2.7	Survey completion time of participants	27
2.8	Linear regression of survey completion time with confidence levels	27
2.9	The distribution of overall engagement across student participants	28
2.10	An example of the changes in electrodermal activity of three different participants, P15, P17 and P20, in the same class (their perceived engagement levels were 4.2, 3.2 and 4.4, respectively)	29
3.1	Temperature and CO ₂ data in R1, R2, R3 (room 1, room 2, room 3) at 11:00 am on 11 Sep 2019. Room 4 is not shown here as it is in another building	42
3.2	Histograms of the Answers. The X axis shows the 5-Likert scale from -2 to 2 which means ‘strongly disagree’ to ‘strongly agree’. The Y axis shows the number of the responses that fall into the specific scale	44
3.3	Calculated class end time with ACC data from 12 student participants	45

3.4	Boxplot of the overall engagement scores for 23 student participants. The red dashed line represents the average score for all participants. The participant ID shown in the figure is randomly generated to maintain the privacy of participants	54
3.5	Prediction error for overall engagement scores for 23 student participants	54
3.6	Prediction performance for the average subject model and general model	59
3.7	Engagement scores on different subjects	59
3.8	Engagement scores with different class time and thermal comfort	61
3.9	Survey completion time for different participants. Each point represents the survey completion time for one response	64
4.1	The screenshot of the self-report survey.	76
4.2	Overall seating distribution of students in the classrooms	77
4.3	Seating location across five different student participants	78
4.4	Three different seating locations (back, right and left) in classrooms	79
4.5	Seating preference for each participant in all courses	79
4.6	Seating preference for each participant in each course	80
4.7	The overview of engagement score across participants and courses	81
4.8	Seating and occurrence information for pairs of students	83
4.9	Overall engagement and seating distribution over four example classes	84
4.10	The distribution of seating preferences between student groups with different physiological arousal	90
4.11	The clustering results of EDA signals for long and short classes	90
5.1	Average Big-5 personality scores	103
5.2	Kernel density distribution of five personality traits in the data set	106
5.3	The number of notifications across all the apps for participant P10	122
5.4	Cumulative distribution of notification response times from the top 10 apps for all participants	123
5.5	Information for different app categories	124

5.6	Cumulative distribution of notification response times from five apps for participant P10	125
5.7	Distribution of arousal and valence for 18 participants	125
5.8	Mood of participants at different levels of interruptibility and various times of the day	126
5.9	Mood of participants over different social roles and days of the week	126
5.10	Prediction results across different regressors for each participant	130
5.11	Feature importance for each participant in the prediction	131
6.1	Six factors affecting thermal comfort (PMV model)	137
6.2	Locations of different studies in ASHRAE RP-884 database, The Scales Project database and Medium US Office dataset	142
6.3	Distribution of thermal sensation over different datasets	143
6.4	Distribution of the indoor air temperature over different domains	144
6.5	Boxplots of thermal sensation and the indoor temperature	145
6.6	Thermal comfort transfer learning system	148
6.7	The architecture for thermal comfort transfer learning	152
6.8	‘Köppen World Map High Resolution’ by Peel, M. C. et al. [1], licenced under Creative Commons Attribution-Share Alike 3.0 Unported [2], Desaturated from original	155
6.9	Confusion matrix on the target domain	159
6.10	Prediction performance with different number of hidden layers	161

List of Tables

2.1	Publicly available datasets in affective computing	17
2.2	Distribution of student participants in different class groups	21
2.3	Collected annotations from the questionnaires	23
3.1	Related work for engagement prediction with sensing data	38
3.2	Self-report items for measuring in-class engagement in online survey	43
3.3	Description of the features computed for different sensors	48
3.4	Prediction performance for emotional, cognitive, behavioural, and overall engagement with all sensing data	53
3.5	The most influential features on multidimensional engagement	55
3.6	Summary of the Prediction performance of multidimensional engagement using different sensor combinations. \mathcal{X}_1 indicates all the wearable data including EDA, HRV, ACC and ST data, and \mathcal{X}_2 means the indoor environmental data including CO ₂ and temperature data	57
3.7	Multidimensional engagement regression result for different subjects	58
4.1	Related works that studied student seating experience and learning engagement, performance and emotion	72
4.2	An overview of the number of wearable signals across all courses	81
4.3	Summary of the Pearson rank correlation results between proximity of seating and physiological synchrony across the courses. The asterisks indicate the statistically significant results: *p < 0.05, **p < 0.01, ***p < 0.004	86
4.4	Description of the features computed for electrodermal activity signals.	88

4.5	The learning engagement of student groups with different courses and class lengths	91
5.1	Overview of the Big Five scores for total/male/female participants	107
5.2	Description of the extracted features	108
5.3	Most useful features to predict personality traits (total population)	109
5.4	Most useful features to predict personality traits (female and male population) .	110
5.5	Prediction performance for total/male/female participants	113
5.6	Extracted features by device. Data marked with (*) were manually reported . . .	119
5.7	Notification and app information for 18 participants	121
5.8	Predictive results with different regressors using mobile data	131
6.1	Information for source dataset and target dataset	142
6.2	Selected features in the Medium US Office dataset	147
6.3	Classification of the ASHRAE RP-884 database for HVAC buildings according to climates	156
6.4	Prediction performance for different algorithms on the target dataset	157
6.5	Prediction performance for different feature sets on the target dataset	160

Abstract

With advances in sensors, wearables and the Internet of Things, it has become more and more convenient to gather information from human daily life, which has promoted the development of human behavioural sensing technology. In general, heterogeneous sensing data (e.g. behavioural, environmental and physiological sensing signals) may come from different sources (e.g. mobile phones, buildings, weather stations and wearables). From this information, it is possible to infer multiple human behaviours and psychological states, such as personality, thermal comfort and learning engagement. Sensing and profiling human behaviours has many advantages, such as supporting medical diagnosis, improving self-awareness, creating supportive study/work environments and taking timely measures to promote human wellbeing.

However, human behaviour sensing is a complex task with some key challenges: (1) Limited sources of sensing data: Previous research has primarily explored one type or a limited number of types of sensing data (e.g. accelerometer data and heart rate signals) to build predictive models rather than incorporating sensing data from multiple sources. (2) Lab-based settings: Most studies have been conducted in environments specifically designed for research; however, field experiments are more likely to reflect real-world human behavioural patterns due to the authenticity of natural settings. (3) Difficulty in validating the ground truth: Self-report surveys are generally considered to be measures of ground truth in human-based research but may be prone to subjectivity and various types of response bias. (4) Difficulty in depicting dynamic behaviours: Human behaviours are dynamic and complex in heterogeneous environments, making it difficult to accurately depict them. (5) Shortage of annotations: Traditional self-report surveys are the most popular way to understand human behaviours, but they are both time consuming and labour-intensive, resulting in insufficient annotations and difficulty

in creating effective models. In this thesis, we address the above challenges and make the following contributions.

First, we address the challenge of the limited sources of sensing data in the wild. Using wearable sensors to log physiological data and daily surveys to query the participants' thermal comfort, learning engagement, emotions and seating behaviours, we will collect data from 23 high school students and six teachers participating in 11 courses (144 classes) over a four-week period. We will then explore the validity of the collected data to ensure that we can reliably profile human behaviours using heterogeneous sensing data collected in the wild.

Second, we will explore wearable and environmental sensing data to understand learning engagement in the wild. With the data previously collected in the wild, we will create a classroom sensing system to automatically measure the multidimensional engagement (i.e. behavioural, emotional and cognitive engagement) of high school students during classes. In particular, we will combine physiological signals, physical activities and indoor environmental data to estimate changes in student engagement levels. To the best of our knowledge, this will be the first system for detecting multidimensional engagement from multiple sensors in the wild.

Third, we will investigate group behaviours to understand social relationships using physiological sensors. We will explore how the group-wise seating experience relates to student engagement by examining the participants' physiological arousal and synchrony. We will investigate whether students sitting close together are more likely to have similar learning engagements and greater physiological synchrony than students sitting far apart. This research has the potential to assist in maximising student engagement by providing more flexible and intelligent seating arrangements in the future.

Fourth, we will employ unobtrusive mobile sensing for dynamic user behavioural modelling. Two real-world tasks (modelling Big Five personality traits and notification response behaviours) will be explored based on the participants' mobile phone usage behaviours. A comprehensive study on a real-world dataset will demonstrate whether it is possible to utilise smartphone usage behaviours to predict users' Big Five personality traits. In order to estimate response times, we will investigate whether the established regression model can accurately

predict the response time to notifications using the user's mood and physiological signals. Our research will shed light on the future intelligent notification management system for mobile users.

Finally, we will model aggregated behaviour (i.e. thermal comfort) using environmental sensing with limited annotations by transferring knowledge from multiple locations to another domain. We will build a transfer learning framework and confirm that thermal comfort sensor data from multiple cities in the same climate zone can be used to improve the small thermal comfort dataset of a target building that has insufficient training data. Extensive experimental results will show that the proposed models outperform the state-of-the-art algorithms for thermal comfort prediction and can be implemented in any building, even if adequate thermal comfort labelled data are not available.

In summary, this thesis provides several contributions to profiling and modelling human behaviours in the wild. This research will exploit various types of sensing data from multiple sources in different real-world tasks to address common challenges in the area of human behavioural modelling. We will also publish the largest and most diverse dataset collected in the wild to better understand participants' behaviour, engagement, emotion and comfort using heterogeneous sensors and wearables. This will benefit building scientists, behavioural psychologists and ubiquitous computing researchers in the future. Overall, we believe this research will provide a significant contribution to human-based sensing and behavioural profiling in the wild that will make researchers, managers and policymakers more aware of occupants/users and more able to adapt to their needs.

Chapter 1

Introduction

A sensor is a device that detects changes in quantities, such as acceleration, pressure, light and temperature. With the advancements in sensing technologies, sensors have become smaller, lighter and more accurate, which allows them provide large amounts of data almost instantaneously from anywhere in the world. Sensors can be attached to fixed objects (e.g. weather stations) or individuals under observation (e.g. in smartphones or wearable sensors) [3]. Time series of digital tracking information are produced from various internet of things (IoT) devices, making it increasingly convenient to gather information from people's daily lives.

The advances in and increased maturity of both the hardware and software involved in sensor technology have facilitated the development of the field of human behaviour sensing [4]. Human behaviour sensing refers to the collection and analysis of data from sensors embedded in daily life with the purpose of inferring human behaviours, feelings, thoughts, traits, etc. from the data collected [5]. Over the past decades, it has become a popular research topic, playing an important role in the fields of education [6, 7], transportation [8], ambient assisted living [9, 10], pervasive and mobile computing [11, 12], etc.

Sensing and monitoring human behaviours to support medical diagnosis, disease surveillance, epidemic outbreak tracking and chronic disease management [13] have been shown to benefit the traditional clinical techniques. In addition, human sensing can improve self-awareness, assist in creating the right study/work environments and help managers or policymakers take timely measures to improve human wellbeing [14, 15]. Notably, with advance-

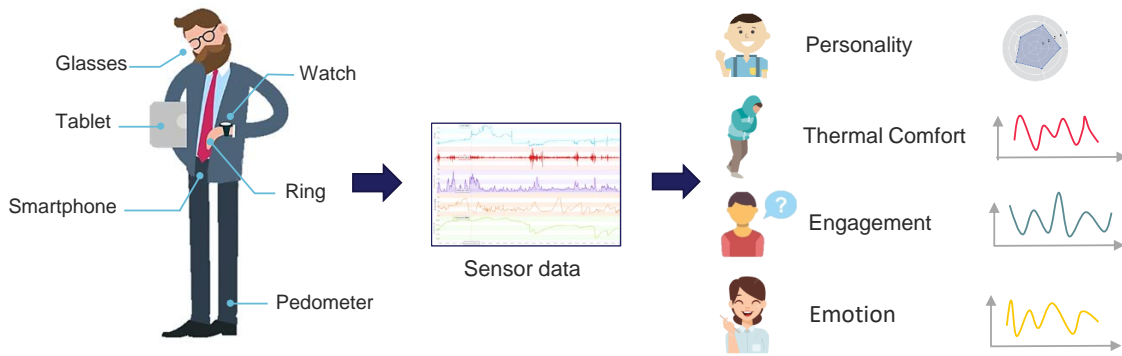


Figure 1.1: An illustration of how human behaviours can be inferred from heterogeneous sensing data

ments in wearable sensors, humans are being encouraged to adopt healthier lifestyles, such as adequately brushing teeth [16], properly washing hands during the COVID-19 pandemic [17], accurately tracking food intake [18, 19] and significantly improving medication compliance [20].

Figure 1.1 illustrates the general process for inferring human behaviours from sensing data. First, the various types of sensing data (e.g. behavioural, environmental and physiological) are collected from multiple sources (e.g. mobile phones, wearables and weather stations). The signals from the sensors are then translated into specific human psychological states and characteristics, such as engagement, emotion, thermal comfort and personality traits. This translation process is achieved by building machine learning (ML) prediction models for human behaviours using sensor data and ground truth data (e.g. self-report survey responses) as the input and output. Once the model is built, human behaviours can be inferred and monitored using only the sensor data.

Despite research efforts over the years, accurate human behaviour sensing in the wild still faces enormous challenges. The effectiveness of human sensing is influenced by multiple factors, such as shortage in annotations, limited sources of sensing data and difficulties validating the ground truth and depicting dynamic behaviours. In addition, most existing research was conducted in the lab, and real-world data are very scarce. To mitigate these issues, in this thesis, we will process various types of pervasive data, including data from wearables, environmental sensors and mobile signals, to analyse human behaviours and discover hidden patterns. Multi-

ple human responses and behaviours (e.g., emotion, seating preferences, notification responses and thermal comfort) will be explored in different real-world scenarios using self-collected data and publicly available datasets. The contributions of this research will verify the importance of sensing and profiling human behaviours in the wild.

1.1 Motivation

Currently, various aspects of people's lives are being continuously measured by sensors embedded in multiple locations (e.g. mobile phones, wearables, buildings and weather stations), promoting the movement of the *quantified self* and improving people's self-awareness and wellbeing. The *quantified-self* seeks self-knowledge through self-tracking. This idea was formed in the 1970s [21] and expanded rapidly from 2015 with advances in consumer-friendly mobile and wearable sensing technologies. The *quantified-self* usually uses life-recording practices and trends for data acquisition, with the aim of improving their physical/mental health and wellbeing. Recently, accurate self-tracking and self-awareness have become possible with the widespread adoption of mobile and wearable fitness/sleep trackers (e.g. Fitbit [22], Apple Watch [23] and Oura Ring [24]) and the increasing popularity of IoT devices in smart homes and buildings (e.g. Nest Thermostat [25] and Netatmo [26]).

The ability to sense and profile human behaviours is necessary for the design of intelligent feedback systems for users or managers in various scenarios. For instance, with the help of a classroom engagement sensing system, teachers can take timely measures to improve the learning experience for students (e.g. plan learning schedules, re-engage students with low engagement or ventilate the classroom to let fresh air in). Such a sensing system has the potential to greatly contribute to improving student achievement and decreasing the school dropout rate. Another example is the thermal comfort controlling system. By sensing the thermal comfort of the occupants of a building accurately, this system can maintain a comfortable environment to optimise the wellbeing of the occupants while minimising energy usage.

1.2 Research Challenges

In this research, we aim to address key research challenges for human behaviour modelling and profiling in the wild. The rapid growth in various sensor technologies and the possibility of mass-producing sensors in an economical manner has made it possible to collect heterogeneous sensing data to enable us to understand complex human behaviours. However, the sensor data collected from various types of sensors in natural settings are usually primitive and heterogeneous in format and storage [3]. Such datasets usually lack descriptions and are ad hoc, making it difficult to share and reuse data from them. When researchers analyse these datasets, they face challenges during data acquisition and processing, such as data sparsity, data heterogeneity, reliability of the self-report responses of users, imbalanced data distribution and limited annotations. Even when a dataset is well structured with sufficient descriptions, it is still difficult to build an effective predictive model due to the complexity and dynamic nature of human behaviours. In summary, some common challenges for human behaviour sensing and profiling are outlined below:

One of the challenges of human behaviour sensing and profiling is gathering heterogeneous sensing data in an unobtrusive way and collecting the *ground truth* of human behaviours and mental states. Collecting sensing data from different modalities at the same time requires taking multiple factors into account (e.g. privacy, storage and battery) and is expensive, especially in the natural environment. Generating ground truth labels for human behaviours and mental states is challenging due to the nature of these phenomena [27]. Unlike traditional tasks, such as human activity recognition or face recognition, it is often difficult or impractical for researchers to identify the real *ground truth* of mental states in human-based studies using traditional methods (e.g. annotations from videos/images, transaction records and GPS trajectories). The most commonly used methods for measuring behavioural traits and mental states is asking participants to respond to self-report surveys [6, 28] or conducting an Ecological Momentary Assessment (EMA). These are generally regarded as the measures of ground truth [29, 6, 30, 31, 32] in prediction models but may be prone to subjectivity and response bias.

Depicting dynamic and complex human behaviours and states in different scenarios using sensing data is a challenging task. Human behaviours and states are affected by multiple

factors, such as social relationships, time and physical spaces. Therefore, they are usually capricious, dynamic and multi-granular [3]. To build an effective predictive model, extracting features from cleaned sensing signals is one of the most important steps. In general, statistical features are not sufficient to describe complex behavioural patterns or emotional arousal from sensing signals, and various high-level features based on curve fitting or domain transformation [33] need to be carefully constructed to accurately depict human behaviour. The establishment of effective features usually requires a large amount of cross-domain background knowledge, especially for physiological signals, such as electrodermal activity (EDA) and photoplethysmography (PPG) signals. In general, there is no proper feature extraction method that has a good practical effect for profiling human behaviours in various contexts. Therefore, an effective method for accurately and effectively extracting robust features from physiological, environmental and behavioural sensing signals is a challenge that has not yet been overcome.

The shortage of annotations is a critical issue that needs to be solved in human sensing studies. In recent years, many researchers have applied data-driven ML techniques to human behaviour modelling (e.g. thermal comfort [14], personality [12] and learning engagement [30]). However, it is usually difficult to obtain sufficient labelled data, especially in human-based studies, due to the limited budget for recruiting participants. The shortage of labelled data undoubtedly limits the performance of data-driven models. Transfer learning allows researchers to create an accurate model from previous tasks [34]. This technique has been applied to many real-world applications involving image/video classification, natural language processing, recommendation systems, etc. Although a few researchers [35, 36, 35] have started to use transfer learning to build human behaviour models, their target datasets are usually collected from laboratory studies. In addition, most researchers have utilised additional sensing devices (e.g., thermal cameras [37], eyeglasses [38] and wristbands [35]). Therefore, it is critical to solve the issue of how to apply transfer learning to human-based studies with limited sensing data.

In summary, the fundamental challenges surrounding human behaviour modelling in the wild are identified as follows:

- Collecting heterogeneous sensing data from multiple modalities and validating the ground

truth of human behaviours, traits and mental states.

- Depicting dynamic and complex human behaviours in different scenarios.
- Understanding human behaviours with limited self-report annotations.
- Modelling real-world human behaviours in natural settings instead of lab-based settings.
- Incorporating sensing data from heterogeneous environments and multiple sources to create a robust prediction model.

1.3 Research Questions

To address the aforementioned research challenges, the following research questions (RQs) are defined, with the goals of performing accurate human behaviour sensing and profiling in the wild.

***RQ-1.** How to capture and validate multidimensional human behaviours and states using heterogeneous sensors in the wild?*

This research question addresses the challenges related to capturing multidimensional behaviours (e.g. seating patterns, engagement, emotion and thermal comfort) of people (e.g. students and teachers) using wearable and environmental sensors in the wild. Specifically, this research question explores the validation of collected data to ensure that human behaviours can be reliably profiled using heterogeneous sensing data collected in the wild.

***RQ-2.** How to model and predict people's emotional, cognitive and behavioural engagement using wearable and environmental sensor data?*

This research question aims to predict user engagement by using wearable and environmental sensing data to build an inference model. We will ask students questions like whether they paid attention in class, pretended to participate in class but actually did not, enjoyed learning new things in class, etc. Using the data collected from **RQ-1**, we will focus on predicting the multiple dimensions of student engagement including emotional, behavioural and cognitive engagement in class. Specifically, we will extract novel features to represent the multidimensional

factors influencing student engagement in various classes across different subjects to predict student learning engagement.

RQ-3. *How to explore the effects of individual and group behaviours (e.g. seating patterns) on people’s perceived and physiologically measured engagement in different courses?*

This research question explores how group-wise seating experiences relate to student engagement in various subjects by understanding their physiological arousal and synchrony. Based on the data from **RQ-1**, we will extract features from the physiological signals that depict student engagement in class. We will then investigate the correlation between group seating behaviours and perceived and physiologically measured engagement.

RQ-4. *How to utilise mobile sensing to profile personality traits and receptivity to interruptions among different user groups?*

In **RQ-1**, **RQ-2** and **RQ-3**, we mainly focus on using environmental and wearable sensing data for human behaviour modelling, which requires the installation of specific sensors or wearing of wristband devices. This research question considers situations in which unobtrusive mobile sensing is used for behaviour modelling. The Big Five personality traits and notification response behaviours of users will be explored and predicted based on their mobile usage behaviours.

RQ-5. *How to model aggregate behaviour (e.g. thermal comfort) from environmental sensing data with limited annotations?*

It is often difficult to obtain sufficient annotations from self-report surveys, which limits the performance of data-driven prediction models of human behaviours. This research question is designed to address this challenge in human-based studies. A transfer learning framework is proposed for accurate thermal comfort modelling with limited labelled data by transferring knowledge from multiple locations.

1.4 Research Contributions

Based on the aforementioned research questions, the contributions of this thesis are as follows:

1. Publishing the largest heterogeneous environmental and affect sensing dataset and dis-

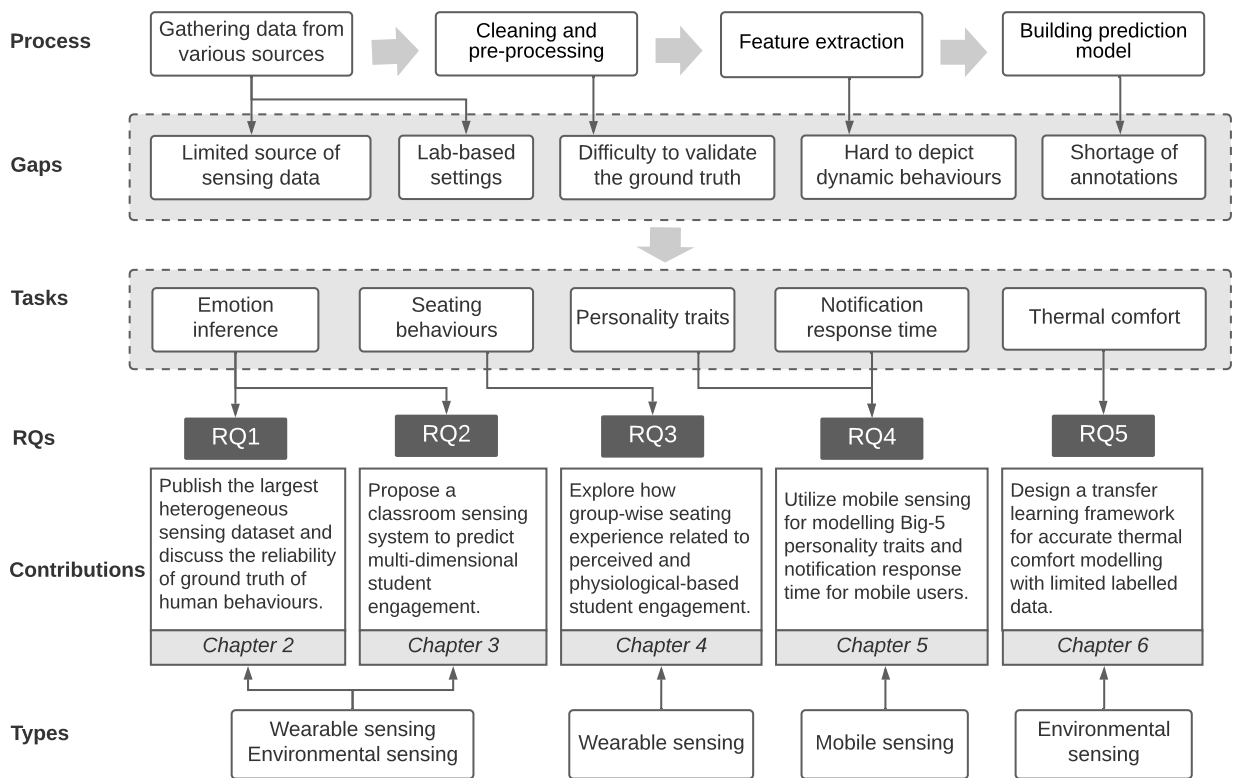


Figure 1.2: Overview of the thesis structure and organisation

cussing the reliability of ground truth for human behaviours.

2. Integrating human behaviour dynamics with domain knowledge in multiple real-world scenarios.
3. Modelling occupant multidimensional engagement with physiological and environmental sensing.
4. Predicting personality traits and response behaviours of people using mobile sensing.
5. Designing a transfer learning model for thermal comfort modelling across different cities.

1.5 Thesis Organisation

The remaining chapters of this thesis are organised as follows:

- **Chapter 2: Data Collection and Ground Truth Validation.** To answer *RQ-1*, this chapter describes the heterogeneous data collection using wearables and environmental sensors on a high school campus and discusses the reliability of self-reported data collected in the wild. The dataset that we collect in this chapter is further used in Chapter 3 and Chapter 4.

Copyright/credit/reuse notice: The contents of this chapter are taken and revised as needed from a paper published as

- Gao, N., Marschall, M., Burry, J., Watkins, S., & Salim, F. D. (2022). *Understanding Occupants' Behaviour, Engagement, Emotion, and Comfort Indoors with Heterogeneous Sensors and Wearables*. *Scientific Data*, 9(1), 1-16. DOI:[10.1038/s41597-022-01347-w](https://doi.org/10.1038/s41597-022-01347-w) [39] (**Impact Factor: 8.501, SJR: Q1**).
 - Gao, N., Rahaman, M. S., Shao, W., & Salim, F. D. (2021). *Investigating the Reliability of Self-report Survey in the Wild: The Quest for Ground Truth*. 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* (pp. 237–242). DOI:[10.1145/3460418.3479338](https://doi.org/10.1145/3460418.3479338) [40] (**Workshop at a Ubicomp 2021**).
- **Chapter 3: Modelling User Engagement Behaviours from Wearable and Environmental Sensors.** Based on *RQ-2*, a classroom engagement sensing system called *n-Gage* is presented. This system can automatically measure the multidimensional engagement (behavioural, emotional and cognitive engagement) of high school students during class. It combines wearable data (physiological signals and physical activities) and indoor environmental data to estimate changes in student engagement levels. Novel features are presented to indicate the physiological and physical synchrony between students, which is useful for predicting student engagement.

Copyright/credit/reuse notice: The contents of this chapter are taken and revised as needed from a paper published as

- Gao, N., Shao, W., Rahaman, M. S., & Salim, F. D. (2020). *n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild*. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3), 1–26. DOI:[10.1145/3411813](https://doi.org/10.1145/3411813) [30] (**Distinguished Paper Award**).

- **Chapter 4: Understanding Classroom Seating Behaviours from Perceived and Physiological-based Student Engagement.** This chapter describes the solution to *RQ-3*. We demonstrate the phenomenon that the individual and group-wise classroom seating experience is associated with perceived student engagement and physiological-based engagement measured from electrodermal activity signals. We also find that students sitting near each other are more likely to have similar physiological arousal and synchrony than students sitting far from each other.

Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from a paper published as:

- Gao, N., Rahaman, M. S., Shao, W., J, K., & Salim, F. D. (2021). Individual and Group-wise Classroom Seating Experience: Effects on Student Engagement in Different Courses (**Under Review in IMWUT, Major Revision**).

- **Chapter 5: Profiling Individual Personality Traits and Response Behaviours using Mobile Sensing Data.** In this chapter, two prediction models are presented in relation to the challenges stated in *RQ-4*. Two real-world case studies are presented to illustrate and explore mobile sensing in different scenarios. The solutions associated with profiling Big Five personality traits and notification response behaviours using mobile sensing are introduced.

Copyright/credit/reuse notice: The contents of this chapter are taken and revised as needed from a paper published as

- Gao, N., Shao, W., & Salim, F. D. (2019). *Predicting Personality Traits From Physical Activity Intensity*. *IEEE Computer*, 52(7), 47–56. DOI:[10.1109/MC.2019.2913751](https://doi.org/10.1109/MC.2019.2913751) [12] (**Impact Factor: 4.419, SJR: Q1**)

- *Heinisch, J. S., Gao, N., Anderson, C., DelDari, S., David, K., & Salim, F. D. (2022). Investigating the Effects of Mood & Usage Behaviour on Notification Response Time [41] (Co-first Authors, To be Submitted to IMWUT).*

- **Chapter 6: Modelling Thermal Comfort using Limited Labelled Data in Smart Buildings.** This chapter presents the solution to *RQ-5*.

The proposed transfer learning framework can deal with the data shortage problem in thermal comfort modelling by transferring the knowledge from similar thermal environments to a target building. In addition, we improve the predictive performance and build meaningful classifiers by using a GAN-based resampling method to imbalance the class distribution of occupants' thermal sensation.

Copyright/credit/reuse notice: The contents of this chapter are taken and revised as needed from a paper published as

- *Gao, N., Shao, W., Rahaman, M. S., Zhai, J., David, K., & Salim, F. D. (2021). Transfer Learning for Thermal Comfort Prediction in Multiple Cities. Building and Environment, 195, 107725. DOI:10.1016/j.buildenv.2021.107725 [14] (Impact Factor: 4.820, SJR: Q1).*

- **Chapter 7: Conclusion.** This chapter concludes the thesis by summarising the main contributions and limitations of the proposed methods. In addition, it discusses the implications and future directions of sensing human behaviours in real-world scenarios.

In summary, the core chapters (Chapter 2–6) contribute to several key research questions in human behaviour sensing and modelling in the wild. The main contributions and connections are detailed in Figure 1.2. Note that the core chapters (excluding Chapter 2) are complete and self-explanatory and include real-world scenarios, tasks and types of sensing data. Therefore, the relevant content is presented in each chapter separately, including an introduction and background for the chapter, related works, extracted features, developed models, experimental settings and results.

Chapter 2

Data Collection and Ground Truth Validation

As discussed in Chapter 1, the intelligent analysis and prediction of human behaviours is very important for a variety of reasons, including improving self-awareness, creating the right study/work environments and adopting healthier lifestyles. However, collecting sensing data from different modalities simultaneously requires the consideration of multiple factors (e.g. privacy, storage and battery) and is very expensive in natural environments. In addition, generating ground truth annotations of human behaviours and mental states is challenging due to the nature of these phenomena.

In addressing these issues and in reference to *RQ-1*, as presented in Section 1.3, this chapter presents heterogeneous data collection using wearables, environmental sensors and self-report tools for human behaviours and mental states. Specifically, we conducted a field study at a K–12 private school in the suburbs of Melbourne, Australia. We tracked 23 students and six teachers in a four-week cross-sectional study, using wearable sensors to log physiological data and self-report surveys to query the participants’ thermal comfort, learning engagement, emotions and seating behaviours. The dataset could be used to analyse students’ behaviours/mental states on campus and provide opportunities for the future design of intelligent feedback systems to benefit both students and staff.

The reliability of the self-report data of human behaviours and mental states is discussed by

studying the participants' confidence levels in the responses and the survey completion time. We find that the physiologically measured and perceived student engagement were not always consistent. This serves as a wake-up call for emotional and mental state sensing research in the ubiquitous computing (ubicomp) community, which usually regards self-report annotations as the ground truth for predicting human mental state.

2.1 Introduction

With advancements in wearables and IoT devices, sensing technologies have been increasingly investigated for predicting human emotional and mental characteristics in terms of mood [32, 42], depression [43, 44], stress [29], engagement [30, 45, 6], concentration [15], personality traits [46, 12] etc. Understanding peoples' engagement, emotions and daily behaviours using sensing technologies has attracted increasing interest to address problems such as low productivity and disaffection and could help in designing intervention strategies to prevent mental health issues and improve people's wellbeing.

In previous studies, various physiological signals, such as EDA and heart rate variability (HRV), and environmental data have been investigated to assess emotional arousal and engagement levels. For example, EDA is generally regarded as a good indicator of psychological arousal, which has been increasingly studied in the affective computing area, such as the detection of engagement [30, 6], emotion [47] and depression[48]. However, existing datasets in affective computing either provide limited scope for understanding emotional responses in real-world settings or only consider a particular type of annotation to meet their research goals (e.g. stress level and mental workload). Therefore, our first aim is to collect a heterogeneous dataset, including data from wearables, environmental sensors and various annotations of human behaviours and mental states in real-world settings. Table 2.1 shows how the proposed dataset *En-Gage* is distinguished from existing related datasets.

In human-based data collection, one of the most commonly used methods for measuring emotions and mental state is asking participants to respond to self-report surveys [12, 6, 28]. An alternative to the self-report survey is the EMA, which is designed to repeatedly collect human responses in real-time in natural settings. When building an ML prediction model, responses to

Name	Year	Par	Type	Modalities	Annotations	Duration	Scenario
<i>Driving-stress</i> [49]	2005	24	Field	ECG, EDA, EMG, RESP	Stress level	>50 mins	Real-world driving tasks
<i>DEAP</i> [50]	2011	32	Lab	Videos, EEG, EDA, BVP, RESP, ST, EMG and EOG	Arousal, valence, like/dislike, dominance, familiarity	40 mins	Watch music videos
<i>Driving-work</i> [51]	2013	10	Field	EDA, HR, TEMP	Mental workload	30 mins	Drive a predefined route
<i>StudentLife</i> [31]	2014	48	Field	Smartphone	Stress, mood, happiness	10 weeks	Real life, student exams
<i>DECAF</i> [52]	2015	30	Lab	ECG, EMG, EOG, MEG, near-infrared face, video	Valence, arousal, dominance	>1 hour	Watch music video and movie clips
<i>Non-EEG</i> [53]	2016	20	Lab	ACC, EDA, HR, TEMP, SpO2	N/A	<1 hours	Four types of stress (physical, emotional, cognitive, none)
<i>ASCERTAIN</i> [54]	2016	58	Lab	ECG, EDA, EEG, facial features	Arousal, valence, engagement, liking, familiarity, personality	90 mins	Watch movie clips
<i>Stress-math</i> [55]	2017	21	Lab	ACC, EDA, HR, TEMP	Anxiety	26 hours (total)	Solve math questions under varying pressure
<i>WESAD</i> [56]	2018	15	Lab	ACC, BVP, ECG, EDA, EMG, RESP, TEMP	Affect, anxiety, stress	2 hours	Neutral, amusement and stress conditions
<i>Snake</i> [57]	2020	23	Lab	ACC, BVP, EDA, TEMP	Cognitive load, personality	>6 mins	Smartphone games with three difficulty levels
<i>CogLoad</i> [57]	2020	23	Lab	ACC, BVP, EDA, TEMP	Cognitive load, personality	N/A	6 cognition load tasks
<i>K-EmoCon</i> [58]	2020	32	Lab	Videos, audio, ACC, EDA, EEG, ECG, BVP, TEMP	Arousal, valence, stress, affect	173 mins (total)	Social interaction scenario involving two people
<i>En-Gage</i>	2021	29	Field	ACC, EDA, BVP, TEMP, In. TEMP, HUMID., CO2, NOISE	Cognitive, behavioral, emotional engagement, thermal comfort, arousal, valence	4 weeks (1416 hours in total)	Real-world courses in a high school

Table 2.1: Publicly available datasets in affective computing

self-report surveys or EMAs are often regarded as a measure of *ground truth* [29, 6, 30, 31, 32] and served as the target variables, while the features extracted from sensing data are used as predictors in ML contexts. The predictor is then mapped to the target variables through the empirical relationship determined by the data. However, Moller et al. [59] pointed out that researchers should not trust self-reports blindly but take into consideration that the responses can be unreliable. Therefore, our second aim is to investigate the reliability of self-report data. We explore the patterns in the reported confidence level and the survey completion time and then use the learning engagement as an example to compare the physiologically measured engagement and the perceived engagement. The main contributions of this chapter are summarised below:

- We propose the *En-Gage* dataset [60, 39, 61], which is the first publicly available dataset created from studying the daily behaviours and engagement of high school students us-

ing heterogeneous methods. It offers a unique opportunity to analyse the relationships between indoor climates and the mental state of school students – not only as it relates to their thermal comfort but also to their emotions, engagement and productivity at school. This dataset has the potential to greatly benefit building scientists, behavioural psychologists and affective computing researchers.

- For the first time, we investigate the reliability of self-report data by studying the confidence levels of the self-reported responses. We then compare the confidence levels of responses with the survey completion time to better understand the reliability of self-report data. Taking the student learning engagement as an example, we show that the perceived student engagement and physiologically measured engagement are not always consistent.
- We note the risk of using subjective annotations as ground truth and discuss the possibility of using physiological signals as objective measures of student engagement.

2.2 Related Work

2.2.1 Inferring Emotions and Mental State using Sensing Technology

In the ubicomp community, many studies have assessed human emotions and mental characteristics (e.g. engagement [30, 45], stress [29], mood [42, 31] and depression [47, 43]) using sensing technologies, which provide an attractive alternative to traditional self-report surveys or EMAs. King et al. [29] proposed a passive sensing framework for detecting stress in pregnant mothers in the wild, with the micro-EMA questions as a measurable ground truth for stress. Similarly, Gao et al. [30] predicted student learning engagement using physiological sensing data, with the adapted In-class Student Engagement Questionnaire (ISEQ) [62] as the ground truth for learning engagement. Wang et al. [43] tracked depression dynamics in college students using mobile and wearable sensing approaches, with the Patient Health Questionnaire (PHQ)-4 [63] and PHQ-8 [64] scores as the ground truth for depression. Zhang et al. [32] detected human compound emotions from smartphone sensing data, with self-report responses as the ground

truth for emotions. It has become common practice to regard subjective responses (e.g. EMA and self-report survey) as the ground truth, and the features extracted from sensing data are fed into the data-driven model for the prediction of emotions and mental state.

2.2.2 Reliability of Self-report Data

Many researchers have worked on designing or adapting psychological questionnaires to achieve higher validity and reliability and mitigate response bias [65, 66, 67, 68, 69]. Clark et al. [68] reviewed recent literature for psychological scale validation, and Huston et al. [69] compared the reliability of different forms of life satisfaction self-reports. Moller et al. [59] explored the reliability of self-report responses under different conditions. They conducted a six-week self-reporting study on smartphone usage. They found that self-reports cannot provide a full view of user behaviours, and participants sometimes significantly overestimated the duration of app usage. Although they demonstrated the inaccuracy of self-reports, they made suggestions for designing a self-report study (e.g. setting reminders and not pressuring participants) instead of offering solutions to evaluating the reliability of self-reports. In addition, they used survey questions related to real-world behaviour (e.g. smartphone usage), which is easier to quantify than subjective attitudes. Wash et al. [26] investigated the agreement between self-reports and behaviours. They found that security research based on self-reports is unreliable for certain behaviours, especially when the behavior involves awareness rather than actions because people are less able to answer those types of questions accurately. They revealed the unreliability of self-reports by comparing the reported data with the actual behaviours, which supported the results of Moller et al [59].

In contrast to the above-mentioned studies, this research has several advantages: (1) We investigate the reliability of self-reported data through the subjective confidence levels provided by participants. (2) We reveal the risks of using self-reported responses as the ground truth, especially for sensing emotions in the ubicomp community, by comparing the physiologically measured engagement and the perceived engagement.

2.3 Data Collection

The data collection was approved by the Science, Engineering and Health College Human Ethics Advisory Network (SEH CHEAN) of RMIT University. SEH CHEAN also reviewed and approved the consent forms for participants and guardians of minors, which included information on the purpose of and procedures for the research, the types of data to be collected, the compensation for involvement and the protocols for privacy protection and data storage. The project was also approved by the principal of the school in which the study was conducted.

2.3.1 Participants and Recruitment

We recruited participants from a K–12 private school in the suburbs of Melbourne (population 700). The recruitment occurred in August and September 2019, and calls for participation were disseminated through information leaflets, recruitment letters and a presentation in the school hall, with the assistance of the director teacher of Year 10 (Year 10 is the eleventh year of compulsory education in Australia). The admission was restricted to Year 10 students and their teachers whose native language was English or who were bilingual. A total of 23 (15–17 years old, 13 female and 10 male) out of 75 Year 10 students and six (33–62 years old, four female and two male) out of 12 teachers met the inclusion criteria, volunteered for the study and signed the consent forms. Since all the student participants were underage, their guardians also provided signed consent forms. Raw data for $n = 23$ student participants were properly recorded and nearly complete (but with different wristband wearing days), constituting the majority of the *En-Gage* dataset.

The volunteers were then asked to complete an online background survey, which was accessible through a web page link that was shared with them. In the survey, we collected information on the participants' age, gender, general thermal comfort and classes. The Year 10 students at the school were taught in separate class groups. They were separated into three *Form* groups for English, Science, Global Politics, Physical Education and Health/Sport courses, three *Maths* groups and four *Language* groups. Asking for each student's class group in the background survey allowed us to determine which classroom they were in at any given time. Among the participating teachers, there were three math teachers, one English teacher,

Group	Room	Participant
Form	R1	P13, P14, P15, P16, P17, P18, P19, P20, P21, P22
	R2	P8, P9, P10, P11, P12, P23
	R3	P1, P2, P3, P4, P5, P6, P7
Math	R1	P2, P4, P5, P10, P11, P14, P18
	R2	P3, P6, P7, P8, P9, P15, P16, P17, P20
	R3	P1, P12, P13, P19, P21, P22, P23
Language	R1	P1, P2, P4, P7, P10, P13, P15, P17, P19, P20, P21, P22, P23
	R2	P9, P14
	R3	P5, P6, P11, P12, P16
	R4	P3, P8 P18

Table 2.2: Distribution of student participants in different class groups

one Japanese teacher and one science teacher. Table 2.2 shows the details of room allocation for participants in different class groups.

As a token of appreciation for their participation, we awarded each participating student with a certificate of participation and four movie vouchers – one for each week of successful participation. Participation in this research project was voluntary, and we communicated to participants that they were free to withdraw from the project at any stage.

2.3.2 Experiment Setup

The study included four weeks of data capture: The first two weeks of data collection started from 2 September 2019, and the second two weeks of data collection started from 28 October 2019. The data capture was based on data from wearable sensors and weather stations.

In the study, we tracked participant data using *Empatica E4*¹ wristbands to measure physiological data and daily surveys to query their thermal comfort, learning engagement and emotions while at school. Overall, we collected 488 survey responses and 1415.56 hours of wearable data from all participants. During the data collection, one student representative was selected in each of the three *Form* classes. Their job was to distribute wristband sensors each morning, collect them after school and remind participants to complete the online surveys at the appropriate times. We anonymised the student data by assigning each student an identity number (ID). Occupancy schedules were obtained from the individual classroom schedules

¹Empatica E4 wristband: <https://www.empatica.com/en-int/research/e4/>

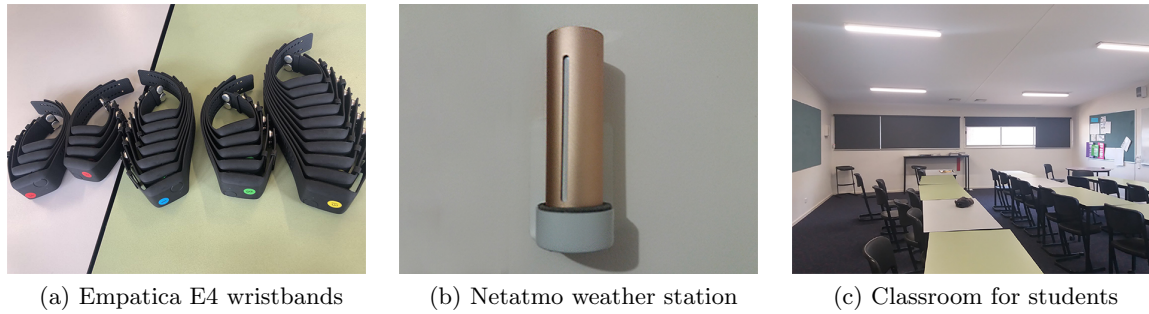


Figure 2.1: Devices and environments for collecting wearable and indoor data

provided by the school. These schedules can be used to represent the actual occupancy patterns of the building, although slight deviations from the planned schedule are to be expected in a school setting due to sickness and other circumstances.

Netatmo Healthy Home Coach. We collected indoor environmental data using Netatmo Healthy Home Coaches² installed in the classrooms. These devices measure indoor temperature, relative humidity, CO₂ levels and noise levels at five-minute intervals. The data was uploaded in real-time via the school’s guest WiFi to the Netatmo cloud platform from which we accessed the data remotely through our Netatmo account login. The ANSI/ASHRAE Standard 55 recommends temperature sensor heights of 0.1, 0.6 and 1.1 m for ankles, waists and heads of seated occupants, respectively. Given these guidelines, since only one device was installed per room in this study and the head height of students is lower than that of adults, we attempted to place the sensors at approximately 0.9 m.

Empatica E4 wristband. These wristband sensors (see Figure 2.1a) were first proposed for use in Garbarino et al. [70]. These watch-like devices have multiple sensors: an EDA sensor, a PPG sensor, a three-axis accelerometer (ACC) and an optical thermometer. EDA refers to constantly fluctuating changes in the electrical properties of the skin at 4 Hz; when the level of sweat increases, the conductivity of the skin increases. PPG sensors measure the blood volume pulse (BVP) at 64 Hz, from which the interbeat interval (IBI) and HRV can be derived. The ACC records in the range of (-2g, 2g) at 32Hz and captures motion-based activity, which has been widely used in smartphones, wearables and other IoT devices [12]. The

²Netatmo Healthy Home Coach: <https://www.netatmo.com/en-eu/aircare/homecoach>

Annotation	Description	Measurement scale
<i>Thermal sensation</i>	ASHRAE thermal sensation [71]	-3: cold, -2: cool, -1: slightly cool, 0: neutral, 1: slightly warm, 2: warm, 3: hot
<i>Thermal preference</i>	ASHRAE thermal preference [71]	Choose one (cooler, no change, warmer)
<i>Clothing level</i>	ASHRAE clothing insulation [71]	Choose multiple
<i>Seating position</i>	Seating position in the classroom	Click one point
<i>Multi-dimensional/engagement</i>	Adapted In-class Student Engagement Questionnaires [62]	-2: strongly disagree, -1: somewhat disagree, 0: neither agree nor disagree, 1: somewhat agree, 2: strongly agree
<i>Arousal/Valence</i>	Affective dimensions from the Photographic Affect Meter [72]	Choose one photo
<i>Confidence level</i>	Confidence level of the response	1: not confident, 2: slightly confident, 3: moderately confident, 4: very confident, 5: extremely confident

Table 2.3: Collected annotations from the questionnaires

optical thermometer reads peripheral skin temperature (ST) at 4 Hz. In recording mode, E4 wristbands can store 60 hours of data in memory, with a battery life of over 32 hours. They are lightweight, comfortable and waterproof and were thus especially suitable for the continuous and unobtrusive monitoring of the participants in our study. Before the data collection, all wristbands were synchronised with the E4 Manager App, using a single laptop to ensure that the internal clocks were accurate. Each student was assigned a wristband sensor marked with their unique study ID. The students were asked to wear the wristband on their non-dominant hand and to avoid pressing the button or performing any unnecessary movements during class. The teacher participants were only required to wear the wristbands while teaching the year 10 classes.

Daily surveys. On each school day, student participants were asked to complete online surveys (either through tablets placed in each classroom or using their own digital devices) at 11:00, 13:25 and 15:35 (directly after the second, fourth and fifth class). The length of the second and fourth class was either 40 min or 80 min, depending on the day of the week, and the fifth class always lasted 80 min. The curriculum in this school had a bi-weekly rhythm, i.e. the first and second weeks had different class schedules, but the first and third weeks were identical, as were the second and fourth weeks. The student representative was tasked with reminding the student participants to complete the online surveys on time, as described in Table 2.3. The online questionnaire included 11 items related to the students' psychological state and

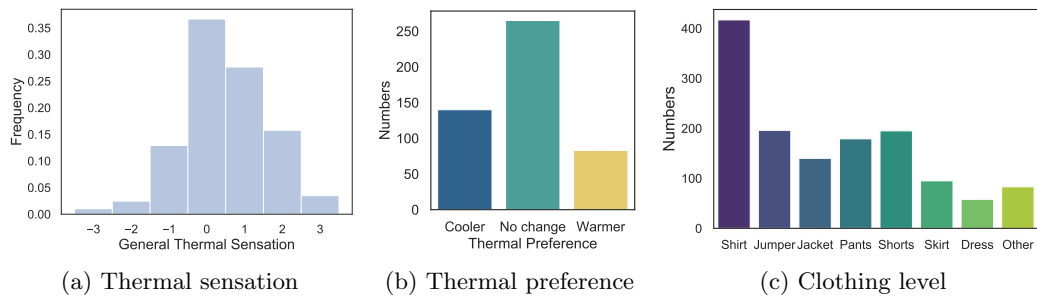


Figure 2.2: Distribution of responses related to thermal comfort

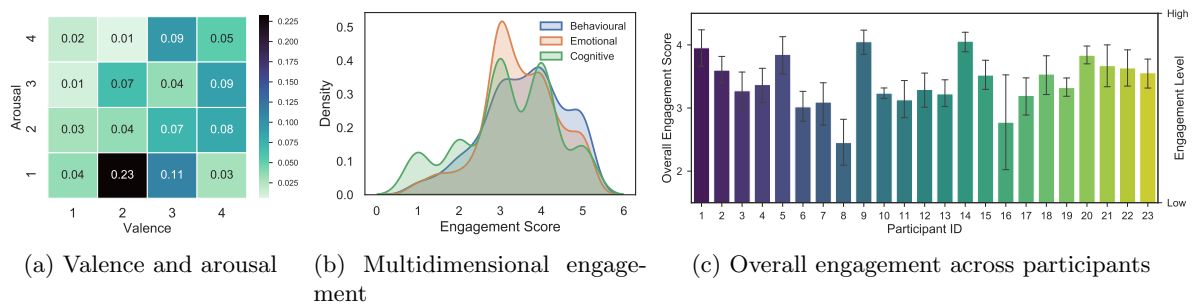


Figure 2.3: Distribution of responses related to engagement and emotions

behaviours (e.g. thermal comfort, student engagement and emotions). All the items (except the seating position and confidence level) were used directly or slightly adapted from validated questionnaires widely used by researchers in this area. Figure 2.2 shows the distribution of responses to the thermal sensation (from -3 to 3), thermal preference and clothing level. The distribution of multidimensional engagement (behavioural, emotional and cognitive) is shown in Figure 2.3b, and the overall engagement across participants is shown in Figure 2.3c. Figure 2.3a shows the distribution of emotions in the valence and arousal dimensions. The numbers indicate the percentage frequencies, and the darker the colour, the higher the frequency of the specific emotion (e.g. arousal = 1 and valence = 2).

Figure 2.4a shows the distribution of survey responses throughout the day. As students were requested to submit their self-reports directly after the second, fourth and fifth class (i.e. 11:00, 13:25 and 15:35, respectively), most responses were submitted 11:00–12:00, 13:00–14:00 and 15:00–16:00. The survey responses that were recorded before the start of the targeted class

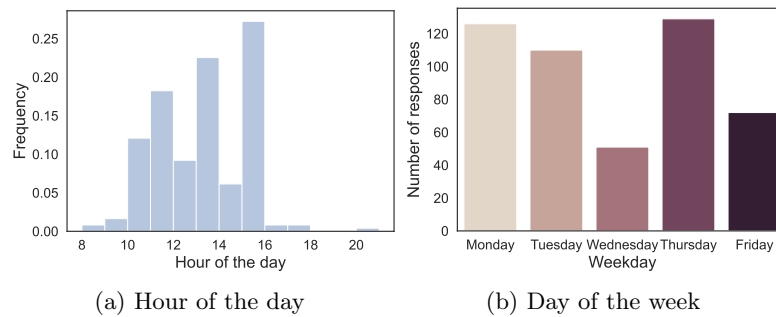


Figure 2.4: Distribution of the survey responses across hours of the day and days of the week

or after the start of the next class should be removed if we aim to explore student engagement in the targeted class. However, if we wish to study thermal comfort, clothing insulation, emotion or confidence level, all the survey responses can be kept. Next, we explored the distribution of the survey responses through the week (see Figure 2.4b). We found that most students submitted their responses on Monday or Thursday. The number of students who submitted self-report data on Wednesday was the lowest. The potential reason for this may be that students forgot to submit their responses, as they took breadth studies on a Wednesday (normal studies on the other weekdays).

2.4 Reliability of Self-Report Data

2.4.1 Confidence Level of Responses

During the data collection process, we collected the participants' confidence levels in the self-reports. Figure 2.5 shows the distribution of the confidence levels of the participants. We can see that most participants have a moderate degree of confidence in their responses, but a small number of participants (whose confidence level is 1 or 2) are not very confident in their responses.

Then, we explored the difference in confidence levels between participants and whether the confidence level of the same participant changes over time. Figure 2.6 shows the box plot of confidence level for each participant. We discovered that different participants tended to have very different confidence levels. For example, some participants (e.g. P1 and P20) were usually

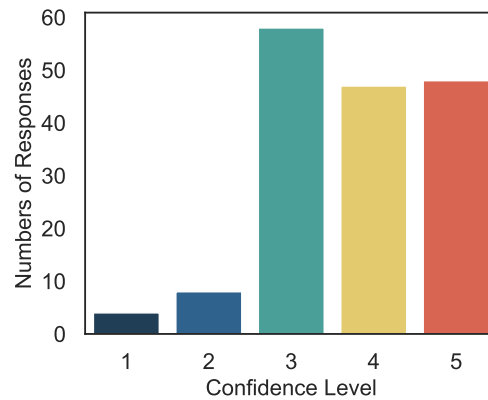


Figure 2.5: Distribution of confidence levels for all self-report responses

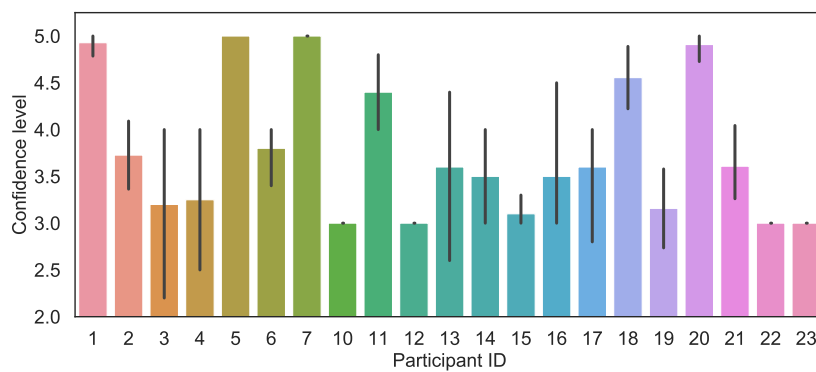


Figure 2.6: Confidence levels across all participants

strongly confident in their self-report responses, while some participants (e.g. P10, P12 and P15) were generally not very confident in their responses. In addition, some participants (e.g. P1, P20 and P15) tended to have similar confidence levels in longitudinal studies, but some participants (e.g. P16 and P3) had very different confidence levels at different times during the data collection process. The above phenomena are in line with our daily experience.

2.4.2 Completion Time and Reliability

Malhotra et al. [73] found that the survey completion time is an indicator of response quality, although it is affected by multiple factors and varies from person to person. For each self-report questionnaire, the completion time was automatically recorded and collected by the *Qualtrics*

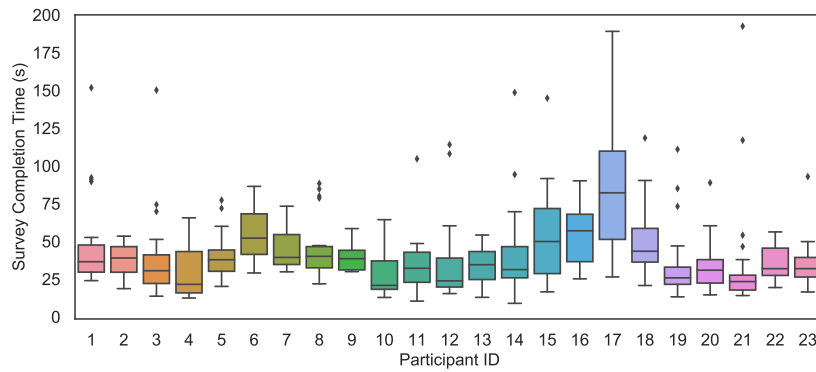


Figure 2.7: Survey completion time of participants

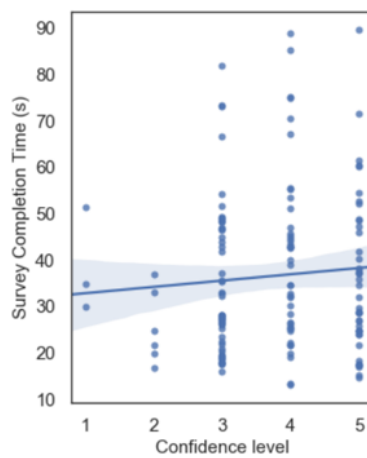


Figure 2.8: Linear regression of survey completion time with confidence levels

timing question, which is a hidden question added to the questionnaire to track the time spent by the respondent on that page.

Figure 2.7 shows the survey completion time for all participants. We can see that different participants had very different survey completion times. Most participants completed the survey in 30–50 s; however, some participants (e.g. P17) spent a lot more time completing the survey, and some participants (e.g. P10 and P12) completed the survey in a very short time.

We then studied whether the survey completion time was correlated with the confidence levels. Figure 2.8 shows that the survey completion time was positively correlated with the confidence level. Participants with longer survey completion times tended to have a higher confidence level in the survey. We also investigated how the confidence levels were correlated

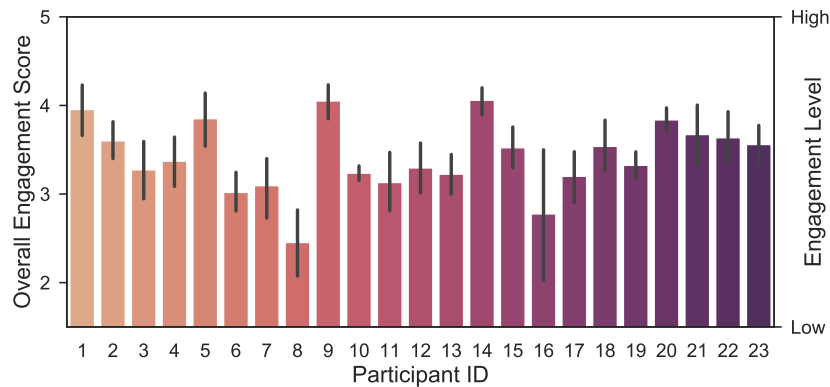


Figure 2.9: The distribution of overall engagement across student participants

with other factors, such as the time of the day and the day of the week, but we did not find any significant correlations. In future research, it would be interesting to use survey completion time as an indicator of survey reliability and assign appropriate weights to self-reported responses for more accurate mental state predictions.

2.4.3 Perceived vs Physiological Measurements

To calculate the perceived engagement scores, we reversed the responses in item 2 and item 4 and calculated an average score based on the 5-point Likert scale for each dimension of engagement (please refer to Table 3.2 for details on questions to measure engagement). We then calculated the overall engagement scores based on all five items, where 1 indicated the lowest level of engagement and 5 the highest. Figure 2.9 shows the distribution of overall perceived engagement across student participants. We can see that different participants tended to have very different levels of perceived engagement. Some participants (e.g. P1, P9 and P14) were usually highly engaged in class while some participants (e.g. P8) had low engagement levels. Gao et al. [30] built an engagement prediction model, with perceived engagement being regarded as the ground truth.

Physiological signals (e.g. EDA, HRV and ST) have been explored in previous studies for predicting student engagement levels [30, 6]. For example, the EDA level is usually considered a good indicator of physiological and psychological arousal (e.g. student engagement [30] and emotional state [6]). Increased heart rate indicates increased effort and is used as an indirect

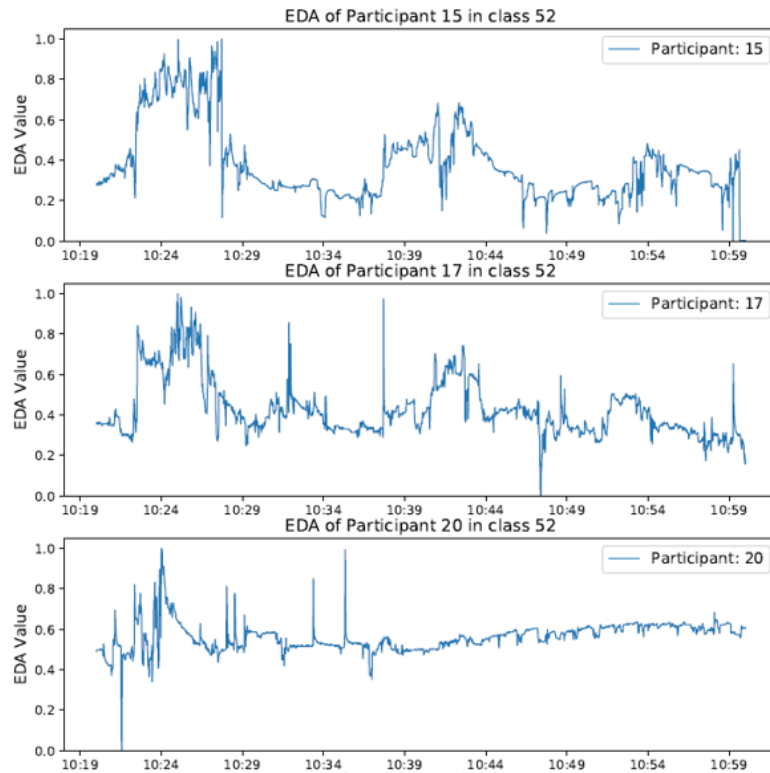


Figure 2.10: An example of the changes in electrodermal activity of three different participants, P15, P17 and P20, in the same class (their perceived engagement levels were 4.2, 3.2 and 4.4, respectively)

measure of engagement [74]. It has been shown that changes in heart rate are related to the intensity of mental efforts and demands of information processing. In addition, changes in skin temperature have been shown to be correlated with social context and mood [75].

Figure 2.10 shows an example of changes in EDA for different participants in the same class. It can be seen that the EDA signals of the first two participants are very similar, and there is strong physiological synchrony [76] between them. Physiological synchrony refers to the association or interdependence of physiological activity between two or more individuals, which has been observed in many scenarios. Physiological synchrony between individuals can be indicative of group engagement [76] and has been used to measure the emotional climate of a classroom [7] and quantify participants' agreement on self-reporting engagements [28].

In Figure 2.10, strong physiological synchrony between P15 and P17 indicates that they had similar engagement patterns. In addition, it is likely that they were both highly engaged:

(1) Their EDA signals have multiple peaks at a similar time, which is a good indicator of emotional arousal. (2) If they were not engaged in class, the changes in their EDA should be more random, instead of being similar. It is likely that participant P20 had lower levels of engagement than participants P15 and P17 since the EDA signal was more random with fewer peaks than the other two participants. However, based on the self-report responses, the engagement score of participants P15, P17 and P20 were 4.2, 3.2 and 4.4, respectively. From this example, it is possible to deduce the following: (1) Participants with very similar physiological patterns may have very different perceived engagement annotations (e.g. P15 and P17). (2) Participants with very similar annotations may have very different physiological patterns (e.g. P15 and P20).

2.4.4 Discussion

Self-reporting is one of the most common ways to study psychological state and attitude in human-based studies. In affective computing, self-report annotations are usually believed to be the ground truth for predicting human mental state using sensing technologies. In recent years, data-driven models have been built using self-reported data as the target variable. However, self-reported data is prone to subjectivity and response bias, making it risky and inaccurate for use as the ground truth in predicting the psychological state (e.g. emotion, depression and engagement) from sensing data.

Using experiments, we investigate the reliability of self-reported data in the wild from two perspectives: (1) For the first time, we study the confidence level of the self-report responses of participants and compare the confidence level with the survey completion time to better understand the reliability of self-reported data. (2) To the best of our knowledge, we are the first researchers to compare the perceived and physiological measures of student engagement. From our experiment, we found that the perceived self-report engagement was not always consistent with the physiologically measured engagement. Participants with similar physiological patterns may report very different perceived engagement and participants with similar self-report annotations may have very different physiological patterns. By contrasting the self-report and physiological measures, we reveal the potential risks of only using subjective annotations as

the ground truth.

2.5 Conclusion

In this chapter, we addressed the challenges of the limited sources of sensing data and the difficulties in validating the ground truth in human behaviour sensing studies. The released dataset *En-Gage* is the first publicly available dataset for studying the daily behaviours and engagement of high school students using heterogeneous sensing. With this dataset, various data mining (e.g. segmentation [77], clustering [78, 79]) and modelling techniques [80, 81, 82]) could be explored to build prediction models for measuring people’s mental state using sensor-based physiological and behavioural recordings in buildings. This could be further used for various applications in future studies: (1) *Monitoring signs of disengagement and negative emotions of students* [30, 15]: Measuring the study engagement and emotions of students is beneficial to both teachers and students. Teachers will be able to improve their teaching strategies to create the right learning environment, improve the learning experience for students and re-engage students with low engagement [7, 30]. Students will be able to self-track their learning engagement and emotions, which could promote their self-regulation and reflective learning. (2) *Studying peer effects in educational settings* [83]: It could be helpful to explore group-wise seating behaviours [84, 85] and their relationship to perceived engagement and physiological synchrony [40]. (3) *Providing comfortable indoor environments for people*: It is possible to mitigate the negative effects of hot weather on student learning by using air conditioning [86], and teachers could ventilate their classrooms timeously to prevent excess carbon dioxide from affecting students’ concentration [87, 88, 89].

In addition, the analysis of the reliability of self-reported data is a very promising step towards validating ground truth in human behavioural studies. It serves as a wake-up call for the emotional and mental sensing research that usually regards self-report annotations as the ground truth for predicting the human mental state. This raises questions, such as, why would students feel more engaged in class if their bodies indicate otherwise? Should we trust their subjective self-report responses more than their objective physiological responses? Is there a better way to understand and model the human mental state than only using self-report

annotations as the ground truth? We hope that more research will explore this issue in the future.

Chapter 3

Modelling User Engagement Behaviours from Wearable and Environmental Sensors

Using the dataset introduced in Chapter 2, this chapter aims to explore wearable and environmental sensing data to understand student engagement in the wild. The study of student engagement has attracted growing interest to address problems such as low academic performance, disaffection and high dropout rates. In relation to *RQ-2*, this chapter investigates whether we can infer and predict engagement on multiple dimensions only using wearable and environmental sensors. We hypothesise that multidimensional student engagement can be translated into physiological responses and activity changes during class and is affected by environmental changes. We present *n-Gage*, a student engagement sensing system using a combination of wearable and environmental sensors to automatically detect students' in-class multidimensional learning engagement. Extensive experimental results have shown that *n-Gage* can accurately predict multidimensional student engagement in real-world scenarios, with an average mean absolute error (MAE) of 0.788 and root mean square error (RMSE) of 0.975 using all the sensors. We will present a set of interesting findings on how different factors (e.g. combinations of sensors, school subjects and CO_2 levels) affect each dimension of student

learning engagement.

3.1 Introduction

In education, *student engagement* refers to the degree of attention, interest, curiosity, and involvement in the learning environment [90]. The study of student engagement has attracted growing interests as a way to address the problems of low academic achievement, high levels of student boredom, disaffection, and high dropout rates in urban areas [91, 92]. Previous research showed that student engagement declines as students progress from elementary to middle school, reaching its lowest levels in high school [93, 94]. Marks et al. [94] estimated that as many as 40-60% of high school students are disengaged (e.g., uninvolved, no interests and not attentive). The consequences of disengagement for high school students are severe. They are less likely to graduate from high school and face limited employment prospects, increasing risks for poverty, poorer health, and involvement in the criminal justice system [95]. In view of the negative impact of disengagement, more and more researchers, educators and policymakers are interested in obtaining data on student engagement and disengagement for needs assessment, diagnosis, and preventive measures [93].

Generally, student engagement is defined as a meta-construct that includes three dimensions [91, 92]: (1) *behavioural engagement* focuses on participation and involvement in academic, social, and co-curricular activities. Some researchers define behavioural engagement with regards to positive conduct, e.g., following the rules, obeying the classroom norms, and the absence of disruptive behavior such as skipping school [92, 96, 97]; (2) *emotional engagement* focuses on the extent of positive and negative reactions to teachers, classmates, academics, and school, which includes a sense of belonging or connectedness to the school [92, 98]; (3) *cognitive engagement* focuses on students' level of investment in learning, which includes being thoughtful, strategic, and willing to put efforts to comprehend complex ideas or master difficult skills [91, 92, 99]. One of the most common method for measuring student engagement is self-report surveys (e.g., Learning Questionnaire (MSLQ) [100], School Engagement Measure (SEM)-MacArthur [101], and Engagement vs. Disaffection with Learning (EvsD) [102]). Though generally reliable, surveys are usually time-consuming and can be a burden for

participants if they need to complete the survey for each class.

Therefore, we want to investigate whether we can infer and predict student multidimensional engagement just using sensors. In particular, we conduct the research around the hypothesis that student multidimensional engagement level can be translated into physiological responses and activity movements during the class, and also be affected by the environmental changes. In previous studies, multiple physiological data (e.g., electrodermal activity (EDA), heart rate variability (HRV), accelerometer (ACC) data, skin temperature (ST)) and environmental data have been explored to assess the emotion arousal and engagement in different scenarios. For instance, EDA is usually considered as a good indication of psychological or physiological arousal (e.g., emotional and cognitive states) [103, 104], which has been increasingly explored in affective computing such as the detection of emotion [47, 105], depression [48], and engagement [6, 106, 107]. Recently, Pflanzner et al. [108] stated that EDA monitoring should be combined with the recording of heart rate and blood pressure because they are all autonomically dependent variables. Heart rate data has been used to predict student engagement in a structured writing activity [109] and the correlation of heart rate and student cognitive and emotional engagement has been found in [62]. As the most commonly used sensors in IoT devices, the accelerometer is proven to be a powerful tool for the quantification of human behavioural patterns [31, 12]. Von et al. [110] used accelerometer sensors to demonstrate how large groups of people moving in sync can enhance group affiliation. Accelerometer-based features have been analyzed to sense children's engagement during a performance using interpersonal movement synchrony [111].

In this chapter, our aim to explore the following questions: 1. *Can we measure the multiple dimensions of student's learning engagement including emotional, behavioural and cognitive engagement in high schools with sensing data in the wild?* 2. *Can we derive the activity, physiological, and environmental factors contributing to the different dimensions of student learning engagement? If yes, which sensors are the most useful in differentiating each dimension of the learning engagement?* To answer the above questions and enable automated engagement detection, we present a new classroom sensing system *n-Gage* to assess the behavioural, emotional and cognitive engagement level of students. The system utilizes sensing data from two

sources: (1) wearable devices capturing physiological and physical signals (e.g., EDA, HRV, ACC); (2) indoor weather stations capturing environmental changes (e.g., temperature, CO₂, sound). The study has been approved by the Human Research Ethics Committee at RMIT University and the high school where it is conducted, and the procedures follow the ethical codes. In summary, the contributions of this work is as below:

- We build *n-Gage*, a classroom sensing system to automatically measure the multidimensional engagement (behavioural, emotional and cognitive engagement) of high school students during the classes. In particular, we combine physiological signals, physical activities, and indoor environmental data together to estimate the changes of student engagement levels. To the best of our knowledge, this is the first system to detect student engagement from multiple sensors in the wild.
- We extract new features to represent the physiological and physical synchrony between students which proved to be useful for the student engagement prediction. We also for the first time extract features from skin temperature and indoor environment for effective engagement estimation.
- We conduct comprehensive experiments to predict the multidimensional student engagement scores with *LightGBM* regressors. Experiment results show that *n-Gage* achieves high accuracy for student engagement. We also derive the different factors and explore the most useful sensors in differentiating each dimension of the learning engagement.
- We show a set of interesting insights into how different factors affect student engagement. For example, CO₂ level in the classroom has negative effects on students' cognitive engagement, which highlights the need to timely ventilate the classroom for improving student engagement.

3.2 Related Work

3.2.1 Traditional Methods for Measuring Engagement

In the education area, there are various methods as follows to study student engagement. (1) *Student self-report* is the most common method to assess student engagement as it is easy to execute in classroom settings. Students are provided with items reflecting different dimensions of engagement and then select the response that best describes them [92]. However, the self-report survey is labour and time-consuming and students may not willing to answer too many questions honestly at a time, leading to low-quality responses [112]. (2) *Experience sampling* [113, 114] allows researchers to collect engagement data at the moment, which reduces the problems of recall-failure and social-desirability bias happened in the self-report surveys. However, it requires a huge time investment from students and the quality of responses largely relies on the students' willingness and ability to answer [92]. (3) *Teacher ratings of students* [92] can be useful for young students with difficulty in completing self-report surveys. Behavior can be observed directly from teachers but emotion engagement is difficult to be observed as students may learn to mask their emotions [92, 115]. (4) *Interviews* can provide a detailed description of the student's performance during the learning process. However, the quality of responses depends on the expert knowledge from the interviewers. (5) *Observation* [92] on the individual students or whole students in the classroom have been developed to assess engagement, which can be time-consuming for the administer and all kinds of academic settings need to be considered to get an accurate picture of student behaviour. The reliability of the observations can be doubtful as they only provide limited information about students.

All the traditional methods for engagement measurement have strengths and limitations in different situations. Overall speaking, traditional methods are usually time-consuming and the quality of answers largely depends on the students, teachers or executor. Recently, with the development of wearables and IoT sensors, some initial progress has been explored to measure student engagement with their physiological data which is more subjective and obtrusive to students.

Table 3.1: Related work for engagement prediction with sensing data

Prediction	Data source	Participants	Data Sessions
Audience Engagement [111]	ACC data	10 children audience in art performance	not stated
Social Engagement [106]	EDA data (wristband)	Children during social interactions	51 sessions
Game Engagement [45]	EDA, PPG data	10 players in 6 mobile games in natural settings	not stated
Audience Engagement [28]	EDA, PPG data	10 attendees and 19 presenters in presentations	40 sessions
Student Engagement [116]	Video, audio data	25 university students in 5 classrooms	not stated
Student Engagement [109]	Video, heart rate data	23 university students in laboratory settings	not stated
Student Engagement [117]	EDA data (hand sensor)	17 undergraduate students in climate science classes	not stated
Student Engagement [118]	EDA data (hand sensor)	17 university students in learning environments	not stated
Emotional Engagement [6]	EDA data (wristband)	27 university students in 41 lectures over 3 weeks	197 sessions
Student Multidimensional Engagement (this work)	EDA, PPG, ST, ACC, CO2, Noise, etc.	23 high school students in 98 classes over 4 weeks	331 sessions

3.2.2 Engagement Prediction with Sensing Technology

Sensing technologies are becoming prevalent to assess people’s mental characteristics (e.g., engagement [6, 106, 111, 119, 45], mood [42, 31], stress [47, 31], personality [12]) and have provided an attractive alternative to traditional self-report surveys. Wang et al. [31] gathered students’ mental health data such as mood and stress from self-report surveys in Dartmouth college. They also recorded students’ activity data from passive sensors and found a significant correlation between the sensor data and mental health. Morshed et al. [42] predicted mood instability only using sensed data from mobile phones and wearable sensors for individuals in situated communities. Wang et al. [120] predicted human personality traits from passive sensing data from mobile phones using within-person variability features such as regularity index of physical activity, the circadian rhythm of location.

Physiological sensors and accelerometers have been explored to assess human’s engagement (see Table 3.1), such as assessing audience engagement during the art performance, social engagement for children during the interaction with adults [106], emotional engagement for

university students during lectures [6].

Ahuja et al. [116] built a classroom sensing system using a distributed array of commodity cameras. Students' and instructor's video and audio were captured for body segmentation and speech detection. Then, the students' engagement levels were analyzed from featured data such as hand raising, smile, sit/stand classification. However, as reported from authors, this system would bring privacy concerns when capturing audio and video data. Similarly, Hutt et al. [119] used the commercial off-the-shelf eye-trackers to automatically detect mind wandering for high school students and Monkaresi et al. [109] used heart rate and video-based estimation of facial expressions to predict the engagement of 23 university students during a structured writing activity in laboratory settings. Besides the privacy concern, the questionnaire is very simple and only asked participants to report whether they are engaged or not.

Only a few studies investigate student engagement in the real-world settings [117, 118, 62, 6]. Mcneal et al. [117] used EDA hand-sensor to measure the engagement from 17 undergraduate students in classrooms emphasizing climate science during a semester. They explored different teaching approaches on a subset of students and reported the statistical results for the mean of the EDA traces. Contrast to their studies, we collected a far more heterogeneous data set and novel features were proposed based on different physiological indices. Wang et al. [118] studied 17 university students' engagement in the distributed learning environment with the EDA hand sensor and they found that EDA sensor measurements were aligned with surveys. Different to us, they only used a very simple question '*how much did you enjoy during the lecture as the ground truth of students' engagement*'.

In recent years, researchers have started to explore the different dimensions of engagement using physiological signals. Lascio et al. [6] predicted university students' emotional engagement from EDA sensors in lectures during 3 continuous week data collection. While in our data collection, we build an in-class multidimensional (behavioural, emotional, cognitive) engagement sensing system including physiological responses (i.e., EDA, HRV, ST), physical movements (ACC) and indoor environmental sensors (i.e., CO₂, temperature, humidity, sound stream) for high school students. Furthermore, the high school classes are very different from lectures at university in [6] (e.g., degree of freedom to choose courses, ability to schedule classes

flexibly, requirements of class attendance, consistency of subjects between different schools), which may lead to very different multi-engagement distribution in high school classes. Another similar study was proposed by Huynh et al. [45] who measure the engagement level of game players using multiple sensors. Though they agreed that user engagement includes behavioural, emotional and cognitive dimensions, they did not differentiate each dimension when predicting the engagement during the game. Nevertheless, in our study, we derive the different factors and most useful sensors contributing to the different dimension of student learning engagement.

In summary, different from the previous efforts, our work has following advantages: (1) we use far more heterogeneous data for engagement prediction (other works only use EDA or heart rate data except [45]); (2) we propose and extract more meaningful features from physiological signals while [118, 117, 62] only use the simple average value of data); (3) to the best of knowledge, we are the first to predict the engagement for all three dimensions based on education research while previous research has either measured the simple general engagement or a single dimension of engagement [6]), and derive the most useful sensors in differentiating each dimension of engagement; (4) we adopt the real-world classroom settings and take the influence of environmental changes into account.

3.3 Dataset

We conducted a field study in a private high school for 4 weeks in 2019. The data collection has been approved by the Human Research Ethics Committee at our University. The details for the data collection have been introduced in Chapter 2. Here, we will briefly describe the participants, procedures and collected data.

3.3.1 Participants and Procedures

In total, we have recruited 23 students (13 females and 10 males, 15-17 years old) and 6 teachers (4 females and 2 males, 33-62 years old) in Year 10. Before the data collection, all wristbands were synchronized with the E4 Manager App. 1 Netatmo weather station was installed and 1 tablet was put on the teacher desk in each classroom. Students were asked not to unplug the

Netatmo stations during the data collection.

The first two weeks of data collection occurred in early September (winter in the southern hemisphere), and the next two weeks of data collection completed in November (spring in the southern hemisphere). We collected data from two different seasons to build a more robust engagement sensing system. As we know, different seasons usually result in different indoor environments (e.g., indoor temperature, humidity), which may affect students' sweat level (EDA, ST) and activity level (ACC, HRV). If we use the data from one season to build the engagement prediction model, the prediction performance can be greatly reduced in another season due to changes in activity, physiological, and environmental data.

During the data collection, student participants were distributed with the wristband at 8:50 before the first class started at 9:00. Then at the end of the school day (i.e., 15:35), student participants were reminded to hand in wristbands. On each school day, student participants were asked to complete the online surveys (either through the public tablets or their own digital devices) at 11:00, 13:25, 15:35 (right after the 2nd, 4th, 5th class). The length of 2nd and 4th class can either be 40 minutes or 80 minutes on the different school day and the 5th class always lasts for 80 minutes. From the class table for Year 10 students, they have the same class schedule on the 1st week and 3rd week, and another class schedule on the 2nd and 4th week. However, considering that it could be a burden for some participants to complete the survey three times a day, we did not require students to complete the survey, which helps us ensure the quality of survey responses. By the end of the 4th week, we had received 488 valid responses in total and the response rate is 35.3%.

3.3.2 Collected Data

3.3.2.1 Physiological and Activity Data

During the school time, we asked participants to wear *Empatica E4*¹ wristbands as shown in Figure 2.1a, first proposed in [70]. E4 wristband is a watch-like device with multiple sensors: electrodermal activity (EDA) sensor, photoplethysmography (PPG) sensor, 3-axis accelerometer (ACC), and optical thermometer. EDA depicts constantly fluctuating changes in skin

¹Empatica E4 wristband: <https://www.empatica.com/en-int/research/e4/>

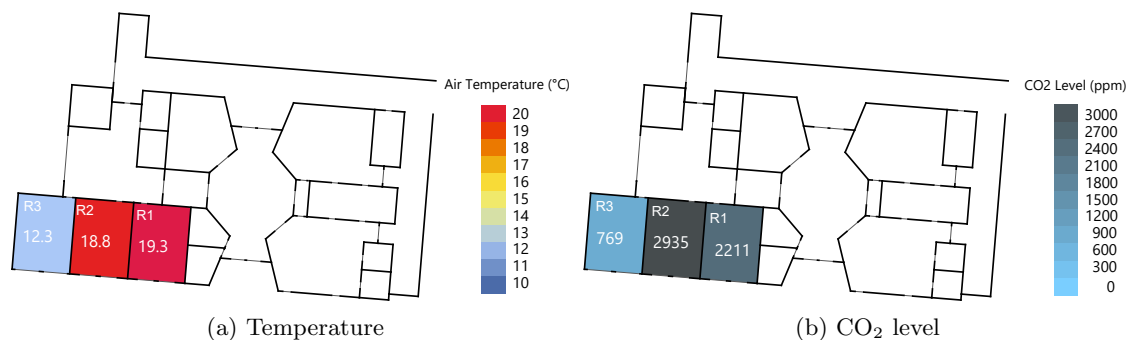


Figure 3.1: Temperature and CO₂ data in R1, R2, R3 (room 1, room 2, room 3) at 11:00 am on 11 Sep 2019. Room 4 is not shown here as it is in another building

electrical properties at 4 Hz. When the level of sweat increases, the conductivity of skin increases. PPG sensor measures the blood volume pulse (BVP) at 64 Hz, from which the inter-beat interval (IBI) and heart rate variability (HRV) can be derived. ACC records 3-axis acceleration in the range of $[-2g, 2g]$ at 32Hz and captures motion-based activity. The optical thermometer reads peripheral skin temperature (ST) at 4 Hz.

3.3.2.2 Indoor Environmental Data

We collected indoor environmental data from the Netatmo Healthy Home Coach² - a smart indoor weather station - installed in the classrooms as shown in Figure 2.1b and Figure 2.1c. The Netatmo station can collect indoor temperature (TEMP), humidity (HUMID), CO₂ and sound (SOUND) in every 5 minutes. Real-time data can be uploaded to the Cloud continuously through the Guest WiFi covered on the campus. Figure 3.1 shows the indoor temperature and CO₂ level in three rooms at 11:00 am on 11 Sep 2019. We can clearly see that the temperature of room 3 is only 12.3 °C and much lower than the comfortable warmth (18 °C) defined by the World Health Organization's standard [121], which may negatively affect student learning in class [122]. Furthermore, CO₂ levels in room 2 and room 3 are beyond 2000 ppm, which has been proved to have a negative influence on the student cognitive load in the classroom [123, 124]. Based on previous studies [125], students may become sleepy and inattentive during the class when the CO₂ level is too high.

²Netatmo Healthy Home Coach: <https://www.netatmo.com/en-eu/aircare/homecoach>

Table 3.2: Self-report items for measuring in-class engagement in online survey

Questions (please describe your engagement in the last class)	Subscales
1. I paid attention in class.	Behavioural
2. I pretended to participate in class but actually not.	Behavioural (-)
3. I enjoyed learning new things in class.	Emotional
4. I felt discouraged when we worked on something.	Emotional (-)
5. I asked myself questions to make sure I understood the class content.	Cognitive

Note: (-) means the reversed score.

3.3.2.3 Ground Truth: Self-report Survey Instrument Data

In this study, we choose to use self-report survey to gather subjective measurements of students' in-class engagement. As discussed in Section 3.2, the self-report survey is the most common way to measure student engagement as they can reflect students' subjective perceptions explicitly. Instead, measures relying on experience sampling, teacher ratings, interviews or observations have been reported to be easily affected by the external factors. The questionnaire includes 5 items related to behavioural, emotional, and cognitive engagement of the validated *In-class Student Engagement Questionnaires* (ISEQ) [62], which has been proved to be effective for measuring multidimensional engagement compared to the traditional long survey. Similar to [6, 45], we slightly adapted survey questions from university lectures to high school class context to make the survey easier for students underage to understand. Moreover, for cognitive engagement measurement, we did not use the original question '*the activities really helped my learning of this topic*' in [62], considering that some classes in high school do not have in-class activities. Instead, we use the well-accepted item '*I asked myself questions to make sure I understood the class content*' [101], which is a good reflection of cognitive engagement. Table 3.2 shows the questionnaire used for measuring multidimensional student engagement in class, where item 1,3 and 5 assess the behavioural, emotional and cognitive engagement, item 2 and 4 indicate the behavioural and emotional disaffection [102, 62].

In the questionnaire, each item ³ is rated with a 5-point Likert-scale from -2 to 2, which indicates 'strongly disagree', 'somewhat disagree', 'neither agree nor disagree', 'somewhat agree'

³In the survey, participants were also asked to report their thermal feelings and mood using the Photographic Affect Meter (PAM) [72]. Nevertheless, this data was not considered in this research.

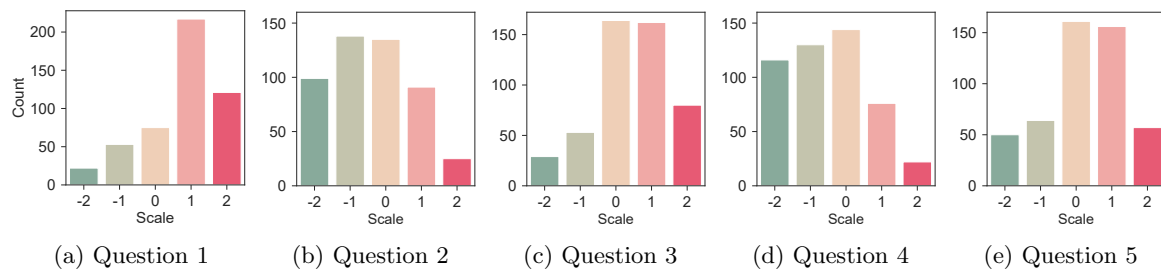


Figure 3.2: Histograms of the Answers. The X axis shows the 5-Likert scale from -2 to 2 which means ‘strongly disagree’ to ‘strongly agree’. The Y axis shows the number of the responses that fall into the specific scale

and ‘strongly agree’. Figure 3.2 shows the distribution of responses for each item from total 488 responses. The online self-report survey is constructed with the external tool named *Qualtrics*⁴. Participants were asked to complete the survey on the public tablets or their digital products with the given survey link generated by *Qualtrics*.

3.4 Data Preprocessing

In this section, we extract class periods based on students’ accelerometer data with the unsupervised time series segmentation method. Then we introduce the data cleaning process and data pre-processing technique for electrodermal activity, blood volume pulse, accelerometer data, and skin temperature data.

For data preparation, we only keep the data between 9:00 am to 15:35 pm, which corresponds to the start time of the first class and the end time of the last class. In addition, some students may have several data recording segments during the same day due to the unexpected closure and re-open of the wristband. We drop the data segments that are less than 15 seconds in length, which is less helpful for extracting useful information. We also discard the data on Tuesday in the last week because students had trip travel and did not have classes on that day.

⁴Qualtrics: <https://www.qualtrics.com/au/>

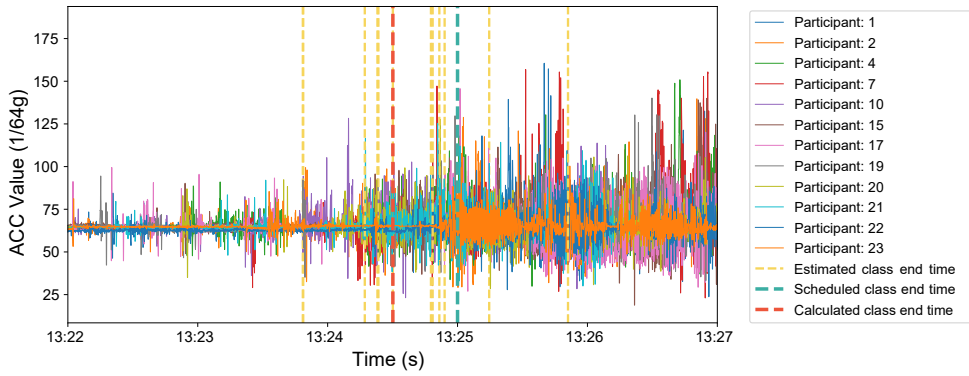


Figure 3.3: Calculated class end time with ACC data from 12 student participants

3.4.1 Class Period Segmentation

As described in Section 3.3, student participants wear wristbands all day along and teachers participants are only asked to wear the wristband at their classes. Participants report their engagement for the 2nd, 4th, 5th classes of the day during recess time, lunchtime and before going home. Though the scheduled class start/end time is already known, teachers may start/finish the class a bit earlier or later than the scheduled time. The accurate class time is significant for wearable data analysis because participants may have very different physiological/movement patterns between in-class and after-class. For instance, increased activity level after class may lead to a higher value of EDA (due to the higher level of sweat) and variation of accelerometer data.

To get the exact class start/end time for meaningful data analysis, we segment the accelerometer data from student participants based on the assumption that students usually have different activity patterns before/after class. *Information-Gain based Temporal Segmentation* (IGTS) [126, 127] is applied on the ACC data to calculate the class start/end time. IGTS is an unsupervised segmentation technique, aims to find the transition times in human activities, which is suitable for dividing the boundary between in-class and out-class [126]. Topdown optimization is adopted in the ACC time-series segmentation. To calculate the class boundary, we choose the ACC time-series from 5 minutes before the class to 5 minutes after the class. Take calculating the actual class end time as an example, from Figure 3.3, there are 12 participants in a class and the scheduled class end time is 13:25 (green vertical dashed

line). Applying IGTS on the ACC data, we can get 12 different estimated class end time from 12 ACC traces. Then, the median time is chosen as the calculated class end time (red vertical dashed line). That is to say, this class finishes early than the scheduled time. We apply IGTS on all the class data and extract the exact class start time and end time for the later data analysis.

3.4.2 Data Cleaning

Before pre-process the collected data, a data cleaning stage needs to be conducted to remove noises from wearable data. As describe in [47, 6, 106, 28, 104], there are several kinds of noises commonly happened in data collection from E4 wristband: (1) flat responses (i.e., 0 micro siemens) due to poor contact between the sensors and the skin. If the contact is not tight enough, the sensor will not measure anything; (2) abrupt signal drops due to the movement of the sensor (e.g., participant bumps the wristband onto the desk); (3) quantization errors. Since the EDA sensor records data through the two electrodes, it is more susceptible to noises compared with ACC, PPG and ST sensors. Therefore, we clean the data set mainly based on the quality of EDA data.

Firstly, we remove the data when students did not wear wristbands during the whole class or closed off the wristbands unintentionally during the class. Similar to [28], we then discard the signals containing a huge number of flat responses, abrupt signal drops and quantization error as suggested in [106, 104]. Finally, we discard the data on students who did not complete the survey. The data cleaning stage leaves us with 331 class data sessions. The final wearable data are gathered from 23 students and 6 teachers in 105 classes. 59 classes are short classes (mean = 39.15 minutes, STD = 1.15 minutes) and 46 classes are long classes with 2 periods (mean = 78.21 minutes, STD = 4.33 minutes).

The data cleaning stage eliminates 157 class data sessions due to the lack of survey data, which takes up to 32.17% of the total data with completed surveys. Though the amount of eliminated data is large, the size of our collected data is comparable and even larger than the previous studies. For instance, Lascio et al. [6] used 197 EDA data sessions after a reduction up to 37%, Gashi et al. [28] used 40 presenter-audience EDA pairs with the elimination of

72 pairs. Hernandez et al. [106] used 51 data sessions with the elimination of 28% from the original data.

3.4.3 Data Pre-processing

The pre-processing procedure is crucial to improve the quality of collected data. For EDA signals, we follow the same pre-processing steps as suggested in [6, 28, 106]. (1) Artifacts removal. To mitigate the influence of motion artifacts (MAs), we apply a median filter on EDA data with a 5-second window as in [6]. (2) Decomposition. EDA signal combines a tonic component and a phasic component [104, 128]. The tonic component varies slowly and reflects the general activity of sweat glands influenced by the body and environmental temperatures. The phasic component indicates rapid changes and is related to the responses to internal and external stimuli. EDA signals are decomposed using convex optimization via the cvxEDA approach [129]. (3) Normalization. The amplitude of the EDA signal varies a lot among different people [128] and thus limits the possibility of comparing the signal directly. We normalize the mixed, tonic and phasic EDA values similar to [28].

PPG data, also known as BVP, is provided by the E4 wristband. Similar to [45], we extract IBI signals by detecting the systolic peak of the heartbeat waveform signals from the raw PPG data (window size = 0.75 seconds). Linear interpolation is applied when the heartbeat intervals can not be detected successfully from the low-quality (e.g., motion artifacts) PPG signal. For the ACC data, we calculate the magnitude of 3-axis accelerations as $|a| = \sqrt{x^2 + y^2 + z^2}$. Then a median filter with 0.2 seconds is applied to the magnitude value. Finally, we apply a median filter on the ST data with 0.5 seconds.

3.5 Feature Extraction

We use various sensing devices to infer multidimensional engagement level of high school students. Table 3.3 summarizes these features. Then, we introduce the computed features and discuss why we explore such sensors and features.

Table 3.3: Description of the features computed for different sensors

<i>Sensors</i>	<i>Feature name</i>	<i>Description of features</i>
EDA	eda/tonic/phasic_avg	Average value for the raw, tonic, phasic data
	eda/tonic/phasic_std	Standard deviation for the raw, tonic, phasic data
	eda/tonic/phasic_n_p	Number of peaks for the raw, tonic, phasic data
	eda/tonic/phasic_a_p	Mean of peak amplitude for the raw, tonic, phasic data
	eda/tonic/phasic_auc	Area under the curve of the raw, tonic, phasic data
	num_arouse	Number of arousing moments during the class
	ratio_arouse	Ratio of arousing and unarousing moments
	level _k	Ratio of the number of level _k and the length of S_k
	eda/tonic/phasic_pcct	Pearson correlation coefficient with teacher
	eda/tonic/phasic_pccs*	Pearson correlation coefficient with average value of students
	eda/tonic/phasic_dtw_t	Dynamic time wrapping distance with teacher
eda/tonic/phasic_dtw_s*	Dynamic time wrapping distance with average value of students	
PPG	hrv_bpm	Average beats per minutes
	hrv_meani	Overall mean of RR intervals (Meani)
	hrv_sdnn	Standard deviation of intervals (SDNN)
	hrv_lf_power	Absolute power of the low-frequency band (0.04–0.15 Hz)
	hrv_hf_power	Absolute power of the high-frequency band (0.15–0.4 Hz)
	hrv_ratio_lf_hf	Ratio of LF-to-HF power
	hrv_rmssd	Root mean square of successive RR interval differences
	hrv_sdsd	Standard deviation of successive RR interval differences
	hrv_pnn50	Percentage of successive interval pairs that differ >50 ms
hrv_pnn20	Percentage of successive interval pairs that differ >20 ms	
ACC	acc_avg	Average physical activity intensity during the class
	acc_std	Standard deviation of physical activity intensity in class
	acc_dtw_t	Dynamic time wrapping distance with teacher
	acc_dtw_s*	Dynamic time wrapping distance with average value of students
	acc_pcc_t	Pearson correlation coefficient with teacher
acc_pcc_s*	Pearson correlation coefficient with average value of students	
ST	sktemp_avg/max/min	Average/maximum/minimum value of skin temperature
CO2	mean/max/min_co2	Average/maximum/minimum value of CO2
TEMP	mean/max/min_temp	Average/maximum/minimum value of indoor temperature
HUMID	mean/max/min_co2	Average/maximum/minimum value of humidity
SOUND	mean/max/min_temp	Average/maximum/minimum value of sound

3.5.1 EDA-based Features

EDA is a common measure of autonomic nervous system activity, with a long history being used in psychological research [130]. Recently, EDA measurements have been increasingly explored in affective computing such as the detection of emotion [47, 105], depression [48], and engagement [6, 106, 107]. From EDA data, we extract statistical features such as the

standard deviation from EDA (mixed, tonic, phasic) data, which reflects the overall general arousal during the class [6]. As suggested in [6], we extract the number of arousing/arousing states, the ratio of arousing states, etc. to show the momentary engagement during the class. The similarity-based method such as Pearson Correlation Coefficient (PCC) [131] and Dynamic Time Wrapping Distance (DTW) [132] are used to evaluate the physiological synchrony [76] of the target student and teacher. Inspired by [111], we also propose some new features (marked with *) to compute physiological synchrony between the target student and the average values of other students, which has proven to be effective in Table 3.5.

3.5.2 HRV-based Features

HRV is controlled by the autonomic nervous system (ANS), which can be used to evaluate human emotional arousal and cognitive performance [133, 134, 135, 136, 137]. With the help of *HeartPy* [138] toolkit, we compute HRV features from IBI signals extracted from the raw PPG data. As suggested in [139, 140, 141], HRV features can be analyzed from time-domain and frequency domain. On the time-domain, we capture features such as the mean/standard deviation of RR intervals (MeanI, SDNN) which estimates the overall HRV. We also extract features such as standard deviation/root mean square of successive RR interval differences (SDSD, RMSSD), number/percentage of successive interval pairs that differ larger than 20/50 ms (NN20, NN50, pNN20, pNN50), which describes the momentary change of HRV. On the frequency-domain where parameters are computed by applying Fast Fourier Transform (FFT) to the time series of RR intervals [141], we compute the absolute power of the low-frequency band (0.04-0.15 Hz) and high-frequency band (0.15-0.4 Hz). Besides, we compute the ratio of LF-to-HF power which reflects the overall balance of the ANS [142].

3.5.3 Accelerometer-based Features

Student behaviour can be inferred from ACC data, which helps us know more about student participation (e.g., team activities) and engagement level in class [111]. For ACC data, we extract features such as the average physical activity and standard deviation, which describes the statistical characteristics of the student movement during the class. Inspired by [111], we

propose the movement synchrony features such as the DTW/PCC between the target student and the average values of the other students.

3.5.4 Other Features

Student learning engagement has been found to be affected by the thermal comfort level of students in the classrooms [122], and thermal comfort is influenced by many factors such as indoor temperature, humidity, skin temperature, sound, CO₂ level, etc [143, 121, 14, 15]. Therefore, statistical features are calculated for indoor temperature, CO₂, sound and humidity, as the overall estimate of the indoor environment during a class. For ST data, statistical features are extracted to estimate the general arousal of student engagement. According to [144], when CO₂ level is higher than 1000 ppm, occupants may complain about the drowsiness and poor air, and when CO₂ level is higher than 2000 ppm, occupants will feel sleepy, headaches and lose attention. Therefore, the above features are selected to study student engagement.

3.6 Prediction Pipeline

Although engagement prediction is usually regarded as a classification problem, where engagement level can be divided into two or three categories [6, 45] based on specific thresholds, it is not a good practice to determine people's psychological characteristics using classification [12]. In this research, we choose regression rather than classification for multidimensional engagement prediction. In order to predict multidimensional engagement scores of students, we set up a regression-based pipeline as described below.

Engagement Score: We assign each student a score for each item in the self-report survey. To achieve this, we first reverse the responses in item 2 and item 4, as shown in Table 3.2. Then, we calculated a score based on the average of the 5-point Likert scale for each dimension of engagement and the overall engagement. Then we rescale the calculated score to 1 to 5, representing the engagement level being low to high. Figure 3.4 shows the calculated overall engagement score for 23 student participants. To save space, we do not display box plots of the distribution of the single-dimensional engagement score.

Regressors: We adopt LightGBM Regressor [145, 146] to predict self-reported multidimensional engagement scores. As one of the most powerful prediction models, LightGBM is an ensemble method combining a set of weak predictors (i.e., regression trees) to make accurate and reliable predictions. It builds the regression tree vertically (leaf-wise) while other algorithms grow trees horizontally (level-wise). It will choose the leaf with max delta loss to grow. When growing the same leaves, LightGBM algorithm can reduce more loss than other tree-based algorithms such as GBRT [147].

Validation: It is natural to use cross-validation to train and test prediction models when we are not in a data-rich situation. The purpose of cross-validation is to estimate the unbiased generalization performance of the prediction model. However, when using the test set for both model selection (hyperparameter tuning) and model estimation, the test data may be overfitted, and the optimistic bias may occur in the model estimation. Therefore, we adopt the nested cross-validation approach [148] with inner loop cross-validation nested in outer loop cross-validation. The inner loop is used for hyperparameter tuning and feature selection, while the outer loop is responsible for evaluating the performance on the test set. In the outer loop, similar to the previous human-centred research [6, 106], we first divide the data into n groups, where n represents the number of participants, i.e., $n = 23$. Each group contains the data for only one participant. Then we apply *k-fold cross-validation* [149] ($k = 5$) and on all student groups. Specifically, data from the same student (group) will not appear in the training and test sets at the same time. In the inner loop, the remaining data groups are split into L ($L = 3$) folds, where each fold serves as a validation set in turn. Then we train (grid search) the hyperparameters on the training set, evaluate them on the validation set, and select the best parameter settings based on the performance recordings over L folds. We use the importance vector generated from LightGBM to reduce the feature dimensionality, which calculates feature importance automatically by averaging the number of times a specific feature used for splitting a branch. Higher values indicate higher feature importance. Top-10 features are selected as the new input features to the LightGBM regressor. The heuristic of choosing 10 features is we find that the prediction error is lowest under this threshold in the experiment.

Similar to [6, 120], we also perform leave-one-subject-out (LOSO) [150] validation to eval-

uate the impact of data from individual participant on the overall prediction error. For both k -fold and LOSO validation approaches, we calculate the average performance score (i.e., MAE and RMSE) of the regressor in each iteration.

Baselines and Metrics: We compare the proposed engagement prediction model with three baselines. The first baseline is the standard linear regressor [151], one of the most widely used regression models. The second baseline takes the average score of each dimension of engagement. The third baseline randomly generates a sample from the distribution of engagement scores and regards it as a predicted value. Similar random baselines have been widely used in previous ubiquitous computing studies such as [120, 6]. To evaluate the prediction performance of the proposed model, we use the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) [152] metrics.

3.7 Results and Discussion

In this section, we conduct extensive experiments to evaluate the prediction performance of *n-Gage*. We answer the first question ‘*Can we measure the multiple dimensions of high school student’s learning engagement including emotional, behavioural and cognitive engagement in high schools with sensing data in the wild?*’ in Section 3.7.1. We answer the second question ‘*Can we derive the activity, physiological, and environmental factors contributing to the different dimensions of student learning engagement? If yes, which sensors are the most useful in differentiating each dimension of the engagement?*’ in Section 3.7.2. We also study how different settings can help improve the performance of *n-Gage*. Unless otherwise stated, the prediction models are built with LightGBM regressors using all sensors and evaluated by k -fold nested cross-validation by default.

3.7.1 Overall Prediction Results

We first evaluate the overall prediction results for *n-Gage* with all sensors available. Table 3.4 displays MAE and RMSE scores of *n-Gage*’s engagement regression in different dimensions. In particular, the overall engagement is calculated by the average of engagement scores from all

Table 3.4: Prediction performance for emotional, cognitive, behavioural, and overall engagement with all sensing data

<i>Dimension</i>	<i>MAE</i>				<i>RMSE</i>			
	LGBM.	LR.	Average	Random	LGBM.	LR.	Average	Random
<i>Emotional</i>	0.675	0.714	0.747	1.059	0.851	0.878	0.928	1.326
<i>Cognitive</i>	0.906	0.921	0.977	1.288	1.113	1.128	1.176	1.658
<i>Behavioural</i>	0.783	0.811	0.871	1.235	0.960	0.980	1.135	1.540
<i>Overall</i>	0.602	0.614	0.641	0.891	0.753	0.769	0.792	1.125

questions related to the engagement, which is commonly used in previous engagement studies [45, 6, 91]. From Table 3.4, we can see that in terms of MAE and RMSE, *n-Gage* achieves higher prediction performance for all dimensions of engagement than all baselines, demonstrating its potential for multidimensional engagement prediction.

Notably, among each dimension of engagement, *n-Gage* works best on predicting emotional engagement. The emotional engagement regression model obtain 0.675 of MAE and 0.851 of RMSE, which is lower than 0.384 (36.26%) and 0.475 (35.82%) of the random baseline. The reasons why *n-Gage* predicts emotional engagement best are possibly two-fold: (1) compared with cognitive and behavioural engagement, emotional engagement is most suitable for evaluation through self-report surveys [91], resulting in a more realistic and stable student emotional engagement measurement (ground truth). (2) emotional engagement is more easily detected by sensors (e.g., EDA and PPG) as it reflects the degree of emotional arousal, thereby producing fluctuations in physiological signals [6, 29, 47].

Although the MAE of cognitive engagement regression is higher than other models, it is still lower than random baseline of 0.382 (29.66%) in MAE and 0.545 (32.87%) in RMSE. The possible reason is that cognitive engagement is more challenging to be assessed by the wearable and indoor sensors than electroencephalography (EEG) sensors [153]. By contrast, *n-Gage* has the lowest prediction error of 0.602 in MAE and 0.753 in RMSE in overall engagement assessment. According to the education research [91, 92], although the multidimensional concept of engagement has been well accepted, the definitions of three dimensions of engagement vary with considerable overlap across components. Therefore, the overall engagement is easier to be evaluated and predicted than the single-dimensional engagement.

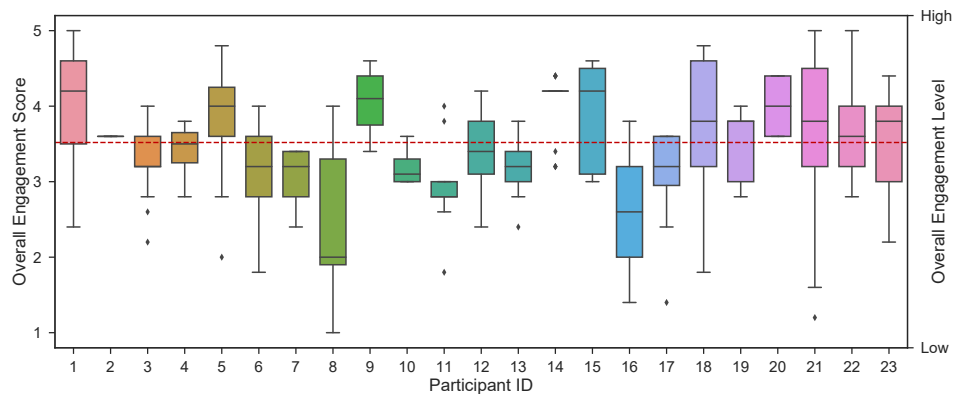


Figure 3.4: Boxplot of the overall engagement scores for 23 student participants. The red dashed line represents the average score for all participants. The participant ID shown in the figure is randomly generated to maintain the privacy of participants

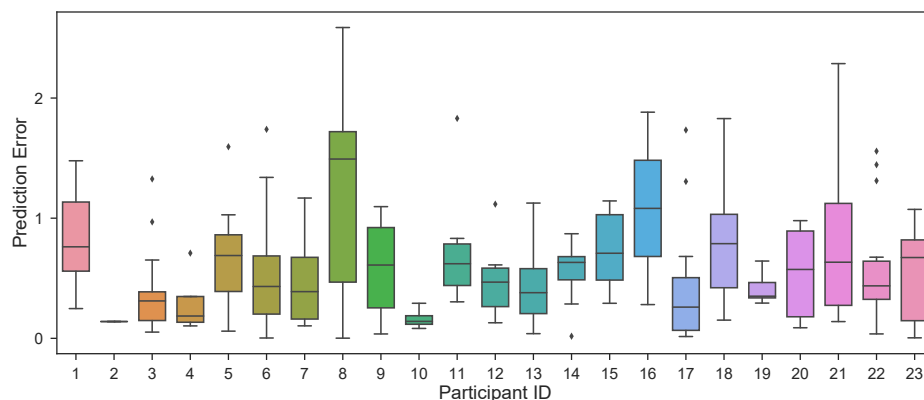


Figure 3.5: Prediction error for overall engagement scores for 23 student participants

We also compare the prediction results with when a standard linear regressor is learned. From Table 3.4, the linear regression model has much higher prediction performance than both average and random baseline models (e.g., 31.09% lower than random baseline model in MAE for overall engagement prediction), indicating the effectiveness of extracted features in engagement prediction. However, the performance of linear regressors is not comparable to the LightGBM in all dimensions. This is because LightGBM has a good ability to capture non-linear feature-target relationships which is more flexible than simple linear regressors. To summarize, we believe that the performance of *n-Gage* is benefited from both the extracted features and powerful non-linear mapping provided by the LightGBM.

Table 3.5: The most influential features on multidimensional engagement

<i>Engagement</i>	<i>Association</i>	<i>Most influential features</i>
<i>Emotional Engagement</i>	(+)	acc_pcc_s, tonic_a_p, eda_pcc_s
	(-)	acc_avg* , sktemp_avg* , eda_dtw_t
<i>Cognitive Engagement</i>	(+)	intemp_min* , level_1, hrv_ratio_lf_hf
	(-)	acc_pcc_s* , co2_max, acc_std
<i>Behavioural Engagement</i>	(+)	acc_std, acc_pcc_s, eda_pcc_avg
	(-)	sktemp_avg* , acc_pcc_t* , acc_dtw_t
<i>Overall Engagement</i>	(+)	level_1, tonic_a_p, intemp_max
	(-)	acc_dtw_t* , sktemp_avg, acc_avg

* indicates p-value ≤ 0.01 .

We then discuss the impact of data from the individual participant on the overall prediction error. We train and test the regressors using the LOSO validation approach which enables us to evaluate the ability of models to accurately predict a new participant not included in the training set. Figure 3.5 shows the boxplot of absolute prediction error per participant. Interestingly, each participant has a very different error distribution. For instance, participants 8 and 16 have the highest median value (1.492 and 1.082) and standard deviation (0.801 and 1.132) of prediction errors. From Figure 3.4, we observe that both participants have a much lower engagement level than the others. Since the regression model is built on the data from all the other participants, it does not work well when the participant (testing set) has a different distribution from the training set. The potential solution is to build participant-wise or groupwise prediction models, as introduced in [154]. In conclusion, we believe the prediction errors come from both the specific participants and overall prediction bias. We will further investigate this issue in future research.

3.7.2 Impact of Sensor Combinations

We will explore the physiological, activity and environmental factors contributing to the different dimensions of student engagement. We compute the Pearson Correlation Coefficient (PCC) between the extracted features and multiple dimensions of engagement, and then list the three most influential features in Table 3.5. We find many EDA features related to the

peaks of tonic EDA signals and physiological synchrony are related to the multidimensional engagement. In previous research, EDA features are generally considered as a good indicator of physiological arousal (e.g., emotional and cognitive states) [103, 104], which have been explored in the detection of engagement [6, 106, 107]. For the HRV features (e.g., ‘hrv_ratio_lf_hf’), they are shown to be correlated with cognitive engagement as HRV is an autonomically dependent variable and has been used to predict student engagement in [109]. Similar to EDA and HRV features, we notice that the average skin temperature (‘sktemp_avg’) are negatively correlated with engagement, as ST reflects the sympathetic nervous activity and attention states [155] which has been used for mind-wandering prediction [156] and stress detection [157].

For activity factors, it is interesting to find that many ACC features are highly correlated with engagement. The accelerometer is a popular and powerful sensor for quantifying human behavioural patterns [31, 12]. ACC features have been utilised to sense audience engagement using interpersonal movement synchrony [111]. In the experiment, we observe that the average physical intensity during class is highly negatively correlated with emotional engagement. This leads us to believe that when students are negatively engaged, they tend to perform more physical movements in the class. As for environmental factors, we find that the maximal CO₂ level is negatively associated with cognitive engagement, while the indoor temperature is positively associated with engagement. This may be because CO₂ has a negative impact on people’s cognitive load [123, 124], and then affects student cognitive engagement. This result highlights the need to ventilate the classroom timely to keep students engaged. Interestingly, we notice that the maximal indoor temperature in the class is positively correlated with overall engagement. One possible explanation is that during the data collection period (winter and spring), the indoor temperature is low and moderately higher indoor temperature makes students feel thermally comfortable [121] and therefore more engaged in learning [122].

Then we investigate the most useful sensors in predicting each dimension of student engagement and explore the performance of *n-Gage* when only a set of sensors available. In this chapter, we use E4 wristbands and Netatmo indoor weather stations for student engagement assessment. However, when other schools want to generalize the system for automatic engagement measurement, it is likely that only a few sensors available considering the types

Table 3.6: Summary of the Prediction performance of multidimensional engagement using different sensor combinations. \mathcal{X}_1 indicates all the wearable data including EDA, HRV, ACC and ST data, and \mathcal{X}_2 means the indoor environmental data including CO₂ and temperature data

<i>Data source</i>	<i>MAE/RMSE</i>			
	<i>Emotional</i>	<i>Cognitive</i>	<i>Behavioural</i>	<i>Overall</i>
<i>EDA</i>	0.697/0.877	0.948/1.149	0.851/1.019	0.637/0.800
<i>HRV</i>	0.714/0.901	0.940/1.140	0.833/1.002	0.659/0.812
<i>EDA+HRV</i>	0.699/0.875	0.949/1.151	0.841/0.989	0.621/0.783
<i>EDA+ACC</i>	0.679/0.860	0.914/1.124	0.816/0.987	0.626/0.789
<i>HRV+ACC</i>	0.691/0.875	0.910/1.125	0.809/0.979	0.641/0.796
<i>EDA+HRV+ACC</i>	0.679/0.860	0.909/1.122	0.800/0.965	0.620/0.778
\mathcal{X}_1^*	0.673 /0.851	0.910/1.126	0.811/0.980	0.619/0.775
$\mathcal{X}_1 + \mathcal{X}_2^*$ (<i>all</i>)	0.675/0.851	0.906 / 1.113	0.783 / 0.960	0.602 / 0.753

* indicates the proposed combination of features for engagement prediction.

of wearables and installation of indoor weather stations. In this experiment, we use different combinations of sensors as shown in Table 3.6 to train the regressors, where \mathcal{X}_1 indicates all the wearable sensors including EDA, HRV, ACC and ST, and \mathcal{X}_2 represents all the environmental sensors containing CO₂, TEMP, HUMID and SOUND sensors. Besides, we predict student engagement using only EDA as in [6], single PPG (HRV) as in [62], and EDA+HRV as in [45]. Since accelerometers are naturally available in wearables and have been used for engagement measurement [111], we add ACC to the above sensor combinations for the first time. Then we utilise all wearable sensors and indoor sensors for more accurate engagement prediction.

For each sensor combination, we use nested cross-validation to train and test the regressors as described in Section 3.6, to achieve optimal feature selection and parameter tuning. Table 3.6 displays the regression result with different sensor combinations. Different combinations are useful for different dimensions of engagement. For instance, a single EDA sensor works well for emotional engagement prediction while less useful in predicting behavioural engagement unless involving ACC together. This is reasonable because EDA is a reflection of emotional arousal, while ACC is capable of quantifying human behavioural patterns [31, 12]. On the other hand, the combination of EDA and HRV sensors has similar prediction performance compared to using a single EDA sensor, which is consistent with the fact that not many HRV features are

Table 3.7: Multidimensional engagement regression result for different subjects

Subject	MAE/RMSE			
	Emotional	Cognitive	Behavioural	Overall
Maths	0.686/0.841	0.841/0.965	0.750/0.891	0.603/0.738
English	0.609/0.779	0.893/1.010	0.694/0.819	0.510/0.629
Language	0.645/0.814	0.829/0.903	0.799/0.900	0.593/0.758
Science	0.646/0.829	0.895/0.941	0.758/0.856	0.575/0.720
Politics	0.674/0.835	0.947/1.057	0.660/0.731	0.525/0.671
Average	0.652/0.820	0.881/0.975	0.732/0.839	0.561/0.703

highly correlated with engagement. When there is no EDA sensor (especially in commercial off-the-shelf smart wristbands), the HRV+ACC combination can achieve similar prediction performance on cognitive and behavioural engagement compared to EDA+HRV+ACC.

Meanwhile, it can be observed that the combination of all wearable sensors (\mathcal{X}_1) has the lowest prediction error for emotional engagement. When considering wearable sensors (\mathcal{X}_1) with indoor sensors (\mathcal{X}_2), *n-Gage* can achieve the best performance on the behavioural, cognitive and overall engagement, and has similar prediction performance in emotional engagement with \mathcal{X}_1 . The underlying reason is that CO₂ and indoor temperature mainly affect students' cognition load and behavioural patterns. For example, students may lose attention (related to behavioural engagement), sleepy (related to cognitive engagement) [144] during class when the CO₂ level is too high (e.g., larger than 2000 ppm), but this does not necessarily mean that students do not like the class (related to emotional engagement). The above results illustrate the importance of taking indoor environmental changes into account for student engagement prediction and creating the optimal environment to keep students engaged in class.

3.7.3 Impact of Class Subjects

Now, we investigate whether considering different school subjects could improve the prediction performance of *n-Gage*. Our assumption here is that different subjects may lead to different learning requirements, thinking styles and emotional preferences. Then, student engagement levels and physiological status may be affected accordingly.

To validate this hypothesis, we establish regression models for each subject (i.e., Language,

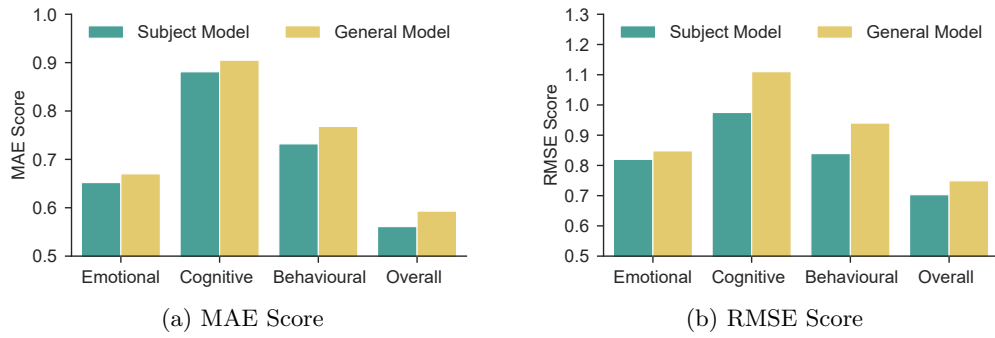


Figure 3.6: Prediction performance for the average subject model and general model

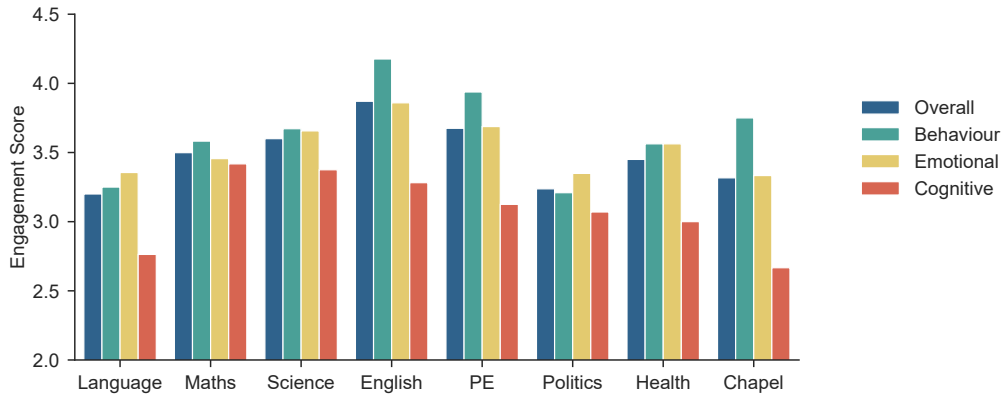


Figure 3.7: Engagement scores on different subjects

Maths, Science, English, PE, Politics, Health, Chapel) to isolate differences in class subjects and engagement assessment. Table 3.7 summarizes MAE and RMSE scores of the regressors over different subjects. We do not consider the Health, Chapel and PE classes because the number of survey responses are limited (less than 30) in those classes which may affect the prediction performance. We also compare the average prediction performance of 5 regression models (i.e., Maths, English, Language, Science, Politics) with the general regressor model in Figure 3.6. The results indicate that, compared with building the general regression model including all subjects, building regression models by school subjects can significantly improve the prediction performance.

To better understand the underlying cause behind the improved regression performance, we review the self-reported engagement scores. Figure 3.7 shows that students have very different

multidimensional engagement scores among different subjects. For instance, while students have the highest behavioural and emotional engagement score in English class, they have the highest cognitive engagement score in Maths class. The possible reason is that students enjoy English classes most and thus like to follow the rules from English teachers. Due to the fact that Math know-how is cumulative and usually contains complex concepts, students may put more effort to comprehend the contents in Maths class, thus leading to a high cognition engagement score. Overall, these observations serve as evidence that building models for each subject can lead to significantly improved prediction performance.

3.7.4 Discussion

We have shown that it is possible to infer multidimensional student engagement by using multiple wearable and environmental sensors. Meantime, we will present the following interesting discussion points.

- **Engagement and class time.** A preliminary study is conducted to investigate the correlation between self-reported student engagement and class time during the school day. Figure 3.8a shows the average engagement scores for the different class time (morning, noon and afternoon). Overall, we observe that classes in the noon show higher engagement levels in all dimensions. Classes in the afternoon (after lunch) have the lowest engagement score, especially in the behavioural and emotional dimensions. Particularly, it is interesting to notice that students have a much higher behavioural and emotional engagement level than the cognition level despite the time of the classes. These observations provide directions for further research in maximizing student engagement by a more reasonable arrangement of class schedule according to the nature of each course.
- **Engagement and thermal comfort.** In the background survey, most students agree that ‘*When I am engaged in class, I could get distracted when the room is too hot or too cold*’ (see Section 3.3.2.2). As another investigative point, Figure 3.8b shows the relationship between self-reported engagement and thermal preference (i.e., warmer, cooler, no change) [143] of the students in class. The results show that students who feel the room

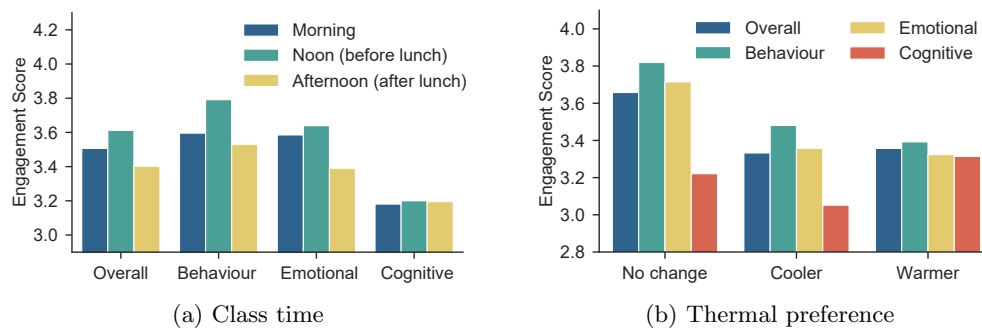


Figure 3.8: Engagement scores with different class time and thermal comfort

thermally comfortable have a higher overall engagement level compared to other groups. In particular, students who prefer a cooler environment usually have the lowest cognition engagement. This reminds us that creating the thermally comfortable environment is necessary to improve student engagement in class, especially considering the individual differences in thermal sensation [158].

- Real-time measurement.** Students' engagement level during a class may vary with the learning content and teaching style. Real-time anonymous engagement tracking can provide teachers with student engagement level and help teachers understand the impact of different teaching contents on student engagement, thereby better adjusting teaching speed and teaching methods. However, the challenge is how to obtain the fine-grained ground truth of student engagement multiple times during the class without disturbing students' studying. One potential approach is ecological momentary assessment (EMA) [29] which repeatedly prompts students to report their engagement level. Though EMA is usually considered a good method of *in situ* data collection, students may be disturbed and distracted in class if they are required to complete the EMA. Overall, ground truth data collection is challenging and more reliable methods need to be investigated.

3.8 Implications and Limitations

This chapter addresses the possibility of automatically predicting students' in-class emotional, behavioural and cognitive engagement using wearable and indoor sensing technology, which provides opportunities for the future design of feedback systems in the classroom. The feedback system has the potential to benefit both teachers and students.

Teachers play an important role in influencing student engagement [91]. With the feedback from students after each class, teachers can evaluate, and, if necessary, adapt or change teaching strategies (e.g., increase time for student thinking, allow students time to write, assign reporters for small groups [159]) for creating the right learning climate to keep students engaged [160]. For instance, when teachers focus more on academics and fail to create a positive social learning environment, students are likely to be emotionally disengaged and worried about making mistakes. Contrarily, when teachers focus more on the social dimension and neglect the intellectual dimension, students possibly experience low cognitive engagement for learning [161, 91]. With such a feedback system, teachers can observe multidimensional student engagement and create the intellectually challenging and socially supportive learning environment.

Further, if this system is deployed, using *n-Gage*, teachers can take timely measures to improve learning experience for students, such as planning learning schedules, re-engaging students with the low engagement, and ventilating the room to let the fresh air in. While overcoming student disengagement is complicated, we do believe teachers can benefit from the engagement feedback of students after every class instead of few times in a term [162, 6], contributing to higher student achievements and protecting students from dropping out of school [91].

Students wearing wristbands are able to self-track their multidimensional in-class engagement, which positively influences academic achievements and is usually regarded as the predictor of learning outcomes [163, 94, 91]. Being conscious of in-class engagement is an effective *quantified-self* [164, 165] approach to promote self-regulation and reflective learning [166] for students. Once students are aware of how much effort they are putting into learning, they can work towards their personal goals by optimizing their study practices and learning strategies

(e.g., practice active listening and thinking, make study plans for different subjects) [165, 167]. Additional strategies such as gamification [168, 169] can also be deployed along with *n-Gage* measurements.

For real-world deployments, the feedback system can still work when only a subset of sensors available (see Section 3.7). For instance, when there are no indoor sensors installed, wearable sensors can be used for accurate engagement prediction especially for the emotional engagement. The system can also allow more sensors to be integrated in the future when becomes available.

The current studies have some limitations that needed to be addressed in future research. Firstly, collecting data from more student participants in the same class may bring new opportunities for data analysis. There are 59 Year 10 students in total, but only 23 students voluntarily became participants and wore wristbands. Compared to students who did not participate, participants may share some similar personality traits and have higher potential to engage in class most of the time.

Secondly, we agree that collecting the ground truth of student engagement is challenging because we need to find a compromise between taking long psychological surveys for more accurate measurement and enabling students to complete surveys faster without affecting their study or rest. Therefore, a more robust way of evaluating multidimensional student engagement needs to be investigated in the future.

Thirdly, the quality of survey responses varies. Online surveys are conducted three times a day, and the total response rate is 35.3%. Since completing surveys multiple times a day may become a burden, students are likely to answer the questions unseriously. Therefore, in this study, we only encourage rather than urge them to complete the survey, which to a certain extent guarantees the quality of responses. Figure 3.9 shows the survey completion time for all responses from participants. Most participants complete the survey in 30 to 50 seconds, but some participants complete the survey in less than 15 seconds. Though the survey completion time may be affected by many factors and varies from person to person, it is still one of the indicators of response quality [73]. In future research, it will be interesting to explore patterns from survey completion time data and assign appropriate weights to survey response for more

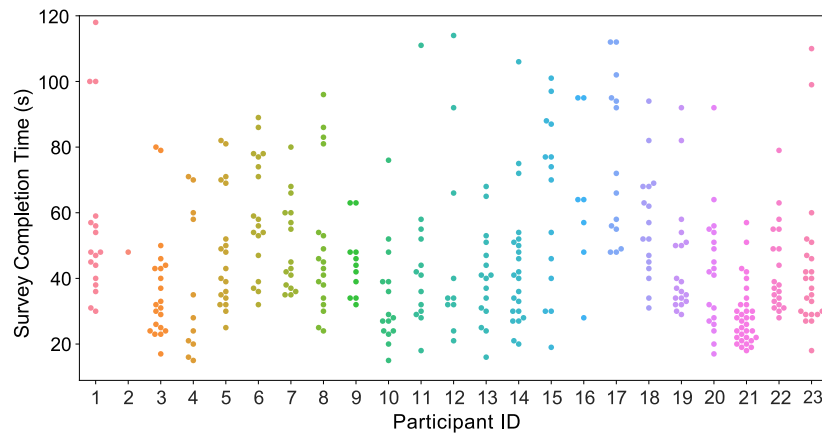


Figure 3.9: Survey completion time for different participants. Each point represents the survey completion time for one response

accurate prediction of student engagement.

During the in-situ data collection, the data recorded by the wristband is not always continuous. For many reasons, we face a considerable loss of data: (1) on each school day, an average of 2 to 4 participants are on sick leave and cannot wear wristbands; (2) 6 participants went abroad to a study program on the second week of data collection; (3) students were curious about the wristbands, especially in the first few days, and they pressed the button again and again out of curiosity. Some students accidentally closed their wristbands, so their data was lost for hours or even the whole day.

Though significant efforts have been made to make the maximal use of the collected data, 32.17% traces must to be removed from the analysis due to the loss of survey data, the incomplete data during the class, the presence of long-time of flat responses, artifacts and quantization errors as discussed in Section 3.4. Despite the fact that we have cleaned and pre-processed wearable data to eliminate noises, collecting physiological data in the wild still faces huge challenges, especially for young students. In our research, one of the main noise is from the poor contact between the sensors and skin, which can be fixed by tightening the wrist strap to the skin. However, this will also increase awareness of wristband during class, resulting in student in-class disengagement and even more motion artifacts.

3.9 Conclusion

In this chapter, we proposed *n-Gage*, an engagement sensing system that can capture students' physiological responses, physical movements, and environmental changes to infer multidimensional engagement (behavioural, emotional and cognitive engagement) level in class. We evaluated the system by combining weather station data and wearable data collected from 23 Year 10 students and 6 teachers over 144 classes in 4 weeks in a high school. Some new features were proposed to characterize different aspects of student engagement. Extensive experiment results showed that *n-Gage* can predict student behavioural, emotional and cognitive engagement score (1 is the lowest score and 5 is the highest score) with an average MAE of 0.788 and RMSE of 0.975. We further demonstrated the most influential features and how different sensor combinations/school subjects affect student engagement. Finally, we have shown some interesting findings that the maximal CO₂ level is highly negatively correlated with student cognitive engagement; class time (morning, noon and afternoon) and thermal preference (warmer, cooler or no change) may affect the level of student engagement, which provides beneficial insights for educators and school managers to improve student learning engagement in high school. Though not perfect, *n-Gage* is still a very promising first-step towards multidimensional in-class engagement tracking for students. As a contribution, *n-Gage* can indicate the future design of feedback systems, assisting students and teachers in a variety of ways such as promoting students' self-regulation and reflective learning, helping teachers create a right learning climate for students.

Chapter 4

Understanding Classroom Seating Behaviours from Perceived and Physiologically Measured Student Engagement

In Chapter 3, we show that student engagement can be inferred from their physiological sensing signals. Some popular features are extracted to represent physiological arousal and synchrony between students. In relation to *RQ-3*, this chapter explores how individual and group-wise classroom seating experiences affect student engagement by using data gathered from wearable physiological sensors. Using the dataset collected in Chapter 2, we will show that the individual and group-wise classroom seating experience is associated with perceived student engagement and physiologically measured engagement, as measured by EDA. We will show that students who sit close together are more likely to have similar learning engagement and tend to have high physiological synchrony.

4.1 Introduction

Anecdotal evidence has suggested that student seating experience has played an important role in affecting student engagement, participation, attention and academic performance in classrooms [170, 171]. Different seating locations provide different degrees of access to learning conditions and resources (e.g. the ability to see and hear the teacher [172]), which impacts students' physical and mental comfort, and their concentration in class [170]. For example, a spacious seat offers physical comfort, a seat near the door results in many distractions, and a seat close to the teacher can result in more attention from the teacher [170, 173]. In addition, peers have an impact on student engagement, most especially a negative impact, as peers may invoke more non-academic interactions and off-task behaviours [170, 174]. However, peers can also have a positive impact by encouraging active learning through discussion [170, 175]. In general, all behaviours or influences related to the seating experience affect student engagement in class, an essential psychological representation. *Student engagement* refers to the degree of student involvement or interest in learning, as well as their degree of connection with the class or each other [176]. The study of student engagement has attracted growing interest to address problems, such as high dropout rates and declining motivation and achievement [91, 92]. Specifically, Fredricks et al. [91] identified three dimensions to student engagement: emotional engagement (i.e. interest, enjoyment and enthusiasm [177, 92]), cognitive engagement (i.e. concentration and comprehension [99, 92]) and behavioural engagement (i.e. effort and determination [96, 91, 6]).

There is a large body of research investigating the impact of seating location on student engagement and classroom experience [178, 174, 179, 180]. Shrenoff et al. [178] conducted a study in a large university lecture hall and found that students who sit in the back report the lowest perceived engagement, and students who sit in the front report the highest engagement. Joshi et al. [174] analyzed the influence of multimedia and seating location on academic engagement, and indicated that students who sit close to the multimedia screen pay more attention than students in the middle row. Both Holliman et al. [180] and Becker et al. [179] found that student performance declined as teacher-to-student seating distance increased.

However, one common limitation of previous studies is that they relied on the self-report

survey or EMA [181] as an engagement measurement, which may be prone to subjectivity and various response biases (e.g. social desirability and extreme rating bias) [40]. By investigating the accuracy of self-report information, Moller et al. [59] pointed out that the self reports should not be trusted blindly and researchers must take into consideration that the responses can be unreliable. To overcome this limitation, sensing physiological signals (e.g. electrodermal activity) could be an alternative or even a better method for understanding the relationship between seating location and student engagement. Electrodermal activity, also known as *Galvanic Skin Response* (GSR), has been used in physiological and psychophysiological research since the 1880s [104]. It refers to changes in the electrical conductance of the skin in response to sweat secretion, which is controlled by the *Sympathetic Nervous System* (SNS) [182]. The primary process of the SNS is to stimulate the body's *fight or flight responses* [183]. When the sympathetic nervous system is highly aroused, the activity of the sweat glands increases, which in turn increases skin conductance [184].

In this way, EDA is widely used to measure arousal in psychology, which is a broad term representing overall activation and is recognised as one of the two dimensions (i.e. arousal and valence) of emotion responses [185]. Although measuring arousal is not exactly the same as measuring emotion, it is an important component of emotion, and it has been found to be a strong predictor of attention, perception and cognitive processing [186, 187]. Previous research has shown that EDA is associated with some psychological constructs, such as arousal, stress and cognition load [188]. As an indirect method for measuring arousal and increased mental workload, EDA has been adopted to evaluate engagement in various domains [28, 6, 30, 45, 106, 111]. However, most researchers studied engagement on the individual level using either behavioural or physiological patterns. To our knowledge, no prior research has explored student engagement at the group level or investigated distinct clusters of physiological signals from students which share similar engagement levels.

On the other hand, in order to facilitate the statistics of the spatial data (seating locations) in questionnaires, most research focused on the general locations (e.g. front, middle, back) in traditional seating arrangement types (e.g. grouped tables or rows-and-columns) rather than the exact seating locations in flexible seating arrangements scenarios [189, 180, 178, 174].

Furthermore, while some research has been carried out on individual seating experiences, only a few studies have investigated the social aspects of these seating experiences. To deal with above issues, in this research, we collect the accurate seating location of students with flexible seating arrangements. In addition to investigating individual experiences, the seating experience and engagement are also investigated in a group-wise manner, i.e. student peers and student groups.

Therefore, in this chapter, we aim to answer the following research questions: 1. *How does seating proximity between students relate to their perceived learning engagement?* 2. *How do students' group seating behaviours relate to their physiologically measured engagement level (i.e. physiological arousal and synchrony)?* This research contributes to empirical evidence on how the classroom seating experience affects student engagement. In this study, seating experience is defined as seating location (e.g. in the front or back of the classroom). We present the results of an in-situ study in a high school in which we collected survey and wearable data from 23 student participants attending 10 different courses over four weeks. We investigate the relationship between the student seating experience and student engagement. The results show that students who sit close together are more likely to have similar engagement levels than those who sit far apart. In summary, the contributions of this research are as follows:

- We investigate how student seating experience affects student engagement by understanding their perceived engagement and physiologically measured engagement, measured by EDA signals. A field study was conducted on a high school campus, with 23 student participants attending 10 courses over hundreds of classes. During the four-week data collection, each participant was asked to wear the E4 wristband during school and report their learning engagement and seating location in the classroom.
- To the best of our knowledge, we are the first to study how individual and group-wise classroom seating experiences relate to student engagement. We analyse student engagement from two different perspectives: perceived engagement and physiologically measured engagement.
- For the first time, we identify statistically significant correlations between student seating behaviours and students' perceived and physiologically measured engagement. Our

results show that students who sit close together are more likely to have similar learning engagement and physiological synchrony than students who sit far apart.

4.2 Related Work

This section describes related works on student engagement and the classroom seating experience in educational settings. A brief summary of the main related works can be found in Table 4.1.

4.2.1 Student Engagement in Educational Research

The concept of student engagement has a history from ten to seventy years [176]. In the 1930s, the educational psychologist Ralph Tyler began to exploring the time students spent at work and its impact on learning. Harper et al. [190] argued that engagement is more than participation or involvement; it requires feelings, sense-making as well as activities. In 2004, drawing on Bloom's research [191], Fredricks et al. [91] identified three dimensions of student engagement: (1) *emotional engagement*. Students who are emotionally engaged would experience affective reactions, such as interest, enjoyment or a sense of belonging [177, 92]; (2) *behavioural engagement*. Behaviourally engaged students usually abide by behavioural norms, such as attendance and participation, and exhibit an absence of destructive or negative behaviour [96, 91]; (3) *cognitive engagement*. Students who engage cognitively would be invested in their learning, show a willingness to go beyond the requirements and exert efforts to comprehend complex ideas [99, 92].

Student engagement at a particular school or university is increasingly recognised as an effective indicator of institutional excellence, rather than traditional characteristics (e.g. the numbers of books in the library or number of Nobel laureates in the faculty). The self-report survey is one of the most widely used tools for measuring student engagement, and some of the most popular are the *National Survey of Student Engagement* (NSSE) [176], *Engagement vs. Disaffection with Learning* (EvsD) [102], *Motivated Strategies for Learning Questionnaire* (MSLQ) [100], and *School Engagement Measure* (SEM) [101]. Recently, some short question-

naires (e.g. *In-class Student Engagement Questionnaires* (ISEQ) [62]) have been designed for *Experience Sampling* [192], which reduces the problem of recall failure and provides instant feedback to inform a cycle of quality improvement.

4.2.2 Classroom Seating Experience and Student Engagement

4.2.2.1 Flexible Seating in the Classroom

Standard classrooms are set up in the traditional linear seating arrangement, with standard desks and chairs facing the podium. However, pedagogical studies [170, 193, 194, 195, 174, 173] have found that when compared with traditional seating, flexible seating provides a more comfortable environment and has multiple benefits that improve educational activities. Flexible seating uses various seating options [173] (e.g. balls or cushions and standing or seating options) or non-linear seating arrangements (e.g. u-shaped or semicircular) [170, 174]. Yang et al. [196] found that in English learning courses, the semicircular arrangement facilitated student engagement by enhancing communication, concentration and the classroom environment when compared with the traditional arrangement. In addition, researchers [170, 197, 198] have suggested that tailoring classroom seating arrangements to educational activities helps manage students and results in better perceptions of student behaviours.

4.2.2.2 Seating Preference and Student Engagement

Most researchers concluded that seating location has an effect on student engagement, attention, involvement and motivation [194, 197, 171]. Ngware et al. [197] showed that students' seating locations were related to their academic abilities. Sitting in the front row led to greater learning gains (5%-27%) when compared with sitting far away from the front. Burda et al. [171] indicated that students who choose the back seats may be more passive and feel more comfortable sitting far away from the teacher to ensure less interaction. These students were often observed disengaging from the class. Gyanendra et al. [174] found that university students who preferred to sit in the first or the last two rows of the classrooms paid greater attention to the multimedia screen and had higher grades than those who sat in the middle.

Table 4.1: Related works that studied student seating experience and learning engagement, performance and emotion

<i>Ref.</i>	<i>Seating Options</i>	<i>Seating Data</i>	<i>Assessed Item</i>	<i>Participants</i>	<i>Data Type</i>
[198]	Rows-and-columns, self-select	Accurate x, y position in the classroom	Academic performance, engagement	182 university students from 4 disciplines	Academic score
[174]	Rows-and-columns, self-select	Rows and groups	Academic performance, attention	25,000+ university students	Attendance, academic score, head-down activity, eye activity
[178]	Rows-and-columns, self-select, semi-circular	Front, middle, back of the classroom	Class experience, engagement, course performance	407 university students	self-reported engagement, attention, classroom experience, course grade
[199]	N/A	N/A	Academic performance, study satisfaction, study usefulness	31 university students (93 sessions)	Self-reported experience, HR, EDA, TEMP, BVP
[7]	N/A	N/A	Class experience, emotion state	24 university students (1008 sessions)	Self-reported experience and emotional state, EDA
[6]	N/A	N/A	Emotional engagement	24 university students (984 sessions)	Self-reported emotional engagement, EDA, BVP

Kalinowski [200] compared academic scores with students' seating preferences and assigned seating locations. They found that students with higher GPAs preferred to sit in the front. This suggests that a correlation between seating location and motivation. Ka et al. [198] examined the relationship between the seating location and academic performance of 182 university students from four disciplines. It highlighted a significant relationship between seating location and academic performance. However, the relationship varied between the academic disciplines. In particular, in fields requiring more active and integrated learning, such as nursing, students who sat in the front performed better academically than those who sat at the back because of their higher level of participation and engagement. It is worth noting that the only the academic scores were compared, which is not sufficient to represent academic performance or engagement.

4.2.2.3 Seating Proximity and Student Engagement

The proximity of the student to the teacher [201, 178] and student groups [202, 203] can affect student engagement and satisfaction in the classroom. Millard et al. [201] found that when undergraduates were periodically moved from one location to another, students' enjoyment and productivity changed significantly in both free and assigned seating settings. They also found that increasing proximity of student and teacher was related to decreasing self-reported motivation, enjoyment, interests and feeling apart. Shernoff et al. [178] suggested that seating location and distance from the teacher are consistently correlated with student engagement, attention and course performance. Although social interactions, such as group discussions, positively impact engagement and encourage active learning [204], there is a risk that non-academic interactions also increase, thereby distracting students. Teachers can assign seats in an attempt to control student interactions, e.g. specifically assigning seats to reduce non-academic peer interactions, which have a negative influence on academic achievement [170]. Gremmen et al. [203] investigated whether near-seated peers influence students' academic engagement and achievement in elementary school settings. They found that students achieved better (worse) scores when near-seated friends scored better (worse).

Social learning theory states that people learn by observing and imitating others [175]. Based on this theory, students should learn by observing peers [175], and peer conditions affect student engagement and motivation. Gyanendra et al. [174] indicated that students preferred to choose seats with similar proximity to the multimedia screen, and the students who sat at a similar proximity to the front had similar distraction rates and performance levels [174], e.g. students with higher grades chose to sit in the front rows. These findings suggest a bidirectional relationship, i.e. performance (or motivations) connects peers, and peer conditions impact performance (or motivation). Although arguably, grades are correlated with seating location [174, 197, 194, 198], it was empirically discovered that study performance and engagement significantly improved when students sat closer [170, 28]. Netware et al. [197] also noted the correlation between seat location and in-class engagement, and suggested that teachers could optimise their teaching effectiveness by monitoring the progress of students sitting in different rows.

4.2.3 Inferring Student Engagement Using Sensing Technologies

Recently, physiological signals, e.g. EDA [30, 6, 118], PPG [109, 30], electroencephalogram (EEG) [205], eye gaze [206] and facial expressions [207], have been used to infer learners' affective states. Especially, there is a thread of research using EDA to indicate engagement levels [28, 6, 30, 45, 106, 111]. Hernandez et al. [106] measured the engagement of the child during interaction using physiological synchrony extracted from EDA sensors. Lascio et al. [6] measured university students' emotional engagement from EDA signals. Huynh et al. [45] developed EngageMon to use EDA together with HRV, touch and vision to indicate game engagement. Gao et al. [30] predicted student multi-dimensional engagement using EDA, PPG and HRV signals.

Physiological synchrony indicates the observed association or interdependence between the physiological processes of two or more people. These physiological signals often reflect connections between people's continuous measurements of the autonomic nervous system [76]. Across various streams of research, physiological synchrony has been shown to be informative for cognitive demands, task difficulties, learning engagement, etc. Therefore, understanding the physiological synchrony between people has attracted attention in the ubicomp community [7, 111, 208, 209]. Gashi et al. [7] proposed using physiological synchrony between students to measure the classroom emotional climate (CEC). They calculated the group physiological synchrony by applying the Dynamic Time Warping (DTW) distance to processed EDA signals and found that the group physiological synchrony of EDA was positively correlated to the CEC. Gashi et al. [28] studied the physiological synchrony of the inter-beat interval (IBI) and EDA signals between a presenter and the audience. They found that these signals could be used as a proxy to quantify participants' agreement on self-reported engagement during presentations. Malmberg et al. [208] found that physiological synchrony occurred within two groups of students who experienced difficulties in collaborative exams and concluded that physiological synchrony may be an indicator of the recognition of meaningful events in computer-supported collaborative learning.

In summary, unlike previous studies, our research has the following advantages: (1) We are the first to examine how student seating behaviour is related to physiologically measured

student engagement rather than simply using traditional measures (i.e. perceived student engagement) as previous studies have done [178, 174, 210, 199]. (2) We explore student seating behaviours using exact seating locations, which provides greater flexibility, whereas most studies only explore traditional seating arrangement styles (e.g. grouped tables or traditional row arrangement) [178, 174, 210]. (3) We explore the student engagement at the group level by utilizing distinct clusters of physiological signals, while previous researchers have studied student engagement at the individual level [6, 30]. (4) For the first time, we identify some statistically significant correlations between student seating behaviours and students' perceived and physiologically measured engagement.

4.3 Data Collection

We collected data from a field study in a high school over four weeks. Detailed information on the dataset can be found in Chapter 2. For simplicity, in this chapter, we will not introduce the data collection but only show the collected data that relates to this chapter, i.e. perceived student engagement, student seating location and physiological signals in class.

4.3.1 Student Multi-dimensional Engagement

We used a self-report survey to collect subjective assessments of student engagement. The self-report tool is the most commonly used method to measure student engagement because it can clearly reflect subjective perceptions, while other methods, such as interviews, teacher ratings and observations, are susceptible to external factors [91, 92]. The student engagement questionnaire included five items from the validated ISEQ. In the questionnaire, each item was rated on a 5-point Likert-scale from 'strongly disagree' to 'strongly agree'.

4.3.2 Seating Location

Classrooms for students in Year 10 are approximately 7.0 m \times 8.9 m in size and can accommodate up to 25 students. This school encourages flexible seating arrangements, therefore seating arrangements will often vary based on teacher needs, teaching style and course content. For

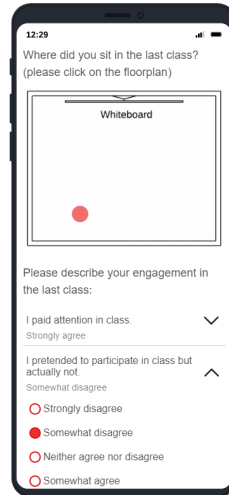


Figure 4.1: The screenshot of the self-report survey.

example, rows and columns are adopted when individual works are preferred, small groups are organised when more interaction is encouraged, semi-circular or u-shape seating arrangements are used when communication is emphasized. Before each class, students are free to choose where they want to sit in the classroom, leaving multiple vacant seats as some students may be absent for various reasons. During the data collection, the seating location of participants was measured by the self-report item ‘*Where did you sit in the last class? (Please click on the floor plan)*’. Participants were shown floor plan pictures ¹ and they could click different locations to report where they sat (see Figure 4.1). The seating location was recorded as the x and y -coordinate (in pixels) for each click, where $x = 0, y = 0$ represented the upper left corner of the floor plan. Compared to traditional methods of asking students about general locations (e.g. back/middle/front of the room [178], districts of multiple rows [174]) or the sequential number of rows and columns [197, 211], reporting the exact location in a classroom enables us to understand seating behaviours in real-world scenarios with flexible seating arrangements. Figure 4.2 shows the heat map of the seating locations from different classrooms in Year 10.

¹All classrooms in Year 10 have the same floor plan.

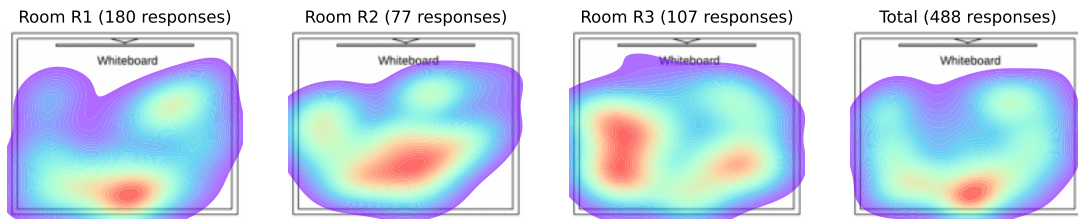


Figure 4.2: Overall seating distribution of students in the classrooms

4.3.3 Physiological Signals

We assess participants' EDA signals using the *Empatica E4* wristbands. The E4 wristband is equipped with multiple sensors designed to gather high-quality data and has an EDA sensor, PPG sensor, an ACC and an optical thermometer. EDA sensors record the constantly fluctuating changes in the electrical properties of the skin at 4 Hz. When the level of sweat increases, the conductivity of the skin increases. For most people, when they experience increased cognitive workload, emotional arousal or physical exertion, the brain will send innervating signals to the skin to increase the sweat production. Therefore, even though they may not feel any sweat on the skin surface, conductivity increases noticeably. Specifically, EDA complex includes two main components: a general tonic component to measure the skin conductance level (SCL) and rapid phasic component to measure the skin conductance response (SCR) resulting from sympathetic neuron activity [212]. The SCL measures the slow-acting and background characteristics of the EDA signal (overall level and slow decline or increase over time), reflecting the influence of autonomic arousal on the general sweat glands. The SCR is usually a sudden increase in the skin's conductance, which is usually associated with short-term events and external/internal stimuli.

4.4 Overview of Seating Experience, Student Engagement and Physiological Patterns

In this section, we explore the patterns of seating experience, student engagement and physiological signals. We divide the student groups based on seating locations and investigate the seating preference of participants over different courses. Then, we display the distribution of

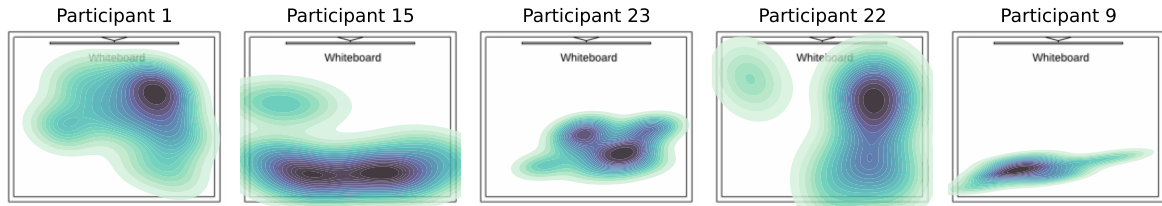


Figure 4.3: Seating location across five different student participants

student engagement and the number of collected wearable signals across all courses.

4.4.1 Seating Locations and Seating Preference

Figure 4.3 displays the seating locations of five different students participants: P1, P15, P23, P22 and P9. Due to space limitations, we have not shown the seating distribution of all participants. We find that different participants tended to have very different seating preferences. For instance, participant P22 typically sat close to the whiteboard, while participant P9 tended to sit at the back of the classroom. To better investigate the seating locations of students, we divide the students' seats into different groups. The simplest and most intuitive ways to do this is to partition the classroom equally from left to right or front to back. For example, sitting on the left-hand side of the classroom centre line is regarded as the left.

However, this simple partitioning method is not applicable in this research for two reasons: 1) The seats in the classroom were not evenly distributed from left to right. For example, the seats were sometimes arranged in two semicircles, with one semicircle positioned towards the center of the classroom and the other towards the side of the classroom. 2) The seats in the classroom were not evenly distributed from front to back. First, the whiteboard occupies a large area in the front of the classroom, which means dividing the classroom evenly from front to back may result in a greater concentration of chairs in the back half of the classroom. More importantly, the seats were not always arranged symmetrically like in traditional seating arrangements (e.g. row-and-column and grouped tables), and many times, they were distributed in a u-shape configuration according to the needs of teacher.

Therefore, we choose to group seating locations using the clustering technique. As one of the most popular clustering methods, we employ the *k-means* algorithm [213] to group similar



Figure 4.4: Three different seating locations (back, right and left) in classrooms

seating locations and discover underlying patterns. However, for the *k-means* algorithm to be effective, the number of the clusters must be predefined. There are two popular ways of determining the optimal number of clusters: the Elbow method and the Silhouette method [214]. After calculating the number of clusters using both methods, we identify that the optimal number of clusters was three. We then run the *k-means* algorithm using the *Scikitlearn* [215] Python package with *n_clusters* = 3 and *random_state* = 0. Figure 4.4 shows the clustering results of the seating locations from 488 responses. From these diagrams, it is clear that there are three different seating preferences in the classes: back, right and left. Therefore, in this research, we mainly focus on those three seating preferences.

An overview of the seating preferences of each participant is shown in Figure 4.5, and different colours indicate the frequency of sitting in each area. It shows that different participants tend to have very different seating preferences. For example, some participants (e.g. P10 and P21) usually sit in the back of the classroom, while some (e.g. P6 and P16) sit the left or right. Interestingly, some participants (e.g. P11 and P12) do not have obvious preferences for where they sit. Figure 4.6 shows how the seating preferences of the participants vary by course. We

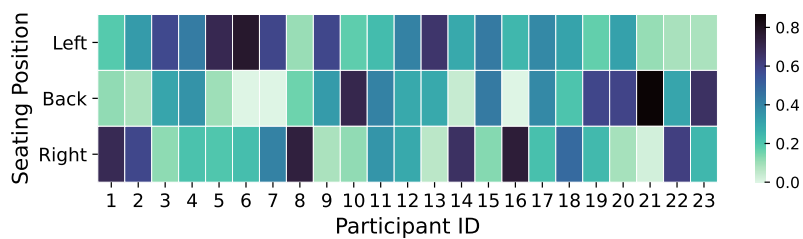


Figure 4.5: Seating preference for each participant in all courses

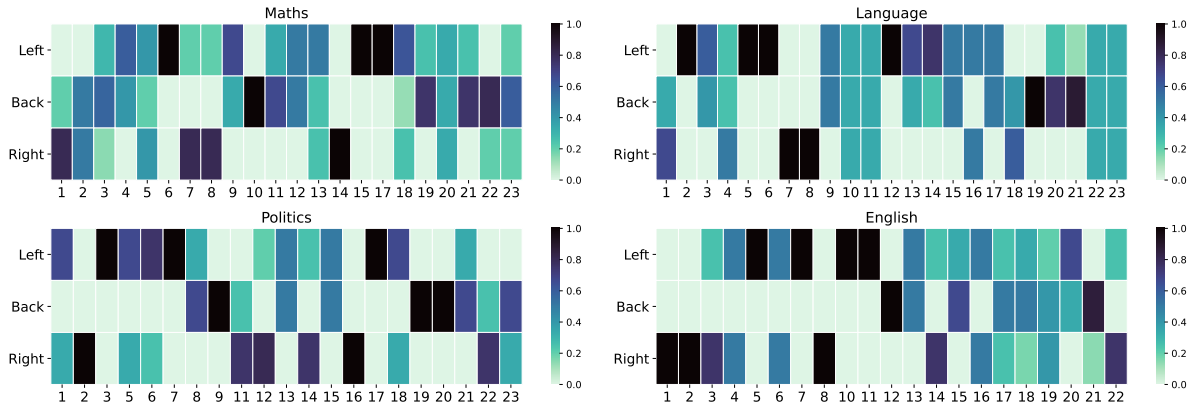


Figure 4.6: Seating preference for each participant in each course

find that participants often have different seating preferences in different courses, e.g. P3 never sits in the back in Politics or English but usually sits in the back in Maths. According to the education research, the potential reasons may be different student motivation and interests [172], territoriality and the desire to feel comfortable in different learning environments [216], and peer conditions within the classrooms [217].

4.4.2 Student Engagement and Physiological Signals

The distribution of overall engagement across student participants is shown in Figure 4.7a. The overall engagement scores are calculated based on the five items in the questionnaire, where 1 = lowest engagement and 5 = highest engagement. We can see that different participants usually have different engagement scores. Some participants (e.g. P5 and P9) tend to be highly engaged in class most time while some participants (e.g. P16) have varying levels of engagement in different classes. Based on the three seating preferences, the engagement score are calculated across different courses in Figure 4.7b.

Then, we calculate the number of wearable signals across all courses in Table 4.2. There are 10 different courses, and students are divided into three groups: Maths, Language and Form groups. ‘All’ indicates that all students are in one big group. Most of the classes are held in rooms $R1$, $R2$ and $R3$. There is an extra room, $R4$ for one language group, $R5$ is the room for Science, $R6$ is for Assembly and $R7$ is for Chapel. ‘Out’ indicates the playground, which is

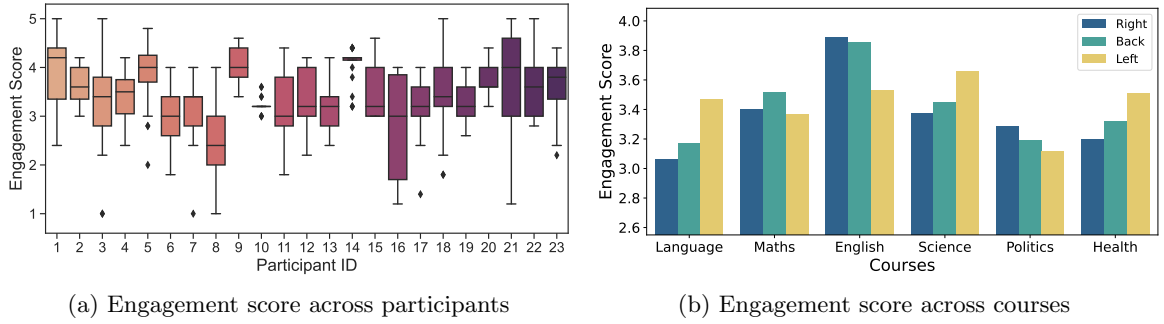


Figure 4.7: The overview of engagement score across participants and courses

Table 4.2: An overview of the number of wearable signals across all courses

<i>Course</i>	<i>Class type</i>	<i>Room</i>	<i>Number</i>	<i>Signal traces</i>
Math	Math groups	R1, R2, R3	36	197
Language	Language groups	R1, R2, R3, R4	44	181
English	Form groups	R1, R2, R3	31	172
Politics	Form groups	R1, R2, R3	37	190
Science	Form groups	R5	32	160
Health	Form groups	R1, R2, R3	18	64
PE	Form groups	Out	13	64
Form	Form groups	R1, R2, R3	6	29
Chapel	All	R7	2	28
Assembly	All	R6	2	35
Total	N/A	R1–R7, Out	221	1,120

used for the physical education course. For the wearable data, there are a total of 1,120 session logs of signal traces that can be used to explore the physiological patterns related to student engagement and seating preferences.

4.5 Results

In this section, we discuss the extensive experiments conducted to understand how individual and group-wise seating experiences affect student engagement. We will answer the first research question ‘*How does seating proximity between students relate to their perceived learning engagement?*’ in Section 4.5.1. We will answer the second research question ‘*How do students’ group seating behaviours relate to their physiologically measured engagement level (i.e. physiological arousal and synchrony)?*’ in Section 4.5.2 and Section 4.5.3. More specifically,

in Section 4.5.2, we investigate the correlation between the seating location and physiological synchrony. Meanwhile, we study the group-wise classroom seating patterns and demonstrate how they affect student engagement in different courses in Section 4.5.3.

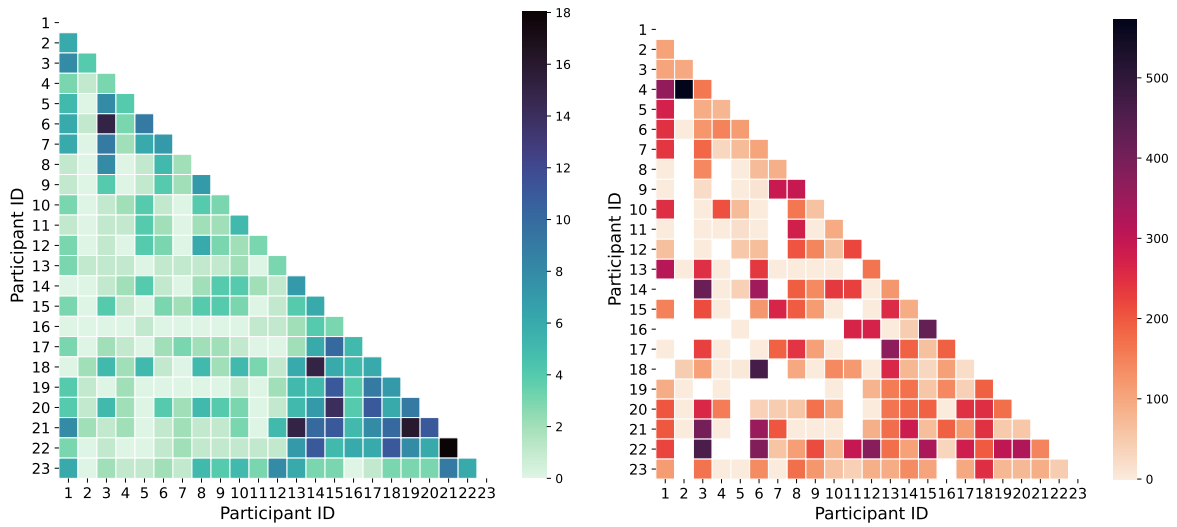
4.5.1 Relationship Between Seating Behaviours and Perceived Student Engagement

As discussed in Section 4.2.2, seating location has been found to be correlated with student engagement. Although some researchers [201, 178] revealed how the seating proximity between the student and teacher influences student engagement, satisfaction and course performance, few studies explored how the proximity of students is correlated with learning engagement. In this research, we define the following two terms: *proximity of students* and *similarity of engagement*. *Proximity of students* is calculated as the seating distance d_s between two students, using the Euclidean distance [218]. Euclidean distance is the most commonly used distance measure, which calculates the straight-line distance between two points on a plane and works well on low-dimensional data. Therefore, the formula to calculate the *proximity of student* is as follows:

$$d_s = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \quad (4.1)$$

In Equation 4.1, x_a and y_a indicate the position where student S_a sat, as marked manually by the student, and the upper left point on the figure has the x, y coordinates (0, 0). To measure the *similarity of engagement*, we computed the Manhattan distance [219] d_e between the engagement score of two students, S_a and S_b , where $d_e = |S_a - S_b|$, $S_a \in [1, 5]$, $S_b \in [1, 5]$. The Manhattan distance works well on discrete/binary variables, and it considers the path that can be realistically taken given the values of the attributes. Therefore, the *similarity of engagement* $\mathcal{E}(d_e)$ is defined as follows:

$$\mathcal{E}(d_e) = 1 - \frac{d_e}{4} = 1 - \frac{|S_a - S_b|}{4} \quad (4.2)$$



(a) The number of times pairs of students appeared in the same class (b) The average seating distance between pairs of students

Figure 4.8: Seating and occurrence information for pairs of students

In Equation 4.2, $\mathcal{E}(d_e)$ is in the range $[0, 1]$, where 1 means the participants' engagements are exactly the same, and 0 means the engagements are very different (i.e. one participant is completely engaged while the other one is not engaged at all).

We then analyse the self-report engagement responses and seating behaviours of participants from 92 out of 115 classes (23 classes with responses from only one participant are removed). After removing duplicated responses and keeping the last response from the same participants in each class, we got 1,123 unique pairs of instances (i.e. each instance had a unique combination of two participants' IDs and a class ID). The overall seating and occurrence information for each pair of students is shown in Figure 4.8. Figure 4.8a shows the number of times the pairs of students appeared in the same class, where a darker colour indicates that the pair usually sat in the same classroom (e.g. P21 and P13). Figure 4.8b shows average seating distance between pairs of students sitting in the same classroom, where a darker colour means that the two students usually sat far apart.

Figure 4.9 displays the distribution of the overall engagement and the seating distribution for four example classes. The points indicate the annotated seating locations in the classroom, as marked by different participants. A darker colour indicates a higher overall engagement

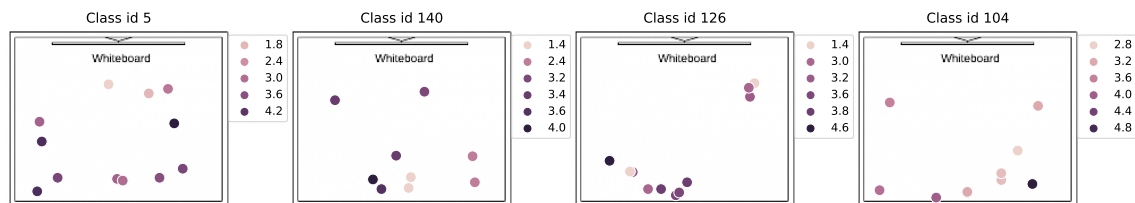


Figure 4.9: Overall engagement and seating distribution over four example classes

score, where 1 is the lowest score and 5 is the highest. In particular, in the sub-figure Class ID 140, we observe that points close together have similar colours, i.e. students who sat close together tended to have similar engagement scores in this class. It is worth noting that while engagement may change during a class, in this study we only asked students to report their overall engagement for the whole class and complete the survey immediately after the class. Typically, submission times are around scheduled times (i.e. 11:00, 13:25 and 15:35), although they vary slightly for each participant. Therefore, all responses submitted before the next class are considered to be feedback for the previous class.

Finally, we calculate the correlation between seating proximity and the similarity of engagement for all participants. The experiment result shows that the similarity of engagement is significantly negatively correlated with the seating distance ($\text{corr} = -0.30$ and $p < 0.05$). This indicates that students who sit close together are more likely to have a similar learning engagement than students who sit far apart. Interestingly, when considering each individual participant, the similarity of engagement is significantly negatively correlated with the seating distance for 18 out of 23 participants ($p < 0.05$), moderately negatively correlated with the seating distance for two participants ($p < 0.1$) and there is no obvious correlation with seating distance for three participants. The possible reason for this result may be due to individual differences, i.e. the engagement of some students may be affected by those around them, while other students may not be affected by those around them.

Notably, one potential threat to the above findings is that the absolute seating position on the whiteboard may affect student engagement [171, 174]. Therefore, we calculate the Pearson correlation between the above two variables, where $\text{corr} = 0.09$ ($p\text{-value} = 0.047$),

indicating that there is a negligible correlation between student engagement and absolute distance to the whiteboard. A potential reason may be that the seating arrangement is ‘u-shaped’ instead of the traditional ‘row-and-column’ most of the time, students sitting far from the whiteboard usually face the whiteboard, while students sitting closer to the whiteboard need to look sideways at the whiteboard. Therefore, the threat of absolute seating position affecting student engagement is minimal in this research.

4.5.2 Relationship Between Seating Behaviours and Physiological Synchrony

Physiological synchrony between individuals can be indicative of group engagement [76], e.g. the synchrony of an audience laughing at the same joke or students in a classroom concentrating on learning a complex math concept. Gashi et al. [28] found that engaged audiences exhibited higher levels of physiological synchrony with the presenter. This can be extended to students as the audience and the teacher as the presenter. The most effective teaching happens when there is a synchrony between the students and the teacher [7].

As discussed in Section 4.2.3, various physiological signals have been explored to infer learning engagement and affective states. From a physiological point of view, time synchronisation occurs when the physiological processes of two or more individuals are correlated with each other [76]. Similar simultaneous changes in people’s physiological signals (e.g. EDA) can provide information about cognition load or task difficulty [208, 220]. Recently, the physiological synchrony between individuals has been widely used in educational settings and has become an effective way to infer group or individual engagement in an activity. In particular, Gashi et al. [28] found that the engaged audiences exhibited higher levels of physiological synchrony with the presenter, which can be extended to students as the audience and the teacher as the presenter. It has also been found that the most effective teaching happens when there is a synchrony between the students and the teacher [7]. Therefore, in this research, physiological synchrony is used to represent the similarity in learning engagement, i.e. the higher the level of physiological synchrony, the greater the similarity in student engagement between individuals.

Data Pre-processing. The physiological synchrony is derived for each pair of students

Table 4.3: Summary of the Pearson rank correlation results between proximity of seating and physiological synchrony across the courses. The asterisks indicate the statistically significant results: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.004$

	All	Maths	Language	English	Politics	Science
<i>EDA_mixed</i>	-0.13*	-0.14	-0.32***	-0.02	-0.11	-0.01
<i>EDA_tonic</i>	-0.12*	-0.24*	-0.27**	-0.05	0.01	0.02
<i>EDA_phasic</i>	-0.02	0.09	-0.08	0.01	-0.21	-0.09

from the beginning until the end of each class. We then apply the following data preprocessing methods to the EDA signals to remove noises in the data (e.g. flat responses, movement artefacts and quantisation errors). First, we remove the incomplete data that was gathered throughout the class and discard signals containing many movement artefacts or flat responses. Second, similar to [7, 30], a median filter with a window of five seconds is applied to the EDA signals, which reduces the artefacts but preserves typical EDA edges. Third, we decompose the EDA signals into two parts: *tonic* and *phasic* [104, 128]. The *tonic* component changes slowly and reflects the general sweat level influenced by the body or environmental temperatures. The *phasic* component indicates rapid changes related to external stimuli. The EDA signals are decomposed using *cvxEDA* [129] with the convex optimisation methods. Finally, we normalise the original values of the EDA signals to the range [0, 1] to make the individual signals comparable.

Correlation Results. After the data preprocessing, the physiological synchrony is quantified using the students' normalised EDA signals. Based on prior research [106, 28, 208], we adopt one of the most popular methods to represent physiological synchrony *Pearson product-moment correlation coefficient*, which measures the linear dependence between two signals. P-values are tested against both the $\alpha = 0.05$ and the corrected threshold $\alpha_c = \frac{\alpha}{n} = 0.004$, where $n = 12$. The latter is known as Bonferroni correction [221], which is applied when n multiple statistical tests are performed simultaneously. We then analyse the relationship between the proximity of students and physiological synchrony using 368 pairs of instances, including the unique combination of two participants' IDs and a class ID. The number of physiological synchrony values (368 pairs) is much lower than the perceived student engagement (1,123 pairs)

for the following reasons: (1) Some students reported their perceived engagement but did not wear the E4 wristbands during the class. (2) Some students wore the E4 wristbands during the class, but the signals were not recorded for the entire class owing to the battery running flat or the wristband turning off accidentally. (3) The quality of some E4 signals was too low (e.g. too many flat responses), and these signals were removed during the preprocessing stage.

Then, we run the correlation analysis separately on the mixed EDA signals, tonic EDA signals and phasic EDA signals. Prior to the Bonferroni correction, there is a significant negative correlation between seating proximity and physiological synchrony (corr = -0.12 and $p = 0.03$). These results suggest that students sitting close together tended to experience higher physiological synchrony. However, after Bonferroni correction, this correlation is not significant. Therefore, it should not be considered as conclusive. One potential reason for this lack of significance is that we did not account for the impact of different courses during the correlation analysis.

Impact of Different Courses. Next, we explore the correlation results across the courses. We only focused on the main courses: *Maths*, *Language*, *English*, *Politics* and *Science*. The other courses, such as *Form*, *Chapel* and *Health* are not considered owing to the limited number of physiological signals. Table 4.3 summarises the results of the *Pearson* rank correlation between the seating distance and physiological synchrony. Before applying Bonferroni correction, seating proximity is only highly correlated with physiological synchrony in the *Language* class (EDA_mixed: corr = -0.32, $p = 0.002$, EDA_tonic: corr = -0.27, $p = 0.009$) and *Maths* class (EDA_tonic: corr = -0.24, $p = 0.03$), and no other significant correlation has been found. These results are interesting, and one possible reason for the results is that the physiological synchrony of students can be depicted more accurately in *Maths* or *Language* classes than in other classes, or students who sit close together are more likely to have physiological synchrony in *Maths* and *Language* classes.

In addition, we observe that statistically significant correlations are mainly reflected in the mixed EDA and the tonic EDA signals. No significant correlation has been found based on the phasic EDA signals. A possible reason is that the phasic EDA signals are related to fine-grained responses to internal and external stimuli, which are usually different among

Table 4.4: Description of the features computed for electrodermal activity signals.

<i>Feature name</i>	<i>Description of features</i>
eda/scl/scr_avg	Average value for the EDA, SCL, SCR
eda/scl/scr_std	Standard deviation for the EDA, SCL, SCR
eda/scl/scr_n_	Number of peaks for the EDA, SCL, SCR
eda/scl/scr_a_p	Mean of peak amplitude for the EDA, SCL, SCR
eda/scl/scr_auc	Area under the curve of the EDA, SCL, SCR
scr_frequency	The frequency of phasic increases in skin conductance
num_arouse	Number of arousing moments during the class
ratio_arouse	Ratio of arousing and unarousing moments
level _k	Ratio of the number of level _k and the length of S_k
eda/tonic/phasic_pcct	Pearson correlation coefficient with teacher
eda/tonic/phasic_pccs	Pearson correlation coefficient with average value of students

students. Despite the small variations, we assume that the calculated physiological synchrony reflects general changes in student engagement during a class. Therefore, it can be seen that calculating physiological synchrony based on the tonic and mixed EDA signal works well, which may be because tonic signals indicate general arousal and learning engagement in the whole class. Since the mixed signal includes both the tonic and phasic signals, it reflects general engagement and preserves some fine-grained information from the phasic component [28].

When Bonferroni correction is taken into account, the correlation between seating proximity and physiological synchrony is only significant when computed using mixed EDA signals in *Language* classes. The results in other courses only exhibit loose statistical significance after applying Bonferroni correction. Since we do not have a prior hypothesis for different courses, the significance before Bonferroni correction may suggest a hypothesis for future exploration.

4.5.3 Relationship Between Seating Behaviours and Physiological Arousal

As suggested in prior research [208, 76, 222], physiological arousal and synchrony are both regarded as effective ways to infer people’s mental activity and cognitive load. Physiological arousal is an activity of the sympathetic nervous system, and it can be measured using EDA signals. Measuring physiological arousal is a useful way to understand people’s emotional and cognitive processes. In general, increases in arousal are related to cognitive demand [223],

attention level [224] and learning engagement [30]. In this research, we extracted some widely used features of EDA signals from previous research [28, 30] to represent the general student engagement level.

Extracted Features. Table 4.4 lists the extracted features from EDA signals based on three categories [28], namely, general engagement, momentary engagement and synchrony: (1) General engagement reflects the overall changes in engagement during a class. In this research, we adopted the features proposed in the literature [28, 225, 30, 188], including the average, standard deviation, number of peaks, average value of peak amplitude and area under the curve, which are calculated from EDA, SCL and SCR signals. (2) Momentary engagement features identify evident increments in physiological arousal [226], including the SCL frequency, number of arousing moments, ratio of arousing and non-arousing moments and the ratio of the number of $level_k$ and S_k as suggested in [28]. (3) For each student, we computed the synchrony features, such as *Pearson* correlation coefficient with the teacher and the average synchrony of students [30].

Group-level Seating Experience. Unlike using physiological synchrony to identify similarities in student engagement, it is difficult to intuitively compare physiological arousal between individuals because of the numerous features that represent physiological arousal (see Table 4.4). Therefore, we consider applying the clustering technique to divide the students into groups, which helps us compare patterns in physiological arousal and understand group-level student behaviours. The groups are built based on the extracted features from all EDA session logs of signal traces.

In the clustering stage, the *k-means* algorithm [213] is adopted for clustering the features extracted from EDA signals. First, we normalise all extracted features to eliminate extreme values and ensure high-quality clusters are generated, which is an essential step because the default *Euclidean* distance metric is very sensitive to changes in the differences [227]. Second, we apply the *k-means* algorithm to the extracted features indicating physiological arousal and identify clusters of students. The *Elbow* and *Silhouette* methods [228] are used together to find the optimal numbers, k , and we find that $k = 3$. Next, we analyse the statistical characteristics (e.g. average value) of each clustered group and calculated students' self-reported responses in

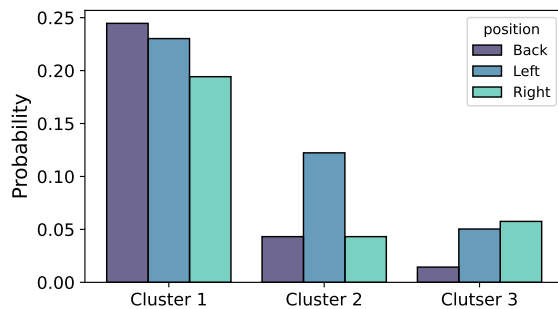


Figure 4.10: The distribution of seating preferences between student groups with different physiological arousal

terms of seating behaviours and learning engagement.

After the clustering, we analyse the distribution of seating preferences (i.e. back, left and right) among different clusters (see Figure 4.10). The clusters are divided based on the similarities in physiological arousal features, and the seating preferences are calculated based on self-report responses, as introduced in Section 4.4. To determine whether the seating preferences and clusters of physiological arousal are likely to be related or not, we adopt the *Pearson's chi-squared test of independence* [229] on above two variables, where $\chi^2 = 15.908$, degrees of freedom = 4, p-value = 0.003. Since the p-value is lower than 0.05, we reject the null hypothesis and reveal that the relation between the clusters of physiological arousal and seating

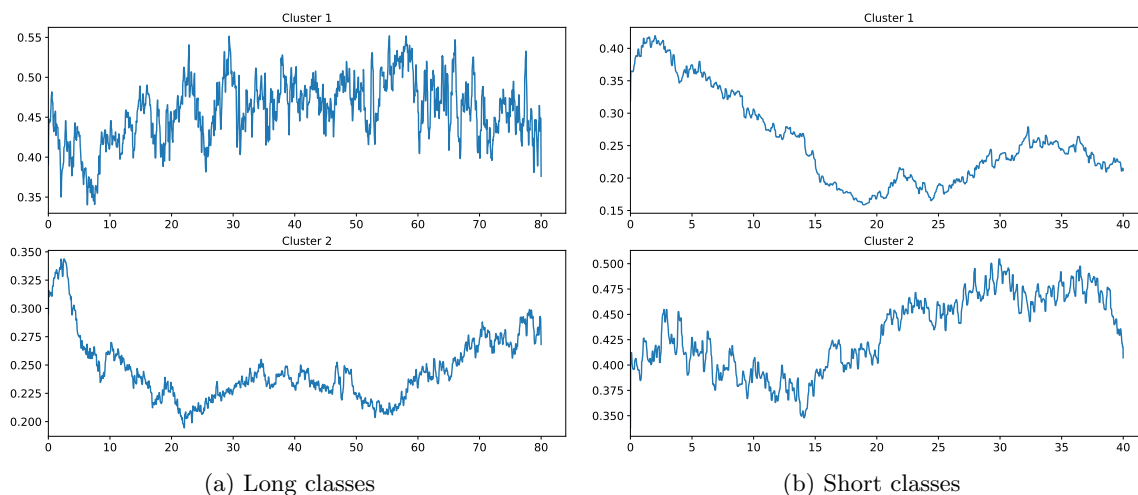


Figure 4.11: The clustering results of EDA signals for long and short classes

Table 4.5: The learning engagement of student groups with different courses and class lengths

	<i>Language</i>	<i>Math</i>	<i>English</i>	<i>Politics</i>	<i>Science</i>
<i>Cluster 1</i> (short)	3.18 (0.87)	3.34 (0.82)	3.81 (0.70)	3.12 (0.75)	N/A
<i>Cluster 2</i> (short)	3.41 (0.90)	3.73 (0.78)	3.78 (0.70)	3.32 (0.84)	N/A
<i>Cluster 1</i> (long)	3.53 (0.81)	3.46 (0.83)	3.72 (0.68)	3.53 (0.75)	3.58 (0.80)
<i>Cluster 2</i> (long)	3.70 (0.75)	3.45 (0.93)	3.93 (0.72)	3.41 (0.73)	3.61 (0.68)

preferences are significant. Figure 4.10 displays that the students in *Cluster 2* tend to sit on the left, while the students in *Cluster 3* seldom sit at the back of the classroom. The students in *Cluster 1* are almost equally likely to sit in any position, but they have a slightly greater tendency to sit at the back of the classroom.

Impact of Course and Class Length. Next, we explore how physiological arousal is related to perceived student engagement. Two factors may affect the clustering results: the length of the class and the course itself. Figure 4.11a and Figure 4.11b show the average value of the EDA signals in different clusters for long and short classes. First, we divide the EDA signals based on class length: long classes (80 min) and short classes (40 min). Then we identify that the optimal number of clustering is 2. For short classes, the statistical characteristics of self-reported overall engagement in *Cluster 1* are mean = 3.64, std = 0.82, and in *Cluster 2* are mean = 3.41, std = 0.73. Results from t-test [230] indicate that two clusters are significantly different in perceived emotional engagement, where t-statistic = -2.01, p-value = 0.04. From Figure 4.11b, we find that there are several peaks in the EDA signals in *Cluster 1*, indicating higher emotion arousal, as discussed in the literature [104]. The perceived engagement scores also support the clustering results of the EDA signals. However, for long classes, we could not find significant differences in any dimension of engagement in the two clusters (p-value > 0.05).

We then calculate the learning engagement of student groups in different courses. We do not consider courses with fewer than 100 signal traces and only focused on the courses with sufficient data (see Table 4.2), i.e. *Language*, *Maths*, *English*, *Politics* and *Science*. The student groups are then generated based on the clustering results of the physiological arousal features, and the optimal number of clusters is 2 for the above courses. Table 4.5 shows the learning engagement of the student groups in different courses and the length of classes. The learning

engagement in short classes for *Science* is *N/A* because all *Science* classes are the same length (i.e. long). Table 4.5 shows a significant difference in learning engagement in the two student groups for short classes in the *Maths* courses (average learning engagement of *Cluster 1* = 3.34 and *Cluster 2* = 3.73. Similar phenomena occurs in the short classes for *Language* courses and long classes for *English* courses. The potential reasons for not finding an obvious difference in the learning engagement in *Politics* and *Science* courses are as follows: (1) Students tend to have similar learning engagement in the above two courses than the other courses, making it difficult to determine to impact of the courses on the student groups. (2) There is a lack of physiological signals or low-quality physiological signals in these classes. Students need to do experiments in *Science* classes, resulting in a high chance of signal noise, such as poor contact between the sensors and the skin.

4.6 Limitations and Implications

Investigating how student seating experiences relate to learning engagement can help educators and policymakers improve student engagement in a variety of ways, such as providing better seating arrangements, organising more effective study groups and addressing problems such as poor academic achievement, student boredom and disaffection. Our study adds quantifiable evidence on the relationship between seating location and student engagement. The use of physiological arousal and physiological synchrony provides an effective method of understanding student engagement in real-time, with less burden on students than traditional questionnaires.

This research addresses possibilities to developing an intelligent seating recommendation system integrated with wearable devices, to optimise student engagement levels during class. This system could analyse student engagement in real-time and provide recommendations for the best or alternative seating locations in the classroom. On one hand, the system may learn to understand students' seating preferences by analysing their previous seating patterns and provide learning tips for students, e.g. if a student's engagement decreases as a result of their seating choices, the system could suggest changing to another seat or increasing participation in class activities. Teachers will be able to help manage students' progress more efficiently

by monitoring behavioral patterns in the classroom. On another hand, capturing student engagement and seating experience may also aid in curriculum design, resulting in improved teaching effectiveness and improved teaching style.

This study has some limitations that need to be overcome in future studies. First, although efforts are made to ensure that the modified five-item questionnaire captured the same constructs as the original questionnaire, validating the modified version was challenging due to the limited number of participants and budgetary constraints. It is also not feasible to adopt the traditional questionnaires (e.g., EvsD [115]) in education, as they often contain dozens of questions which would create a burden on students if they needed to be filled out after each class. However, the risk of a lack of validity of the questionnaire is minimal because we only slightly adapt it to suit the high school context, and the adapted questions are almost identical to those used in widely accepted questionnaires [115] in the educational context. Second, the study into group-wise experiences is limited to studying student peers and groups based on clustering. Future research could investigate different psychological patterns within and between student groups.

Third, the accuracy of the EDA signals measured by the E4 wristbands is limited. Menghini et al. [231] evaluated the accuracy of the E4 wristbands compared with the gold-standard sensors and found that similar accuracy could not be achieved using EDA signals from the wrists and fingers. A promising solution to improve the accuracy of the E4 may be lead wire extension, which would allow EDA recordings to be taken from the finger or palm rather than the wrist, thereby eliminating any potential site difference. However, in this study, the E4 wristbands are the best choice to obtain data without disturbing students in the classroom.

Furthermore, since participants were required to wear wristbands during school, sweat accumulation may have affected our results, especially given the hours of recording and the use of dry electrodes. In our data collection, it is challenging to control the factors related to the learning environment such as the humidity and temperature. However, the students' classrooms and teaching buildings were equipped with central air conditioners, so that the indoor temperature was not too high/too low, and the students did not sweat too much. After each class (40 minutes or 80 minutes), students would take breaks and walk, which is conducive

to the evaporation of sweat from their body surface. When analyzing EDA signals, the tonic signal changes slowly and reflects the general sweat level influenced by the body or environment temperatures, while the phasic component indicates the rapid changes related to the external stimuli. Despite the influence of sweat accumulation, the phasic EDA may still indicate student engagement.

Last but not least, in our research, some procedures for collecting and processing EDA data may not have strictly followed the best standards of EDA practice as suggested by Babaei et al. [188] (e.g. caffeine control, medication control, counter-balancing of external factors), thus threatening the validity our results to some extent. However, with our data collection spanning four weeks, it was not feasible to control for various external factors without burdening students and affecting their daily learning. In addition, due to the relatively low quality of the collected EDA signals compared to laboratory studies, it was challenging for us to apply standard settings to signal processing in the recommended way [188]. Therefore, we have followed some EDA signal processing methods that have proven effective in similar data collection environments [6, 30]. In the future, more community standards of EDA practices could be explored for in-situ studies to improve the validity of research.

4.7 Conclusion

In this chapter, we explored how student seating experiences were related to their emotional and behavioural engagement by understanding their physiological and behavioural patterns during a four-week data collection. The results showed that the individual and group-wise classroom seating experience is statistically correlated with both perceived and physiologically measured student engagement (physiological synchrony and physiological arousal). We found that students who sat close together were more likely to have similar learning engagement and higher physiological synchrony than students who sat far apart.

Chapter 5

Profiling Individual Personality Traits and Response Behaviours from Mobile Sensing Data

In Chapter 3 and 4, we used wearable sensing and environmental sensing data for human behaviour modelling. Capturing this data usually requires specific sensors to be installed or worn. In relation to *RQ-4*, this chapter explores using unobtrusive mobile sensing for user behavioural profiling. We will consider two real-world user behaviour prediction scenarios. The first scenario relates to modelling users' mental characteristics (i.e. Big Five personality traits). We will propose some important features that allow us to describe human activities based on mobile phone logs, call logs and accelerometer data and use these for the first time to predict human personality traits. The results reveal that the predicted personality scores were close to the ground truth, with an observable reduction in errors in predicting the Big Five personality traits in both male and female participants. The second scenario will investigate the effect of individuals' smartphone usage behaviour and mood on notification response time. We conduct an *in-the-wild* study with more than 18 participants for five weeks. The proposed regression model accurately predicts the notification response times using the users' current mood and physiological signals.

5.1 Introduction

Globally, smartphones are widely used and contain a wealth of sensors that can be used to easily collect large amounts of data relating to user behaviours (e.g. communication, location, media consumption and notification responses) in an unobtrusive and timely manner. These digital footprints derived from smartphones can reveal users' psychological characteristics, such as personality traits and emotions, which can help researchers, developers and managers better understand the interests and needs of their mobile users. Past research [232, 233, 120] has shown that it is possible to predict a person's personality through historical mobile usage data, such as calls, messages, app usage and location logs. To predict personality traits using mobile phones, researchers have mainly focused on exploring phone activities or app usage. However, to date, nobody has taken advantage of combining data on mobile usage behaviours with data on physical activity intensity from accelerometer sensors.

Accelerometers have been widely applied in various devices, such as mobile phones and fitness wristbands, to detect the intensity of human physical activity [234], which has been proved to be significantly correlated with personality [235]. In this chapter, we combine the physical activity intensity with phone usage behaviours to predict human personality traits. We propose several important metrics based on *diversity*, *dispersion* and *regularity*, which are defined in Section 5.3.1.2. Then we categorise these features based on different temporal factors and genders and apply support vector regression (SVR) to build a predictive model for human personality traits. The results of the experiment showed that using data relating to physical activity intensity based on accelerometer data can improve the predictive accuracy of the model. However, the improvement in predictive performance was different between males and females when the physical activity intensity was considered.

In addition to understanding users' psychological characteristics using smartphone sensing data, mobile computing can benefit users by providing intelligent interruption management. Smartphones frequently send users notifications, such as emails, messages, news and app update information. Inappropriate interruptions can have multiple effects on users, such as causing annoyance, increasing anxiety levels [236], decreasing productivity [237], negatively affecting task performance [238] or impacting emotional state [236]. For instance, Perlow et al. [239]

found that software engineers in a technology company had difficulties meeting deadlines due to frequent interruptions, which highlights the importance of interruption management to reduce distractions.

The notification is the prevailing mechanism on smartphones to convey timely and important information. They demand attention and can cause stress and frustration when delivered at inappropriate times. An interruption during a task can split the user's attention between two interactive tasks [240]. People need to decide whether the benefits of the interruption offset the loss of attention to the original task. Different actions can be taken to deal with interruptions, such as ignoring the interruption, postponing the required processing to a more convenient time or immediately resolving the interruption. These measures may delay resuming the original task and reduce task performance to varying degrees [241]. *Receptivity* refers to a user's reaction to an interruption, which may indicate the level of interruptibility of the user and their experience of the interruption [242]. In some cases, even though the notification is interruptive, the user can still be receptive to the notification. Previous studies have shown that users' receptivity to notifications is influenced by many factors: (1) informational qualities of the notifications, e.g. interest, entertainment, relevance and actionability [242]; (2) mobile usage, such as the time of the interruption and the type of app pushing the notification [243, 242]; (3) demographics, such as personality traits [244]; and (4) personal dynamics, such as location [245], transitions between activities [246] and social roles [247].

However, we propose a system to manage the automatic pop-up notifications of frequently used apps, which has not been attempted by other researchers before. Users' receptivity varies based on physical, psychological and affective conditions, and the accuracy of existing systems in addressing these conditions is still relatively low [248]. The difficulty of including these conditions can be explained by an example: Users may get annoyed (psychological) if an email from their boss suddenly pops up while they are concentrating on writing and are in a state of 'flow' (physical). However, it is not clear how the user would feel (affect) if this email notification were postponed. On the one hand, they may be relieved at not being disturbed, but on the other hand, it could cause stress if they were waiting for important information to help them with a problem they are experiencing.

Therefore, this chapter investigates the effect of individuals' emotions and smartphone usage behaviours on notification response times. We conduct an *in-the-wild* study with more than 18 participants for five weeks. The results of extensive experiments showed that the proposed regression model can accurately predict the response time to notifications using the user's current mood and physiological signals. We also investigate how physiological signals (collected from E4 wristbands) can be used as an indicator of mood and discuss individual differences in app usage and categories of smartphone apps and their impact on notification response times.

In summary, the contributions of this chapter are as follows:

- For the first time, we predict Big Five personality traits by combining physical activity intensity data with traditional phone activity data. Several novel metrics are proposed based on various categories: *diversity*, *dispersion* and *regularity*. We also identify significant associations between mobile phone usage behaviours and self-reported personality traits. We found that the features describing physical activity intensity from mobile accelerometer signals can improve the performance of personality prediction, with observable reductions in errors in both males and females.
- We conducted an in-situ study with 27 participants over a five-week period. In total, we collected 42,270 notifications with 3,236 ESM responses and more than 5,920 hours of physiological signals from Empatica E4 wristbands. We then explored diverse notification response behaviours of different participants and investigated the relationships between multiple factors (e.g. mood and apps) and notification response times. We found a statistically significant correlation between response time and in-use apps.
- We conducted extensive experiments to predict the notification response time for each participant. The experiment results showed that the proposed model achieved high predictive performance. We then derived the most useful features for each participant to achieve a meaningful and personalised prediction of notification response. In addition, we investigated how the mood-related features improved the predictive performance by utilising the ESM responses and physiological signals.

5.2 Related Work

5.2.1 Inferring Personality Traits through Mobile Sensing

Machine learning techniques have been applied successfully on sensor data to predict human mobility [249], identity [250], activities, transportation modes and complex behaviours [251]. Users' personality traits can be predicted using various media apps. Online social networking sites have been used to reflect user personalities, e.g. Facebook profiles [252] and Facebook messages [253]. Nhi et al. proposed a personality mining framework to exploit information from videos (e.g. YouTube clips), which include visual, auditory and textual perspectives [254]. Xin et al. demonstrated the relationships between active users' micro-blogging behaviours and personality traits [255]. Other works have shown that it is possible to predict the personality traits of users by exploring their mobile usage behaviours, which can be inferred from mobile data, such as call logs, app logs, Bluetooth logs and message logs [256, 233].

Cabrera-Quiros et al. [257] recognised self-assessed personality during crowd mingling scenarios using accelerometers and proximity sensors embedded in wearable devices alone. Although the physical activity of each person was considered, people were all required to wear the same wristband in specified scenarios, which is not representative of daily life. Recently, Weichen et al. [120] predicted personality traits through mobile sensings, such as ambient sound, ambient voice, physical activity and phone activity. However, they only computed the sedentary duration in each hour to represent the pattern of physical activity, which is simple and naïve, since they did not consider the entire physical activity intensity distribution. To estimate phone activity, they merely used the number of phone lock/unlock events and unlock duration but did not consider the diversity, regularity or dispersion of phone contact.

Mobile logs (phone and message activity) are easily accessible and have been used for efficient personality prediction [233]. To the best of our knowledge, no research has attempted to infer personality by combining traditional mobile activity with physical activity intensity, which has been proved to be strongly associated with personality [235]. Physical activity intensity can be estimated using data from accelerometers [234], which have been widely deployed in multiple devices, such as mobile phones and fitness wristbands.

5.2.2 Interruptibility Management and Receptivity

In this chapter, we define response time as the time that elapses between receiving a notification and opening the corresponding app. Okoshi et al. [258] presented a system to detect opportune moments for interruptions based on click rate gain using mobile sensing and ML methods. They calculated the users' click response times by measuring the time between a notification's arrival and the response to the notification, i.e. clicking on the notification. This data was logged along with contextual information from the smartphone and the data were evaluated. A trained linear regression model then identified whether a moment in time was an opportune moment to display a notification based on the extracted features. The adaptive notification component then delayed the presentation of notifications to the user until an opportune moment was detected. This breakpoint-based notification scheduling system resulted in increased click rates and quicker response times from users.

Saikia et al. [259] developed an optimisation process to reduce the reaction time and increase the opening rate of notifications for a mobile news app. Similar to Okoshi et al. [258], they defined the response time as the time between receiving and opening the notification and gathered additional contextual data, such as the category of the notification, time of the day and location. The notification opening rate, which is similar to the click rate [258], was used to optimise the opening rate and minimise the response time. Saikia et al. reduced reaction time by 13.3% and improved opening rates by 65.24%. Westermann et al. [260] studied the impact of the contextual factor time, i.e. the time of the day and the day of the week, on receptivity to notifications on Android smartphones. They sent advertisement notifications about popular brochures, and the response times were recorded as the time between receiving a notification and opening the app. The results exhibited notable variations in response times and notification-triggered app launch numbers on different days of the week and at different times of the day.

Fortin et al. [261] highlighted the correlation between SCR and the prediction of perceptions of smartphone notifications. To study the impact of user activity on the determined signals, the participants were asked to perform an inactive and active task during the measurement. They were directed to note the stimulus that caused them to perceive the notification and

press the corresponding button on a Pebble smartwatch placed next to them. The experiments showed that notifications perceived because of their tactile properties (vibration) stimulated greater SCRs with higher amplitudes than those perceived through auditory properties (sound). A logistic regression model was trained to examine whether a perception prediction method based on skin conductance could aid notifications, including the smartphone's ringer mode as a predictor variable. This model successfully identified perception in 75% of true cases when participants perceived the notification and 38% of missed notifications.

Mehrotra et al. [262] investigated the factors that make a smartphone notification disruptive and the impact of this on response time. An Android app called 'My phone and me' was created. The app uses Android's Notification Listener Service to access notifications and Google's Activity Recognition API and ESSensorManager to receive context information. The app also triggered questionnaires every four hours between 8 am and 8 pm. Reaction time was considered the time from the notification arrival to the time it was acted upon. The modes of identifying the notification (i.e. ringer or vibration) and the user's personality traits were also noted. The results showed that users responded to high-priority notifications much faster than other notifications, and those from less frequently contacted contacts were responded to the slowest. Notifications were considered most disruptive when the user was performing a task and least disruptive before they started a new task or while they were idle.

Zueger et al. [263] predicted the interruptibility of 13 software developers based on computer interaction and heart-, sleep- and physical activity-related data. They found that the interaction with a computer gave more information about interruptibility than the biometric data. However, using both types of data produced the best results.

5.2.3 Mood Sensing Approaches

Because the term mood is frequently used in this chapter, it is important to define it. Mood is a diffuse affective state that describes an individual's subjective feeling over time. Unlike emotions, mood lasts for hours or days, and its intensity is usually low. Most often, it is not possible to assign a specific trigger to or reason for our mood. Nonetheless, mood influences our behaviours and experiences [264].

Changes in activities, moods and behaviours of users provide valuable insights for providing context-aware services and minimising unwanted interruptions. According to recent research in psychology, the frequency of change or the rate of instability in various characteristics can affect the interruptibility of users [265]. In the field of attention management, various consequences have been investigated, including the influence of interruptions on mood.

For example, Zijlstra [266] showed that interruptions cause negative emotions. However, mood is also an internal stimulus that results from our insights and influences our interruptibility [267, 268]. Therefore, emotions and stress are not only consequences of interruptions but also influence our interruptibility. Yuan et al. [244] proposed using personality traits to group similar users in addition to considering contextual information, such as location, changes in the state of the user, time, transition state, current activity and mood to predict reactions to interruptions and the level of interruptibility.

Khan et al. [269] proposed a new approach for automated mood recognition (AMR) in the smart office environment, which reduces computational requirements by requiring fewer mood models. This was done by clustering physiological signals by groups of people who sense emotions in the same way. They used ML models for classification and regression, which were trained based on the extracted features of users in common perception clusters by recognising mood. Eight different categories of moods were recognised, each with three different levels denoting low, medium and high intensities. The proposed approach seems to be a trade-off between the requirement for a large number of personalised mood models and the insufficient performance of generalised models for AMR. The results showed average F1 scores of 0.76 and 0.79 for perception clusters and personalised AMR, respectively.

Current approaches in the field of attention management concentrate on notifications and their impact on human behaviour and wellbeing. It is known that receiving notifications can negatively impact our mood and trigger stress. It has also been shown that the reverse is true, i.e. our mood influences our behaviour towards notifications and our interruptibility. In this study, we want to go one step further and consider the effects of our mood on notification response times. For this purpose, we extend the current state of research by adding physiological signals to moods captured via ESM. We aim to identify whether mood directly affects response

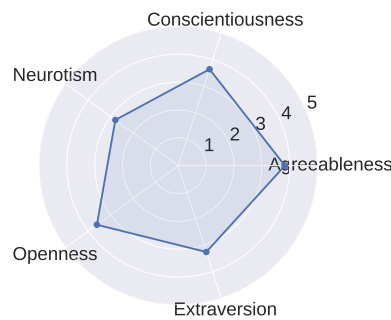


Figure 5.1: Average Big-5 personality scores

time. Using individual regression models, we aim to predict the receptivity of each user.

5.3 Inferring Personality Traits with Mobile Computing

5.3.1 Methodology

5.3.1.1 Participants and Procedure

In this research, we exploited a dataset including 55 participants living in a residential community, consisting primarily of young families, adjacent to a major research university in North America between March 2010 and July 2011 [270]. Each participant was equipped with a mobile phone running the Android OS and the sensing software Funf [270], which is designed to periodically collect mobile data. The software operates in a passive way and does not affect users' normal usage of the mobile phone. At the initial stage of data collection, each participant completed a personality survey, and Big Five scores were calculated [271].

The Big Five personality framework is one of the most important measures of personality traits [271, 272] and consists of five dimensions: extraversion, agreeableness, conscientiousness, neuroticism and openness to experience (openness). Extraversion reflects the degree to which a person is energetic, sociable and talkative. Openness represents the tendency to be curious and inventive. Agreeableness usually relates to the potential to be friendly and compassionate rather than suspicious and hostile to others. Conscientiousness represents the tendency to be organised, efficient and careful. Neuroticism is the tendency to be nervous and sensitive rather

than confident and secure. Figure 5.1 shows the average scores for the five different personality traits in our dataset, where 1 is the lowest score and 5 is the highest score.

After removing participants who did not fully complete the Big Five survey, our final sample comprised 52 participants (27 female and 25 male). For this study, we mainly focused on users' activity data, including phone activity and physical activity. Phone activity data, such as calls and text messages received or sent, have been widely used in personality prediction [233]. Physical activity data inferred from accelerometers has been proven to have a strong association with personality [235]. Therefore, in this research, we limited the scope of the study to the participants' call logs, text message logs and accelerometer logs, which are easily accessible for future mobile data collection.

For accelerometer logs, raw three-axis measurements were sampled at rate of 5 Hz for 15 sec every 2 min. Participants were not constrained in the way they could carry the phone. For call logs and message logs, the human-readable texts were captured as hashed identifiers. For more details about the dataset, please see [270].

5.3.1.2 Activity Behaviour Metrics

Personality can be evaluated using the Big Five model, which consists of five major dimensions of personality traits: Openness, Extraversion, Agreeableness, Conscientiousness and Neuroticism. To better understand the daily patterns of human activity, we computed several metrics to meaningfully distinguish between personality traits. The metrics are divided into three categories: dispersion, diversity and regularity. We used these metrics to evaluate the participants' phone activity and physical activity. Phone activities include calls and message interactions, which were computed separately based on the metrics. For physical activity, we first partitioned the raw accelerometer data into 24-hour periods and processed it in hourly increments. We then used the hourly mean amplitude deviation (MAD) [273, 274] to assess the intensity of physical activity.

$$\text{MAD} = \frac{1}{n} \sum |r_i - \bar{r}|, \quad (5.1)$$

where n is the number of accelerometer data samples in each time period, r_i is the resultant

acceleration at the i th time stamp and \bar{r} represents the mean resultant value for the time period. r_i can be calculated using the following formula:

$$r_i = \sqrt{x_i^2 + y_i^2 + z_i^2}, \quad (5.2)$$

where x_i , y_i and z_i represent the x, y and z directions of the raw acceleration signal. Next, we computed metrics to assess activity behaviours as follows.

Dispersion of activity behaviours depicts how sporadic a behaviour is. In our research, *SD* was used to evaluate the dispersion of phone activity and physical activity intensity. As people tend to have different activity patterns at different times (i.e. more physical activity over weekends and fewer phone calls at night), we computed the SD separately for three time periods (i.e. day, evening and night) on weekdays and weekends during the data collection period.

Diversity of activity behaviours refers to the level of diversity of users' activities. *Shannon entropy* measures the amount of disorder in a system, which can be used to measure the diversity of users' contacts using the following equation:

$$S = - \sum_{i=1}^n \mathcal{F}_i \log \mathcal{F}_i, \quad (5.3)$$

where \mathcal{F}_i means the frequency with which a user s interacts with i of all contacts n . Higher entropy means that the user s interacts equally with a lot of contacts, and lower entropy happens when the user mostly interacts with a few specific contacts. Shannon entropy was used to evaluate the diversity of phone activity in this study.

Regularity of activity behaviours refers to regular patterns of activity. We used the *regularity index* (RI) [120] to calculate the difference between the time periods T on two different days. First, we rescaled the data for each participant to fit the range $[-1,1]$, where -1 corresponds to the minimum value in the original data and 1 corresponds to the maximum value. The regularity index is positive if the values are close together and negative if they are not. We then defined the RI of the time period t between day i and day j as follows:

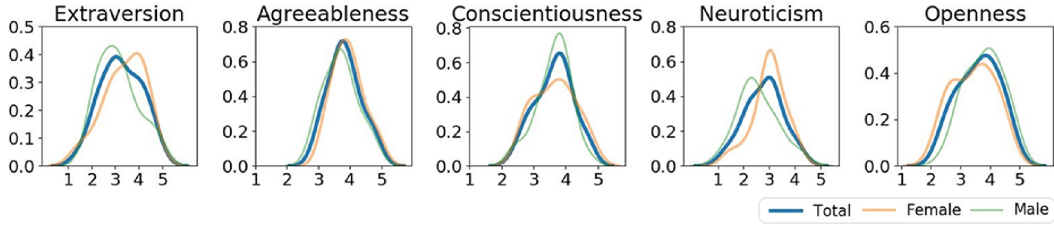


Figure 5.2: Kernel density distribution of five personality traits in the data set

$$\forall (i, j) \in S, \text{RI}_{i,j}^T = \frac{1}{T} \sum_{t=1}^T x_t^i x_t^j, \quad (5.4)$$

where S is set to the two time period pairs and x_t^i and x_t^j refer to the rescaled values at hour t in the time period T . We computed the average RI values from every possible pair in the following sets: (1) all days, (2) weekdays, (3) weekends, (4) weekday days, (5) weekday evenings, (6) weekday nights, (7) weekend days, (8) weekend evenings and (9) weekend nights. The RI was used to evaluate the regularity of phone activity and physical activity of the study participants.

To prove the significance of the advantages of the extracted physical activity features and to aid in comparison with prior research, we obtained traditional phone activity features based on the prior literature [233], including average inter-event time, variance in inter-event time, response rate, response latency, percentage during the night and percentage initiated. Table 5.2 summarises the features used in our study.

5.3.1.3 Big Five Personality Ground Truth

We used the self-reported Big Five results from the participants as the ground truth for various personality traits. The scores were computed based on 52 questions related to personality traits [275], with scores in the range 1–5, where 1 is the lowest score and 5 is the highest score for the personality trait. Figure 5.2 shows the distribution of the five personality traits by gender.

The descriptive statistics (i.e. mean, SD, median, minimum and maximum) are shown in Table 1 for the entire population and by gender. For the entire population, the average score for all personality traits was close to 3. The average score for agreeableness was approx-

Table 5.1: Overview of the Big Five scores for total/male/female participants

Gender	Personality Traits	Mean	Std. Dev.	Median	Min	Max
Total	Extraversion	3.26	0.86	3.13	1.50	4.88
	Agreeableness	3.83	0.52	3.78	2.78	5.00
	Conscientiousness	3.64	0.58	3.78	2.44	4.67
	neuroticism	2.79	0.74	2.88	1.13	4.25
	Openness	3.61	0.70	3.70	2.20	4.90
Female	Extraversion	3.38	0.87	3.63	1.50	4.63
	Agreeableness	3.95	0.50	3.89	3.11	5.00
	Conscientiousness	3.65	0.64	3.67	2.67	4.67
	neuroticism	3.00	0.65	3.00	1.38	4.13
	Openness	3.44	0.72	3.50	2.20	4.60
Male	Extraversion	3.13	0.84	3.00	2.00	4.88
	Agreeableness	3.71	0.54	3.67	2.78	4.78
	Conscientiousness	3.62	0.53	3.78	2.44	4.67
	neuroticism	2.56	0.77	2.38	1.13	4.25
	Openness	3.80	0.64	3.90	2.50	4.90

imately 4, which was the highest score, followed by conscientiousness, openness, extraversion and neuroticism in descending order. Agreeableness had the lowest SD, which means that the agreeableness scores for all participants were close to the mean.

Interestingly, we found that female and male participants had different distribution patterns in the five personality traits. Females scored higher on neuroticism than males (t-test p-value = 0.03), which suggests that the females in this population sample were more sensitive and emotional than the males. In addition, the males scored higher on openness than females, which suggests that the males in this population sample were more likely to be curious, and the females were more likely to be cautious.

5.3.2 Feature Analysis

We extracted features based on the introduced metrics and time spans in Section 5.3.1.2 (see Table 5.2). We defined the daytime period as 9:00 am to 6:00 pm, the evening period as 6:00 pm to 12:00 am and the night period as 12:00 am to 9:00 am.

Since most features, except for the entropy metrics, were strongly positively skewed, we applied log transformation before conducting the correlation analysis. The Pearson correlation

Table 5.2: Description of the extracted features

Feature	Features computed	Data
<i>Dispersion</i>	STD on number of interactions for all days	call, message, c&m
	STD on physical activity intensity for all days (daytime, evening, night)	accelerometer data
	STD on physical activity intensity for all weekdays (daytime, evening, night)	accelerometer data
	STD on physical activity intensity for all weekends (daytime, evening, night)	accelerometer data
	STD on physical activity magnitude for all days	accelerometer data
<i>Diversity</i>	Entropy of total contacts for all days	call, message, c&m
	Entropy of total contacts for weekdays	call, message, c&m
	Entropy of contacts in sent box for all days	call, message, c&m
	Entropy of contacts in sent box for weekdays	call, message, c&m
<i>Regularity</i>	Average RI of number of interactions for all days	call, message, c&m
	Average RI of physical activity intensity	accelerometer data
	Variance of RI for number of interactions (daytime, evening, night)	call, message, c&m
	Variance of RI for physical activity intensity (daytime, evening, night)	accelerometer data
<i>Basic</i>	Total number of interactions for all days	call, message, c&m
	Total number of interactions for weekdays	call, message, c&m
	Average physical activity intensity for all days (daytime, evening, night)	accelerometer data
	Average physical activity intensity for weekdays (daytime, evening, night)	accelerometer data
	Average physical activity intensity for weekends (daytime, evening, night)	accelerometer data

coefficient, which is widely applied to measure the correlation between variables in the field of psychology, was calculated between the extracted activity features and Big Five personality traits. The value of the Pearson correlation coefficient is in the range $[-1, 1]$, where 1 represents an exact positive linear correlation, 0 means no linear correlation and -1 indicates an exact negative linear correlation. Table 5.3 shows the three features that were identified as the most useful predictors of the Big Five personality scores for all, female and male participants, where (+) represents a positive correlation and (-) represents a negative correlation. Table 3 also shows the Pearson correlation coefficient for each useful feature. A discussion of these features follows.

Extraversion. The regularity index of physical activity intensity on weekday evenings was negatively correlated with the extraversion trait, which suggests that people with a higher

Table 5.3: Most useful features to predict personality traits (total population)

Personality	Top-3 Features
Extraversion	(−0.30) RI of physical activity intensity on weekday evenings (+0.26) Entropy of contacts (call & messages) (−0.23) STD of physical activity intensity on weekday daytime
Agreeableness	(−0.33) RI of physical activity intensity on weekday evenings (+0.26) Average physical activity intensity on weekends (+0.23) Average physical activity intensity on weekday evenings
Conscientiousness	(+0.44) Entropy of call & messages (+0.27) Total number of messages (+0.18) Average physical activity intensity on weekday evenings
Neuroticism	(+0.27) Entropy of contacts (calls) (−0.24) STD of physical activity intensity on weekend daytime (−0.21) Average physical activity intensity on all days
Openness	(−0.32) Total number of calls (+0.21) Average Inter-event time of calls (+0.15) STD of physical activity intensity on weekday daytime

extraversion score do not follow similar patterns on weekday nights. The high entropy of contacts means that they tend to interact with different people randomly, which is in accordance with our experience in daily life.

Agreeableness. Similar to extraversion, people who scored high on agreeableness tended to have a low regularity index for physical activity on weekday evenings, as they may be socialising. They also tended to have greater physical activity intensity on weekends and weekday evenings than those with a low agreeableness score. It is highly likely that a friendly and compassionate female will have more outgoing calls than one who is less friendly and compassionate.

Conscientiousness. We found that females and males with a high conscientiousness score tended to have high entropy of contacts, which suggests that people who are more organised and efficient tend to contact different people and do not usually connect with the same people. In addition, organised people may have a high activity intensity on weekend evenings because they plan their activities in advanced and are well prepared.

Neuroticism. We found that the regularity index of physical activity intensity on weekday and weekend nights for females was positively correlated with neuroticism, which suggests that females who are sensitive tend to participate in regular physical activity at night (after 12:00

Table 5.4: Most useful features to predict personality traits (female and male population)

Personality	Gender	Top-3 Features
Extraversion	Female	(-) Average physical activity intensity on weekend evening (-) Average physical activity intensity on all days (-) Average inter-event time (messages)
	Male	(+) Average physical activity intensity on weekend evenings (-) Average inter-event time (messages) (-) RI of physical activity intensity on weekday evenings
Agreeableness	Female	(+) Number of outgoing calls (-) RI of physical activity intensity on all days (-) RI of physical activity intensity on weekday nights
	Male	(-) RI of physical activity intensity on weekdays (-) RI of physical activity intensity on all days (+) Total number of incoming calls
Conscientiousness	Female	(+) RI of physical activity intensity on weekend daytime (+) Entropy of contacts (calls) (+) Average physical activity intensity on weekend evenings
	Male	(+) Entropy of call & messages (+) RI of physical activity intensity on weekend daytime (-) STD of physical activity intensity on weekday daytime
Neuroticism	Female	(+) Entropy of contacts (calls) (+) RI of physical activity intensity on weekend nights (+) RI of physical activity intensity on weekday nights
	Male	(+) RI of physical activity intensity on weekend nights (+) Entropy of call & messages (+) RI of physical activity intensity on weekday nights
Openness	Female	(-) Total number of calls (-) RI of physical activity intensity on weekday evenings (-) Average physical activity intensity on weekday nights
	Male	(-) Total number of calls (-) RI of physical activity intensity on weekday evenings (+) STD of physical activity intensity on weekday evenings

am). Interestingly, in males, these same features were negatively correlated with neuroticism, which suggests a difference between men and women.

Openness. We found the total number of calls was negatively correlated with the openness trait. In addition, the average inter-event time for calls was positively correlated with the openness score, i.e. individuals who have fewer phone calls and longer periods between each call tend to be more inventive and curious.

5.3.3 Predictive Analysis

Personality prediction is commonly regarded as a regression problem, and the scores range from 1 (lowest) to 5 (highest) for each personality trait. Although the personality score can be divided into several classes (e.g. high, medium and low) using certain thresholds, researchers have shown that this is not good practice for determining people’s psychological characteristics. Most classification models exhibit a low prediction accuracy of around 49%–63% [233]. Thus, in this study, we used the regression model to predict personality traits.

The SVR method with a radial basis function kernel was chosen to predict the Big Five personality scores. The SVR method has been applied in various fields and can handle high-dimensional data and automatically model non-linear relationships. Since no noticeable dissimilarities existed in the personality scores between the genders, and the key features were different, we conducted the prediction by choosing the best regressors for the entire population and for the males and females separately.

Baseline and Evaluation. Through the literature review, we found that most researchers used the random chance or majority class selection method as the baseline for Big Five personality prediction [120, 233]. However, in our research, we aim to improve prediction performance by combining human physical activity features and traditional phone features. Therefore, it did not make sense to compare our model with the random chance or majority class selection, as it is difficult to predict personality traits from only one type of data. Personality prediction using only phone activity data (call logs and message logs) and state-of-the-art metrics (introduced in Section 3.2) was considered as the baseline model in the experiment.

For evaluation, we adopted the leave-one-out validation method because it usually has the best performance when estimating a model from a small dataset. Using the leave-one-out method, we calculated the average value for the MAE and mean squared error (MSE) for each personality trait.

We validated our model with the *MAE* and *MSE*.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_{true} - y_{pred}| \quad (5.5)$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_{true} - y_{pred})^2 \quad (5.6)$$

where n represents the number of samples, y_{true} is the true personality score and y_{pred} is the predictive value of the personality score. The MAE and MSE can describe the accuracy of the predictions when compared with the ground truth of the personality score. The closer the MAE and MSE are to 0, the more successful the model is at forecasting the personality traits.

Discussion. Table 5.5 displays the performance of our predictive model based on the extracted features from call logs, message logs and raw accelerometer logs. With the observable reduction in errors, our model has a better performance than the baseline model for all personality traits. The predicted Big Five scores were highly correlated with the ground truth.

Based on the comparison of the MAE and MSE between our model and the baseline model, it is interesting to note that conscientiousness, neuroticism and extraversion were the personality traits that were predicted best by our predictive model. For the entire population, in predicting conscientiousness, the model achieved MAE = 0.249, which is 0.148 (37.28%) lower than the baseline model. For females, in predicting neuroticism, the model achieved MSE = 0.425, which is 0.129 (23.29%) lower than the baseline model, and the MSE for extraversion in the female group was 0.128 (17.56%) lower than the baseline.

The predictive performance of the model for neuroticism was better in the gender-specific model than in the population model, which may be due to different key features for males and females. According to the explanations in Section 4.1, males and females with high neuroticism scores may exhibit very different regularity of activity intensity at night. However, if we do not consider the gender difference, the regularity of activity intensity was not a key feature in the entire population. This phenomenon addresses the importance of building the gender-specific predictive models for the neuroticism personality trait.

Our model was less effective at predicting the openness trait, which may result from physical activity intensity not being strongly associated with the openness trait. In other words, in daily life, it is also difficult to tell if someone is inventive or curious by considering their patterns of activity intensity.

The current research had some limitations that need to be addressed in the future. First, the sample size of our adopted dataset ($n = 52$) is relatively small, which may limit the performance

Table 5.5: Prediction performance for total/male/female participants

Group	Big-5 Traits	MAE		MSE	
		Baseline	Proposed Method	Baseline	Proposed Method
Total	Extraversion	0.685	0.655	0.730	0.692
	Agreeableness	0.444	0.399	0.298	0.262
	Conscientiousness	0.397	0.249	0.249	0.240
	Neuroticism	0.620	0.617	0.565	0.551
	Openness	0.623	0.621	0.519	0.518
Female	Extraversion	0.624	0.576	0.734	0.605
	Agreeableness	0.395	0.387	0.261	0.243
	Conscientiousness	0.561	0.492	0.415	0.334
	Neuroticism	0.612	0.532	0.554	0.425
	Openness	0.709	0.709	0.625	0.625
Male	Extraversion	0.691	0.661	0.736	0.734
	Agreeableness	0.424	0.419	0.270	0.268
	Conscientiousness	0.407	0.393	0.293	0.275
	Neuroticism	0.571	0.525	0.536	0.463
	Openness	0.521	0.521	0.400	0.400

of the personality prediction. Further research is needed to explore a larger dataset to prove the effectiveness of the physical activity features for personality prediction. Second, the evaluation method used was relatively simple, and a more comprehensive evaluation method is needed to allow for better comparison with the extant literature. Finally, the existence of bias in the Big Five self-report data, such as sampling bias or response bias (e.g. misunderstanding the measurement, social desirability or the need to ‘look good’ in the survey) may affect predictive performance. Further work is needed to recognise and mitigate such bias.

5.4 Inferring Response Behaviours with Mobile Computing

5.4.1 Data Collection

5.4.1.1 Overview

We performed an *in-the-wild* study to gather user behaviours relating to smartphone notification arrival and response time combined with contextual information and mood via smartphones, desktop computers and physiological signals. By advertising our study on our websites and networks, we acquired 27 participants for the field study. The data collection began at the end of January 2020 and continued for five weeks. The participants were asked, if possible, to

install the apps *Balance for Android* and *Balance for Desktop* on their smartphones and desktop computers, respectively. Both apps facilitate continuous background sensing and ESMs [276]. In addition, the participants were free to choose whether to have their physiological signals measured via an E4 wristband. This part of the measurement was coordinated and supervised by a contact person in the participant's country of origin.

By installing the apps or wearing the E4 wristband, the participants were eligible to receive information about the study and the data collected. Participants were made aware of the privacy protection measures and their rights (e.g. the right to request to have their data erased). Our privacy department and ethics committee approved the consent forms and data collection procedures. Before the participants were given a short tutorial on using the apps and handling the E4 wristbands, they were required to give their informed consent. The study design called for contextual information to be recorded in the background without the participants' input, such as running apps, logging physical activities or recording locations.

As part of our mixed-method approach, participants were presented with questionnaires every 90 min. We asked them about their mood, social role, interruptibility and the type of task they were working on over the preceding 15 min. In addition, we implemented an event-based approach to present users with the questionnaire, which was activated when the participant had interacted with their phone for more than 10 min. These scheduled questionnaires were limited to the time period 7 am to 10 pm, and there was a minimum 30 min between questionnaires. In addition, a questionnaire was not sent if a participant still had a pending questionnaire. Through these restrictions, we addressed the strain of responding to questionnaires to ensure the quality of the data [277, 278]. All the approaches used are well known in ESM-based studies to capture contextual information *in-situ* [277].

5.4.1.2 Participants

In our experiments, we focus on response time regarding smartphone notifications. Therefore, we used data from 18 out of the 27 participants. The remaining 9 were removed because there was insufficient data from them on the ESM questionnaires or technical problems that affected the data collected from them. Our participants were between 25 and 41 years old.

There were 18 Android, 11 Windows and 7 macOS users, and 15 participants installed both the smartphone and the desktop app. The data were regularly transmitted to a server hosted at our university and stored in an internal database. The upload and the data were encrypted. The overall response rate (26.20%) was comparable to similar ESM-based studies in the field of interruptibility [279]. A total of 3,236 out of 12,352 questionnaires were answered.

5.4.1.3 Collected data

Balance for Android. The balance for Android regularly uploaded encrypted recorded data to the university server. The main focus of the design was low battery consumption, minimal resource consumption and the seamless recording of data in the background. Using background services, we kept track of interactions with apps and notifications, location updates and the phone's state (e.g. screen status and ringer modes). The phone's last known location was processed using a fused location provider¹, which is an API that estimates location information and manages Wi-Fi, mobile communication services and GPS while improving battery performance and resource consumption. In addition, we gathered information on physical activities by using the Google recognition API². This API offers to report recognised physical activities and optimise battery performance. The optimised battery performance is achieved by reducing updates when the device is idle and using low-power sensors until activity is reported.

Balance for Windows & macOS. We decided to use a multiplatform app to cover the broadest possible range of users, including users on the Windows and macOS operating system. The access to foreground apps, information on their title bars, and keyboard and mouse events were provided by the libraries *pywin32*³ and *pyobjc*⁴ on Windows and macOS, respectively. Both libraries are wrappers to low-level native operating system interfaces that allow direct access to system information, peripheral devices and functions. Our apps also relied on the *psutil*⁵ and *subprocess32*⁶ libraries. We used the cross-platform library *psutil* to

¹See: <https://developers.google.com/location-context/fused-location-provider/>

²See: <https://developers.google.com/location-context/activity-recognition/>

³See: <https://pypi.org/project/pywin32/>

⁴See: <https://pypi.org/project/pyobjc/>

⁵See: <https://pypi.org/project/psutil/>

⁶See: <https://pypi.org/project/subprocess32/>

abstract information about system load and access to running processes. This included, among other things, retrieving battery information, such as the remaining charge and power state. With the *subprocess32* library and native system calls, we parsed and scanned nearby Wi-Fi networks.

Applications & Notifications. Accessibility services⁷ or notification listeners⁸ are common methods to gather data on apps and notifications on Android phones in the field of interruption management [280, 281]. We used the accessibility service to gather the name and the package identifier of the used app. This information is always recorded when the window or its state changes. Another integrated service is the notification listener, which intercepts the reception and removal of notifications and accesses their underlying representation. This helped us obtain information, such as the time of arrival of the notification, the contact and group names the notification came from and the length of the notification’s content. To extract the contacts and group names, we added some apps to a white list to process their notifications on the smartphone directly. As we were only interested in contacts, we only added popular messaging apps, such as WhatsApp, Outlook, Twitter, Facebook, Microsoft Teams, Slack and Telegram, to the list.

In order to infer the responsibility of the user and distinguish between the notifications, we asked the users for their relationships to the senders. The users could choose *family*, *friend*, *work* or *none*, and multiple choices were possible. As pseudonyms were used to transmit sender information for data protection reasons, it was not possible to detect if a sender had a different name in different messenger apps or was part of a group chat. Therefore, we could not avoid sending multiple relationship questionnaires relating to one sender with different names. These additional questionnaires did not negatively influence the response rate, and a certain minimum number of correspondences with a sender was required to trigger the questionnaire.

Physiological data. During the data collection, participants were asked to wear *Empatica E4*⁹ wristband. The E4 wristband was first proposed by [70] and has multiple sensors: an EDA sensor, ACC, PPG sensor and optical thermometer. Another term for EDA is GSR or SCR,

⁷See: <https://developer.android.com/reference/android/accessibilityservice/AccessibilityService>

⁸See: <https://developer.android.com/reference/android/service/notification/NotificationListenerService>

⁹Empatica E4 wristband: <https://www.empatica.com/en-int/research/e4/>

which measures the continuous variation in skin electrical characteristics at 4 Hz. The ACC records the acceleration in three axes at 32 Hz in the range [-2 g, 2 g] and captures the physical activity of users. The PPG data is optically obtained and can be used to measure BVP at 64 Hz. The HR and IBI are derived from BVP signals by the wristband. The optical thermometer measures the peripheral ST at 4 Hz. The E4 wristband is lightweight and comfortable, making it suitable for continuous and unobtrusive monitoring in this study. It has long been known that emotions are related to the autonomic nervous system and are accompanied by changes in physiological signals [282, 283]. By measuring a person's physiological signals, changes can be recognised, and emotions can be assigned. We conducted a correlation analysis between mood and the features extracted from physiological signals.

ESM questionnaire. In this study, participants were asked to rate their mood over the preceding hour. We used the ESM from Bradley and Lang [284] to uncover the arousal and valence states. The arousal scale ranges from relaxed to excited, and the valence scale ranges from positive to negative. In addition, we gathered the dominant social role the person had been in for the preceding 15 min. Ashforth et al. [285] described a social role as a mental construct that individuals maintain to organise their surroundings. Therefore, we investigated *work* and *private* as domains, with their labelled social roles to characterise different behaviours. In contrast to existing studies [286, 287], we decided not to be more granular regarding the different roles, although *family*, *work* and *social* are reported as the most universal social behaviours. The focus of our study was based on the work–life balance, and the distinction between *social* and *family* seemed redundant, especially given the relationships that were assigned to contacts. Finally, we asked the participants for whom they were interruptible, i.e. contacts from the *work* or *private* domain, nobody or everybody (i.e. *both* domains).

5.4.2 Methodology

5.4.2.1 Pre-processing Approaches

At the beginning of the ML process, it is necessary to clean the data to eliminate noise and create a homogeneous dataset. This preparation helps with data processing in the steps that follow. One task during this stage was to standardise the app names across the platforms (i.e.

Windows, macOS and Android), e.g. changing the name microsoft-powerpoint to PowerPoint or removing system-specific endings. In addition, we parsed the Google Play Store websites according to the mobile apps used by our participants to extract the relevant app categories. The Google Recognition API returned all recognised physical activities and their corresponding confidence ratings. To reduce the data, we chose the activities with the highest confidence rating and forwarded the last known activity for all following events.

Upsampling was used for other data, such as the ringer mode, last known location and screen status. The *Plus Codes*¹⁰ software package from Google was used to extract more valuable place information and returned a code that gave us a description of a rectangular area, including the given longitude and latitude information. The accuracy of the location information generated by Plus Codes depends on the length of time it is used.

5.4.2.2 Extracted features

We prepared the data according to our needs for the regression model. It was decided that the best method for this investigation was to calculate the features on the data before the notification arrived. All extracted features are shown in Table 5.6.

Features extracted from Smartphone Data. We first examined the current context of the user. For this purpose, we analysed the apps used in the 5–30 min preceding the arrival of the notification. From this data, it was possible to deduce whether the user was interruptible and, accordingly, whether they would react immediately to an incoming notification. We discovered the top k smartphone apps by counting the frequency of appearance of the apps per user. Assume user X_1 has an app set $\mathcal{A} = \{A_1, A_2, \dots, A_N\}$, where the apps are sorted by the number of notifications received from the app in the training dataset, i.e. app A_1 received the most notifications and A_n received the fewest notifications. In this research, we only studied the top k apps where $k = 10$. We will explain the k in detail in Section 5.4.3.

There were five main indicators that were considered important in the process of finding the opportune moments to send notifications to the user to minimise notification response times. First, the sender–recipient relationship was closely related to the notification response

¹⁰See: <https://maps.google.com/pluscodes/>

Table 5.6: Extracted features by device. Data marked with (*) were manually reported

Feature	Description	Contextual Information
<i>Smartphone Data</i>		
topk_x_unique	Top k applications in the last $x \in 5, 10, 15, 20, 25, 30$ minutes.	Foreground application
phone_apps_X	Number of used smartphone applications in the last $x \in 5, 10, 15, 20, 25, 30$ minutes, extracted from the name and the package identifier of the current foreground application	Foreground application
physical_activity_X	Number of unique physical activities reported by the Google Recognition API	Physical activity
place_top_x, place_other	Top three ($x \in 1, 2, 3$) frequently visited places and all other places. Category of the location according to Google Geocoding API.	Location (Android)
screen_on, screen_off, screen	The current state of the screen.	Screen state
notification_length	Length of the text within the notification.	Notification content
Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday	Day of the week.	Notification arrival time
morning, afternoon, evening, midnight	Time of the day: morning (from 6 a.m. to 12 p.m.), afternoon (from 12 p.m. to 6 p.m.), evening (from 6 p.m. to 0 a.m.), and midnight (from 0 a.m. to 6 a.m.)	Notification arrival time
is_weekend	Binary value describing, whether it is weekend or not.	Notification arrival time
loc_8, loc_10	Longitude and latitude information of the device as Plus Code	Location
relation_x	The participants relationship to the extracted contact and/or group. Participants could choose between family, friend, work, and none. Multiple selections are possible (e.g., work and friend).	Relationship*
contact	Hashed contact and/or group name extracted from notification titles	Contact*
<i>Experience Sampling Method Data</i>		
valence, arousal	The affective state of the last 60 minutes	Mood*
private, work, both, none	Interruptibility preferences of the last 15 minutes.	Interruptibility*
private, work, both	Social role of the person in the last 15 minutes.	Social role*
<i>Physiological Signals</i>		
μ, σ^2, σ	Mean, Variance, Standard Deviation	EDA, SCR, SCL, BVP, HR, IBI, ST
min, max	Min and max value	EDA, SCL, SCR, BVP, HR, ST
rms	Root mean square	HR
f_{slope}	The absolute value of the slope of the linear regression line	EDA, SCL, HR, ST
$f_{\sqrt{slope}}$	The square root of the absolute values of the slope of the linear regression line	EDA, SCL, HR, ST
$f_{1intercept}$	The square root of the absolute value of the intercept of the linear regression line	EDA, SCL, HR, ST
$f_{2intercept}$	The third power of the square root of the absolute value of the intercept of the linear regression line	EDA, SCL, HR, ST
nmi_50/20, pnmi_50/20, nmi_20, pnmi_20	Number, and percentage of interval differences of successive RR-intervals greater than 50ms and 20 ms, respectively	IBI
vlsf, lf, hf, lf_hf_ratio	Power in HRV in the very low/low/high frequency. Power of lf/hf	IBI
sdsd, range_nmi	The standard deviation of differences between adjacent RR-intervals. Difference between the maximum and minimum nn_interval	IBI
cvsd, cvnni	Coefficient of variation, of successive differences (cvsd), equal to the ratio of rmssd / sdn divided by mean_nmi.	IBI
triangular_index	The HRV triangular index measurement is the integral of the density distribution divided by the maximum of the density distribution.	IBI

rate [288, 262]. Mehrotra et al. reported that the users' perception of the interruption depends on the sender of a notification, and chat notifications from family members have the highest acceptance rates. We considered both the length of the notification and the sender of the message. If the contact was known, we noted the relationship of the contact to the user. Second, notification response time was closely linked with whether the user was active on their smartphone at the time the notification was delivered. To gather this information, we queried whether the screen was on or not.

Third, breakpoints in physical activities have been proven to mark opportune moments for interruptions. Okoshi et al. [289, 290, 291] examined breakpoints in physical activities and app usage and found that notifications delivered at breakpoints, denoted as transitions between apps and physical activities, could lower individuals' mental burden. Ho and Intille [246] also suggested that notifications delivered during activity transitions produced more favourable outcomes than those delivered randomly. As described earlier, we used the Android Google API to record the current physical activities of the participants. The number of different activities detected was also used as a feature in the first stage classification.

Fourth, the location of the participants was considered important in determining their notification response time. We used Plus Codes to represent the current location of the user. The most frequently visited locations for each participant were set as features. For this purpose, we first determined the three locations that each participant visited most frequently during the measurement process, and these locations represented their top three locations. All other Plus Codes were assigned to the category 'other'. The location of the user before receiving the notification was noted by setting one of the top three locations or the category 'other' to true.

Fifth, the time of day was considered important in determining users' notification response times. Several previous studies have investigated the relationship between times of the day and notification responses [291, 258, 259]. Okoshi et al. and Saikia et al. found that sending notifications at opportune times greatly reduced response times. Therefore, we noted the day and the time of day to represent the time a notification was received. As previously described, we split the day into four parts, i.e. midnight (from 12 am to 6 am), morning (from 6 am to 12 pm), afternoon (from 12 pm to 6 pm) and evening (from 6 pm to 12 am) and the week into

two parts, i.e. week days and weekends.

Features from ESM Data. In addition to the features already mentioned, we used the ESM questionnaire data, which described the users’ moods, interruptibility and current social role. Mood was measured on two scales: valence and arousal. These scales represent different types of feelings on a scale of 1–5: from unhappy to happy and from calm to excited, respectively. We used the features that contained contextual information about interruptibility and the social role. We applied one-hot encoding to represent this nominal data.

Features for Physiological Signals. We decided to extract statistical features on all physiological signals that are commonly used for mood recognition. As suggested by Heinisch et al. [292], we added features based on the linear regression, as these features had been shown to be robust influencing factors of physical activity. The EDA signal can be divided into two components, the SCR and the SCL. The SCR contains high-frequency components of the signal, reflecting rapid changes in the signal in response to a stimulus. In contrast, the SCL contains low-frequency components of the EDA, representing the long-term or baseline conductance. We used the Python tool suggested by Greco et al. [129] to split the EDA signal into these two components.

Table 5.7: Notification and app information for 18 participants

	Min	Max	Median	Mean
Number of apps	18	47	26	30
Number of notifications	363	6213	1914	2362
Percentage of notifications sent by top 10 apps	84.67%	99.57%	94.88%	94.30%
Percentage of notifications sent by top 5 apps	65.56%	97.90%	84.17%	83.33%

5.4.3 Understanding the Mood, Usage Behaviours and Notification Response Time of Participants

5.4.3.1 Understanding Notification Response Times for Different Participants

In total, we have received 3,236 ESM responses and 42,270 notifications from 18 participants during the data collection. We explore the notification response time from top- k apps where $k = 10$ because on average, the top ten apps sent 94.30% of the notifications (out of 2,362 notifications), while the other apps only sent 5.70% of the notifications (see Table 5.7). If we only study the top five apps, we would miss 16.67% of the notifications, which is almost three times the number of missed notifications from studying the top ten apps. For instance, Figure 5.3 displays the number of notifications across all the apps for one participant P10 during the data collection. We find that P10 received 96.04% notifications from top ten apps and 86.16% from top five apps. Therefore, in this research, we did not consider the apps receiving only a few notifications ($k > 10$) because the relatively small data set would not offer a robust representation of the notification response times for modelling. In real-world scenarios, k can be set to any values based on the categories of apps being explored.

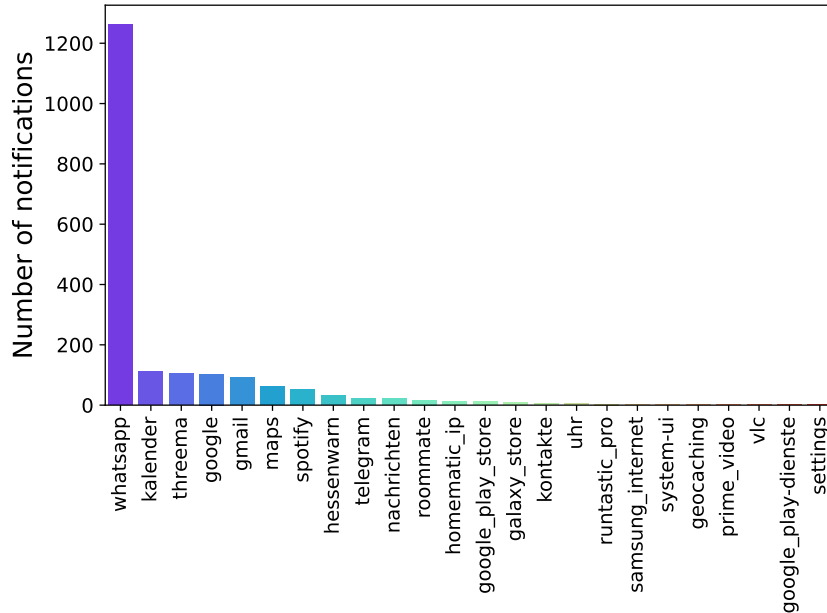


Figure 5.3: The number of notifications across all the apps for participant P10

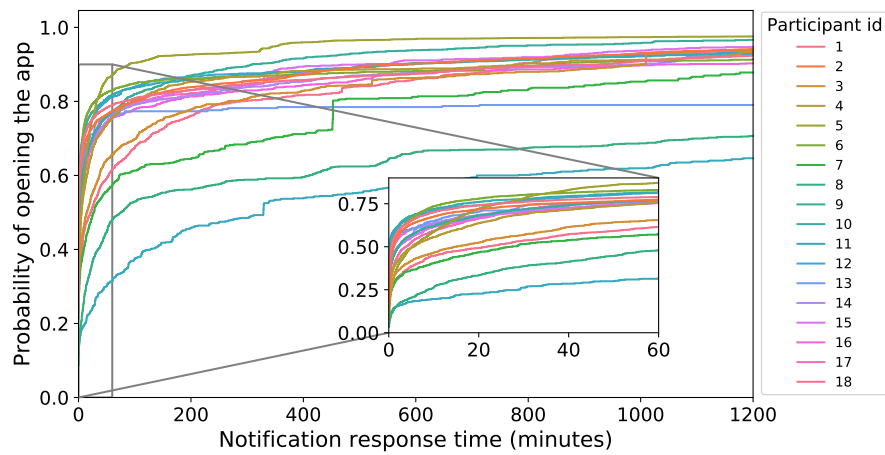


Figure 5.4: Cumulative distribution of notification response times from the top 10 apps for all participants

To understand the notification response time for all participants, we show the cumulative distribution of notification response times from top ten apps for each participant in Figure 5.4. It is obvious that the response time to most notifications is short, but the response time of some notifications is long. Specifically, out of 40,290 notifications received by 18 participants, the response time was within five minutes for 54.32% of the notifications, within one hour for 75.86% of the notifications, and within one day for 93.90% of the notifications. However, if we look at the response times for different participants, we find that each participant has their own patterns and trends for responding to notifications. For instance, participant P5 responded to 49.37% of their notifications within five minutes and 86.96% within one hour, while participant P11 responded to notifications much more slowly, only responding to 18.46% within five minutes and 32.36% within one hour. Hence, studying the participant-wise notification response time is necessary, as the general model may be inaccurate due to individual differences.

5.4.3.2 App Categories and Response Time

Figure 5.5a displays the number of notifications across the app categories, showing that the *communication* apps receive much more notifications than all the other app categories. In total, *communication* apps receive six times more notifications than the app category that was ranked second (i.e. *Productivity* apps). Figure 5.5b shows the average response times for

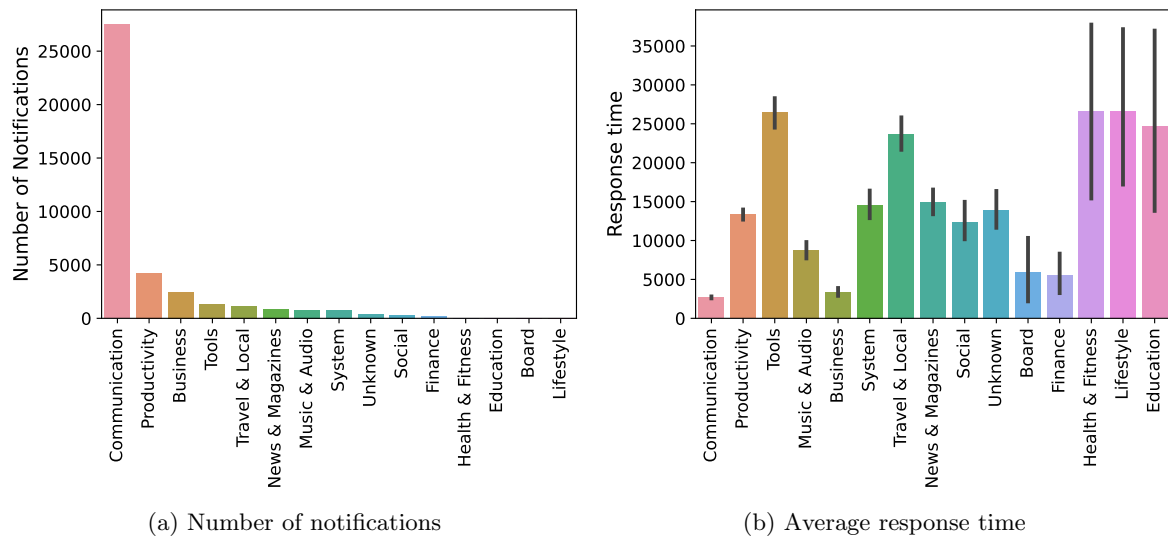


Figure 5.5: Information for different app categories

each app category (black vertical line indicates the error bar, with a 95% confidence interval). Since 93.90% of the notifications from all participants are responded to in one day, we focus on analysing those notifications and have removed the notifications with a response time of more than one day. Messages that have not been responded to more than 24 hours may be due to various reasons, such as the user forgot or has already responded on other platforms. We believe that it is more meaningful to focus on the notifications that users reply in a timely manner, and the small number of notifications unanswered for a long time will be explored in our future research. We find that the response times varied significantly between the app categories. If we aim to predict response time across all categories, the prediction performance would be unreliable due to the extreme variations in the number of notifications and the average notification response time between app categories. Therefore, in this research, we focus on predicting users' response behaviours for *communication* apps.

5.4.3.3 Impact of Applications on Notification Response Times

We already know that each participant has their own patterns for responding to notifications. However, we also investigate whether each participant responds to different apps in different ways. Here we explore the influence of apps on notification response times. Figure 5.6 shows

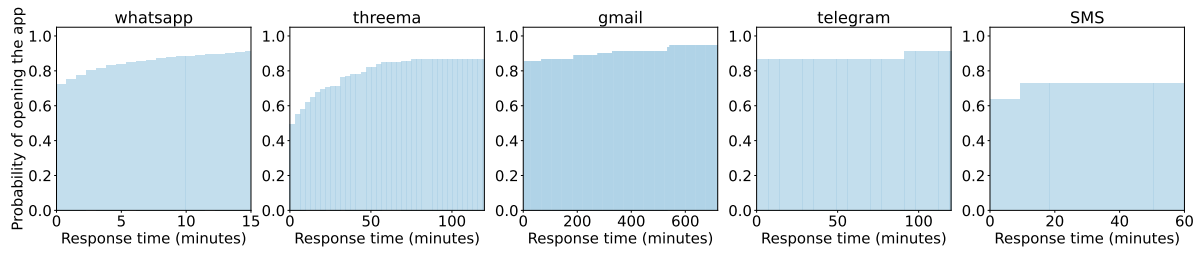


Figure 5.6: Cumulative distribution of notification response times from five apps for participant P10

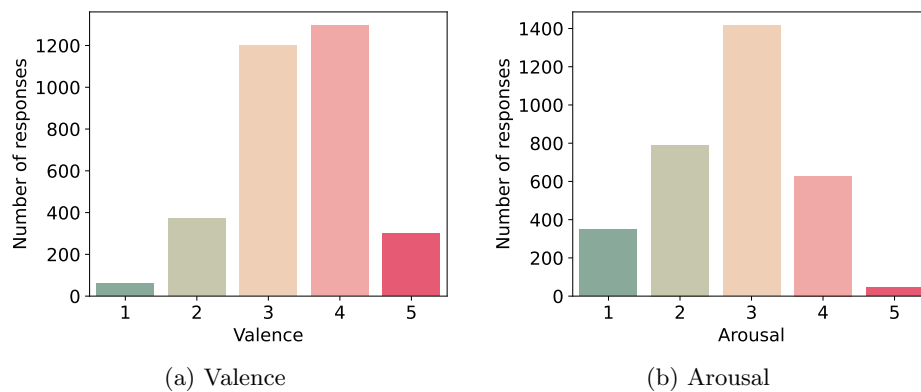


Figure 5.7: Distribution of arousal and valence for 18 participants

the cumulative distribution of the notification response times for five popular apps for participant P10. It clearly shows that even for the same participant, the notification response times vary from app to app. For example, this participant usually responded quickly to *whatsapp*, *gmail* and *telegram* but much more slowly to *threema*. Specifically, within five minutes, this participant responded to 83.53% of notifications from *whatsapp* but 53.33% from *threema*. Therefore, it is necessary to consider the impact of the apps to meaningfully model the notification response times.

5.4.3.4 The Mood of Users and Notification Response time

We calculate the overall distribution of mood in Figure 5.7, where 1 to 5 indicates a low to high value of valence/arousal. Generally, participants usually reported positive valence (mean = 3.44) and low arousal (mean = 2.77), meaning that they were relaxed, clam, and comfortable

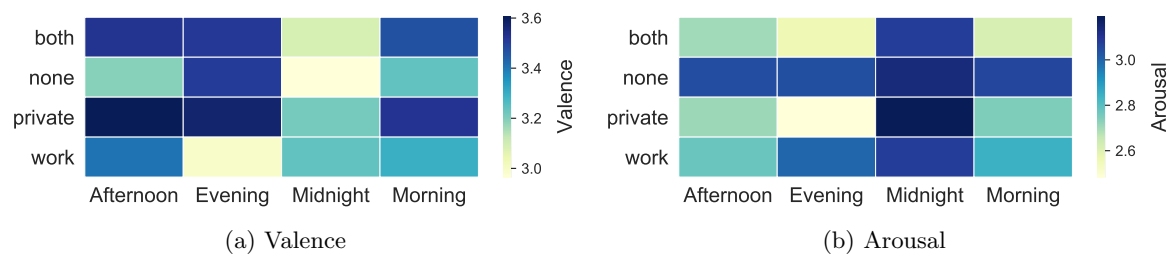


Figure 5.8: Mood of participants at different levels of interruptibility and various times of the day

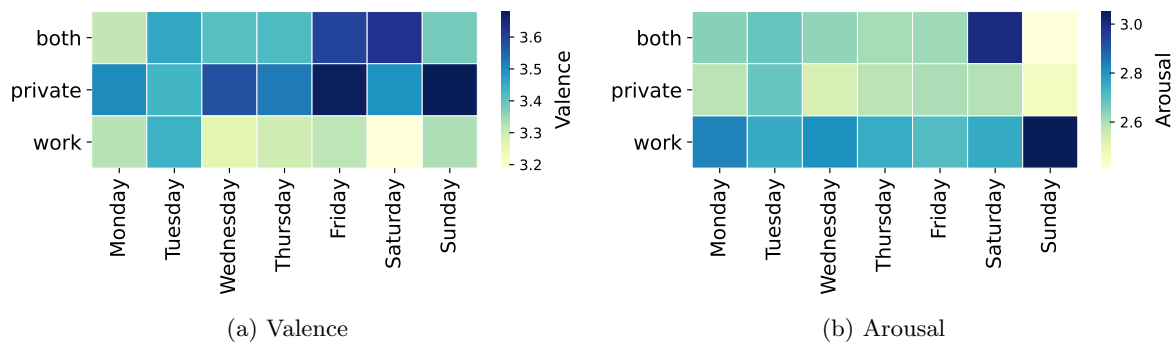


Figure 5.9: Mood of participants over different social roles and days of the week

[185] most of the time. We also explore how the mood is related to factors such as daytime and interruptibility. As shown in Figure 5.8a and Figure 5.8b, participants usually experienced the highest valence (mean = 3.56) and lowest arousal (mean = 2.58) in the evening (6pm - 12am). In contrast, they usually experienced the lowest valence (mean = 3.28) and highest arousal (mean = 2.97) in the midnight (12am-6am). We also found that when the participants did not want to be interrupted by either work or private affairs (i.e. interruptibility was ‘none’), they were usually experiencing lowest valence (mean = 3.21) and highest arousal (mean = 3.03). Interestingly, when the participants experienced positive mood (high valence), they were more likely to be amenable to interruptions relating to private, or private and work (i.e. both) affairs. In general, the participants experienced varying mood with different levels of interruptibility at different times.

We also investigate how the mood changed based on social roles and the day of the week (see Figure 5.9a and Figure 5.9b). We found that participants usually experienced high valence

(mean = 3.55) and low arousal (mean = 2.67) when they were busy with private issues and tended to experience low valence (mean = 3.31) and high arousal (mean = 2.88) when they were at work. Our participants had the highest valence in the private role on Friday (mean = 3.66) and Sunday (mean = 3.66) and the lowest valence values (mean = 3.25) at work on Saturday. Saturday and Sunday were also different in the arousal scale, as the social roles ‘both’ (mean = 3.08) and ‘work’ (mean = 3.22) had the highest values, respectively. Interestingly enough, being in the role of private or both made our participants feel the lowest arousal (mean = 2.5) on Sunday.

5.4.4 Experiment

As introduced in Section 5.4.1.2, we focus on predicting the response times of 18 participants who installed the smartphone app. In this research, we built the regression model for predicting the users’ notification response times. First, we introduce the experimental setting and prediction pipeline. Then we show the overall results of the predictions and the impact of the study on the mood-related features. Finally, we investigate how individual differences and categories of apps influence response times.

5.4.4.1 Prediction Pipeline

We have adopted the regression model for predicting notification response times. The prediction pipeline is described below.

Regression models. In the prediction model, we adopted several commonly used regression models, such as *Standard Linear Regression* [151], *SVR* [293], *Gradient Boosting Regression* (GBR), *Random Forest Regression* [294] and *Bayesian Ridge Regression* [295]. Linear regression is one of the most widely used regression models. The *support-vector machine* in regression problems is usually known as SVR, which is one of the most commonly used regression models. The GBR model is a powerful prediction model, and it is an ensemble method combining a set of weak predictors to achieve reliable and accurate predictions. *Random Forest Regression* uses the idea of a random forest, and it can estimate the importance of various features in a model. *Bayesian Ridge Regression* conducts linear regression using probability distributors

rather than point estimates, which provides a natural mechanism to create predictive models when data are insufficient or poorly distributed.

Validation. Cross-validation is a common practice for training and testing prediction models and is used to estimate the unbiased generalisation performance of models. However, cross-validation may lead to the optimistically biased evaluation of prediction performance when the same cross-validation process is chosen to both tune and select the model. Similar to previous ubiquitous computational studies [30, 6], we adopted *nested cross-validation* [148], which performs two iterations over the data. The outer loop is used to evaluate the performance of the regressors, and the inner loop is used for optimisation of hyper-parameters and feature selection. After performing this cross-validation, we then applied *k-fold cross-validation* ($k = 5$) on both loops for each participant. In the outer loop, once the training set and testing set were defined, we standardised features by removing the mean and scaling the data to unit variance. In the inner loop, we optimised the hyper-parameters using a grid search. We then selected features according to the K highest scores based on *f-regression* [215] (f-value between the label/feature for regression tasks). The top eight features ($K = 8$) were selected as the input features for each regression model because we found that this resulted in the lowest prediction error.

Baselines. In human-centred research, it is usually difficult to compare the prediction results with state-of-art baselines. The main reason is that the types of data collected, the demographics of participants and the natural environment vary widely across studies, it is not fair or applicable to compare the prediction performance between different studies. Additionally, to our knowledge, we have not found any research that attempts to predict the notification response time for mobile users. As a result, similar to previous human-centred studies [30, 120], we have adopted simple baselines to compare the modelling performance. In particular, we compare the proposed models with two baselines: *Mean* baseline and *Median* baseline. As one of the most widely used simple baselines to compare with other regressors, *Mean* baseline always predicts the mean of the training set. *Median* baseline always predicts the median of the training set. The reason why we choose Median baseline is that the distribution of notification response time is highly skewed (see Figure 5.4), whereas the Median baseline is most

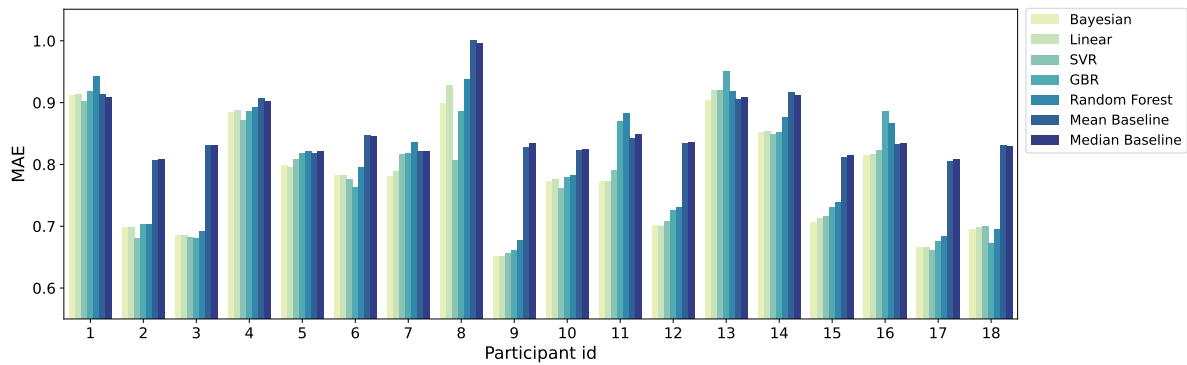
informative for skewed distributions or distributions with outliers.

Evaluation Metrics. To evaluate the performance of notification response time, the *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE) metrics are applied for evaluating the prediction performance. The $MAE = \frac{1}{n} \sum_{i=1}^n |y_{true} - y_{pred}|$ and $RMSE = \frac{1}{n} \sum_{i=1}^n (y_{true} - y_{pred})^2$, where n indicates the number of samples, y_{true} means the actual notification response time and y_{pred} means the predicted response time. The MAE and MSE describe the goodness of predictions compared with the ground truth of notification response time. The closer the MAE and MSE are to 0, the better the performance of the prediction model.

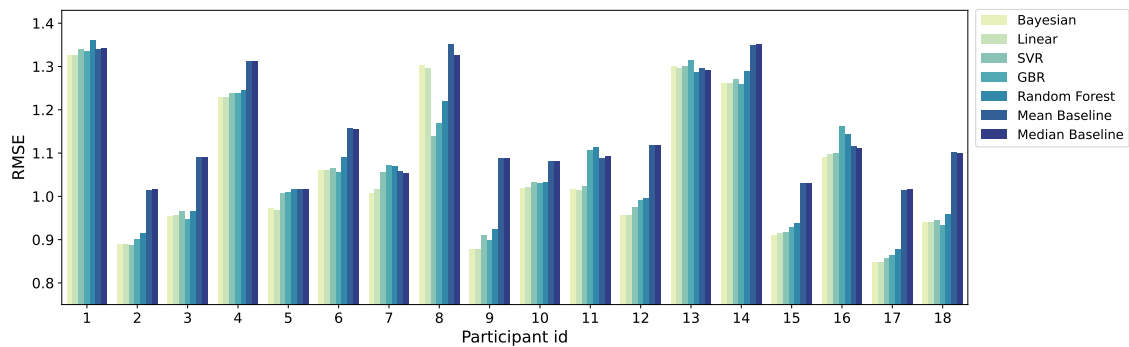
5.4.4.2 Prediction Result with Mobile Data

As discussed in Section 5.4.3, the notification response behaviours were very different between the participants (see Figure 5.4). Therefore, in the experiment, we built participant-wise regression models instead of a general model for all participants. Figure 5.10a and Figure 5.10b show the MAE and RMSE results across different regressors for each participant. We found that the regression models achieved much better predictive performance than both baselines for most participants (i.e. P2, P3, P9, P12, P17 and P18). For example, for participant P9, the *Bayesian* regression model had the best predictive performance (MAE = 0.6505 and RMSE = 0.8779), with MAE = 0.1828 (21.94%) and RMSE = 0.2101 (19.31%) lower than the *median* baseline model.

However, for some particular participants (e.g. P1 and P13), only a small number of regressors achieved a lower MAE and RMSE than the baseline models. The possible reasons why some regressors did not work well on a small number of participants are twofold: (1) The notification response behaviours of these participants were more random and changeable than others, which makes them difficult to predict. These individual differences in mobile usage behaviours have been discussed in prior research [296]. (2) These participants had very different notification response behaviours when using different apps, which is difficult to represent in one regression model. However, it was not practical to build a predictive model for each app due to the limited number of notifications.



(a) MAE result



(b) RMSE result

Figure 5.10: Prediction results across different regressors for each participant

Next, we calculated the overall predictive performance for all participants by averaging the MAE and RMSE values from the participant-wise models. Table 5.8 shows the overall predictive result for all participants. It shows that all regression models had better predictive performance than the two baseline models in terms of MAE and RMSE, demonstrating the models' potential for predicting notification response times for ordinary people. The *Bayesian* model achieved the best predictive performance of all the regression models and obtained MAE = 0.7764 and RMSE = 1.0527, which was 0.078 (9.10%) and 0.093 (8.09%) lower than the mean baseline MAE and RMSE, respectively. Although the overall predictive performance does not sound particularly good, the predictive performance was very high for most individuals (see Figure 5.10).

Figure 5.11 shows the importance of each feature for each participant, which was calculated using the f-regression score in the *scikit-learn* python package. Higher values indicate more

Table 5.8: Predictive results with different regressors using mobile data

	<i>Bayesian.</i>	<i>Linear.</i>	<i>SVR.</i>	<i>GBR.</i>	<i>R. Forest.</i>	<i>Mean Baseline</i>	<i>Median Baseline</i>
MAE	0.7764	0.7797	0.7770	0.7797	0.8014	0.8541	0.8544
RMSE	1.0527	1.0533	1.0601	1.066	1.0798	1.1454	1.1441

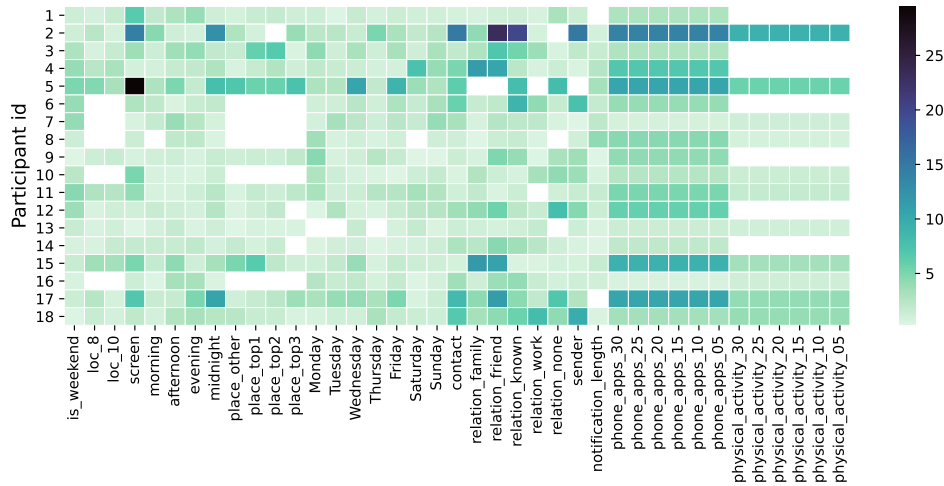


Figure 5.11: Feature importance for each participant in the prediction

important features. Understanding the importance of a feature is significant in helping us better understand a problem and can lead to better predictive performance through feature selection. In Figure 5.11, we can see obvious individual differences in feature importance for predicting notification response times. For example, the response time for some participants (e.g. P3, P5 and P15) was significantly affected by location, while some participants' (e.g. P12 and P13) were not affected by location. Many participants' response times were influenced by the time of the day (i.e. weekends or week days), screen status, relationship with senders or the number of apps used in the past 5, 10, 15, 20, 25 or 30 minutes. The above phenomena are in line with our daily experience and may be due to the various personalities or usage habits of mobile users [12, 260].

5.4.4.3 Impact of Mood-related Features

We also explored the impact of mood-related features on predicting notification response times. The mood-related features were divided into two groups: ESM features and E4 features. For

ESM features, we mainly focused on the perceived arousal and valence, based on ESM questionnaires. For E4 features, we mainly focused on the features extracted from physiological signals (i.e. EDA, HRV and ACC) from the E4 wristbands.

ESM features. We built regression models with two different sets of features: (1) mobile features and ESM features and (2) mobile features only. Since we only had a limited number of ESM responses, we removed the data without corresponding arousal and valence values. To achieve a fair comparison, we used the exact same rows of data (8,408 data instances) in each of the above two models. The results of the experiment showed that all the regression models using the second set of features had higher MAE and RMSE values than those using the first set of features, where the MAE/RMSE of baseline models are exactly the same. The findings indicated that the ESM features improved the predictive performance of the model for notification response times.

E4 features. To study the impact of the E4 features, we built regression models with two different sets of features: (1) mobile features and E4 features and (2) mobile features only. After removing the NaN values in the whole dataset, 1,491 rows of data remained, which were used to build the regression models using the two sets of features, as mentioned above. The results of the experiment showed that most of the regression models (except *Bayesian* regression) achieved better predictive performance with the first set of features, i.e. mobile features and E4 features. A possible reason may be the small number of E4 data instances, e.g. participant P11 only had 19 rows of data, and P9 only had 27 rows of data, which makes it difficult to make meaningful predictions.

5.4.5 Implications and Limitations

This research addressed the relationship between mood and interruptibility and investigated the possibility of automatically predicting notification response times and actions based on users' moods. Our research also provides opportunities for the future design of intelligent notification management systems for mobile or desktop devices, which could benefit the wellbeing and productivity of users. In our research, we analysed the impact of mood, as measured by ESM questionnaires, and physiological data, as measured by E4 wristbands, on notification response

times. We found that affective data can help to improve regression models to assist in the handling of smartphone notifications.

ESM data. One limitation of our study is that some data, such as mood, was gathered using an ESM questionnaire pushed either every 90 minutes or after the user had been using their smartphone for 10 min. This kind of questionnaire must be seen as an interruption itself. In addition, the questionnaire popped up on the smartphone as a notification, which may have caused the participants to interact with their smartphones more often than they would normally have. However, this method of data collection is very common in the field of interruption management, and as the data were used to develop the individual regression models, we believe that these initial results are valuable for further research. We are aware that a follow-up in-the-wild study is needed to validate the models developed.

Mood. Another limitation is the use of the ESM questionnaire to capture the participant's mood. It is important to note that many people struggle to identify or name their moods correctly [283], and the reliability of self-report data can be influenced by various response biases [40]. To compensate for this weakness, we added physiological signals to the ESM data, which also conveys information about human affective states. Even though these are not free of external influences (e.g. external temperature and physical movement), they form a basis for the research in combination with the ESM data.

Data Distribution. There was minimal diversity in terms of age and gender, and there were only a small number of participants. In particular, the number of participants wearing the E4 wristband needs to be increased in future research to reduce the potential for bias. In addition, the data were very unbalanced because of the number of different apps used by each participant and the number of notifications. There was significant variation in how the subjects behaved and the apps that they used. Some users interacted frequently with many apps, while some users interacted very frequently with a few apps and rarely with many other apps. These factors mainly influenced the results of the regression analysis, making it almost impossible to create a generalised model. After pre-processing, we also recognised that for some participants the quantity of data recorded was very low.

5.5 Conclusion

In this research, we first demonstrated that it is possible to combine human physical activity intensity data with traditional phone activity data to estimate the Big Five personality traits. We proposed a set of important metrics based on dispersion, diversity, regularity, etc. and found some interesting associations between human activity patterns and personality traits. We used SVR to predict participants' personality scores. The results of the experiment showed that our predictive model was highly correlated with the ground truth and outperformed the baseline model. We also found that the performance of the predictive model differed between females and males, with an observable reduction in errors when the predictive model was split by gender (when compared with the model for all participants). These results present a significant step in passive human personality prediction using smartphone activity data.

Understanding the notification response behaviours of users is of vital importance to developing the next generation of mobile management systems to improve users' overall productivity and wellbeing. In this research, we predicted notification response times by understanding people's mobile usage behaviours, moods and physiological patterns. We conducted an *in-the-wild* study of more than 18 participants with mobile devices and wearables over a five-week period. We developed multiple regression models to predict the notification response times for each participant. The experimental results showed that the proposed model achieved greater predictive performance than all the baseline models. We found that the use of both the mood data from the ESM questionnaires and physiological signals (e.g. EDA and HRV) improved the predictive ability of the models significantly. In addition, we identified the most significant features affecting the accurate prediction of notification response times for each participant, and we discussed various factors affecting the predictive performance, such as individual differences between users and categories of apps. The research showed that notification response times can be predicted accurately using smartphone data (e.g. location and app usage), and the predictive performance can be significantly improved by utilising mood-related information from ESM data or physiological signals. This result is a significant step toward achieving an attention management system that combines human wellbeing and behaviours.

Chapter 6

Modelling Thermal Comfort with Limited Labelled Data in Smart Buildings

Previously in Chapter 3, 4 and 5, we utilized heterogeneous sensing data to predict human behaviours and mental states. However, it is usually difficult to obtain sufficient labelled data in human-based studies for accurate data-driven modelling. In relation to *RQ-5*, this chapter aggregates behaviour (i.e., thermal comfort) from environmental sensing with limited annotations by transferring knowledge from multiple locations to another domain. We present a transfer learning-based multilayer perceptron model from the same climate zone (TL-MLP-C*) for accurate thermal comfort prediction. Extensive experimental results on the ASHRAE RP-884, Scales Project and Medium US Office datasets show that the performance of the proposed TL-MLP-C* exceeds the performance of state-of-the-art methods in accuracy and F1-score.

6.1 Introduction

Recently, Internet of Things (IoT) devices have been widely used in urban environments. In addition, sensors have become the backbones of smart cities that enable spatial and situational

awareness of real-time dynamic phenomena, e.g., pedestrian movement [297], parking events [298], and energy consumption [299, 300]. As one of the most important parts of cities, buildings account for approximately 40% of the global energy usage and 60% of the worldwide electricity usage [301]. Large proportions of these usages are contributed by buildings' HVAC systems [302]. The main goal of the HVAC system is to maintain the indoor occupant comfort at minimal energy usage. To achieve overall satisfaction with an indoor environment, thermal comfort is considered to be the most influential factor compared with visual and acoustic comfort [303].

Thermal comfort is the state of mind that expresses satisfaction with the thermal environment [304]. Thermal discomfort not only affects occupant productivity, work performance and engagement [30, 15], but it also has a negative influence on lifelong health. Hence, it is important to maintain a thermally comfortable environment for the well-being of occupants while minimizing buildings' energy usage. A crucial step towards this goal is to create an accurate model for thermal comfort. The *Predicted Mean Vote* (PMV) model proposed by Fanger et al. [305] developed with principles of human heat balance and adopted by the ASHRAE Standard 55, is one of the most prevalent models. It relates the thermal comfort scale with six different factors (see Figure 6.1).

However, some researchers revealed the discrepancy between the predicted mean vote and occupant-reported thermal sensation votes [306]. This discrepancy is likely because a variety of parameters such as time factors (e.g., hour, day, and season) [307, 308], personal information (e.g., heart rate, age, and gender) [309], environmental factors (e.g., colour, light, and outdoor climates) [310], culture (e.g., dress code and economic status) [311], short- and long-term thermal exposure [308], etc. may affect thermal comfort. Therefore, a data-driven method is a better choice than the traditional PMV model since more parameters could be utilised to improve the performance of thermal comfort prediction.

Some researchers have applied data-driven machine learning techniques for thermal comfort prediction for a specified group of people. However, it is usually difficult to obtain sufficient labelled data, which limits the performance of data-driven models. Recently, various thermal comfort studies have been conducted worldwide; and several databases, including databases

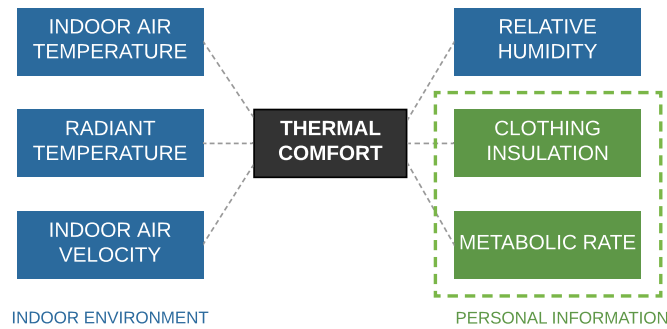


Figure 6.1: Six factors affecting thermal comfort (PMV model)

covering multiple cities and climate zones, are currently available online (see Section 6.3). Since the sensor data inferred from different cities may have very divergent patterns caused by building materials, construction requirements and climate changes, previous research has mainly focused on investigating how people living in specific cities react to their thermal environment (e.g., the hot-arid climate in Kalgoorlie-Boulder, Australia [312] and the humid subtropical climate in Brisbane Australia [313]).

We aim to explore whether we can utilise sensor data from multiple cities to benefit a target building. We hypothesize that the performance of thermal comfort modelling can be boosted by conducting transfer learning using data from multiple cities. Therefore, we wish to answer the following research questions: *Can we predict occupants' thermal comfort accurately by learning from multiple buildings in the same climate zone when we do not have enough data? If so, which features contribute the most to effective thermal comfort transfer learning?*

In this chapter, we present the *transfer learning-based multilayer perceptron* (TL-MLP) model and *transfer learning-based multilayer perceptron from the same climate zone* (TL-MLP-C*) model for predicting occupants' thermal sensation with insufficient labelled data. ASHRAE RP-884 [305] and the Scales Project [314] are chosen as the source datasets, and the Medium US Office [315] is used as the target dataset. Extensive experiments on these three public databases show that the proposed thermal comfort models outperform the popular knowledge-driven and data-driven models. To summarize, the contributions are as follows:

- To the best of our knowledge, we are the first to transfer the knowledge from similar

thermal environments (climate zones) to a target building for effective thermal comfort modelling. We propose the TL-MLP and TL-MLP-C* thermal comfort models and confirm that the thermal comfort sensor data from multiple cities in the same climate zone can benefit the small thermal comfort dataset of a target building in another city, with insufficient training data.

- Extensive experimental results show that the proposed TL-MLP and TL-MLP-C* models outperform the popular knowledge-driven and data-driven models for thermal comfort prediction and can be implemented in buildings without adequate thermal comfort labelled data.
- We identify the significant feature sets for effective thermal comfort transfer learning. We also find that the combination of age, gender, outdoor environmental features and the six factors from the PMV model can lead to the best prediction performance for transfer learning-based thermal comfort modelling.

6.2 Related Work

First, we list the previous literature for traditional thermal comfort modelling methods and transfer learning applications. Then, we discuss the current gaps and identify the advantages of this work.

6.2.1 Traditional Thermal Comfort Modelling Methods

The PMV model developed by Fanger et al. [305] and adaptive model developed by De Dear et al. [313] are the most famous knowledge-driven thermal comfort models. The adaptive model is based on the idea that occupant can adapt to different temperatures at different times and that outdoor weather affects indoor comfort. Occupants can achieve their comfort through personal adjustments such as clothing changes or window adjustments [316]. Clear et al. [317] explored how adaptive thermal comfort could be supported by new ubiquitous computing technologies. They noted that IoT sensing technologies can help build a more sustainable

environment where people are more active in maintaining and pursuing their thermal comfort, which is less energy-intensive and less tightly controlled.

In recent years, data-driven thermal comfort modelling has become increasingly more popular and huge efforts have been made to apply machine learning to thermal comfort modelling [37, 38, 309, 318, 319, 320, 321]. Ran et al. [37] used rotation forests to predict occupants' thermal comfort using thermographic imaging information. Similarly, Ghahramani et al. [38] used a hidden Markov model (HMM) based method to predict thermal comfort using the infrared thermography of faces. Chaudhuri et al. [309] established a random forest-based model for different genders using physiological signals (e.g., skin conductance and blood pressure). However, all the thermal comfort models mentioned above require the installation of additional devices (individual thermal cameras, smart eyeglasses, and physiological sensors) and may lead to privacy concerns.

The performance of traditional machine learning algorithms on thermal comfort prediction has been discussed in [318]. Researchers compared nine widely used machine learning algorithms for thermal sensation prediction using the ASHRAE Comfort Database II. They found that ML-based thermal sensation prediction models generally have higher accuracy than traditional PMV models and that the random forest has the best performance compared to other ML algorithms.

As the non-traditional machine learning algorithms, artificial neural networks have been increasingly used in thermal comfort modelling. Ferreira et al. [319] controlled an HVAC system to achieve the desired thermal comfort level and energy savings. They applied several neural network models to calculate the PMV index for model-based thermal comfort prediction. Hu et al. [320] implemented a black-box MLP neural network for thermal comfort modelling, which obtained better prediction performance than the PMV model and traditional white-box machine learning models. Compared to most previous research using a coarse-grained neural network architecture (link input attributes and thermal comfort score directly), Zhang et al. [321] used the MLP neural network to model the relationship between controlling building operations and thermal comfort factors. Their proposed fine-grained DNN approach for thermal comfort modelling outperforms the coarse-grained modelling and other popular machine

learning algorithms.

6.2.2 Transfer Learning Applications

Although great contributions have been made to improve the prediction accuracy of thermal comfort through various machine learning techniques, there is still a main bottleneck for data-driven thermal comfort modelling - the accessibility of sufficient thermal comfort data. Transfer learning allows researchers to learn an accurate model using only a tiny amount of new data and a large amount of data from a previous task [34].

Transfer learning has been applied to many real-world applications involving Figures/video classification, natural language processing (NLP), recommendation systems, etc. For instance, transfer learning has been used for children's Automatic Speech Recognition (ASR) task [322]. Researchers learn from adult models to child models through a Deep Neural Network (DNN) framework. They investigated the transfer learning techniques between adult and child ASR systems in acoustic variability (layers near the input) and pronunciation variability (layers near the output), updated both the top-most and bottom-most layers and kept the rest of the layers fixed.

Some existing work has focused on transfer learning using sensor data. Wang et al. [323] proposed a transfer learning based-framework for cross-domain activity recognition. First, they used the majority voting technique to obtain the pseudo label of the target domain. Intra-class knowledge transfer was interactively performed to convert two domains into the same feature subspace. Then, the labels of the target domain can be ignored by the second annotation. Ye et al. [324] learned human activity labels by leveraging annotations across multiple datasets with the same feature space, even though the datasets may have different sensing deployments, sensing technologies and different users.

Recently, a transfer active learning framework was proposed to predict thermal comfort [36]. They considered thermal comfort prediction as inductive transfer learning where labelled data are available in both source and target domains but users do not have access to all labelled data in the target domain. They used the parameters transferred from the source domain to the target domain. The biggest disadvantage of their method is that they assume the feature

spaces in both domains must be the same, which is not applicable in daily life as there may be unique useful features in the target dataset.

Similarly, Hu et al. [35] adopted transfer learning for thermal comfort modelling and assumed that the feature space of the source domain is a subset of that of the target domain. They connected the classifiers from the source domain and target domain and then built a new classifier to obtain knowledge from the source domain; however, they did not explain why the network structure works well. Besides, they trained the thermal comfort model for a lab study and learned knowledge from the data from buildings all over the world in the ASHRAE RP-884 dataset, but they did not consider the differences in the thermal environments in different climate zones.

Overall, there are several advantages of our work: (1) We are the first to transfer the knowledge from similar thermal environments (climate zones) to the target building for effective thermal comfort modelling. Most previous research has focused on building a thermal comfort model for one target building [320, 321, 37, 38, 309, 318, 319]. Although a few researchers [35] have started to use transfer learning for building thermal comfort models, their target datasets are collected from laboratory studies and do not consider the influences of different climate zones. (2) Unlike some research that uses data collected from laboratory studies [37, 38, 309, 35], we build thermal comfort models using data from field studies in both the target and source domains, which is much more meaningful in real-world scenarios. (3) Compared with some research utilising additional devices (e.g., thermal cameras in [37], eyeglasses in [38], and wristbands in [35]), our research is easier and cheaper to conduct, and better protects the privacy of occupants.

6.3 Data Sets Introduction

6.3.1 Overview

ASHRAE RP-884 Database [304] is one of the most popular public databases for human thermal comfort research [325, 326]. It was initially collected to develop De Dear's adaptive model, involving more than 25,000 observations collected from 52 studies and 26 cities over

Table 6.1: Information for source dataset and target dataset

Dataset	ASHRAE RP-884	The Scales Project	Medium US Office
Instances	25,623	8225	2,497
Participants	Unknown (48% M, 52% F)	8225 (53% M, 46% F)	24 (33% M, 67% F)
Indoor AT Range (C)	6.2 - 42.7	13.2 - 34.2	17.9 - 27.8
Indoor RH Range (%)	2.0 - 97.8	18.0 - 82.4	15.7 - 72.4
Indoor AV Range (m/s)	0.01 - 1.71	0.00 - 0.70	0.02 - 0.19
MR Range (Met)	0.64 - 6.82	N/A	1.00 - 6.80
CL Range (Clo)	0.04 - 2.29	N/A	0.21 - 1.73

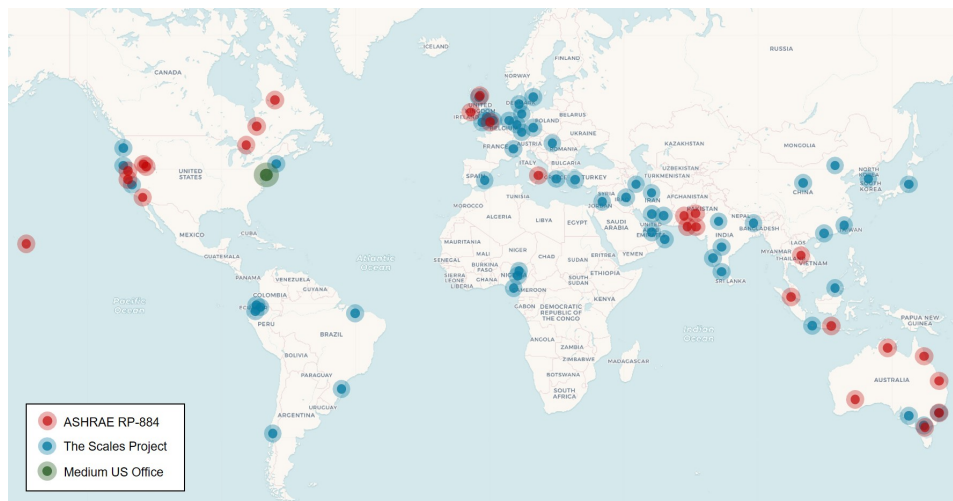


Figure 6.2: Locations of different studies in ASHRAE RP-884 database, The Scales Project database and Medium US Office dataset

different climate zones all over the world. We adopt this public database as one of the source datasets in our research.

The Scales Project Dataset [314] released in 2019 and it contains thermal comfort responses from 57 cities in 30 countries for 8225 participants. This dataset aims at exploring participants' thermal comfort, thermal sensation, thermal acceptances and to investigate the validity of assumptions regarding the interpretation of responses from the survey. This public dataset is used as one of the source datasets in the research.

Medium US Office Dataset [315] is a popular dataset used by many thermal comfort studies [327, 321]. It collected data from 24 participants (16 females and 8 males) in the Friends Center Office building in Philadelphia city, USA. Longitudinal thermal comfort surveys were distributed online three times daily (morning, mid-day and afternoon) for a continuous 2-week

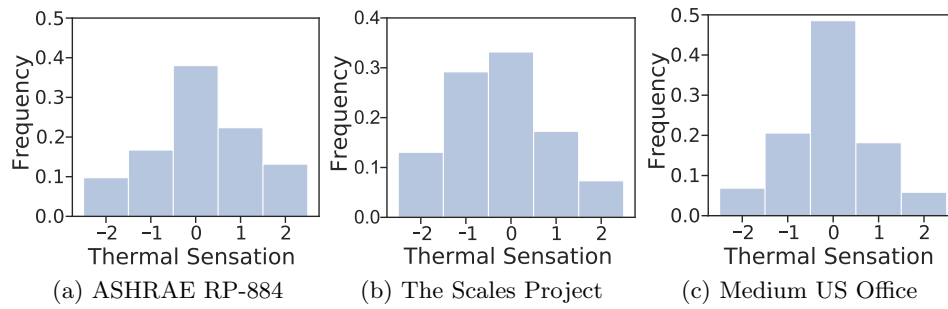


Figure 6.3: Distribution of thermal sensation over different datasets

period in each of the four project seasons between July 2012 and August 2013. Data types varied from daily surveys to sensor data including but not limited to the indoor air temperature, air velocity, relative humidity, CO₂ concentration and illuminance. This public dataset is used as the target dataset in the research.

The locations of all cities with data used in the study are displayed in Figure 6.2. The red points represent the 26 cities in the ASHRAE RP-884 database, the blue points indicate the 57 cities in the Scales Project dataset, and the green point indicates Philadelphia in the Medium US Office dataset. In this chapter, we aim to learn the knowledge from data in cities indicated by red points and blue points to benefit one building in Philadelphia (green point).

Table 6.1 shows the basic information for the three datasets. The first two datasets have different building types (HVAC, naturally ventilated and mixed ventilated) while there is only one HVAC building in the Medium US Office dataset. Since the ASHRAE and Scales datasets include different climate zones all over the world, they have wider indoor air temperature ranges than the Friends Center building in the Medium US Office ($17.9^{\circ}C$ - $27.8^{\circ}C$). Different from the first two datasets, the Medium US Office dataset has much smaller groups of participants. Besides, the ranges of the indoor relative humidity (Indoor RH), indoor air velocity (Indoor AV), metabolic rate (MR), and clothing level (Clo) in the Medium US Office dataset are smaller than those in the ASHRAE dataset.

6.3.2 Preliminary Analytics

Figure 6.3 shows the distribution of thermal sensation for the ASHRAE RP-884 dataset, the Scales Project and the Medium US Office dataset. Since the numbers of instances of the sensation scale for +3 (Hot) and -3 (cold) are far less than those of the other instances in both data sets, we merged +3 (hot) and +2 (warm) into one class, and -3 (cold) and -2 (cool) into one class. In the office environment, indoor environmental factors such as temperature are generally maintained at a relatively comfortable level (17.9°C - 27.8°C in the Medium US dataset), and people can also choose to adjust their clothing level and behaviour (e.g., open the heater vents and have hot drinks) if they are too cold or too hot.

Although the regression model is effective in many time-series problems [298, 12], the classification method still dominates the thermal comfort area. Therefore, in this chapter, we choose classifiers rather than regressors for effective thermal comfort prediction. Besides, based on the previous discussion, thermal sensation scales are classified into 5 categories (i.e., cold or cool, slightly cool, neutral, slightly warm, hot or warm).

The above three datasets have similar thermal sensation distributions, and occupants feel neutral towards the thermal environment most of the time. We can observe that there are more responses for feeling slightly warm or cool than feeling warm/cool or hot/cold, which accords with our thermal comfort feelings in daily life. Meanwhile, the thermal sensation distributions in the ASHRAE dataset and the Scales Project dataset are more uniform than the distribution

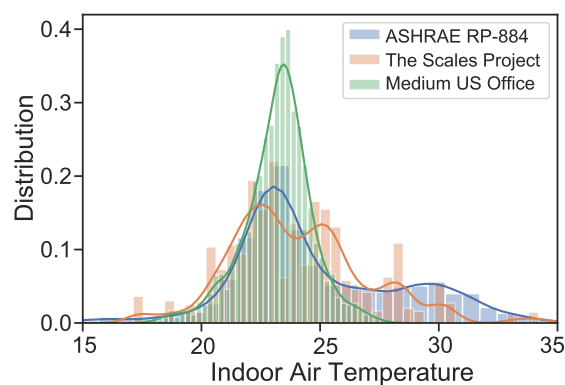


Figure 6.4: Distribution of the indoor air temperature over different domains

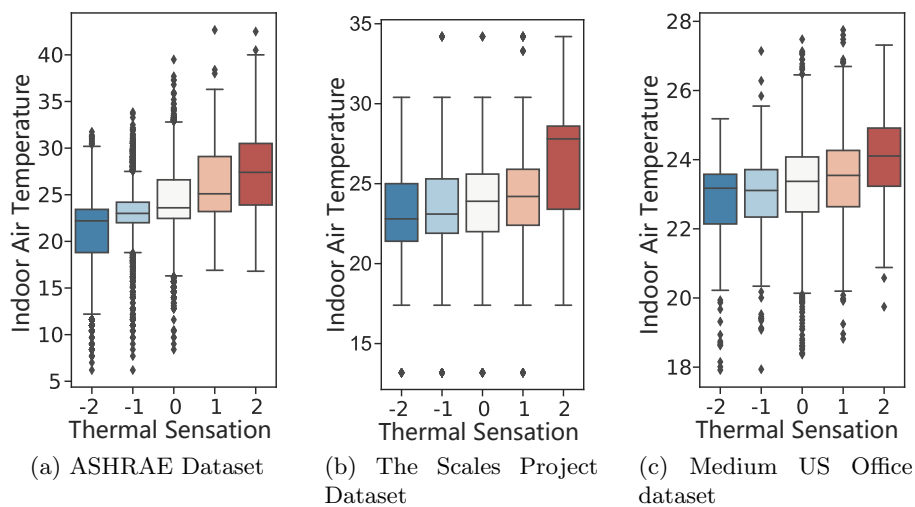


Figure 6.5: Boxplots of thermal sensation and the indoor temperature

of the Medium US Office dataset. This is because the ASHRAE dataset and the Scales Project dataset consist of a variety of data from different climate zones all over the world while the Medium US Office dataset includes data from only one building.

Indoor air temperature is one of the most significant factors affecting occupants' thermal feelings. Figure 6.4 shows the distribution of the indoor air temperature for the three datasets. Most temperature values range from 22°C – 24°C . However, there are also some differences between these three distributions. The ASHRAE and the Scales Project datasets have higher indoor air temperatures because some thermal sensation responses are from hot climate areas. In contrast, in the Medium US Office dataset, the indoor temperature distribution seems to be centred at approximately 20°C to 27°C .

From Figure 6.5, we can see the relationship between the indoor air temperature and thermal sensation scale. Usually, a higher indoor air temperature indicates a higher thermal sensation scale for all three datasets. Interestingly, in the Medium US Office dataset, the average indoor air temperature for feeling cold or cool is slightly higher than that for feeling slightly cool. This phenomenon may be due to there being too few subjects (24 participants in total) in the Medium US Office dataset. Additionally, the other factors such as the relative humidity, age, gender, and outdoor weather will affect the thermal sensation. This is the reason

why we use as many features as possible to build a more accurate and robust thermal comfort prediction model.

From the above analysis, there are observable differences between the ASHRAE, Scales Project and Medium US Office datasets. One of the reasons is that buildings in these three datasets are located in various climate zones, where climate variability can lead to a different working environment, occupant cognition and behaviour, therefore affecting occupants' thermal sensation in different buildings. Considering that the three datasets share many similarities in occupant thermal comfort and that the number of instances in the target dataset is very limited, we explore occupants' thermal comfort by learning from multiple buildings in the same climate zone with similar climate conditions. We will then introduce the proposed thermal comfort modelling in Section 6.4.4.

6.4 Methodology

6.4.1 Problem Definition

To learn sensor data from multiple datasets for thermal comfort modelling, some notations need to be defined. Firstly, we give the definition of a 'task' and a 'domain'. A domain \mathcal{D} can be represented as $\mathcal{D} = \{\mathcal{X}, P(X)\}$, which contains two parts: the feature space \mathcal{X} and the marginal probability distribution $P(X)$, where $X = \{x_1, x_2, \dots, x_n\} \in \mathcal{X}$. The task \mathcal{T} can be represented as $\mathcal{T} = \{y, f(\cdot)\}$, which contains two components: the label space y and a target prediction function $f(\cdot)$. $f(\cdot)$ can not be observed but can be learnt from the training data, which could also be considered as a conditional function $P(y|x)$.

In the context of traditional machine learning, the common assumption is that the training and test data share exactly the same feature space and data distribution [328]. However, once the new task \mathcal{T} arrives and its data distribution $P(X)$ is different from the previous task, the new model must be rebuilt from the beginning using the current data. This method requires extra effort and is very expensive in most cases. Compared with traditional machine learning methods, transfer learning can tolerate differences in data distribution and utilise knowledge from other sources to target tasks.

Table 6.2: Selected features in the Medium US Office dataset

Category	Data Source	Feature Name	Description	Units
<i>Indoor</i>	HOBO Datalogger (15 mins)	Indoor_AT	Indoor temperature	$^{\circ}C$
		Indoor_RH	Indoor relative humidity	%
		Indoor_AV	Indoor air velocity	m/s
		Indoor_AMRT	Indoor radiant temperature	$^{\circ}C$
<i>Outdoor</i>	Weather Analytics (15 mins)	Outdoor_AT	Outdoor temperature	$^{\circ}C$
		Outdoor_RH	Outdoor humidity	%
<i>Personal</i>	Daily Survey (3 times/day)	CL	Clothing insulation	<i>clo</i>
		MR	Metabolic rate	<i>Met</i>
	Background Survey (once)	Age	Participants' age	<i>Years</i>
		Gender	Participants' gender	<i>N/A</i>

In this chapter, we transfer the knowledge from the source domain (RP-884 and the Scales Project datasets) to benefit thermal comfort prediction in the target domain (Medium US Office dataset). Although both domains have different features, they share several common features such as the indoor air temperature, indoor relative humidity, indoor air velocity, indoor mean radiant temperature, clothing level, metabolic rate, and occupants' age and gender. Therefore, predicting thermal comfort falls under *transductive transfer learning* [329], which can be formally defined as follows: given a source domain \mathcal{D}_s and the corresponding learning task \mathcal{T}_s , a target domain \mathcal{D}_t and the corresponding learning task \mathcal{T}_t , we aim to improve the performance of the prediction function $f(\cdot)_t$ in \mathcal{T}_t by discovering the knowledge from \mathcal{D}_s and \mathcal{T}_s , where $\mathcal{D}_s \neq \mathcal{D}_t$ and $\mathcal{T}_s = \mathcal{T}_t$.

Figure 6.6 shows the thermal comfort transfer learning system in which we could use the transfer learning method to learn knowledge from the source datasets and benefit the target dataset in a specified city.

6.4.2 Feature Selection

Human thermal sensation is influenced by a variety of factors such as time factors [307], personal information [309], environmental changes [310], and culture [311]. In this chapter, several features are chosen for thermal comfort transfer learning based on the following criteria: (1) the features were commonly studied in previous thermal comfort research and (2) the features are easy calculate or collect by using passive sensing or self-reported responses. In summation,

we divide the features into three broad categories: indoor environmental features, outdoor environmental features and personal features. Table 6.2 displays the selected features in the Medium US Office dataset.

Indoor Environmental Features. The indoor environment affects occupants' thermal comfort directly, and we adopt the following basic indoor environmental features derived from Fanger's PMV model [305] for thermal comfort prediction: the air temperature, mean radiant temperature, air velocity and relative humidity. The air temperature is the average temperature of the air surrounding the occupant at a location and time. The radiant temperature indicates the radiant heat transferred from a surface, and the mean radiant temperature is affected by the emissivity and temperature of the surrounding surfaces, viewing angles, etc. The air velocity is the average speed of air with respect to the direction and time. The relative humidity is the ratio of the amount of water vapour in the air to the amount of water vapour that the air can hold at a specified pressure and temperature.

Outdoor Environmental Features. Outdoor weather conditions can have physiological effects on individuals thermal perception and clothing preference in different seasons [306, 330]. For instance, in summer people tend to choose lightweight clothing, which will influence their indoor thermal comfort. The most popular measurements of the outdoor environment include the outdoor air temperature and outdoor humidity, which will also be adopted in this research.

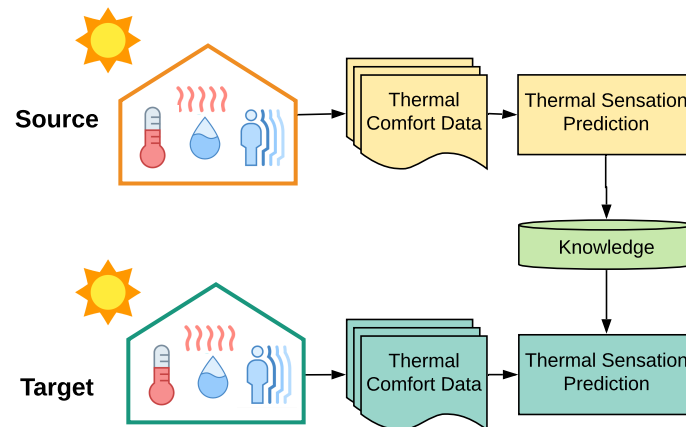


Figure 6.6: Thermal comfort transfer learning system

Personal Features. Studying personal features is crucial for effective thermal comfort modelling because thermal sensation is a subjective measurement and different individuals perceive the same environment differently. In this chapter, we selected the following personal features: clothing insulation, metabolic rate, age and gender. Clothing insulation has a major impact on the thermal comfort level because it affects heat loss and thus the heat balance. Previous research shows the relationship between age and thermal sensation [331, 311]. Besides, Sami et al. [332] found a significant gender difference in thermal comfort: females tend to prefer a higher room temperature than males and feel both uncomfortably hot and uncomfortably cold more often than males. Hence, gender and age are considered to be the features for thermal comfort modelling.

The features in a source domain can be considered as a subset in the target domain. The ASHRAE dataset shares eight features with the Medium US Office dataset while the Scales Project dataset only shares six features with the target dataset. Although there are various other features in these three datasets such as occupant behaviour data (e.g., adjusting heaters/curtains/ thermostats) and background survey data (e.g., acceptable temperature), we simplify the thermal comfort prediction and therefore do not show the other features.

6.4.3 Imbalance Class Distribution

As the thermal sensation scale has 5-point values, we regard thermal comfort prediction as a classification task. Fig. 6.3 shows the distributions of the ASHRAE RP-884, Scales Project and Medium US Office datasets. It is clear that the three distributions are imbalanced, and the number of thermal sensation instances for -1 (cool) to 1 (warm) far exceeds the number of other instances. To train a fair classifier, we must address this class imbalance issue in thermal comfort data. Take the binary classification as an example. If class M is 95% and class N is 5% in the dataset, we can simply reach an accuracy of 95% by predicting class M each time, which provides a useless classifier for our purpose. In this chapter, we assume that the survey responses are ‘correct’. Although there may be some biases (e.g., rating bias, anchoring bias, and social desirability bias) in self-reported data, we will not discuss them in this chapter.

To address an imbalanced dataset, oversampling and undersampling are efficient techniques

to adjust the class distribution of the data set. Under-sampling (e.g., *clustering*, *edited nearest neighbours* [333] and *Tomek links* [334]) can balance the dataset by reducing the size of the majority class. However, undersampling methods are usually used when we have sufficient data. Oversampling (e.g., the *synthetic minority oversampling technique* [335] and *adaptive synthetic sampling* [336]) aims to balance the dataset by increasing the number of minority classes, which can be applied when the data are insufficient.

Generative Adversarial Networks (GANs) have been successfully applied in various fields to learn the probability distribution of a dataset and synthesize samples from the distribution [337, 338]. A GAN uses a generator G to capture the underlying data distribution of a dataset and a discriminator D to estimate the probability that a given sample comes from the original dataset rather than being created by G . Some techniques such as the *TableGAN* [339] and *TabularGAN* [340] have been proposed to handle the imbalance of tabular data. In particular, Quintana et al. [341] used the *TabularGAN* to synthesize a small thermal comfort dataset. They found that when the amount of synthesized data is no larger than the amount of real data, the thermal comfort dataset can achieve similar performance to the real samples.

In the thermal comfort classification problem, labelled thermal comfort responses are usually few. Therefore, in this chapter, we synthesize survey responses to handle the imbalance of thermal sensation classes. The *TabularGAN*¹ is used in this research to generate tabular data based on the generative adversarial network. It can learn each column's marginal distribution by minimizing the KL divergence, which is more suitable for thermal comfort classification problems compared with other methods such as the *TableGAN*, edited nearest neighbours [333], SMOTE [335], etc. The reason why we did not adopt the *TableGAN* is that it optimizes the prediction accuracy on synthetic data by minimizing the cross entropy loss while *TabularGAN* focuses more on the marginal distribution. The *TabularGAN* learns each column's marginal distribution by minimizing the KL divergence, which is more suitable for the thermal comfort classification problem.

¹Python package for TabularGAN:<https://pypi.org/project/tgan/>

6.4.4 Thermal Comfort Modelling

Traditional algorithms for thermal comfort modelling is isolated and occurs purely based on specific buildings in the same climate zone. No thermal comfort knowledge is retained that can be transferred from one thermal comfort model to another. Recently, the transfer learning technique has been intensively studied in different applications [36, 322]. It aims to leverage knowledge from source tasks and then apply them to the target task. There are various transfer learning techniques that can be roughly grouped into three categories: *inductive transfer learning*, *unsupervised transfer learning* and *transductive transfer learning* [342]. Inductive transfer learning [343] aims to improve performance on the current task after having learned a different but related skill or concept on a previous task. Unsupervised transfer learning [344] focuses on solving unsupervised learning tasks in the target domain such as dimensionality reduction, clustering, and density. Transductive transfer learning aims to utilize the knowledge from the source domain to improve the performance of the prediction task in the target domain.

Transductive transfer learning can exploit the different levels of information captured from different layers in the neural network. Generally, layers close to the input data capture specific characteristics in the dataset while deeper layers capture information more relevant to the tasks (e.g., object types in image recognition and thermal sensation labels in thermal comfort prediction). The Medium US Office dataset, as described in Section 6.3.1, differs in cities and climate zones from the ASHRAE dataset and the Scales Project dataset. In different climate zones, there are various factors possibly contributing to thermal comfort, e.g., climate characteristics and occupants' perceptions and tolerance. This motivates us to investigate transfer learning between the ASHRAE/Scales Project datasets and Medium US office dataset in climate variability, which is close to the layers near the input.

We assume that climate variability affects the lower-level neural network only. Therefore, these layers need to be adapted to better represent the Friends Center office building in the target dataset. This can be regarded as retaining the knowledge of higher-level mappings from the source dataset. Hence, we retain the last hidden layer of the models on the ASHRAE and Scales Project datasets as shown in Figure 6.7. Then, the thermal comfort neural network will be retrained with the Medium US Office dataset until convergence to find the optimal

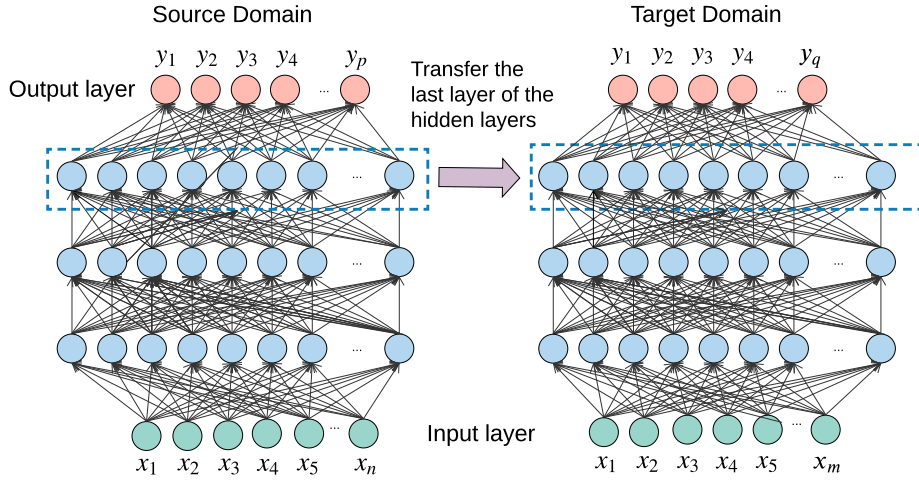


Figure 6.7: The architecture for thermal comfort transfer learning

parameters for the lower hidden layers.

6.5 Experiment

In this section, we conduct experiments on the proposed thermal comfort transfer learning models and compare the performance with the state-of-the-art techniques and different configurations. We address the two research questions: *Can we predict occupants' thermal comfort accurately by learning from multiple buildings in the same climate zone when we do not have enough data? If so, which features contribute the most to effective thermal comfort transfer learning?* Specifically, we explore how the numbers of hidden layers and sample size of the training set in the target building affect thermal comfort transfer learning performance.

6.5.1 Experimental Setup

In our research, the source domain (ASHRAE RP-884 dataset and the Scales Project dataset) and the target domain (Medium US Office dataset) share some common features, which include four indoor environmental variables (air temperature, indoor relative humidity, mean radiant temperature, and indoor air velocity), two environmental variables (air temperature and humidity) and two personal variables (age and gender). In addition, the ASHRAE RP-

884 and Medium US Office datasets share two other personal variables (clothing insulation and metabolic rate). The shared features make it possible to transfer knowledge to the target domain from the source domain.

Preprocessing. As discussed in Section 6.3.2, we first merge the minority classes and reclassify the thermal sensation into five categories. Then, we standardize the features by scaling them to unity variance for better classification performance. Considering that the thermal sensation classes are extremely imbalanced, in order to train a meaningful classifier, the *TabularGAN* [340] technique is applied for synthesizing the samples in all the classes except the majority class in the training set. Here, 50% of the samples in each class were synthesized while ensuring that the number of samples per category did not exceed the number of samples in the majority class.

Taking the Medium US Office dataset as an example, there are 2497 instances in the original dataset. After removing the null values and categorizing the thermal sensation responses, there were 1090 ‘neutral’ responses, 462 ‘slightly cool’ responses, 408 ‘slightly warm’ responses, 154 ‘cool or cold’ responses and 131 ‘warm or hot’ responses. After synthesizing the data using the *TabularGAN*, there were 981 ‘neutral’ responses, 624 ‘slightly cool’ responses, 551 ‘slightly warm’ responses, 208 ‘cool or cold’ responses and 177 ‘warm or hot’ responses in the training set (90% of the dataset).

Architecture. In this research, we choose the multilayer perception (MLP) neural network as the classifier for the source domain and target domain. Each neural network consists of two hidden layers with 64 neurons in each layer. The Relu function is used as the activation function in hidden layers. Then, the softmax function is applied to the output layer as the activation function. We train the classifier with the categorical cross-entropy loss function and the Adam optimizer with learning rate = 0.001. The batch size is set to 200 and the max epoch has been set to 500. Besides, the fixed random seed is chosen for dataset shuffling and training.

Evaluation. Similar to previous thermal comfort studies [320, 37, 309, 38], the accuracy and weighted F1-score are chosen as the performance metrics. Accuracy reflects the overall performance of the thermal comfort model. Since our priority goal is to correctly predict the thermal sensation for as many occupants as possible to achieve overall thermal comfort/energy

savings in the building, accuracy is used as the main evaluation metric in this problem. We also adopt the weighted F1-score as the best metric to assess the accuracy of capturing performance across imbalanced classes. The F1-score considers both false positives and false negatives to strike a balance between the precision and recall. The ‘weighted-average’ calculates the metrics for each class and finds their average weighted by the number of true instances for each class. Compared with the ‘macro-average’ method, the ‘weighted average’ considers class imbalances. The weighted F1-score is helpful for evaluating thermal sensation classifiers as it considers all imbalanced classes. That is, it evaluates the classifiers for different user groups with different thermal sensation levels instead of all occupants globally.

Baselines. For the baseline, three different categories of baselines are selected for comparison with our proposed method: random guessing, the PMV model and multiple traditional machine learning models. Random guessing generates the sample from the distribution of thermal comfort and regards it as a predicted value. Similar random baselines have been widely used in previous thermal comfort studies such as [345, 35]. The PMV model is the most prevalent thermal comfort model worldwide. In the experiment, we will only use the four indoor environmental variables, the metabolic rate and clothing insulation to calculate the PMV score p_s according to the formula in [346] for the target dataset. Then, the thermal sensation class $\mathcal{C}(p_s)$ is calculated using Equation 6.1.

$$\mathcal{C}(p_s) = \begin{cases} -2, & \text{if } p_s \leq -1.5 \\ -1, & \text{if } -1.5 < p_s \leq -0.5 \\ 0, & \text{if } -0.5 < p_s \leq 0.5 \\ 1, & \text{if } 0.5 < p_s \leq 1.5 \\ 2, & \text{if } p_s \geq 1.5 \end{cases} \quad (6.1)$$

For the multiple traditional machine learning models, we choose K-nearest Neighbors [347], Naive Bayes [348], Support Vector Machine (with Linear, RBF and Polynomial kernel) [349], Decision Tree [350], Random Forest [294], AdaBoost [351] as baselines. Naive Bayes [348] from Bayes family methods is chosen due to its fast speed and working well with high dimensions. Support Vector Machine [349] technique is efficient for handling high dimensional spaces. Different from algorithms like SVM, AdaBoost [351] is fast, simple and easy to use

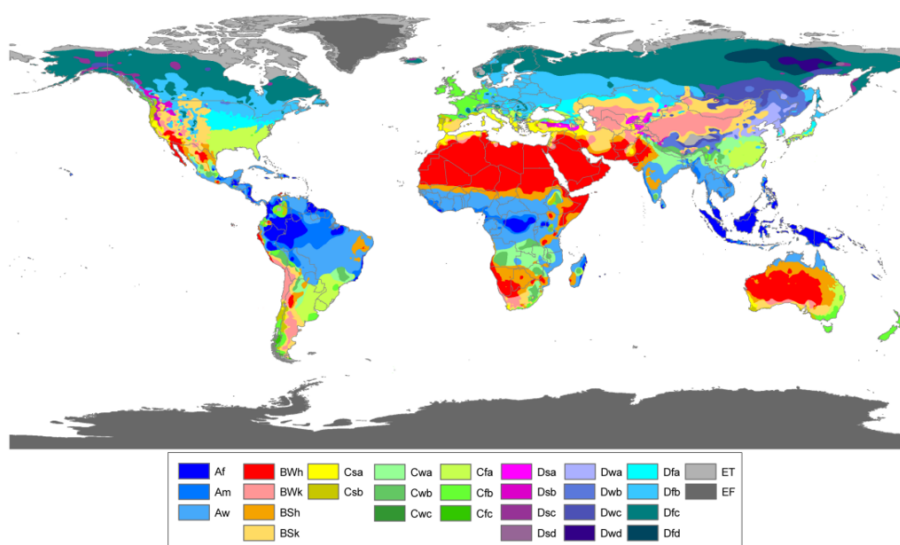


Figure 6.8: ‘Köppen World Map High Resolution’ by Peel, M. C. et al. [1], licenced under Creative Commons Attribution-Share Alike 3.0 Unported [2], Desaturated from original

with less need for tuning parameters. K-nearest Neighbors [347] is a simple method storing all available instances and classifying data instances according to a similarity measure, which has been widely used in the pattern recognition and statistical prediction area. Random Forest [294] is an ensemble learning method for classification operated by building multiple decision trees. It can cope with high-dimensional features and judge the feature importance.

Compared to the PMV model using six factors for thermal comfort prediction, the multiple machine learning algorithms use ten features as input features (see Table 6.2). Besides, all three of the above baselines build a thermal comfort classification model using the Medium US dataset.

Cross-validation. We apply the *k-fold cross-validation* [352] ($k = 10$) method for effective thermal comfort classification. The advantage of 10-fold cross-validation is that it estimates the unbiased generalization performance of the thermal comfort prediction model. In the experiment, the data from the target domain (US Medium Office dataset) are randomly partitioned into 10 folds, each fold serves as the testing data iteratively, and the remaining 9 folds are used as the training data. The cross-validation process is repeated 10 times, and the prediction results (accuracy and weighted F1-score) are averaged to produce a single estimation.

Climate zone divisions. We adopt the Köppen climate classification updated by Peel et al. [1], which is one of the most widely used climate classification systems in the world. As shown in Figure 6.8, the Köppen climate classification divides climates into five main climate zones: A (tropical), B (dry), C (temperate), D (continental), and E (polar). Each large climate zone is then divided into several small subzones based on temperature patterns and seasonal precipitation. All specific climates are assigned a main group of climate zones (the first letter).

In our study, the target domain (Philadelphia in the US) belongs to the 'temperate' climate zone. In the source domain, the Scales Project dataset includes 8225 instances from 57 cities in total, and 5411 instances from 32 cities (e.g., Yokohama, Sydney, and Cambridge) were located in the 'temperate' climate zone. The ASHRAE RP-884 database consists of 25623 thermal comfort responses from 26 cities in total, where 12 cities (e.g., Berkeley, Athens, and Chester) [353] are situated in the same climate zone as Philadelphia.

We run the proposed TL-MLP model and TL-MLP-C* models with the ASHRAE database and the Scales Project database as the source domain and the Medium US Office dataset as the target domain. In particular, for both proposed models, we only use the data from buildings with HVAC systems in all datasets. For the TL-MLPC* model, we use the data from the buildings with HVAC systems in the same climate zone as the source domain and the Friends Center building as the target domain.

Table 6.3: Classification of the ASHRAE RP-884 database for HVAC buildings according to climates

Climate	Number of cities	Instances
<i>Tropical</i>	5 (Townsville, Jakarta, Darwin, Bangkok, and Singapore)	3826
<i>Dry</i>	6 (Honolulu, Kalgoorlie-Boulder, Karachi, Quettar, Multan, and Peshawar)	3290
<i>Temperate</i>	12 (Brisbane, Melbourne, Athens, South Wales, Sydney, San Francisco, Merseyside, San Ramon, Antioch, Auburn, Oxford, and Saidu)	3512
<i>Continental</i>	3 (Ottawa, Montreal, and Grand Rapids)	2808
<i>All</i>	26	13436

Table 6.4: Prediction performance for different algorithms on the target dataset

Algorithm	Accuracy (%)	F1-score (%)
PMV	33.35 (2.40)	32.45 (2.35)
Random	27.23 (1.30)	29.30 (1.40)
KNN	41.43 (2.95)	41.93 (2.85)
SVM (Linear)	29.44 (5.19)	30.92 (4.84)
SVM (RBF)	37.93 (3.86)	40.91 (4.04)
SVM (Poly)	34.02 (4.59)	37.66 (5.15)
Decision Tree	43.33 (4.94)	43.34 (4.87)
Random Forest	51.41 (3.03)	52.93 (3.69)
Naive Bayes	40.43 (4.10)	39.40 (3.97)
AdaBoost	42.94 (3.22)	42.41 (3.94)
MLP	50.35 (3.81)	50.67 (4.51)
TL-MLP	50.76 (4.31)	53.60 (4.43)
TL-MLP-C*	54.50 (4.16)	55.12 (4.14)

Besides, we classify the HVAC buildings in the ASHRAE RP-884 database into different climates (see Table 6.3). The table shows that in the ASHRAE RP-884 database, there are 13436 observations from buildings with HVAC systems in total and 3512 such observations in the 'temperate' climate zone. Since the Scales Project dataset recorded the Köppen climate and HVAC status information during the data collection, after calculation, there were 4621 observations from buildings with HVAC systems in total and 3245 observations collected from buildings with HVAC systems located in the 'temperate' climate zone.

6.5.2 Overall Prediction Result

Table 6.4 shows the performance of different thermal comfort modelling algorithms. We use all ten features described in Section 6.4.2 on most algorithms except for the PMV model. From Table 6.4, we can see that the PMV model performs better than only the random baseline and SVM classifiers (kernel = 'Linear') in accuracy. The F1-score of the linear SVM is still higher than that of the PMV model. This may be because we use more features in machine learning classifiers while the PMV model only has six factors. We will discuss the prediction performance with different feature sets later in Section 6.5.3.

Table 6.4 shows that the random forest algorithm performs the best on all metrics compared

with the PMV model, random baseline and other data-driven models including eight traditional machine learning classifiers. This may be because the random forest is usually regarded as the best classification algorithm for small datasets [35] and has been proven to have the highest prediction accuracy for thermal sensation [318].

Most importantly, we find that the TL-MLP has a higher F1-score for thermal comfort classification than other machine learning methods without using transfer learning. Although the TL-MLP has better prediction performance than the MLP on all metrics, the prediction accuracy of the TL-MLP is slightly lower than that of the random forest. The potential reason is that the TL-MLP transfers knowledge from all HVAC buildings in the world regardless of the different climate zones, leading to lower prediction accuracy than that of the random forest. Excitingly, the TL-MLP-C* model works better than all of the state-of-the-art algorithms on both metrics (accuracy and F1-score), indicating the effectiveness of the proposed approach.

To further investigate how the proposed TL-MLP-C* improves the prediction performance compared to the MLP, we show the confusion matrixes for the MLP and TL-MLP-C* in Figure 6.9. The figure shows that the MLP model can predict label 0 (neutral) with the highest probability of 0.61, which is similar to the 0.62 of the TL-MLP-C*. However, it still has high chances to misclassify labels 1 (slightly warm) to 0 (neutral). Instead, the transfer learning-based thermal comfort model TL-MLP-C* can predict labels more accurately than the traditional MLP model, especially for the minority classes (-2, -1, 1). It can predict 67% of the label -2 (cool or cold) instances and 40% of the label 1 (slightly warm) instances correctly and achieves an average accuracy of 54.50% for all classes from -2 to 2.

In summary, our proposed transfer learning-based models (TL-MLP and TL-MLP-C*) achieve remarkable performance for thermal comfort prediction compared with the random baseline, traditional PMV model and data-driven algorithms without transfer learning. In particular, the TL-MLP-C* model outperforms the state-of-the-art algorithms on both metrics (accuracy and F1-score). Furthermore, the improved prediction performance of the TL-MLP-C* is significant compared to that of the standard MLP model.

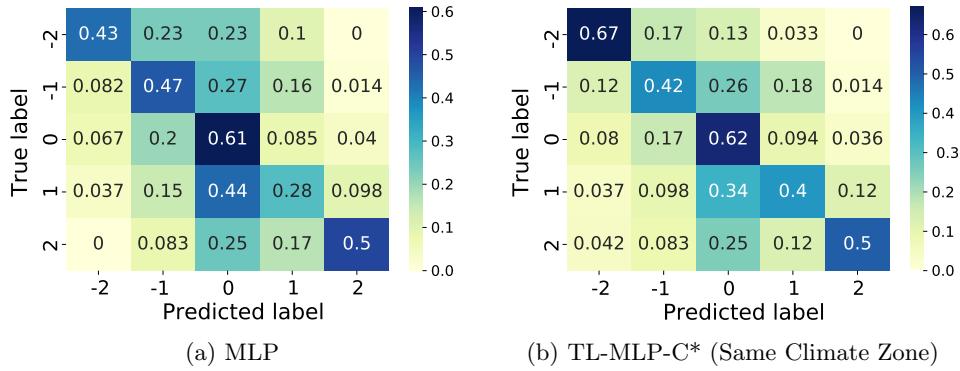


Figure 6.9: Confusion matrix on the target domain

6.5.3 Impact of Different Feature Combinations

We will now explore how accurately the proposed TL-MLP and TL-MLP-C* models work when only a set of features is available. Usually, indoor sensors are inexpensive and unobtrusive and have been installed in many buildings with HVAC systems. However, some features may be unavailable due to factors such as privacy, costs, etc. For instance, occupants may not be willing to report their age, which reflects their metabolism level and influence their thermal comfort feelings. Besides, it is somewhat inconvenient to install outdoor weather stations outside a building to capture outdoor environmental changes (e.g., outdoor air temperature and humidity) more accurately than the official weather stations used for local weather forecasting.

Hence, in the experiment, we will divide our features into 3 different sets \mathcal{X}_a , \mathcal{X}_b , and \mathcal{X}_c based on PMV factors, personal factors and outdoor environmental factors, respectively; and then compare the different sets and explore which features contribute the most to effective thermal comfort transfer learning. The feature sets are as follows:

- \mathcal{X}_a : Six basic factors introduced in the PMV model: indoor air temperature, indoor air velocity, indoor relative humidity, indoor radiant temperature, clothing insulation and metabolic rate. This is the most common feature set for thermal comfort modelling used in previous studies [35].
- \mathcal{X}_b : Six factors from \mathcal{X}_a and two personal factors: age and gender. Personal factors such as gender and age can be easily collected through background surveys.

Table 6.5: Prediction performance for different feature sets on the target dataset

Sets	Algorithm	Accuracy (%)	F1-score (%)
\mathcal{X}_a	PMV	33.35	32.45
	Random Forest	34.77	34.92
	MLP	33.18	34.06
	TL-MLP	33.53	35.90
	TL-MLP-C*	33.98	39.32
\mathcal{X}_b	Random Forest	43.43	43.18
	MLP	42.96	45.31
	TL-MLP	44.10	45.88
	TL-MLP-C*	47.10	51.15
\mathcal{X}_c	Random Forest	51.41	52.93
	MLP	50.35	50.67
	TL-MLP	50.76	53.60
	TL-MLP-C*	54.50	55.12

- \mathcal{X}_c : Eight factors from \mathcal{X}_b and two outdoor environmental factors including the outdoor air temperature and outdoor relative humidity. The above two outdoor environmental features need to be accessed from the outdoor weather station near the target building.

For different feature sets, we use the same oversampling methods and fixed random seeds in neural network training. Table 6.5 shows the prediction performance for different feature sets on the target dataset. The random forest and MLP algorithms are chosen for comparison with the TL-MLP and TL-MLP-C* algorithms due to their relatively high performance, as shown in Table 6.4. For the \mathcal{X}_a , \mathcal{X}_b , and \mathcal{X}_c feature sets, we can observe that the performance of the TL-MLP and TL-MLP-C* models increases as the number of features increases. In addition, the TL-MLP-C* model has the highest accuracy and F1-score in each feature set.

For feature set \mathcal{X}_a , the PMV model works slightly better than the MLP model in accuracy but worse in F1-score. The random forest algorithm achieves the best performance in accuracy while TL-MLP-C* achieves the highest F1-score. With transfer learning from source datasets, the TL-MLP and TL-MLP-C* have similar prediction accuracies to the traditional PMV model. This shows that the advantages of the proposed TL-MLP and TL-MLP-C* models cannot be fully utilized when the number of features is limited.

In feature set \mathcal{X}_b , all data-driven models achieve better prediction performance than using

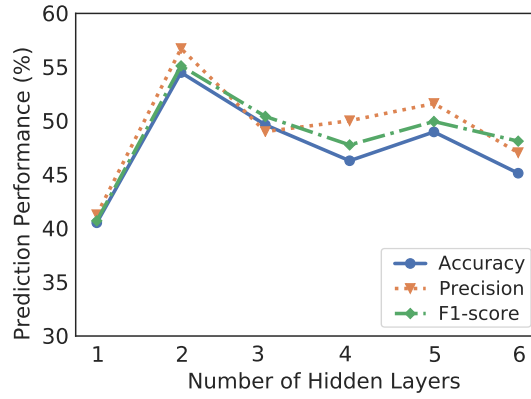


Figure 6.10: Prediction performance with different number of hidden layers

only feature set \mathcal{X}_a . This shows that personal information (age and gender) could effectively improve thermal comfort prediction. Moreover, the TL-MLP-C* model has the best prediction performance compared with the other methods in both metrics when considering personal factors.

In comparison to feature sets \mathcal{X}_a and \mathcal{X}_b , the random forest, MLP, TL-MLP and TL-MLP-C* work best among all metrics on the feature set \mathcal{X}_c . This proves that outdoor environmental changes can affect occupants' thermal sensation in HVAC buildings and shows the necessity to consider outdoor features for effective thermal comfort modelling.

6.5.4 Impact of the Number of Hidden Layers

We also conduct adaption experiments by using different numbers of hidden layers in the TL-MLP-C* model. Figure 6.10 shows the prediction accuracy and F1-score for TL-MLP-C* with different numbers of hidden layers. We can observe that the prediction performance is worst in all metrics with only one hidden layer. Since our proposed method transfers the last layer of the hidden layer, if we set only one hidden layer, the target dataset will have little contribution to the prediction model. When the number of hidden layers is set to 2, the proposed TL-MLP-C* model has the highest prediction performance in accuracy and F1-score. As the number of hidden layers continues to increase, the prediction performance tends to decrease, which may be due to the model being overfitted with more trainable parameters.

Finally, although our proposed TL-MLP-C* model has better thermal comfort prediction

performance than the state-of-the-art methods, the achieved accuracy (54.50%) is still not remarkably high. There are several potential reasons: (1) We adopt the *TabularGAN* to resample the minority classes for meaningful classification. Fifty percent of the instances in each class were synthesized while ensuring that the number of samples per category did not exceed the number of samples in the majority class. Although some previous works achieve slightly higher accuracy for thermal comfort prediction (e.g., 63.09% in [35] and 62% in [318]), they only assigned slightly higher weights to the instances in the minority classes, which cannot handle the class imbalance problem as well as our method. (2) Predicting thermal comfort is challenging since many factors affect occupants' thermal sensation (as discussed in Section 6.1). There may also be many response biases during the survey. Therefore, the classification accuracy in most previous research is also not good and rarely higher than 60%, even for personal thermal comfort modelling. (3) It could be better to regard the thermal comfort prediction as a regression problem instead of a classification problem. For example, classifying '-2' (cool) to '-1' (slightly cool) should be more acceptable than classifying '-2' (cool) to '+2' (warm). We will study the thermal comfort regression in future work.

6.6 Conclusion

A huge amount of sensor data has been generated in cities worldwide. Recently, utilising such data from multiple cities to benefit a target city has become a critical issue. In this research, we applied the idea of transfer learning to the thermal comfort area and proposed two transfer learning-based thermal comfort prediction models: TL-MLP and TL-MLP-C*. For the first time, we transferred the knowledge from similar thermal environments to a target building for effective thermal comfort modelling. Furthermore, we improved the prediction performance and built meaningful classifiers by using a GAN-based resampling method (i.e., *TabularGAN*) to imbalance the class distribution of occupants' thermal sensation.

By retaining the last hidden layer of the neural network from the source domain (ASHARE RP-884 and Scales Project datasets), we trained the thermal comfort model for the Friends Center building from the Medium US Office dataset and found the optimal parameter settings for lower hidden layers. Extensive experimental results showed that the proposed TL-MLP and

TL-MLP-C* models outperform the state-of-the-art algorithms for thermal comfort prediction. Interestingly, the most significant feature sets are identified for effective thermal comfort transfer learning.

This chapter shows the possibility of building thermal comfort models with limited data. The publicly available thermal comfort data from similar climate zones can be used to benefit the thermal comfort modelling in the target building. However, the current studies have some limitations that needed to be addressed in future research: (1) First, we only used the Friends Center Office as the target building which is located in the ‘temperate’ climate zone. The performance of transfer learning on more target buildings in the same or different climate zones should be explored in the future. (2) Although the proposed method can benefit the target building with a small amount of labelled data, the prediction model will achieve the best performance when at least six factors are provided. In real-world scenarios, if there are only several factors (lower than six) in the target building, our method can still work by setting the values of missing factors to the same distribution as the source domain, but the prediction performance for the target building will be affected. (3) We only investigated the MLP model because it is one of the most classical types of neural networks, which is suitable for classification with tabular datasets. More advanced transfer learning architectures can be explored in the future to find transferable representations between the source domain and target domain in future studies.

Chapter 7

Conclusion

Research into sensing and profiling human behaviours has been a popular topic over the past decades and continues to grow. With advancements in the IoT, a huge amount of sensing data can be generated, stored, processed and analysed from various devices within an acceptable time, which is revolutionary in the field of data science and ML.

The aim of this research was to build a human behaviour sensing framework with predictive capabilities in multiple real-world scenarios for both single domain modelling and domain adaptation. The core chapters of the thesis addressed key challenges related to data acquisition in natural settings, the representation of dynamic human behaviours and mental states, and the shortage of annotations in human studies. Five research questions were constructed to derive specific solutions to these research issues for different tasks, including inferring user engagement, seating behaviours, personality traits, response time and thermal comfort. The benefits of this study are extensive, ranging from helping individuals improve their level of self-awareness and adopt healthier lifestyles to assisting managers and policymakers in creating the right study/work environments to improve human wellbeing.

This thesis tackled several issues around sensing and profiling human behaviours and made the following contributions to the field:

- We publish the largest heterogeneous indoor environmental and affect sensing dataset for understanding various human behaviours and discussing the reliability of self-report

data as the ground truth in human behaviour studies.

- We integrate individual/group human behavioural dynamics with domain knowledge in multiple real-world scenarios (i.e. seating patterns, personality traits, notification response behaviours and thermal comfort).
- We incorporate sensing data from various environments and multiple sources (i.e. wearable, mobile and environmental sensing) to create a robust predictive model.
- We predict human behaviours of a domain using scarce data based on knowledge learned from available data from previously modelled domains.

The research presented in this thesis focused on profiling and modelling human behaviours using sensing technology in the wild. It explored different ML solutions for physiological and behavioural sensing data in multiple real-world scenarios. Some of the experiments were carried out on the self-collected dataset *En-Gage*, as introduced in Chapter 2, and some on other popular publicly released datasets. The proposed frameworks and predictive models in each chapter were consistently connected from the viewpoint of the research process for sensing human behaviours. The developed techniques reduced the rate of false predictions and were designed to assist developers, managers and policymakers to improve mental health, wellbeing and productivity of individuals using the prediction results.

7.1 Research Questions and Answers

RQ-1. How to capture and validate multidimensional human behaviours and states using heterogeneous sensors in the wild?

RQ-1 was addressed in Chapter 2. We presented the heterogeneous data collection using data from wearable sensors, environmental sensors and self-report responses. We conducted a field study at a private school in the suburbs of Melbourne, Australia in which we tracked 23 students and 6 teachers in a four-week cross-sectional study, using wearable sensors to log physiological data and self-report surveys to query the occupants' thermal comfort, learning engagement, emotions and seating behaviours. The released dataset is the largest and most

heterogeneous indoor environmental and affect sensing dataset and can be used to further analyse human behaviours and mental states, study peer effects on students and create comfortable indoor environments. In addition, we discussed the reliability of self-report data for human behaviours by studying the confidence level of responses and survey completion time. We found that the physiologically measured student engagement and perceived student engagement were not always consistent, which serves as a wake-up call for the emotional and mental sensing research, which usually regards self-report annotations as the ground truth for predicting human behaviours.

***RQ-2.** How to model and predict people's emotional, cognitive and behavioural engagement using wearable and environmental sensor data?*

To address **RQ-2**, Chapter 3 explored the usefulness of wearable and environmental sensing data in modelling user engagement in the wild. Using the data collected in Chapter 2, a classroom sensing system *n-Gage* was proposed to detect students' in-class emotional, behavioural and cognitive engagement. In particular, we combined physiological signals, behavioural data and indoor environmental data to estimate changes in student engagement levels. Novel features were proposed to represent the physiological and physical synchrony between students, which proved to be useful for predicting student engagement. Comprehensive experiments were conducted to predict the multidimensional student engagement scores using LightGBM regressors. Experiment results showed that n-Gage achieved a high degree of accuracy for predicting student engagement. In addition, various factors were derived, and the most useful sensors were explored to differentiate between the dimensions of learning engagement.

***RQ-3.** How to explore the effects of individual and group behaviours (e.g. seating patterns) on people's perceived and physiologically measured engagement in different courses?*

Chapter 4 focused on tackling **RQ-3** and explored how individual and group-wise classroom seating experiences affect students' perceived engagement and physiologically measured engagement. In Chapter 3, we showed that student engagement could be inferred from their physiological sensing signals. Therefore, using the dataset from Chapter 2, we investigated the physiologically measured engagement by examining students' physiological arousal and synchrony. We identified statistically significant correlations between student seating behaviours

and students' perceived and physiologically measured engagement. Experimental result showed that students who sat close together were more likely to have similar learning engagement and had higher physiological synchrony than students who sat far apart.

***RQ-4.** How to utilise mobile sensing to profile personality traits and receptivity to interruptions among different user groups?*

In Chapter 5, **RQ-4** was addressed. We explored the use of unobtrusive mobile sensing for user behaviour profiling. Two real-world scenarios were considered for user behaviour prediction scenarios. The first scenario involved modelling users' mental characteristics (i.e. Big Five personality traits). Based on the proposed novel metrics (i.e. diversity, dispersion and regularity), some important features were extracted from mobile phone logs, call logs and accelerometer data to represent human activities. Experimental results showed that the predicted personality scores were close to the ground truth, with an observable reduction in errors in predicting the Big Five personality traits for both males and females. In the second scenario, we explored the effect of individuals' smartphone usage behaviors and moods on notification response times. An in-the-wild study was conducted with more than 18 participants over a five-week period. In total, we have collected 42,270 notifications, 3,553 self-report responses and more than 5,920 hours of physiological signals from Empatica E4 wristbands. We found a statistically significant correlation between response time and in-use apps. Extensive experiments showed that the proposed regression model achieved high predictive performance for notification response times. We also investigated how the mood-related features improve the predictive performance by utilising the self-report responses and physiological signals.

***RQ-5.** How to model aggregate behaviour (e.g., thermal comfort) from environmental sensing with limited annotations by transferring knowledge from multiple locations to another domain?*

RQ-5 was covered in Chapter 6. We aggregated occupant thermal comfort data from environmental sensors with limited annotations by transferring knowledge from multiple locations to a separate domain. We proposed two transfer learning-based thermal comfort prediction models: TL-MLP and TL-MLP-C*. For the first time, we transferred knowledge from similar thermal environments to a target building for effective thermal comfort modelling. In addition,

we improved the predictive performance and built meaningful classifiers by using a GAN-based resampling method (i.e. TabularGAN) to imbalance the class distribution of occupants' thermal sensation. Extensive experimental results on the popular publicly available datasets (i.e. ASHRAE RP-884, Scales Project and Medium US Office) showed that the performance of the proposed TL-MLP-C* model exceeded the performance of state-of-the-art methods in both accuracy and F1-score.

7.2 Future Research Directions

The proposed methods outperformed existing models in terms of predictive performance, heterogeneity of the sensing data and diversity of the real-world scenarios. However, some improvements could be considered in several areas: the proposed approaches, better predictive performance and real-world deployments. From this research, several directions for future research on human behaviour sensing using extensive sensor data and machine learning techniques have emerged. The most important aspects are data acquisition and processing, human behaviour itself, and models and evaluations [3].

Currently, sensor data generated from by sensors in the wild are heterogeneous in format and storage. Such datasets usually lacks descriptions and are ad hoc, making it difficult to share and reuse them. When researchers are able to analyse these datasets, they face challenges during data acquisition and processing. In the future, it would be useful to explore the possibility of creating more datasets with sufficient descriptions and structured, real-world data on human behaviour gathered from heterogeneous sensors, as this would benefit researchers in a variety of disciplines. In addition, self-report data are prone to subjectivity and responses bias, making it risky and inaccurate to use it as the ground truth for human psychological states. Therefore, more research should be done to investigate the reliability of self-report annotations as ground truth and explore ways to combine subjective self-report responses and objective physiological sensing data for more effective predictive models.

Human behaviours and states are usually affected by various factors (e.g. demographic diversity, social relationships, time and physical spaces) and have characteristics such as capriciousness, dynamics and multi-granularity [3]. Therefore, it is challenging to perceive and

analyse human behaviours and states due to the difficulty in quantifying influencing factors. In addition, human behaviours and states are usually complex and can be affected by interactions with other people, e.g. a group of students interact with each other during in-class activities, or an individual's mental state can be influenced by others due to their ability to empathise. The methods and approaches presented in this thesis may not be directly applicable to the above scenarios. Therefore, more complex human behaviours (e.g. group behaviours, abnormal behaviours and criminal behaviors) should be explored in various real-world situations in future studies.

A variety of models have been proposed for effectively profiling human behaviours and states. The transfer learning approach is a promising area when limited labels are available in human behavioural studies. However, transfer learning research is still in its infancy, and research needs to branch out in more directions, such as incremental learning and unsupervised domain adaptation without any labeling. In this area of research, it is difficult to evaluate and compare results relating to predicted human behaviours with results from prior studies because human behaviour varies between subjects in real-world scenarios. To establish standards for perceiving human behaviour, it is important to clarify the processing techniques, baselines and experimental settings in future research.

Finally, though sensing human behaviours has achieved good performance and is widely adopted, it is usually intrusive and has raised issues such as ethics, privacy, and deployment cost [3]. In general, the ethical evaluation of sensing technology is highly dependent on the application areas and contexts of use, and there are certain obligations to be respected [354]. For instance, anyone working in the area should abide by the ethics that govern human research and data privacy; they should uphold ethical values that make sensing technologies involving humans more likely to have positive effects and less likely to have negative effects; they should ensure the system does nothing that the users would object to, and let users understand what's going on. Additionally, researchers should work closely with non-experts to form a realistic assessment of the capabilities of the systems and the risks they might pose. In future research, we need to pay close attention to issues of ethics and privacy, as well as the gap between real-world system deployment and theoretical models, so as to help sensing technology better

improve human well-being.

Bibliography

- [1] Murray C Peel, Brian L Finlayson, and Thomas A McMahon. Updated world map of the köppen-geiger climate classification. 2007. [Cited on pages xv, 155, and 156]
- [2] Creative commons license deed. [Cited on pages xv and 155]
- [3] Zhiwen Yu and Zhu Wang. *Human behavior analysis: sensing and understanding*. Springer Nature, 2020. [Cited on pages 4, 7, 8, 168, and 169]
- [4] Maitri Vaghela and Kalyan Sasidhar. Analyzing the human behavior using pervasive sensing system. In *Proceedings of the 21st International Conference on Distributed Computing and Networking*, pages 1–5, 2020. [Cited on page 4]
- [5] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017. [Cited on page 4]
- [6] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Unobtrusive assessment of students’ emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–21, 2018. [Cited on pages 4, 7, 16, 17, 28, 35, 38, 39, 40, 43, 46, 47, 48, 49, 50, 51, 52, 53, 56, 57, 62, 67, 68, 72, 74, 75, 94, and 128]
- [7] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. Using students’ physiological synchrony to quantify the classroom emotional climate. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and*

- Ubiquitous Computing and Wearable Computers*, pages 698–701, 2018. [Cited on pages 4, 29, 31, 72, 74, 85, and 86]
- [8] John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In *International Conference on Ubiquitous Computing*, pages 243–260. Springer, 2006. [Cited on page 4]
- [9] Geetika Singla, Diane J Cook, and Maureen Schmitter-Edgecombe. Recognizing independent and joint activities among multiple residents in smart environments. *Journal of ambient intelligence and humanized computing*, 1(1):57–63, 2010. [Cited on page 4]
- [10] Tim van Kasteren and Ben Krose. Bayesian activity recognition in residence for elders. In *2007 3rd IET International Conference on Intelligent Environments*, pages 209–212. IET, 2007. [Cited on page 4]
- [11] Mark Weiser. The computer for the 21st century. *ACM SIGMOBILE mobile computing and communications review*, 3(3):3–11, 1999. [Cited on page 4]
- [12] Nan Gao, Wei Shao, and Flora D Salim. Predicting personality traits from physical activity intensity. *Computer*, 52(7):47–56, 2019. [Cited on pages 4, 8, 13, 16, 22, 35, 38, 50, 56, 57, 131, and 144]
- [13] Varun Mishra. From sensing to intervention for mental and behavioral health. 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 388–392, 2019. [Cited on page 4]
- [14] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, Jun Zhai, Klaus David, and Flora D Salim. Transfer learning for thermal comfort prediction in multiple cities. *Building and Environment*, 195:107725, 2021. [Cited on pages 4, 8, 14, and 50]
- [15] Mohammad Saiedur Rahaman, Jonathan Liono, Yongli Ren, Jeffrey Chan, Shaw Kudo, Tim Rawling, and Flora D Salim. An ambient-physical system to infer concentration in

- open-plan workplace. *IEEE Internet of Things Journal*, 2020. [Cited on pages 4, 16, 31, 50, and 136]
- [16] Sayma Akther, Nazir Saleheen, Mithun Saha, Vivek Shetty, and Santosh Kumar. mTeeth: Identifying brushing teeth surfaces using wrist-worn inertial sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–25, 2021. [Cited on page 5]
- [17] Asrat Gedefa Yadessa and Ayodeji Olalekan Salau. Low cost sensor based hand washing solution for covid-19 prevention. In *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 93–97. IEEE, 2021. [Cited on page 5]
- [18] Edith Talina Luhanga, Akpa Akpro Elder Hippocrate, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. Towards proactive food diaries: A participatory design study. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1263–1266, 2016. [Cited on page 5]
- [19] Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Food diary coaching chatbot. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1676–1680, 2018. [Cited on page 5]
- [20] Rodrigo De Oliveira, Mauro Cherubini, and Nuria Oliver. MoviPill: improving medication compliance for elders using a mobile persuasive social game. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 251–260, 2010. [Cited on page 5]
- [21] Margreet Riphagen, Marco van Hout, and G Gootjes. Learning tomorrow: visualising student and staff’s daily activities and reflect on it. *ICERIE2013*, 2013. [Cited on page 6]
- [22] Keith M Diaz, David J Krupka, Melinda J Chang, James Peacock, Yao Ma, Jeff Goldsmith, Joseph E Schwartz, and Karina W Davidson. Fitbit®: An accurate and reli-

- able device for wireless physical activity tracking. *International Journal of Cardiology*, 185:138, 2015. [Cited on page 6]
- [23] Paul Dempsey. The teardown: Apple watch. *Engineering & Technology*, 10(6):88–89, 2015. [Cited on page 6]
- [24] Heli Koskimäki, Hannu Kinnunen, Teemu Kurppa, and Juha Röning. How do we sleep: a case study of sleep duration and quality using data from oura ring. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 714–717, 2018. [Cited on page 6]
- [25] Grant Hernandez, Orlando Arias, Daniel Buentello, and Yier Jin. Smart nest thermostat: A smart spy in your home. *Black Hat USA*, (2015), 2014. [Cited on page 6]
- [26] Rick Wash, Emilee Rader, and Chris Fennell. Can people self-report security accurately? agreement between self-report and behavioral measures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2228–2232, 2017. [Cited on pages 6 and 19]
- [27] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J Oedegaard, and Jim Tørresen. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51:1–26, 2018. [Cited on page 7]
- [28] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–19, 2019. [Cited on pages 7, 16, 29, 38, 46, 47, 68, 73, 74, 85, 86, 88, and 89]
- [29] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. micro-Stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant

- mothers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–22, 2019. [Cited on pages 7, 16, 17, 18, 53, and 61]
- [30] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. n-Gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–26, 2020. [Cited on pages 7, 8, 13, 16, 17, 18, 28, 31, 68, 74, 75, 86, 89, 94, 128, and 136]
- [31] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014. [Cited on pages 7, 17, 18, 35, 38, 56, and 57]
- [32] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–30, 2018. [Cited on pages 7, 16, 17, and 18]
- [33] Jia Yan, Fengchun Tian, Qinghua He, Yue Shen, Shan Xu, Jingwei Feng, and Kadri Chaibou. Feature extraction from sensor data for detection of wound pathogen based on electronic nose. *Sensors and Materials*, 24(2):57–73, 2012. [Cited on page 8]
- [34] Wenyuan Dai, Ou Jin, Gui-Rong Xue, Qiang Yang, and Yong Yu. Eigentransfer: A unified framework for transfer learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 193–200. ACM, 2009. [Cited on pages 8 and 140]
- [35] Weizheng Hu, Yong Luo, Zongqing Lu, and Yonggang Wen. Heterogeneous transfer learning for thermal comfort modeling. In *Proceedings of the 6th ACM International*

- Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 61–70, 2019. [Cited on pages 8, 141, 154, 158, 159, and 162]
- [36] Emil Laftchiev Annamalai Natarajan. A transfer active learning framework to predict thermal comfort. *International Journal of Prognostics and Health Management*, 10:13, 2019. [Cited on pages 8, 140, and 151]
- [37] Juhi Ranjan and James Scott. ThermalSense: determining dynamic thermal comfort preferences using thermographic imaging. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1212–1222. ACM, 2016. [Cited on pages 8, 139, 141, and 153]
- [38] Ali Ghahramani, Guillermo Castro, Simin Ahmadi Karvigh, and Burcin Becerik-Gerber. Towards unsupervised learning of thermal comfort using infrared thermography. *Applied Energy*, 211:41–49, 2018. [Cited on pages 8, 139, 141, and 153]
- [39] Nan Gao, Max Marschall, Jane Burry, Simon Watkins, and Flora D Salim. Understanding occupants’ behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data*, 9(1):1–16, 2022. [Cited on pages 12 and 17]
- [40] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, and Flora D Salim. Investigating the reliability of self-report data in the wild: The quest for ground truth. 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 237–242, 2021. [Cited on pages 12, 31, 68, and 133]
- [41] Judith S Heinisch, Nan Gao, Christoph Anderson, Shohreh Deldari, Klaus David, and Flora D Salim. Investigating the effects of mood and usage behaviour on notification response time. *arXiv preprint arXiv:2207.03405*, 2022. [Cited on page 14]
- [42] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K D’Mello, Munmun De Choudhury, Gregory D Abowd, and Thomas Plötz. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):75, 2019. [Cited on pages 16, 18, and 38]

- [43] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018. **[Cited on pages 16 and 18]**
- [44] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–33, 2019. **[Cited on page 16]**
- [45] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. EngageMon: Multi-modal engagement sensing for mobile games. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–27, 2018. **[Cited on pages 16, 18, 38, 40, 43, 47, 50, 53, 57, 68, and 74]**
- [46] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–21, 2018. **[Cited on page 16]**
- [47] Jorn Bakker, Mykola Pechenizkiy, and Natalia Sidorova. What’s your current stress level? detection of stress patterns from GSR sensor data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 573–580. IEEE, 2011. **[Cited on pages 16, 18, 35, 38, 46, 48, and 53]**
- [48] Marco Sarchiapone, Carla Gramaglia, Miriam Iosue, Vladimir Carli, Laura Mandelli, Alessandro Serretti, Debora Marangon, and Patrizia Zeppegno. The association between electrodermal activity (EDA), depression and suicidal behaviour: A systematic review and narrative synthesis. *BMC Psychiatry*, 18(1):22, 2018. **[Cited on pages 16, 35, and 48]**

- [49] 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. [Cited on page 17]
- [50] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011. [Cited on page 17]
- [51] Stefan Schneegass, Bastian Pfleging, Nora Broy, Frederik Heinrich, and Albrecht Schmidt. A data set of real world driving to assess driver workload. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 150–157, 2013. [Cited on page 17]
- [52] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015. [Cited on page 17]
- [53] Javad Birjandtalab, Diana Cogan, Maziyar Baran Pouyan, and Mehrdad Nourani. A non-ecg biosignals dataset for assessment and visualization of neurological status. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, pages 110–114. IEEE, 2016. [Cited on page 17]
- [54] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2016. [Cited on page 17]
- [55] 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. [Cited on page 17]

- [56] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 400–408, 2018. **[Cited on page 17]**
- [57] Martin Gjoreski, Tine Kolenik, Timotej Knez, Mitja Luštrek, Matjaž Gams, Hristijan Gjoreski, and Veljko Pejović. Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences*, 10(11):3843, 2020. **[Cited on page 17]**
- [58] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leon-tios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):1–16, 2020. **[Cited on page 17]**
- [59] Andreas Möller, Matthias Kranz, Barbara Schmid, Luis Roalter, and Stefan Diewald. Investigating self-reporting behavior in long-term studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2931–2940, 2013. **[Cited on pages 17, 19, and 68]**
- [60] Nan Gao, Max Marschall, Jane Burry, Simon Watkins, and Flora Salim. In-Gauge and En-Gage datasets. *figshare* <https://doi.org/10.25439/rmt.14578908>, May 2021. **[Cited on page 17]**
- [61] Bing Dong, Yapan Liu, Wei Mu, Zixin Jiang, Pratik Pandey, Tianzhen Hong, Bjarne Olesen, Thomas Lawrence, Zheng O’Neil, Clinton Andrews, et al. A global building occupant behavior database. *Scientific Data*, 9(1):1–15, 2022. **[Cited on page 17]**
- [62] Kathryn A Fuller, Nilushi S Karunaratne, Som Naidu, Betty Exintaris, Jennifer L Short, Michael D Wolcott, Scott Singleton, and Paul J White. Development of a self-report instrument for measuring in-class student engagement reveals that pretending to engage is a significant unrecognized problem. *PLoS ONE*, 13(10):e0205828, 2018. **[Cited on pages 18, 23, 35, 39, 40, 43, 57, and 71]**

- [63] Kurt Kroenke, Robert L. Spitzer, Janet B.W. Williams, and Bernd Löwe. An ultra-brief screening scale for anxiety and depression: The PHQ-4. *Psychosomatics*, 50(6):613–621, 2009. [Cited on page 18]
- [64] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1):163–173, 2009. [Cited on page 18]
- [65] Stephen Barclay, Chris Todd, Ilora Finlay, Gunn Grande, and Penny Wyatt. Not another questionnaire! maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of gps. *Family Practice*, 19(1):105–111, 2002. [Cited on page 19]
- [66] Robin C. Jackson, Hayley Barton, Kelly J. Ashford, and Bruce Abernethy. Steppers and signal detection: Response sensitivity and bias in the differentiation of genuine and deceptive football actions. *Frontiers in Psychology*, 9:2043, 2018. [Cited on page 19]
- [67] Eric van Sonderen, Robbert Sanderma, and James C. Coyne. Ineffectiveness of reverse wording of questionnaire items: Let’s learn from cows in the rain. *PLoS ONE*, 8(7), 2013. [Cited on page 19]
- [68] Lee Anna Clark and David Watson. Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12):1412–1427, 2019. [Cited on page 19]
- [69] Nathan W. Hudson, Ivana Anusic, Richard E. Lucas, and M. Brent Donnellan. Comparing the reliability and validity of global self-report measures of subjective well-being with experiential day reconstruction measures. *Assessment*, 27(1):102–116, 2020. PMID: 29254354. [Cited on page 19]
- [70] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. Empatica e3—a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile*

- Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, pages 39–42. IEEE, 2014. [Cited on pages 22, 41, and 116]
- [71] ASHRAE Handbook-Fundamentals. American society of heating. *Refrigerating and Air-Conditioning Engineers*, 2009. [Cited on page 23]
- [72] John P Pollak, Phil Adams, and Geri Gay. Pam: A photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 725–734, 2011. [Cited on pages 23 and 43]
- [73] Neil Malhotra. Completion time and response order effects in web surveys. *Public Opinion Quarterly*, 72(5):914–934, 2008. [Cited on pages 26 and 63]
- [74] Daniel C Richardson, Nicole K Griffin, Lara Zaki, Auburn Stephenson, Jiachen Yan, Thomas Curry, Richard Noble, John Hogan, Jeremy I Skipper, and Joseph T Devlin. Engagement in video and audio narratives: Contrasting self-report and physiological measures. *Scientific Reports*, 10(1):1–8, 2020. [Cited on page 29]
- [75] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, 51(10):951–963, 2014. [Cited on page 29]
- [76] Richard V Palumbo, Marisa E Marraccini, Lisa L Weyandt, Oliver Wilder-Smith, Heather A McGee, Siwei Liu, and Matthew S Goodwin. Interpersonal autonomic physiology: A systematic review of the literature. *Personality and Social Psychology Review*, 21(2):99–141, 2017. [Cited on pages 29, 49, 74, 85, and 88]
- [77] Shohreh Deldari, Daniel V Smith, Amin Sadri, and Flora Salim. Espresso: Entropy and shape aware time-series segmentation for processing heterogeneous sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–24, 2020. [Cited on page 31]

- [78] Wei Shao, Flora D Salim, Andy Song, and Athman Bouguettaya. Clustering big spatiotemporal-interval data. *IEEE Transactions on Big Data*, 2(3):190–203, 2016. [Cited on page 31]
- [79] Wei Shao, Flora D Salim, Jeffrey Chan, Kai Qin, Jiaman Ma, and Bradley Feest. Onlineairtrajclus: An online aircraft trajectory clustering for tarmac situation awareness. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 192–201. IEEE, 2019. [Cited on page 31]
- [80] Flora D Salim, Bing Dong, Mohamed Ouf, Qi Wang, Ilaria Pigliautile, Xuyuan Kang, Tianzhen Hong, Wenbo Wu, Yapan Liu, Shakila Khan Rumi, et al. Modelling urban-scale occupant behaviour, mobility, and energy in buildings: A survey. *Building and Environment*, 183:106964, 2020. [Cited on page 31]
- [81] Salvatore Carlucci, Marilena De Simone, Steven K Firth, Mikkel B Kjærgaard, Romana Markovic, Mohammad Saiedur Rahaman, Masab Khalid Annaqeeb, Silvia Biandrate, Anooshmita Das, Jakub Wladyslaw Dziedzic, et al. Modeling occupant behavior in buildings. *Building and Environment*, 174:106768, 2020. [Cited on page 31]
- [82] Mikkel B Kjærgaard, Omid Ardakanian, Salvatore Carlucci, Bing Dong, Steven K Firth, Nan Gao, Gesche Margarethe Huebner, Ardeshir Mahdavi, Mohammad Saiedur Rahaman, Flora D Salim, et al. Current practices and infrastructure for open data based research on occupant-centric design and operation of buildings. *Building and Environment*, 177:106848, 2020. [Cited on page 31]
- [83] Bruce Sacerdote. Experimental and quasi-experimental analysis of peer effects: two steps forward? *Annual Review of Economics*, 6(1):253–272, 2014. [Cited on page 31]
- [84] Mohammad Saiedur Rahaman, Shaw Kudo, Tim Rawling, Yongli Ren, and Flora D Salim. Seating preference analysis for hybrid workplaces. *arXiv preprint arXiv:2007.15807*, 2020. [Cited on page 31]

- [85] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, Kaixin Ji, and Flora D Salim. Individual and group-wise classroom seating experience: Effects on student engagement in different courses. *ArXiv Preprint ArXiv:2112.12342*, 2021. **[Cited on page 31]**
- [86] R Jisung Park, Joshua Goodman, Michael Hurwitz, and Jonathan Smith. Heat and learning. *American Economic Journal: Economic Policy*, 12(2):306–39, 2020. **[Cited on page 31]**
- [87] Irvan B Arief-Ang, Flora D Salim, and Margaret Hamilton. DA-HOC: semi-supervised domain adaptation for room occupancy prediction using CO2 sensor data. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, pages 1–10, 2017. **[Cited on page 31]**
- [88] Irvan B Arief-Ang, Margaret Hamilton, and Flora D Salim. RUP: Large room utilisation prediction with carbon dioxide sensor. *Pervasive and Mobile Computing*, 46:49–72, 2018. **[Cited on page 31]**
- [89] Irvan B Arief-Ang, Margaret Hamilton, and Flora D Salim. A scalable room occupancy prediction with transferable time series decomposition of CO2 sensor data. *ACM Transactions on Sensor Networks (TOSN)*, 14(3-4):1–28, 2018. **[Cited on page 31]**
- [90] James E Groccia. What is student engagement? *New Directions for Teaching and Learning*, 2018(154):11–20, 2018. **[Cited on page 34]**
- [91] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1):59–109, 2004. **[Cited on pages 34, 53, 62, 67, 70, and 75]**
- [92] Jennifer A Fredricks and Wendy McColskey. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of Research on Student Engagement*, pages 763–782. Springer, 2012. **[Cited on pages 34, 37, 53, 67, 70, and 75]**

- [93] Jonathan Martin and Amada Torres. What is student engagement and why is it important. *Retrieved May, 4:2018*, 2016. **[Cited on page 34]**
- [94] Helen M Marks. Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American educational research journal*, 37(1):153–184, 2000. **[Cited on pages 34 and 62]**
- [95] National Research Council et al. *Engaging Schools: Fostering High School Students' Motivation to Learn*. National Academies Press, 2003. **[Cited on page 34]**
- [96] Jeremy D Finn, Gina M Pannozzo, and Kristin E Voelkl. Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *The Elementary School Journal*, 95(5):421–434, 1995. **[Cited on pages 34, 67, and 70]**
- [97] Jeremy D Finn and Donald A Rock. Academic success among students at risk for school failure. *Journal of applied psychology*, 82(2):221, 1997. **[Cited on page 34]**
- [98] Jeremy D Finn. Withdrawing from school. *Review of Educational Research*, 59(2):117–142, 1989. **[Cited on page 34]**
- [99] Lyn Corno and Ellen B Mandinach. The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist*, 18(2):88–108, 1983. **[Cited on pages 34, 67, and 70]**
- [100] Paul R Pintrich and Elisabeth V De Groot. Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, 82(1):33, 1990. **[Cited on pages 34 and 70]**
- [101] KA Moore and L Lippman. Conceptualizing and measuring indicators of positive development: What do children need to flourish, 2005. **[Cited on pages 34, 43, and 70]**
- [102] Ellen A Skinner, Thomas A Kindermann, and Carrie J Furrer. A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's

- behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*, 69(3):493–525, 2009. [Cited on pages 34, 43, and 70]
- [103] Hugo D Critchley. Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2):132–142, 2002. [Cited on pages 35 and 56]
- [104] Wolfram Boucsein. *Electrodermal activity*. Springer Science & Business Media, 2012. [Cited on pages 35, 46, 47, 56, 68, 86, and 91]
- [105] Filipe Canento, Ana Fred, Hugo Silva, Hugo Gamboa, and André Lourenço. Multimodal biosignal sensor data handling for emotion recognition. In *SENSORS, 2011 IEEE*, pages 647–650. IEEE, 2011. [Cited on pages 35 and 48]
- [106] Javier Hernandez, Ivan Riobo, Agata Rozga, Gregory D Abowd, and Rosalind W Picard. Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 307–317, 2014. [Cited on pages 35, 38, 46, 47, 48, 51, 56, 68, 74, and 86]
- [107] Celine Latulipe, Erin A Carroll, and Danielle Lottridge. Love, hate, arousal and engagement: Exploring audience responses to performing arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1845–1854. ACM, 2011. [Cited on pages 35, 48, and 56]
- [108] Richard Pfanzer and W McMullen. Galvanic skin response and the polygraph. *BIOPAC Systems, Inc. Retrieved*, 5, 2013. [Cited on page 35]
- [109] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D’Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2016. [Cited on pages 35, 38, 39, 56, and 74]

- [110] Jorina von Zimmermann, Staci Vicary, Matthias Sperling, Guido Orgs, and Daniel C Richardson. The choreography of group affiliation. *Topics in Cognitive Science*, 10(1):80–94, 2018. [Cited on page 35]
- [111] Jamie A Ward, Daniel Richardson, Guido Orgs, Kelly Hunter, and Antonia Hamilton. Sensing interpersonal synchrony between actors and autistic children in theatre using wrist-worn accelerometers. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pages 148–155. ACM, 2018. [Cited on pages 35, 38, 49, 56, 57, 68, and 74]
- [112] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. Measuring cognitive and psychological engagement: Validation of the student engagement instrument. *Journal of school psychology*, 44(5):427–445, 2006. [Cited on page 37]
- [113] David J Shernoff and Jennifer A Schmidt. Further evidence of an engagement - achievement paradox among us high school students. *Journal of Youth and Adolescence*, 37(5):564–580, 2008. [Cited on page 37]
- [114] David J Shernoff, Mihaly Csikszentmihalyi, Barbara Schneider, and Elisa Steele Shernoff. Student engagement in high school classrooms from the perspective of flow theory. In *Applications of flow in human development and education*, pages 475–494. Springer, 2014. [Cited on page 37]
- [115] Ellen Skinner, Carrie Furrer, Gwen Marchand, and Thomas Kindermann. Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of educational psychology*, 100(4):765, 2008. [Cited on pages 37 and 93]
- [116] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):71, 2019. [Cited on pages 38 and 39]
- [117] Karen S McNeal, Jacob M Spry, Ritayan Mitra, and Jamie L Tipton. Measuring student engagement, knowledge, and perceptions of climate change in an introductory environ-

- mental geology course. *Journal of Geoscience Education*, 62(4):655–667, 2014. [Cited on pages 38, 39, and 40]
- [118] Chen Wang and Pablo Cesar. Physiological measurement on students’ engagement in a distributed learning environment. *PhyCS*, 10:0005229101490156, 2015. [Cited on pages 38, 39, 40, and 74]
- [119] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R Brockmole, and Sidney K D’Mello. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29(4):821–867, 2019. [Cited on pages 38 and 39]
- [120] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):141, 2018. [Cited on pages 38, 51, 52, 96, 99, 105, 111, and 128]
- [121] World Health Organization et al. Housing, energy and thermal comfort: a review of 10 countries within the who european region. 2007. [Cited on pages 42, 50, and 56]
- [122] Han Jiang, Matthew Iandoli, Steven Van Dessel, Shichao Liu, and Jacob Whitehill. Measuring students’ thermal comfort and its impact on learning. *Educational Data Mining*, 2019. [Cited on pages 42, 50, and 56]
- [123] Ulla Haverinen-Shaughnessy, DJ Moschandreas, and RJ Shaughnessy. Association between substandard classroom ventilation rates and students’ academic achievement. *Indoor Air*, 21(2):121–131, 2011. [Cited on pages 42 and 56]
- [124] Usha Satish, Mark J Mendell, Krishnamurthy Shekhar, Toshifumi Hotchi, Douglas Sullivan, Siegfried Streufert, and William J Fisk. Is CO₂ an indoor pollutant? direct effects of low-to-moderate CO₂ concentrations on human decision-making performance. *Environmental health perspectives*, 120(12):1671–1677, 2012. [Cited on pages 42 and 56]

- [125] Wargocki Pawel, Alí Porras-Salazar José, and William P. Bahnfleth. Quantitative relationships between classroom CO2 concentration and learning in elementary schools. *8th AIVC Conference "Ventilating healthy low-energy buildings"*, 2017. [Cited on page 42]
- [126] Amin Sadri, Yongli Ren, and Flora D Salim. Information gain-based metric for recognizing transitions in human activities. *Pervasive and Mobile Computing*, 38:92–109, 2017. [Cited on page 45]
- [127] Shohreh Deldari, Jonathan Liono, Flora D Salim, and Daniel V Smith. Inferring work routines and behavior deviations with life-logging sensor data. 2019. [Cited on page 45]
- [128] John T Cacioppo, Louis G Tassinary, and Gary Berntson. *Handbook of psychophysiology*. Cambridge University Press, 2007. [Cited on pages 47 and 86]
- [129] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Scilingo, and Luca Citi. cvxEDA: a convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, pages 1–1, 2016. [Cited on pages 47, 86, and 121]
- [130] Wendy Berry Mendes. Assessing autonomic nervous system activity. *Methods in social neuroscience*, pages 118–147, 2009. [Cited on page 48]
- [131] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009. [Cited on page 49]
- [132] Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008. [Cited on page 49]
- [133] Nutan D Ahuja, Amit K Agarwal, Ninad M Mahajan, Naresh H Mehta, and Hatim N Kapadia. GSR and HRV: Its application in clinical diagnosis. In *16th IEEE Sympo-*

- sium Computer-Based Medical Systems, 2003. Proceedings.*, pages 279–283. IEEE, 2003. [Cited on page 49]
- [134] Bradley M Appelhans and Linda J Luecken. Heart rate variability as an index of regulated emotional responding. *Review of general psychology*, 10(3):229–240, 2006. [Cited on page 49]
- [135] Kwang-Ho Choi, Junbeom Kim, O Sang Kwon, Min Ji Kim, Yeon Hee Ryu, and Ji-Eun Park. Is heart rate variability (HRV) an adequate tool for evaluating human emotions?—a focus on the use of the international affective picture system (iaps). *Psychiatry research*, 251:192–196, 2017. [Cited on page 49]
- [136] Peter Nickel and Friedhelm Nachreiner. Sensitivity and diagnosticity of the 0.1-hz component of heart rate variability as an indicator of mental workload. *Human factors*, 45(4):575–590, 2003. [Cited on page 49]
- [137] Antonio Luque-Casado, Mikel Zabala, Esther Morales, Manuel Mateo-March, and Daniel Sanabria. Cognitive performance and heart rate variability: the influence of fitness level. *PloS one*, 8(2):e56935, 2013. [Cited on page 49]
- [138] Paul van Gent, Haneen Farah, Nicole van Nes, and Bart van Arem. HeartPy: A novel heart rate algorithm for the analysis of noisy signals. *Transportation research part F: traffic psychology and behaviour*, 66:368–378, 2019. [Cited on page 49]
- [139] Bruce Mehler, Bryan Reimer, and Ying Wang. A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. *Proceedings of the 6th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design : driving assessment 2011*, 2011. [Cited on page 49]
- [140] A John Camm, Marek Malik, J Thomas Bigger, Günter Breithardt, Sergio Cerutti, Richard J Cohen, Philippe Coumel, Ernest L Fallen, Harold L Kennedy, RE Kleiger, et al. Heart rate variability: Standards of measurement, physiological interpretation

- and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. 1996. **[Cited on page 49]**
- [141] Fred Shaffer and JP Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in public health*, 5:258, 2017. **[Cited on page 49]**
- [142] H Nagendra, Vinod Kumar, and Shaktidev Mukherjee. Cognitive behavior evaluation based on physiological parameters among young healthy subjects with yoga as intervention. *Computational and mathematical methods in medicine*, 2015, 2015. **[Cited on page 49]**
- [143] Richard J De Dear and Gail S Brager. Thermal comfort in naturally ventilated buildings: Revisions to ASHRAE standard 55. *Energy and Buildings*, 34(6):549–561, 2002. **[Cited on pages 50 and 60]**
- [144] Kane. What are safe levels of CO and CO2 in rooms?, 2020. Accessed 2020-07-08. **[Cited on pages 50 and 58]**
- [145] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017. **[Cited on page 51]**
- [146] Wei Shao, Arian Prabowo, Sichen Zhao, Siyu Tan, Piotr Koniusz, Jeffrey Chan, Xinhong Hei, Bradley Feest, and Flora D Salim. Flight delay prediction using airport situational awareness map. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 432–435, 2019. **[Cited on page 51]**
- [147] Jane Elith, John R Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008. **[Cited on page 51]**
- [148] Andreas C Müller, Sarah Guido, et al. *Introduction to machine learning with python: A guide for data scientists.* ” O’Reilly Media, Inc.”, 2016. **[Cited on pages 51 and 128]**

- [149] Trevor S Wiens, Brenda C Dale, Mark S Boyce, and G Peter Kershaw. Three way k-fold cross-validation of resource selection functions. *Ecological Modelling*, 212(3-4):244–255, 2008. [Cited on page 51]
- [150] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001. [Cited on page 51]
- [151] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012. [Cited on pages 52 and 127]
- [152] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014. [Cited on page 52]
- [153] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(5):B231–B244, 2007. [Cited on page 53]
- [154] Manoranjan Pal and Premananda Bharati. *Applications of regression techniques*. Springer, 2019. [Cited on page 55]
- [155] John L Andreassi. *Psychophysiology: Human behavior and physiological response*. Psychology Press, 2010. [Cited on page 56]
- [156] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D’Mello. Automated physiological-based detection of mind wandering during learning. In *International Conference on Intelligent Tutoring Systems*, pages 55–60. Springer, 2014. [Cited on page 56]
- [157] Katherine A Herborn, James L Graves, Paul Jerem, Neil P Evans, Ruedi Nager, Dominic J McCafferty, and Dorothy EF McKeegan. Skin temperature reveals the intensity of acute stress. *Physiology & behavior*, 152:225–230, 2015. [Cited on page 56]

- [158] Asma Ahmad Farhan, Krishna Pattipati, Bing Wang, and Peter Luh. Predicting individual thermal comfort using machine learning algorithms. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 708–713. IEEE, 2015. **[Cited on page 61]**
- [159] Kimberly D Tanner. Structure matters: Twenty-one teaching strategies to promote student engagement and cultivate classroom equity. *CBE—Life Sciences Education*, 12(3):322–331, 2013. **[Cited on page 62]**
- [160] John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007. **[Cited on page 62]**
- [161] Deborah Stipek. Good instruction is motivating. In *Development of achievement motivation*, pages 309–332. Elsevier, 2002. **[Cited on page 62]**
- [162] Steven Cantrell and Thomas J Kane. Ensuring fair and reliable measures of effective teaching: Culminating findings from the met project’s three-year study. *MET Project Research Paper*, 2013. **[Cited on page 62]**
- [163] James Patrick Connell, Margaret Beale Spencer, and J Lawrence Aber. Educational risk and resilience in african-american youth: Context, self, action, and outcomes in school. *Child development*, 65(2):493–506, 1994. **[Cited on page 62]**
- [164] Verónica Rivera-Pelayo, Valentin Zacharias, Lars Müller, and Simone Braun. Applying quantified self approaches to support reflective learning. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 111–114, 2012. **[Cited on page 62]**
- [165] Rebecca Eynon. The quantified self for learning: Critical questions for education, 2015. **[Cited on pages 62 and 63]**
- [166] Paul Black and Dylan Wiliam. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21(1):5, 2009. **[Cited on page 62]**

- [167] Kimberly E Arnold, Brandon Karcher, Casey V Wright, and James McKay. Student empowerment, awareness, and self-regulation through a quantified-self student tool. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 526–527, 2017. [Cited on page 63]
- [168] Marguerite Cronk. Using gamification to increase student engagement and participation in class discussion. In *EdMedia+ Innovate Learning*, pages 311–315. Association for the Advancement of Computing in Education (AACE), 2012. [Cited on page 63]
- [169] Patrick Buckley and Elaine Doyle. Gamification and student motivation. *Interactive learning environments*, 24(6):1162–1175, 2016. [Cited on page 63]
- [170] Amanda Careena Fernandes, Jinyan Huang, and Vince Rinaldo. Does where a student sits really matter? the impact of seating locations on student classroom learning. *International Journal of Applied Educational Studies*, 10(1), 2011. [Cited on pages 67, 71, and 73]
- [171] Jeffrey M Burda and Charles I Brooks. College classroom seating position and changes in achievement motivation over a semester. *Psychological Reports*, 78(1):331–336, 1996. [Cited on pages 67, 71, and 84]
- [172] Mary Ellen Benedict and John Hoag. Seating location in large lectures: Are seating preferences or location related to course performance? *The Journal of Economic Education*, 35(3):215–231, 2004. [Cited on pages 67 and 80]
- [173] Pepper Anne Grimm. *Teacher perceptions on flexible seating in the classroom: effects on student engagement and student achievement*. PhD thesis, William Woods University, 2020. [Cited on pages 67 and 71]
- [174] Gyanendra Prasad Joshi, Sudan Jha, Seongsoo Cho, Changho Seo, Le Son, and Thong Pham. Influence of multimedia and seating location in academic engagement and grade performance of students. *Computer Applications in Engineering Education*, 28, 12 2019. [Cited on pages 67, 68, 71, 72, 73, 75, 76, and 84]

- [175] Albert Bandura and David C McClelland. *Social learning theory*, volume 1. Englewood cliffs Prentice Hall, 1977. **[Cited on pages 67 and 73]**
- [176] Rick D Axelson and Arend Flick. Defining student engagement. *Change: The magazine of higher learning*, 43(1):38–43, 2010. **[Cited on pages 67 and 70]**
- [177] Vicki Trowler. Student engagement literature review. *The higher education academy*, 11(1):1–15, 2010. **[Cited on pages 67 and 70]**
- [178] David J Shernoff, Alexander J Sannella, Roberta Y Schorr, Lina Sanchez-Wall, Erik A Ruzek, Suparna Sinha, and Denise M Bressler. Separate worlds: The influence of seating location on student engagement, classroom experience, and performance in the large university lecture hall. *Journal of Environmental Psychology*, 49:55–64, 2017. **[Cited on pages 67, 68, 72, 73, 75, 76, and 82]**
- [179] Franklin D Becker, Robert Sommer, Joan Bee, and Bart Oxley. College classroom ecology. *Sociometry*, pages 514–525, 1973. **[Cited on page 67]**
- [180] William B Holliman and Howard N Anderson. Proximity and student density as ecological variables in a college classroom. *Teaching of Psychology*, 13(4):200–203, 1986. **[Cited on pages 67 and 68]**
- [181] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, 2008. **[Cited on page 68]**
- [182] Wolfram Boucsein, Don C Fowles, Sverre Grimnes, Gershon Ben-Shakhar, Walton T Roth, Michael E Dawson, and Diane L Filion. Publication recommendations for electrodermal measurements. *Psychophysiology*, 49(8):1017–1034, 2012. **[Cited on page 68]**
- [183] Per Brodal. *The central nervous system: structure and function*. oxford university Press, 2004. **[Cited on page 68]**
- [184] Neil R Carlson and Neil R Carlson. *Physiology of behavior*. Pearson Boston, 2007. **[Cited on page 68]**

- [185] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. **[Cited on pages 68 and 126]**
- [186] Justin Storbeck and Gerald L Clore. Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and personality psychology compass*, 2(5):1824–1843, 2008. **[Cited on page 68]**
- [187] Nilgun Turkileri, David T Field, Judi A Ellis, and Michiko Sakaki. Emotional arousal enhances the impact of long-term memory in attention. *Journal of Cognitive Psychology*, 33(2):119–132, 2021. **[Cited on page 68]**
- [188] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. A critique of electrodermal activity practices at CHI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021. **[Cited on pages 68, 89, and 94]**
- [189] Michael D Meeks, Tami L Knotts, Karen D James, Felice Williams, John A Vassar, and Amy Oakes Wren. The impact of seating location and seating type on student performance. *Education Sciences*, 3(4):375–386, 2013. **[Cited on page 68]**
- [190] Shaun R Harper. Institutional seriousness concerning black male student engagement: Necessary conditions and collaborative partnerships. *Student Engagement in Higher Education: Theoretical Perspectives and Practical Approaches for Diverse Populations*, pages 137–156, 2009. **[Cited on page 70]**
- [191] Benjamin S Bloom et al. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, 20:24, 1956. **[Cited on page 70]**
- [192] Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. In *Flow and the foundations of positive psychology*, pages 21–34. Springer, 2014. **[Cited on page 71]**
- [193] Samantha Burgeson. Flexible seating influencing student engagement. 2017. **[Cited on page 71]**

- [194] Daniel R Montello. Classroom seating location and its effect on course achievement, participation, and attitudes. *Journal of Environmental Psychology*, 8(2):149–157, 1988. [Cited on pages 71 and 73]
- [195] Callie Allen. Flexible seating: Effects of student seating type choice in the classroom. Master’s thesis, Western Illinois University, January 2018. [Cited on page 71]
- [196] Xiaoming Yang, Xing Zhou, and Jie Hu. Students’ preferences for seating arrangements and their engagement in cooperative learning activities in college english blended learning classrooms in higher education. *Higher Education Research & Development*, 0(0):1–16, 2021. [Cited on page 71]
- [197] Moses Waithanji Ngware, James Ciera, Peter K Musyoka, Moses Oketch, et al. The influence of classroom seating position on student learning gains in primary schools in kenya. *Creative Education*, 4(11):705, 2013. [Cited on pages 71, 73, and 76]
- [198] Ka Long Chan, David C.W. Chin, Man Sing Wong, Roy Kam, Benedict Shing Bun Chan, Chun-Ho Liu, Frankie Kwan Kit Wong, Lorna K.P. Suen, Lin Yang, Simon Ching Lam, Wallace Wai lok Lai, and Xiaolin Zhu. Academic discipline as a moderating variable between seating location and academic performance: implications for teaching. *Higher Education Research & Development*, 0(0):1–15, 2021. [Cited on pages 71, 72, and 73]
- [199] Michail N. Giannakos, Kshitij Sharma, Sofia Papavlasopoulou, Ilias O. Pappas, and Vassilis Kostakos. Fitbit for learning: Towards capturing the learning experience using wearable sensing. *International Journal of Human-Computer Studies*, 136:102384, 2020. [Cited on pages 72 and 75]
- [200] Steven Kalinowski and Mark L Toper. The effect of seat location on exam grades and student perceptions in an introductory biology class. *Journal of College Science Teaching*, 36(4), 2007. [Cited on page 72]
- [201] Richard J Millard and David V Stimpson. Enjoyment and productivity as a function of classroom seating location. *Perceptual and Motor Skills*, 1980. [Cited on pages 73 and 82]

- [202] Zhe Dong, Haiyan Liu, and Xinqi Zheng. The influence of teacher-student proximity, teacher feedback, and near-seated peer groups on classroom engagement: An agent-based modeling approach. *PLoS ONE*, 16(1):e0244935, 2021. [Cited on page 73]
- [203] Mariola C Gremmen, Yvonne HM Van den Berg, Christian Steglich, René Veenstra, and Jan Kornelis Dijkstra. The importance of near-seated peers for elementary students' academic engagement and achievement. *Journal of Applied Developmental Psychology*, 57:42–52, 2018. [Cited on page 73]
- [204] Beth Hurst, Randall R Wallace, and Sarah B Nixon. The impact of social interaction on student learning. *Reading Horizons*, 2013. [Cited on page 73]
- [205] Daniel Szafir and Bilge Mutlu. ARTFul: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1001–1010, 2013. [Cited on page 74]
- [206] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398, 2012. [Cited on page 74]
- [207] Beverly Woolf, Winslow Bursleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009. [Cited on page 74]
- [208] Jonna Malmberg, Eetu Haataja, Tapio Seppänen, and Sanna Järvelä. Are we together or not? the temporal interplay of monitoring, physiological arousal and physiological synchrony during a collaborative exam. *International Journal of Computer-Supported Collaborative Learning*, 14(4):467–490, 2019. [Cited on pages 74, 85, 86, and 88]
- [209] Ivo Stuldreher. Multimodal physiological synchrony as measure of attentional engagement. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 718–722, 2020. [Cited on page 74]

- [210] Martha J. Sanders. Classroom design and student engagement. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1):496–500, 2013. **[Cited on page 75]**
- [211] Qifeng Lyu, Yunhong Jiang, and Junyong Wu. Relations between university students' academic achievement and their seating positions in classrooms. In *2021 7th International Conference on Education and Training Technologies*, pages 36–43, 2021. **[Cited on page 76]**
- [212] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034, 2013. **[Cited on page 77]**
- [213] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. **[Cited on pages 78 and 89]**
- [214] Chunhui Yuan and Haitao Yang. Research on k-value selection method of k-means clustering algorithm. *J—Multidisciplinary Scientific Journal*, 2(2):226–235, 2019. **[Cited on page 79]**
- [215] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011. **[Cited on pages 79 and 128]**
- [216] Naz Kaya and Brigitte Burgess. Territoriality: Seat preferences in different types of classroom arrangements. *Environment and Behavior*, 39(6):859–876, 2007. **[Cited on page 80]**
- [217] Robert R Weaver and Jiang Qi. Classroom organization and participation: College students' perceptions. *The Journal of Higher Education*, 76(5):570–601, 2005. **[Cited on page 80]**

- [218] John Clifford Gower. Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985. [Cited on page 82]
- [219] Jie Yu, Jaume Amores, Nicu Sebe, Petia Radeva, and Qi Tian. Distance learning for similarity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):451–462, 2008. [Cited on page 82]
- [220] Dan Mønster, Dorte Døjbak Håkonsson, Jacob Kjær Eskildsen, and Sebastian Wallo. Physiological evidence of interpersonal dynamics in a cooperative production task. *Physiology & behavior*, 156:24–34, 2016. [Cited on page 85]
- [221] Matthew A Napierala. What is the bonferroni correction? *Aaos Now*, pages 40–41, 2012. [Cited on page 86]
- [222] Hugo D Critchley, Jessica Eccles, and Sarah N Garfinkel. Interaction between cognition, emotion, and the autonomic nervous system. In *Handbook of clinical neurology*, volume 117, pages 59–77. Elsevier, 2013. [Cited on page 88]
- [223] Stephen H Fairclough, Louise Venables, and Andrew Tattersall. The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology*, 56(2):171–184, 2005. [Cited on page 88]
- [224] Tali Sharot and Elizabeth A Phelps. How arousal modulates memory: Disentangling the effects of attention and retention. *Cognitive, Affective, & Behavioral Neuroscience*, 4(3):294–306, 2004. [Cited on page 89]
- [225] Adrian Colomer Granero, Félix Fuentes-Hurtado, Valery Naranjo Ornedo, Jaime Guixeres Provinciale, Jose M Ausín, and Mariano Alcañiz Raya. A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents. *Frontiers in computational neuroscience*, 10:74, 2016. [Cited on page 89]

- [226] Ryan Cain and Victor R Lee. Measuring electrodermal activity to capture engagement in an afterschool maker program. In *Proceedings of the 6th Annual Conference on Creativity and Fabrication in Education*, pages 78–81, 2016. [Cited on page 89]
- [227] Deepali Virmani, Shweta Taneja, and Geetika Malhotra. Normalization based k means clustering algorithm. *arXiv preprint arXiv:1503.00900*, 2015. [Cited on page 89]
- [228] Rosa Lletí, M Cruz Ortiz, Luis A Sarabia, and M Sagrario Sánchez. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1):87–100, 2004. [Cited on page 89]
- [229] Robin L Plackett. Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale De Statistique*, pages 59–72, 1983. [Cited on page 90]
- [230] Tae Kyun Kim. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6):540, 2015. [Cited on page 91]
- [231] Luca Menghini, Evelyn Gianfranchi, Nicola Cellini, Elisabetta Patron, Mariaelena Tagliabue, and Michela Sarlo. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology*, 56(11):e13441, 2019. [Cited on page 93]
- [232] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, 2013. [Cited on page 96]
- [233] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. Predicting personality using novel mobile phone-based metrics. In *International Conference on Social Computing, Behavioral-cultural Modeling, and Prediction*, pages 48–55. Springer, 2013. [Cited on pages 96, 99, 104, 106, and 111]
- [234] Jiawei Bai, Chongzhi Di, Luo Xiao, Kelly R Evenson, Andrea Z LaCroix, Ciprian M Crainiceanu, and David M Buchner. An activity index for raw accelerometry data and its comparison with other activity metrics. *PloS One*, 11(8):e0160644, 2016. [Cited on pages 96 and 99]

- [235] Magdalena I Tolea, Antonio Terracciano, Eleanor M Simonsick, E Jeffrey Metter, Paul T Costa Jr, and Luigi Ferrucci. Associations between personality traits, physical activity level, and muscle strength. *Journal of research in personality*, 46(3):264–270, 2012. **[Cited on pages 96, 99, and 104]**
- [236] Brian P Bailey, Joseph A Konstan, and John V Carlis. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Interact*, volume 1, pages 593–601, 2001. **[Cited on page 96]**
- [237] Karen Renaud, Judith Ramsay, and Mario Hair. "you've got e-mail!"... shall i deal with it now? electronic mail from the recipient's perspective. *International Journal of Human-Computer Interaction*, 21(3):313–332, 2006. **[Cited on page 96]**
- [238] Brian P Bailey, Joseph A Konstan, and John V Carlis. Measuring the effects of interruptions on task performance in the user interface. In *SMC 2000 conference proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions'*, volume 2, pages 757–762. IEEE, 2000. **[Cited on page 96]**
- [239] Leslie A Perlow. The time famine: Toward a sociology of work time. *Administrative science quarterly*, 44(1):57–81, 1999. **[Cited on page 96]**
- [240] SeungJun Kim, Jaemin Chun, and Anind K Dey. Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 487–496, 2015. **[Cited on page 97]**
- [241] J Gregory Trafton, Erik M Altmann, Derek P Brock, and Farilee E Mintz. Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5):583–603, 2003. **[Cited on page 97]**
- [242] J. E. Fischer, N. Yee, V. Bellotti, N. Good, S. Benford, and C. Greenhalgh. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proceedings of the*

- 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '10, pages 103–112, New York, NY, USA, 2010. ACM. [Cited on page 97]
- [243] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3055–3064, Toronto, Ontario, Canada, 2014. ACM. [Cited on page 97]
- [244] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. How busy are you? predicting the interruptibility intensity of mobile users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5346–5360, Denver, Colorado, USA, 2017. ACM. [Cited on pages 97 and 102]
- [245] Anja Exler, Marcel Braith, Andrea Schankin, and Michael Beigl. Preliminary investigations about interruptibility of smartphone users at specific place types. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1590–1595, 2016. [Cited on page 97]
- [246] Joyce Ho and Stephen S. Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 909–918, New York, NY, USA, 2005. ACM. [Cited on pages 97 and 120]
- [247] Christoph Anderson, Clara Heissler, Sandra Ohly, and Klaus David. Assessment of social roles for interruption management: A new concept in the field of interruptibility. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 1530–1535, New York, NY, USA, 2016. Association for Computing Machinery. [Cited on page 97]
- [248] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. My phone and me: understanding people's receptivity to mobile notifications. In *Pro-*

- ceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, pages 1021–1032, New York, NY, USA, 2016. ACM. [Cited on page 97]
- [249] Amin Sadri, Flora D Salim, Yongli Ren, Wei Shao, John C Krumm, and Cecilia Mascolo. What will you do for the rest of the day? An approach to continuous trajectory prediction. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(4):1–26, 2018. [Cited on page 99]
- [250] Wei Shao, Thuong Nguyen, Kai Qin, Moustafa Youssef, and Flora D Salim. Bleedorguard: a device-free person identification framework using bluetooth signals for door access. *IEEE Internet of Things Journal*, 5(6):5227–5239, 2018. [Cited on page 99]
- [251] Jonathan Liono, Zahraa S Abdallah, A Kai Qin, and Flora D Salim. Inferring transportation mode and human activity from mobile sensing in daily life. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 342–351, 2018. [Cited on page 99]
- [252] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21(3):372–374, 2010. [Cited on page 99]
- [253] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013. [Cited on page 99]
- [254] Nhi NY Vo, Shaowu Liu, Xuezhong He, and Guandong Xu. Multimodal mixture density boosting network for personality mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 644–655. Springer, 2018. [Cited on page 99]
- [255] Xin Fang, Bing Li, and Wei-Hao Chang. Analyzing personality traits based on user behaviours of using microblog. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 773–779. Springer, 2014. [Cited on page 99]

- [256] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 321–330. ACM, 2012. **[Cited on page 99]**
- [257] Laura Cabrera-Quiros, Ekin Gedik, and Hayley Hung. Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 238–242, 2016. **[Cited on page 99]**
- [258] Tadashi Okoshi, Kota Tsubouchi, and Hideyuki Tokuda. Real-world product deployment of adaptive push notification scheduling on smartphones. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2792–2800, Anchorage AK USA, July 2019. ACM. **[Cited on pages 100 and 120]**
- [259] Prasanta Saikia, Ming Cheung, James She, and Soochang Park. Effectiveness of mobile notification delivery. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 21–29, Daejeon, South Korea, May 2017. IEEE. **[Cited on pages 100 and 120]**
- [260] Tilo Westermann, Ina Wechsung, and Sebastian Möller. Smartphone notifications in context: a case study on receptivity by the example of an advertising service. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2355–2361, San Jose California USA, May 2016. ACM. **[Cited on pages 100 and 131]**
- [261] Pascal E. Fortin, Elisabeth Sulmont, and Jeremy Cooperstock. Detecting perception of smartphone notifications using skin conductance responses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–9, New York, NY, USA, 2019. Association for Computing Machinery. **[Cited on page 100]**

- [262] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. Prefminer: Mining user's preferences for intelligent mobile notification management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '16*, pages 1223–1234, New York, NY, USA, 2016. ACM. **[Cited on pages 101 and 120]**
- [263] Manuela Züger, Sebastian C. Müller, André N. Meyer, and Thomas Fritz. Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, pages 591:1–591:14, New York, NY, USA, 2018. ACM. **[Cited on page 101]**
- [264] Klaus R Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005. **[Cited on page 101]**
- [265] Cyril Couffe and George A Michael. Failures due to interruptions or distractions: A review and a new framework. *American Journal of Psychology*, 130(2):163–181, 2017. **[Cited on page 102]**
- [266] Fred R. H. Zijlstra, Robert A. Roe, Anna B. Leonora, and Irene Krediet. Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology*, 72(2):163–185, 1999. **[Cited on page 102]**
- [267] Yoshiro Miyata and Donald A. Norman. Psychological issues in support of multiple activities. In Donald A. Norman and S. W. Draper, editors, *User centered system design: New perspectives on human-computer interaction*, pages 265–284. Lawrence Erlbaum, Hillsdale, N.J. USA, 1986. **[Cited on page 102]**
- [268] Laura Dabbish, Gloria Mark, and Víctor M. González. Why do i keep interrupting myself? environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 3127–3130, New York, NY, USA, 2011. ACM. **[Cited on page 102]**
- [269] Aftab Khan, Alexandros Zenonos, Georgios Kalogridis, Yaowei Wang, Stefanos Vatsikas, and Mahesh Sooriyabandara. Perception clusters: Automated mood recognition using a

- novel cluster-driven modelling system. *ACM Transactions on Computing for Healthcare*, 2(1):1–16, January 2021. [Cited on page 102]
- [270] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011. [Cited on pages 103 and 104]
- [271] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2(1999):102–138, 1999. [Cited on page 103]
- [272] Timothy A Judge, Chad A Higgins, Carl J Thoresen, and Murray R Barrick. The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, 52(3):621–652, 1999. [Cited on page 103]
- [273] Henri Vähä-Ypyä, Tommi Vasankari, Pauliina Husu, Jaana Suni, and Harri Sievänen. A universal, accurate intensity-based classification of different physical activities using raw data of accelerometer. *Clinical Physiology and Functional Imaging*, 35(1):64–70, 2015. [Cited on page 104]
- [274] Minna Aittasalo, Henri Vähä-Ypyä, Tommi Vasankari, Pauliina Husu, Anne-Mari Jusila, and Harri Sievänen. Mean amplitude deviation calculated from raw acceleration data: a novel method for classifying the intensity of adolescents’ physical activity irrespective of accelerometer brand. *BMC Sports Science, Medicine and Rehabilitation*, 7(1):18, 2015. [Cited on page 104]
- [275] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212, 2007. [Cited on page 106]
- [276] Joel Hektner, Jennifer Schmidt, and Mihaly Csikszentmihalyi. *Experience sampling method*. SAGE Publications, Inc., Thousand Oaks, CA, USA, 2007. [Cited on page 114]

- [277] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. The experience sampling method on mobile devices. *ACM Comput. Surv.*, 50(6), December 2017. [Cited on page 114]
- [278] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio, and Vassilis Kostakos. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies*, 125:118 – 128, 2019. [Cited on page 114]
- [279] Veljko Pejovic and Mirco Musolesi. Interruptme: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14*, pages 897–908, New York, NY, USA, 2014. ACM. [Cited on page 115]
- [280] Aku Visuri, Niels van Berkel, Tadashi Okoshi, Jorge Goncalves, and Vassilis Kostakos. Understanding smartphone notifications’ user interactions and content importance. *International Journal of Human-Computer Studies*, 128:72–85, August 2019. [Cited on page 116]
- [281] Dominik Weber, Alexandra Voit, and Niels Henze. Clear all: A large-scale observational study on mobile notification drawers. In *Proceedings of Mensch und Computer 2019, MuC'19*, pages 361–372, Hamburg, Germany, September 2019. Association for Computing Machinery. [Cited on page 116]
- [282] Robert W Levenson. Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. *Social psychophysiology: Theory and clinical applications*, pages 17–42, 1988. [Cited on page 117]
- [283] Sylvia D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3):394–421, July 2010. [Cited on pages 117 and 133]
- [284] MM BRADLEY and PJ LANG. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994. [Cited on page 117]

- [285] Blake E. Ashforth, Glen E. Kreiner, and Mel Fugate. All in a day's work: Boundaries and micro role transitions. *Academy of Management Review*, 25(3):472–491, July 2000. [Cited on page 117]
- [286] Christena E. Nippert-Eng. *Home and work: Negotiating boundaries through everyday life*. University of Chicago Press, 60th Street, Chicago, IL, USA, July 1996. [Cited on page 117]
- [287] Fatih Kursat Ozenc and Shelly D. Farnham. Life "modes" in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 561–570, New York, NY, USA, 2011. Association for Computing Machinery. [Cited on page 117]
- [288] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 813–824, New York, NY, USA, 2015. ACM. [Cited on page 120]
- [289] Tadashi Okoshi, Julian Ramos, H. Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones. In *2015 IEEE International Conference on Pervasive Computing and Communications*, pages 96–104, St. Louis, MO, USA, 2015. IEEE. [Cited on page 120]
- [290] Tadashi Okoshi, Julian Ramos, H. Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. Reducing users' perceived mental effort due to interruptive notifications in multi-device mobile environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 475–486, New York, NY, USA, 2015. ACM. [Cited on page 120]
- [291] Tadashi Okoshi, Kota Tsubouchi, Masaya Taji, Takanori Ichikawa, and Hideyuki Tokuda. Attention and engagement-awareness in the wild: A large-scale study with adaptive No-

- tifications. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 100–110, Big Island, HI, USA, 2017. IEEE. [Cited on page 120]
- [292] Judith Heinisch, Christoph Anderson, and Klaus David. Angry or climbing stairs? Towards physiological emotion recognition in the wild. In *Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications Workshops*, Kyoto, Japan, March 2019. IEEE. [Cited on page 121]
- [293] Gavin C Cawley and Nicola LC Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17(10):1467–1475, 2004. [Cited on page 127]
- [294] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. [Cited on pages 127, 154, and 155]
- [295] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001. [Cited on page 127]
- [296] Marcia Cassitas Hino and Maria Alexandra Cunha. Study of individual differences in the behavior of mobile technology users in the context of urban mobility. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019. [Cited on page 129]
- [297] Xianjing Wang, Jonathan Liono, Will Mcintosh, and Flora D Salim. Predicting the city foot traffic with pedestrian sensor data. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 1–10. ACM, 2017. [Cited on page 136]
- [298] Wei Shao, Yu Zhang, Bin Guo, Kai Qin, Jeffrey Chan, and Flora D Salim. Parking availability prediction with long short term memory model. In *International Conference on Green, Pervasive, and Cloud Computing*, pages 124–137. Springer, 2018. [Cited on pages 136 and 144]

- [299] Hui Song, AK Qin, and Flora D Salim. Evolutionary model construction for electricity consumption prediction. *Neural Computing and Applications*, pages 1–18, 2019. [**Cited on page 136**]
- [300] Sicheng Zhan and Adrian Chong. Building occupancy and energy consumption: Case studies across building types. *Energy and Built Environment*, 2020. [**Cited on page 136**]
- [301] Waselul H Sadid, Saad A Abobakr, and Guchuan Zhu. Discrete-event systems-based power admission control of thermal appliances in smart buildings. *IEEE Transactions on Smart Grid*, 8(6):2665–2674, 2017. [**Cited on page 136**]
- [302] MM Rahman, MG Rasul, and Mohammad Masud Kamal Khan. Energy conservation measures in an institutional building in sub-tropical climate in australia. *Applied Energy*, 87(10):2994–3004, 2010. [**Cited on page 136**]
- [303] Monika Frontczak and Pawel Wargocki. Literature survey on how different factors influence human comfort in indoor environments. *Building and Environment*, 46(4):922–937, 2011. [**Cited on page 136**]
- [304] American Society of Heating, Refrigerating, Air-Conditioning Engineers, and American National Standards Institute. *Thermal environmental conditions for human occupancy*, volume 55. American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2004. [**Cited on pages 136 and 141**]
- [305] Poul O Fanger et al. Thermal comfort: Analysis and applications in environmental engineering. 1970. [**Cited on pages 136, 137, 138, and 148**]
- [306] R Becker and M Paciuk. Thermal comfort in residential buildings—failure to predict by standard model. *Building and Environment*, 44(5):948–960, 2009. [**Cited on pages 136 and 148**]

- [307] Frederik Aufferberg, Sebastian Stein, and Alex Rogers. A personalised thermal comfort model using a bayesian network. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. **[Cited on pages 136 and 147]**
- [308] Chungyoon Chun, Alison Kwok, Teruaki Mitamura, Norie Miwa, and Akihiro Tamura. Thermal diary: Connecting temperature history to indoor comfort. *Building and Environment*, 43(5):877–885, 2008. **[Cited on page 136]**
- [309] Tanaya Chaudhuri, Deqing Zhai, Yeng Chai Soh, Hua Li, and Lihua Xie. Random forest based thermal comfort prediction from gender-specific physiological parameters using wearable sensing technology. *Energy and Buildings*, 166:391–406, 2018. **[Cited on pages 136, 139, 141, 147, and 153]**
- [310] OA Seppänen, WJ Fisk, and MJ Mendell. Association of ventilation rates and CO2 concentrations with health and other responses in commercial and institutional buildings. *Indoor Air*, 9(4):226–252, 1999. **[Cited on pages 136 and 147]**
- [311] Madhavi Indraganti and Kavita Daryani Rao. Effect of age, gender, economic group and tenure on thermal comfort: a field study in residential buildings in hot and dry climate with seasonal variations. *Energy and Buildings*, 42(3):273–281, 2010. **[Cited on pages 136, 147, and 149]**
- [312] Krzysztof Cena and Richard J de Dear. Field study of occupant comfort and office thermal environments in a hot, arid climate. *ASHRAE Transactions*, 105:204, 1999. **[Cited on page 137]**
- [313] Richard De Dear and Gail Schiller Brager. Developing an adaptive model of thermal comfort and preference. 1998. **[Cited on pages 137 and 138]**
- [314] Marcel Schweiker, Amar Abdul-Zahra, Maíra André, Farah Al-Atrash, Hanan Al-Khatri, Rea Risky Alprianti, Hayder Alsaad, Rucha Amin, Eleni Ampatzi, Alpha Yacob Arsano, et al. The scales project: A cross-national dataset on the interpretation of thermal perception scales. *Scientific Data*, 6(1):1–10, 2019. **[Cited on pages 137 and 142]**

- [315] Jared Langevin, Patrick L Gurian, and Jin Wen. Tracking the human-building interaction: A longitudinal field study of occupant behavior in air-conditioned offices. *Journal of Environmental Psychology*, 42:94–115, 2015. **[Cited on pages 137 and 142]**
- [316] Leonidas Bourikas, Enrico Costanza, Stephanie Gauthier, PAB James, Jacob Kittley-Davies, Carmine Ornaghi, Alex Rogers, Elham Saadatian, and Yitong Huang. Camera-based window-opening estimation in a naturally ventilated office. *Building Research & Information*, 46(2):148–163, 2018. **[Cited on page 138]**
- [317] Adrian K Clear, Janine Morley, Mike Hazas, Adrian Friday, and Oliver Bates. Understanding adaptive thermal comfort: New directions for UbiComp. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 113–122. ACM, 2013. **[Cited on page 138]**
- [318] Maohui Luo, Jiaqing Xie, Yichen Yan, Zhihao Ke, Peiran Yu, Zi Wang, and Jingsi Zhang. Comparing machine learning algorithms in predicting thermal sensation with ashrae comfort database ii. *Energy and Buildings*, page 109776, 2020. **[Cited on pages 139, 141, 158, and 162]**
- [319] PM Ferreira, AE Ruano, S Silva, and EZE Conceicao. Neural networks based predictive control for thermal comfort and energy savings in public buildings. *Energy and Buildings*, 55:238–251, 2012. **[Cited on pages 139 and 141]**
- [320] Weizheng Hu, Yonggang Wen, Kyle Guan, Guangyu Jin, and King Jet Tseng. itcm: Toward learning-based thermal comfort modeling via pervasive sensing for smart buildings. *IEEE Internet of Things Journal*, 5(5):4164–4177, 2018. **[Cited on pages 139, 141, and 153]**
- [321] Wei Zhang, Weizheng Hu, and Yonggang Wen. Thermal comfort modeling for smart buildings: A fine-grained deep learning approach. *IEEE Internet of Things Journal*, 2018. **[Cited on pages 139, 141, and 142]**


- [322] Prashanth Gurunath Shivakumar and Panayiotis Georgiou. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer Speech & Language*, 63:101077, 2020. [Cited on pages 140 and 151]
- [323] Jindong Wang, Yiqiang Chen, Lisha Hu, Xiaohui Peng, and S Yu Philip. Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2018. [Cited on page 140]
- [324] Juan Ye. Slearn: Shared learning human activity labels across multiple datasets. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–10. IEEE, 2018. [Cited on page 140]
- [325] Doris Hooi Chyee Toe and Tetsu Kubota. Development of an adaptive thermal comfort equation for naturally ventilated buildings in hot–humid climates using ASHRAE RP-884 database. *Frontiers of Architectural Research*, 2(3):278–291, 2013. [Cited on page 141]
- [326] Jörn von Grabe and Stefan Winter. The correlation between PMV and dissatisfaction on the basis of the ASHRAE and the mcintyre scale—towards an improved concept of dissatisfaction. *Indoor and Built Environment*, 17(2):103–121, 2008. [Cited on page 141]
- [327] Marcel Schweiker and Andreas Wagner. The effect of occupancy on perceived control, neutral temperature, and behavioral patterns. *Energy and Buildings*, 117:246–259, 2016. [Cited on page 142]
- [328] Mohsen Kaboli. A review of transfer learning algorithms. 2017. [Cited on page 146]
- [329] Andrew Arnold, Ramesh Nallapati, and William W Cohen. A comparative study of methods for transductive transfer learning. In *ICDM Workshops*, pages 77–82, 2007. [Cited on page 147]

- [330] Lili Zhang, Dong Wei, Yuyao Hou, Junfei Du, Zu'an Liu, Guomin Zhang, Long Shi, et al. Outdoor thermal comfort of urban park: A case study. *Sustainability*, 12(5):1961, 2020. [Cited on page 148]
- [331] Madhavi Indraganti, Ryoza Ooka, and Hom B Rijal. Thermal comfort in offices in india: behavioral adaptation and the effect of age and gender. *Energy and Buildings*, 103:284–295, 2015. [Cited on page 149]
- [332] Sami Karjalainen. Gender differences in thermal comfort and use of thermostats in everyday thermal environments. *Building and Environment*, 42(4):1594–1603, 2007. [Cited on page 149]
- [333] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421, 1972. [Cited on page 150]
- [334] Ivan Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man and Cybernetics*, 6:769–772, 1976. [Cited on page 150]
- [335] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. [Cited on page 150]
- [336] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008. [Cited on page 150]
- [337] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27:2672–2680, 2014. [Cited on page 150]
- [338] Nan Gao, Hao Xue, Wei Shao, Sichen Zhao, Kyle Kai Qin, Arian Prabowo, Mohammad Saiedur Rahaman, and Flora D Salim. Generative adversarial networks for spatio-

- temporal data: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–25, 2022. **[Cited on page 150]**
- [339] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018. **[Cited on page 150]**
- [340] Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*, 2018. **[Cited on pages 150 and 153]**
- [341] Matias Quintana and Clayton Miller. Towards class-balancing human comfort datasets with gans. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 391–392, 2019. **[Cited on page 150]**
- [342] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009. **[Cited on page 151]**
- [343] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006. **[Cited on page 151]**
- [344] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Self-taught clustering. In *Proceedings of the 25th International conference on Machine Learning*, pages 200–207. ACM, 2008. **[Cited on page 151]**
- [345] Mohamed Abouelenien, Mihai Burzo, Rada Mihalcea, Kristen Rusinek, and David Van Alstine. Detecting human thermal discomfort via physiological signals. In *Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments*, pages 146–149, 2017. **[Cited on page 154]**
- [346] Manuel Carlos Gameiro da Silva, JN Pires, A Loureiro, LD Pereira, P Neto, A Gaspar, DX Viegas, N Soares, M Oliveira, and J Costa. Spreadsheets for the calculation of thermal comfort indices PMV and PPD. *ResearchGate*, 2014. **[Cited on page 154]**

- [347] Keinosuke Fukunaga and Patrenahalli M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, 100(7):750–753, 1975. [Cited on pages 154 and 155]
- [348] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volume 3, pages 41–46, 2001. [Cited on page 154]
- [349] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999. [Cited on page 154]
- [350] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991. [Cited on page 154]
- [351] Xue-zhu ZHAO, Xiu WANG, and Xue-feng ZHU. Research of samples selection in eye detection based on Adaboost algorithm. *Computer Technology and Development*, 2, 2010. [Cited on page 154]
- [352] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5(Sep):1089–1105, 2004. [Cited on page 155]
- [353] Tay Lee Yong and Harimi Djamila. Exploring köppen-geiger climate classification of the ASHRAE RP-884 database. *International Journal of Recent Technology and Engineering*, 7(6):854–860, 2019. [Cited on page 156]
- [354] Roddy Cowie. Ethical issues in affective computing. *The Oxford Handbook of Affective Computing*, pages 334–348, 2015. [Cited on page 169]

Appendix A: Ethics Approval Documents

College Human Ethics Advisory Network (CHEAN)
College of Science, Engineering & Health (SEH)
NHMRC Code: EC00237

Notice of Approval

Date: **14 August 2019**

Project number: **SEHAPP 55-18**

Project title: **'Sensor and indoor comfort data collection at Cornish College'**

Risk classification: **Low Risk**

Investigator(s): **A/Prof Flora Salim, Nan Gao, Max Marschall, Prof Simon Watkins, Prof Jane Burry, David Tennent, Dr Abdulghani Mohamed, Awnili Shabnam, Dr Wei Shao, Dr Mohammad Saiedur Rahaman**

Approval period: From: **03/12/2019** To: **31/07/2019**

I am pleased to advise that your amendment/extension request has been granted ethics approval by the Science, Engineering and Health College Human Ethics Advisory Network (SEH CHEAN), as a sub-committee of the RMIT Human Research Ethics Committee (HREC). Ethics approval is extended until **31/07/2020**.

Terms of approval:

1. Responsibilities of investigator

It is the responsibility of the above investigator/s to ensure that all other investigators and staff on a project are aware of the terms of approval and to ensure that the project is conducted as approved by the CHEAN. Approval is only valid whilst the investigator/s holds a position at RMIT University.

2. Amendments

Approval must be sought from the CHEAN to amend any aspect of a project including approved documents. To apply for an amendment please use the 'Request for Amendment Form' that is available on the RMIT website. Amendments must not be implemented without first gaining approval from CHEAN.

3. Adverse events

You should notify HREC immediately of any serious or unexpected adverse effects on participants or unforeseen events affecting the ethical acceptability of the project.

4. Participant Information Sheet and Consent Form (PISCF)

The PISCF and any other material used to recruit and inform participants of the project must include the RMIT university logo. The PISCF must contain a complaints clause.

5. Annual reports

Continued approval of this project is dependent on the submission of an annual report. This form can be located online on the human research ethics web page on the RMIT website.

6. Final report

A final report must be provided at the conclusion of the project. CHEAN must be notified if the project is discontinued before the expected date of completion.

7. Monitoring

Projects may be subject to an audit or any other form of monitoring by HREC at any time.

8. Retention and storage of data





College Human Ethics Advisory Network (CHEAN)
College of Science, Engineering & Health (SEH)
NHMRC Code: EC00237

The investigator is responsible for the storage and retention of original data pertaining to a project for a minimum period of five years.

Please quote the project number and project title in any future correspondence.

On behalf of the SEH College Human Ethics Advisory Network, I wish you well in your research.

Yours sincerely

Associate Professor Barbara Polus
Chair, Science Engineering & Health
College Human Ethics Advisory Network





STEM College
College Human Ethics Advisory
Network (CHEAN)
Email: humanethics@rmit.edu.au
Tel: [61 3] 9925 4620

Notice of Approval

Date: **24 September 2020**

Project number: **23721**

Project title: **Social Roles and Interruptibility (original Project ID: 23632)**

Risk classification: **Negligible/Low**

Chief investigator: **Associate Professor Flora Salim**

Status: **Approved**

Approval period: From: **24/092020** To: **24/09/2021**

The following documents have been reviewed and approved:

Title	Version	Date
Risk Assessment and Application Form	4	9/09/2020
Participant Information Sheet and Consent Form	4	9/09/2020
Research Instruments and Protocol	4	9/09/2020
Risk Management Protocol	4	9/09/2020

The above application has been approved by the RMIT University CHEAN as it meets the requirements of the *National Statement on Ethical Conduct in Human Research* (NHMRC, 2007).

Terms of approval:

1. Responsibilities of chief investigator

It is the responsibility of the above chief investigator to ensure that all other investigators and staff on a project are aware of the terms of approval and to ensure that the project is conducted as approved by CHEAN. Approval is valid only whilst the chief investigator holds a position at RMIT University.

2. Amendments

Approval must be sought from CHEAN to amend any aspect of a project. To apply for an amendment, use the request for amendment form, which is available on the HREC website and submitted to the CHEAN secretary. Amendments must not be implemented without first gaining approval from CHEAN.



3. Adverse events

You should notify the CHEAN immediately (within 24 hours) of any serious or unanticipated adverse effects of their research on participants, and unforeseen events that might affect the ethical acceptability of the project.

4. Annual reports

Continued approval of this project is dependent on the submission of an annual report. Annual reports must be submitted by the anniversary of approval of the project for each full year of the project. If the project is of less than 12 months duration, then a final report only is required.

5. Final report

A final report must be provided within six months of the end of the project. CHEAN must be notified if the project is discontinued before the expected date of completion.

6. Monitoring

Projects may be subject to an audit or any other form of monitoring by the CHEAN at any time.

7. Retention and storage of data

The investigator is responsible for the storage and retention of original data according to the requirements of the *Australian Code for the Responsible Conduct of Research (R22)* and relevant RMIT policies.

8. Special conditions of approval

Nil.

In any future correspondence please quote the project number and project title above.

Yours faithfully,

Professor Falk Scholer
Chair, Science Engineering & Health
College Human Ethics Advisory Network

Cc Student Investigator/s: Ms Nan Gao
Other Investigator/s: Ms Shohreh Deldari, Mr Christoph Anderson

Appendix B: Credits

Portions of the materials used in this thesis have previously appeared or under consideration in the following scientific publications:

- **Gao, N.**, Marschall, M., Burry, J., Watkins, S., & Salim, F. D. (2022). Understanding Occupants' Behaviour, Engagement, Emotion, and Comfort Indoors with Heterogeneous Sensors and Wearables. *Scientific Data*, 9(1), 1-16. (**Impact Factor: 8.501, SJR: Q1**) - **Chapter 2**
- **Gao, N.**, Rahaman, M. S., Shao, W., & Salim, F. D. (2021). Investigating the Reliability of Self-report Survey in the Wild: The Quest for Ground Truth. 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* (pp. 237-242). (**Workshop at Ubicomp 2021**) - **Chapter 2**
- **Gao, N.**, Shao, W., Rahaman, M. S., & Salim, F. D. (2020). n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3), 1-26. (**Distinguished Paper Award**) - **Chapter 3**
- **Gao, N.**, Rahaman, M. S., Shao, W., Ji, K., & Salim, F. D. (2021). Individual and Group-wise Classroom Seating Experience: Effects on Student Engagement in Different Courses. (**Under Review in IMWUT, Major Revision**) - **Chapter 4**
- **Gao, N.**, Shao, W., & Salim, F. D. (2019). Predicting Personality Traits From Physical Activity Intensity. *IEEE Computer*, 52(7), 47-56. (**Impact Factor: 4.419, SJR: Q1**)

- Chapter 5

- Heinisch, J.S., **Gao, N.**, Anderson, C., DelDari, S., David, K., & Salim, F. D. (2022). Investigating the Effects of Mood & Usage Behaviour on Notification Response Time. **(Co-first Authors, To be Submitted to IMWUT) - Chapter 5**
- **Gao, N.**, Shao, W., Rahaman, M. S., Zhai, J., David, K., & Salim, F. D. (2021). Transfer Learning for Thermal Comfort Prediction in Multiple Cities. *Building and Environment*, 195, 107725. **(Impact Factor: 4.820, SJR: Q1) - Chapter 6**

The research is supported by the Australian Government through the Australian Research Council's Linkage Projects funding scheme (project LP150100246). It is also supported by the Tuition Fee Scholarship and Higher Degree Research Support Grant by the School of Computing Technologies, RMIT University.